Variable projection framework for the reduced-rank matrix approximation problem by weighted least-squares

Pascal Terray*†

May 7, 2025

Abstract

In this monograph, we review and develop variable projection Gauss-Newton, Levenberg-Marquardt and Newton methods for the Weighted Low-Rank Approximation (WLRA) problem, which has now an increasing number of applications in many scientific fields. Particular attention is drawn at the robustness, efficiency and scalability of these variable projection second-order algorithms such that they can be used also on larger datasets now commonly found in many practical problems for which only first-order algorithms based on sequential repetitions of local optimization (e.g., majorization, Expectation-Maximization or alternating least-squares methods) or variations of gradient descent (e.g., conjugate, proximal or stochastic gradient descent methods), or hybrid algorithms from these two classes of methods, were only feasible due to their lower cost and memory requirement per iteration.

In parallel with this review of variable projection algorithms, we develop new formulae for the Jacobian and Hessian matrices involved in these variable projection methods and demonstrate their very specific properties such as the uniform rank deficiency of the Jacobian matrix or the rank deficiency of the Hessian matrix at the (local) minimizers of the cost function associated with the WLRA problem. These systematic deficiencies must be taken into account in any practical implementations of the algorithms. These different properties and the very particular geometry of the WLRA problem have not been well appreciated in the past and have been the main obstacles in the development of robust variable projection second-order algorithms for solving the WLRA problem.

In addition, we demonstrate that the variable projection framework gives original insights on the solvability, the landscape and the non-smoothness of the WLRA problem. It also helps to describe the tight links between previously unrelated methods, which have been proposed to solve it. Specifically, we illustrate the closed links between the variable projection framework and Riemannian optimization on the Grassmann manifold for the WLRA problem. We expect that software's developers and practitioners in different fields such as computer vision, signal processing, recommender systems, machine learning, multivariate statistics and geophysical sciences will benefit from the results in this monograph in order to devise more robust and accurate algorithms to solve the WLRA problem.

*Email: pascal.terray@locean.ipsl.fr

†Affiliation: Laboratoire d'Océanographie et du Climat: Expérimentations et Approches Numériques, Institut Pierre-Simon Laplace, Sorbonne Université/CNRS/IRD/MNHN, Paris, France

Contents

1 Introduction

Let $\mathbf{X}$ be a $p \times n$ real matrix and $\mathbf{W}$ be a $p \times n$ nonnegative real (weight) matrix (e.g., $\mathbf{W}_{ij} \geq 0$) associated with $\mathbf{X}$. This monograph is about the Weighted Low-Rank Approximation (WLRA) problem:

$$\min_{\mathbf{Y} \in \mathbb{R}^{p \times n}_{\leq k}} \quad \varphi(\mathbf{Y}) = \frac{1}{2} \sum_{j=1}^{n} \sum_{i=1}^{p} \mathbf{W}_{ij}.(\mathbf{X}_{ij} - \mathbf{Y}_{ij})^2 = \frac{1}{2} \|\sqrt{\mathbf{W}} \odot (\mathbf{X} - \mathbf{Y})\|_F^2 \,, \qquad \text{(P0)}$$

where $\mathbb{R}^{p \times n}_{\leq k} = \big\{\mathbf{Y} \in \mathbb{R}^{p \times n} \text{ and } \mathit{rank}(\mathbf{Y}) \leq k\big\}$ and we assume that $k \leq \mathit{rank}(\mathbf{X}) \leq \mathit{min}(p, n)$, $\odot$ denotes the Hadamard product (e.g., element-wise product) of two $p \times n$ matrices and $\|\|_F$ is the Frobenius norm, i.e., the matrix norm induced by the standard inner product $\langle \mathbf{Y}, \mathbf{X} \rangle = \mathit{trace}(\mathbf{Y}^T \mathbf{X})$ on the Hilbert space of $p \times n$ real matrices. The factor $\frac{1}{2}$ in the definition of $\varphi(.)$ has no effect on the minimizers of $\varphi(.)$, it is introduced only for notational convenience. Without this factor, we would have got an annoying factor of 2 in many expressions of this monograph. Thus, a solution of the WLRA problem in its formulation (P0), if it exists, is a $p \times n$ real matrix $\hat{\mathbf{X}}$ with $\mathit{rank}(\hat{\mathbf{X}}) \leq k \leq \mathit{rank}(\mathbf{X})$. If $k = \mathit{rank}(\hat{\mathbf{X}}) = \mathit{rank}(\mathbf{X})$ and $\mathbf{W}$ is a binary matrix, e.g., $\mathbf{W}_{ij} \in \big\{0, 1\big\}$, the WLRA problem is simply the so-called low-rank matrix completion problem (e.g., the problem of recovering matrices of low-rank when a large fraction of its elements are missing), which has been extensively studied in the past decades [140][44]. In a slightly more general scenario, i.e., when $k = \mathit{rank}(\hat{\mathbf{X}}) < \mathit{rank}(\mathbf{X})$ and $\mathbf{W}$ is a binary matrix, a solution of the WLRA problem can be viewed as a robust generalisation of Principal Component Analysis (PCA) to incomplete, noisy or corrupted observations [92][91][35][138]. In an even more general scenario when $\mathbf{W}$ is a general nonnegative matrix, a solution of the WLRA problem is very useful for denoising and revealing low-dimensional structures in incomplete and noisy datasets [155][180]. Thus, in its general form, the WLRA problem can be considered as a robust generalization of (truncated) Singular Value Decomposition (SVD) analysis and extends significantly the usefulness and versatility of the classical low-rank approximation problem for many interesting applications arising from different fields including statistics [72][35][181], computer vision [15][27][28][81], machine learning for recommender systems [104], signal processing and system identification[125][127][182] and physical sciences [155][179][180][20], to name a few.

Using general weights in the cost function $\varphi(.)$ allows us to take into account different confidence or sampling levels among the entries of the elements in $\mathbf{X}$ beyond the simple case of missing values, which corresponds to binary weights. As the error estimates of data are often widely varying, this is often better suited for many problems [155]. Thus, weighted low-rank approximations of $\mathbf{X}$ can be used to deal with non-i.i.d. Gaussian noise in the data [180][125][27] and to design robust versions of many multivariate statistical methods, which hinge on the classical low-rank matrix approximation in the Frobenius norm and are heavily used in data sciences. If the weight matrix $\mathbf{W}$ takes carefully into account the sampling properties of the dataset $\mathbf{X}$, the resulting weighted low-rank approximation $\hat{\mathbf{X}}$ is then defined to emphasize the better-observed aspects of the data [155][180]. In other words, the nonnegative weights $\mathbf{W}_{ij}$ allow for a differential weighting of the accuracy of the measurements $\mathbf{X}_{ij}$ as well as for missing data if $\mathbf{W}_{ij} = 0$. In particular, for the extreme case of zero sample size, an entry of the data matrix $\mathbf{X}$ should play no role in fitting the low-rank model; this can be done by assigning zero weight to such element of $\mathbf{X}$.

Note, that we implicitly assume throughout the monograph that the weight matrix $\mathbf{W}$ is such that

$$\sum_{i=1}^{p} \mathbf{W}_{ij} > 0 \text{ for } j = 1, \cdots, n \text{ and } \sum_{j=1}^{n} \mathbf{W}_{ij} > 0 \text{ for } i = 1, \cdots, p \,.$$

Stated more simply, these last two conditions imply that there is at least one nonzero weight in each column and row of $\mathbf{W}$ as otherwise the WLRA problem is not well-posed and tractable. Furthermore, we will demonstrate later that it is sometimes useful and necessary to impose stronger conditions on $\mathbf{W}$ such that each column and row of $\mathbf{W}$ have at least k nonzero weights in order to

avoid overfitting and obtain a meaningful approximate solution of the WLRA problem. In addition, as for the matrix completion problem, the WLRA problem may suffer from non-identifiability issues and is ill-posed without any incoherence type of conditions on the data matrix $\mathbf{X}$ [30][187]. As an illustration, with a sparse matrix $\mathbf{X}$, the matrix $\mathbf{W} \odot \mathbf{X}$ is likely to be a zero matrix if the number of non-zero weights $\mathbf{W}_{ij}$ is very small, and, in this case, the WLRA problem owns the zero matrix as a trivial solution, which obviously has no interest and is far from being optimal. To prevent this pathological case to occur, we need to impose some incoherent conditions on $\mathbf{X}$ with respect to the set of sparse matrices and assume that the number of samples is large enough, see [30] or [187] for more formal definitions of these so-called low incoherence hypotheses, which provide reliable recoveries of the data matrix $\mathbf{X}$ in the context of robust PCA, the matrix completion or WLRA problems.

If all the elements of $\mathbf{W}$ are all equal to 1 (or more generally are all equal to a strictly positive real number), we have $\varphi(\mathbf{Y}) = \frac{1}{2}\|\mathbf{X} - \mathbf{Y}\|_F^2$ up to a scaling constant, and this problem is well known and easily solved as the SVD theory provides the best rank-k approximation $\hat{\mathbf{X}}$ of a given $p \times n$ real matrix $\mathbf{X}$ in terms of the Frobenius norm and also characterizes when this solution is unique or not (see Theorem 2.1 below and [71] or [8] for details). Thus, in the simple case when all the elements of $\mathbf{W}$ are equal, but different from zero, it follows that once the SVD of $\mathbf{X}$ is available, its best rank-k approximation $\hat{\mathbf{X}}$ is readily computed. Moreover, if we are only interested in some $\hat{\mathbf{X}}$ with $k \ll \mathit{min}(p, n)$, many less expensive alternatives than the computation of the complete SVD of $\mathbf{X}$ are available for computing $\hat{\mathbf{X}}$ [71][178], including very fast and accurate randomized algorithms [85][119][128]. Furthermore, under ideal conditions, i.e., $\mathbf{X}$ has no-missing values and the noise in all its elements can be modeled as zero-mean, independent and identically distributed (i.i.d.) Gaussian variables, the truncated SVD solution is the maximum likelihood solution and is, thus, the optimal one. However, this optimal property does not hold for non-i.i.d. Gaussian noise.

The more general case of uneven noisy observations (e.g., non-i.i.d. Gaussian noise) is in fact a particular instance of a WLRA problem in which we may assume that there is a ground truth low-rank matrix $\hat{\mathbf{X}}$, which we are trying to reconstruct and which is perturbed by non-i.i.d. Gaussian noise. Thus, implicit in the WLRA problem, is the statistical hypothesis that the input data consists of the observed (and also perturbed) data and weight matrices, $\mathbf{X}$ and $\mathbf{W}$, such that

$$\mathbf{X} = \mathbf{M} \odot (\hat{\mathbf{X}} + \mathbf{E}) \,, \tag{1.1}$$

where $\mathbf{M}$ is a boolean mask that indicates the observed elements of $\mathbf{X}$ (e.g., $\mathbf{M}_{ij} = 0$ if $\mathbf{W}_{ij} = 0$ and $\mathbf{M}_{ij} = 1$ otherwise), $\mathbf{E}$ is a noise matrix such that $\mathbf{E}_{ij} \sim \mathcal{N}(0, \alpha_{ij}^2)$ (e.g., $\mathbf{E}_{ij}$ is a Gaussian noise term) and $\mathbf{W}_{ij}$ is assumed to be modeled as a monotonically decreasing function of α_{ij}, the noise level for each of the observed elements of $\mathbf{X}$. See [155][180], [181] and [27] for examples, respectively, in the physical sciences, statistics and computer vision community on how such weight matrix $\mathbf{W}$ can be constructed in the case of non-i.i.d. Gaussian noise.

However, for a general choice of the weight matrix $\mathbf{W}$ and, even in the simple and very common case in which the weights are all 0 or 1 (e.g., the missing value or matrix completion problems [93][91][140]), the SVD of the masked observed matrix (e.g., set $\mathbf{X}_{ij} = 0$ if $\mathbf{W}_{ij} = 0$) may provide a useful and simple heuristic [135], but does not give the desired closest fit to $\mathbf{X}$ in weighted 2-norm (or semi-norm if some weights are equal to zero) and the minimum of $\varphi(.)$. When general nonnegative weights are introduced, the problem of finding $\mathbf{Y} \in \mathbb{R}^{p \times n}_{\leq k}$ so that $\varphi(\mathbf{Y})$ is minimized is a NonLinear Least-Squares (NLLS) optimization problem in a finite-dimensional Hilbert space.

However, in its general setting, the WLRA problem is not convex (but only bi-convex) because of the nonconvexity and discontinuity of the rank function [80], has no closed-form solution because of the low-rank requirement, is known to be NP-hard [62] and is, thus, not well understood [167]. Furthermore, for some matrices $\mathbf{X}$ and $\mathbf{W}$ and some integers k, the WLRA problem has no solution

at all [62] and, in other cases, the cost function $\varphi(.)$ may have several local minima [171], a situation which can not occur in the classical low-rank approximation problem [171][75]. This hardness of the WLRA problem can be partly alleviated and some algorithms with provable guarantees have been proposed in the machine learning literature by making very strong assumptions such as incoherence of the ground truth low-rank matrix $\hat{\mathbf{X}}$, randomly sampled missing (or observed) entries in $\mathbf{X}$ or that the weight matrix $\mathbf{W}$ is spectrally closed to the all ones matrix [30][101][94][16][114]. See the book by Vidal et al. [187] for a good introduction and discussion of these assumptions and the related algorithms in the case of binary weights (e.g., the matrix completion problem). However, for many applications these assumptions are unrealistic and violated, especially the assumption of randomly missing entries in the physical sciences, in which the statistical model (1.1) is a more realistic framework, but does not provide any proven guarantees of success or provable time bounds for current WLRA algorithms. Taking into account this challenging background, the main objective of this paper is to discuss various efficient (pseudo) second-order iterative techniques for minimizing $\varphi(.)$, which exploit explicitly the separable properties of this cost function [158][190][166][182][81], and to show how to adapt standard NLLS algorithms to the special structure and geometry of the WLRA problem in its separable formulation.

The structure of the monograph is the following. Section 2 describes the notation used in the paper and gives an overview of some important definitions and preliminary results on linear algebra, multilinear algebra and differentiation of vector and matrix functions relevant to the WLRA problem. In Section 3, we study the geometry of the WLRA problem, the existence of solutions for it and we review several of its alternative formulations, which have been used in the literature, demonstrate their equivalence, which has not always been well appreciated in past studies and, finally, show that the WLRA problem can be reformulated as a separable NLLS problem [63][166][87][182][81]. This result was first used by Ruhe [158] for solving WLRA problems with binary weights and $k = 1$ despite this separable formulation of the WLRA problem is often erroneously attributed to Wiberg [190] in the computer vision literature [176][147][150][81]. In fact, Wiberg [190] (who was a student of A. Ruhe) has extended the results of Ruhe [158] for an arbitrary integer k and a slightly different component model specifically designed to the problem of estimating a principal components model when missing values are present in the data; see also [147][187] for more details on this slightly different factor model used in [190]. Again in the computer vision literature, the separable NLLS algorithm originally proposed by Ruhe [158] and Wiberg [190] has been confused with the simplest Alternating Least-Squares (ALS) method [176][15][187] as first noted by Okatani et al. [147]. As a preamble to the variable projection algorithms, Section 4 gives a modern description of the block variant of this ALS method and its recent extensions.This ALS algorithm was perhaps the oldest and simplest method used to solve the WLRA problem in the statistical literature [191][192][93][72] and can be interpreted as a particular instance of the cyclic block-coordinate descent method for the WLRA problem. The ancestor of this block ALS algorithm is the Nonlinear Iterative PArtial Least Squares (NIPALS) method devised originally by Wold and his collaborators for the missing value problem in PCA, i.e., in the case where the weight matrix is binary [191][192][93][91]. Generalizations of the NIPALS algorithm to arbitrarily weighted least-squares have been first discussed in Gabriel and Zamir [72] and is now the topic of many recent papers in different fields [171][14][167][23][181][25][50]. However, nearly all the proposed algorithms dealing with general positive weights are first-order methods, excepted for the optimization approaches on the Grassmann manifold (e.g., the submanifold of fixed-rank matrices embedded in $\mathbb{R}^{p\times n}$) detailed in [125][27][14]. Section 5 is devoted to a detailed study of variable projection NLLS methods for solving the general WLRA problem, which use explicitly the separable property of this WLRA problem [63][158]. Variable projection methods originate from numerical analysis and are efficient methods for solving separable NLLS problems in which some variables of the problem occur linearly and other nonlinearly; see Subsection 2.4 for a more formal definition. Explicit formulations of the gradient vector, Jacobian and Hessian matrices used in these variable projection second-order methods are given and their very specific mathematical properties are also derived in this Section 5. Separable NLLS algorithms have a long history in applied mathematics

and excellent reviews are offered in [166][65][87]. The closed relationships between the variable projection NLLS method and Riemannian optimization on the Grassmann manifold in the context of the WLRA problem are also explored in this Section 5, extending and clarifying the results of Hong and Fitzgibbon [81][82] who have focused on the binary weights case. Templates and implementation aspects of these variable projection second-order algorithms are detailed in Section 6. Finally, a summary of our contribution and perspectives for further advancing our understanding of the WLRA problem and methods for solving it are given in Section 7.

2 Definitions and preliminaries

We first collect in this section some basic notations, definitions and results concerning linear algebra, multilinear algebra, differentiation of vector and matrices and nonlinear optimization problems, which will be used frequently in the following sections.

Throughout this monograph, we have tried to adhere to the following conventions: bold capital letters will denote matrices and bold lower-case letters will indicate vectors. A lower-case letter in italic, but not in boldface, will indicate a scalar. The symbols $\mathbb{R}^p$ and $\mathbb{R}^{p\times n}$ denote, respectively, the linear spaces of the real p-vectors and of the real $p \times n$ matrices. In some occasions, the sizes of the vectors or the shapes of the matrices will be given as an upperscript. As an illustration, for $a \in \mathbb{R}$, the symbols $\mathbf{a}^p$ and $\mathbf{a}^{p\times n}$ represent, respectively, the p-vector and the $p \times n$ matrix composed of all a. For $\mathbf{u} \in \mathbb{R}^p$, the symbol $\mathit{diag}(\mathbf{u})$ is used to represent a diagonal $p \times p$ matrix with diagonal elements, $[\mathit{diag}(\mathbf{u})]_{ii} = \mathbf{u}_i$ for $i = 1, \cdots, p$. For any $\mathbf{C}$ matrix, the symbol $\mathbf{C}_{.j}$ is used to represent the j^{th} column vector of $\mathbf{C}$ and the symbol $\mathbf{C}_{i.}$ is used to represent the i^{th} row vector of $\mathbf{C}$. The symbol $\mathbf{I}_p$ is used to denote the identity matrix of order p.

2.1 Linear algebra

For a matrix $\mathbf{C} \in \mathbb{R}^{p\times n}$, we denote the transpose, the range and the null space of $\mathbf{C}$ by $\mathbf{C}^T$, $\mathit{ran}(\mathbf{C})$ and $\mathit{null}(\mathbf{C})$, respectively:

$$\mathbf{C}^T_{ij} = \mathbf{C}_{ji}, \mathit{ran}(\mathbf{C}) = \big\{\mathbf{y} \in \mathbb{R}^p \;/\; \exists \mathbf{x} \in \mathbb{R}^n \text{ with } \mathbf{y} = \mathbf{C}\mathbf{x}\big\}, \mathit{null}(\mathbf{C}) = \big\{\mathbf{x} \in \mathbb{R}^n \;/\; \mathbf{C}\mathbf{x} = \mathbf{0}^p\big\}.$$

$\mathit{ran}(\mathbf{C})$ and $\mathit{null}(\mathbf{C})$ are vector subspaces of $\mathbb{R}^p$ and $\mathbb{R}^n$, respectively. The rank of a matrix $\mathbf{C}$ is then defined by the dimension of the vector space $\mathit{ran}(\mathbf{C})$, i.e., $\mathit{rank}(\mathbf{C}) = \mathit{dim}(\mathit{ran}(\mathbf{C}))$. Equivalently, the rank of a $p \times n$ matrix $\mathbf{C}$ can be defined as the smallest integer $k = \mathit{rank}(\mathbf{C})$ such that it exists $\mathbf{A} \in \mathbb{R}^{p\times k}$ and $\mathbf{B} \in \mathbb{R}^{k\times n}$ such that $\mathbf{C} = \mathbf{A}\mathbf{B}$. From this definition, it is not difficult to show that $\mathit{rank}(\mathbf{C}^T) = \mathit{rank}(\mathbf{C})$. Then, it can been shown that

$$\mathit{dim}\big(\mathit{null}(\mathbf{C})\big) + \mathit{dim}\big(\mathit{ran}(\mathbf{C})\big) = \mathit{dim}\big(\mathit{null}(\mathbf{C})\big) + \mathit{rank}(\mathbf{C}) = n\ , \tag{2.1}$$

which is known as the rank-nullity theorem or relationship, and also that

$$\mathit{rank}(\mathbf{A}\mathbf{B}) \leq min\big(\mathit{rank}(\mathbf{A}), \mathit{rank}(\mathbf{B})\big)\ , \tag{2.2}$$

if the number of columns of $\mathbf{A}$ is equal to the number of rows of $\mathbf{B}$, and, finally, that

$$\mathit{rank}(\mathbf{A} + \mathbf{B}) \leq \mathit{rank}(\mathbf{A}) + \mathit{rank}(\mathbf{B})\ , \tag{2.3}$$

when $\mathbf{A}$ and $\mathbf{B}$ are matrices of the same dimensions. We further assume the following equalities

$$\mathit{null}(\mathbf{C}^T) = \mathit{ran}(\mathbf{C})^\perp \text{ and } \mathit{ran}(\mathbf{C}^T) = \mathit{null}(\mathbf{C})^\perp\ , \tag{2.4}$$

where $\mathit{ran}(\mathbf{C})^\perp$ and $\mathit{null}(\mathbf{C})^\perp$ denote, respectively, the orthogonal complements of the range and null spaces of $\mathbf{C}$ with respect to the standard Euclidean inner products in $\mathbb{R}^p$ and $\mathbb{R}^n$, respectively.

We will use mostly the Euclidean norm for vectors and the Frobenius norm for matrices, i.e.,

$$\|\mathbf{u}\|_2 = \Big(\sum_{i=1}^{p} \mathbf{u}_i^2\Big)^{\frac{1}{2}} \text{ for } \mathbf{u} \in \mathbb{R}^p \text{ and } \|\mathbf{C}\|_F = \Big(\sum_{i=1}^{p}\sum_{j=1}^{n} \mathbf{C}_{ij}^2\Big)^{\frac{1}{2}} \text{ for } \mathbf{C} \in \mathbb{R}^{p\times n}\,, \tag{2.5}$$

which are, respectively, associated to the Euclidean inner product in $\mathbb{R}^p$

$$\langle \mathbf{u}, \mathbf{v}\rangle_2 = \sum_{i=1}^{p} \mathbf{u}_i \mathbf{v}_i\,, \tag{2.6}$$

and to the Frobenius inner product in $\mathbb{R}^{p\times n}$, defined for matrices $\mathbf{U}$ and $\mathbf{V}$ of identical sizes, by

$$\langle \mathbf{U}, \mathbf{V}\rangle_F = \mathrm{Tr}\left(\mathbf{U}^T\mathbf{V}\right) = \sum_{i=1}^{p}\sum_{j=1}^{n} \mathbf{U}_{ij}\mathbf{V}_{ij}\,, \tag{2.7}$$

where for squared matrices $\mathrm{Tr}\left(\mathbf{W}\right) = \sum_{i=1}^{p} \mathbf{W}_{ii}$. When we do not specify it, we implicitly mean these standard norms and inner products for vectors and matrices. Occasionally, especially in Section 3, we will also use the spectral norm for matrices, which is the natural norm on the set of $p \times n$ real matrices induced by the Euclidean norm for vectors. For $\mathbf{C} \in \mathbb{R}^{p\times n}$, its spectral norm $\|\mathbf{C}\|_S$ can be computed as the squared root of the greatest eigenvalue of the matrix product $\mathbf{C}^T\mathbf{C}$ [71], i.e.,

$$\|\mathbf{C}\|_S = \max_{\mathbf{x}\in\mathbb{R}^n \text{ and } \mathbf{x}\neq\mathbf{0}^n} \frac{\|\mathbf{C}\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = (\text{maximum eigenvalue of } \mathbf{C}^T\mathbf{C})^{\frac{1}{2}}\,. \tag{2.8}$$

A matrix $\mathbf{Q} \in \mathbb{R}^{p\times p}$ is said to be orthogonal if $\mathbf{Q}\mathbf{Q}^T = \mathbf{Q}^T\mathbf{Q} = \mathbf{I}_p$. It is easily verified that the product of two orthogonal matrices is also an orthogonal matrix. A matrix norm $|||$ on $\mathbb{R}^{p\times n}$ is called unitarily invariant if $\|\mathbf{C}\| = \|\mathbf{Q}\mathbf{C}\mathbf{P}\|$ for all orthogonal matrices $\mathbf{Q}$ and $\mathbf{P}$ of order p and n, respectively, and the Frobenius and spectral norms are unitarily invariant.

If $\mathbf{C} \in \mathbb{R}^{p\times n}$, then $\mathbf{C}^+ \in \mathbb{R}^{n\times p}$ denotes the Moore-Penrose inverse (or pseudo-inverse) of $\mathbf{C}$ and is defined as the unique matrix which verifies the equalities

$$\mathbf{C}\mathbf{C}^+\mathbf{C} = \mathbf{C}, \mathbf{C}^+\mathbf{C}\mathbf{C}^+ = \mathbf{C}^+, (\mathbf{C}\mathbf{C}^+)^T = \mathbf{C}\mathbf{C}^+ \text{ and } (\mathbf{C}^+\mathbf{C})^T = \mathbf{C}^+\mathbf{C}\,. \tag{2.9}$$

If $\mathbf{C}$ is of full column rank, it is easy to verify that

$$\mathbf{C}^+ = (\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T\,.$$

In addition, every matrix $\mathbf{C}^- \in \mathbb{R}^{n\times p}$ satisfying only the two equalities

$$\mathbf{C}\mathbf{C}^-\mathbf{C} = \mathbf{C} \text{ and } (\mathbf{C}\mathbf{C}^-)^T = \mathbf{C}\mathbf{C}^- \tag{2.10}$$

is called a symmetric generalized inverse of $\mathbf{C}$.

An explicit formulation of the Moore-Penrose inverse $\mathbf{C}^+$ may be obtained with the help of the Singular Value Decomposition (SVD) of the matrix $\mathbf{C}$

$$\mathbf{C} = \mathbf{U}\Sigma\mathbf{V}^T\,, \tag{2.11}$$

where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices of order p and n, respectively, and

$$\Sigma = \begin{pmatrix} \sigma_1 & 0 & \dots & 0 & 0 \\ 0 & \sigma_2 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & \sigma_{n-1} & 0 \\ 0 & 0 & \dots & 0 & \sigma_n \\ \vdots & \vdots & \ddots & \vdots & \ddots \\ 0 & 0 & \dots & 0 & 0 \end{pmatrix}\,,$$

where we have assumed for notational convenience that $p \geq n$. The existence of the SVD can be proved using the spectral theorem for symmetric matrices [71][8]. $\mathbf{U}$ and $\mathbf{V}$ consist of the orthonormal eigenvectors of $\mathbf{C}\mathbf{C}^T$ and of $\mathbf{C}^T\mathbf{C}$, respectively. $\mathbf{U}$ and $\mathbf{V}$ are called, respectively, the left and right singular vectors of $\mathbf{C}$. The diagonal elements of Σ are called the singular values $\mathbf{C}$ and will always be taken to be nonnegative and ordered such that

$$\sigma_1 \geq \sigma_2 \geq \cdots \sigma_{min(p,n)} \geq 0 \, .$$

These singular values are the non-negative square roots of the eigenvalues of $\mathbf{C}^T\mathbf{C}$ or $\mathbf{C}\mathbf{C}^T$. Then, in exact arithmetic, if $\mathit{rank}(\mathbf{C}) = k < n$, we have $\sigma_{k+1} = \sigma_{k+2} = \cdots = \sigma_n = 0$ and it is easy to verify that

$$\mathbf{C}^+ = \mathbf{V}\Lambda\mathbf{U}^T \, , \tag{2.12}$$

where Λ is the $n \times p$ diagonal matrix with $\Lambda_{ii} = \sigma_i^{-1}$ for $i = 1, \cdots, k$ and $\Lambda_{ii} = 0$ for $i = k+1, \cdots, n$. We also assume the following important property of the Moore-Penrose inverse $\mathbf{C}^+$ for all matrices $\mathbf{C}$:

$$\mathit{null}(\mathbf{C}^+) = \mathit{null}(\mathbf{C}^T) \, .$$

A matrix $\mathbf{P} \in \mathbb{R}^{p \times p}$ is an orthogonal projector if the following two conditions are satisfied:

$$\mathbf{P}\mathbf{P} = \mathbf{P} \text{ and } \mathbf{P}^T = \mathbf{P} \tag{2.13}$$

and, given an orthogonal projector $\mathbf{P} \in \mathbb{R}^{p \times p}$, its associated complementary projector is defined as $\mathbf{P}^\perp = \mathbf{I}_p - \mathbf{P}$ and is also an orthogonal projector. As for any matrix, an (orthogonal) projector $\mathbf{P}$ maps vectors into its range $\mathit{ran}(\mathbf{P})$. However, an interesting and special property of any matrix $\mathbf{P}$ verifying $\mathbf{P}\mathbf{P} = \mathbf{P}$ is that it maps vectors of its range $\mathit{ran}(\mathbf{P})$ to themselves. In addition, given an orthogonal projector $\mathbf{P} \in \mathbb{R}^{p \times p}$ and a vector $\mathbf{x} \in \mathbb{R}^p$, the vector $\mathbf{P}\mathbf{x} \in \mathbb{R}^p$ uniquely solves the linear least-squares optimization problem

$$\mathbf{P}\mathbf{x} = \text{Arg} \min_{\mathbf{z} \in \mathit{ran}(\mathbf{P})} \|\mathbf{x} - \mathbf{z}\|_2 \, . \tag{2.14}$$

In words, $\mathbf{P}\mathbf{x}$ is the unique closest point to $\mathbf{x}$ in $\mathit{ran}(\mathbf{P})$. Note that $\mathit{ran}(\mathbf{P}^\perp) = \mathit{ran}(\mathbf{P})^\perp$, i.e., the range of $\mathbf{P}^\perp$ is the orthogonal complement of the range of $\mathbf{P}$. Given a linear subspace V of $\mathbb{R}^p$, we can decompose uniquely any vector $\mathbf{x} \in \mathbb{R}^p$ into the sum of one vector in V and one vector in $V^\perp$. This is easily verified as, given the (unique) orthogonal projector $\mathbf{P}$ onto V, we have immediately for any $\mathbf{x} \in \mathbb{R}^p$,

$$\mathbf{x} = \mathbf{P}\mathbf{x} + (\mathbf{I}_p - \mathbf{P})\mathbf{x} = \mathbf{P}\mathbf{x} + \mathbf{P}^\perp\mathbf{x} \, ,$$

where $\mathbf{P}\mathbf{x} \in V$ and $\mathbf{P}^\perp\mathbf{x} \in V^\perp$. In such a case, we say that $\mathbb{R}^p$ is the direct sum of V and $V^\perp$ and we write $\mathbb{R}^p = V \oplus V^\perp$. Finally, if the columns of $\mathbf{W} \in \mathbb{R}^{p \times k}$ form an orthonormal basis of V, it is not difficult to verify that

$$\mathbf{P} = \mathbf{W}\mathbf{W}^T \text{ and } \mathbf{P}^\perp = \mathbf{I}_p - \mathbf{W}\mathbf{W}^T \, .$$

Thus, provided that we have an orthonormal basis of V, we can also immediately project onto $V^\perp$ without constructing a basis for it. Furthermore, if we have such an orthonormal basis of V, we note that we have also a quick and efficient way of applying orthogonal projectors to vectors as

$$\mathbf{P}\mathbf{x} = \mathbf{W}(\mathbf{W}^T\mathbf{x}) \text{ and } \mathbf{P}^\perp\mathbf{x} = \mathbf{x} - \mathbf{W}(\mathbf{W}^T\mathbf{x}) \, .$$

The Moore-Penrose inverse and the SVD are also particularly useful to define and compute orthogonal projectors associated with the range of a matrix, especially if this matrix is rank deficient [71][8]. If the rank of the matrix $\mathbf{C} \in \mathbb{R}^{p \times n}$ is equal to k (and looking at the distribution of the singular values of $\mathbf{C}$ is the best way to determine its numerical rank), the matrix

$$\mathbf{P}_\mathbf{C} = \mathbf{C}\mathbf{C}^+ = \mathbf{U} \begin{bmatrix} \mathbf{I}_k & \mathbf{0}^{k \times (p-k)} \\ \mathbf{0}^{(p-k) \times k} & \mathbf{0}^{(p-k) \times (p-k)} \end{bmatrix} \mathbf{U}^T \, ,$$

where $\mathbf{U}$ are the left singular vectors of $\mathbf{C}$, is the orthogonal projector onto *ran*$(\mathbf{C})$. Furthermore, the matrix

$$\mathbf{P}_{\mathbf{C}}^{\perp} = \mathbf{I}_p - \mathbf{C}\mathbf{C}^{+} = \mathbf{U} \begin{bmatrix} \mathbf{0}^{k \times k} & \mathbf{0}^{k \times (p-k)} \\ \mathbf{0}^{(p-k) \times k} & \mathbf{I}_{p-k} \end{bmatrix} \mathbf{U}^T$$

is the orthogonal projector onto the orthogonal complement of *ran*$(\mathbf{C})$ (e.g., *ran*$(\mathbf{C})^{\perp}$). It is easy to show that if $\mathbf{x} \in \mathbb{R}^p$ then $\mathbf{P}_{\mathbf{C}}\mathbf{x} \in$ *ran*$(\mathbf{C})$ and $\mathbf{P}_{\mathbf{C}}^{\perp}\mathbf{x} \in$ *ran*$(\mathbf{C})^{\perp}$. In the same conditions, the matrix

$$\mathbf{P}_{\mathbf{C}^T} = \mathbf{C}^{+}\mathbf{C} = \mathbf{V} \begin{bmatrix} \mathbf{I}_k & \mathbf{0}^{k \times (n-k)} \\ \mathbf{0}^{(n-k) \times k} & \mathbf{0}^{(n-k) \times (n-k)} \end{bmatrix} \mathbf{V}^T$$

is the orthogonal projector onto the row space of $\mathbf{C}$, e.g., *ran*$(\mathbf{C}^T) =$ *null*$(\mathbf{C})^{\perp}$ and the matrix

$$\mathbf{P}_{\mathbf{C}^T}^{\perp} = \mathbf{I}_n - \mathbf{C}^{+}\mathbf{C} = \mathbf{V} \begin{bmatrix} \mathbf{0}^{k \times k} & \mathbf{0}^{k \times (n-k)} \\ \mathbf{0}^{(n-k) \times k} & \mathbf{I}_{n-k} \end{bmatrix} \mathbf{V}^T$$

is the orthogonal projector onto *ran*$(\mathbf{C}^T)^{\perp} =$ *null*$(\mathbf{C})$. If *rank*$(\mathbf{C}) = n \leq p$, then $\mathbf{P}_{\mathbf{C}^T} = \mathbf{I}_n$ and $\mathbf{P}_{\mathbf{C}^T}^{\perp}$ is the $n \times n$ zero matrix $\mathbf{0}^{n \times n}$.

The Moore-Penrose inverse and SVD of a matrix are particularly useful for solving (rank-deficient) linear least-squares problems [111][8]. For $\mathbf{C} \in \mathbb{R}^{p \times n}$ and $\mathbf{y} \in \mathbb{R}^p$, consider the linear least-squares problem

$$\min_{\mathbf{x} \in \mathbb{R}^n} \|\mathbf{y} - \mathbf{C}\mathbf{x}\|_2 = \|\mathbf{r}(\mathbf{x})\|_2 \text{ with } \mathbf{r}(\mathbf{x}) = \mathbf{y} - \mathbf{C}\mathbf{x} \, .$$

The unique vector $\widehat{\mathbf{x}}$ of minimum Euclidean norm minimizing $\|\mathbf{r}(\mathbf{x})\|_2$ is given by $\mathbf{C}^{+}\mathbf{y}$ as $\mathbf{P}_{\mathbf{C}}\mathbf{y} = \mathbf{C}\mathbf{C}^{+}\mathbf{y} = \mathbf{C}\widehat{\mathbf{x}}$ is the unique closest point to $\mathbf{y}$ in the range of $\mathbf{C}$. Even if $\mathbf{C}$ is of full column rank, in which case $\widehat{\mathbf{x}} = (\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T\mathbf{y}$, we use the pseudo-inverse notation $\mathbf{C}^{+}$ for $(\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T$ to indicate that the normal equations shall not be used to compute the solution of linear least-squares problems, especially if $\mathbf{C}$ is badly conditioned [111][71][8].

Note that linear least-squares problems can also be solved and orthogonal projectors be evaluated with the help of symmetric generalized inverses defined above [64][87]. The advantage is that these symmetric generalized inverses can be computed much more cheaply than the SVD or the pseudo-inverse of $\mathbf{C}$ with the help of other matrix decompositions such as the standard QR decomposition with Column Pivoting (QRCP) [111][64][71][87][8]. According to this decomposition, there exist a $p \times p$ orthogonal matrix $\mathbf{Q}$ and a $n \times n$ permutation matrix $\mathbf{P}$ such that, for a given $p \times n$ matrix $\mathbf{C}$ of rank k,

$$\mathbf{Q}\mathbf{C}\mathbf{P} = \begin{bmatrix} \mathbf{R} & \mathbf{S} \\ \mathbf{0}^{(p-k) \times k} & \mathbf{0}^{(p-k) \times (n-k)} \end{bmatrix} \, , \tag{2.15}$$

where $\mathbf{R}$ is a $k \times k$ nonsingular upper triangular matrix (with diagonal elements of decreasing absolute magnitude) and $\mathbf{S}$ an $k \times (n-k)$ full matrix, which is vacuous if $k = n$. Several procedures are available to compute this QRCP, but the usual one is based on Householder transformations (e.g., elementary orthogonal reflectors), which are orthogonal matrices of the form

$$\mathbf{H}(i) = \mathbf{I}_p - 2\mathbf{v}(i)\mathbf{v}(i)^T \, , \tag{2.16}$$

where the p-vector $\mathbf{v}(i)$ has a 2-norm equal to one [111][71]. Premultiplication by $\mathbf{H}(i)$ is frequently used to zero out a sequence of entries in a given column p-vector. Thus, in order to compute the QR or QRCP decomposition, $\mathbf{C}$ is successively pre-multiplied by at most $\min(n, p)$ Householder transformations $\mathbf{H}(i)$, permuting the columns of $\mathbf{C}$ if necessary (thus determining the permutation matrix $\mathbf{P}$). Also, the orthogonal matrix $\mathbf{Q}$ can be compactly stored (as only the vectors $\mathbf{v}(i)$ need to be stored) and explicitly computed as a product of k elementary reflectors

$$\mathbf{Q} = \mathbf{H}(k) \cdots \mathbf{H}(2)\mathbf{H}(1) \, .$$

For more details concerning Householder transformations, see [111][71][8].

Furthermore, the rank k of $\mathbf{C}$ can be efficiently estimated in an additional step from the upper triangular factor $\mathbf{R}$ computed during the QRCP, but we omit the details here [111][71][87][8]. Note that the QRCP is not unique as the permutation matrix $\mathbf{P}$ is not unique. However, with the help of a QRCP of $\mathbf{C}$, the orthogonal projectors $\mathbf{P}_{\mathbf{C}}$ and $\mathbf{P}_{\mathbf{C}}^{\perp}$ can be efficiently computed as

$$\mathbf{P}_{\mathbf{C}} = \mathbf{C}\mathbf{C}^{-} = \mathbf{Q}^T \begin{bmatrix} \mathbf{I}_k & \mathbf{0}^{k\times(p-k)} \\ \mathbf{0}^{(p-k)\times k} & \mathbf{0}^{(p-k)\times(p-k)} \end{bmatrix} \mathbf{Q} \tag{2.17}$$

and

$$\mathbf{P}_{\mathbf{C}}^{\perp} = \mathbf{I}_p - \mathbf{C}\mathbf{C}^{-} = \mathbf{Q}^T \begin{bmatrix} \mathbf{0}^{k\times k} & \mathbf{0}^{k\times(p-k)} \\ \mathbf{0}^{(p-k)\times k} & \mathbf{I}_{p-k} \end{bmatrix} \mathbf{Q} \,. \tag{2.18}$$

Furthermore, a symmetric generalized inverse of $\mathbf{C}$ defined by the equations (2.10) can be represented as

$$\mathbf{C}^{-} = \mathbf{P} \begin{bmatrix} \mathbf{R}^{-1} & \mathbf{0}^{k\times(p-k)} \\ \mathbf{0}^{(n-k)\times k} & \mathbf{0}^{(n-k)\times(p-k)} \end{bmatrix} \mathbf{Q} \,. \tag{2.19}$$

Note that this particular symmetric generalized inverse also satisfies the additional equation

$$\mathbf{C}^{-}\mathbf{C}\mathbf{C}^{-} = \mathbf{C}^{-} \,.$$

Furthermore, if $k = n$ then $\mathbf{C}^{-} = \mathbf{C}^{+}$. Finally, the vector $\mathbf{C}^{-}\mathbf{y}$ is also a solution of the linear least-squares problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{y} - \mathbf{C}\mathbf{x}\|_2 = \|\mathbf{r}(\mathbf{x})\|_2 \,,$$

but not the solution of minimum Euclidean norm if $k < n$ [111][64][87][8]. In other words, the pseudo-inverse $\mathbf{C}^{+}$ singles out the least-squares solution of minimum Euclidean length, which is not the case of $\mathbf{C}^{-}$.

If $k < n$, by applying additional Householder transformations (or, alternatively, Givens rotations) on the right of the QRCP to annihilate the submatrix $\mathbf{S}$, it is possible to obtain a Complete Orthogonal Decomposition (COD) of the matrix $\mathbf{C}$ of rank k (see Chapter 5 of [71] or Theorems 29 and 30 of [87] and also [64]). More precisely, by applying these additional Householder transformations, we obtain the following expression

$$\mathbf{QCP} = \begin{bmatrix} \mathbf{R} & \mathbf{S} \\ \mathbf{0}^{(p-k)\times k} & \mathbf{0}^{(p-k)\times(n-k)} \end{bmatrix} = \begin{bmatrix} \mathbf{T} & \mathbf{0}^{k\times(n-k)} \\ \mathbf{0}^{(p-k)\times k} & \mathbf{0}^{(p-k)\times(n-k)} \end{bmatrix} \mathbf{Z} \,,$$

where $\mathbf{Z}$ is an $n \times n$ orthogonal matrix and is again implicitly represented by the product of elementary Householder matrices and $\mathbf{T}$ is an $k \times k$ upper triangular matrix of full rank (which is different from the triangular factor $\mathbf{R}$ in the QRCP). The COD of $\mathbf{C}$ is then defined as

$$\mathbf{QCO} = \mathbf{QC}(\mathbf{PZ}^T) = \begin{bmatrix} \mathbf{T} & \mathbf{0}^{k\times(n-k)} \\ \mathbf{0}^{(p-k)\times k} & \mathbf{0}^{(p-k)\times(n-k)} \end{bmatrix} \,, \tag{2.20}$$

where $\mathbf{Q}$ is the same $p \times p$ orthogonal matrix as in the QRCP, $\mathbf{T}$ is an $k \times k$ nonsingular upper triangular matrix and $\mathbf{O} = \mathbf{PZ}^T$ is an $n \times n$ orthogonal matrix as the product of two orthogonal matrices. With the help of a COD of $\mathbf{C}$, its pseudo-inverse can be represented by

$$\mathbf{C}^{+} = \mathbf{O} \begin{bmatrix} \mathbf{T}^{-1} & \mathbf{0}^{k\times(p-k)} \\ \mathbf{0}^{(n-k)\times k} & \mathbf{0}^{(n-k)\times(p-k)} \end{bmatrix} \mathbf{Q} \,. \tag{2.21}$$

It is easily checked that this $n \times p$ matrix verifies the four equations (2.9) defining the pseudo-inverse of $\mathbf{C}$ and since, for any matrix $\mathbf{C}$, there is only one matrix having these four properties, the above matrix is the pseudo-inverse of $\mathbf{C}$. This demonstrates that there is no need to compute a more costly SVD of $\mathbf{C}$ for this purpose. Importantly, with a COD, we also get the orthogonal projectors on the row space of $\mathbf{C}$ and its orthogonal complement as

$$\mathbf{P}_{\mathbf{C}^T} = \mathbf{C}^{+}\mathbf{C} = \mathbf{O} \begin{bmatrix} \mathbf{I}_k & \mathbf{0}^{k\times(n-k)} \\ \mathbf{0}^{(n-k)\times k} & \mathbf{0}^{(n-k)\times(n-k)} \end{bmatrix} \mathbf{O}^T$$

and

$$\mathbf{P}_{\mathbf{C}^T}^{\perp} = \mathbf{I}_n - \mathbf{C}^{+}\mathbf{C} = \mathbf{O} \begin{bmatrix} \mathbf{0}^{k\times k} & \mathbf{0}^{k\times(n-k)} \\ \mathbf{0}^{(n-k)\times k} & \mathbf{I}_{n-k} \end{bmatrix} \mathbf{O}^T \ .$$

Finally, if we assume that $\mathbf{C}$ is of full column rank $k = n < p$, there is no need to compute a QRCP or COD of $\mathbf{C}$ to get the pseudo-inverse and the orthogonal projectors on the row or column spaces of $\mathbf{C}$ as a simple QR decomposition will do the job.

The SVD theory also provides the characterization of the best rank-k approximation of a given $p \times n$ real matrix in terms of the Frobenius norm [71]. As the Frobenius norm is unitarily invariant, we first note that

$$\|\mathbf{C}\|_F = \|\Sigma\|_F = \Big(\sum_{l=1}^{min(p,n)} \sigma_l^2 \Big)^{\frac{1}{2}} , \tag{2.22}$$

which shows that the Frobenius norm of a matrix is entirely defined by its singular values. From the SVD of a $p \times n$ matrix $\mathbf{C} = \mathbf{U}\Sigma\mathbf{V}^T$, we can also obtain directly its spectral norm as

$$\|\mathbf{C}\|_S = \sigma_1 = \mathbf{U}_{.1}^T \mathbf{C} \mathbf{V}_{.1} \ . \tag{2.23}$$

Then, the following theorem is the reason for the importance of the SVD for applications involving low-rank approximation of matrices:

Theorem 2.1. Let the SVD of $\mathbf{C} \in \mathbb{R}^{p\times n}$ be $\mathbf{C} = \mathbf{U}\Sigma\mathbf{V}^T$ with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{min(p,n)}$. In addition, for k such that $1 \leq k \leq min(p,n)$, defined the truncated SVD of $\mathbf{C}$ by

$$\mathbf{C}_k = \mathbf{U}_k \Sigma_k \mathbf{V}_k^T \ ,$$

where $\mathbf{U}_k$ and $\mathbf{V}_k$ are the submatrices formed by the k first columns of $\mathbf{U}$ and $\mathbf{V}$, respectively, and $\Sigma_k = \mathit{diag}([\sigma_1, \cdots, \sigma_k])$. Then, $\mathbf{C}_k$ provides a matrix of rank at most k that is closest in Frobenius norm to $\mathbf{C}$ and this minimum distance is given by

$$\|\mathbf{C} - \mathbf{C}_k\|_F = \min_{\mathbf{B}\in\mathbb{R}^{p\times n} \text{ with } \mathit{rank}(\mathbf{B})\leq k} \|\mathbf{C} - \mathbf{B}\|_F = \Big(\sum_{l=k+1}^{min(p,n)} \sigma_l^2 \Big)^{\frac{1}{2}} .$$

If $\sigma_k > \sigma_{k+1}$ or if $\sigma_k = 0$ then $\mathbf{C}_k$ is the unique best approximation of rank at most k of $\mathbf{C}$.

□

Theorem 2.1 is often called the Eckart-Young Theorem and is in fact valid in any unitarily invariant norm, see [71].

2.2 Multilinear algebra

In the next sections, we also need some operators and results from multilinear algebra [124]. These tools will be particularly useful when we need to manipulate matrices as elements of a linear vector space and for computing derivatives of matrices (or matrix-matrix products) with respect to another matrix.

For any $\mathbf{C}$ and $\mathbf{D}$ matrices of the same dimensions, the expression $\mathbf{C} \odot \mathbf{D}$ is used to mean the element-wise product of the $\mathbf{C}$ and $\mathbf{D}$ matrices (e.g., the Hadarmard product of two matrices):

$$\big[\mathbf{C} \odot \mathbf{D}\big]_{ij} = \mathbf{C}_{ij}.\mathbf{D}_{ij} \ . \tag{2.24}$$

The following property holds for matrices $\mathbf{B}$, $\mathbf{C}$ and $\mathbf{D}$ of the same shapes:

$$\langle \mathbf{B} \odot \mathbf{C} \ , \ \mathbf{D} \rangle_F = \langle \mathbf{C} \ , \ \mathbf{B} \odot \mathbf{D} \rangle_F \ .$$

Let $\mathbf{C} \in \mathbb{R}^{q\times r}$ and $\mathbf{C}_{.j}$ denotes the j^{th} column of $\mathbf{C}$, then the *vec*$(.)$ function maps the $q \times r$ matrix $\mathbf{C}$ into a $q.r \times 1$ column vector by "stacking" the columns of $\mathbf{C}$ below one another

$$\mathbf{C} \in \mathbb{R}^{q\times r} \Longrightarrow \textit{vec}(\mathbf{C}) = \begin{bmatrix} \mathbf{C}_{.1} \\ \vdots \\ \mathbf{C}_{.r} \end{bmatrix} \in \mathbb{R}^{q.r} \ . \tag{2.25}$$

The *vec*$(.)$ operator is an element of $\pounds(\mathbb{R}^{q\times r}, \mathbb{R}^{q.r})$, e.g., is a continuous linear mapping from $\mathbb{R}^{q\times r}$ into $\mathbb{R}^{q.r}$ and is also a bijection. The *mat*$(.)$ operator is then the inverse mapping of *vec*$(.)$, which is a continuous linear bijection from $\mathbb{R}^{q.r}$ into $\mathbb{R}^{q\times r}$ such that

$$\textit{mat}\big(\textit{vec}(\mathbf{C})\big) = \mathbf{C} \ , \forall \mathbf{C} \in \mathbb{R}^{q\times r} \ . \tag{2.26}$$

When it is not obvious from the context what is the shape of the image matrix for a given vector $\mathbf{c} \in \mathbb{R}^{q.r}$, we will use the notation $\textit{mat}_{q\times r}(.)$ instead.

A useful property involving the *vec*$(.)$ and Hadamard operators is that the vectorized form of the Hadamard product of two matrices of the same dimensions can be written as a matrix-vector product

$$\textit{vec}(\mathbf{C} \odot \mathbf{D}) = \textit{diag}\big(\textit{vec}(\mathbf{C})\big)\textit{vec}(\mathbf{D}) \ . \tag{2.27}$$

Let further $\mathbf{D} \in \mathbb{R}^{s\times t}$, then the Kronecker product $\mathbf{C} \otimes \mathbf{D}$ is the $q.s \times r.t$ block matrix, whose ij^{th} block is defined by

$$\big[\mathbf{C} \otimes \mathbf{D}\big]^{ij} = \mathbf{C}_{ij}\mathbf{D} \text{ for } i = 1, \cdots, q \text{ and } j = 1, \cdots, r \ . \tag{2.28}$$

The Kronecker product is a bilinear operator meaning that

$$\begin{aligned} (\mathbf{C} + \mathbf{D}) \otimes \mathbf{E} &= (\mathbf{C} \otimes \mathbf{E}) + (\mathbf{D} \otimes \mathbf{E}) \ , \\ \mathbf{E} \otimes (\mathbf{C} + \mathbf{D}) &= (\mathbf{E} \otimes \mathbf{C}) + (\mathbf{E} \otimes \mathbf{D}) \ , \\ \alpha(\mathbf{C} \otimes \mathbf{D}) &= (\alpha\mathbf{C}) \otimes \mathbf{D} = \mathbf{C} \otimes (\alpha\mathbf{D}) \ , \end{aligned} \tag{2.29}$$

where $\alpha \in \mathbb{R}$, $\mathbf{E}$ is any matrix, and $\mathbf{C}$ and $\mathbf{D}$ are two matrices of the same dimensions. We assume that the reader is familiar with the basic properties of Kronecker products (see Chapter 2 of [124] for details). For easy reference, we only state the following relations for any matrices $\mathbf{C}$ and $\mathbf{D}$:

$$\begin{aligned} (\mathbf{C} \otimes \mathbf{D})^T &= \mathbf{C}^T \otimes \mathbf{D}^T \\ \textit{rank}(\mathbf{C} \otimes \mathbf{D}) &= \textit{rank}(\mathbf{C}).\textit{rank}(\mathbf{D}) \ ; \end{aligned} \tag{2.30}$$

for partitioned matrices:

$$\begin{bmatrix} \mathbf{C}_1 & \mathbf{C}_2 \end{bmatrix} \otimes \mathbf{D} = \begin{bmatrix} \mathbf{C}_1 \otimes \mathbf{D} & \mathbf{C}_2 \otimes \mathbf{D} \end{bmatrix} \ ; \tag{2.31}$$

and for conforming matrices $\mathbf{C}$, $\mathbf{D}$, $\mathbf{E}$ and $\mathbf{F}$:

$$\begin{aligned} (\mathbf{C} \otimes \mathbf{D})(\mathbf{E} \otimes \mathbf{F}) &= \mathbf{CE} \otimes \mathbf{DF} \ , \\ \textit{vec}(\mathbf{CDE}) &= (\mathbf{E}^T \otimes \mathbf{C})\textit{vec}(\mathbf{D}) \ . \end{aligned} \tag{2.32}$$

This last equality is particularly useful to rearrange a matrix-matrix product as a simple matrix-vector product:

$$\begin{aligned} \textit{vec}(\mathbf{CD}) &= \textit{vec}(\mathbf{CDI}) = (\mathbf{I} \otimes \mathbf{C})\textit{vec}(\mathbf{D}) \ , \\ \textit{vec}(\mathbf{CD}) &= \textit{vec}(\mathbf{ICD}) = (\mathbf{D}^T \otimes \mathbf{I})\textit{vec}(\mathbf{C}) \ , \end{aligned} \tag{2.33}$$

where $\mathbf{I}$ is the identity matrix of appropriate order. These two relationships illustrate that we can evaluate the derivative of a matrix-matrix product with respect to one of the matrices by reshaping

the different matrices as vectors and computing the Jacobian matrix (see Subsection 2.4 for more details). Thus, we will use these two relations very frequently in the next sections, without explicit citation, when we need to compute derivatives of some matrices.

Let $\mathbf{X} \in \mathbb{R}^{p\times n}$. We now introduce the $p.n \times p.n$ permutation matrix $\mathbf{K}_{(p,n)}$ uniquely defined by the relation

$$\mathbf{K}_{(p,n)} vec(\mathbf{X}) = vec(\mathbf{X}^T) \ . \tag{2.34}$$

This permutation matrix is well-known in statistics where it is called the commutation matrix, see Chapter 3 of [124] for details. Its explicit form is given by

$$\mathbf{K}_{(p,n)} = \sum_{i=1}^{p}\sum_{j=1}^{n} \mathbf{H}(i,j) \otimes \mathbf{H}(i,j)^T \ ,$$

where $\mathbf{H}(i,j)$ is an $p \times n$ matrix with a 1 in its ij^{th} position and zeroes elsewhere. Important properties of the commutation matrix for our application are

$$\mathbf{K}_{(n,p)} = \mathbf{K}_{(p,n)}^T \text{ and } \mathbf{K}_{(p,n)}^T \mathbf{K}_{(p,n)} = \mathbf{K}_{(p,n)} \mathbf{K}_{(p,n)}^T = \mathbf{I}_{p.n} \ . \tag{2.35}$$

In other words, $\mathbf{K}_{(p,n)}$ is an orthogonal matrix and its transpose and inverse is $\mathbf{K}_{(n,p)}$. Another useful property of the commutation matrix is that it can be used to reverse the order of a Kronecker product

$$\mathbf{K}_{(s,p)}(\mathbf{X} \otimes \mathbf{Y}) = (\mathbf{Y} \otimes \mathbf{X})\mathbf{K}_{(t,n)} \text{ where } \mathbf{X} \in \mathbb{R}^{p\times n} \text{ and } \mathbf{Y} \in \mathbb{R}^{s\times t} \ . \tag{2.36}$$

This property will be also used frequently in the calculation of matrix derivatives. The following Lemma will also be useful later:

Lemma 2.2. Let $\mathbf{X} \in \mathbb{R}^{p\times n}$, then

$$\begin{aligned} \mathbf{K}_{(p,n)} diag(vec(\mathbf{X}))\mathbf{K}_{(n,p)} &= diag(vec(\mathbf{X}^T)) \ , \\ diag(vec(\mathbf{X}))\mathbf{K}_{(n,p)} &= \mathbf{K}_{(n,p)} diag(vec(\mathbf{X}^T)) \ . \end{aligned}$$

Proof. Omitted.

□

2.3 Topology of Euclidean vector or Frobenius matrix spaces

For the sake of convenience, we define some notations first. The following subsets of $\mathbb{R}^{p\times n}$, the set of $p \times n$ real matrices, and $\mathbb{R}$ will be used frequently in the following sections.

Definition 2.1. Let $p, n, k \in \mathbb{N}_*$ with $k \leq \min(p,n)$, then

$$\begin{aligned} \mathbb{R}_k^{p\times n} &= \left\{\mathbf{Y} \in \mathbb{R}^{p\times n} \text{ and } rank(\mathbf{Y}) = k\right\} \ , \\ \mathbb{R}_{\leq k}^{p\times n} &= \left\{\mathbf{Y} \in \mathbb{R}^{p\times n} \text{ and } rank(\mathbf{Y}) \leq k\right\} \ , \\ \mathbb{R}_{>k}^{p\times n} &= \left\{\mathbf{Y} \in \mathbb{R}^{p\times n} \text{ and } rank(\mathbf{Y}) > k\right\} \ , \\ \mathbb{R}_{+}^{p\times n} &= \left\{\mathbf{Y} \in \mathbb{R}^{p\times n} \text{ and } \mathbf{Y}_{ij} \geq 0\right\} \ , \\ \mathbb{R}_{+*}^{p\times n} &= \left\{\mathbf{Y} \in \mathbb{R}^{p\times n} \text{ and } \mathbf{Y}_{ij} > 0\right\} \ , \\ \mathbb{O}^{p\times k} &= \left\{\mathbf{U} \in \mathbb{R}^{p\times k} \ / \ \mathbf{U}^T\mathbf{U} = \mathbf{I}_k\right\} \ , \\ \mathbb{O}_t^{k\times n} &= \left\{\mathbf{U} \in \mathbb{R}^{k\times n} \ / \ \mathbf{U}\mathbf{U}^T = \mathbf{I}_k\right\} \ , \end{aligned}$$

and

$$\mathbb{R}_* = \left\{\mathbf{x} \in \mathbb{R} \ / \ \mathbf{x} \neq 0\right\} , \mathbb{R}_+ = \left\{\mathbf{x} \in \mathbb{R} \ / \ \mathbf{x} \geq 0\right\} \text{ and } \mathbb{R}_{+*} = \left\{\mathbf{x} \in \mathbb{R} \ / \ \mathbf{x} > 0\right\} .$$

The following definitions will also be useful:

Definition 2.2. Given $\mathbf{X} \in \mathbb{R}^{p\times n}$, $r \in \mathbb{R}_{+*}$, the open ball with center $\mathbf{X}$ and radius r of $\mathbb{R}^{p\times n}$ is the set of $p \times n$ matrices defined by

$$B_{p\times n}(\mathbf{X}, r) = \left\{\mathbf{Y} \in \mathbb{R}^{p\times n} \text{ and } \|\mathbf{X} - \mathbf{Y}\| < r\right\}$$

and the closed ball with center $\mathbf{X}$ and radius r is the set

$$\bar{B}_{p\times n}(\mathbf{X}, r) = \left\{\mathbf{Y} \in \mathbb{R}^{p\times n} \text{ and } \|\mathbf{X} - \mathbf{Y}\| \leq r\right\}.$$

Note that in these definitions, $\|\|$ can be the Frobenius norm or any norm defined on $\mathbb{R}^{p\times n}$ since $\mathbb{R}^{p\times n}$ is a finite-dimensional vector space over $\mathbb{R}$ and, in this case, all norms on $\mathbb{R}^{p\times n}$ are equivalent and induce the same topology [12][26]. Similarly, given $\mathbf{x} \in \mathbb{R}^n$, $r \in \mathbb{R}_{+*}$, the open ball with center $\mathbf{x}$ and radius r of $\mathbb{R}^n$ is the set of n-dimensional vectors defined by

$$B_n(\mathbf{x}, r) = \left\{\mathbf{y} \in \mathbb{R}^n \text{ and } \|\mathbf{x} - \mathbf{y}\| < r\right\}$$

and the closed ball with center $\mathbf{x}$ and radius r is the set

$$\bar{B}_n(\mathbf{x}, r) = \left\{\mathbf{y} \in \mathbb{R}^n \text{ and } \|\mathbf{x} - \mathbf{y}\| \leq r\right\}.$$

Again, here, $\|\|$ can be the Euclidean norm or any norm defined on $\mathbb{R}^n$.

A set $U \subset \mathbb{R}^{p\times n}$ is open if every point of U is contained in an open ball included in U. A set $U \subset \mathbb{R}^{p\times n}$ is closed if and only if its complement in $\mathbb{R}^{p\times n}$ is open. An arbitrary union of open sets is open and an arbitrary intersection of closed sets is closed. A finite union of closed sets is also closed. The closure of a set $U \subset \mathbb{R}^{p\times n}$ is the smallest closed set (in the sense of inclusion) of $\mathbb{R}^{p\times n}$ which contains U and is denoted $\bar{U}$. On the other hand, the interior of a set $U \subset \mathbb{R}^{p\times n}$ is the largest open set (in the sense of inclusion) of $\mathbb{R}^{p\times n}$ which is included in U and is denoted $\mathring{U}$. A set N is called a neighborhood of $\mathbf{X}$ in $\mathbb{R}^{p\times n}$ if there is an open set $U \subset N$ with $\mathbf{X} \in U$. A point $\mathbf{X}$ is a boundary point of a set A if every neighborhood of $\mathbf{X}$ contains a point of A and a point of its complement B in $\mathbb{R}^{p\times n}$. The set of boundary points of A is denoted $bd(A)$ and we have $bd(A) = \bar{A} \cap \bar{B}$. Thus, $bd(A)$ is closed as the intersection of two closed sets. The term frontier refers to the set of points of $bd(A)$ which are not in A (e.g., $\bar{A}/A$). As an illustration, the boundary of both the open ball $B_{p\times n}(\mathbf{X}, r)$ and the closed ball $\bar{B}_{p\times n}(\mathbf{X}, r)$ is the sphere $S_{p\times n}(\mathbf{X}, r) = \left\{\mathbf{Y} \in \mathbb{R}^{p\times n} \text{ and } \|\mathbf{X} - \mathbf{Y}\| = r\right\}$, but the frontier of $B_{p\times n}(\mathbf{X}, r)$ is equal to $bd\big(B_{p\times n}(\mathbf{X}, r)\big)$ while the frontier of $\bar{B}_{p\times n}(\mathbf{X}, r)$ is empty. Let now A and B be two subsets of $\mathbb{R}^{p\times n}$ such that $A \subset B$. We say that A is dense in B if $B \subset \bar{A}$ and we say that A is dense everywhere if $\bar{A} = \mathbb{R}^{p\times n}$. Similar definitions hold for $\mathbf{x}$ in $\mathbb{R}^n$. In a finite-dimensional vector (or matrix) space over $\mathbb{R}$, a closed and bounded set is compact and all closed balls are compact. The preimage of a closed (open) set by a continuous function is a closed (open) set. The image of a compact set by a continuous function is compact.

Next, we collect some important topological results concerning certain subsets of $\mathbb{R}^{p\times n}$ in the following theorem that we will also use frequently in the following sections.

Theorem 2.3. Let $p, n, k \in \mathbb{N}_*$ with $k \leq \min(p, n)$. The sets $\mathbb{O}^{p\times k}$ and $\mathbb{O}_t^{k\times n}$ are compact in $\mathbb{R}^{p\times k}$ and $\mathbb{R}^{k\times n}$, respectively. The set $\mathbb{R}_k^{p\times k}$ is open in $\mathbb{R}^{p\times k}$. If $k \neq n$, the interior of $\mathbb{R}_k^{p\times n}$ is empty and $\mathbb{R}_k^{p\times n}$ is not closed or open in $\mathbb{R}^{p\times n}$. The sets $\mathbb{R}_{>k}^{p\times n}$ and $\mathbb{R}_{\leq k}^{p\times n}$ are, respectively, open and closed in $\mathbb{R}^{p\times n}$. In all cases, the sets $\mathbb{R}_{\leq k}^{p\times n}$ and $\mathbb{R}_{<k}^{p\times n}$ are, respectively, the closure and the frontier of $\mathbb{R}_k^{p\times n}$ in $\mathbb{R}^{p\times n}$, and $\mathbb{R}_k^{p\times n}$ is dense in $\mathbb{R}_{\leq k}^{p\times n}$. Furthermore, the set $\mathbb{R}_k^{p\times k}$ is an open subset dense everywhere in $\mathbb{R}^{p\times k}$.

Proof. Omitted. See Section 3 and Theorem 2 of [80], Proposition 2.1 of [142] and Section 3.1.5 of [3] for some details. □

Importantly, since the *rank*(.) function defined on $\mathbb{R}^{p\times n}$ is an integer-valued and lower-semicontinuous function, an important result is that the rank function does not decrease in a sufficiently small neighborhood of any matrix $\mathbf{X} \in \mathbb{R}^{p\times n}$ [80]. On the other hand, with the help of the SVD and Theorem (2.1), it is not difficult to see that if *rank*$(\mathbf{X}) = k < \min(p, n)$ then any neighborhood of $\mathbf{X}$ contains matrices of rank $k+1, k+2, \ldots, \min(p, n)$.

To close this subsection, we finally recall the definition of a convex set for later reference. A subset $\mathcal{C}$ of a normed vector space $\mathcal{X}$ is called convex, if

$$\lambda.\mathbf{x} + (1-\lambda).\mathbf{y} \in \mathcal{C} \text{ whenever } \mathbf{x}, \mathbf{y} \in \mathcal{C} \text{ and } 0 \leq \lambda \leq 1 \ . \tag{2.37}$$

Geometrically, a subset of a normed vector space is convex, if and only if, it contains the line segment $\{\lambda.\mathbf{x} + (1-\lambda).\mathbf{y} \ / \ 0 \leq \lambda \leq 1\}$ joining each pair of its points $\mathbf{x}, \mathbf{y}$. As an illustration, the open and closed balls of $\mathcal{X}$ and the subspaces of $\mathcal{X}$ are all convex. Note, on the other hand, that the subsets $\mathbb{R}_k^{p\times n}$ and $\mathbb{R}_{\leq k}^{p\times n}$ of $\mathbb{R}^{p\times n}$ are not convex if $k \neq n$, which makes solving the WLRA problem (P0) challenging.

For more discussion about the topology of $\mathbb{R}^{p\times n}$ or arbitrary normed vector spaces, we refer the reader to [148][12][26].

2.4 Differential calculus, variational geometry and optimization

We also assume that the reader has some familiarity with differentiation in a Euclidean space and derivatives of vectors and matrices, and their properties. Useful references on these topics are [148][26][124].

Let $\mathcal{X}$ be a Euclidean space, e.g., a real vector space of finite dimension, say k, equipped with a scalar product $\langle ., . \rangle_{\mathcal{X}}$ and the vector norm $\|.\|_{\mathcal{X}}$ induced by this scalar product. Let now $\phi(.)$ be a function from an open set $\Omega \subset \mathcal{X}$ to some other Euclidean space, say $\mathcal{Y}$. In the following, we may have $\mathcal{Y} = \mathbb{R}$, $\mathcal{Y} = \mathbb{R}^p$ or $\mathcal{Y} = \mathbb{R}^{p\times n}$. We say that $\phi(.)$ is $o(\|\mathbf{h}\|_{\mathcal{X}})$ if

$$\forall \varepsilon \in \mathbb{R}_{+*}, \exists \delta \in \mathbb{R}_{+*} \text{ such that } \|\mathbf{h}\|_{\mathcal{X}} \leq \delta \Longrightarrow \|\phi(\mathbf{h})\|_{\mathcal{Y}} \leq \varepsilon \|\mathbf{h}\|_{\mathcal{X}} \ .$$

Similarly, we say that $\phi(.)$ is $\mathcal{O}(\|\mathbf{h}\|_{\mathcal{X}})$ if

$$\exists \lambda, \eta \in \mathbb{R}_{+*} \text{ such that } \|\mathbf{h}\|_{\mathcal{X}} \leq \eta \Longrightarrow \|\phi(\mathbf{h})\|_{\mathcal{Y}} \leq \lambda \|\mathbf{h}\|_{\mathcal{X}} \ .$$

Notations like $o(\|\mathbf{h}\|_{\mathcal{X}}^{\alpha})$ or $\mathcal{O}(\|\mathbf{h}\|_{\mathcal{X}}^{\alpha})$, for $\alpha \in \mathbb{N}_*$, will also be used to distinguish functions tending to zero faster than $\|\mathbf{h}\|_{\mathcal{X}}^{\alpha}$ instead of faster than $\|\mathbf{h}\|_{\mathcal{X}}$.

With these notations, a function $\phi(.)$ from the open set $\Omega \subset \mathcal{X}$ to the Euclidean space $\mathcal{Y}$ is said to be differentiable at $\mathbf{a} \in \Omega$, if there exists a linear operator $\phi^{'}(\mathbf{a})$ from $\mathcal{X}$ to $\mathcal{Y}$ such that

$$\phi(\mathbf{a}+\mathbf{h}) = \phi(\mathbf{a}) + \phi^{'}(\mathbf{a})(\mathbf{h}) + o(\|\mathbf{h}\|_{\mathcal{X}}) \ .$$

The set of (continuous) linear operators from $\mathcal{X}$ to $\mathcal{Y}$ is denoted by $\pounds(\mathcal{X}, \mathcal{Y})$. If $\mathcal{Y} = \mathbb{R}$, then $\phi(.)$ is a real-valued function, $\phi^{'}(\mathbf{a})$ is a linear form and it can be represented by an unique element of $\mathcal{X}$, called the gradient of $\phi(.)$ at $\mathbf{a}$ and denoted by $\nabla\phi(\mathbf{a})$, which verifies

$$\phi^{'}(\mathbf{a})(\mathbf{h}) = \langle \nabla\phi(\mathbf{a}), \mathbf{h} \rangle_{\mathcal{X}} \ , \forall \mathbf{h} \in \mathcal{X} \ .$$

In the same conditions, e.g., when $\mathcal{Y} = \mathbb{R}$, we can consider the function from Ω into $\pounds(\mathcal{X}, \mathbb{R})$, which at $\mathbf{a} \in \Omega$ associates the linear form $\phi^{'}(\mathbf{a})$. If this new function is itself differentiable, we get the second-order differential of $\phi(.)$ at $\mathbf{a}$, which is denoted by $\phi^{''}(\mathbf{a})$ and is an element of $\pounds(\mathcal{X}, \pounds(\mathcal{X}, \mathbb{R})) \simeq \pounds(\mathcal{X}, \mathcal{X}; \mathbb{R})$. In other words, $\phi^{''}(\mathbf{a})$ can be identified with an unique bilinear form, also noted $\phi^{''}(\mathbf{a}) \in \pounds(\mathcal{X}, \mathcal{X}; \mathbb{R})$ by an abuse of notation, and defined by

$$[\phi^{''}(\mathbf{a})(\mathbf{h})](\mathbf{k}) = \phi^{''}(\mathbf{a})(\mathbf{h}, \mathbf{k}) \ , \forall (\mathbf{h}, \mathbf{k}) \in \mathcal{X} \times \mathcal{X} \ .$$

This bilinear form is also symmetric and yields the following second-order approximation of $\phi(.)$ at $\mathbf{a}$

$$\phi(\mathbf{a}+\mathbf{h}) = \phi(\mathbf{a}) + \phi^{'}(\mathbf{a})(\mathbf{h}) + \phi^{''}(\mathbf{a})(\mathbf{h},\mathbf{h}) + o(\|\mathbf{h}\|_{\mathcal{X}}^2) \; .$$

Again, using the Euclidean structure associated with $\mathcal{X}$, the symmetric bilinear form $\phi^{''}(\mathbf{a})$ can be associated with an unique symmetric linear operator from $\mathcal{X}$ to $\mathcal{X}$, called the Hessian of $\phi(.)$ at $\mathbf{a}$, denoted by $\nabla^2\phi(\mathbf{a})$, and defined by

$$\phi^{''}(\mathbf{a})(\mathbf{h},\mathbf{k}) = \langle \nabla^2\phi(\mathbf{a})(\mathbf{h}),\mathbf{k}\rangle_{\mathcal{X}} = \langle \mathbf{h}, \nabla^2\phi(\mathbf{a})(\mathbf{k})\rangle_{\mathcal{X}} \; , \forall (\mathbf{h},\mathbf{k}) \in \mathcal{X}\times\mathcal{X} \; .$$

Note that both $\nabla\phi(\mathbf{a})$ and $\nabla^2\phi(\mathbf{a})$ depend on the scalar product $\langle .,. \rangle_{\mathcal{X}}$, while $\phi^{'}(\mathbf{a})$ and $\phi^{''}(\mathbf{a})$ do not. When $\mathcal{X} = \mathbb{R}^k$ and is equipped with the standard Euclidean inner product defined in Subsection 2.1 and the canonical basis of $\mathbb{R}^k$ is used to represent vectors in $\mathbb{R}^k$, the self-adjoint linear operator $\nabla^2\phi(\mathbf{a})$ is represented by a $k\times k$ symmetric real matrix, which is known as the Schwarz's theorem [26]. Then, by a slight abuse of notation, we will also use the symbol $\nabla^2\phi(\mathbf{a})$ to represent this $k\times k$ symmetric matrix and we can write

$$\phi^{''}(\mathbf{a})(\mathbf{h},\mathbf{k}) = \langle \nabla^2\phi(\mathbf{a})(\mathbf{h}),\mathbf{k}\rangle_2 = \mathbf{h}^T\nabla^2\phi(\mathbf{a})\mathbf{k} \; , \forall (\mathbf{h},\mathbf{k}) \in \mathbb{R}^k\times\mathbb{R}^k \; .$$

In the following sections, instead of the generic notations $\phi^{'}(\mathbf{a})$ and $\phi^{''}(\mathbf{a})$ for the first- and second-order derivatives of a (twice) differentiable function $\phi(.)$ from $\Omega \subset \mathcal{X}$ to $\mathcal{Y}$ at a point $\mathbf{a}\in\Omega$, the symbols D, J, ∇, ∇^2 will be used for the (Euclidean) derivative of a real matrix, the Jacobian matrix (e.g., derivative) of a real vector function, the gradient (e.g., first derivative) and Hessian (e.g., second derivative) of a real functional, respectively.

As a first illustration, a $q\times r$ matrix function $\mathbf{C}(\mathbf{a})$ for $\mathbf{a}\in\Omega=\mathbb{R}^k$ can be interpreted as a (nonlinear) mapping from the linear space of parameters, $\mathbb{R}^k$, into the space of linear transformations $\pounds(\mathbb{R}^r,\mathbb{R}^q) = \mathcal{Y}$, which can be identified to the linear space $\mathbb{R}^{q\times r}$ [26]. Consequently, the derivative of the matrix function $\mathbf{C}(.)$ at a point $\mathbf{a}\in\mathbb{R}^k$ is an element of $\pounds(\mathbb{R}^k,\pounds(\mathbb{R}^r,\mathbb{R}^q))$, or equivalently $\pounds(\mathbb{R}^k,\mathbb{R}^{q\times r}) \simeq \mathbb{R}^{q\times r\times k}$, and can be interpreted as the tridimensional tensor $D(\mathbf{C}(\mathbf{a}))\in\mathbb{R}^{q\times r\times k}$ defined by

$$\big[D(\mathbf{C}(\mathbf{a}))\big]_{ijl} = \frac{\partial \mathbf{C}_{ij}(\mathbf{a})}{\partial \mathbf{a}_l} \text{ for } i=1,\cdots,q \; ; \; j=1,\cdots,r \; ; \; l=1,\cdots,k \; , \tag{2.38}$$

following [63].

On the other hand, the first derivative of a real q-vector function $\mathbf{r}(.)$ at a point $\mathbf{a}\in\mathbb{R}^k$ is an element of $\pounds(\mathbb{R}^k,\mathbb{R}^q) \simeq \mathbb{R}^{q\times k}$. If $\mathbb{R}^k$ and $\mathbb{R}^q$ are equipped with their usual Euclidean inner products and the canonical bases of $\mathbb{R}^k$ (e.g., the columns of the identity matrix $\mathbf{I}_k$) and $\mathbb{R}^q$ (e.g., the columns of the identity matrix $\mathbf{I}_q$) are used to represent vectors in these two linear spaces, the first derivative of $\mathbf{r}(.)$ at a point $\mathbf{a}\in\mathbb{R}^k$ can be identified to the Jacobian matrix $J(\mathbf{r}(\mathbf{a}))\in\mathbb{R}^{q\times k}$ defined by

$$\big[J(\mathbf{r}(\mathbf{a}))\big]_{ij} = \frac{\partial \mathbf{r}_i(\mathbf{a})}{\partial \mathbf{a}_j} \text{ for } i=1,\cdots,q \; ; \; j=1,\cdots,k \; , \tag{2.39}$$

where $\mathbf{r}_i(\mathbf{a})$ is the i^{th} component of the real q-vector $\mathbf{r}(\mathbf{a})$. Note that each i^{th} row of the Jacobian matrix $J(\mathbf{r}(\mathbf{a}))\in\mathbb{R}^{q\times k}$ is equal to the transpose of the gradient of the real function $\mathbf{r}_i(.)$ at the point $\mathbf{a}\in\mathbb{R}^k$,

$$\big[J(\mathbf{r}(\mathbf{a}))\big]_{i.} = \nabla\mathbf{r}_i(\mathbf{a})^T \; .$$

In addition, if $\mathbf{r}(.)$ is continuously differentiable at a point $\mathbf{a}\in\mathbb{R}^k$, we have the following first-order Taylor expansion

$$\mathbf{r}(\mathbf{a}+d\mathbf{a}) = \mathbf{r}(\mathbf{a}) + J(\mathbf{r}(\mathbf{a}))d\mathbf{a} + \mathcal{O}(\|d\mathbf{a}\|_2^2) \; . \tag{2.40}$$

As another illustration, if $\mathbb{R}^k$ is again equipped with its usual Euclidean inner product and its canonical basis, the gradient of a real functional $\phi(.)$ at a point $\mathbf{a}\in\mathbb{R}^k$ forms a $k\times 1$ column vector, i.e.,

$$\big[\nabla\phi(\mathbf{a})\big]_i = \frac{\partial\phi(\mathbf{a})}{\partial\mathbf{a}_i} \text{ for } i=1,\cdots,k \; , \tag{2.41}$$

and the Hessian of a real functional $\phi(.)$ at a point $\mathbf{a} \in \mathbb{R}^k$ can be identified with a $k \times k$ symmetric matrix, $\nabla^2\phi(\mathbf{a})$, defined by

$$\big[\nabla^2\phi(\mathbf{a})\big]_{ij} = \frac{\partial^2\phi(\mathbf{a})}{\partial\mathbf{a}_i\partial\mathbf{a}_j} = \frac{\partial^2\phi(\mathbf{a})}{\partial\mathbf{a}_j\partial\mathbf{a}_i} = \big[\nabla^2\phi(\mathbf{a})\big]_{ji} \text{ for } i = 1, \cdots, k \, ; j = 1, \cdots, k \, . \tag{2.42}$$

Finally, if $\phi(.)$ is at least twice continuously differentiable at a point $\mathbf{a} \in \mathbb{R}^k$, we have the following second-order Taylor expansion

$$\begin{aligned} \phi(\mathbf{a} + d\mathbf{a}) &= \phi(\mathbf{a}) + \langle d\mathbf{a}, \nabla\phi(\mathbf{a})\rangle_2 + \frac{1}{2}\langle\nabla^2\phi(\mathbf{a})d\mathbf{a}, d\mathbf{a}\rangle_2 + \mathcal{O}(\|d\mathbf{a}\|_2^3) \\ &= \phi(\mathbf{a}) + d\mathbf{a}^T\nabla\phi(\mathbf{a}) + \frac{1}{2}d\mathbf{a}^T\nabla^2\phi(\mathbf{a})d\mathbf{a} + \mathcal{O}(\|d\mathbf{a}\|_2^3) \, . \end{aligned} \tag{2.43}$$

Let $\mathcal{K}$ be a nonempty subset of $\mathbb{R}^k$ or more generally a nonempty subset of an arbitrary normed vector space. We recall that a point $\hat{\mathbf{a}} \in \mathcal{K}$ is a global minimizer of a real function $\phi(.)$ defined over $\mathcal{K}$, if and only if, $\forall \mathbf{a} \in \mathcal{K}$, we have $\phi(\mathbf{a}) \geq \phi(\hat{\mathbf{a}})$. On the other hand, $\hat{\mathbf{a}}$ is a local minimizer of $\phi(.)$ over $\mathcal{K}$, if and only if, $\exists r \in \mathbb{R}_{+*}$ such that $\forall \mathbf{a} \in \mathcal{K}$ and $\|\mathbf{a} - \hat{\mathbf{a}}\|_2 < r$ imply $\phi(\mathbf{a}) \geq \phi(\hat{\mathbf{a}})$. Similar definitions hold for strict global and local minimizers of $\phi(.)$ over $\mathcal{K}$.

Let now Ω be an open subset of $\mathbb{R}^k$ or more generally an open subset of a normed vector space of finite dimension. A necessary condition for a point $\hat{\mathbf{a}} \in \Omega$ to minimize a real function $\phi(.)$ defined and assumed to be twice continuously differentiable on Ω is that the gradient of $\phi(.)$ at $\hat{\mathbf{a}}$ is equal to the zero-vector of the ambient linear space, i.e.,

$$\nabla\phi(\hat{\mathbf{a}}) = \mathbf{0}^k \, , \tag{2.44}$$

and this condition defines the first-order Karush-Kuhn-Tucker (KKT) condition [148][26]. If such KKT condition is satisfied then $\hat{\mathbf{a}}$ is said to be a first-order stationary or critical point of $\phi(.)$. However, first-order critical points of $\phi(.)$ can be minimizers, but also maximizers or saddle points (e.g., points for which the Hessian matrix has both positive and negative eigenvalues). A necessary condition for a first-order stationary point $\hat{\mathbf{a}}$ to be a local minimizer of $\phi(.)$ is that the Hessian (bilinear form or matrix) $\nabla^2\phi(\hat{\mathbf{a}})$ is positive semi-definite [148][26]:

$$\nabla^2\phi(\hat{\mathbf{a}})\big(d\mathbf{a}, d\mathbf{a}\big) = \langle\nabla^2\phi(\hat{\mathbf{a}})d\mathbf{a}, d\mathbf{a}\rangle_2 \geq 0 \, , \forall d\mathbf{a} \in \mathbb{R}^k \, . \tag{2.45}$$

Such first-order critical points for which the Hessian is positive semi-definite are called second-order stationary or critical points of $\phi(.)$. On the other hand, a sufficient condition for a first-order stationary point $\hat{\mathbf{a}}$ to be a strict local minimizer of $\phi(.)$ is that $\nabla^2\phi(\hat{\mathbf{a}})$ is positive definite (second-order KKT condition). These assertions can be derived by noting that the second-order Taylor expansion of $\phi(.)$ at a first-order stationary point $\widehat{\mathbf{a}}$ reduces to

$$\phi(\widehat{\mathbf{a}} + d\mathbf{a}) = \phi(\widehat{\mathbf{a}}) + \frac{1}{2}\langle\nabla^2\phi(\hat{\mathbf{a}})d\mathbf{a}, d\mathbf{a}\rangle_2 + \mathcal{O}(\|d\mathbf{a}\|_2^3) \, ,$$

see [148][26] for details.

We now consider the case, where we seek to minimize a real function $\phi(.)$ on a linear subspace $\Upsilon \subset \Omega$ of dimension s, where Ω is an open subset of $\mathbb{R}^k$ on which $\phi(.)$ is defined and twice-differentiable. Obviously, we must have $dim(\Upsilon) \leq k$. In these conditions, we can consider $\phi(.)$ as a function from Ω to $\mathbb{R}$, but we can also consider its restriction to Υ, $\phi_\Upsilon(.)$. If $\phi(.)$ is differentiable on Ω than $\phi_\Upsilon(.)$ will be differentiable on Υ as well and their differentials verify

$$\phi^{'}(\mathbf{a})(\mathbf{b}) = \phi^{'}_\Upsilon(\mathbf{a})(\mathbf{b}) \, , \forall \mathbf{a}, \mathbf{b} \in \Upsilon \, .$$

In other words, the linear form $\phi^{'}_\Upsilon(\mathbf{a})$ is nothing else than the restriction of the linear form $\phi^{'}(\mathbf{a})$ to Υ. Furthermore, if we equip both $\mathbb{R}^k$ and its linear subspace Υ with the same Euclidean structure

induced by $\mathbb{R}^k$, we have, by definition, $\langle \mathbf{a}, \mathbf{b} \rangle_\Upsilon = \langle \mathbf{a}, \mathbf{b} \rangle_2, \forall \mathbf{a}, \mathbf{b} \in \Upsilon$. In these conditions, the linear forms $\phi^{'}(\mathbf{a})$ and $\phi^{'}_\Upsilon(\mathbf{a})$ can be both represented by their own gradients, $\nabla\phi(\mathbf{a})$ and $\nabla\phi_\Upsilon(\mathbf{a})$, which are, respectively, elements of the linear spaces $\mathbb{R}^k$ and Υ such that

$$\phi^{'}(\mathbf{a})(\mathbf{b}) = \langle \nabla\phi(\mathbf{a}), \mathbf{b} \rangle_2 \text{ and } \phi^{'}_\Upsilon(\mathbf{a})(\mathbf{b}) = \langle \nabla\phi_\Upsilon(\mathbf{a}), \mathbf{b} \rangle_\Upsilon \ , \forall \mathbf{a}, \mathbf{b} \in \Upsilon \ .$$

Since the linear forms $\phi^{'}(\mathbf{a})$ and $\phi^{'}_\Upsilon(\mathbf{a})$ coincide on Υ, we deduce immediately that

$$\langle \nabla\phi(\mathbf{a}), \mathbf{b} \rangle_2 = \langle \nabla\phi_\Upsilon(\mathbf{a}), \mathbf{b} \rangle_\Upsilon, \forall \mathbf{a}, \mathbf{b} \in \Upsilon \ .$$

Next, remember that $\nabla\phi(\mathbf{a}) \in \mathbb{R}^k$, while $\nabla\phi_\Upsilon(\mathbf{a}) \in \Upsilon$, but we can easily expressed $\nabla\phi_\Upsilon(\mathbf{a})$ as a function of $\nabla\phi(\mathbf{a})$. More precisely, using the two complementary orthogonal projectors on Υ and $\Upsilon^\perp$ (considered here as $k \times k$ symmetric and idempotent matrices rather than linear operators as discussed in Subsection 2.1), denoted, respectively, by $\mathbf{P}_\Upsilon$ and $\mathbf{P}^\perp_\Upsilon$, and defined on $\mathbb{R}^k$, we have

$$\nabla\phi(\mathbf{a}) = \mathbf{P}_\Upsilon\nabla\phi(\mathbf{a}) + \mathbf{P}^\perp_\Upsilon\nabla\phi(\mathbf{a}), \text{ with } \mathbf{P}_\Upsilon\nabla\phi(\mathbf{a}) \in \Upsilon \text{ and } \mathbf{P}^\perp_\Upsilon\nabla\phi(\mathbf{a}) \in \Upsilon^\perp \ ,$$

and this implies immediately that

$$\langle \nabla\phi(\mathbf{a}), \mathbf{b} \rangle_2 = \langle \mathbf{P}_\Upsilon\nabla\phi(\mathbf{a}), \mathbf{b} \rangle_2 = \langle \mathbf{P}_\Upsilon\nabla\phi(\mathbf{a}), \mathbf{b} \rangle_\Upsilon = \langle \nabla\phi_\Upsilon(\mathbf{a}), \mathbf{b} \rangle_\Upsilon, \forall \mathbf{a}, \mathbf{b} \in \Upsilon \ .$$

Then, by the unicity of the gradient of $\phi_\Upsilon(.)$ at $\mathbf{a} \in \Upsilon$, we get, by identification, the vector equality

$$\nabla\phi_\Upsilon(\mathbf{a}) = \mathbf{P}_\Upsilon\nabla\phi(\mathbf{a}), \forall \mathbf{a} \in \Upsilon \ . \tag{2.46}$$

In words, the gradient of $\phi_\Upsilon(.)$ at $\mathbf{a} \in \Upsilon$ is simply the orthogonal projection on Υ of the gradient of $\phi(.)$ at $\mathbf{a}$, considered as an element of $\mathbb{R}^k$ instead of Υ. The interpretation of this result is simple and is that it is not necessary to check all the feasible directions in $\mathbb{R}^k$ to satisfy the first-order stationary condition for a point $\mathbf{a} \in \Upsilon$ if the search space is reduced to Υ, only those belonging to Υ matter in that case. Similarly, it is not too difficult to verify that the linear operators $\nabla^2\phi_\Upsilon(\mathbf{a})$ and $\nabla^2\phi(\mathbf{a})$ are related by

$$\nabla^2\phi_\Upsilon(\mathbf{a})[\mathbf{b}] = \mathbf{P}_\Upsilon\big(\nabla^2\phi(\mathbf{a})[\mathbf{b}]\big), \forall \mathbf{b} \in \Upsilon \ , \tag{2.47}$$

where $\mathbf{P}_\Upsilon$ is now interpreted as an orthogonal projector operator rather than as a matrix. However, keep in mind that, in the above formulae, $\nabla^2\phi_\Upsilon(\mathbf{a})$ is expressed as a non-symmetric linear operator from $\mathbb{R}^k$ to $\mathbb{R}^k$ rather than as a symmetric linear operator from Υ to Υ (or a $s \times s$ symmetric matrix), but both operators coincide on Υ. Finally, Υ being a linear space of dimension s, the definitions of the first- and second-order stationary points of $\phi_\Upsilon(.)$ are exactly similar to those stated above. Namely, the first-order KKT condition is met if $\mathbf{P}_\Upsilon\nabla\phi(\mathbf{a}) = \mathbf{0}^k$ and the second-order stationary (KKT) condition is equivalent to say that the bilinear form associated with the self-adjoint linear operator $\mathbf{P}_\Upsilon o\nabla^2\phi(\mathbf{a})$ is positive semi-definite (positive definite) over Υ, for $\mathbf{a} \in \Upsilon$.

Next, we consider the problem of the minimization of a smooth function $\phi(.)$ defined on a smooth submanifold $\mathcal{M}$ of dimension r embedded in $\mathbb{R}^k$. This problem enters in the domain of differential geometry and optimization on Riemannian manifolds described comprehensively in [3][164][11]. We first precise what we mean by a smooth function and a smooth embedded manifold in $\mathbb{R}^k$ in the following definitions, which will be sufficient for our purpose.

Definition 2.3. Given any set $Q \subset \mathbb{R}^k$ and a mapping $\phi(.)$ from Q to $\mathbb{R}^m$, we say that $\phi(.)$ is C^p smooth if, $\forall \mathbf{a} \in Q$, there is a neighborhood U of $\mathbf{a}$ in $\mathbb{R}^k$ and a C^p differentiable mapping $\hat{\phi}(.)$ from U to $\mathbb{R}^m$ that agrees with $\phi(.)$ on $U \cap Q$. Here we assume that p lies in $\mathbb{N}_* \cup \{\infty\}$.

Definition 2.4. Let $\mathcal{M}$ be a nonempty subset of $\mathbb{R}^k$. We say that $\mathcal{M}$ is a C^p embedded submanifold of dimension r of $\mathbb{R}^k$, with $r \leq k$, if for each point $\mathbf{a} \in \mathcal{M}$, there is an open neighborhood U around $\mathbf{a}$ in $\mathbb{R}^k$ such that $U \cap \mathcal{M} = f^{-1}(\mathbf{0}^{k-r})$ for some C^p differentiable map $f(.)$ from U to $\mathbb{R}^{k-r}$, with its differential at $\mathbf{a}$, $f^{'}(\mathbf{a})$, being a surjective linear operator, which is equivalent to say that $f^{'}(\mathbf{a})$ has full rank equal to $k - r$.

The mapping $f(.)$ is called a local defining function for $\mathcal{M}$ at $\mathbf{a}$. This definition implies that, locally around $\mathbf{a}$, a smooth embedded submanifold (of dimension r) looks like a subspace of dimension r of $\mathbb{R}^k$, which is also the dimension of the kernel of $f^{'}(\mathbf{a})$, see Theorem 3.12 of Boumal [11] or Theorem 2.1.10 of Robbin and Salomon [164] for details. More precisely, if $\mathcal{M}$ is a smooth submanifold of $\mathbb{R}^k$, it admits a tangent space noted $\mathcal{T}_{\mathbf{a}}\mathcal{M}$, which is nothing else than the kernel of $f^{'}(\mathbf{a})$, where $f(.)$ is any local defining function for $\mathcal{M}$ at $\mathbf{a}$ (see Theorem 3.15 of Boumal [11] or Theorem 2.2.3 of Robbin and Salomon [164]) and this tangent space can be interpreted as a vector subspace of $\mathbb{R}^k$ that approximates the smooth submanifold locally. Thus, a smooth (sub)manifold of dimension r is defined as a set that locally looks like a r-dimensional space, but can be very different globally.

Next, we clarify what we call a tangent vector to an arbitrary subset C of a general vector space $\mathcal{X}$ at a point $\mathbf{a} \in C$ in the following definition, which will also be sufficient for our purpose:

Definition 2.5. Let $\mathcal{X}$ be a normed vector space and C a nonempty subset of $\mathcal{X}$. A vector $\mathbf{d} \in \mathcal{X}$ is a tangent vector to C at $\mathbf{a} \in C$, if and only if, it exists $\alpha > 0$ and a mapping $\varepsilon(.)$ from $[-\alpha, \alpha]$ to $\mathcal{X}$ such that

$$\mathbf{a} + t.\mathbf{d} + t.\varepsilon(t) \in C, \forall t \in [-\alpha, \alpha], \text{ and } \lim_{t \to 0} \varepsilon(t) = 0 \ ,$$

or, equivalently, if it exists an open interval I of $\mathbb{R}$ containing $t = 0$ and a function $\mathcal{E} : I \to C$ such that $\mathcal{E}(.)$ is derivable at $t = 0$ with $\mathbf{d} = \mathcal{E}^{'}(0)$ and $\mathcal{E}(0) = \mathbf{a}$.

The set of all tangent vectors to C at $\mathbf{a} \in C$ is noted $\mathcal{T}_{\mathbf{a}}C$. If $\mathcal{T}_{\mathbf{a}}C$ is a linear subspace of $\mathcal{X}$, it is called the tangent space to C at $\mathbf{a}$.

Note that, in this definition, it is only required that the function $\mathcal{E}(.)$ is derivable at $t = 0$, not on all I and this definition is sufficient for many results stated in [11] or [164] for a C^p or C^∞ differentiable function $\mathcal{E}(.)$ on all I. Furthermore, keep in mind that if $\mathcal{M}$ is a C^p embedded submanifold in the sense of Definition 2.4, all the elements of its tangent space at a given point $\mathbf{a} \in \mathcal{M}$, defined as the kernel of $f^{'}(\mathbf{a})$ for any given local defining function $f(.)$ for $\mathcal{M}$ at $\mathbf{a}$, verify Definition 2.5 and the terminology is thus consistent [11][164].

Let $\mathcal{M}$ be a C^p embedded submanifold of dimension r of $\mathbb{R}^k$ in the sense of Definition (2.4). If we now endow $\mathbb{R}^k$ with its standard Euclidean inner product, $\mathcal{T}_{\mathbf{a}}\mathcal{M}$, which is a linear subspace of $\mathbb{R}^k$, admits an orthogonal supplementary subspace in $\mathbb{R}^k$, which is called the normal space of $\mathcal{M}$ at $\mathbf{a}$ and is denoted by $\mathcal{N}_{\mathbf{a}}\mathcal{M}$ in the following. Both $\mathcal{T}_{\mathbf{a}}\mathcal{M}$ and $\mathcal{N}_{\mathbf{a}}\mathcal{M}$ are linear subspaces of $\mathbb{R}^k$ and we have the identity: $\mathbb{R}^k = \mathcal{T}_{\mathbf{a}}\mathcal{M} \oplus \mathcal{N}_{\mathbf{a}}\mathcal{M}$, which is equivalent to say that any vector of $\mathbb{R}^k$ can be written uniquely as the sum of an element of $\mathcal{T}_{\mathbf{a}}\mathcal{M}$ and an element of $\mathcal{N}_{\mathbf{a}}\mathcal{M}$.

Suppose now that we want to minimize a C^p smooth function $\phi(.)$ from a smooth submanifold $\mathcal{M} \subset \mathbb{R}^k$ to $\mathbb{R}$. To define and also analyze Riemannian optimization methods on $\mathcal{M}$ for solving this kind of problems, we need to define the notions of the Riemannian gradient and Hessian, which will be obviously different from their Euclidean analogs as $\mathcal{M}$ is only locally homeomorphic to an Euclidean vector space. First, similarly to the standard case of a differentiable function from an open set U to $\mathbb{R}$, the smooth function $\phi(.)$ admits a differential at $\mathbf{a} \in \mathcal{M}$, which is a linear mapping from $\mathcal{T}_{\mathbf{a}}\mathcal{M}$ to $\mathbb{R}$ denoted also by $\phi^{'}(\mathbf{a})$ [3][164][11]. If, $\forall \mathbf{a} \in \mathcal{M}$, we equip $\mathcal{T}_{\mathbf{a}}\mathcal{M}$ with the standard Euclidean inner product induced by $\mathbb{R}^k$, e.g.,

$$\langle ., . \rangle_{\mathcal{T}_{\mathbf{a}}\mathcal{M}} = \langle ., . \rangle_2, \forall \mathbf{a} \in \mathcal{M} \ ,$$

$\mathcal{M}$ is then, by definition, equipped with a smoothly varied inner product on all its tangent spaces and $(\mathcal{M}, \langle ., . \rangle_2)$ is a Riemannian manifold [3][164][11]. In this setting, the Riemannian gradient of $\phi(.)$ at $\mathbf{a} \in \mathcal{M}$, denoted here by $\nabla_R \phi(\mathbf{a})$, is then defined as the unique vector in $\mathcal{T}_{\mathbf{a}}\mathcal{M}$ satisfying

$$\phi^{'}(\mathbf{a})(\mathbf{b}) = \langle \nabla_R \phi(\mathbf{a}), \mathbf{b} \rangle_{\mathcal{T}_{\mathbf{a}}\mathcal{M}} = \langle \nabla_R \phi(\mathbf{a}), \mathbf{b} \rangle_2 \ , \forall \mathbf{b} \in \mathcal{T}_{\mathbf{a}}\mathcal{M} \ ,$$

where $\phi^{'}(\mathbf{a})$ is the differential of the smooth mapping $\phi(.)$ at $\mathbf{a}$ in the sense defined above. We can also define the Riemannian Hessian of the smooth mapping $\phi(.)$ at $\mathbf{a}$, denoted by $\nabla^2_R\phi(\mathbf{a})$, which is a self-adjoint linear operator from $\mathcal{T}_{\mathbf{a}}\mathcal{M}$ to $\mathcal{T}_{\mathbf{a}}\mathcal{M}$ defined by

$$\nabla^2_R\phi(\mathbf{a})[\mathbf{b}] = \tilde{\nabla}_{\mathbf{b}}\nabla_R\phi(\mathbf{a}), \forall \mathbf{b} \in \mathcal{T}_{\mathbf{a}}\mathcal{M} \ ,$$

where $\tilde{\nabla}_{(.)}(.)$ denotes the so-called Levi-Civita connection on $\mathcal{M}$. The Levi-Civita connection $\tilde{\nabla}_{\eta_{\mathbf{a}}}\xi_{\mathbf{a}}$ on the Riemannian manifold $\mathcal{M}$ acting on two vector fields, $\eta_{\mathbf{a}}$ and $\xi_{\mathbf{a}}$, in the tangent bundle of $\mathcal{M}$ (the tangent bundle is the disjoint union of all the tangent spaces of the manifold $\mathcal{M}$, see Definition 3.42 in Boumal [11]) is a generalization of the notion of directional derivative of a vector field on the manifold $\mathcal{M}$. In this way, the Levi-Civita connection $\tilde{\nabla}_{\eta_{\mathbf{a}}}\xi_{\mathbf{a}}$ can be interpreted as the directional derivative of the vector field $\xi_{\mathbf{a}} \in \mathcal{T}_{\mathbf{a}}\mathcal{M}$ in the direction of $\eta_{\mathbf{a}} \in \mathcal{T}_{\mathbf{a}}\mathcal{M}$. Note further that the Riemannian gradient $\nabla_R\phi(.)$ defined for all $\mathbf{a} \in \mathcal{M}$ is a vector field from $\mathcal{M}$ to its tangent bundle and, in this condition, the Riemannian Hessian $\nabla^2_R\phi(\mathbf{a})[\mathbf{b}]$ can thus be interpreted as the directional derivative of the Riemannian gradient of $\phi(.)$ at $\mathbf{a} \in \mathcal{M}$ in the direction of $\mathbf{b} \in \mathcal{T}_{\mathbf{a}}\mathcal{M}$. See Section 3.5 of Boumal [11], Chapter 3 of Robbin and Salomon [164] or Section 5.3 of Absil et al. [3] for more information.

Furthermore, if $\phi(.)$ is a C^p smooth mapping, it can be extended to a C^p differentiable function $\hat{\phi}(.)$ on an open neighborhood U of $\mathbb{R}^k$ such that $\mathcal{M} \subset U$ (see Proposition 3.31 of Boumal [11]) and if, in addition, we equip the submanifold $\mathcal{M}$ with the Euclidean metric of the ambient linear space on all its tangent spaces, we have the following relationships between the Riemannian gradient and Hessian of $\phi(.)$ with the Euclidean gradient of $\hat{\phi}(.)$, respectively:

$$\nabla_R\phi(\mathbf{a}) = \mathbf{P}_{\mathcal{T}_{\mathbf{a}}\mathcal{M}}\big(\nabla\hat{\phi}(\mathbf{a})\big), \forall \mathbf{a} \in \mathcal{M} \ . \tag{2.48}$$

and

$$\begin{aligned}\nabla^2_R\phi(\mathbf{a})[\mathbf{b}] &= \mathbf{P}_{\mathcal{T}_{\mathbf{a}}\mathcal{M}}\Big(J\big(\nabla_R\phi(\mathbf{a})\big)[\mathbf{b}]\Big) \\ &= \mathbf{P}_{\mathcal{T}_{\mathbf{a}}\mathcal{M}}\Big(J\big(\mathbf{P}_{\mathcal{T}_{\mathbf{a}}\mathcal{M}}\nabla\hat{\phi}(\mathbf{a})\big)[\mathbf{b}]\Big), \forall \mathbf{a} \in \mathcal{M}, \forall \mathbf{b} \in \mathcal{T}_{\mathbf{a}}\mathcal{M} \ ,\end{aligned} \tag{2.49}$$

where $\mathbf{P}_{\mathcal{T}_{\mathbf{a}}\mathcal{M}}$ denotes the orthogonal projector operator onto $\mathcal{T}_{\mathbf{a}}\mathcal{M}$ in $\mathbb{R}^k$ and $J\big(\nabla_R\phi(\mathbf{a})\big)$ is the usual Euclidean derivative (e.g., Jacobian matrix operator) of the Riemannian gradient of $\phi(.)$ at $\mathbf{a}$. In words, if the metric on $\mathcal{M}$ is inherited from the ambient Euclidean space, the Riemannian gradient is just the tangent space projection of the embedded gradient in the ambient space and the Levi-Civita connection on $\mathcal{M}$ is the tangent space projection of the Levi-Civita connection on the ambient space, which is equivalent to the Euclidean (directional) derivative.

Alternatively, again in the case of an embedded submanifold, the Riemannian Hessian of $\phi(.)$ can be defined by means of so-called second-order retractions, which are second-order approximations of the exponential map, see Propositions 5.5.4 and 5.5.5 in [3] and Proposition 3 in [1] for details. This also allows to derive the Riemannian Hessian of a cost function defined on an embedded submanifold in terms of standard Euclidean derivatives as in equation (2.49). See Appendix A of [185] for an illustration with the Riemannian Hessian of the the cost function $\varphi(.)$ used in the formulation (P0) of the WLRA problem in the case of binary weights and also Proposition 2 in [116] for a generalization to an arbitrary twice differentiable cost function $\varphi(.)$ defined on the smooth matrix submanifold $\mathbb{R}^{p\times n}_k$ embedded in $\mathbb{R}^{p\times n}$. These results are useful in our WLRA context and will be used later, see equation (3.9) in Subsection 3.2.

Finally, the first- and second-order stationary conditions for a C^p smooth real function $\phi(.)$ defined on a submanifold $\mathcal{M} \subset \mathbb{R}^k$ are exactly similar to their standard Euclidean counterparts when the search space is reduced to a linear subspace embedded in $\mathbb{R}^k$ [89] : a vector $\hat{\mathbf{a}} \in \mathcal{M}$ is a first-order critical point for $\phi(.)$ if the vector $\nabla_R\phi(\hat{\mathbf{a}}) \in \mathcal{T}_{\hat{\mathbf{a}}}\mathcal{M}$ is equal to the zero-vector. Using equation (2.48), this is equivalent to say that the usual Euclidean gradient of the differentiable extension $\hat{\phi}(.)$ at $\hat{\mathbf{a}}$, $\nabla\hat{\phi}(\hat{\mathbf{a}})$, is orthogonal to $\mathcal{T}_{\hat{\mathbf{a}}}\mathcal{M}$, e.g., that $\nabla\hat{\phi}(\hat{\mathbf{a}}) \in \mathcal{N}_{\hat{\mathbf{a}}}\mathcal{M}$. Thus, $\hat{\mathbf{a}} \in \mathcal{M}$ is a

first-order stationary point of the C^p smooth real function $\phi(.)$ if one of the following equivalent conditions are satisfied

$$\nabla\hat{\phi}(\hat{\mathbf{a}}) \in \mathcal{N}_{\hat{\mathbf{a}}}\mathcal{M} \iff \mathbf{P}_{\mathcal{T}_{\hat{\mathbf{a}}}\mathcal{M}}\big(\nabla\hat{\phi}(\hat{\mathbf{a}})\big) = \mathbf{0}^k \iff \|\mathbf{P}_{\mathcal{T}_{\hat{\mathbf{a}}}\mathcal{M}}\big(\nabla\hat{\phi}(\hat{\mathbf{a}})\big)\|_2 = 0\,, \tag{2.50}$$

where $\hat{\phi}(.)$ is a differentiable extension in the ambient linear space of the smooth function $\phi(.)$ at $\hat{\mathbf{a}}$. On the other hand, a vector $\hat{\mathbf{a}} \in \mathcal{M}$ is a second-order critical point for $\phi(.)$ if it is a first-order critical point for $\phi(.)$ and if, in addition, the self-adjoint operator $\nabla^2_R\phi(\hat{\mathbf{a}})$ defines a (symmetric) positive semi-definite bilinear form on $\mathcal{T}_{\hat{\mathbf{a}}}\mathcal{M}$. Finally, a vector $\hat{\mathbf{a}} \in \mathcal{M}$ is a strict (local) minimum of the smooth real function $\phi(.)$ if it is a first-order critical point for $\phi(.)$ and if, in addition, the self-adjoint operator $\nabla^2_R\phi(\hat{\mathbf{a}})$ defines a (symmetric) positive definite bilinear form on $\mathcal{T}_{\hat{\mathbf{a}}}\mathcal{M}$.

In the following, we will also be concerned with the minimization of a smooth real mapping $\phi(.)$ defined over a smooth submanifold $\mathcal{M} \subset \mathbb{R}^k$ (or $\mathcal{M} \subset \mathbb{R}^{p\times k}$), where $\phi(.)$ is invariant under the action of a certain group $\mathcal{G}$, which allows us to define an equivalence relation $\sim$ in the total computational space $\mathcal{M}$. In these conditions, all the elements of a given equivalence class of $\sim$ have the same value for $\phi(.)$. The quotient $\mathcal{M}/\sim$ generated by this equivalence relation consists of elements that are equivalence classes. If $\mathring{\mathbf{a}} \in \mathcal{M}/\sim$ then its vector representation in $\mathcal{M}$ is $\mathbf{a}$. Because of the invariance property, we want to minimize $\phi(.)$ over the set of equivalence classes $\mathcal{M}/\sim$ instead on $\mathcal{M}$. This leads to the notion of smooth and Riemannian quotient manifolds if some conditions on the group $\mathcal{G}$ are satisfied [3][11]. An important example of quotient manifolds is the Grassmann manifold which is the collection of all linear subspaces of a given dimension k in a particular Euclidean space of dimension $n > k$ and is denoted by $\mathrm{Gr}(n,k)$; see Chapter 9 of [11] for a comprehensive overview of general quotient manifolds and $\mathrm{Gr}(n,k)$. More precisely, each of these linear subspaces can be represented by a $n \times k$ matrix of rank k whose columns form a basis of this given subspace and all the $n \times k$ matrices of rank k which are associated with the same subspace of rank k form obviously an equivalence class, which can be identified with each subspace of rank k embedded in the Euclidean space of dimension n. See Section 3 for more concrete examples of Grassmann manifolds in the context of the WLRA problem.

On such smooth quotient manifolds, the concept of tangent space to the quotient manifold $\mathcal{M}/\sim$ at $\mathring{\mathbf{a}} \in \mathcal{M}/\sim$ can be also defined and this abstract tangent space will be denoted by $\mathcal{T}_{\mathring{\mathbf{a}}}\mathcal{M}/\sim$ or simply by $\mathcal{T}_{\mathring{\mathbf{a}}}\mathcal{M}$ by an abuse of notation. Furthermore, the notions of Riemannian gradient and Hessian of the smooth mapping $\phi(.)$ defined (again with a slight abuse of notation) on $\mathcal{M}/\sim$ and such that $\phi(\mathring{\mathbf{a}}) = \phi(\mathbf{a}), \forall \mathring{\mathbf{a}} \in \mathcal{M}/\sim$ with $\mathbf{a} \in \mathring{\mathbf{a}} \subset \mathcal{M}$, can be extended. First- and second-order optimality conditions of $\phi(.)$ for an element of $\mathcal{M}/\sim$ can also be formulated. More detailed information on the related backgrounds can be found in Section 3.4 of Absil et al. [3] or in Section 9.8 of Boumal [11]. Comprehensive introduction to these abstract notions are also provided in [130][132][133][14]. Fortunately, when $\mathcal{M}$ is an embedded submanifold of $\mathbb{R}^k$ (or $\mathbb{R}^{p\times k}$) and inherits of the Euclidean (or Frobenius) metric of the ambient linear space, each abstract element of $\mathcal{T}_{\mathring{\mathbf{a}}}\mathcal{M}/\sim$ (where $\mathring{\mathbf{a}} \in \mathcal{M}/\sim$ and $\mathbf{a} \in \mathcal{M}$) can be uniquely represented by an element of the tangent space $\mathcal{T}_{\mathbf{a}}\mathcal{M}$ whose direction in the total space $\mathcal{M}$ does not induce a displacement (from $\mathbf{a}$) along the equivalence class $\mathring{\mathbf{a}}$. This is achieved by decomposing the tangent space $\mathcal{T}_{\mathbf{a}}\mathcal{M}$ to the total space $\mathcal{M}$ at $\mathbf{a}$ in the following complementary and orthogonal direct sum

$$\mathcal{T}_{\mathbf{a}}\mathcal{M} = \mathcal{H}_{\mathbf{a}}\mathcal{M} \oplus \mathcal{V}_{\mathbf{a}}\mathcal{M}\,,$$

where $\mathcal{H}_{\mathbf{a}}\mathcal{M}$ and $\mathcal{V}_{\mathbf{a}}\mathcal{M}$ are orthogonal (with respect to the inner product of the ambient linear space) linear subspaces of $\mathcal{T}_{\mathbf{a}}\mathcal{M}$. $\mathcal{V}_{\mathbf{a}}\mathcal{M}$ is called the vertical space of $\mathcal{M}$ at $\mathbf{a}$ and is the set of tangent vectors to $\mathcal{M}$ at $\mathbf{a}$, which do induce a displacement along the equivalence class $\mathring{\mathbf{a}}$. The horizontal space $\mathcal{H}_{\mathbf{a}}\mathcal{M}$ is the orthogonal complement of $\mathcal{V}_{\mathbf{a}}\mathcal{M}$ and provides a valid and one-to-one representation of the abstract tangent vectors to the quotient space $\mathcal{M}/\sim$ at $\mathring{\mathbf{a}}$; see Section 9.4 of [11] for more information. Displacements in the vertical space leave the vector $\mathbf{a}$, representing the equivalence class $\mathring{\mathbf{a}}$, unchanged. This justifies to restrict both tangent vectors and metric to the horizontal space $\mathcal{H}_{\mathbf{a}}\mathcal{M}$ [3][11].

Provided that the inherited Euclidean metric defined in the total space $\mathcal{M}$ is invariant along the equivalence classes in $\mathcal{M}/\sim$, the quotient space $\mathcal{M}/\sim$ endowed with this (Riemannian) metric is called a Riemannian quotient manifold of $\mathcal{M}$ [3][11]. For such Riemannian quotient manifold $\mathcal{M}/\sim$ whose total space $\mathcal{M}$ is a submanifold embedded in a Euclidean space, we can then obtain convenient practical representations for the abstract Riemannian gradient and Hessian of $\mathring{\phi}(.)$ (defined on $\mathcal{M}/\sim$) at $\mathring{\mathbf{a}}$ by simply replacing the tangent space $\mathcal{T}_{\mathbf{a}}\mathcal{M}$ by its horizontal space $\mathcal{H}_{\mathbf{a}}\mathcal{M}$ in expressions (2.48) and (2.49):

$$\nabla_R \phi(\mathring{\mathbf{a}}) \simeq \mathbf{P}_{\mathcal{H}_{\mathbf{a}}\mathcal{M}}\big(\nabla \hat{\phi}(\mathbf{a})\big), \forall \mathbf{a} \in \mathcal{M} \; . \tag{2.51}$$

and

$$\nabla^2_R \phi(\mathring{\mathbf{a}})[\bar{\mathbf{b}}] \simeq \mathbf{P}_{\mathcal{H}_{\mathbf{a}}\mathcal{M}}\Big(J\big(\mathbf{P}_{\mathcal{H}_{\mathbf{a}}\mathcal{M}} \nabla \hat{\phi}(\mathbf{a})\big)[\mathbf{b}]\Big), \forall \mathbf{a} \in \mathcal{M}, \forall \mathbf{b} \in \mathcal{H}_{\mathbf{a}}\mathcal{M} \; , \tag{2.52}$$

where $\mathbf{P}_{\mathcal{H}_{\mathbf{a}}\mathcal{M}}$ denotes now the orthogonal projector operator onto $\mathcal{H}_{\mathbf{a}}\mathcal{M}$ in the ambient linear space $\mathbb{R}^k$, $\bar{\mathbf{b}}$ is an abstract tangent vector of the quotient manifold $\mathcal{M}/\sim$ at $\mathring{\mathbf{a}} \in \mathcal{M}/\sim$, which is uniquely represented by the so-called horizontal lift $\mathbf{b} \in \mathcal{H}_{\mathbf{a}}\mathcal{M}$, and $\hat{\phi}(.)$ is a C^p differentiable extension of the smooth function $\phi(.)$ defined on $\mathcal{M}$ to an open neighborhood U of $\mathbb{R}^k$ (or of $\mathbb{R}^{p \times k}$) such that $\mathcal{M} \subset U$. See Mishra et al. [130][132][133] or Boumal and Absil [14] for concrete illustrations of these abstract objects in the context of the WLRA problem. Importantly, the first- and second-order critical conditions for $\phi(.)$ on the quotient manifold $\mathcal{M}/\sim$ can now be expressed and evaluated concretely in terms of $\mathbf{P}_{\mathcal{H}_{\mathbf{a}}\mathcal{M}}\big(\nabla \hat{\phi}(\mathbf{a})\big)$ and $\mathbf{P}_{\mathcal{H}_{\mathbf{a}}\mathcal{M}}\Big(J\big(\mathbf{P}_{\mathcal{H}_{\mathbf{a}}\mathcal{M}} \nabla \hat{\phi}(\mathbf{a})\big)[\mathbf{b}]\Big)$ as for a "standard" submanifold embedded in a Euclidean linear space. As an illustration, the first-order stationary condition for $\mathring{\mathbf{a}} \in \mathcal{M}/\sim$ becomes

$$\mathbf{P}_{\mathcal{H}_{\mathbf{a}}\mathcal{M}}\big(\nabla \hat{\phi}(\mathbf{a})\big) = \mathbf{0}^k \iff \|\mathbf{P}_{\mathcal{H}_{\mathbf{a}}\mathcal{M}}\big(\nabla \hat{\phi}(\mathbf{a})\big)\|_2 = 0 \; . \tag{2.53}$$

The minimization of a real function $\phi(.)$ over a nonempty (arbitrary) set $\mathcal{K} \subset \mathbb{R}^k$ (or more generally over a subset of a Euclidean or Frobenius linear space) is more involved than solving the same problem over the whole linear space $\mathbb{R}^k$, or over one of its linear subspaces or one of its embedded smooth submanifolds described above. The first difficulty comes up in characterizing the optimality of feasible solutions itself, e.g., the necessary first- and second-order conditions for a point $\hat{\mathbf{a}} \in \mathcal{K}$ to be a local minimizer of $\phi(.)$ over $\mathcal{K}$. Here, the feasible set $\mathcal{K}$ and its topological properties play a role as important as the properties of the function $\phi(.)$ itself as it is first necessary to characterize which search directions are admissible around $\hat{\mathbf{a}} \in \mathcal{K}$. It is well known now that these admissible directions are related to the notions of tangent and normal cones to the set $\mathcal{K}$ at $\hat{\mathbf{a}}$, see Chapter 6 of Rockafellar and Wets [165] and also Ruszczynski [160]. Moreover, minimizing an even C^∞ differentiable real function $\phi(.)$ over a nonempty subset $\mathcal{K} \subset \mathbb{R}^k$ leads to different and confusing notions of stationarity [165][84][112][144][151].

We now introduce the required elements of variational geometry to characterize the first- and second-order stationarity conditions of a possible local solution $\hat{\mathbf{a}}$ to the minimization of $\phi(.)$ over a nonempty (arbitrary) set $\mathcal{K} \subset \mathbb{R}^k$.

Definition 2.6. A subset $\mathcal{C} \subset \mathbb{R}^k$ is called a cone if it contains the zero-vector and contains with each of its vectors, all positive multiples of that vector, e.g., if $\mathbf{a} \in \mathcal{C} \Rightarrow \lambda.\mathbf{a} \in \mathcal{C}$, $\forall \lambda \in \mathbb{R}_{+*}$.

As an illustration, the set consisting of a nonzero-vector $\mathbf{a} \in \mathbb{R}^k$ and all of its positive multiples $\lambda.\mathbf{a}$ (with $\lambda \geq 0$) is a particular cone, which is called a ray. In other words, a cone, which is distinct from $\{\mathbf{0}^k\}$, is therefore composed of the union of the rays it contains.

Next, if $\mathcal{K}$ is a nonempty subset of $\mathbb{R}^k$, the (dual) polar of $\mathcal{K}$, noted $\mathcal{K}^o$, is the set

$$\mathcal{K}^o := \Big\{ \mathbf{b} \in \mathbb{R}^k \,/\, \langle \mathbf{a}, \mathbf{b} \rangle_2 \leq 0 \; , \forall \mathbf{a} \in \mathcal{K} \Big\} \; . \tag{2.54}$$

First of all, we see that the polar of $\mathcal{K}$ depends on the scalar product used in $\mathbb{R}^k$, if we changed this scalar product then $\mathcal{K}^o$ is also changed. Geometrically, $\mathcal{K}^o$ is the set of all vectors in $\mathbb{R}^k$, which

have an angle of at least 90° with every vector in $\mathcal{K}$. Next, note that $\mathcal{K}^o$ is a cone, as it obviously contains $\mathbf{0}^k$, but also $\lambda.\mathbf{b}$ for any $\lambda \geq 0$ if $\mathbf{b} \in \mathcal{K}^o$. $\mathcal{K}^o$ is further convex and closed in $\mathbb{R}^k$ as the above definition of $\mathcal{K}^o$ expresses $\mathcal{K}^o$ as the intersection of a family of closed half-spaces, which are also all convex:

$$\mathcal{K}^o = \bigcap_{\mathbf{a} \in \mathcal{K}} \left\{ \mathbf{b} \in \mathbb{R}^k \,/\, \langle \mathbf{a}, \mathbf{b} \rangle_2 \leq 0 \right\} .$$

If $\mathcal{K}$ is a nonempty subset of $\mathbb{R}^k$, its orthogonal complement is the set:

$$\mathcal{K}^{\perp} := \mathcal{K}^o \cap (-\mathcal{K})^o = \left\{ \mathbf{b} \in \mathbb{R}^k \,/\, \langle \mathbf{a}, \mathbf{b} \rangle_2 = 0 \,, \forall \mathbf{a} \in \mathcal{K} \right\} . \tag{2.55}$$

We deduce immediately that $\mathcal{K}^{\perp}$ is a closed convex cone of $\mathbb{R}^k$ as the intersection of two closed convex cones. Obviously, $\mathcal{K}^{\perp}$ is also a linear subspace of $\mathbb{R}^k$. Interestingly, if $\mathcal{K}$ is a linear subspace of $\mathbb{R}^k$, we have $\mathcal{K} = -\mathcal{K}$ and, consequently, $\mathcal{K}^{\perp} = \mathcal{K}^o$. Thus, polarity generalises the notion of orthogonality between linear subspaces discussed in Subsection 2.1 to arbitrary nonempty subsets of $\mathbb{R}^k$. If $\mathcal{K}_1$ and $\mathcal{K}_2$ are two nonempty cones of $\mathbb{R}^k$, we have

$$(\mathcal{K}_1 \cup \mathcal{K}_2)^o = (\mathcal{K}_1 + \mathcal{K}_2)^o = \mathcal{K}_1^o \cap \mathcal{K}_2^o \,.$$

In addition, if $\mathcal{K}_1$ and $\mathcal{K}_2$ are two closed convex cones then

$$(\mathcal{K}_1 \cap \mathcal{K}_2)^o = \mathcal{K}_1^o + \mathcal{K}_2^o \,,$$

and, finally, the property $\mathcal{K}^{oo} = \mathcal{K}$ is true if $\mathcal{K}$ is a closed convex cone. See Deutsch [39] for more details on (convex) cones and their polars.

We now introduce the general concepts of tangent and normal vectors at a nonempty set $\mathcal{K} \subset \mathbb{R}^k$, which generalize the notions of tangent and normal vectors at a smooth submanifold of $\mathbb{R}^k$ introduced above, following [165]; see also [79] or [160] for a more gentle introduction to these concepts.

Definition 2.7. For a nonempty set $\mathcal{K} \subset \mathbb{R}^k$ and a point $\bar{\mathbf{a}} \in \mathcal{K}$, a vector $\mathbf{d} \in \mathbb{R}^k$ is said to be tangent to $\mathcal{K}$ at $\bar{\mathbf{a}}$, when there exists a sequence $(\mathbf{a}_k)_{k \in \mathbb{N}_*}$ in $\mathcal{K}$ tending to $\bar{\mathbf{a}}$ and a sequence $(\mathbf{t}_k)_{k \in \mathbb{N}_*}$ in $\mathbb{R}_{+*}$ tending to zero (e.g., decreasing to zero) such that the vectors $\mathbf{b}_k = \frac{(\mathbf{a}_k - \bar{\mathbf{a}})}{\mathbf{t}_k}$ tend to $\mathbf{d}$, e.g., if

$$\lim_{k \to \infty} \frac{(\mathbf{a}_k - \bar{\mathbf{a}})}{\mathbf{t}_k} = \mathbf{d} \,.$$

Note that, if $\lim_{k \to \infty} \frac{(\mathbf{a}_k - \bar{\mathbf{a}})}{\mathbf{t}_k} = \mathbf{d}$, it is implicit that the sequence $(\mathbf{a}_k)_{k \in \mathbb{N}_*}$ tends to $\bar{\mathbf{a}}$, as otherwise the above limit does not exist as the sequence $(\mathbf{t}_k)_{k \in \mathbb{N}_*}$ tends to zero. Consequently, some authors define a tangent vector without the condition that the sequence $(\mathbf{a}_k)_{k \in \mathbb{N}_*}$ tends to $\bar{\mathbf{a}}$. Furthermore, different, but equivalent, definitions of a tangent vector are also used in the literature, see Guignard [60], Equation 2.2 of Schneider and Uschmajew [173] and Section 5.1 of Hiriart-Urruty and Le Marechal [79] for details.

This new definition of tangency generalizes the classical Definition 2.5 in which a tangent vector $\mathbf{d}$ to $\mathcal{K}$ at $\bar{\mathbf{a}}$ is the derivative at $\bar{\mathbf{a}}$ of some curve drawn on $\mathcal{K}$. This classical definition is not relevant here as $\mathcal{K}$ can be a subset of $\mathbb{R}^k$ of discrete type and also because half-derivatives are key here instead of full-derivatives as in standard differential geometry.

We observe immediately that $\mathbf{0}^k$ is always a tangent vector at $\mathcal{K}$ for any $\bar{\mathbf{a}} \in \mathcal{K}$: it suffices to take $\mathbf{a}_k = \bar{\mathbf{a}}, \forall k \in \mathbb{N}_*$. Furthermore, if $\mathbf{d}$ is a tangent vector to $\mathcal{K}$ at $\bar{\mathbf{a}}$, then $\alpha.\mathbf{d}$ for $\alpha > 0$ is also a tangent vector to $\mathcal{K}$ at $\bar{\mathbf{a}}$ since it suffices to change $\mathbf{t}_k$ to $\frac{\mathbf{t}_k}{\alpha}$ in the Definition 2.7 of a tangent vector. In other words, the set of all tangent vectors to $\mathcal{K}$ at $\bar{\mathbf{a}}$ in the sense of Definition 2.7 is a cone. The next theorem further shows that the set of all tangent vectors to $\mathcal{K}$ at $\bar{\mathbf{a}}$ is in fact a closed cone, which is called the tangent cone (or the contingent or Bouligand's cone) to $\mathcal{K}$ at $\bar{\mathbf{a}}$ and is denoted by $\mathcal{T}_{\bar{\mathbf{a}}}^{\mathcal{B}} \mathcal{K}$.

Theorem 2.4. Let $\mathcal{K}$ be a nonempty subset of $\mathbb{R}^k$ and let $\bar{\mathbf{a}} \in \mathcal{K}$. The set $\mathcal{T}_{\bar{\mathbf{a}}}^{\mathcal{B}}\mathcal{K}$ of all tangent directions for $\mathcal{K}$ at $\bar{\mathbf{a}}$ in the sense of Definition 2.7 is a closed cone.

Proof. Omitted. See Lemma 3.12 of [160] or Proposition 5.1.3 of [79]. □

Furthermore, it is not difficult to see that if $\bar{\mathbf{a}}$ is an interior point of $\mathcal{K}$ (e.g., $\bar{\mathbf{a}} \in \mathring{\mathcal{K}}$), we have $\mathcal{T}_{\bar{\mathbf{a}}}^{\mathcal{B}}\mathcal{K} = \mathbb{R}^k$. Thus, "the interesting" points are those on $bd(\mathcal{K})$, the boundary of $\mathcal{K}$. We next define the notion of normal vectors or directions to a set $\mathcal{K}$ in the regular sense following [165]:

Definition 2.8. For a nonempty set $\mathcal{K} \subset \mathbb{R}^k$ and a point $\bar{\mathbf{a}} \in \mathcal{K}$, a vector $\mathbf{d} \in \mathbb{R}^k$ is said to be normal to $\mathcal{K}$ at $\bar{\mathbf{a}}$ in the regular sense, or a regular normal, if

$$\langle \mathbf{d}, \mathbf{a} - \bar{\mathbf{a}} \rangle_2 \leq o(\|\mathbf{a} - \bar{\mathbf{a}}\|_2) \, , \forall \mathbf{a} \in \mathcal{K} \, ,$$

where we denote by $o(\|\mathbf{a} - \bar{\mathbf{a}}\|_2)$, for $\mathbf{a} \in \mathcal{K}$, a term with the property that $\frac{o(\|\mathbf{a}-\bar{\mathbf{a}}\|_2)}{\|\mathbf{a}-\bar{\mathbf{a}}\|_2}$ tends to zero when $\mathbf{a}$ tends to $\bar{\mathbf{a}}$ in $\mathcal{K}$, with $\mathbf{a} \neq \bar{\mathbf{a}}$.

The set of normal vectors to $\mathcal{K}$ at $\bar{\mathbf{a}}$ in the regular sense is called the Frechet normal cone to $\mathcal{K}$ at $\bar{\mathbf{a}}$ and is denoted by $\mathcal{N}_{\bar{\mathbf{a}}}^{\mathcal{F}}\mathcal{K}$.

This name is justified by the following result, which provides a more comprehensive interpretation of the set of normal vectors in the regular sense to $\mathcal{K}$ at $\bar{\mathbf{a}}$.

Theorem 2.5. Let $\mathcal{K}$ be a nonempty subset of $\mathbb{R}^k$ and let $\bar{\mathbf{a}} \in \mathcal{K}$. The set $\mathcal{N}_{\bar{\mathbf{a}}}^{\mathcal{F}}\mathcal{K}$ of all regular normal vectors is characterized by

$$\mathbf{d} \in \mathcal{N}_{\bar{\mathbf{a}}}^{\mathcal{F}}\mathcal{K} \Longleftrightarrow \langle \mathbf{d}, \mathbf{a} \rangle_2 \leq 0 \, , \forall \mathbf{a} \in \mathcal{T}_{\bar{\mathbf{a}}}^{\mathcal{B}}\mathcal{K} \, .$$

In other words, we have $\mathcal{N}_{\bar{\mathbf{a}}}^{\mathcal{F}}\mathcal{K} = (\mathcal{T}_{\bar{\mathbf{a}}}^{\mathcal{B}}\mathcal{K})^o$ and the Frechet normal cone to $\mathcal{K}$ at $\bar{\mathbf{a}}$ is the polar of the Bouligand tangent cone to $\mathcal{K}$ at $\bar{\mathbf{a}}$ and is, thus, a closed convex cone.

Proof. Omitted. See Propostion 6.5 in Rockafellar and Wets [165]. □

Thus, the normal vectors to $\mathcal{K}$ at $\bar{\mathbf{a}}$ in the regular sense, apart from $\mathbf{0}^k$, are simply the vectors $\mathbf{d}$ of $\mathbb{R}^k$ that make a right or obtuse angle with every tangent vector $\mathbf{a}$ to $\mathcal{K}$ at $\bar{\mathbf{a}}$. Importantly, if the subset $\mathcal{K}$ is an embedded submanifold of $\mathbb{R}^k$, the Bouligand tangent and Frechet normal cones to $\mathcal{K}$ at $\bar{\mathbf{a}}$ reduce, respectively, to the tangent and normal spaces to $\mathcal{K}$ at $\bar{\mathbf{a}}$ [165], e.g.,

$$\mathcal{T}_{\bar{\mathbf{a}}}^{\mathcal{B}}\mathcal{K} = \mathcal{T}_{\bar{\mathbf{a}}}\mathcal{K} \text{ and } \mathcal{N}_{\bar{\mathbf{a}}}^{\mathcal{F}}\mathcal{K} = \mathcal{N}_{\bar{\mathbf{a}}}\mathcal{K} \, .$$

Thus, in a sense, the notions of Bouligand tangent and Frechet normal cones generalize the concepts of tangent and normal spaces to a smooth submanifold, described above, to an arbitrary nonempty set $\mathcal{K}$ embedded in a given Euclidean vector or Frobenius matrix space. Furthermore, we will see now that the first- and second-order optimality conditions for mimimizing a real function $\phi(.)$ over $\mathcal{K}$ can also be interpreted as an extension of the first- and second-order optimality conditions required over a smooth submanifold discussed above.

The motivation and interest for the above paragraphs about cones, tangent and normal directions are related to this task and come from the following Theorem 2.6, which provides a first basic first-order necessary condition for a vector $\hat{\mathbf{a}}$ to be a solution of the minimization of a real function $\phi(.)$ over a nonempty (arbitrary) subset $\mathcal{K} \subset \mathbb{R}^k$ (or more generally a subset of a normed vector space of finite dimension).

To be more precise, consider a nonempty set $\mathcal{K} \subset \mathbb{R}^k$, a differentiable function $\phi(.) : \Omega \longrightarrow \mathbb{R}$, where Ω is open in $\mathbb{R}^k$ and such that $\mathcal{K} \subset \Omega$, and the constrained optimization problem $\min_{\mathbf{a} \in \mathcal{K}} \phi(\mathbf{a})$. Note that we don't assume here that $\mathcal{K}$ is open, so if the constrained problem has a (local) solution $\hat{\mathbf{a}}$, this solution $\hat{\mathbf{a}}$ can be a boundary point of the feasible set $\mathcal{K}$, in which case the necessary conditions of optimality formulated above in equations (2.44) and (2.45) do not have

to be satisfied because the perturbations $d\mathbf{a}$ to the vector $\hat{\mathbf{a}}$ such that $\hat{\mathbf{a}} + d\mathbf{a} \notin \mathcal{K}$ do not have to be taken into account and therefore they may correspond to a decrease of the cost function $\phi(.)$. In order to obtain a correct first-order necessary condition for optimality in a such case, the next theorem shows that we can restrict the set of possible perturbations $d\mathbf{a}$ to the tangent directions to $\mathcal{K}$ at $\hat{\mathbf{a}}$ in the sense of Definition 2.7, e.g., to the elements of the Bouligand's cone to $\mathcal{K}$ at $\hat{\mathbf{a}}$.

Theorem 2.6. Let $\mathcal{K}$ be a nonempty subset of $\mathbb{R}^k$ and assume that $\phi(.)$ is a differentiable real function from an open subset Ω of $\mathbb{R}^k$ to $\mathbb{R}$ such that $\mathcal{K} \subset \Omega$. If $\phi(.)$ has a local minimum over $\mathcal{K}$ at $\hat{\mathbf{a}}$, then $\phi(.)$ has not descent vector $\mathbf{d} \in \mathcal{T}_{\hat{\mathbf{a}}}^{\mathcal{B}}\mathcal{K}$, i.e.,

$$\langle \nabla\phi(\hat{\mathbf{a}}), \mathbf{d} \rangle_2 \geq 0 \, , \forall \mathbf{d} \in \mathcal{T}_{\hat{\mathbf{a}}}^{\mathcal{B}}\mathcal{K} \, , \tag{2.56}$$

which is equivalent to say that

$$-\nabla\phi(\hat{\mathbf{a}}) \in \mathcal{N}_{\hat{\mathbf{a}}}^{\mathcal{F}}\mathcal{K} = (\mathcal{T}_{\hat{\mathbf{a}}}^{\mathcal{B}}\mathcal{K})^o \, . \tag{2.57}$$

In words, if a vector $\hat{\mathbf{a}}$ is a local minimizer of $\phi(.)$ over $\mathcal{K}$, the anti-gradient $-\nabla\phi(\hat{\mathbf{a}})$ is a normal vector in the regular sense to $\mathcal{K}$ at $\hat{\mathbf{a}}$, which is equivalent to say that $-\nabla\phi(\hat{\mathbf{a}})$ is an element of the Frechet normal cone to $\mathcal{K}$ at $\hat{\mathbf{a}}$.

Proof. See Theorem 3.24 of Ruszczyinski [160], Theorem 6.12 of Rockafellar and Wets [165] or Theorem 1 of Guignard [60] for a proof. □

Thus, Theorem 2.6 and equation (2.57) provides a first-order optimality condition for the problem of minimizing $\phi(.)$ over $\mathcal{K}$ at a point $\hat{\mathbf{a}}$ and we will say that $\hat{\mathbf{a}}$ is a Frechet first-order stationarity point for this minimizaion problem if such condition is fulfilled. However, beware that other first-order optimality conditions have been proposed in the literature by replacing the Frechet normal cone $\mathcal{N}_{\hat{\mathbf{a}}}^{\mathcal{F}}\mathcal{K}$ in equation (2.57) by other cones like the so-called Mordukhovich or Clarke normal cones depending on the assumed properties for the function $\phi(.)$; see [165][84][116][151][144] for more information. However, if we only assume that $\phi(.)$ is a continuously differentiable or twice continuously differentiable function, the above Frechet stationarity provides the strongest necessary condition [116][144] and this is the first-order optimality condition we shall use in this monograph.

We now derive a more convenient expression to check that a given point $\hat{\mathbf{a}} \in \mathcal{K}$ is a Frechet first-order stationary point based on the notion of metric projection onto an arbitrary nonempty subset $\mathcal{K} \subset \mathbb{R}^k$ (or more generally a subset of an arbitrary normed vector space), which generalizes the concept of an orthogonal projection operator onto a linear subspace introduced in Subsection 2.1.

Let first $\mathcal{K}$ be a linear subspace of $\mathbb{R}^k$ and denote by $\text{Proj}_{\mathcal{K}}(.)$ the orthogonal projector mapping onto the subspace $\mathcal{K}$. $\text{Proj}_{\mathcal{K}}(.)$ is linear, idempotent ($\text{Proj}_{\mathcal{K}} o \text{Proj}_{\mathcal{K}} = \text{Proj}_{\mathcal{K}}$), non-expansive ($\|\text{Proj}_{\mathcal{K}}(\mathbf{a})\|_2 \leq \|\mathbf{a}\|_2 \, , \, \forall \mathbf{a} \in \mathbb{R}^k$) and it defines a direct sum of $\mathbb{R}^k$ as $\mathbf{a} = \text{Proj}_{\mathcal{K}}(\mathbf{a}) + \text{Proj}_{\mathcal{K}^\perp}(\mathbf{a}) \, , \, \forall \mathbf{a} \in \mathbb{R}^k$.

We now generalize this operator to the case where $\mathcal{K}$ is only a nonempty closed and, eventually, convex set in $\mathbb{R}^k$. We will also see that, if $\mathcal{K}$ is in addition a cone in the sense of Definition 2.6, almost all the above properties of an orthogonal projector can be conserved or extended to the metric projection operator. Let us first define precisely the metric projection operator with the following definition.

Definition 2.9. Let $\mathcal{K}$ be a nonempty subset of $\mathbb{R}^k$ and $\mathbf{a} \in \mathbb{R}^k$. An element $\mathbf{b} \in \mathcal{K}$ is called a nearest point to $\mathbf{a}$ from $\mathcal{K}$ if

$$\|\mathbf{a} - \mathbf{b}\|_2 = d(\mathbf{a}, \mathcal{K}) \, ,$$

where $d(\mathbf{a}, \mathcal{K}) := \inf_{\mathbf{d} \in \mathcal{K}} \|\mathbf{a} - \mathbf{d}\|_2$. The number $d(\mathbf{a}, \mathcal{K})$ always exists and is called the distance from $\mathbf{a}$ to $\mathcal{K}$. Next, the possibly empty, discrete or infinite set of all nearest points from $\mathbf{a}$ to $\mathcal{K}$ is denoted by $P_{\mathcal{K}}(\mathbf{a})$. In other words,

$$P_{\mathcal{K}}(\mathbf{a}) := \left\{ \mathbf{b} \in \mathcal{K} \, / \, \|\mathbf{a} - \mathbf{b}\|_2 = d(\mathbf{a}, \mathcal{K}) \right\} \, .$$

This defines a mapping $P_{\mathcal{K}}(.)$ from $\mathbb{R}^k$ to the subsets of $\mathcal{K}$ called the metric projection onto $\mathcal{K}$.

If each $\mathbf{a} \in \mathbb{R}^k$ has at least (respectively, exactly) one nearest point in $\mathcal{K}$, then $\mathcal{K}$ is called a proximinal (respectively, Chebyshev) set [39]. In other words, $\mathcal{K}$ is proximinal if $P_{\mathcal{K}}(\mathbf{a}) \neq \emptyset$, $\forall \mathbf{a} \in \mathbb{R}^k$ and is Chebyshev, if and only if, $P_{\mathcal{K}}(\mathbf{a}) = \{\mathbf{b}\}$, with $\mathbf{b} \in \mathcal{K}$, $\forall \mathbf{a} \in \mathbb{R}^k$. In this last case, $P_{\mathcal{K}}(.)$ can be viewed simply as a mapping from $\mathbb{R}^k$ to $\mathcal{K}$ in the usual sense. This will be for example the case if $\mathcal{K}$ is a linear subspace of $\mathbb{R}^k$ (in which case $P_{\mathcal{K}}(.)$ is simply the orthogonal projector $\text{Proj}_{\mathcal{K}}(.)$) or, more generally, if $\mathcal{K}$ is a closed convex set, as we will show shortly.

First, if we assume that $\mathcal{K}$ is a nonempty closed subset of $\mathbb{R}^k$ then all points $\mathbf{a} \in \mathbb{R}^k$ have at least one nearest point in $\mathcal{K}$. To see this, define a real function $f_{\mathbf{a}}(.)$ from $\mathbb{R}^k$ to $\mathbb{R}$, $\forall \mathbf{a} \in \mathbb{R}^k$, by

$$f_{\mathbf{a}}(\mathbf{b}) = \|\mathbf{b} - \mathbf{a}\|_2 \ , \ \forall \mathbf{b} \in \mathbb{R}^k \ ,$$

take a point $\mathbf{c} \in \mathcal{K}$ and define the sublevel set

$$S_{\mathbf{c}} = \{\mathbf{b} \in \mathbb{R}^k \ / \ f_{\mathbf{a}}(\mathbf{b}) \leq f_{\mathbf{a}}(\mathbf{c})\} \ .$$

$S_{\mathbf{c}}$ is a compact set of $\mathbb{R}^k$ as $f_{\mathbf{a}}(.)$ is continuous, $[-\infty \ , \ f_{\mathbf{a}}(\mathbf{c})]$ is closed in $\mathbb{R}$ and $S_{\mathbf{c}}$ is bounded by definition. Then, we have obviously

$$d(\mathbf{a}, \mathcal{K}) = \inf_{\mathbf{b} \in \mathcal{K} \cap S_{\mathbf{c}}} f_{\mathbf{a}}(\mathbf{b}) \ ,$$

which has a solution in $\mathcal{K}$ as $f_{\mathbf{a}}(.)$ is continuous and $\mathcal{K} \cap S_{\mathbf{c}}$ is compact (since $\mathcal{K} \cap S_{\mathbf{c}}$ is closed and bounded) in $\mathbb{R}^k$. This implies, the existence of, at least, one nearest point in $\mathcal{K}$ to $\mathbf{a}$ for all $\mathbf{a} \in \mathbb{R}^k$ if $\mathcal{K}$ is closed.

On the other hand, if $\mathcal{K}$ is a convex subset of $\mathbb{R}^k$, then $\forall \mathbf{a} \in \mathbb{R}^k$, $\mathbf{a}$ has at most one nearest point in $\mathcal{K}$. To demonstrate this claim suppose that $\mathcal{K}$ is convex and that $\mathbf{a} \in \mathbb{R}^k$ has two distinct nearest points in $\mathcal{K}$, say $\mathbf{b}_1$ and $\mathbf{b}_2$. By using the parallelogram law with $\mathbf{d}_1 = \mathbf{b}_1 - \mathbf{a}$ and $\mathbf{d}_2 = \mathbf{b}_2 - \mathbf{a}$, we get

$$\begin{aligned} \|\mathbf{d}_1 + \mathbf{d}_2\|_2^2 + \|\mathbf{d}_1 - \mathbf{d}_2\|_2^2 &= 2.\|\mathbf{d}_1\|_2^2 + 2.\|\mathbf{d}_2\|_2^2 \\ &\Longrightarrow \|\frac{\mathbf{b}_1 + \mathbf{b}_2}{2} - \mathbf{a}\|_2^2 = \|\mathbf{b}_1 - \mathbf{a}\|_2^2 - \frac{1}{4}\|\mathbf{b}_1 - \mathbf{b}_2\|_2^2 \ . \end{aligned}$$

Since $\mathcal{K}$ is convex, $\frac{\mathbf{b}_1 + \mathbf{b}_2}{2}$ belongs to $\mathcal{K}$ and we have $\|\frac{\mathbf{b}_1 + \mathbf{b}_2}{2} - \mathbf{a}\|_2^2 < \|\mathbf{b}_1 - \mathbf{a}\|_2^2$, which contradicts the fact that $\mathbf{b}_1$ is a nearest point to $\mathbf{a}$ in $\mathcal{K}$.

In summary, if $\mathcal{K}$ is a nonempty closed and convex subset of $\mathbb{R}^k$, $P_{\mathcal{K}}(\mathbf{a}) = \{\mathbf{c}\}$ with $\mathbf{c} \in \mathcal{K}$, $\forall \mathbf{a} \in \mathbb{R}^k$, and, by an abuse of notation, the metric projection defines effectively a simple metric projection mapping $P_{\mathcal{K}}(\mathbf{a}) = \mathbf{c}$, which to each $\mathbf{a} \in \mathbb{R}^k$ associates its unique nearest point in $\mathcal{K}$. Interestingly, when $\mathcal{K}$ is a nonempty closed and convex set, the point $P_{\mathcal{K}}(\mathbf{a}) = \mathbf{c}$ is equivalently characterized by the following property:

$$P_{\mathcal{K}}(\mathbf{a}) = \mathbf{c} \Longleftrightarrow \langle \mathbf{a} - \mathbf{c}, \mathbf{b} - \mathbf{c} \rangle_2 \leq 0 \ , \forall \mathbf{b} \in \mathcal{K} \ , \tag{2.58}$$

see Theorem 3.1.1 of Hiriart-Urruty and Le Marechal [79] for a proof. This equivalence can be obviously restated with the help of the polar cone of the set $(\mathcal{K} - \mathbf{c})$ as

$$P_{\mathcal{K}}(\mathbf{a}) = \mathbf{c} \Longleftrightarrow \mathbf{a} - \mathbf{c} \in (\mathcal{K} - \mathbf{c})^o \ ,$$

which generalizes the property $\mathbf{a} - \text{Proj}_{\mathcal{K}}(\mathbf{a}) \in \mathcal{K}^{\perp}$ when $\mathcal{K}$ is a subspace of $\mathbb{R}^k$ and $\text{Proj}_{\mathcal{K}}(.)$ is the orthogonal projector onto $\mathcal{K}$. Furthermore, when $\mathcal{K}$ is a nonempty closed and convex subset of $\mathbb{R}^k$, we have the following additional properties [79][39]:

- the set $\{\mathbf{a} \in \mathbb{R}^k \ / \ P_{\mathcal{K}}(\mathbf{a}) = \mathbf{a}\}$ of fixed points of $P_{\mathcal{K}}(.)$ is $\mathcal{K}$ itself;

- the metric projection mapping is idempotent, e.g., $P_{\mathcal{K}} o P_{\mathcal{K}} = P_{\mathcal{K}}$ and this justifies the term metric projection for $P_{\mathcal{K}}(.)$;

- The metric projection mapping $P_{\mathcal{K}}(.)$ is nonexpansive in the sense that $\|P_{\mathcal{K}}(\mathbf{a}) - P_{\mathcal{K}}(\mathbf{b})\|_2 \leq \|\mathbf{a}-\mathbf{b}\|_2$, $\forall \mathbf{a}, \mathbf{b} \in \mathbb{R}^k$, implying that the metric projection mapping $P_{\mathcal{K}}(.)$ is uniformly continuous on $\mathbb{R}^k$. Furthermore, if $\mathcal{K}$ is also a cone, $\mathbf{0}^k \in \mathcal{K}$ and we have $\|P_{\mathcal{K}}(\mathbf{a})\|_2 \leq \|\mathbf{a}\|_2$, $\forall \mathbf{a} \in \mathbb{R}^k$, as for the orthogonal projector $\text{Proj}_{\mathcal{K}}(.)$ when $\mathcal{K}$ is a subspace of $\mathbb{R}^k$;

- and, finally, $P_{\mathcal{K}}(.)$ is a linear operator if and only if $\mathcal{K}$ is a linear subspace of $\mathbb{R}^k$.

Suppose now that $\mathcal{K}$ is a nonempty subspace of $\mathbb{R}^k$. Then, $\mathcal{K}$ is a closed convex set and the metric projection operator $P_{\mathcal{K}}(.)$ is well defined as a mapping from $\mathbb{R}^k$ to $\mathcal{K}$. However, we also know from the results of Subsection 2.1 that

$$\inf_{\mathbf{d}\in\mathcal{K}} \|\mathbf{a}-\mathbf{d}\|_2 = \min_{\mathbf{d}\in\mathcal{K}} \|\mathbf{a}-\mathbf{d}\|_2 = \|\mathbf{a} - \text{Proj}_{\mathcal{K}}(\mathbf{a})\|_2 \ , \ \forall \mathbf{a} \in \mathbb{R}^k \ ,$$

where $\mathcal{K}$ is a nonempty linear subspace of $\mathbb{R}^k$ and $\text{Proj}_{\mathcal{K}}(.)$ is the unique orthogonal projector operator onto $\mathcal{K}$. Consequently, as the metric projection operator $P_{\mathcal{K}}(.)$ also solves uniquely this minimization problem in $\mathcal{K}$, we deduce immediately that $\text{Proj}_{\mathcal{K}}(\mathbf{a}) = P_{\mathcal{K}}(\mathbf{a}), \forall \mathbf{a} \in \mathbb{R}^k$. Thus, when $\mathcal{K}$ is a linear subspace, the metric projection mapping $P_{\mathcal{K}}(.)$ is nothing else than the orthogonal projector operator onto $\mathcal{K}$, $\text{Proj}_{\mathcal{K}}(.)$, suggesting again that we can interpret the metric projection mapping as an extension of the orthogonal projector mapping.

All these different properties confirm that we can somehow interpret the metric projection mapping as an extension of an orthogonal projector mapping when the set of fixed points is a closed and convex subset rather than a linear subspace. Furthermore, we come even closer to an orthogonal projector, if we further assume that $\mathcal{K}$ is also cone, since in that case we have

$$\mathbf{a} = P_{\mathcal{K}}(\mathbf{a}) + P_{\mathcal{K}^o}(\mathbf{a}) \text{ with } \langle P_{\mathcal{K}}(\mathbf{a}), P_{\mathcal{K}^o}(\mathbf{a})\rangle_2 = 0 \ , \ \forall \mathbf{a} \in \mathbb{R}^k \ ,$$

which generalizes the canonical orthogonal decomposition $\mathbf{a} = \text{Proj}_{\mathcal{K}}(\mathbf{a}) + \text{Proj}_{\mathcal{K}^\perp}(\mathbf{a})$ when $\mathcal{K}$ is a subspace, see Section 3.2 of [79] for details.

Since, we will mainly use the metric projection to project onto closed cones (e.g., the Bouligand's tangent cone to $\mathcal{K}$ at $\mathbf{a}$ when $\mathcal{K}$ is a nonempty, eventually closed, subset of $\mathbb{R}^k$), we focus now specifically on the properties of the metric projection operator, which are still valid in this case.

First, note that if $\mathcal{C}$ is a closed subset of $\mathbb{R}^k$, the distance function defined as $d_{\mathcal{C}}(\mathbf{a}) = d(\mathbf{a}, \mathcal{C})$ from $\mathbb{R}^k$ to $\mathbb{R}$ is well-defined (since $\mathcal{C}$ is closed) and continuous on $\mathbb{R}^k$, see example 1.20 of Rockafellar and Wets [165] for a proof. Next, $\forall \mathbf{a} \in \mathbb{R}^k$, the set $P_{\mathcal{C}}(\mathbf{a}) = d_{\mathcal{C}}^{-1}(\mathbf{a})$ is nonempty (as shown above), bounded and closed, and thus compact in $\mathbb{R}^k$. It is closed as the reciprocal image of the singleton $\{d(\mathbf{a}, \mathcal{C})\}$ of $\mathbb{R}$ by the continuous distance function $d_{\mathcal{C}}(.)$. It is bounded, because if we take a fixed point $\mathbf{c} \in \mathcal{C}$, we have, $\forall \mathbf{b} \in P_{\mathcal{C}}(\mathbf{a})$, by definition, the inequality $\|\mathbf{a}-\mathbf{b}\|_2 \leq \|\mathbf{a}-\mathbf{c}\|_2$ and the distance of $\mathbf{b}$ to $\mathbf{a}, \forall \mathbf{b} \in P_{\mathcal{C}}(\mathbf{a})$ is bounded by $\|\mathbf{a}-\mathbf{c}\|_2$.

We next state the following Lemma, which will be useful to prove our next Theorem:

Lemma 2.7. Let $\mathcal{C}$ be a closed cone in $\mathbb{R}^k$. $\forall \mathbf{a} \in \mathbb{R}^k$ and $\forall \mathbf{b} \in P_{\mathcal{C}}(\mathbf{a})$, we have

$$\|\mathbf{b}\|_2 = \max\big(0, \max_{\mathbf{c}\in\mathcal{C}, \|\mathbf{c}\|_2=1} \langle \mathbf{a}, \mathbf{c}\rangle_2\big) = \sqrt{\langle \mathbf{a}, \mathbf{b}\rangle_2} \ .$$

Proof. Omitted. See Proposition A.6 of Levin et al. [112] for a proof. □

Theorem 2.8. Let $\mathcal{C}$ be a closed cone in $\mathbb{R}^k$. $\forall \mathbf{a} \in \mathbb{R}^k$ and $\forall \mathbf{b} \in P_{\mathcal{C}}(\mathbf{a})$, we have

$$\|\mathbf{b}\|_2^2 = \|\mathbf{a}\|_2^2 - d(\mathbf{a}, \mathcal{C})^2 \ ,$$

implying that all the elements of $P_{\mathcal{C}}(\mathbf{a})$ have the same length, and

$$\mathbf{a} \in \mathcal{C}^o \iff P_{\mathcal{C}}(\mathbf{a}) = \{\mathbf{0}^k\} \ .$$

In words, if $\mathbf{a}$ belongs to the polar of the closed cone $\mathcal{C}$, its metric projection over $\mathcal{C}$, $P_{\mathcal{C}}(\mathbf{a})$, is reduced to the zero-vector of the ambiant linear space and reciprocally.

Proof. First, we have the equalities

$$\begin{aligned}\|\mathbf{b}\|_2^2 &= \|\mathbf{a}-(\mathbf{a}-\mathbf{b})\|_2^2 \\ &= \|\mathbf{a}\|_2^2 + \|\mathbf{a}-\mathbf{b}\|_2^2 - 2\langle \mathbf{a}, \mathbf{a}-\mathbf{b}\rangle_2 \\ &= \|\mathbf{a}\|_2^2 + \|\mathbf{a}-\mathbf{b}\|_2^2 - 2\|\mathbf{a}\|_2^2 + 2\langle \mathbf{a}, \mathbf{b}\rangle_2 \\ &= \|\mathbf{a}-\mathbf{b}\|_2^2 - \|\mathbf{a}\|_2^2 + 2\langle \mathbf{a}, \mathbf{b}\rangle_2 \ .\end{aligned}$$

Now, since $\mathcal{C}$ is a closed cone by hypothesis, using Lemma (2.7), we have $\|\mathbf{b}\|_2^2 = \langle \mathbf{a}, \mathbf{b}\rangle_2$, from which we get

$$\|\mathbf{b}\|_2^2 = \|\mathbf{a}-\mathbf{b}\|_2^2 - \|\mathbf{a}\|_2^2 + 2\|\mathbf{b}\|_2^2 \ ,$$

which is equivalent after simplification to

$$\|\mathbf{b}\|_2^2 = \|\mathbf{a}\|_2^2 - \|\mathbf{a}-\mathbf{b}\|_2^2 = \|\mathbf{a}\|_2^2 - d(\mathbf{a}, \mathcal{C})^2 \ ,$$

as claimed in the theorem.

We now demonstrate the implication $\mathbf{a} \in \mathcal{C}^o \Rightarrow P_{\mathcal{C}}(\mathbf{a}) = \{\mathbf{0}^k\}$, $\forall \mathbf{a} \in \mathbb{R}^k$. If $\mathbf{a} \in \mathcal{C}^o$, for $\mathbf{c} \in \mathcal{C}$, we have first

$$\|\mathbf{a}-\mathbf{c}\|_2^2 = \|\mathbf{a}\|_2^2 + \|\mathbf{c}\|_2^2 - 2\langle \mathbf{a}, \mathbf{c}\rangle_2$$

and, as $\mathbf{a} \in \mathcal{C}^o$, also the inequality $\langle \mathbf{a}, \mathbf{c}\rangle_2 \leq 0$. This implies that the term $\|\mathbf{c}\|_2^2 - 2\langle \mathbf{a}, \mathbf{c}\rangle_2$ is strictly positive, $\forall \mathbf{c} \in \mathcal{C}\backslash\{\mathbf{0}^k\}$, and we get the inequality

$$\|\mathbf{a}-\mathbf{c}\|_2^2 > \|\mathbf{a}\|_2^2 \ , \ \forall \mathbf{c} \in \mathcal{C}\backslash\{\mathbf{0}^k\} \ ,$$

and also

$$\|\mathbf{a}-\mathbf{c}\|_2 > \|\mathbf{a}\|_2 \ , \ \forall \mathbf{c} \in \mathcal{C}\backslash\{\mathbf{0}^k\} \ ,$$

after simplification. In other words, $\mathbf{0}^k \in \mathcal{C}$ is the unique nearest point in $\mathcal{C}$ to $\mathbf{a}$, e.g., $P_{\mathcal{C}}(\mathbf{a}) = \{\mathbf{0}^k\}$ if $\mathbf{a} \in \mathcal{C}^o$, as claimed above.

Reciprocally, we now demonstrate the implication $\mathbf{0}^k \in P_{\mathcal{C}}(\mathbf{a}) \Rightarrow \mathbf{a} \in \mathcal{C}^o$, $\forall \mathbf{a} \in \mathbb{R}^k$. First, as $\mathcal{C}$ is a closed cone by hypothesis, using the first assertion of the Theorem demonstrated above, we have $\|\mathbf{a}\|_2 = d(\mathbf{a}, \mathcal{C})$ and, $\forall \mathbf{c} \in P_{\mathcal{C}}(\mathbf{a})$, we have $\|\mathbf{c}\|_2 = 0$ and, thus, $P_{\mathcal{C}}(\mathbf{a}) = \{\mathbf{0}^k\}$. In other words, $\mathbf{0}^k$ is the unique nearest point to $\mathbf{a}$ in $\mathcal{C}$. Furthermore, from Lemma (2.7), we have also

$$\max_{\mathbf{c}\in\mathcal{C}, \|\mathbf{c}\|_2=1} \langle \mathbf{a}, \mathbf{c}\rangle_2 \leq 0 \ .$$

In order to demonstrate that $\mathbf{a} \in \mathcal{C}^o$, e.g., that $\langle \mathbf{a}, \mathbf{d}\rangle_2 \leq 0$, $\forall \mathbf{d} \in \mathcal{C}$, we now proceed by contradiction. Suppose that it exists $\mathbf{d} \in \mathcal{C}$ such that $\langle \mathbf{a}, \mathbf{d}\rangle_2 > 0$. Then, $\mathbf{d} \neq \mathbf{0}^k$ and $\|\mathbf{d}\|_2 \neq 0$, and we have

$$\max_{\mathbf{c}\in\mathcal{C}, \|\mathbf{c}\|_2=1} \langle \mathbf{a}, \mathbf{c}\rangle_2 \leq 0 < \langle \mathbf{a}, \frac{\mathbf{d}}{\|\mathbf{d}\|_2}\rangle_2 \ ,$$

which is a contradiction, since $\frac{\mathbf{d}}{\|\mathbf{d}\|_2} \in \mathcal{C}$, because $\mathcal{C}$ is a cone, and $\|\frac{\mathbf{d}}{\|\mathbf{d}\|_2}\|_2 = 1$.

Summarizing, we have, $\forall \mathbf{a} \in \mathbb{R}^k$,

$$\mathbf{a} \in \mathcal{C}^o \iff P_{\mathcal{C}}(\mathbf{a}) = \{\mathbf{0}^k\} \ ,$$

as claimed in the Theorem and we are done. □

Now, we can come back to our problem of reformulating the first-order stationarity condition (2.57) for a point $\hat{\mathbf{a}} \in \mathcal{K}$, $\mathcal{K}$ being an arbitrary subset of $\mathbb{R}^k$, to be a (local) minimizer of a real function $\phi(.)$ differentiable over an open neighborhood of $\mathcal{K}$. An application of Theorem 2.8 to the anti-gradient $-\nabla\phi(\hat{\mathbf{a}})$ and the tangent Bouligand's cone $\mathcal{T}^{\mathcal{B}}_{\hat{\mathbf{a}}}\mathcal{K}$, which is a closed cone, leads to the following equivalent first-order critical conditions

$$-\nabla\phi(\hat{\mathbf{a}}) \in \mathcal{N}^{\mathcal{F}}_{\hat{\mathbf{a}}}\mathcal{K} = (\mathcal{T}^{\mathcal{B}}_{\hat{\mathbf{a}}}\mathcal{K})^o \iff P_{\mathcal{T}^{\mathcal{B}}_{\hat{\mathbf{a}}}\mathcal{K}}(-\nabla\phi(\hat{\mathbf{a}})) = \{\mathbf{0}^k\} \, . \tag{2.59}$$

Moreover, by a small abuse of notation, we can write $P_{\mathcal{T}^{\mathcal{B}}_{\hat{\mathbf{a}}}\mathcal{K}}(-\nabla\phi(\hat{\mathbf{a}})) = \mathbf{0}^k$ and the first-order stationarity condition for $\hat{\mathbf{a}} \in \mathcal{K}$ to be a (local) minimizer becomes

$$\|P_{\mathcal{T}^{\mathcal{B}}_{\hat{\mathbf{a}}}\mathcal{K}}(-\nabla\phi(\hat{\mathbf{a}}))\|_2 = 0 \, , \tag{2.60}$$

where $P_{\mathcal{T}^{\mathcal{B}}_{\hat{\mathbf{a}}}\mathcal{K}}(-\nabla\phi(\hat{\mathbf{a}}))$ designs now any of its elements since they have all the same length according to Theorem 2.8. Note the similarity of this first-order condition (2.59) or (2.60) with the one stated above in equation (2.50) in the case where $\mathcal{K}$ is an embedded smooth submanifold of $\mathbb{R}^k$.

To conclude these paragraphs on optimality conditions for a real function $\phi(.)$ at a point $\hat{\mathbf{a}} \in \mathcal{K}$, where $\mathcal{K}$ is an nonempty arbitrary subset of $\mathbb{R}^k$, we now recall in the following theorem the necessary second-order condition for a point $\hat{\mathbf{a}} \in \mathcal{K}$ to be a (local) minimizer over $\mathcal{K}$ of a cost function $\phi(.)$ twice continuously differentiable over an open neighborhood of $\mathcal{K}$ in $\mathbb{R}^k$.

Theorem 2.9. Let $\mathcal{K}$ be a nonempty subset of $\mathbb{R}^k$ and assume that $\phi(.)$ is a real function twice continuously differentiable over an open neighborhood of $\mathcal{K}$ in $\mathbb{R}^k$ and that $\hat{\mathbf{a}} \in \mathcal{K}$ is a (local) minimizer of $\phi(.)$ over $\mathcal{K}$. Then, for every $\mathbf{d} \in \mathcal{T}^{\mathcal{B}}_{\hat{\mathbf{a}}}\mathcal{K}$ satisfying $\langle \nabla\phi(\hat{\mathbf{a}}), \mathbf{d}\rangle_2 = 0$ we have

$$\langle \nabla\phi(\hat{\mathbf{a}}), \mathbf{c}\rangle_2 + \langle \big[\nabla^2\phi(\hat{\mathbf{a}})\big](\mathbf{d}), \mathbf{d}\rangle_2 \geq 0 \, , \; \forall \mathbf{c} \in \mathcal{T}^{\mathcal{B}}_{(\hat{\mathbf{a}},\mathbf{d})}\mathcal{K} \, , \tag{2.61}$$

where $\mathcal{T}^{\mathcal{B}}_{\hat{\mathbf{a}}}\mathcal{K}$ is the Bouligand tangent cone to $\mathcal{K}$ at $\hat{\mathbf{a}}$ and $\mathcal{T}^{\mathcal{B}}_{(\hat{\mathbf{a}},\mathbf{d})}\mathcal{K}$ is the second-order (Bouligand) tangent set to $\mathcal{K}$ at $\hat{\mathbf{a}}$ in the direction of $\mathbf{d} \in \mathcal{T}^{\mathcal{B}}_{\hat{\mathbf{a}}}\mathcal{K}$ (see Definition 3.41 in Ruszczynski [160] for a precise definition of this second-order tangent set). Note, however, that $\mathcal{T}^{\mathcal{B}}_{(\hat{\mathbf{a}},\mathbf{d})}\mathcal{K}$ is not a cone in general, nor it is convex.

Proof. Omitted. See Theorem 3.45 of Ruszczynski [160] for a proof. □

Using Theorem 2.9, we will say that $\hat{\mathbf{a}} \in \mathcal{K}$ is a (Frechet) second-order stationarity point of $\phi(.)$ over $\mathcal{K}$ if it is a (Frechet) first-order stationarity point for $\phi(.)$ and if, in addition, the condition 2.61 is fulfilled.

In the following sections, we will also manipulate (differentiable) scalar, vector or matrix functions with a matrix argument $\mathbf{A} \in \mathbb{R}^{p\times n}$. As an illustration, let $\phi(.)$ be a scalar function defined on $\mathbb{R}^{p\times n}$. If $\mathbb{R}^{p\times n}$ is equipped with its usual Frobenius inner product, the gradient of $\phi(.)$ at a matrix variable $\mathbf{A} \in \mathbb{R}^{p\times n}$ is also a $p \times n$ matrix, i.e.,

$$\big[\nabla\phi(\mathbf{A})\big]_{ij} = \frac{\partial\phi(\mathbf{A})}{\partial\mathbf{A}_{ij}} \text{ for } i = 1, \cdots, p \, ; j = 1, \cdots, n \, . \tag{2.62}$$

Alternatively, we can interpret this gradient as a linear form $\big(\nabla\phi(\mathbf{A})\big) \in \pounds(\mathbb{R}^{p\times n}, \mathbb{R})$ defined by

$$\big(\nabla\phi(\mathbf{A})\big)(\mathbf{C}) = \langle \nabla\phi(\mathbf{A}) \, , \, \mathbf{C}\rangle_F = \mathrm{Tr}\left(\nabla\phi(\mathbf{A})^T\mathbf{C}\right) = \sum_{i=1}^{p}\sum_{j=1}^{n}\frac{\partial\phi(\mathbf{A})}{\partial\mathbf{A}_{ij}}\mathbf{C}_{ij} \, , \forall \mathbf{C} \in \mathbb{R}^{p\times n} \, .$$

On the other hand, the Hessian of $\phi(.)$ at $\mathbf{A} \in \mathbb{R}^{p\times n}$ can be viewed as a 4^{th} order tensor of dimension $p\times n\times p\times n$, instead of a symmetric matrix (see equation (2.42)) as in the case of a vector argument, which is equal to

$$\big[\nabla^2\phi(\mathbf{A})\big]_{ijkl} = \frac{\partial^2\phi(\mathbf{A})}{\partial\mathbf{A}_{ij}\partial\mathbf{A}_{kl}} \text{ for } i = 1,\cdots,p\,;\, j = 1,\cdots,n\,;\, k = 1,\cdots,p\,;\, l = 1,\cdots,n\,. \tag{2.63}$$

Equivalently, we can view $\nabla^2\phi(\mathbf{A})$ as a bilinear form $\big(\nabla^2\phi(\mathbf{A})\big)$, from $\mathbb{R}^{p\times n}\times\mathbb{R}^{p\times n}$ to $\mathbb{R}$, defined by

$$\big(\nabla^2\phi(\mathbf{A})\big)(\mathbf{C},\mathbf{D}) = \sum_{i,j,k,l}\frac{\partial^2\phi(\mathbf{A})}{\partial\mathbf{A}_{ij}\partial\mathbf{A}_{kl}}\mathbf{C}_{ij}\mathbf{D}_{kl}\,,\,\forall\mathbf{C},\mathbf{D}\in\mathbb{R}^{p\times n}\,.$$

Finally, another very useful representation of $\nabla^2\phi(\mathbf{A})$, implicit in the preceding one, is as a huge $p.n\times p.n$ symmetric matrix

$$\big[\nabla^2\phi(\mathbf{A})\big]_{ij} = \frac{\partial^2\phi(\mathbf{A})}{\partial\mathbf{a}_i\partial\mathbf{a}_j} \text{ for } i = 1,\cdots,p.n\,;\, j = 1,\cdots,p.n\,,$$

where $\mathbf{a}_i$ is the i^{th} element of a vectorized form of $\mathbf{A}$, e.g., $\mathbf{a} = vec(\mathbf{A})$ or $\mathbf{a} = vec(\mathbf{A}^T)$. For example, in Subsection 5.3 we will derive the Hessian of a real (variable projection) functional $\psi(.)$ of the matrix variable $\mathbf{A}\in\mathbb{R}^{p\times k}$ (defined in the next section) using this specific representation.

The first and second derivatives of a matrix function $f(.)$ from $\mathbb{R}^{p\times n}$ to $\mathbb{R}^{q\times r}$ can also be viewed as higher order tensors. However, it is generally more convenient to represent them as linear or multilinear operators [26]. For example, the first derivative of $f(.)$ at $\mathbf{A}\in\mathbb{R}^{p\times n}$ is a linear operator from $\mathbb{R}^{p\times n}$ to $\mathbb{R}^{q\times r}$, e.g., $Df(\mathbf{A})\in\pounds(\mathbb{R}^{p\times n},\mathbb{R}^{q\times r})$, and the second derivative of $f(.)$ at $\mathbf{A}$, $D^2f(\mathbf{A})$, is an element of $\pounds\big(\mathbb{R}^{p\times n},\pounds(\mathbb{R}^{p\times n},\mathbb{R}^{q\times r})\big)$, which is isomorphic to $\pounds(\mathbb{R}^{p\times n},\mathbb{R}^{p\times n};\mathbb{R}^{q\times r})$, the set of bilinear maps from $\mathbb{R}^{p\times n}$ into $\mathbb{R}^{q\times r}$ [26]. Thus, $D^2f(\mathbf{A})$ can be interpreted as a bilinear operator from $\mathbb{R}^{p\times n}\times\mathbb{R}^{p\times n}$ to $\mathbb{R}^{q\times r}$. In this way, the Hessian of a scalar function $\phi(.)$ with a matrix argument $\mathbf{A}\in\mathbb{R}^{p\times n}$ discussed above is the first derivative of its gradient, which is a mapping from $\mathbb{R}^{p\times n}$ to $\mathbb{R}^{p\times n}$, and, thus, this Hessian can be viewed as a mapping from $\mathbb{R}^{p\times n}$ to $\pounds(\mathbb{R}^{p\times n},\mathbb{R}^{p\times n})$, e.g., for $\mathbf{A}\in\mathbb{R}^{p\times n}$, $\big[\nabla^2\phi(\mathbf{A})\big]\in\pounds(\mathbb{R}^{p\times n},\mathbb{R}^{p\times n})$ and is a linear operator from $\mathbb{R}^{p\times n}$ to $\mathbb{R}^{p\times n}$. Furthermore, we can identify the bilinear form $\big(\nabla^2\phi(\mathbf{A})\big)$ with $\big[\nabla^2\phi(\mathbf{A})\big]$ [26] and they verify the equality

$$\langle\big[\nabla^2\phi(\mathbf{A})\big](\mathbf{C})\,,\,\mathbf{D}\rangle_F = \big(\nabla^2\phi(\mathbf{A})\big)(\mathbf{C},\mathbf{D})\,,\forall\mathbf{C},\mathbf{D}\in\mathbb{R}^{p\times n}.$$

This identification of $\big(\nabla^2\phi(\mathbf{A})\big)$ with $\big[\nabla^2\phi(\mathbf{A})\big]$ can be very useful in practice as evaluating directly $\big[\nabla^2\phi(\mathbf{A})\big](\mathbf{C})$ (e.g., the directional derivative of the gradient of $\phi(.)$ in the direction of $\mathbf{C}$) can be much cheaper and efficient than computing analytically the full Hessian $\nabla^2\phi(\mathbf{A})$. This is for example the approach followed by Boumal and Absil [13][14] in their Newton Riemannian trust-region method for solving the WLRA problem in a Grassmann manifold framework (recall that a Grassmann manifold is the collection of all linear subspaces of a given dimension in a particular Euclidean space as already discussed above).

Keep also in mind that all the above notions of a smooth function, smooth manifold, tangent space to a smooth manifold, tangent and normal cones to an arbitrary subset and metric projection onto an arbitrary subset can be defined without any difficulties in the case when the ambient linear space is $\mathbb{R}^{p\times k}$ instead of $\mathbb{R}^p$ if the linear space $\mathbb{R}^{p\times k}$ is equipped with the standard Frobenius inner product [116]. Moreover, the linear spaces $\mathbb{R}^{p\times k}$ and $\mathbb{R}^{p.k}$ are isomorphic and the Frobenius metric on $\mathbb{R}^{p\times k}$ is equivalent to the standard Euclidean metric on $\mathbb{R}^{p.k}$ thanks to this isomorphism.

We conclude that preliminary section by a few more definitions about nonlinear optimization, which will be useful for our next sections.

A function $\phi(.)$ is said to be nonlinear in some scalar parameter α, vector parameter $\mathbf{a}$ or matrix parameter $\mathbf{A}$ if the derivatives $\frac{\partial\phi(.)}{\partial\alpha}$, $\frac{\partial\phi(.)}{\partial\mathbf{a}}$ and $\frac{\partial\phi(.)}{\partial\mathbf{A}}$ are functions of α, $\mathbf{a}$ and $\mathbf{A}$, respectively [87].

As an illustration, let $\mathbf{r}(.)$ be a real-vector function from $\mathbb{R}^k$ into $\mathbb{R}^q$ and further assume that $\mathbf{r}(.)$ is at least twice continuously differentiable. Then, the real function $\phi(.)$ from $\mathbb{R}^k$ into $\mathbb{R}$ defined by

$$\phi(\mathbf{a}) = \frac{1}{2}\|\mathbf{r}(\mathbf{a})\|_2^2 = \frac{1}{2}\mathbf{r}(\mathbf{a})^T\mathbf{r}(\mathbf{a}) \text{ for } \mathbf{a} \in \mathbb{R}^k \tag{2.64}$$

is called a Non-Linear Least-Squares (NLLS) functional. If we differentiate $\phi(.)$ with respect to $\mathbf{a} \in \mathbb{R}^k$ (e.g., we compute its gradient at $\mathbf{a}$) and equate the derivative to zero, this leads to the following equation

$$\nabla\phi(\mathbf{a}) = J\big(\mathbf{r}(\mathbf{a})\big)^T\mathbf{r}(\mathbf{a}) = \mathbf{0}^k \ , \tag{2.65}$$

which may be used in practice to test the convergence of NLLS iterative algorithms employed for minimizing $\phi(.)$ over $\mathbb{R}^k$ [148][45][123]. This last equation shows that the vector $\mathbf{r}(\mathbf{a})$ is orthogonal to *ran*$(J(\mathbf{r}(\mathbf{a})))$, the linear subspace spanned by the columns of the Jacobian matrix of the real q-vector function $\mathbf{r}(.)$ at $\mathbf{a}$, if $\mathbf{a}$ is a stationary point of $\phi(.)$. Furthermore, if $\phi(.)$ is a NLLS functional then its Hessian matrix is

$$\nabla^2\phi(\mathbf{a}) = J\big(\mathbf{r}(\mathbf{a})\big)^T J\big(\mathbf{r}(\mathbf{a})\big) + \sum_{l=1}^{q} \mathbf{r}_l(\mathbf{a})\nabla^2\mathbf{r}_l(\mathbf{a}) \ , \tag{2.66}$$

where $\nabla^2\mathbf{r}_l(\mathbf{a})$ is the Hessian matrix of the l^{th} component of the q-vector function $\mathbf{r}(.)$ at $\mathbf{a}$ (i.e., $\mathbf{r}_l(\mathbf{a})$) given by

$$\big[\nabla^2\mathbf{r}_l(\mathbf{a})\big]_{ij} = \frac{\partial^2\mathbf{r}_l(\mathbf{a})}{\partial\mathbf{a}_i\partial\mathbf{a}_j} \text{ for } i = 1, \cdots, k \text{ and } j = 1, \cdots, k \ ,$$

for $l = 1, \cdots, q$. Note that the factor $\frac{1}{2}$ in the definition 2.64 of the NLLS functional $\phi(.)$ has been introduced here only for notational convenience as without it a factor 2 will appear in the two preceding equations defining $\nabla\phi(.)$ and $\nabla^2\phi(.)$ and in many equations of this paper. Furthermore, the second-order Taylor expansion of the NLLS functional $\phi(.)$ at a point $\mathbf{a} \in \mathbb{R}^k$ has the following form

$$\begin{aligned}\phi(\mathbf{a} + d\mathbf{a}) = \phi(\mathbf{a}) + d\mathbf{a}^T J(\mathbf{r}(\mathbf{a}))^T\mathbf{r}(\mathbf{a}) + \\ \frac{1}{2}d\mathbf{a}^T\left(J(\mathbf{r}(\mathbf{a}))^T J(\mathbf{r}(\mathbf{a})) + \sum_{l=1}^{q}\mathbf{r}_l(\mathbf{a})\nabla^2\mathbf{r}_l(\mathbf{a})\right)d\mathbf{a} + \mathcal{O}(\|d\mathbf{a}\|_2^3) \ .\end{aligned}$$

These special forms of the gradient, Hessian and Taylor expansion of $\phi(.)$ are exploited by methods for solving NLLS problems, see Subsection 5.1 and [45][123][87] for details.

Finally, we give the following definition, which will be also useful in the next sections:

Definition 2.10. Let $m, n, p, k \in \mathbb{N}_*$ (e.g., the set of strictly positive integers). A NLLS problem associated with a cost function $\phi(.)$ from $\mathbb{R}^k$ into $\mathbb{R}$ and a residual real-vector function $\mathbf{r}(.)$ from $\mathbb{R}^k$ into $\mathbb{R}^m$ is said to be separable if the parameter vector $\mathbf{a} \in \mathbb{R}^k$ can be partitioned as

$$\mathbf{a} = \begin{bmatrix}\mathbf{b}\\ \mathbf{c}\end{bmatrix} \text{ with } \mathbf{b} \in \mathbb{R}^n, \mathbf{c} \in \mathbb{R}^p \text{ and } n + p = k \ ,$$

in such a way that the subproblem

$$\min_{\mathbf{c}\in\mathbb{R}^p} \phi\Big(\begin{bmatrix}\mathbf{b}\\ \mathbf{c}\end{bmatrix}\Big) = \frac{1}{2}\Big\|\mathbf{r}\Big(\begin{bmatrix}\mathbf{b}\\ \mathbf{c}\end{bmatrix}\Big)\Big\|_2^2$$

is easy to solve numerically for every fixed $\mathbf{b} \in \mathbb{R}^n$ [63][166][87].

In the following, we will be particularly interested in the particular case when $\mathbf{r}(\begin{bmatrix}\mathbf{b}\\ \mathbf{c}\end{bmatrix})$ is linear in $\mathbf{c} \in \mathbb{R}^p$, i.e.,

$$\mathbf{r}(\begin{bmatrix}\mathbf{b}\\ \mathbf{c}\end{bmatrix}) = \mathbf{F}(\mathbf{b})\mathbf{c} - \mathbf{g}(\mathbf{b}) \text{ with } \mathbf{F}(\mathbf{b}) \in \mathbb{R}^{q\times p} \text{ and } \mathbf{g}(\mathbf{b}) \in \mathbb{R}^q \, .$$

Let $\mathbf{c}(\mathbf{b})$ denotes one solution of the above subproblem for a given $\mathbf{b} \in \mathbb{R}^n$ and formulate the problem

$$\min_{\mathbf{b}\in\mathbb{R}^n} \psi(\mathbf{b}) = \phi(\begin{bmatrix}\mathbf{b}\\ \mathbf{c}(\mathbf{b})\end{bmatrix}) = \frac{1}{2}\|\mathbf{r}(\begin{bmatrix}\mathbf{b}\\ \mathbf{c}(\mathbf{b})\end{bmatrix})\|_2^2 \, .$$

In doing that we have replaced our initial k-dimensional NLLS minimization problem by a n-dimensional one and we have separated the vector variables $\mathbf{b}$ and $\mathbf{c}$ [166][65]. This definition is also valid for a cost function $\phi(.)$ from $\mathbb{R}^{p\times k}$ into $\mathbb{R}$ and a residual real-matrix function $\mathbf{r}(.)$ from $\mathbb{R}^{p\times k}$ into $\mathbb{R}^{n\times m}$. Algorithms for minimizing a separable real function $\psi(.)$ with a vector or matrix argument are called variable projection methods [63][95][96][166][10][149].

3 Alternative and separable forms of the weighted low-rank approximation problem

In this section, we first provide some theoretical insights into the WLRA problem and the existence of solutions for it. Of course, some information on the subject is already available in the literature [125][33][171][62][167], but further investigations are clearly needed both theoretically and numerically, especially about the solvability of the WLRA problem. Moreover, the WLRA problem in its general form is much less well understood that the matrix completion or low-rank approximation problems [62][167]. We also explain how the WLRA problem can be reformulated in several different, but related, ways such that variable projection algorithms for separable NLLS problems [63][96][166][10] can be used to solve it efficiently even when the number of missing entries in the input matrix is high. Finally, we highlight the closed links between variable projection methods and Riemannian optimization on Grassmann manifolds [3][11], which are two seemingly different approaches often used to solve the WLRA problem numerically. Despite the similarity of the two frameworks has already been highlighted in some studies (e.g., [82]), the near equivalence of these two approaches (from a numerical point of view) in the context of the WLRA problem has not been well appreciated in the literature, probably because these two approaches have been developed in different communities [51][125][28][14][81][88].

3.1 Nonconvex formulations of the WLRA problem

A reasonable and efficient way to tackle the low-rank constraint in the formulation (P0) of the WLRA problem is to introduce a bilinear factorization model of the low-rank matrix solution as $\mathbf{Y} = \mathbf{AB}$ with $\mathbf{A} \in \mathbb{R}^{p\times k}$ and $\mathbf{B} \in \mathbb{R}^{k\times n}$ [57][72][171]. This non-convex bilinear formulation has a very long history in statistics [191][192][93] and has been revitalized recently for solving similar semi-definite problems [18]. This re-parametrization technique is justified by the fact that any matrix $\mathbf{Y}$ of rank at most k can be written as $\mathbf{Y} = \mathbf{AB}$, with $\mathbf{A} \in \mathbb{R}^{p\times k}$ and $\mathbf{B} \in \mathbb{R}^{k\times n}$ and that, reciprocally, any such matrix product $\mathbf{AB}$ is of rank at most k (see Subsection 2.1 for details). Note that a similar multiplicative formulation holds for the (Eckart-Young) Theorem 2.1, which solves the WLRA problem in the simple case where all elements of $\mathbf{W}$ are equal to one [57]. In recent decades, this bilinear factorization approach for low-rank matrix decomposition (often called the Burer-Monteiro factorization in the machine learning literature [18]) has also been the subject of intense research (for efficiency reasons) in solving large-scale convex optimization problems as this (nonconvex) reformulation of the original convex problems allows to drastically reduce the number of optimization variables from $p.n$ to $(p+n).k$, when k is small (e.g., $k \ll min(p,n)$), and, thus, allowing it to scale to problems with thousands or even millions of variables [86][156][117]. However, as we will illustrate below, this increased efficiency comes with a price as the intrinsic

bilinearity of the multiplicative (Burer-Monteiro) formulation makes the landscape and geometry of the factored objective functions much more complicated than the original (convex) ones with additional first-order critical and solution points that are not global optima of the factored optimization problems, which can be also badly-conditioned matrices [117].

We begin with the following well-known and simple result:

Theorem 3.1. For $\mathbf{X} \in \mathbb{R}^{p\times n}$, $\mathbf{W} \in \mathbb{R}_+^{p\times n}$ (i.e., $\mathbf{W}_{ij} \geq 0$), $\sqrt{\mathbf{W}} \in \mathbb{R}_+^{p\times n}$ with $\sqrt{\mathbf{W}}_{ij} = \sqrt{\mathbf{W}_{ij}}$ and any fixed integer $k \leq \mathit{rank}(\mathbf{X}) \leq \min(p, n)$, the problem (P0) is equivalent to the problem (P1):

$$\min_{\mathbf{A}\in\mathbb{R}^{p\times k},\, \mathbf{B}\in\mathbb{R}^{k\times n}} \qquad \varphi^*(\mathbf{A}, \mathbf{B}) = \frac{1}{2}\|\sqrt{\mathbf{W}} \odot (\mathbf{X} - \mathbf{AB})\|_F^2 \,. \tag{P1}$$

In other words, if we consider the range of $\varphi(.)$

$$\mathrm{C}_\varphi = \left\{ y \in \mathbb{R}_+ \,/\, \exists \mathbf{Y} \in \mathbb{R}_{\leq k}^{p\times n} \text{ with } y = \varphi(\mathbf{Y}) \right\} ,$$

and the range of $\varphi^*(.)$

$$\mathrm{C}_{\varphi^*} = \left\{ y \in \mathbb{R}_+ \,/\, \exists (\mathbf{A}, \mathbf{B}) \in \mathbb{R}^{p\times k} \times \mathbb{R}^{k\times n} \text{ with } y = \varphi^*(\mathbf{A}, \mathbf{B}) \right\} ,$$

these two subsets of $\mathbb{R}$ have the same infimum (e.g., greatest lower bound) and if this infimum is a minimum for one subset, the other subset also admits a minimum and these two minima are equal.

Proof. Since elements of the ranges C_φ and C_{φ^*} are sums of squares, they are bounded below by zero and both C_φ and C_{φ^*} admit an infimum greater or equal to zero, say $\bar{\mathbf{c}}_\varphi$ and $\bar{\mathbf{c}}_{\varphi^*}$, respectively. Now, we will demonstrate the stronger result $\mathrm{C}_\varphi = \mathrm{C}_{\varphi^*}$ in which case the assertions in the theorem are obvious.

Suppose first that $y \in \mathrm{C}_\varphi$. Then, $\exists \mathbf{Y} \in \mathbb{R}_{\leq k}^{p\times n}$ such that $y = \varphi(\mathbf{Y})$. Now let

$$\mathbf{Y} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$$

be the SVD of $\mathbf{Y}$, where it is assumed that $\mathbf{\Sigma}$ is a diagonal matrix with the singular values of $\mathbf{Y}$ arranged in decreasing order of magnitude in the diagonal. Since $\mathbf{Y}$ is of rank less than or equal to k, this SVD will have no more than k singular triplets with a singular value distinct from zero. Thus,

$$\mathbf{Y} = \mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^T \,,$$

where $\mathbf{U}_k$ and $\mathbf{V}_k$ stand for submatrices formed by the first k columns of $\mathbf{U}$ and $\mathbf{V}$, respectively, and $\mathbf{\Sigma}_k$ is the submatrix defined by the first k columns and rows of $\mathbf{\Sigma}$. Defining $\mathbf{A} = \mathbf{U}_k$ and $\mathbf{B} = \mathbf{\Sigma}_k\mathbf{V}_k^T$, $\mathbf{Y}$ can be factorized as

$$\mathbf{Y} = \mathbf{AB} \qquad \text{with } \mathbf{A} \in \mathbb{R}^{p\times k} \text{ and } \mathbf{B} \in \mathbb{R}^{k\times n} \,.$$

However, the equation $\mathbf{Y} = \mathbf{AB}$ implies that $y = \varphi(\mathbf{Y}) = \varphi^*(\mathbf{A}, \mathbf{B})$ and, thus, $y \in \mathrm{C}_{\varphi^*}$.

Reciprocally, assume that $y \in \mathrm{C}_{\varphi^*}$. Then, it exists $(\mathbf{A}, \mathbf{B}) \in \mathbb{R}^{p\times k} \times \mathbb{R}^{k\times n}$ such that $y = \varphi^*(\mathbf{A}, \mathbf{B})$. If we define $\mathbf{Y} = \mathbf{AB}$, we have $\mathit{rank}(\mathbf{Y}) \leq \min\big(\mathit{rank}(\mathbf{A}), \mathit{rank}(\mathbf{B})\big) \leq k$ according to equation (2.2) and we conclude that $\mathbf{Y} \in \mathbb{R}_{\leq k}^{p\times n}$. In these conditions, $y = \varphi^*(\mathbf{A}, \mathbf{B}) = \varphi(\mathbf{Y})$ and $y \in \mathrm{C}_\varphi$ and we are done.

□

Remark 3.1. Since any $p \times n$ matrix $\mathbf{Y}$ of rank at most k can also be written as $\mathbf{Y} = \mathbf{AB}$ with

$$\begin{aligned}
&1) \quad \mathbf{A} \in \mathbb{R}_k^{p\times k}, \mathbf{B} \in \mathbb{R}^{k\times n} \,,\\
&2) \quad \mathbf{A} \in \mathbb{R}^{p\times k}, \mathbf{B} \in \mathbb{R}_k^{k\times n} \,,\\
&3) \quad \mathbf{A} \in \mathbb{O}^{p\times k}, \mathbf{B} \in \mathbb{R}^{k\times n} \,,\\
&4) \quad \mathbf{A} \in \mathbb{R}^{p\times k}, \mathbf{B} \in \mathbb{O}_t^{k\times n} \,,
\end{aligned}$$

and, reciprocally, any of these $\mathbf{AB}$ matrix products is also of rank at most k and the range of $\varphi^*(.)$ is also equal to

$$\begin{aligned}\mathrm{C}_{\varphi^*} &= \big\{\mathbf{y} \in \mathbb{R}_+ \,/\, \exists(\mathbf{A},\mathbf{B}) \in \mathbb{R}_k^{p\times k} \times \mathbb{R}^{k\times n} \text{ and } \mathbf{y} = \varphi^*(\mathbf{A},\mathbf{B})\big\} \\ &= \big\{\mathbf{y} \in \mathbb{R}_+ \,/\, \exists(\mathbf{A},\mathbf{B}) \in \mathbb{R}^{p\times k} \times \mathbb{R}_k^{k\times n} \text{ and } \mathbf{y} = \varphi^*(\mathbf{A},\mathbf{B})\big\} \\ &= \big\{\mathbf{y} \in \mathbb{R}_+ \,/\, \exists(\mathbf{A},\mathbf{B}) \in \mathbb{O}^{p\times k} \times \mathbb{R}^{k\times n} \text{ and } \mathbf{y} = \varphi^*(\mathbf{A},\mathbf{B})\big\} \\ &= \big\{\mathbf{y} \in \mathbb{R}_+ \,/\, \exists(\mathbf{A},\mathbf{B}) \in \mathbb{R}^{p\times k} \times \mathbb{O}_t^{k\times n} \text{ and } \mathbf{y} = \varphi^*(\mathbf{A},\mathbf{B})\big\} \,.\end{aligned}$$

In these conditions, it is immediate that the problems (P0) and (P1) are also equivalent to the problems:

$$\begin{aligned} 1) \quad & \min_{\mathbf{A}\in\mathbb{R}_k^{p\times k},\, \mathbf{B}\in\mathbb{R}^{k\times n}} \quad \varphi^*(\mathbf{A},\mathbf{B}) \,, \\ 2) \quad & \min_{\mathbf{A}\in\mathbb{R}^{p\times k},\, \mathbf{B}\in\mathbb{R}_k^{k\times n}} \quad \varphi^*(\mathbf{A},\mathbf{B}) \,, \\ 3) \quad & \min_{\mathbf{A}\in\mathbb{O}^{p\times k},\, \mathbf{B}\in\mathbb{R}^{k\times n}} \quad \varphi^*(\mathbf{A},\mathbf{B}) \,, \\ 4) \quad & \min_{\mathbf{A}\in\mathbb{R}^{p\times k},\, \mathbf{B}\in\mathbb{O}_t^{k\times n}} \quad \varphi^*(\mathbf{A},\mathbf{B}) \,, \end{aligned}$$

where $\varphi^*(\mathbf{A},\mathbf{B}) = \frac{1}{2}\|\sqrt{\mathbf{W}} \odot (\mathbf{X} - \mathbf{AB})\|_F^2$ and we will use these alternative forms to demonstrate some important properties of the WLRA problem in this section and the followings. ■

Remark 3.2. By using the rank-nullity relationship (2.1) in Subsection 2.1, another way to tackle the low-rank constraint in the WLRA problem is to impose this low-rank constraint on the dimensions of the null space of $\mathbf{Y}$ (or $\mathbf{Y}^T$) instead on the range of $\mathbf{Y}$ (or $\mathbf{Y}^T$) as in the formulation (P0) [51][125][127][182]. Since, from equation (2.4), we have

$$\textit{null}(\mathbf{Y}) = \textit{ran}(\mathbf{Y}^T)^\perp \text{ and } \textit{null}(\mathbf{Y}^T) = \textit{ran}(\mathbf{Y})^\perp \,.$$

This is equivalent to impose the low-rank constraint on the dimensions of the orthogonal complements of $\textit{ran}(\mathbf{Y})$ or $\textit{ran}(\mathbf{Y}^T)$ and leads to what we will call the formulation (P2) of the WLRA problem, which has the following form if the low-rank constraint is imposed on the dimension of $\textit{ran}(\mathbf{Y})^\perp$

$$\min_{\mathbf{N}\in\mathbb{R}_{p-k}^{p\times(p-k)},\, \mathbf{Y}\in\mathbb{R}^{p\times n} \text{ with } \mathbf{N}^T\mathbf{Y}=\mathbf{0}^{(p-k)\times n}} \quad \varphi^{**}(\mathbf{N},\mathbf{Y}) = \frac{1}{2}\|\sqrt{\mathbf{W}} \odot (\mathbf{X} - \mathbf{Y})\|_F^2 \,, \qquad \text{(P2)}$$

or its transpose formulation (P2t), if the low-rank constraint is imposed on the dimension of $\textit{null}(\mathbf{Y}) = \textit{ran}(\mathbf{Y}^T)^\perp$,

$$\min_{\mathbf{N}\in\mathbb{R}_{n-k}^{n\times(n-k)},\, \mathbf{Y}\in\mathbb{R}^{p\times n} \text{ with } \mathbf{YN}=\mathbf{0}^{p\times(n-k)}} \quad \varphi^{**}(\mathbf{N},\mathbf{Y}) = \frac{1}{2}\|\sqrt{\mathbf{W}} \odot (\mathbf{X} - \mathbf{Y})\|_F^2 \,. \qquad \text{(P2t)}$$

If $p < n$, the formulation (P2) should be preferred as the number of parameters to be estimated is reduced and vice versa if $p > n$. Here, the rank constraint is imposed by the equalities

$$\mathbf{N}^T\mathbf{Y} = \mathbf{0}^{(p-k)\times n} \text{ and } \mathbf{YN} = \mathbf{0}^{p\times(n-k)} \,,$$

which are, respectively, equivalent to

$$\textit{dim}\big(\textit{null}(\mathbf{Y}^T)\big) \geq p - k \text{ and } \textit{dim}\big(\textit{null}(\mathbf{Y})\big) \geq n - k \,,$$

since all the columns of $\mathbf{N}$ belong to the null space of $\mathbf{Y}^T$, or $\mathbf{Y}$ in the second case, and $\mathbf{N}$ is of full column rank in both cases. Obviously, since by the rank-nullity relationship (2.1) we have

$$\textit{dim}\big(\textit{null}(\mathbf{Y}^T)\big) + \textit{rank}\big(\mathbf{Y}^T\big) = p, \textit{dim}\big(\textit{null}(\mathbf{Y})\big) + \textit{rank}\big(\mathbf{Y}\big) = n \text{ and } \textit{rank}(\mathbf{Y}^T) = \textit{rank}(\mathbf{Y}) \,,$$

this is equivalent in both cases to the rank constraint *rank*$(\mathbf{Y}) \leq k$, which is used in the formulation (P0) of the WLRA problem. Further inspection along the same lines of Theorem 3.1 will demonstrate that this formulation (P2) is also equivalent to the formulations (P0) and (P1). When $\mathbf{W} \in \mathbb{R}_{+*}^{p\times n}$, Edelman et al. [51] and Manton et al. [125] have proposed a Grassmann manifold framework to solve problem (P2) as the solution of this problem depends only on the span of the columns of $\mathbf{N}$. A Grassmann manifold is the collection of all linear subspaces of a given dimension in a particular Euclidean or Frobenius space, see Subsection 2.4 and [11] for a good introduction on manifolds and optimization on manifolds. Furthermore, they have described a large variety of first- and second-order algorithms for minimizing the cost function $\varphi^{**}(.)$ in this framework. As we will illustrate below, the solutions of the problem (P1) also do not depend on the individual elements of the matrices $\mathbf{A}$ and $\mathbf{B}$, but only on the range of $\mathbf{A}$ and, thus, can also be formulated as an optimization problem on the Grassmann manifold [47][14].

In these conditions, it is not difficult to recognize that each algorithm develops for minimizing $\varphi^{**}(.)$ (when $\mathbf{W} \in \mathbb{R}_{+*}^{p\times n}$) has a dual formulation for minimizing $\varphi^{*}(.)$ and vice versa, as determining the range of $\mathbf{A}$ leads implicitly to determine its orthogonal complement. In practical applications, the choice between an algorithm to minimize $\varphi^{*}(.)$ or its dual version to minimize $\varphi^{**}(.)$ will depend on the values of k, p and n. For small values of k, the formulation (P1) is likely to be more efficient as the size of the matrix variables will be smaller and, conversely, the formulation (P2) can be a better choice for large values of k as we will deal with smaller matrix variables when minimizing $\varphi^{**}(.)$. We will come back to these alternatives in the next sections. Finally, we mention that it is probably possible to extend the algorithms proposed by Manton et al. [125] to minimize the cost function $\varphi^{**}(.)$ to the case where $\mathbf{W} \in \mathbb{R}_{+}^{p\times n}$ instead of $\mathbb{R}_{+*}^{p\times n}$, see [28] for work in this direction. But, this is not pursued here, as in most applications, we use values of k which are much more smaller than $\min(p, n)$ for which the formulation (P1) is likely more economical. ■

Remark 3.3. A popular way to tackle the WLRA problem is also to consider the simpler problems:

$$\min_{\mathbf{Y}\in\mathbb{R}_k^{p\times n}} \quad \varphi(\mathbf{Y}) = \frac{1}{2}\|\sqrt{\mathbf{W}} \odot (\mathbf{X} - \mathbf{Y})\|_F^2 \,, \tag{3.1}$$

or

$$\min_{\mathbf{A}\in\mathbb{R}_k^{p\times k},\, \mathbf{B}\in\mathbb{R}_k^{k\times n}} \quad \varphi^{*}(\mathbf{A}, \mathbf{B}) = \frac{1}{2}\|\sqrt{\mathbf{W}} \odot (\mathbf{X} - \mathbf{AB})\|_F^2 \,, \tag{3.2}$$

which are equivalent (as *rank*$(\mathbf{AB}) = k$ if *rank*$(\mathbf{A}) =$ *rank*$(\mathbf{B}) = k$, see Subsection 2.1) and are also frequently solved by Riemannian optimization methods applied to smooth fixed-rank matrix manifolds [186][130][132] as the cost functions $\varphi(.)$ and $\varphi^{*}(.)$ are infinitely differentiable (e.g., of class C^{∞}) and the set $\mathbb{R}_k^{p\times n}$ is a smooth (C^{∞}) embedded submanifold of $\mathbb{R}^{p\times n}$ of dimension $(p + n - k)k$ (see Proposition 1.14 in Chap. 5 of [77], Example 8.14 of [106] or Section 7.5 in Chap. 7 of [11]). This approach is justified by the fact that $\mathbb{R}_k^{p\times n}$ is dense and open in $\mathbb{R}_{\leq k}^{p\times n}$ (see Theorem 2.3) meaning that with an initial guess in $\mathbb{R}_k^{p\times n}$, an iterate belonging to $\mathbb{R}_{<k}^{p\times n}$ or a non-smooth point of $\varphi(.)$ are both unlikely to occur in practice.

However, these two simpler problems are not mathematically equivalent to (P0) and (P1) for any choice of the weight matrix $\mathbf{W}$ as the submanifold $\mathbb{R}_k^{p\times n}$ is not closed in $\mathbb{R}^{p\times n}$ and a solution of these simpler problems may be on the frontier of $\mathbb{R}_k^{p\times n}$, which is $\mathbb{R}_{<k}^{p\times n}$, as stated in Theorem 2.3. This implies that these simpler problems may not admit a global minimizer, while such global minimizer will exist for problems (P0) and (P1) [33]. Furthermore, closedness of the domain is important in (non-convex) nonlinear optimization to garantee that the limit point of the iterative sequence is still in the domain of interest. As the set $\mathbb{R}_k^{p\times n}$ is not closed, some matrices in $\mathbb{R}_{<k}^{p\times n}$ can be the limit points of the iterative sequences in $\mathbb{R}_k^{p\times n}$ leading to so-called spurious critical points which do not belong to the smooth fixed-rank manifold $\mathbb{R}_k^{p\times n}$ [112]. Similarly, a sequence might also cross the frontier of $\mathbb{R}_k^{p\times n}$ at a certain iterate and the rank might fall below k breaking the sequence. For all these reasons, it is better to solve the WLRA problem over $\mathbb{R}_{\leq k}^{p\times n}$ rather than over $\mathbb{R}_k^{p\times n}$. Note, on the

other hand, that optimization algorithms on smooth fixed-rank manifolds are not strictly applicable on $\mathbb{R}^{p\times n}_{\leq k}$ as this set is a (non-smooth) real algebraic variety, not an embedded smooth submanifold of $\mathbb{R}^{p\times n}$ (see Proposition 1.1 in [22], Lecture 9 of [74] or [11][173] for details). More precisely, $\mathbf{Y} \in \mathbb{R}^{p\times n}_{\leq k}$ is, by definition, a space of matrices with a given upper bound on their ranks and is naturally an algebraic variety as the rank condition on a matrix is equivalent to the vanishing of all of its $(k+1, k+1)$-minors, which are polynomials of degree $k+1$. $\mathbb{R}^{p\times n}_{\leq k}$ is then defined as the solution set of polynomial equations therefore a so-called real determinantal variety [74]. Extending (first-order) optimization algorithms developed for smooth fixed-rank manifolds to real determinantal varieties like $\mathbb{R}^{p\times n}_{\leq k}$ is a very active area of research recently [173][112][145][143], but variable projection techniques, which are the focus of this monograph, can also be used for that purpose. ■

Thus, it is equivalent to minimize $\varphi(\mathbf{Y})$ or $\varphi^*(\mathbf{A}, \mathbf{B})$ for solving the WLRA problem. However, the WLRA problem (e.g., in the formulations (P0) and (P1)) has no known closed form solution in the general case and is known to be NP-hard [62] as already discussed in the Introduction 1. For certain classes of weighting matrices, a globally optimal solution can be found and one such class is obviously the unweighted case (e.g., $\mathbf{W}_{i,j} = 1$), since in that case the solution of the WLRA problem is given by the Eckart-Young Theorem 2.1. Another very important specialization of this is the case where all the elements of $\mathbf{W}$ are greater than 0 in which case it is possible to demonstrate that the WLRA problem has a well-defined solution as demonstrated in Theorem 3.3 below. Moreover, in the case where all the elements of $\mathbf{W}$ are greater than 0 and the rank of $\mathbf{W}$ is equal to 1, the solution of the WLRA problem can also be found via a generalization of the SVD in which we use diagonal metrics and scalar products different from the identity matrix in $\mathbb{R}^p$ and $\mathbb{R}^n$ (see Theorem 3 of [125] and also [62][167]). Finally, if $k = \mathit{rank}(\mathbf{X})$, the WLRA problem is equivalent to the $consistent$ matrix completion problem, which is to find one matrix $\widehat{\mathbf{X}}$ of rank at most k consistent with the observed entries (e.g., $\mathbf{W}_{ij} \neq 0$) of $\mathbf{X} \in \mathbb{R}^{p\times n}_{k}$ (e.g., the problem of recovering large matrices of low rank when most of the entries are unknown). In this case, the problem is also well-posed since $\mathbf{X}$ is obviously a solution to the consistent completion problem and we have $\varphi(\widehat{\mathbf{X}}) = 0$ for all solution matrices $\widehat{\mathbf{X}}$ [46][47]. In the general case, a very large variety of iterative methods have been previously suggested to solve the WLRA problem or convex and smooth proxies of it, especially in the framework of low-rank matrix completion, which is also NP-hard [30], and is the focus of lot of recent research [157][177][30][32][188][100][86][140]. Both the WLRA and matrix completion problems are also frequently recast as an optimization problem on smooth matrix manifolds as already noted above [3][125][169][47][14][11].

The formulation (P0) of the WLRA problem is well suited to derived theoretical properties of the WLRA problem such as the existence of solutions for this problem. On the other hand, the interest of the alternative formulation (P1) and its variants (see Remark 3.1), is that smaller matrices are manipulated and the introduction of the (non-unique) parameterization $\mathbf{Y} = \mathbf{AB}$ allows us to recast the WLRA problem as a standard unconstrained NLLS minimization problem as we will show below. This is particularly useful to derive practical algorithms to solve the WLRA problem as we will illustrate in the next sections.

Remark 3.4. The problem (P1) or its variants is over-parameterized. More precisely, if $\mathbf{C}$ is a $k \times k$ invertible matrix, we have

$$\mathbf{AB} = \mathbf{A}(\mathbf{C}\mathbf{C}^{-1})\mathbf{B} = (\mathbf{AC})(\mathbf{C}^{-1}\mathbf{B}) \quad \text{and} \quad \varphi^*(\mathbf{A}, \mathbf{B}) = \varphi^*(\mathbf{AC}, \mathbf{C}^{-1}\mathbf{B}) \,.$$

Consequently, the set of global minimizers of $\varphi^*(.)$ can be empty or infinite, but never finite or an isolated minimum implying that the Hessian of $\varphi^*(.)$ is at best positive semi-definite, but never positive definite, see Subsection2.4 for details. This can severely degrade the performance of standard optimization algorithms, which are mostly developed for isolated optima [45][139]. Furthermore, this scaling ambiguity tends to make the cost function $\varphi^*(.)$ of problem (P1) badly-conditioned, especially when the matrix $\mathbf{C}$ or its inverse is nearly singular. To overcome this difficulty, many authors have proposed to add different regularizers to $\varphi^*(.)$ as we will discussed later in this section.

Notice also that, if $\mathbf{A} \in \mathbb{R}_k^{p\times k}$ and $\mathbf{B} \in \mathbb{R}_k^{k\times n}$, these two full rank matrices have $p.k$ and $k.n$ degrees of freedom, respectively. However, specifying the matrix product $\mathbf{AB}$ in $\varphi^*(.)$ is equivalent to use the matrix product $(\mathbf{AC})(\mathbf{C}^{-1}\mathbf{B})$ for any $k \times k$ matrix $\mathbf{C}$ of rank k, which is equivalent to specify the column space of $\mathbf{AB}$. Hence, the matrix product $\mathbf{AB}$, or its column space, has only $p.k + k.n - k.k = (p + n - k).k$ degrees of freedom in general, which is consistent with the fact that the set $\mathbb{R}_k^{p\times n}$ is a smooth submanifold of $\mathbb{R}^{p\times n}$ of dimension $(n + p - k).k$ as already noted in Remark 3.3 above.

More generally, as all the matrix products $(\mathbf{AC})(\mathbf{C}^{-1}\mathbf{B})$ share the same column space, possibly remedies for the implicit over-parameterization in the formulation (P1) can be to recast the WLRA problem as an optimization problem on a Grassmann manifold [47][28][14][125][130][132] as discussed in Remark 3.3 or to use variable projection methods [158][27][150][147]. Moreover, these two seemingly different approaches for solving the WLRA problem are in fact tightly related as we will illustrate below. ■

The cost functions $\varphi(.)$ and $\varphi^*(.)$ are the composition of several infinitely differentiable functions on their respective domain of definition and, consequently, are also infinitely differentiable as smoothness is preserved by composition thanks to the standard chain rule [26]. Since $\varphi(.)$ and $\varphi^*(.)$ are smooth, they are also continuous on their respective domains. However, in the next theorem, we give a direct demonstration of the continuity of $\varphi(.)$ and $\varphi^*(.)$ by making clear that the WLRA problem differs from the standard low-rank approximation problem only by the choice of a different metric than the standard Frobenius metric on $\mathbb{R}^{p\times n}$. This metric is derived from the norm or seminorm induced by the choice of the weight matrix $\mathbf{W} \in \mathbb{R}_+^{p\times n}$.

Theorem 3.2. Using the same notations and definitions as in Theorem 3.1, the objective function defined in problem (P0)

$$\varphi : \mathbb{R}^{p\times n} \longrightarrow \mathbb{R} : \mathbf{Y} \mapsto \varphi(\mathbf{Y}) = \frac{1}{2}\|\sqrt{\mathbf{W}} \odot (\mathbf{X} - \mathbf{Y})\|_F^2 \,,$$

and the objective function defined in problem (P1)

$$\varphi^* : \mathbb{R}^{p\times k} \times \mathbb{R}^{k\times n} \longrightarrow \mathbb{R} : (\mathbf{A}, \mathbf{B}) \mapsto \varphi^*(\mathbf{A}, \mathbf{B}) = \frac{1}{2}\|\sqrt{\mathbf{W}} \odot (\mathbf{X} - \mathbf{AB})\|_F^2 \,,$$

are continuous on their respective domains of definition.

Proof. We first define a weighted norm or seminorm (if some of elements of $\mathbf{W}$ are equal to zero) of an $p \times n$ real matrix $\mathbf{Y}$ as

$$\|\mathbf{Y}\|_{\mathbf{W}}^2 = vec(\mathbf{Y})^T diag\big(vec(\mathbf{W})\big) vec(\mathbf{Y}) \,,$$

where $vec(\mathbf{Y})$ stands for the vectorized form of $\mathbf{Y}$, i.e., a vector formed by stacking the consecutive columns of $\mathbf{Y}$ in one $p.n$-dimensional vector (see equation (2.25) in Subsection 2.2). If none of the elements of $\mathbf{W}$ is equal to zero, $\|\|_{\mathbf{W}}$ is obviously a norm on $\mathbb{R}^{p\times n}$ and, as $\mathbb{R}^{p\times n}$ is a finite-dimensional vector space over $\mathbb{R}$, all norms on $\mathbb{R}^{p\times n}$ are equivalent, induce the same topology and are continuous functions on $\mathbb{R}^{p\times n}$ with respect to this topology [26][12]. On the other hand, if some of the elements of $\mathbf{W}$ are equal to zero, $\|\|_{\mathbf{W}}$ is only a seminorm on $\mathbb{R}^{p\times n}$, e.g., $\|\|_{\mathbf{W}}$ is a real-valued function : $\mathbb{R}^{p\times n} \longrightarrow \mathbb{R}$, which verifies, for all $\mathbf{Y}, \mathbf{Z} \in \mathbb{R}^{p\times n}$ and $\alpha \in \mathbb{R}$,

$$\begin{aligned}
&\|\mathbf{Y}\|_{\mathbf{W}} \geq 0 \,, \\
&\|\alpha\mathbf{Y}\|_{\mathbf{W}} = |\alpha|\|\mathbf{Y}\|_{\mathbf{W}} \,, \\
&\|\mathbf{Y} + \mathbf{Z}\|_{\mathbf{W}} \leq \|\mathbf{Y}\|_{\mathbf{W}} + \|\mathbf{Z}\|_{\mathbf{W}} \,.
\end{aligned}$$

However, even if $\|\|_{\mathbf{W}}$ is only a seminorm, it is still continuous with the respect to the unique topology on $\mathbb{R}^{p\times n}$ as demonstrated by Goldberg [58].

Now, $\varphi(\mathbf{Y})$ may be expressed as

$$\begin{aligned}\varphi(\mathbf{Y}) &= \frac{1}{2}\|\sqrt{\mathbf{W}} \odot (\mathbf{X} - \mathbf{Y})\|_F^2 \\ &= \frac{1}{2}\|vec\big(\sqrt{\mathbf{W}} \odot (\mathbf{X} - \mathbf{Y})\big)\|_2^2 \\ &= \frac{1}{2}\|vec(\sqrt{\mathbf{W}}) \odot vec(\mathbf{X} - \mathbf{Y})\|_2^2 \\ &= \frac{1}{2}vec(\mathbf{X} - \mathbf{Y})^T diag\big(vec(\mathbf{W})\big)vec(\mathbf{X} - \mathbf{Y}) \\ &= \frac{1}{2}\|\mathbf{X} - \mathbf{Y}\|_{\mathbf{W}}^2 \, .\end{aligned}$$

In other words, $\varphi(.)$ is the composition of the residual matrix function: $\mathbf{Y} \mapsto \mathbf{X} - \mathbf{Y}$, the norm or seminorm: $\mathbf{Z} \mapsto \|\mathbf{Z}\|_{\mathbf{W}}$ and the square function: $y \mapsto y^2$. As all these functions are continuous on their respective domain of definition, we conclude that $\varphi(.)$ is also continuous on $\mathbb{R}^{p\times n}$.

Similarly, $\varphi^*(\mathbf{A}, \mathbf{B})$ may be expressed as

$$\begin{aligned}\varphi^*(\mathbf{A}, \mathbf{B}) &= \frac{1}{2}\|\sqrt{\mathbf{W}} \odot (\mathbf{X} - \mathbf{AB})\|_F^2 \\ &= \frac{1}{2}vec(\mathbf{X} - \mathbf{AB})^T diag\big(vec(\mathbf{W})\big)vec(\mathbf{X} - \mathbf{AB}) \\ &= \frac{1}{2}\|\mathbf{X} - \mathbf{AB}\|_{\mathbf{W}}^2\end{aligned}$$

and $\varphi^*(.)$ is also the composition of several continuous functions on their respective domain of definition and, consequently, $\varphi^*(.)$ is also continuous on $\mathbb{R}^{p\times k} \times \mathbb{R}^{k\times n}$.

□

As $\varphi(.)$ is continuous on its domain of definition, it is not difficult to show that the problem (P0) has a well-defined solution when all the elements of the weight matrix $\mathbf{W}$ are strictly positive as stated in the next theorem.

Theorem 3.3. For $\mathbf{X} \in \mathbb{R}^{p\times n}$ different of the zero matrix of $\mathbb{R}^{p\times n}$ and $\mathbf{W} \in \mathbb{R}_{+*}^{p\times n}$ (i.e., $\mathbf{W}_{ij} > 0$), and any fixed integer $k \leq rank(\mathbf{X}) \leq \min(p, n)$, the set of global minimizers of $\varphi(\mathbf{Y})$ on $\mathbb{R}_{\leq k}^{p\times n}$ is nonempty and compact.

Proof. This theorem is a direct consequence of Theorem 3.1 stated without proof in Chu et al. [33], but we give a direct proof for completeness.

As $\mathbf{W} \in \mathbb{R}_{+*}^{p\times n}$ by hypothesis, we first observe that $\|\|_{\mathbf{W}}$ defines a norm on $\mathbb{R}^{p\times n}$. Let us now consider the closed ball with center $\mathbf{X}$ and radius $r = \|\mathbf{X}\|_{\mathbf{W}}$ with respect to this norm in $\mathbb{R}^{p\times n}$:

$$\bar{B}_{p\times n}(\mathbf{X}, r) = \big\{\mathbf{Y} \in \mathbb{R}^{p\times n} \text{ and } \|\mathbf{X} - \mathbf{Y}\|_{\mathbf{W}} \leq r\big\} \, .$$

$\bar{B}_{p\times n}(\mathbf{X}, r)$ is not empty as the zero matrix of $\mathbb{R}^{p\times n}$, which is also an element of $\mathbb{R}_{\leq k}^{p\times n}$, is in this closed ball. As $\mathbb{R}^{p\times n}$ is a finite-dimensional vector space, this closed ball is also a compact set (as it is by definition a bounded set). Furthermore, as $\mathbb{R}_{\leq k}^{p\times n}$ is closed in $\mathbb{R}^{p\times n}$ (see Theorem 2.3), the intersection of $\mathbb{R}_{\leq k}^{p\times n}$ and $\bar{B}_{p\times n}(\mathbf{X}, r)$ is also closed and bounded and, thus, compact in $\mathbb{R}^{p\times n}$. Now, as $\varphi(.)$ is continuous on $\mathbb{R}^{p\times n}$ and the image of a compact set by a continuous function is also compact, we conclude that $\varphi\big(\mathbb{R}_{\leq k}^{p\times n} \cap \bar{B}_{p\times n}(\mathbf{X}, r)\big) \subset \mathsf{C}_\varphi$ is a compact set in $\mathbb{R}$ and, thus, a closed and bounded interval of $\mathbb{R}$. Thus, $\varphi(.)$ attains its infimum on $\mathbb{R}_{\leq k}^{p\times n} \cap \bar{B}_{p\times n}(\mathbf{X}, r)$. In other words, it exists $\widehat{\mathbf{Y}} \in \mathbb{R}_{\leq k}^{p\times n} \cap \bar{B}_{p\times n}(\mathbf{X}, r)$ such that

$$\varphi(\widehat{\mathbf{Y}}) \leq \varphi(\mathbf{Y}) \, , \forall \mathbf{Y} \in \mathbb{R}_{\leq k}^{p\times n} \cap \bar{B}_{p\times n}(\mathbf{X}, r) \, .$$

It remains to show that $\varphi(\widehat{\mathbf{Y}}) = \bar{\mathbf{c}}_\varphi$ where $\bar{\mathbf{c}}_\varphi$ is the infimum of $\varphi(.)$ on $\mathbb{R}^{p\times n}_{\leq k}$, i.e., that $\widehat{\mathbf{Y}}$ is also a global minimizer of $\varphi(.)$ on $\mathbb{R}^{p\times n}_{\leq k}$. By definition of $\bar{\mathbf{c}}_\varphi$, we already have $\bar{\mathbf{c}}_\varphi \leq \varphi(\widehat{\mathbf{Y}})$ and it is sufficient to show that $\varphi(\widehat{\mathbf{Y}}) \leq \bar{\mathbf{c}}_\varphi$ to demonstrate the theorem.
Suppose on the contrary that $\varphi(\widehat{\mathbf{Y}}) > \bar{\mathbf{c}}_\varphi$, then it exists $\mathbf{Y} \in \mathbb{R}^{p\times n}_{\leq k}$ such that $\varphi(\widehat{\mathbf{Y}}) > \varphi(\mathbf{Y}) \geq \bar{\mathbf{c}}_\varphi$ by definition of $\bar{\mathbf{c}}_\varphi$. However, this implies that

$$\frac{1}{2}\|\sqrt{\mathbf{W}} \odot (\mathbf{X} - \mathbf{Y})\|_F^2 = \varphi(\mathbf{Y}) < \varphi(\widehat{\mathbf{Y}}) = \frac{1}{2}\|\sqrt{\mathbf{W}} \odot (\mathbf{X} - \widehat{\mathbf{Y}})\|_F^2 \,,$$

and it follows that

$$\|\mathbf{X} - \mathbf{Y}\|_{\mathbf{W}} < \|\mathbf{X} - \widehat{\mathbf{Y}}\|_{\mathbf{W}} \leq \|\mathbf{X}\|_{\mathbf{W}} = r \,.$$

In other words, $\mathbf{Y} \in \mathbb{R}^{p\times n}_{\leq k} \cap \bar{B}_{p\times n}(\mathbf{X}, r)$ and $\varphi(\mathbf{Y}) < \varphi(\widehat{\mathbf{Y}})$, which contradicts the assertion that $\widehat{\mathbf{Y}}$ is a minimizer of φ on $\mathbb{R}^{p\times n}_{\leq k} \cap \bar{B}_{p\times n}(\mathbf{X}, r)$ and we are done.

□

Remark 3.5. Using the equivalence between problems (P0) and (P1) stated in Theorem 3.1 above, we conclude that the set of global minimizers of $\varphi^*(.)$, when the weight matrix is strictly positive, is also nonempty. However, in the formulation (P1) of the WLRA problem, an important point to keep in mind is that, if the solution set is not empty, problem (P1) has an infinity of solutions as $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{B}}$ are not determined uniquely and we can normalize them in an arbitrary manner without changing the value of $\varphi^*(\widehat{\mathbf{A}}, \widehat{\mathbf{B}})$ (see Remark 3.4 above). Moreover, if $\alpha \in \mathbb{R}_*$, $(\alpha.\widehat{\mathbf{A}}, \frac{1}{\alpha}.\widehat{\mathbf{B}})$ is also a solution of (P1), which shows that the set of solutions in $\mathbb{R}^{p\times k} \times \mathbb{R}^{k\times n}$ is unbounded and, thus, not compact despite the set of global minimizers of $\varphi(.)$ is compact in $\mathbb{R}^{p\times n}$. ■

Remark 3.6. In the unweighted case (and with no missing values), the WLRA problem has an unique global minimum and all critical points of $\varphi(.)$ or $\varphi^*(.)$ which are not global minimizers are saddle points (e.g., critical points whose every neighborhood contains both "higher" and "smaller" points for $\varphi(.)$ or $\varphi^*(.)$), see Section 2.1 of [171] and Theorem 1.14 of [75] for details. In other words, $\varphi(.)$ or $\varphi^*(.)$ do not admit local minima in the unweighted case despite they are not convex functions.
While Theorem 3.3 shows that the WLRA problem has still well defined solutions when $\mathbf{W} \in \mathbb{R}^{p\times n}_{+*}$ because $\|\|_{\mathbf{W}}$ is a norm, several authors have illustrated by examples that $\varphi(.)$ or $\varphi^*(.)$ can have multiple local minima in addition to saddle points when the weights are all different of zero, but not uniform (see Section 2.1 of [171] and Example 1 of [62]). Such local minima emerge especially when the weights become significantly non-uniform (see Figure 1 of [171] for illustration). When $\mathbf{W}$ has zero entries, the situation is even worse as $\varphi(.)$ or $\varphi^*(.)$ may have multiple local minima [91], but the infimum of $\varphi(.)$ or $\varphi^*(.)$ can also be unattained, see Example 2 of [62] for illustration. ■

An alternative and insightful demonstration of the above theorem can also be given using the notion of the level sets of a continuous real function as defined in Chapter 4 of Ortega and Rheinboldt [148]. More precisely, for $\gamma \in \mathbb{R}$, the level set of $\varphi(.)$ at level γ is the set $L(\gamma) = \left\{\mathbf{Y} \in \mathbb{R}^{p\times n}_{\leq k} \,/\, \varphi(\mathbf{Y}) \leq \gamma\right\}$. In other words, $L(\gamma)$ is the subset of $\mathbb{R}^{p\times n}_{\leq k}$ whose elements $\mathbf{Y}$ verify the inequality $\varphi(\mathbf{Y}) \leq \gamma$. Obviously, $L(\gamma)$ is empty if $\gamma < \bar{\mathbf{c}}_\varphi$, where $\bar{\mathbf{c}}_\varphi$ is the infimum of $\varphi(.)$, and is the set of the global minimizers of $\varphi(.)$ if $\gamma = \bar{\mathbf{c}}_\varphi$ (which can be also empty in the general case where $\mathbf{W} \in \mathbb{R}^{p\times n}_{+}$ as discussed above).
As $\varphi(.)$ is continuous on the closed set $\mathbb{R}^{p\times n}_{\leq k}$ and the range of $\varphi(.)$, C_φ, is included in the non-negative half-space of $\mathbb{R}$, then every level set of $\varphi(.)$ at level γ for $\gamma \geq \bar{\mathbf{c}}_\varphi$ is closed in $\mathbb{R}^{p\times n}$ as the reciprocal image of the closed interval $[\bar{\mathbf{c}}_\varphi, \gamma]$ by a continuous and real function is also closed. Under these conditions, a necessary and sufficient condition for the set of global minimizers of $\varphi(.)$ to be nonempty and compact is that $\varphi(.)$ has a nonempty and bounded level set $L(\gamma)$ as this implies that $L(\gamma)$ is compact in $\mathbb{R}^{p\times n}$ (see Propositions 4.2.2 and 4.3.1 in Chap. 4 of [148]). However,

since $L(\gamma)$ is simply the intersection of $\mathbb{R}^{p\times n}_{\leq k}$ and the closed ball with center $\mathbf{X}$ and radius $\sqrt{2.\gamma}$ (with respect to the norm $\|\|_{\mathbf{W}}$) if all the elements of $\mathbf{W}$ are strictly positive, $L(\gamma)$ is nonempty and bounded by definition for all $\gamma > \bar{\mathbf{c}}_\varphi$. This also proves that the set of global minimizers of $\varphi(.)$ is nonempty and compact if all the elements of $\mathbf{W}$ are strictly positive as stated in Theorem 3.3.

In the more difficult case, where some elements of $\mathbf{W}$ are equal to zero, $\varphi(.)$ is still continuous as $\|\|_{\mathbf{W}}$ defines a seminorm on $\mathbb{R}^{p\times n}$ and every level set of $\varphi(.)$ is also automatically closed and the question of the existence of a global minimizer of $\varphi(.)$ reduces again to the existence of a bounded level set $L(\gamma)$ according to the previous discussion. However, in the case where some of the elements of $\mathbf{W}$ are equal to zero, the seminorm $\|\|_{\mathbf{W}}$ does not define the topology of $\mathbb{R}^{p\times n}$ [58] and the level set $L(\gamma)$ is not automatically bounded, so that the question of the existence of a nonempty and compact set of global minimizers is still unanswered in that case.

In order to discuss in more details, the existence of a nonempty and compact set of global minimizers of $\varphi(.)$ when some elements of $\mathbf{W}$ are equal to zero, let $\Omega \subset [p] \times [n]$ be the set of indices of the elements of $\mathbf{W}$ such that $\mathbf{W}_{ij} \neq 0$, where $[L] = [1, 2, ..., L]$. With this definition, from a weight matrix $\mathbf{W}$ with some zero elements and any $\lambda \in \mathbb{R}_{+*}$ (e.g., $\lambda > 0$), we can define a new $p \times n$ weight matrix $\mathbf{W}_\lambda$ as follows

$$\big[\mathbf{W}_\lambda\big]_{ij} = \begin{cases} \mathbf{W}_{ij} & \text{if } (i,j) \in \Omega \\ \lambda & \text{if } (i,j) \notin \Omega \end{cases} .$$

This new weight matrix $\mathbf{W}_\lambda$ induces a norm $\|\|_{\mathbf{W}_\lambda}$ on $\mathbb{R}^{p\times n}$, which is closely related to the seminorm $\|\|_{\mathbf{W}}$. More precisely, for any $\lambda \in \mathbb{R}_{+*}$ and $\mathbf{Y} \in \mathbb{R}^{p\times n}$, we have $\|\mathbf{Y}\|_{\mathbf{W}} \leq \|\mathbf{Y}\|_{\mathbf{W}_\lambda}$, which provides another simpler and different proof that $\|\|_{\mathbf{W}}$ is a continuous real-valued function on $\mathbb{R}^{p\times n}$ (see Theorem 3.2), and also

$$\lim_{\lambda\to 0} \|\mathbf{Y}\|_{\mathbf{W}_\lambda} = \|\mathbf{Y}\|_{\mathbf{W}} .$$

Furthermore, for any $\gamma \in \mathbb{R}_{+*}$ with $\gamma \geq \bar{\mathbf{c}}_\varphi$, we have the implications

$$\|\mathbf{X} - \mathbf{Y}\|_{\mathbf{W}_\lambda} \leq \sqrt{2.\gamma} \Rightarrow \|\mathbf{X} - \mathbf{Y}\|_{\mathbf{W}} \leq \sqrt{2.\gamma} \Rightarrow \varphi(\mathbf{Y}) \leq \gamma .$$

This shows that $\mathbb{R}^{p\times n}_{\leq k} \cap \bar{B}_{p\times n}(\mathbf{X}, \sqrt{2.\gamma}) \subset L(\gamma)$ where $\bar{B}_{p\times n}(\mathbf{X}, \sqrt{2.\gamma})$ is the closed ball of center $\mathbf{X}$ and radius $\sqrt{2.\gamma}$ with respect to the norm $\|\|_{\mathbf{W}_\lambda}$ on $\mathbb{R}^{p\times n}$. While the reciprocal inclusion $L(\gamma) \subset \bar{B}_{p\times n}(\mathbf{X}, \sqrt{2.\gamma})$ is obviously false in general, the fact that $\lim_{\lambda\to 0} \|\mathbf{Y}\|_{\mathbf{W}_\lambda} = \|\mathbf{Y}\|_{\mathbf{W}}$ suggests that for some weight matrices $\mathbf{W}$, it may still exist $\gamma \geq \bar{\mathbf{c}}_\varphi$ and $\lambda \in \mathbb{R}_{+*}$ sufficiently small such that $L(\gamma) \subset \bar{B}_{p\times n}(\mathbf{X}, \sqrt{2.\gamma})$ so that $L(\gamma) = \mathbb{R}^{p\times n}_{\leq k} \cap \bar{B}_{p\times n}(\mathbf{X}, \sqrt{2.\gamma})$ because of the imposed rank constraint on the $p \times n$ matrix $\mathbf{Y}$ in the formulation (P0). In such cases, $\varphi(.)$ will have a bounded level set and, consequently, the set of global minimizers of $\varphi(.)$ will be nonempty and compact.

3.2 Landscape connections of formulations P0 and P1 of the WLRA problem

As noted above, the cost functions $\varphi(.)$ and $\varphi^*(.)$ are obviously infinitely differentiable as they are polynomial functions of the entries of $\mathbf{Y}$ or $(\mathbf{A}, \mathbf{B})$, respectively. In these conditions, a natural and more modest question to ask, in addition of the existence of an absolute minimum of these cost functions, is the following: is there a connection between the first- and second-order critical points of $\varphi(.)$ and $\varphi^*(.)$?

To begin with, we first derive the gradient of $\varphi(.)$ at $\mathbf{Y} \in \mathbb{R}^{p\times n}$. We have the following differentiation rule for a differentiable function $g(.)$ defined from $\mathbb{R}^{p\times n}$ to $\mathbb{R}^{p\times n}$ and $\forall \mathbf{Y}, \mathbf{H} \in \mathbb{R}^{p\times n}$:

$$D\big(\mathbf{Y} \rightarrowtail \frac{1}{2}\|g(\mathbf{Y})\|_F^2\big)(\mathbf{Y})[\mathbf{H}] = \langle Dg(\mathbf{Y})[\mathbf{H}], g(\mathbf{Y})\rangle_F .$$

Here, we have $\varphi(\mathbf{Y}) = \frac{1}{2}\|g(\mathbf{Y})\|_F^2$ with $g(\mathbf{Y}) = \sqrt{\mathbf{W}} \odot (\mathbf{X} - \mathbf{Y})$, and we get

$$\begin{aligned} D\varphi(\mathbf{Y})[\mathbf{H}] &= \left\langle Dg(\mathbf{Y})[\mathbf{H}], g(\mathbf{Y}) \right\rangle_F \\ &= \left\langle \sqrt{\mathbf{W}} \odot -\mathbf{H}, \sqrt{\mathbf{W}} \odot (\mathbf{X} - \mathbf{Y}) \right\rangle_F \\ &= \left\langle \mathbf{W} \odot (\mathbf{Y} - \mathbf{X}), \mathbf{H} \right\rangle_F . \end{aligned}$$

By the unicity of the Frobenius gradient of $\varphi(.)$, this implies that

$$\nabla\varphi(\mathbf{Y}) = \mathbf{W} \odot (\mathbf{Y} - \mathbf{X}) \ , \ \forall \mathbf{Y} \in \mathbb{R}^{p\times n} . \tag{3.3}$$

In particular, the gradient of $\varphi(.)$ at $\mathbf{X}$ is $\nabla\varphi(\mathbf{X}) = \mathbf{W} \odot (\mathbf{X} - \mathbf{X}) = \mathbf{0}^{p\times n}$, which implies that $\mathbf{X}$ is a first-order critical point of $\varphi(.)$ if the feasible set is the whole linear space $\mathbb{R}^{p\times n}$. However, in most cases, especially when $\mathbf{W} \in \mathbb{R}_{+*}^{p\times n}$, $\mathbf{X}$ is the unique first-order critical point of $\varphi(.)$ considered as a function defined on the whole linear space $\mathbb{R}^{p\times n}$. In other words, and as expected from Subsection 2.4, for $\mathbf{Y} \in \mathbb{R}_{\leq k}^{p\times n}$, $\nabla\varphi(\mathbf{Y})$ cannot be used alone as a test of the optimality of $\mathbf{Y}$ in solving the WLRA problem (P0) because perturbations of $\mathbf{Y}$ which take it out of the feasible set $\mathbb{R}_{\leq k}^{p\times n}$ are not allowed and they may correspond to a decrease of the cost function $\varphi(.)$.

In general term, $\nabla^2\varphi(\mathbf{Y})$ is a 4^{th} order tensor of dimension $p \times n \times p \times n$, but $\nabla^2\varphi(\mathbf{Y})$ can also be viewed as a bilinear form $\big(\nabla^2\varphi(\mathbf{Y})\big)$ from $\mathbb{R}^{p\times n} \times \mathbb{R}^{p\times n}$ to $\mathbb{R}$ and also as a self-adjoint linear operator $\big[\nabla^2\varphi(\mathbf{Y})\big]$ from $\mathbb{R}^{p\times n}$ to $\mathbb{R}^{p\times n}$ (see Subsection 2.4 for details), and we have the equality

$$\big(\nabla^2\varphi(\mathbf{Y})\big)\big(\mathbf{C}, \mathbf{D}\big) = \langle \big[\nabla^2\varphi(\mathbf{Y})\big](\mathbf{C}), \mathbf{D}\rangle_F \ , \ \forall \mathbf{C}, \mathbf{D} \in \mathbb{R}^{p\times n} .$$

Taking into account the particular form of $\nabla\varphi(\mathbf{Y})$ derived in equation (3.3), we have simply

$$\big[\nabla^2\varphi(\mathbf{Y})\big](\mathbf{C}) = \mathbf{W} \odot \mathbf{C} \ , \ \forall \mathbf{C} \in \mathbb{R}^{p\times n} ,$$

and, thus, the bilinear form of $\nabla^2\varphi(\mathbf{Y})$ is defined by

$$\big(\nabla^2\varphi(\mathbf{Y})\big)\big(\mathbf{C}, \mathbf{D}\big) = \left\langle \mathbf{W} \odot \mathbf{C}, \mathbf{D} \right\rangle_F \ , \ \forall \mathbf{C}, \mathbf{D} \in \mathbb{R}^{p\times n} .$$

In particular, the Hessian quadratic form $\big(\nabla^2\varphi(\mathbf{Y})\big)$ for any $p \times n$ matrices $\mathbf{Y}$ and $\mathbf{C}$ is simply given by

$$\big(\nabla^2\varphi(\mathbf{Y})\big)\big(\mathbf{C}, \mathbf{C}\big) = \|\sqrt{\mathbf{W}} \odot \mathbf{C}\|_F^2 \geq 0 . \tag{3.4}$$

Thus, $\big(\nabla^2\varphi(\mathbf{Y})\big)$ is always positive semi-definite and is even always positive definite when $\mathbf{W} \in \mathbb{R}_{+*}^{p\times n}$.

In summary, for $\mathbf{Y} \in \mathbb{R}_{\leq k}^{p\times n}$, $\nabla\varphi(\mathbf{Y})$ and $\nabla^2\varphi(\mathbf{Y})$ cannot be used alone as test conditions for the global or local optimality of $\mathbf{Y}$ in solving the WLRA problem (P0) because in most settings the unconstrained local or global minimizers of $\varphi(.)$ do not satisfy the rank constrained *rank*$(\mathbf{Y}) \leq k$ and, also, for a given matrix $\mathbf{Y}$ of rank less than k, not all the search directions or perturbations have to be taken into account for determining the criticality conditions only those for which the rank constraint will be satisfied.

Thus, to continue with, we now characterize precisely the critical points of the rank-constrained minimization problem (P0) over the real-algebraic variety $\mathbb{R}_{\leq k}^{p\times n}$, which is a closed subset of the matrix space $\mathbb{R}^{p\times n}$ as stated in Theorem (2.3). To this end, we first identify $\mathbb{R}^{p\times n}$ and $\mathbb{R}^{p.n}$ with the two isomorphisms *vec*(.) and *mat*(.), defined in equations (2.25) and (2.26). Next, we note that the Euclidean scalar product in $\mathbb{R}^{p.n}$ and the Frobenius inner product in $\mathbb{R}^{p\times n}$ are intimately related since

$$\langle \mathbf{C}, \mathbf{D}\rangle_F = \mathrm{Tr}\left(\mathbf{C}^T\mathbf{D}\right) = \left\langle \mathit{vec}(\mathbf{C}), \mathit{vec}(\mathbf{D}) \right\rangle_2 \ , \ \forall \mathbf{C}, \mathbf{D} \in \mathbb{R}^{p\times n}$$

and, reciprocally,

$$\langle \mathbf{c}, \mathbf{d}\rangle_2 = \left\langle \mathit{mat}(\mathbf{c}), \mathit{mat}(\mathbf{d}) \right\rangle_F \ , \ \forall \mathbf{c}, \mathbf{d} \in \mathbb{R}^{p.n} .$$

Based on these considerations, it is rather straightforward to extend the notions of tangent vectors, Bouligand tangent and Frechet normal cones, and metric projection in $\mathbb{R}^{p.n}$ summarized in Subsection 2.4, especially, Theorem (2.6) and equations (2.59), to the case of the matrix space $\mathbb{R}^{p\times n}$.

Thus, a matrix $\mathbf{D} \in \mathbb{R}^{p\times n}$ is said to be tangent to $\mathbb{R}^{p\times n}_{\leq k}$ at $\bar{\mathbf{Y}} \in \mathbb{R}^{p\times n}_{\leq k}$ if there exist a matrix sequence $(\mathbf{Y}_i)_{i\in\mathbb{N}_*}$ in $\mathbb{R}^{p\times n}_{\leq k}$ tending to $\bar{\mathbf{Y}}$ and a real sequence $(\mathbf{t}_i)_{i\in\mathbb{N}_*}$ in $\mathbb{R}_{+*}$ tending to zero such that

$$\lim_{i\to\infty} \frac{(\mathbf{Y}_i - \bar{\mathbf{Y}})}{\mathbf{t}_i} = \mathbf{D}.$$

The set of all tangent matrices to $\mathbb{R}^{p\times n}_{\leq k}$, at $\bar{\mathbf{Y}}$ is a closed cone (see Theorem (2.4) for details), also called the Bouligand tangent cone to $\mathbb{R}^{p\times n}_{\leq k}$ at $\bar{\mathbf{Y}}$, and denoted by $\mathcal{T}^{\mathcal{B}}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{\leq k}$, similarly to the case of the vector space $\mathbb{R}^{p.n}$ discussed in Subsection 2.4. Its polar is defined by

$$(\mathcal{T}^{\mathcal{B}}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{\leq k})^o = \left\{\mathbf{D} \in \mathbb{R}^{p\times n} \;/\; \langle \mathbf{D}, \mathbf{Y}\rangle_F \leq 0 \;,\; \forall \mathbf{Y} \in \mathcal{T}^{\mathcal{B}}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{\leq k}\right\}$$

and is also a closed convex cone called the Frechet normal cone to $\mathbb{R}^{p\times n}_{\leq k}$ at $\bar{\mathbf{Y}}$, noted as $\mathcal{N}^{\mathcal{F}}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{\leq k}$, again similarly to the case of the vector space $\mathbb{R}^{p.n}$ discussed in Subsection 2.4.

Finally, a point $\bar{\mathbf{Y}} \in \mathbb{R}^{p\times n}_{\leq k}$ is a Frechet first-order stationary point for the WLRA problem (P0) if one of the following equivalent conditions is satisfied

$$\begin{aligned}
&\langle \nabla\varphi(\bar{\mathbf{Y}}), \mathbf{Y}\rangle_F \geq 0 \;,\; \forall \mathbf{Y} \in \mathcal{T}^{\mathcal{B}}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{\leq k} \;,\\
&-\nabla\varphi(\bar{\mathbf{Y}}) \in \mathcal{N}^{\mathcal{F}}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{\leq k} \;,\\
&P_{\mathcal{T}^{\mathcal{B}}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{\leq k}}(-\nabla\varphi(\bar{\mathbf{Y}})) = \{\mathbf{0}^{p\times n}\} \;,
\end{aligned} \tag{3.5}$$

where $\nabla\varphi(\bar{\mathbf{Y}})$ is given by equation (3.3) and $P_{\mathcal{T}^{\mathcal{B}}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{\leq k}}(-\nabla\varphi(\bar{\mathbf{Y}}))$ is the metric projection of the antigradient $-\nabla\varphi(\bar{\mathbf{Y}})$ onto $\mathcal{T}^{\mathcal{B}}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{\leq k}$ defined by

$$P_{\mathcal{T}^{\mathcal{B}}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{\leq k}}(-\nabla\varphi(\bar{\mathbf{Y}})) = \text{Arg}\min_{\mathbf{Y}\in\mathcal{T}^{\mathcal{B}}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{\leq k}} \|-\nabla\varphi(\bar{\mathbf{Y}}) - \mathbf{Y}\|^2_F \;.$$

Note that the set $P_{\mathcal{T}^{\mathcal{B}}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{\leq k}}(-\nabla\varphi(\bar{\mathbf{Y}}))$ is always nonempty as $\mathcal{T}^{\mathcal{B}}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{\leq k}$ is a closed cone, but it is not necessarily reduced to a singleton as $\mathcal{T}^{\mathcal{B}}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{\leq k}$ is not convex in $\mathbb{R}^{p\times n}$, see Subsection 2.4 for details.

However, $\forall \mathbf{Z} \in P_{\mathcal{T}^{\mathcal{B}}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{\leq k}}(-\nabla\varphi(\bar{\mathbf{Y}}))$, we have

$$\|\mathbf{Z}\|_F = \sqrt{\|-\nabla\varphi(\bar{\mathbf{Y}})\|^2_F - d(-\nabla\varphi(\bar{\mathbf{Y}}), \mathcal{T}^{\mathcal{B}}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{\leq k})^2} \;,$$

where the distance from $-\nabla\varphi(\bar{\mathbf{Y}})$ to $\mathcal{T}^{\mathcal{B}}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{\leq k}$ is given by

$$\begin{aligned}
d(-\nabla\varphi(\bar{\mathbf{Y}}), \mathcal{T}^{\mathcal{B}}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{\leq k}) &= \inf_{\mathbf{T}\in\mathcal{T}^{\mathcal{B}}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{\leq k}} \|-\nabla\varphi(\bar{\mathbf{Y}}) - \mathbf{T}\|_F\\
&= \min_{\mathbf{T}\in\mathcal{T}^{\mathcal{B}}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{\leq k}} \|-\nabla\varphi(\bar{\mathbf{Y}}) - \mathbf{T}\|_F \;,
\end{aligned}$$

again because $\mathcal{T}^{\mathcal{B}}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{\leq k}$ is a closed set. In other words, all elements of $P_{\mathcal{T}^{\mathcal{B}}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{\leq k}}(-\nabla\varphi(\bar{\mathbf{Y}}))$ have the same Frobenius norm and by the same small abuse of notation as used in equation (2.60) of Subsection 2.4, the Frechet first-order stationary condition for $\varphi(.)$ can be expressed as

$$\|P_{\mathcal{T}^{\mathcal{B}}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{\leq k}}(-\nabla\varphi(\bar{\mathbf{Y}}))\|_F = 0 \;,$$

where $P_{\mathcal{T}^{\mathcal{B}}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{\leq k}}(-\nabla\varphi(\bar{\mathbf{Y}}))$ designs now any of its elements. However, to use these results, we first need to find convenient practical expressions for $\mathcal{T}^{\mathcal{B}}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{\leq k}$ and the metric projection operator onto this closed set.

In order to derive a more convenient way for checking if a matrix $\bar{\mathbf{Y}} \in \mathbb{R}^{p\times n}_{\leq k}$ is a Frechet first-order stationary point for $\varphi(.)$, we first note that the set $\mathbb{R}^{p\times n}_{\leq k}$ stratifies into the set $\mathbb{R}^{p\times n}_{s}$ for $s = 1, \cdots, k$, e.g.,

$$\mathbb{R}^{p\times n}_{\leq k} = \bigcup_{s=1}^{k} \mathbb{R}^{p\times n}_{s} \ .$$

Furthermore, it is well-known, that each set $\mathbb{R}^{p\times n}_{s}$ is a smooth submanifold of dimension $(p+n-s).s$ embedded in $\mathbb{R}^{p\times n}$ and that its tangent space at $\mathbf{Y} \in \mathbb{R}^{p\times n}_{s}$ is given by

$$\begin{aligned}
\mathcal{T}_{\mathbf{Y}}\mathbb{R}^{p\times n}_{s} &= \big\{\mathbf{U_Y M V_Y^T} + \mathbf{U V_Y^T} + \mathbf{U_Y V}^T \ / \ \mathbf{M} \in \mathbb{R}^{s\times s}, \mathbf{U} \in \mathbb{R}^{p\times s}, \mathbf{V} \in \mathbb{R}^{n\times s} \\
&\qquad \text{with } \mathbf{U_Y^T U} = \mathbf{V_Y^T V} = \mathbf{0}^{s\times s}\big\} \\
&= \big\{[\mathbf{U_Y U_Y^\perp}] \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{0}^{(p-s)\times(n-s)} \end{bmatrix} [\mathbf{V_Y V_Y^\perp}]^T \\
&\qquad \text{with } \mathbf{A} \in \mathbb{R}^{s\times s}, \mathbf{B} \in \mathbb{R}^{s\times(n-s)} \text{ and } \mathbf{C} \in \mathbb{R}^{(p-s)\times s}\big\} \ ,
\end{aligned}$$

where $\mathbf{Y} = \mathbf{U_Y}\Sigma_{\mathbf{Y}}\mathbf{V_Y^T}$ is the thin SVD of $\mathbf{Y}$ with $\mathbf{U_Y^T U_Y} = \mathbf{V_Y^T V_Y} = \mathbf{I}_s$ and $\Sigma_{\mathbf{Y}}$ is a $s \times s$ diagonal matrix with strictly positive diagonal elements (e.g., the singular values of $\mathbf{Y}$), and $[\mathbf{U_Y U_Y^\perp}]$ and $[\mathbf{V_Y V_Y^\perp}]$ are, respectively, $p \times p$ and $n \times n$ orthogonal matrices. See Example 8.14 of Lee [106], Section 7.5 of Boumal [11] or Proposition 4.1 of Helmke and Shayman [78] for proofs and further details. Furthermore, the equivalence of the two definitions of the tangent space $\mathcal{T}_{\mathbf{Y}}\mathbb{R}^{p\times n}_{s}$ can be easily verified by direct computations.

Interestingly, if each $\mathbf{Y} \in \mathbb{R}^{p\times n}_{s}$ is identified by its singular triplets $(\mathbf{U_Y}, \Sigma_{\mathbf{Y}}, \mathbf{V_Y})$, then the first formulation of $\mathcal{T}_{\mathbf{Y}}\mathbb{R}^{p\times n}_{s}$ shows that, to represent an element of $\mathcal{T}_{\mathbf{Y}}\mathbb{R}^{p\times n}_{s}$, we only need to store the small matrices $\mathbf{M}$, $\mathbf{U}$ and $\mathbf{V}$. Furthermore, this formulation also shows that the elements of $\mathcal{T}_{\mathbf{Y}}\mathbb{R}^{p\times n}_{s}$ have a rank of at most $2.s$. On the other hand, the second formulation is useful for deriving the normal space to $\mathbb{R}^{p\times n}_{s}$ at $\mathbf{Y}$, noted $\mathcal{N}_{\mathbf{Y}}\mathbb{R}^{p\times n}_{s}$, which is the orthogonal complement of $\mathcal{T}_{\mathbf{Y}}\mathbb{R}^{p\times n}_{s}$ in $\mathbb{R}^{p\times n}$ with respect to the Frobenius inner product:

$$\mathcal{N}_{\mathbf{Y}}\mathbb{R}^{p\times n}_{s} = (\mathcal{T}_{\mathbf{Y}}\mathbb{R}^{p\times n}_{s})^{\perp} \ ,$$

and also the orthogonal projectors on both $\mathcal{T}_{\mathbf{Y}}\mathbb{R}^{p\times n}_{s}$ and $\mathcal{N}_{\mathbf{Y}}\mathbb{R}^{p\times n}_{s}$ as we will see now.

First, the second formulation reveals immediately the dimension of $\mathcal{T}_{\mathbf{Y}}\mathbb{R}^{p\times n}_{s}$ as

$$\mathit{dim}(\mathcal{T}_{\mathbf{Y}}\mathbb{R}^{p\times n}_{s}) = s.s + s.(p-s) + s.(n-s) = s.(p+n-s) \ .$$

Next, it is obvious from this formulation of $\mathcal{T}_{\mathbf{Y}}\mathbb{R}^{p\times n}_{s}$ that $\mathcal{N}_{\mathbf{Y}}\mathbb{R}^{p\times n}_{s}$ is equal to

$$\mathcal{N}_{\mathbf{Y}}\mathbb{R}^{p\times n}_{s} = \Big\{\mathbf{U_Y^\perp N}(\mathbf{V_Y^\perp})^T \ \text{ with } \mathbf{N} \in \mathbb{R}^{(p-s)\times(n-s)}\Big\} \ . \tag{3.6}$$

Obviously and as expected, we have

$$\mathit{dim}(\mathcal{N}_{\mathbf{Y}}\mathbb{R}^{p\times n}_{s}) = (p-s).(n-s) = p.n - s.(p+n-s) = \mathit{dim}(\mathbb{R}^{p\times n}) - \mathit{dim}(\mathcal{T}_{\mathbf{Y}}\mathbb{R}^{p\times n}_{s})$$

and the maximum rank of the matrix elements of $\mathcal{N}_{\mathbf{Y}}\mathbb{R}^{p\times n}_{s}$ is $\mathit{min}(p,n) - s$ according to equation (2.2). Next, by definition, the orthogonal projection of an arbitrary $\mathbf{Z} \in \mathbb{R}^{p\times n}$ onto $\mathcal{T}_{\mathbf{Y}}\mathbb{R}^{p\times n}_{s}$ satisfies both

$$\mathbf{Z} - \mathbf{P}_{\mathcal{T}_{\mathbf{Y}}\mathbb{R}^{p\times n}_{s}}(\mathbf{Z}) = \mathbf{U_Y^\perp N}(\mathbf{V_Y^\perp})^T \ ,$$

for some $\mathbf{N} \in \mathbb{R}^{(p-s)\times(n-s)}$, and

$$\mathbf{P}_{\mathcal{T}_{\mathbf{Y}}\mathbb{R}^{p\times n}_{s}}(\mathbf{Z}) = \mathbf{U_Y M V_Y^T} + \mathbf{U V_Y^T} + \mathbf{U_Y V}^T \ ,$$

for some $\mathbf{M} \in \mathbb{R}^{s\times s}, \mathbf{U} \in \mathbb{R}^{p\times s}$ and $\mathbf{V} \in \mathbb{R}^{n\times s}$ with $\mathbf{U}_\mathbf{Y}^T\mathbf{U} = \mathbf{V}_\mathbf{Y}^T\mathbf{V} = \mathbf{0}^{s\times s}$. Combined, these two statements imply that

$$\mathbf{Z} = \mathbf{U}_\mathbf{Y}\mathbf{M}\mathbf{V}_\mathbf{Y}^T + \mathbf{U}\mathbf{V}_\mathbf{Y}^T + \mathbf{U}_\mathbf{Y}\mathbf{V}^T + \mathbf{U}_\mathbf{Y}^\perp\mathbf{N}(\mathbf{V}_\mathbf{Y}^\perp)^T \, .$$

If we define now the orthogonal projectors associated with the column and row spaces of $\mathbf{Y}$ and their orthogonal complements

$$\mathbf{P}_\mathbf{U} = \mathbf{U}_\mathbf{Y}\mathbf{U}_\mathbf{Y}^T \, , \; \mathbf{P}_\mathbf{V} = \mathbf{V}_\mathbf{Y}\mathbf{V}_\mathbf{Y}^T \, , \; \mathbf{P}_\mathbf{U}^\perp = \mathbf{I}_p - \mathbf{P}_\mathbf{U} \text{ and } \mathbf{P}_\mathbf{V}^\perp = \mathbf{I}_n - \mathbf{P}_\mathbf{V} \, ,$$

we have, using orthogonal relationships,

$$\begin{aligned}
\mathbf{P}_\mathbf{U}\mathbf{Z}\mathbf{P}_\mathbf{V} &= \big(\mathbf{U}_\mathbf{Y}\mathbf{M}\mathbf{V}_\mathbf{Y}^T + \mathbf{U}_\mathbf{Y}\mathbf{V}^T\big)\mathbf{P}_\mathbf{V} = \mathbf{U}_\mathbf{Y}\mathbf{M}\mathbf{V}_\mathbf{Y}^T \, , \\
\mathbf{P}_\mathbf{U}^\perp\mathbf{Z}\mathbf{P}_\mathbf{V} &= \big(\mathbf{U}\mathbf{V}_\mathbf{Y}^T + \mathbf{U}_\mathbf{Y}^\perp\mathbf{N}(\mathbf{V}_\mathbf{Y}^\perp)^T\big)\mathbf{P}_\mathbf{V} = \mathbf{U}\mathbf{V}_\mathbf{Y}^T \, , \\
\mathbf{P}_\mathbf{U}\mathbf{Z}\mathbf{P}_\mathbf{V}^\perp &= \big(\mathbf{U}_\mathbf{Y}\mathbf{M}\mathbf{V}_\mathbf{Y}^T + \mathbf{U}_\mathbf{Y}\mathbf{V}^T\big)\mathbf{P}_\mathbf{V}^\perp = \mathbf{U}_\mathbf{Y}\mathbf{V}^T \, .
\end{aligned}$$

Using these results, we deduce that the orthogonal projector onto $\mathcal{T}_\mathbf{Y}\mathbb{R}_s^{p\times n}$ is given, equivalently, by

$$\begin{aligned}
\mathbf{P}_{\mathcal{T}_\mathbf{Y}\mathbb{R}_s^{p\times n}}(\mathbf{Z}) &= \mathbf{P}_\mathbf{U}\mathbf{Z}\mathbf{P}_\mathbf{V} + \mathbf{P}_\mathbf{U}^\perp\mathbf{Z}\mathbf{P}_\mathbf{V} + \mathbf{P}_\mathbf{U}\mathbf{Z}\mathbf{P}_\mathbf{V}^\perp \\
&= \mathbf{Z}\mathbf{P}_\mathbf{V} + \mathbf{P}_\mathbf{U}\mathbf{Z}\mathbf{P}_\mathbf{V}^\perp = \mathbf{Z}(\mathbf{V}_\mathbf{Y}\mathbf{V}_\mathbf{Y}^T) + (\mathbf{U}_\mathbf{Y}\mathbf{U}_\mathbf{Y}^T)\mathbf{Z}(\mathbf{I}_n - \mathbf{V}_\mathbf{Y}\mathbf{V}_\mathbf{Y}^T) \\
&= \mathbf{P}_\mathbf{U}\mathbf{Z} + \mathbf{P}_\mathbf{U}^\perp\mathbf{Z}\mathbf{P}_\mathbf{V} = (\mathbf{U}_\mathbf{Y}\mathbf{U}_\mathbf{Y}^T)\mathbf{Z} + (\mathbf{I}_p - \mathbf{U}_\mathbf{Y}\mathbf{U}_\mathbf{Y}^T)\mathbf{Z}(\mathbf{V}_\mathbf{Y}\mathbf{V}_\mathbf{Y}^T) \, ,
\end{aligned}$$

from which, we can also derive the orthogonal projector onto $\mathcal{N}_\mathbf{Y}\mathbb{R}_s^{p\times n}$ as

$$\begin{aligned}
\mathbf{P}_{\mathcal{N}_\mathbf{Y}\mathbb{R}_s^{p\times n}}(\mathbf{Z}) &= \mathbf{P}^\perp_{\mathcal{T}_\mathbf{Y}\mathbb{R}_s^{p\times n}}(\mathbf{Z}) = \mathbf{Z} - \mathbf{P}_{\mathcal{T}_\mathbf{Y}\mathbb{R}_s^{p\times n}}(\mathbf{Z}) \\
&= \mathbf{Z} - (\mathbf{U}_\mathbf{Y}\mathbf{U}_\mathbf{Y}^T)\mathbf{Z} - (\mathbf{I}_p - \mathbf{U}_\mathbf{Y}\mathbf{U}_\mathbf{Y}^T)\mathbf{Z}(\mathbf{V}_\mathbf{Y}\mathbf{V}_\mathbf{Y}^T) \\
&= (\mathbf{I}_p - \mathbf{U}_\mathbf{Y}\mathbf{U}_\mathbf{Y}^T)\mathbf{Z}(\mathbf{I}_n - \mathbf{V}_\mathbf{Y}\mathbf{V}_\mathbf{Y}^T) = \mathbf{P}_\mathbf{U}^\perp\mathbf{Z}\mathbf{P}_\mathbf{V}^\perp \, .
\end{aligned}$$

In these conditions, if $\mathbf{Z} \in \mathcal{N}_\mathbf{Y}\mathbb{R}_s^{p\times n}$, we have $\mathbf{P}_{\mathcal{N}_\mathbf{Y}\mathbb{R}_s^{p\times n}}(\mathbf{Z}) = \mathbf{Z}$, which implies that

$$\begin{aligned}
\mathbf{U}_\mathbf{Y}^T\mathbf{Z} &= \mathbf{U}_\mathbf{Y}^T(\mathbf{I}_p - \mathbf{U}_\mathbf{Y}\mathbf{U}_\mathbf{Y}^T)\mathbf{Z}(\mathbf{I}_n - \mathbf{V}_\mathbf{Y}\mathbf{V}_\mathbf{Y}^T) = \mathbf{0}^{s\times n} \, , \\
\mathbf{Z}\mathbf{V}_\mathbf{Y} &= (\mathbf{I}_p - \mathbf{U}_\mathbf{Y}\mathbf{U}_\mathbf{Y}^T)\mathbf{Z}(\mathbf{I}_n - \mathbf{V}_\mathbf{Y}\mathbf{V}_\mathbf{Y}^T)\mathbf{V}_\mathbf{Y} = \mathbf{0}^{p\times s} \, .
\end{aligned}$$

Reciprocally, if $\mathbf{Z} \in \mathbb{R}^{p\times n}$ with $\mathbf{U}_\mathbf{Y}^T\mathbf{Z} = \mathbf{0}^{s\times n}$ and $\mathbf{Z}\mathbf{V}_\mathbf{Y} = \mathbf{0}^{p\times s}$, we have

$$\begin{aligned}
\mathbf{P}_{\mathcal{T}_\mathbf{Y}\mathbb{R}_s^{p\times n}}(\mathbf{Z}) &= (\mathbf{Z}\mathbf{V}_\mathbf{Y})\mathbf{V}_\mathbf{Y}^T + \mathbf{U}_\mathbf{Y}(\mathbf{U}_\mathbf{Y}^T\mathbf{Z})(\mathbf{I}_n - \mathbf{V}_\mathbf{Y}\mathbf{V}_\mathbf{Y}^T) \\
&= \mathbf{0}^{p\times n} + \mathbf{0}^{p\times n} = \mathbf{0}^{p\times n}
\end{aligned}$$

and $\mathbf{Z} \in \mathcal{N}_\mathbf{Y}\mathbb{R}_s^{p\times n}$. In other words, we get an alternative formulation of $\mathcal{N}_\mathbf{Y}\mathbb{R}_s^{p\times n}$ as

$$\mathcal{N}_\mathbf{Y}\mathbb{R}_s^{p\times n} = \big\{\mathbf{Z} \in \mathbb{R}^{p\times n} \text{ with } \mathbf{U}_\mathbf{Y}^T\mathbf{Z} = \mathbf{0}^{s\times n} \text{ and } \mathbf{Z}\mathbf{V}_\mathbf{Y} = \mathbf{0}^{p\times s}\big\} \, ,$$

where the columns of $\mathbf{U}_\mathbf{Y}$ and $\mathbf{V}_\mathbf{Y}$ are, respectively, the leading s left and right singular vectors of $\mathbf{Y}$, which is of rank s.

Armed with these various results on the smooth manifold $\mathbb{R}_s^{p\times n}$ embedded in $\mathbb{R}^{p\times n}$, we can now reformulate the definitions of the Bouligand tangent cone to $\mathbb{R}_{\leq k}^{p\times n}$ at a matrix $\mathbf{Y}$ of rank $s \leq k$ and of the metric projection onto that closed set as follow.

Theorem 3.4. Let $\mathbf{Y} \in \mathbb{R}_{\leq k}^{p\times n}$ with $\mathit{rank}(\mathbf{Y}) = s \leq k$, the Bouligand tangent cone to $\mathbb{R}_{\leq k}^{p\times n}$ at $\mathbf{Y}$ is given by

$$\mathcal{T}_\mathbf{Y}^\mathcal{B}\mathbb{R}_{\leq k}^{p\times n} = \mathcal{T}_\mathbf{Y}\mathbb{R}_s^{p\times n} \oplus \big(\mathcal{N}_\mathbf{Y}\mathbb{R}_s^{p\times n} \cap \mathbb{R}_{\leq k-s}^{p\times n}\big) \, ,$$

where $\oplus$ stands for a direct orthogonal sum with respect to the Frobenius inner product in $\mathbb{R}^{p\times n}$.

In addition, the metric projection of an arbitrary $\mathbf{Z} \in \mathbb{R}^{p\times n}$ onto $\mathcal{T}_{\mathbf{Y}}^{\mathcal{B}}\mathbb{R}_{\leq k}^{p\times n}$ is given by

$$\mathbf{P}_{\mathcal{T}_{\mathbf{Y}}^{\mathcal{B}}\mathbb{R}_{\leq k}^{p\times n}}(\mathbf{Z}) = \mathbf{P}_{\mathcal{T}_{\mathbf{Y}}\mathbb{R}_{s}^{p\times n}}(\mathbf{Z}) + \mathbf{P}_{\mathbb{R}_{\leq k-s}^{p\times n}}\left(\mathbf{P}_{\mathcal{N}_{\mathbf{Y}}\mathbb{R}_{s}^{p\times n}}(\mathbf{Z})\right) ,$$

where $\mathbf{P}_{\mathcal{T}_{\mathbf{Y}}\mathbb{R}_{s}^{p\times n}}(.)$ and $\mathbf{P}_{\mathcal{N}_{\mathbf{Y}}\mathbb{R}_{s}^{p\times n}}(.)$ are the two unique complementary orthogonal projectors onto the linear subspaces $\mathcal{T}_{\mathbf{Y}}\mathbb{R}_{s}^{p\times n}$ and $\mathcal{N}_{\mathbf{Y}}\mathbb{R}_{s}^{p\times n}$, which are orthogonal to each other with respect to the Frobenius inner product in $\mathbb{R}^{p\times n}$, and where $\mathbf{P}_{\mathbb{R}_{\leq k-s}^{p\times n}}(.)$ is the metric projection onto the closed set $\mathbb{R}_{\leq k-s}^{p\times n}$.

Proof. For a proof, see Theorem 3.2 and Corollary 3.3 of Schneider and Uschmajew [173], Theorem 6.1 of Cason et al. [31] or Example 20.5 of Harris [74]. □

First note that, in Theorem 3.4, $\mathbf{P}_{\mathcal{T}_{\mathbf{Y}}^{\mathcal{B}}\mathbb{R}_{\leq k}^{p\times n}}(\mathbf{Z})$ is always a nonempty set as $\mathcal{T}_{\mathbf{Y}}^{\mathcal{B}}\mathbb{R}_{\leq k}^{p\times n}$ is a closed cone, but it is not neccessarily reduced to a singleton as $\mathcal{T}_{\mathbf{Y}}^{\mathcal{B}}\mathbb{R}_{\leq k}^{p\times n}$ is not convex. More precisely, the cardinality of $\mathbf{P}_{\mathcal{T}_{\mathbf{Y}}^{\mathcal{B}}\mathbb{R}_{\leq k}^{p\times n}}(\mathbf{Z})$ relies on the cardinality of the set $\mathbf{P}_{\mathbb{R}_{\leq k-s}^{p\times n}}\left(\mathbf{P}_{\mathcal{N}_{\mathbf{Y}}\mathbb{R}_{s}^{p\times n}}(\mathbf{Z})\right)$.

For an arbitrary $\mathbf{T} \in \mathbb{R}^{p\times n}$, the metric projection of $\mathbf{T}$ onto $\mathbb{R}_{\leq k-s}^{p\times n}$ is the set defined by

$$\mathbf{P}_{\mathbb{R}_{\leq k-s}^{p\times n}}\left(\mathbf{T}\right) = \text{Arg}\min_{\mathbf{Z}\in\mathbb{R}_{\leq k-s}^{p\times n}} \|\mathbf{T}-\mathbf{Z}\|_F .$$

Thus, the elements of $\mathbf{P}_{\mathbb{R}_{\leq k-s}^{p\times n}}(\mathbf{T})$ are easily determined with the help of the Eckart-Young Theorem 2.1 and are the best approximation of rank at most $k-s$ of $\mathbf{T}$ with respect to the Frobenius norm. In other words, $\mathbf{P}_{\mathbb{R}_{\leq k-s}^{p\times n}}\left(\mathbf{P}_{\mathcal{N}_{\mathbf{Y}}\mathbb{R}_{s}^{p\times n}}(\mathbf{Z})\right)$ is single-valued when

$$\sigma_{k-s}\left(\mathbf{P}_{\mathcal{N}_{\mathbf{Y}}\mathbb{R}_{s}^{p\times n}}(\mathbf{Z})\right) > \sigma_{k-s+1}\left(\mathbf{P}_{\mathcal{N}_{\mathbf{Y}}\mathbb{R}_{s}^{p\times n}}(\mathbf{Z})\right) ,$$

in which case its unique element is given by the truncated SVD of rank $k-s$ of $\mathbf{P}_{\mathcal{N}_{\mathbf{Y}}\mathbb{R}_{s}^{p\times n}}(\mathbf{Z})$ according to Theorem 2.1), or, when,

$$\sigma_{k-s}\left(\mathbf{P}_{\mathcal{N}_{\mathbf{Y}}\mathbb{R}_{s}^{p\times n}}(\mathbf{Z})\right) = 0 ,$$

in which case

$$\mathbf{P}_{\mathbb{R}_{\leq k-s}^{p\times n}}\left(\mathbf{P}_{\mathcal{N}_{\mathbf{Y}}\mathbb{R}_{s}^{p\times n}}(\mathbf{Z})\right) = \mathbf{P}_{\mathcal{N}_{\mathbf{Y}}\mathbb{R}_{s}^{p\times n}}(\mathbf{Z}) .$$

In the above equations, $\sigma_i(\mathbf{T})$ denotes the i^{th} largest singular value of the matrix $\mathbf{T}$. Furthermore, when $\mathbf{P}_{\mathbb{R}_{\leq k-s}^{p\times n}}\left(\mathbf{P}_{\mathcal{N}_{\mathbf{Y}}\mathbb{R}_{s}^{p\times n}}(\mathbf{Z})\right)$ is single-valued then $\mathbf{P}_{\mathcal{T}_{\mathbf{Y}}^{\mathcal{B}}\mathbb{R}_{\leq k}^{p\times n}}(\mathbf{Z})$ is also single-valued according to Theorem 3.4.

In order to clarify the practical meaning of Theorem 3.4, it is now useful to distinguish the two cases *rank*$(\mathbf{Y}) = s = k$ and *rank*$(\mathbf{Y}) = s < k$.

Obviously, in the first case, when $s = k$, we get

$$\mathcal{T}_{\mathbf{Y}}^{\mathcal{B}}\mathbb{R}_{\leq k}^{p\times n} = \mathcal{T}_{\mathbf{Y}}\mathbb{R}_{k}^{p\times n}$$

and

$$\mathbf{P}_{\mathcal{T}_{\mathbf{Y}}^{\mathcal{B}}\mathbb{R}_{\leq k}^{p\times n}}(.) = \mathbf{P}_{\mathcal{T}_{\mathbf{Y}}\mathbb{R}_{k}^{p\times n}}(.) .$$

In words, when *rank*$(\mathbf{Y}) = k$, the Bouligand tangent cone to $\mathbb{R}_{\leq k}^{p\times n}$ at $\mathbf{Y}$ coincides with the tangent linear space to $\mathbb{R}_{k}^{p\times n}$ at $\mathbf{Y}$. Furthermore, the metric projection onto this Bouligand tangent cone is nothing else then the orthogonal projector onto the tangent linear space to $\mathbb{R}_{k}^{p\times n}$ at $\mathbf{Y}$. Finally, from these results, we deduce immediately that the Frechet normal cone to $\mathbb{R}_{\leq k}^{p\times n}$ at $\mathbf{Y}$, which is defined as the polar of $\mathcal{T}_{\mathbf{Y}}^{\mathcal{B}}\mathbb{R}_{\leq k}^{p\times n}$, also coincides with the normal space to $\mathbb{R}_{k}^{p\times n}$ at $\mathbf{Y}$ (e.g., the orthogonal

complement of $\mathcal{T}_{\mathbf{Y}}\mathbb{R}_k^{p\times n}$ in $\mathbb{R}^{p\times n}$ with respect to the Frobenius inner product) when $\mathit{rank}(\mathbf{Y}) = k$. If $\mathbf{Z} \in \mathcal{N}_{\mathbf{Y}}^{\mathcal{F}}\mathbb{R}_{\leq k}^{p\times n} = (\mathcal{T}_{\mathbf{Y}}^{\mathcal{B}}\mathbb{R}_{\leq k}^{p\times n})^o = (\mathcal{T}_{\mathbf{Y}}\mathbb{R}_k^{p\times n})^o$ then, by definition,

$$\langle \mathbf{Z}, \mathbf{Q}\rangle_F \leq 0 \ , \ \forall \mathbf{Q} \in \mathcal{T}_{\mathbf{Y}}\mathbb{R}_k^{p\times n}.$$

However, since $\mathcal{T}_{\mathbf{Y}}\mathbb{R}_k^{p\times n}$ is a linear space, if $\mathbf{Q} \in \mathcal{T}_{\mathbf{Y}}\mathbb{R}_k^{p\times n}$ then $-\mathbf{Q}$ also belongs to $\mathcal{T}_{\mathbf{Y}}\mathbb{R}_k^{p\times n}$, from which we deduce

$$\langle \mathbf{Z}, \mathbf{Q}\rangle_F \geq 0 \ , \ \forall \mathbf{Q} \in \mathcal{T}_{\mathbf{Y}}\mathbb{R}_k^{p\times n}$$

and we get the equivalences

$$\mathbf{Z} \in \mathcal{N}_{\mathbf{Y}}^{\mathcal{F}}\mathbb{R}_{\leq k}^{p\times n} \Longleftrightarrow \langle \mathbf{Z}, \mathbf{Q}\rangle_F = 0 \ , \ \forall \mathbf{Q} \in \mathcal{T}_{\mathbf{Y}}\mathbb{R}_k^{p\times n} \Longleftrightarrow \mathbf{Z} \in (\mathcal{T}_{\mathbf{Y}}\mathbb{R}_k^{p\times n})^{\perp} = \mathcal{N}_{\mathbf{Y}}\mathbb{R}_k^{p\times n} \ .$$

Summarizing the preceding results, when $\mathbf{Y} \in \mathbb{R}_k^{p\times n}$ and $\mathbf{Z} \in \mathbb{R}^{p\times n}$, we have

$$\begin{aligned}
\mathcal{T}_{\mathbf{Y}}^{\mathcal{B}}\mathbb{R}_{\leq k}^{p\times n} &= \mathcal{T}_{\mathbf{Y}}\mathbb{R}_k^{p\times n} \\
&= \big\{\mathbf{U}_{\mathbf{Y}}\mathbf{M}\mathbf{V}_{\mathbf{Y}}^T + \mathbf{U}\mathbf{V}_{\mathbf{Y}}^T + \mathbf{U}_{\mathbf{Y}}\mathbf{V}^T \ / \ \mathbf{M} \in \mathbb{R}^{k\times k}, \mathbf{U} \in \mathbb{R}^{p\times k}, \mathbf{V} \in \mathbb{R}^{n\times k} \\
&\qquad \text{with } \mathbf{U}_{\mathbf{Y}}^T\mathbf{U} = \mathbf{V}_{\mathbf{Y}}^T\mathbf{V} = \mathbf{0}^{k\times}\big\}, \\
\mathcal{N}_{\mathbf{Y}}^{\mathcal{F}}\mathbb{R}_{\leq k}^{p\times n} &= (\mathcal{T}_{\mathbf{Y}}\mathbb{R}_k^{p\times n})^{\perp} = \mathcal{N}_{\mathbf{Y}}\mathbb{R}_k^{p\times n} \\
&= \big\{\mathbf{T} \in \mathbb{R}^{p\times n} \text{ with } \mathbf{U}_{\mathbf{Y}}^T\mathbf{T} = \mathbf{0}^{k\times n} \text{ and } \mathbf{T}\mathbf{V}_{\mathbf{Y}} = \mathbf{0}^{p\times k}\big\} \ , \qquad (3.7) \\
\mathbf{P}_{\mathcal{T}_{\mathbf{Y}}^{\mathcal{B}}\mathbb{R}_{\leq k}^{p\times n}}(\mathbf{Z}) &= \mathbf{P}_{\mathcal{T}_{\mathbf{Y}}\mathbb{R}_k^{p\times n}}(\mathbf{Z}) \\
&= \mathbf{Z}(\mathbf{V}_{\mathbf{Y}}\mathbf{V}_{\mathbf{Y}}^T) + (\mathbf{U}_{\mathbf{Y}}\mathbf{U}_{\mathbf{Y}}^T)\mathbf{Z}(\mathbf{I}_n - \mathbf{V}_{\mathbf{Y}}\mathbf{V}_{\mathbf{Y}}^T) \\
&= (\mathbf{U}_{\mathbf{Y}}\mathbf{U}_{\mathbf{Y}}^T)\mathbf{Z} + (\mathbf{I}_p - \mathbf{U}_{\mathbf{Y}}\mathbf{U}_{\mathbf{Y}}^T)\mathbf{Z}(\mathbf{V}_{\mathbf{Y}}\mathbf{V}_{\mathbf{Y}}^T) \ , \\
\mathbf{P}_{\mathcal{N}_{\mathbf{Y}}^{\mathcal{F}}\mathbb{R}_{\leq k}^{p\times n}}(\mathbf{Z}) &= \mathbf{Z} - \mathbf{P}_{\mathcal{T}_{\mathbf{Y}}\mathbb{R}_k^{p\times n}}(\mathbf{Z}) = \mathbf{P}_{\mathcal{N}_{\mathbf{Y}}\mathbb{R}_k^{p\times n}}(\mathbf{Z}) \\
&= (\mathbf{I}_p - \mathbf{U}_{\mathbf{Y}}\mathbf{U}_{\mathbf{Y}}^T)\mathbf{Z}(\mathbf{I}_n - \mathbf{V}_{\mathbf{Y}}\mathbf{V}_{\mathbf{Y}}^T) \ ,
\end{aligned}$$

where the columns of $\mathbf{U}_{\mathbf{Y}}$ and $\mathbf{V}_{\mathbf{Y}}$ are, respectively, the leading k left and right singular vectors of $\mathbf{Y}$, which is of rank k.

These results are further consistent with the more general result that, when $\mathcal{M}$ is an arbitrary submanifold embedded in $\mathbb{R}^{p\times n}$ or $\mathbb{R}^p$, its tangent and normal spaces at an arbitrary $\mathbf{Z} \in \mathcal{M}$ coincide exactly with the Bouligand tangent and Frechet normal cones to $\mathcal{M}$ at $\mathbf{Z}$, see Example 6.8 of Rockafellar and Wets [165] or Theorem 3.15 of Ruszczynski [160] for details.

Furthermore, from the above results, we see that $\bar{\mathbf{Y}} \in \mathbb{R}_k^{p\times n}$ is a Frechet first-order stationary point for the WLRA problem in its formulation (P0) if it satisfies one of the following equivalent conditions:

$$(1)\langle \nabla\varphi(\bar{\mathbf{Y}}), \mathbf{Z}\rangle_F \geq 0 \ , \forall \mathbf{Z} \in \big\{\mathbf{U}_{\bar{\mathbf{Y}}}\mathbf{M}\mathbf{V}_{\bar{\mathbf{Y}}}^T + \mathbf{U}\mathbf{V}_{\bar{\mathbf{Y}}}^T + \mathbf{U}_{\bar{\mathbf{Y}}}\mathbf{V}^T / \mathbf{M} \in \mathbb{R}^{k\times k}, \mathbf{U} \in \mathbb{R}^{p\times k}, \mathbf{V} \in \mathbb{R}^{n\times k}\big\} \ ,$$

$$(2)\mathbf{U}_{\bar{\mathbf{Y}}}^T\nabla\varphi(\bar{\mathbf{Y}}) = \mathbf{0}^{k\times n} \text{ and } \nabla\varphi(\bar{\mathbf{Y}})\mathbf{V}_{\bar{\mathbf{Y}}} = \mathbf{0}^{p\times k} \ ,$$

$$(3)\nabla\varphi(\bar{\mathbf{Y}})(\mathbf{V}_{\bar{\mathbf{Y}}}\mathbf{V}_{\bar{\mathbf{Y}}}^T) + (\mathbf{U}_{\bar{\mathbf{Y}}}\mathbf{U}_{\bar{\mathbf{Y}}}^T)\nabla\varphi(\bar{\mathbf{Y}})(\mathbf{I}_n - \mathbf{V}_{\bar{\mathbf{Y}}}\mathbf{V}_{\bar{\mathbf{Y}}}^T) = \mathbf{0}^{p\times n} \ ,$$

where the columns of $\mathbf{U}_{\bar{\mathbf{Y}}}$ and $\mathbf{V}_{\bar{\mathbf{Y}}}$ are, respectively, the leading k left and right singular vectors of $\bar{\mathbf{Y}}$, which is of rank k.

Obviously, the second condition is the more convenient for our purpose and, as $\nabla\varphi(\bar{\mathbf{Y}}) = \mathbf{W} \odot (\bar{\mathbf{Y}} - \mathbf{X})$ according to equation (3.3), it translates to the simple statement

$$\mathbf{U}_{\bar{\mathbf{Y}}}^T\big(\mathbf{W} \odot (\bar{\mathbf{Y}} - \mathbf{X})\big) = \mathbf{0}^{k\times n} \text{ and } \big(\mathbf{W} \odot (\bar{\mathbf{Y}} - \mathbf{X})\big)\mathbf{V}_{\bar{\mathbf{Y}}} = \mathbf{0}^{p\times k} \ . \qquad (3.8)$$

We now consider the case where $\bar{\mathbf{Y}} \in \mathbb{R}^{p\times n}_{<k}$, e.g., when $\mathit{rank}(\bar{\mathbf{Y}}) = s < k$. In that case, we deduce from Theorem 3.4 that the structure of $\mathcal{T}^{\mathcal{B}}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{\leq k}$ is more complex as it contains $\mathcal{T}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{s}$, but also matrices of rank less or equal to $k-s$ which intersect orthogonally $\mathbb{R}^{p\times n}_{\leq k}$ (with respect the Frobenius inner product) and also sum of elements belonging to each of these two sets.

A key-remark for deriving a simple condition of Frechet first-order stationarity for $\varphi(.)$ at a point $\bar{\mathbf{Y}} \in \mathbb{R}^{p\times n}_{<k}$ is the following. Assume that $\mathit{rank}(\bar{\mathbf{Y}}) = s < k$ and consider an arbitrary matrix $\mathbf{Z} \in \mathbb{R}^{p\times n}_{\leq k-s}$. We have

$$\mathbf{Z} = \mathbf{P}_{\mathcal{T}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{s}}(\mathbf{Z}) + \mathbf{P}_{\mathcal{N}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{s}}(\mathbf{Z}) \, ,$$

since $\mathbf{P}_{\mathcal{T}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{s}}(.)$ and $\mathbf{P}_{\mathcal{N}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{s}}(.)$ are two complementary orthogonal projectors with respect to the Frobenius inner product in $\mathbb{R}^{p\times n}$. Clearly, by definition, $\mathbf{P}_{\mathcal{T}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{s}}(\mathbf{Z}) \in \mathcal{T}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{s}$ and $\mathbf{P}_{\mathcal{N}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{s}}(\mathbf{Z}) \in \mathcal{N}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{s}$. Furthermore, as the orthogonal projector $\mathbf{P}_{\mathcal{N}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{s}}(.)$ never increases the rank of a matrix, we also have $\mathbf{P}_{\mathcal{N}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{s}}(\mathbf{Z}) \in \mathbb{R}^{p\times n}_{\leq k-s}$ as $\mathbf{Z} \in \mathbb{R}^{p\times n}_{\leq k-s}$, and we conclude that

$$\mathbf{P}_{\mathcal{N}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{s}}(\mathbf{Z}) \in \mathcal{N}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{s} \cap \mathbb{R}^{p\times n}_{\leq k-s} \, ,$$

which implies finally that $\mathbf{Z} \in \mathcal{T}^{\mathcal{B}}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{\leq k}$. In other words, we have the inclusion $\mathbb{R}^{p\times n}_{\leq k-s} \subset \mathcal{T}^{\mathcal{B}}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{\leq k}$. From this relationship and Theorem 3.4, it is not difficult to see that an equivalent formulation of $\mathcal{T}^{\mathcal{B}}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{\leq k}$ is

$$\mathcal{T}^{\mathcal{B}}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{\leq k} = \mathcal{T}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{s} + \mathbb{R}^{p\times n}_{\leq k-s} \, ,$$

where the direct orthogonal sum $\oplus$ is now replaced by an ordinary sum, see Hosseini et al. [84] for more details.

Now, if $\mathbf{Z} \in \mathbb{R}^{p\times n}_{\leq k-s}$, $-\mathbf{Z}$ also belongs to $\mathbb{R}^{p\times n}_{\leq k-s}$ and, thus, any element of $\mathcal{N}^{\mathcal{F}}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{\leq k}$ needs to be orthogonal to $\mathbf{Z}, \forall \mathbf{Z} \in \mathbb{R}^{p\times n}_{\leq k-s}$. Next, if $\mathbf{T} \in \mathcal{N}^{\mathcal{F}}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{\leq k}$ and $\mathbf{T} \neq \mathbf{0}^{p\times n}$, this implies that $\mathbf{T}$ must be orthogonal (with respect to the Frobenius inner product in $\mathbb{R}^{p\times n}$) to its best approximation of rank $k-s$ given by the Eckart-Young Theorem 2.1, which is absurd, and we conclude that

$$\mathcal{N}^{\mathcal{F}}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{\leq k} = \{\mathbf{0}^{p\times n}\} \, .$$

In this condition, if $\bar{\mathbf{Y}} \in \mathbb{R}^{p\times n}_{<k}$, the Frechet first-order stationary condition for $\varphi(.)$ at $\bar{\mathbf{Y}} \in \mathbb{R}^{p\times n}_{<k}$ reduces to

$$\nabla\varphi(\bar{\mathbf{Y}}) = \mathbf{0}^{p\times n} \, ,$$

which translates to the simple matrix equality $\mathbf{W} \odot (\bar{\mathbf{Y}} - \mathbf{X}) = \mathbf{0}^{p\times n}$, using equation (3.3).

Collecting all the above developments, we have demonstrated the following theorem, which is used without proof in Ha et al. [83] in a slightly larger setting where $\varphi(.)$ is a continuously differentiable function instead of the objective function associated with the formulation (P0) of the WLRA problem.

Theorem 3.5. Let $\bar{\mathbf{Y}} \in \mathbb{R}^{p\times n}_{\leq k}$, with $\mathit{rank}(\bar{\mathbf{Y}}) = s \leq k$. Then $\bar{\mathbf{Y}}$ is a Frechet first-order stationary point for $\varphi(.)$ if

$$\mathbf{U}^{T}_{\bar{\mathbf{Y}}}\big(\mathbf{W} \odot (\bar{\mathbf{Y}} - \mathbf{X})\big) = \mathbf{0}^{k\times n} \text{ and } \big(\mathbf{W} \odot (\bar{\mathbf{Y}} - \mathbf{X})\big)\mathbf{V}_{\bar{\mathbf{Y}}} = \mathbf{0}^{p\times k} \, ,$$

when $\mathit{rank}(\bar{\mathbf{Y}}) = s = k$, or if

$$\mathbf{W} \odot (\bar{\mathbf{Y}} - \mathbf{X}) = \mathbf{0}^{p\times n} \, ,$$

when $\mathit{rank}(\bar{\mathbf{Y}}) = s < k$ and the columns of $\mathbf{U}_{\bar{\mathbf{Y}}}$ and $\mathbf{V}_{\bar{\mathbf{Y}}}$ are, respectively, the leading s left and right singular vectors of $\bar{\mathbf{Y}}$, which is of rank s.
□

Any local minimizer $\bar{\mathbf{Y}}$ of $\varphi(.)$ in the set $\mathbb{R}^{p\times n}_{\leq k}$ must satisfy the first-order conditions stated in Theorem 3.5, though these conditions are not sufficient in general, see Theorem 6.12 in Rockafellar and Wets [165] and also Ha et al. [83] for more details. However, in the case where *rank*$(\bar{\mathbf{Y}}) < k$ and $\mathbf{W} \odot (\bar{\mathbf{Y}} - \mathbf{X}) = \mathbf{0}^{p\times n}$, we deduce immediately that

$$\sqrt{\mathbf{W}} \odot (\bar{\mathbf{Y}} - \mathbf{X}) = \mathbf{0}^{p\times n}$$

and $\bar{\mathbf{Y}}$ is obviously a global minimizer of $\varphi(.)$ and a solution of the WLRA problem in this particular case.

We now derive a more convenient expression than the one given in Theorem 2.9 to verify that a matrix $\bar{\mathbf{Y}} \in \mathbb{R}^{p\times n}_{\leq k}$ is a Frechet second-order stationarity point of $\varphi(.)$ over $\mathbb{R}^{p\times n}_{\leq k}$. First, we observe that, in the case where $\bar{\mathbf{Y}} \in \mathbb{R}^{p\times n}_{<k}$ is a Frechet first-order stationarity point of $\varphi(.)$, $\nabla\varphi(\bar{\mathbf{Y}}) = \mathbf{0}^{p\times n}$ and $\bar{\mathbf{Y}} \in \mathbb{R}^{p\times n}_{<k}$ is a global minimum of $\varphi(.)$ over $\mathbb{R}^{p\times n}$ and, thus, $\big(\nabla^2\varphi(\bar{\mathbf{Y}})\big)$ is a positive semi-definite quadratic form over $\mathbb{R}^{p\times n}$ (note, alternatively, that $\big(\nabla^2\varphi(\bar{\mathbf{Y}})\big)$ is always positive semi-definite according to equation (3.4)). From these results, when *rank*$(\bar{\mathbf{Y}}) < k$ is a Frechet first-order stationarity point of $\varphi(.)$, we deduce immediately that the condition (2.61) in Theorem 2.9 is verified and consistently $\bar{\mathbf{Y}}$ is also a Frechet second-order stationarity point of $\varphi(.)$ in the sense of Theorem 2.9.

Next, in the case where $\bar{\mathbf{Y}} \in \mathbb{R}^{p\times n}_{k}$, we first recall from equations (3.7) that

$$\mathcal{T}^{\mathcal{B}}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{\leq k} = \mathcal{T}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{k} \text{ and } \mathcal{N}^{\mathcal{F}}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{\leq k} = \mathcal{N}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{k}$$

and the Frechet first-order condition is thus equivalent to

$$\nabla\varphi(\bar{\mathbf{Y}}) \in \mathcal{N}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{k} \, ,$$

which is exactly similar to the statement that the Riemannian gradient of $\varphi(.)$ at $\bar{\mathbf{Y}} \in \mathbb{R}^{p\times n}_{k}$ is equal to zero by equation (2.50). In other words, in the case *rank*$(\bar{\mathbf{Y}}) = k$, the Frechet first-order condition for $\varphi(.)$, considered as a function defined on $\mathbb{R}^{p\times n}_{\leq k}$, at $\bar{\mathbf{Y}}$ stated in equation (3.8) is equivalent to the Riemannian first-order condition for the restriction of $\varphi(.)$ over the embeddded smooth submanifold $\mathbb{R}^{p\times n}_{k}$ at $\bar{\mathbf{Y}}$ stated in equation (2.50).

Furthermore, when *rank*$(\bar{\mathbf{Y}}) = k$ and $\bar{\mathbf{Y}}$ is a Frechet first-order stationary point, equation (2.61) in Theorem (2.9), specialized to the case of $\mathbb{R}^{p\times n}_{\leq k}$, simplifies to

$$\langle\nabla\varphi(\bar{\mathbf{Y}}), \mathbf{Z}\rangle_F + \langle\big[\nabla^2\varphi(\bar{\mathbf{Y}})\big](\mathbf{D}), \mathbf{D}\rangle_F \geq 0 \, , \; \forall \mathbf{D} \in \mathcal{T}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{k}, \forall \mathbf{Z} \in \mathcal{T}_{(\bar{\mathbf{Y}},\mathbf{D})}\mathbb{R}^{p\times n}_{k} \, .$$

This condition is strictly equivalent to the statement that the Riemannian Hessian of $\varphi(.)$ at $\bar{\mathbf{Y}}$, $\big(\nabla^2_R\varphi(\bar{\mathbf{Y}})\big)$, is positive semi-definite over $\mathcal{T}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{k}$, as noted in [197] and [110]. Next, using the explicit formulation (in terms of standard Euclidean derivatives) of this Riemannian Hessian of the smooth function $\varphi(.)$ defined on the smooth submanifold $\mathbb{R}^{p\times n}_{k}$, derived in Proposition 2.2 of [186] and Proposition 2 of [115], the statement that $\big(\nabla^2_R\varphi(\bar{\mathbf{Y}})\big)$ is positive semi-definite is equivalent to

$$\big(\nabla^2_R\varphi(\bar{\mathbf{Y}})\big)[\mathbf{D},\mathbf{D}] = \big(\nabla^2\varphi(\bar{\mathbf{Y}})\big)[\mathbf{D},\mathbf{D}] + 2.\langle\nabla\varphi(\bar{\mathbf{Y}}), \mathbf{U}^{\perp}_{\bar{\mathbf{Y}}}\mathbf{C}\Sigma^{-1}_{\bar{\mathbf{Y}}}\mathbf{B}(\mathbf{V}^{\perp}_{\bar{\mathbf{Y}}})^T\rangle_F \geq 0 \, , \tag{3.9}$$

$\forall \mathbf{D} \in \mathcal{T}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{k}$, and where the thin SVD of $\bar{\mathbf{Y}} \in \mathbb{R}^{p\times n}_{k}$ is given by $\bar{\mathbf{Y}} = \mathbf{U}_{\bar{\mathbf{Y}}}\Sigma_{\bar{\mathbf{Y}}}\mathbf{V}^T_{\bar{\mathbf{Y}}}$ with $\mathbf{U}^T_{\bar{\mathbf{Y}}}\mathbf{U}_{\bar{\mathbf{Y}}} = \mathbf{V}^T_{\bar{\mathbf{Y}}}\mathbf{V}_{\bar{\mathbf{Y}}} = \mathbf{I}_k$ and $\Sigma_{\bar{\mathbf{Y}}}$ is a $k \times k$ diagonal matrix with strictly positive diagonal elements (e.g., the singular values of $\bar{\mathbf{Y}}$) and

$$\mathbf{D} = [\mathbf{U}_{\bar{\mathbf{Y}}}\mathbf{U}^{\perp}_{\bar{\mathbf{Y}}}]\begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{0}^{(p-k)\times(n-k)}\end{bmatrix}[\mathbf{V}_{\bar{\mathbf{Y}}}\mathbf{V}^{\perp}_{\bar{\mathbf{Y}}}]^T \in \mathcal{T}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_{k} \, ,$$

where $\mathbf{A} \in \mathbb{R}^{k\times k}, \mathbf{B} \in \mathbb{R}^{k\times(n-k)}, \mathbf{C} \in \mathbb{R}^{(p-k)\times k}$ and $[\mathbf{U}_{\bar{\mathbf{Y}}}\mathbf{U}^{\perp}_{\bar{\mathbf{Y}}}]$ and $[\mathbf{V}_{\bar{\mathbf{Y}}}\mathbf{V}^{\perp}_{\bar{\mathbf{Y}}}]$ are, respectively, $p \times p$ and $n \times n$ orthogonal matrices.

Using equations (3.3) and (3.4), the previous discussion leads to the following theorem, which characterizes more explicitly the Frechet second-order stationarity points of $\varphi(.)$.

Theorem 3.6. Let $\bar{\mathbf{Y}} \in \mathbb{R}^{p\times n}_{\leq k}$, with $\mathit{rank}(\bar{\mathbf{Y}}) \leq k$. Then $\bar{\mathbf{Y}}$ is a Frechet second-order stationary point for $\varphi(.)$ if it is a Frechet first-order stationary point for $\varphi(.)$ and if, in addition, in the case of $\mathit{rank}(\bar{\mathbf{Y}}) = k$, if

$$\|\sqrt{\mathbf{W}} \odot \mathbf{D}\|_F^2 \geq -2.\langle \mathbf{W} \odot (\bar{\mathbf{Y}} - \mathbf{X}), \mathbf{U}_{\bar{\mathbf{Y}}}^{\perp}\mathbf{C}\Sigma_{\bar{\mathbf{Y}}}^{-1}\mathbf{B}(\mathbf{V}_{\bar{\mathbf{Y}}}^{\perp})^T\rangle_F \ , \ \forall \mathbf{D} \in \mathcal{T}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_k \ , \tag{3.10}$$

where the thin SVD of $\bar{\mathbf{Y}} \in \mathbb{R}^{p\times n}_k$ is given by $\bar{\mathbf{Y}} = \mathbf{U}_{\bar{\mathbf{Y}}}\Sigma_{\bar{\mathbf{Y}}}\mathbf{V}_{\bar{\mathbf{Y}}}^T$, the columns of $\mathbf{U}_{\bar{\mathbf{Y}}}^{\perp} \in \mathbb{R}^{p\times(p-k)}$ and $\mathbf{V}_{\bar{\mathbf{Y}}}^{\perp} \in \mathbb{R}^{n\times(n-k)}$ form, respectively, orthonormal bases of $\mathit{ran}(\bar{\mathbf{Y}})^{\perp}$ and $\mathit{ran}(\bar{\mathbf{Y}}^T)^{\perp}$ and

$$\mathbf{D} = [\mathbf{U}_{\bar{\mathbf{Y}}}\mathbf{U}_{\bar{\mathbf{Y}}}^{\perp}]\begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{0}^{(p-k)\times(n-k)}\end{bmatrix}[\mathbf{V}_{\bar{\mathbf{Y}}}\mathbf{V}_{\bar{\mathbf{Y}}}^{\perp}]^T \in \mathcal{T}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_k \ ,$$

where $\mathbf{A} \in \mathbb{R}^{k\times k}, \mathbf{B} \in \mathbb{R}^{k\times(n-k)}, \mathbf{C} \in \mathbb{R}^{(p-k)\times k}$.
□

Interestingly, observe that, using equation (3.6), when $\bar{\mathbf{Y}} \in \mathbb{R}^{p\times n}_k$ is a Frechet first-order stationary point for $\varphi(.)$, both $\nabla\varphi(\bar{\mathbf{Y}}) = \mathbf{W} \odot (\bar{\mathbf{Y}} - \mathbf{X})$ and $\mathbf{U}_{\bar{\mathbf{Y}}}^{\perp}\mathbf{C}\Sigma_{\bar{\mathbf{Y}}}^{-1}\mathbf{B}(\mathbf{V}_{\bar{\mathbf{Y}}}^{\perp})^T$ are elements of $\mathcal{N}_{\bar{\mathbf{Y}}}\mathbb{R}^{p\times n}_k$ as $\mathbf{C}\Sigma_{\bar{\mathbf{Y}}}^{-1}\mathbf{B} \in \mathbb{R}^{(n-k)\times(n-k)}$.

We now characterize the critical points of the factorized cost function $\varphi^*(.)$, which is used in the formulation (P1) of the WLRA problem. $\varphi^*(.)$ is defined on the product space $\mathbb{R}^{p\times k} \times \mathbb{R}^{k\times n}$, which is a "standard" Euclidean linear (product) space. In other words, the gradient, Hessian and critical points of $\varphi^*(.)$ are defined in the usual way (see Subsection 2.4) as there are no additional constraints on the matrix variables $\mathbf{A} \in \mathbb{R}^{p\times k}$ and $\mathbf{B} \in \mathbb{R}^{k\times n}$. Thus, the pair $(\mathbf{A}, \mathbf{B})$ is a first-order stationary point of $\varphi^*(.)$, if and only if,

$$\nabla\varphi^*(\mathbf{A}, \mathbf{B}) = (\mathbf{0}^{p\times k}, \mathbf{0}^{k\times n}) \ ,$$

and a second-order stationary point of $\varphi^*(.)$ if, in addition,

$$\big(\nabla^2\varphi^*(\mathbf{A}, \mathbf{B})\big)\big((\mathbf{C}, \mathbf{D}), (\mathbf{C}, \mathbf{D})\big) \geq 0 \ , \ \forall (\mathbf{C}, \mathbf{D}) \in \mathbb{R}^{p\times k} \times \mathbb{R}^{k\times n} \ ,$$

where the second derivative (Hessian) $\big(\nabla^2\varphi^*(\mathbf{A}, \mathbf{B})\big)$ is a (symmetric) quadratic form mapping from $(\mathbb{R}^{p\times k} \times \mathbb{R}^{k\times n}) \times (\mathbb{R}^{p\times k} \times \mathbb{R}^{k\times n})$ to $\mathbb{R}$.

By definition, $\varphi^*(.)$ is the composition of $\varphi(.)$, from $\mathbb{R}^{p\times n}$ to $\mathbb{R}$, with the bilinear mapping, from $\mathbb{R}^{p\times k} \times \mathbb{R}^{k\times n}$ to $\mathbb{R}^{p\times n}$, defined by $(\mathbf{A}, \mathbf{B}) \longrightarrow \mathbf{AB}$, $\forall (\mathbf{A}, \mathbf{B}) \in \mathbb{R}^{p\times k} \times \mathbb{R}^{k\times n}$. Furthermore, $\varphi(.)$ and this bilinear mapping are C^{∞} differentiable. Thus, using the standard chain rule on the differential of the composition of two differentiable functions, we can easily obtain the two partial derivatives of $\varphi^*(.)$ since, $\forall (\mathbf{C}, \mathbf{D}) \in \mathbb{R}^{p\times k} \times \mathbb{R}^{k\times n}$, we have, using properties of the $\mathrm{Tr}(.)$ operator stated in Subsection 2.1,

$$\begin{aligned}
D\varphi^*_{\mathbf{A}}(\mathbf{A}, \mathbf{B})(\mathbf{C}) &= \big\langle \nabla\varphi(\mathbf{AB}), \mathbf{CB}\big\rangle_F \\
&= \mathrm{Tr}\left(\nabla\varphi(\mathbf{AB})^T\mathbf{CB}\right) \\
&= \mathrm{Tr}\left(\mathbf{CB}\nabla\varphi(\mathbf{AB})^T\right) \\
&= \mathrm{Tr}\left(\mathbf{C}\big(\nabla\varphi(\mathbf{AB})\mathbf{B}^T\big)^T\right) \\
&= \mathrm{Tr}\left(\big(\nabla\varphi(\mathbf{AB})\mathbf{B}^T\big)^T\mathbf{C}\right) \\
&= \big\langle \nabla\varphi(\mathbf{AB})\mathbf{B}^T, \mathbf{C}\big\rangle_F
\end{aligned}$$

and, similarly,

$$\begin{aligned}
D\varphi^*_{\mathbf{B}}(\mathbf{A}, \mathbf{B})(\mathbf{D}) &= \big\langle \nabla\varphi(\mathbf{AB}), \mathbf{AD}\big\rangle_F \\
&= \mathrm{Tr}\left(\nabla\varphi(\mathbf{AB})^T\mathbf{AD}\right) \\
&= \mathrm{Tr}\left(\big(\mathbf{A}^T\nabla\varphi(\mathbf{AB})\big)^T\mathbf{D}\right) \\
&= \big\langle \mathbf{A}^T\nabla\varphi(\mathbf{AB}), \mathbf{D}\big\rangle_F \ .
\end{aligned}$$

Thus, by the unicity of the Frobenius gradients of the partial functions $\varphi^*_{\mathbf{A}}(.)$ and $\varphi^*_{\mathbf{B}}(.)$, and equation (3.3), we get

$$\begin{aligned}\nabla\varphi^*_{\mathbf{A}}(\mathbf{A},\mathbf{B}) &= \nabla\varphi(\mathbf{AB})\mathbf{B}^T = \big(\mathbf{W}\odot(\mathbf{AB}-\mathbf{X})\big)\mathbf{B}^T \in \mathbb{R}^{p\times k}\,,\\ \nabla\varphi^*_{\mathbf{B}}(\mathbf{A},\mathbf{B}) &= \mathbf{A}^T\nabla\varphi(\mathbf{AB}) = \mathbf{A}^T\big(\mathbf{W}\odot(\mathbf{AB}-\mathbf{X})\big) \in \mathbb{R}^{k\times n}\,,\end{aligned} \tag{3.11}$$

and, finally, we obtain the gradient of $\varphi^*(.)$ at any pair $(\mathbf{A},\mathbf{B}) \in \mathbb{R}^{p\times k}\times\mathbb{R}^{k\times n}$ as

$$\begin{aligned}\nabla\varphi^*(\mathbf{A},\mathbf{B}) &= \big(\nabla\varphi^*_{\mathbf{A}}(\mathbf{A},\mathbf{B}), \nabla\varphi^*_{\mathbf{B}}(\mathbf{A},\mathbf{B})\big)\\ &= \Big(\big(\mathbf{W}\odot(\mathbf{AB}-\mathbf{X})\big)\mathbf{B}^T, \mathbf{A}^T\big(\mathbf{W}\odot(\mathbf{AB}-\mathbf{X})\big)\Big)\,.\end{aligned} \tag{3.12}$$

Consequently, the pair $(\mathbf{A},\mathbf{B})$ is a first-order stationary point of $\varphi^*(.)$ if

$$\big(\mathbf{W}\odot(\mathbf{AB}-\mathbf{X})\big)\mathbf{B}^T = \mathbf{0}^{p\times k} \text{ and } \mathbf{A}^T\big(\mathbf{W}\odot(\mathbf{AB}-\mathbf{X})\big) = \mathbf{0}^{k\times n}\,.$$

We now derive a convenient expression for the quadratic form $\big(\nabla^2\varphi^*(\mathbf{A},\mathbf{B})\big)$ in order to characterize the second-order stationary points of $\varphi^*(.)$, which are defined by the conditions

$$\nabla\varphi^*(\mathbf{A},\mathbf{B}) = (\mathbf{0}^{p\times k}, \mathbf{0}^{k\times n})$$

and

$$\big(\nabla^2\varphi^*(\mathbf{A},\mathbf{B})\big)\big((\mathbf{C},\mathbf{D}),(\mathbf{C},\mathbf{D})\big) \geq 0\,,\ \forall(\mathbf{C},\mathbf{D}) \in \mathbb{R}^{p\times k}\times\mathbb{R}^{k\times n}\,.$$

This will be useful to determine the relationships between the critical points of $\varphi(.)$ and $\varphi^*(.)$ in Theorem 3.7 below.

$\forall(\mathbf{C},\mathbf{D}) \in \mathbb{R}^{p\times k}\times\mathbb{R}^{k\times n}$, we have by the bilinearity of $\nabla^2\varphi^*(\mathbf{A},\mathbf{B})$

$$\begin{aligned}\big(\nabla^2\varphi^*(\mathbf{A},\mathbf{B})\big)\big((\mathbf{C},\mathbf{D}),(\mathbf{C},\mathbf{D})\big) &= \big(\nabla^2\varphi^*(\mathbf{A},\mathbf{B})\big)\big((\mathbf{C},\mathbf{0}^{k\times n}) + (\mathbf{0}^{p\times k},\mathbf{D}),\\ &\qquad (\mathbf{C},\mathbf{0}^{k\times n}) + (\mathbf{0}^{p\times k},\mathbf{D})\big)\\ &= \big(\nabla^2\varphi^*(\mathbf{A},\mathbf{B})\big)\big((\mathbf{C},\mathbf{0}^{k\times n}),(\mathbf{C},\mathbf{0}^{k\times n})\big)\\ &\quad + \big(\nabla^2\varphi^*(\mathbf{A},\mathbf{B})\big)\big((\mathbf{0}^{p\times k},\mathbf{D}),(\mathbf{0}^{p\times k},\mathbf{D})\big)\\ &\quad + 2.\big(\nabla^2\varphi^*(\mathbf{A},\mathbf{B})\big)\big((\mathbf{0}^{p\times k},\mathbf{D}),(\mathbf{C},\mathbf{0}^{k\times n})\big)\,.\end{aligned} \tag{3.13}$$

The last equality resulting from the fact that $\nabla^2\varphi^*(\mathbf{A},\mathbf{B})$ can also be considered as a self-adjoint (e.g., symmetric) mapping from $\mathbb{R}^{p\times k}\times\mathbb{R}^{k\times n}$ to $\mathbb{R}^{p\times k}\times\mathbb{R}^{k\times n}$ with respect to the inner product in $\mathbb{R}^{p\times k}\times\mathbb{R}^{k\times n}$ (see Subsection 2.4 for details).

Next, for the same reason, using the expression for $\nabla\varphi^*_{\mathbf{A}}(\mathbf{A},\mathbf{B})$ given in equation (3.11) and properties of the $\mathrm{Tr}(.)$ operator, notice that

$$\begin{aligned}\big(\nabla^2\varphi^*(\mathbf{A},\mathbf{B})\big)\big((\mathbf{C},\mathbf{0}^{k\times n}),(\mathbf{C},\mathbf{0}^{k\times n})\big) &= \big\langle[\nabla^2\varphi^*(\mathbf{A},\mathbf{B})]\big((\mathbf{C},\mathbf{0}^{k\times n})\big),(\mathbf{C},\mathbf{0}^{k\times n})\big\rangle_{\mathbb{R}^{p\times k}\times\mathbb{R}^{k\times n}}\\ &= \big(\nabla^2\varphi^*_{\mathbf{A}}(\mathbf{A},\mathbf{B})\big)(\mathbf{C},\mathbf{C})\\ &= \big\langle[\nabla^2\varphi^*_{\mathbf{A}}(\mathbf{A},\mathbf{B})](\mathbf{C}),\mathbf{C}\big\rangle_F\\ &= \big\langle\big(\mathbf{W}\odot\mathbf{CB}\big)\mathbf{B}^T,\mathbf{C}\big\rangle_F\\ &= \mathrm{Tr}\Big(\big((\mathbf{W}\odot\mathbf{CB})\mathbf{B}^T\big)^T\mathbf{C}\Big)\\ &= \mathrm{Tr}\Big(\mathbf{C}\big((\mathbf{W}\odot\mathbf{CB})\mathbf{B}^T\big)^T\Big)\\ &= \mathrm{Tr}\Big(\mathbf{CB}\big(\mathbf{W}\odot\mathbf{CB}\big)^T\Big)\\ &= \mathrm{Tr}\Big(\big(\mathbf{W}\odot\mathbf{CB}\big)^T\mathbf{CB}\Big)\\ &= \big\langle\mathbf{W}\odot\mathbf{CB},\mathbf{CB}\big\rangle_F\\ &= \big(\nabla^2\varphi(\mathbf{AB})\big)(\mathbf{CB},\mathbf{CB})\,,\end{aligned}$$

where the last equality results from equation (3.4). Similarly, we have

$$\begin{aligned}
\big(\nabla^2\varphi^*(\mathbf{A},\mathbf{B})\big)\big((\mathbf{0}^{p\times k},\mathbf{D}),(\mathbf{0}^{p\times k},\mathbf{D})\big) &= \big\langle[\nabla^2\varphi^*(\mathbf{A},\mathbf{B})]\big((\mathbf{0}^{p\times k},\mathbf{D})\big),(\mathbf{0}^{p\times k},\mathbf{D})\big\rangle_{\mathbb{R}^{p\times k}\times\mathbb{R}^{k\times n}} \\
&= \big(\nabla^2\varphi^*_{\mathbf{B}}(\mathbf{A},\mathbf{B})\big)(\mathbf{D},\mathbf{D}) \\
&= \big\langle[\nabla^2\varphi^*_{\mathbf{B}}(\mathbf{A},\mathbf{B})](\mathbf{D}),\mathbf{D}\big\rangle_F \\
&= \big\langle\mathbf{A}^T\big(\mathbf{W}\odot\mathbf{AD}\big),\mathbf{D}\big\rangle_F \\
&= \mathrm{Tr}\Big(\big(\mathbf{A}^T(\mathbf{W}\odot\mathbf{AD})\big)^T\mathbf{D}\Big) \\
&= \mathrm{Tr}\Big(\big(\mathbf{W}\odot\mathbf{AD}\big)^T\mathbf{AD}\Big) \\
&= \big\langle\mathbf{W}\odot\mathbf{AD},\mathbf{AD}\big\rangle_F \\
&= \big(\nabla^2\varphi(\mathbf{AB})\big)(\mathbf{AD},\mathbf{AD})\ .
\end{aligned}$$

We now reformulate similarly the last factor in the right-hand side of equation (3.13) in terms of $\nabla\varphi(\mathbf{AB})$ and $\big(\nabla^2\varphi(\mathbf{AB})\big)$:

$$\begin{aligned}
\big(\nabla^2\varphi^*(\mathbf{A},\mathbf{B})\big)\big((\mathbf{0}^{p\times k},\mathbf{D}),(\mathbf{C},\mathbf{0}^{k\times n})\big) &= \Big\langle D_{\mathbf{B}}\big(\nabla\varphi^*_{\mathbf{A}}(\mathbf{A},\mathbf{B})\big)(\mathbf{D}),\mathbf{C}\Big\rangle_F \\
&= \Big\langle D_{\mathbf{B}}\big((\mathbf{W}\odot(\mathbf{AB}-\mathbf{X}))\mathbf{B}^T\big)(\mathbf{D}),\mathbf{C}\Big\rangle_F \\
&= \Big\langle\big(\mathbf{W}\odot(\mathbf{AB}-\mathbf{X})\big)\mathbf{D}^T+(\mathbf{W}\odot\mathbf{AD})\mathbf{B}^T,\mathbf{C}\Big\rangle_F \\
&= \Big\langle\big(\mathbf{W}\odot(\mathbf{AB}-\mathbf{X})\big)\mathbf{D}^T,\mathbf{C}\Big\rangle_F \\
&\quad+\Big\langle(\mathbf{W}\odot\mathbf{AD})\mathbf{B}^T,\mathbf{C}\Big\rangle_F \\
&= \mathrm{Tr}\Big(\mathbf{D}\big(\mathbf{W}\odot(\mathbf{AB}-\mathbf{X})\big)^T\mathbf{C}\Big)+\mathrm{Tr}\Big(\mathbf{B}(\mathbf{W}\odot\mathbf{AD})^T\mathbf{C}\Big) \\
&= \mathrm{Tr}\Big(\mathbf{CD}\big(\mathbf{W}\odot(\mathbf{AB}-\mathbf{X})\big)^T\Big)+\mathrm{Tr}\Big(\mathbf{CB}(\mathbf{W}\odot\mathbf{AD})^T\Big) \\
&= \big\langle\mathbf{W}\odot(\mathbf{AB}-\mathbf{X}),\mathbf{CD}\big\rangle_F+\big\langle\mathbf{W}\odot\mathbf{AD},\mathbf{CB}\big\rangle_F \\
&= \big\langle\nabla\varphi(\mathbf{AB}),\mathbf{CD}\big\rangle_F+\big(\nabla^2\varphi(\mathbf{AB})\big)(\mathbf{AD},\mathbf{CB})\ .
\end{aligned}$$

Summarizing the preceding results, we have

$$\begin{aligned}
\big(\nabla^2\varphi^*(\mathbf{A},\mathbf{B})\big)\big((\mathbf{C},\mathbf{0}^{k\times n}),(\mathbf{C},\mathbf{0}^{k\times n})\big) &= \big(\nabla^2\varphi^*_{\mathbf{A}}(\mathbf{A},\mathbf{B})\big)(\mathbf{C},\mathbf{C}) = \big(\nabla^2\varphi(\mathbf{AB})\big)(\mathbf{CB},\mathbf{CB}), \\
\big(\nabla^2\varphi^*(\mathbf{A},\mathbf{B})\big)\big((\mathbf{0}^{p\times k},\mathbf{D}),(\mathbf{0}^{p\times k},\mathbf{D})\big) &= \big(\nabla^2\varphi^*_{\mathbf{B}}(\mathbf{A},\mathbf{B})\big)(\mathbf{D},\mathbf{D}) = \big(\nabla^2\varphi(\mathbf{AB})\big)(\mathbf{AD},\mathbf{AD}), \\
\big(\nabla^2\varphi^*(\mathbf{A},\mathbf{B})\big)\big((\mathbf{0}^{p\times k},\mathbf{D}),(\mathbf{C},\mathbf{0}^{k\times n})\big) &= \big\langle\nabla\varphi(\mathbf{AB}),\mathbf{CD}\big\rangle_F+\big(\nabla^2\varphi(\mathbf{AB})\big)(\mathbf{AD},\mathbf{CB})\ ,
\end{aligned} \tag{3.14}$$

and this implies, finally, using the symmetry and bilinearity of the bilinear form $\big(\nabla^2\varphi^*(\mathbf{A},\mathbf{B})\big)$ that

$$\begin{aligned}
\big(\nabla^2\varphi^*(\mathbf{A},\mathbf{B})\big)\big((\mathbf{C},\mathbf{D}),(\mathbf{C},\mathbf{D})\big) &= \big(\nabla^2\varphi(\mathbf{AB})\big)(\mathbf{CB},\mathbf{CB})+\big(\nabla^2\varphi(\mathbf{AB})\big)(\mathbf{AD},\mathbf{AD}) \\
&\quad+2.\big(\nabla^2\varphi(\mathbf{AB})\big)(\mathbf{AD},\mathbf{CB})+2.\big\langle\nabla\varphi(\mathbf{AB}),\mathbf{CD}\big\rangle_F \\
&= \big(\nabla^2\varphi(\mathbf{AB})\big)(\mathbf{CB},\mathbf{CB}+\mathbf{AD}) \\
&\quad+\big(\nabla^2\varphi(\mathbf{AB})\big)(\mathbf{AD},\mathbf{AD}+\mathbf{CB}) \\
&\quad+2.\big\langle\nabla\varphi(\mathbf{AB}),\mathbf{CD}\big\rangle_F \\
&= \big(\nabla^2\varphi(\mathbf{AB})\big)(\mathbf{CB}+\mathbf{AD},\mathbf{CB}+\mathbf{AD}) \\
&\quad+2.\big\langle\nabla\varphi(\mathbf{AB}),\mathbf{CD}\big\rangle_F\ .
\end{aligned} \tag{3.15}$$

Thus, the second-order stationary condition for $\varphi^*(.)$ at $(\mathbf{A},\mathbf{B})$, e.g., that the quadratic form $\big(\nabla^2\varphi^*(\mathbf{A},\mathbf{B})\big)$ is positive semi-definite, is equivalent to the inequality

$$\big(\nabla^2\varphi(\mathbf{AB})\big)(\mathbf{CB}+\mathbf{AD},\mathbf{CB}+\mathbf{AD})\geq-2.\big\langle\nabla\varphi(\mathbf{AB}),\mathbf{CD}\big\rangle_F\ ,$$

or, using equations (3.3) and (3.4), to the more convenient inequality

$$\|\sqrt{\mathbf{W}} \odot (\mathbf{CB} + \mathbf{AD})\|_F \geq -2.\big\langle \mathbf{W} \odot (\mathbf{AB} - \mathbf{X}), \mathbf{CD} \big\rangle_F \ , \ \forall (\mathbf{C}, \mathbf{D}) \in \mathbb{R}^{p \times k} \times \mathbb{R}^{k \times n} \ .$$

Note that, while these inequalities based on the quadratic expression of $\nabla^2 \varphi(\mathbf{AB})$ and the gradient $\nabla \varphi(\mathbf{AB})$ are sufficient for our purpose in this section, it is rather straightforward to obtain the general bilinear form of $\nabla^2 \varphi^*(\mathbf{A}, \mathbf{B})$ since

$$\begin{aligned} \big(\nabla^2 \varphi^*(\mathbf{A}, \mathbf{B})\big)\big((\mathbf{C}, \mathbf{D}), (\mathbf{E}, \mathbf{F})\big) &= \big(\nabla^2 \varphi^*(\mathbf{A}, \mathbf{B})\big)\big((\mathbf{C} + \mathbf{E}, \mathbf{D} + \mathbf{F}), (\mathbf{C} + \mathbf{E}, \mathbf{D} + \mathbf{F})\big) \\ &\quad - \frac{1}{4}\big(\nabla^2 \varphi^*(\mathbf{A}, \mathbf{B})\big)\big((\mathbf{C} - \mathbf{E}, \mathbf{D} - \mathbf{F}), (\mathbf{C} - \mathbf{E}, \mathbf{D} - \mathbf{F})\big) \ . \end{aligned}$$

Furthermore, a vectorized formulation of the symmetric bilinear mapping $\big(\nabla^2 \varphi^*(\mathbf{A}, \mathbf{B})\big)$ will be also derived later in Section 4.

We are now in the position to characterize more precisely the connections between the critical points of $\varphi(.)$ and $\varphi^*(.)$ in the following theorem, which is a reformulation and a slight extension in our WLRA context of results first given in Ha et al. [83] and later refined in Levin et al. [113] and Luo et al. [115].

Theorem 3.7. Let $(\mathbf{A}, \mathbf{B}) \in \mathbb{R}^{p \times k} \times \mathbb{R}^{k \times n}$. Then:

(1) If $\mathbf{AB} \in \mathbb{R}^{p \times n}_{\leq k}$ is a Frechet first-order stationary point of $\varphi(.)$ in the sense of Theorem 3.5 then $(\mathbf{A}, \mathbf{B})$ is a first-order stationary point of $\varphi^*(.)$.

(2) Reciprocally, if $(\mathbf{A}, \mathbf{B})$ is a first-order stationary point of $\varphi^*(.)$ such that $\mathbf{AB} \in \mathbb{R}^{p \times n}_{k}$ then $\mathbf{AB}$ is a Frechet first-order stationary point of $\varphi(.)$ in the sense of Theorem 3.5.

(3) Moreover, if $(\mathbf{A}, \mathbf{B})$ is a second-order stationary point of $\varphi^*(.)$ such that $\mathbf{AB} \in \mathbb{R}^{p \times n}_{<k}$, then $\mathbf{AB}$ is a Frechet first-order stationary point of $\varphi(.)$ in the sense of Theorem 3.5 and, thus, also a Frechet second-order stationary point of $\varphi(.)$ and even a solution of the WLRA problem in its formulation (P0).

(4) Reciprocally, if $\mathbf{AB} \in \mathbb{R}^{p \times n}_{<k}$ is a Frechet second-order stationary point of $\varphi(.)$ then $(\mathbf{A}, \mathbf{B})$ is a second-order stationary point of $\varphi^*(.)$ and also a solution of the WLRA problem in its formulation (P1).

(5) Finally, if $(\mathbf{A}, \mathbf{B})$ is a second-order stationary point of $\varphi^*(.)$ such that $\mathbf{AB} \in \mathbb{R}^{p \times n}_{k}$, then $\mathbf{AB}$ is a Frechet second-order stationary point of $\varphi(.)$ in the sense of Theorem 3.5.

(6) Reciprocally, if $\mathbf{AB} \in \mathbb{R}^{p \times n}_{k}$ is a Frechet second-order stationary point of $\varphi(.)$ then $(\mathbf{A}, \mathbf{B})$ is a second-order stationary point of $\varphi^*(.)$.

Proof. (1) : In order to prove the first assertion, we assume that $\mathbf{AB} \in \mathbb{R}^{p \times n}_{\leq k}$ is a Frechet first-order stationary point of $\varphi(.)$ and we consider separately the two cases *rank*$(\mathbf{AB}) < k$ and *rank*$(\mathbf{AB}) = k$.

If *rank*$(\mathbf{AB}) < k$, according to Theorem 3.5, we have $\nabla \varphi(\mathbf{AB}) = \mathbf{0}^{p \times n}$ and we deduce immediately that

$$\begin{aligned} \nabla \varphi^*_{\mathbf{A}}(\mathbf{A}, \mathbf{B}) &= \nabla \varphi(\mathbf{AB}) \mathbf{B}^T = \mathbf{0}^{p \times k} \ , \\ \nabla \varphi^*_{\mathbf{B}}(\mathbf{A}, \mathbf{B}) &= \mathbf{A}^T \nabla \varphi(\mathbf{AB}) = \mathbf{0}^{k \times n} \ . \end{aligned}$$

In other words, the pair $(\mathbf{A}, \mathbf{B})$ is a first-order critical point of $\varphi^*(.)$.

On the other hand, if *rank*$(\mathbf{AB}) = k$, again according to Theorem 3.5, we have

$$\nabla \varphi(\mathbf{AB})^T \mathbf{U}_{AB} = \mathbf{0}^{n \times k} \ \text{ and } \ \nabla \varphi(\mathbf{AB}) \mathbf{V}_{AB} = \mathbf{0}^{p \times k} \ ,$$

where the columns of $\mathbf{U}_{AB}$ and $\mathbf{V}_{AB}$ are, respectively, the first k left and right singular vectors of the matrix product $\mathbf{AB}$ in its thin SVD, e.g., $\mathbf{AB} = \mathbf{U}_{AB} \Sigma_{AB} \mathbf{V}_{AB}$.

As $\mathit{rank}(\mathbf{AB}) = \mathit{rank}(\mathbf{A}) = \mathit{rank}(\mathbf{B}) = k$, we have $\mathit{ran}(\mathbf{AB}) = \mathit{ran}(\mathbf{A})$ and $\mathit{ran}(\mathbf{B}^T\mathbf{A}^T) = \mathit{ran}(\mathbf{B}^T)$, and also

$$\begin{aligned}\mathit{ran}(\mathbf{U}_{AB}) &= \mathit{ran}(\mathbf{AB}) = \mathit{ran}(\mathbf{A}),\\ \mathit{ran}(\mathbf{V}_{AB}) &= \mathit{ran}(\mathbf{B}^T\mathbf{A}^T) = \mathit{ran}(\mathbf{B}^T)\ .\end{aligned}$$

This implies that it exists $\mathbf{C} \in \mathbb{R}^{k\times k}$ and $\mathbf{D} \in \mathbb{R}^{k\times k}$ such that

$$\mathbf{A} = \mathbf{U}_{AB}\mathbf{C} \ \text{ and } \ \mathbf{B}^T = \mathbf{V}_{AB}\mathbf{D}\ .$$

In these conditions, we have

$$\begin{aligned}\nabla\varphi^*_{\mathbf{A}}(\mathbf{A},\mathbf{B}) &= \nabla\varphi(\mathbf{AB})\mathbf{B}^T = \big(\nabla\varphi(\mathbf{AB})\mathbf{V}_{AB}\big)\mathbf{D} = \mathbf{0}^{p\times k}\ ,\\ \nabla\varphi^*_{\mathbf{B}}(\mathbf{A},\mathbf{B}) &= \mathbf{A}^T\nabla\varphi(\mathbf{AB}) = \mathbf{C}^T\big(\mathbf{U}_{AB}^T\nabla\varphi(\mathbf{AB})\big) = \mathbf{0}^{k\times n}\ ,\end{aligned}$$

as $\mathbf{AB}$ is is a first-order critical point of $\varphi(.)$. In other words, we have $\nabla\varphi^*(\mathbf{A},\mathbf{B}) = (\mathbf{0}^{p\times k}, \mathbf{0}^{k\times n})$ and the pair $(\mathbf{A},\mathbf{B})$ is a first-order stationary point of $\varphi^*(.)$.

(2) : Reciprocally, if the pair $(\mathbf{A},\mathbf{B})$ is a first-order stationary point of $\varphi^*(.)$ such that $\mathbf{AB} \in \mathbb{R}^{p\times n}_k$, we have also $\mathit{rank}(\mathbf{AB}) = \mathit{rank}(\mathbf{A}) = \mathit{rank}(\mathbf{B}^T) = k$, which implies again that $\mathbf{U}_{AB}$ and $\mathbf{A}$ span the same column space and that their columns form two bases of $\mathit{ran}(\mathbf{U}_{AB}) = \mathit{ran}(\mathbf{A})$. Similarly, $\mathbf{V}_{AB}$ and $\mathbf{B}^T$ span the same column space and their columns form two bases of $\mathit{ran}(\mathbf{V}_{AB}) = \mathit{ran}(\mathbf{B}^T)$. In these conditions, it exist $\mathbf{C} \in \mathbb{R}^{k\times k}$ and $\mathbf{D} \in \mathbb{R}^{k\times k}$ such that

$$\mathbf{U}_{AB} = \mathbf{AC} \ \text{ and } \ \mathbf{V}_{AB} = \mathbf{B}^T\mathbf{D}\ .$$

Using the first-order optimality conditions of $(\mathbf{A},\mathbf{B})$ for $\varphi^*(.)$, we have

$$\nabla\varphi(\mathbf{AB})\mathbf{B}^T = \mathbf{0}^{p\times k} \ \text{ and } \ \mathbf{A}^T\nabla\varphi(\mathbf{AB}) = \mathbf{0}^{k\times n}\ ,$$

which implies that

$$\begin{aligned}\nabla\varphi(\mathbf{AB})\mathbf{V}_{AB} &= \big(\nabla\varphi(\mathbf{AB})\mathbf{B}^T\big)\mathbf{D} = \mathbf{0}^{p\times k}\ ,\\ \nabla\varphi(\mathbf{AB})^T\mathbf{U}_{AB} &= \big(\nabla\varphi(\mathbf{AB})^T\mathbf{A}\big)\mathbf{C} = \big(\mathbf{A}^T\nabla\varphi(\mathbf{AB})\big)^T\mathbf{C} = \mathbf{0}^{k\times n}\ ,\end{aligned}$$

and the matrix product $\mathbf{AB}$ is a first-order critical point of $\varphi(.)$ in the sense of Theorem 3.5.

(3) : To demonstrate the next claim of the theorem, let $\mathbf{u}_1 \in \mathbb{R}^p$, $\mathbf{v}_1 \in \mathbb{R}^n$ and $\sigma_1 \in \mathbb{R}_+$ be, respectively, the first left and right singular vectors and the first singular value of $\nabla\varphi(\mathbf{AB}) \in \mathbb{R}^{p\times n}$. We first recall from equation (2.23) in Subsection 2.1 that the spectral norm of $\nabla\varphi(\mathbf{AB})$ is given by

$$\|\nabla\varphi(\mathbf{AB})\|_S = \sigma_1 = \mathbf{u}_1^T\nabla\varphi(\mathbf{AB})\mathbf{v}_1\ .$$

Moreover, as demonstrated just before Theorem 3.7, the hypothesis that the pair $(\mathbf{A},\mathbf{B})$ is a second-order stationary point of $\varphi^*(.)$ is equivalent to the inequality

$$\|\sqrt{\mathbf{W}} \odot (\mathbf{AD} + \mathbf{CB})\|_F \geq -2.\big\langle\nabla\varphi(\mathbf{AB}), \mathbf{CD}\big\rangle_F\ ,\ \forall(\mathbf{C},\mathbf{D}) \in \mathbb{R}^{p\times k}\times\mathbb{R}^{k\times n}\ .$$

Now, suppose that $\mathbf{AB} \in \mathbb{R}^{p\times n}_{<k}$ then $\mathbf{A}$ or $\mathbf{B}$ are not of full rank since $k \leq min(p,n)$. Without loss of generality suppose that $\mathit{rank}(\mathbf{A}) < k$. By the rank-nullity theorem (2.1), this implies that it exists a unit vector $\mathbf{w} \in \mathbb{R}^k$ such that $\mathbf{Aw} = \mathbf{0}^p$. Let

$$(\mathbf{C}_c, \mathbf{D}_c) = (-\mathbf{u}_1\mathbf{w}^T, c.\mathbf{w}\mathbf{v}_1^T) \in \mathbb{R}^{p\times k}\times\mathbb{R}^{k\times n}\ ,\ \forall c \in \mathbb{R}_{+*}\ .$$

By hypothesis, the pair $(\mathbf{A},\mathbf{B})$ is a second-order stationary point of $\varphi^*(.)$, which implies that

$$\|\sqrt{\mathbf{W}} \odot (\mathbf{AD}_c + \mathbf{C}_c\mathbf{B})\|_F \geq -2.\big\langle\nabla\varphi(\mathbf{AB}), \mathbf{C}_c, \mathbf{D}_c\big\rangle_F\ .$$

Now, we have

$$\mathbf{A}\mathbf{D}_c + \mathbf{C}_c\mathbf{B} = c.\mathbf{A}\mathbf{w}\mathbf{v}_1^T - \mathbf{u}_1\mathbf{w}^T\mathbf{B} = -\mathbf{u}_1\mathbf{w}^T\mathbf{B} \ ,$$

since $\mathbf{A}\mathbf{w} = \mathbf{0}^p$. Furthermore, as $\|\mathbf{w}\|_2^2 = \mathbf{w}^T\mathbf{w} = 1$, $\|\nabla\varphi(\mathbf{AB})\|_S = \mathbf{u}_1^T\nabla\varphi(\mathbf{AB})\mathbf{v}_1 = \sigma_1$ and

$$\mathrm{Tr}\left(\mathbf{EFG}\right) = \mathrm{Tr}\left(\mathbf{GEF}\right) \ , \ \forall \mathbf{E} \in \mathbb{R}^{p\times n}, \mathbf{F} \in \mathbb{R}^{n\times m}, \mathbf{G} \in \mathbb{R}^{m\times p} \ ,$$

we deduce that

$$\begin{aligned}
\left\langle \nabla\varphi(\mathbf{AB}), \mathbf{C}_c, \mathbf{D}_c \right\rangle_F &= \left\langle \nabla\varphi(\mathbf{AB}), -c.\mathbf{u}_1\mathbf{w}^T\mathbf{w}\mathbf{v}_1^T \right\rangle_F \\
&= -c.\left\langle \nabla\varphi(\mathbf{AB}), \mathbf{u}_1\mathbf{v}_1^T \right\rangle_F \\
&= -c.\ \mathrm{Tr}\left(\nabla\varphi(\mathbf{AB})^T\mathbf{u}_1\mathbf{v}_1^T\right) \\
&= -c.\ \mathrm{Tr}\left(\mathbf{v}_1^T\nabla\varphi(\mathbf{AB})^T\mathbf{u}_1\right) \\
&= -c.\mathbf{v}_1^T\nabla\varphi(\mathbf{AB})^T\mathbf{u}_1 \\
&= -c.\mathbf{u}_1^T\nabla\varphi(\mathbf{AB})\mathbf{v}_1^T \\
&= -c.\sigma_1 \ .
\end{aligned}$$

Using these different results, the preceding inequality simplifies to

$$\|\sqrt{\mathbf{W}} \odot (\mathbf{u}_1\mathbf{w}^T\mathbf{B})\|_F \geq 2.c.\sigma_1 = 2.c.\|\nabla\varphi(\mathbf{AB})\|_S \ ,$$

which holds for any $c > 0$. On the other hand, since the left-hand side of the last inequality is the Frobenius norm of a fixed element of $\mathbb{R}^{p\times n}$, which is not a function of c, it must be finite and this implies that $\|\nabla\varphi(\mathbf{AB})\|_S = \sigma_1 = 0$, i.e., $\nabla\varphi(\mathbf{AB}) = \mathbf{0}^{p\times n}$. Consequently, since $\mathit{rank}(\mathbf{AB}) < k$ by hypothesis, $\mathbf{AB}$ is a Frechet first-order stationary point of $\varphi(.)$ in the sense of Theorem 3.5 and even a solution of the WLRA problem in its formulation (P0).

(4) : if $\mathbf{AB} \in \mathbb{R}_{<k}^{p\times n}$ is a Frechet second-order stationary point of $\varphi(.)$ then this pair is a fortiori a Frechet first-order stationary point of $\varphi(.)$ and, according to Theorem 3.5, also a solution of the WLRA problem in its formulation (P1). By an application of Theorem 3.1, we deduce immediately that the pair $(\mathbf{A}, \mathbf{B})$ is a solution of the WLRA problem in its formulation (P1) and, thus, also a second-order stationary point of $\varphi^*(.)$.

(5) and (6) : the proofs of these two assertions can be found in Luo et al. [115], especially their Corollary 2, and we omit them here.

□

On the other hand, we highlight that, if the pair $(\mathbf{A}, \mathbf{B})$ is a first-order stationary point of $\varphi^*(.)$ such that $\mathbf{AB} \in \mathbb{R}_{<k}^{p\times n}$, then $\mathbf{AB}$ is not necessarily a Frechet first-order critical point of $\varphi(.)$, as noted by Ha et al. [83]. As an illustration, consider the pair $(\mathbf{0}^{p\times k}, \mathbf{0}^{k\times n})$. Obviously, this pair is a first-order critical point of $\varphi^*(.)$, but $\mathbf{0}^{p\times k}\mathbf{0}^{k\times n} = \mathbf{0}^{p\times n}$ is not a Frechet first-order critical point of $\varphi(.)$ in the sense of Theorem 3.5 as $\nabla\varphi(\mathbf{0}^{p\times n}) = -\mathbf{W} \odot \mathbf{X}$, which is not equal to $\mathbf{0}^{p\times n}$ as soon as we have for some pair of integers (i, j), $\mathbf{X}_{ij} \neq 0$ and $\mathbf{W}_{ij} > 0$. Thus, in general, $\mathbf{0}^{p\times n}$ is not a Frechet first-order stationary point of $\varphi(.)$ and is obviously not a solution of of the WLRA problem in its formulation (P0).

In addition, it is also possible to demonstrate that if the pair $(\mathbf{A}, \mathbf{B})$ is not a second-order stationary point of $\varphi^*(.)$, then $\mathbf{AB} \in \mathbb{R}_{\leq k}^{p\times n}$ is not a (local) minimizer of $\varphi(.)$ over $\mathbb{R}_{\leq k}^{p\times n}$, see Ha et al. [83] and Levin et al. [113] for details.

3.3 Approximate and regularized forms of the WLRA problem

In practice, instead of an exact solution of the WLRA problem, which can even not exist if missing values are present as noted above, one often seeks an approximation of $\mathbf{X}$ such that

$$\|\sqrt{\mathbf{W}} \odot (\mathbf{X} - \widehat{\mathbf{X}})\|_F^2 \leq (1 + \varepsilon)\bar{\mathbf{c}}_\varphi \ ,$$

where $\widehat{\mathbf{X}} \in \mathbb{R}^{p \times n}_{\leq k}$ denotes the approximation, $\bar{\mathbf{c}}_\varphi$ is the infimum of $\varphi(.)$ and $\varepsilon \in (0,1)$ is a tolerance parameter called the approximation error. In such framework, Razenshteyn et al. [167] recently show that in the case that $\mathbf{W}$ has at most r distinct rows and r distinct columns, there is an algorithm solving the above approximate version of the WLRA problem in $2^{O(k^2.r/\varepsilon)}\text{poly}(n)$ time with probability of success at least $9/10$. In the case that $\mathbf{W}$ has at most r distinct columns, but any number of distinct rows, there is also an algorithm solving the approximate version of the WLRA problem in $2^{O(k^2.r^2/\varepsilon)}\text{poly}(n)$ time with probability $9/10$. These bounds imply that for constant r and ε, even if r is as large as $\Theta\big(\log(n)\big)$ in the first case, and $\Theta\big(\sqrt{\log(n)}\big)$ in the second case, the corresponding algorithms are polynomial time. Razenshteyn et al. [167] also consider the case when the rank of the weight matrix $\mathbf{W}$ is at most r, which includes as special cases the two above cases, and devise an $n^{O(k^2.r/\varepsilon)}$ time algorithm for this more general case again with probability $9/10$. In other words, assuming that $\mathbf{W}$ has low rank, the algorithms of [167] achieve a $1+\varepsilon$ multiplicative approximation to the infimum of $\varphi(.)$.

Alternatively, some authors have recently developed simple and greedy algorithms with additive error bounds for the WLRA problem which do not require any structural assumption on $\mathbf{W}$, see Bhaskara et al. [25] for general weights and also Musco et al. [135] in the case of binary weights. In such approach, one seeks an approximation $\widehat{\mathbf{X}}$ of $\mathbf{X}$ such that

$$\|\sqrt{\mathbf{W}} \odot (\mathbf{X} - \widehat{\mathbf{X}})\|_F^2 \leq \bar{\mathbf{c}}_\varphi + \varepsilon \|\mathbf{X}\|_F^2 \text{ or } \|\sqrt{\mathbf{W}} \odot (\mathbf{X} - \widehat{\mathbf{X}})\|_F^2 \leq \gamma + \varepsilon \|\mathbf{X}\|_F^2 \,,$$

where γ is a (small) real constant of the order of $\bar{\mathbf{c}}_\varphi$ and the rank of $\widehat{\mathbf{X}}$ is of the order of k. Such methods with additive guarantees are interesting in applications (e.g., give sufficient matrix compression) when $\bar{\mathbf{c}}_\varphi$ is only a small fraction of the squared Frobenius norm of $\mathbf{X}$.

However, as these different algorithms with provable guarantees are inherently slow due the hardness of the WLRA problem and it is an open problem to determine when the WLRA problem has a closed form solution in general when some of the weights are zero, several authors have also proposed to minimize other related cost functions, which are convex, more smooth, and with a well-defined, nonempty and compact set of global minimizers, instead of problems (P0) or (P1) to address these issues [47][14][129][131][157][177][100][101][23].

As a first illustration, [129][131] have proposed the following convex relaxation to the rank constraint imposed in the formulation (P0):

$$\min_{\mathbf{Y} \in \mathbb{R}^{p \times n}} \quad \varphi_\lambda(\mathbf{Y}) = \frac{1}{2}\|\sqrt{\mathbf{W}} \odot (\mathbf{X} - \mathbf{Y})\|_F^2 + \lambda \|\mathbf{Y}\|_* \,.$$

Here $\|\mathbf{Y}\|_*$ is the nuclear norm (also called the trace norm), which is equal to the sum of the singular values of the $p \times n$ matrix $\mathbf{Y}$ and $\lambda \in \mathbb{R}_{+*}$ is a regularization parameter controlling the nuclear norm of the minimizer $\widehat{\mathbf{Y}}(\lambda)$ of this Lagrange form of (P0). $\varphi_\lambda(.)$ defines a convex function of its argument so that the above problem as an unique solution. Furthermore, it can be demonstrated that the rank of $\widehat{\mathbf{Y}}(\lambda)$ tends to zero when λ grows unbounded so that this proxy can provide suboptimal low-rank minimizers of problem (P0) when this Lagrange form of (P0) is solved for a range of values of λ [129][131]. Moreover, as the rank of $\widehat{\mathbf{Y}}(\lambda)$ increases when λ decreases, if this problem is solved for a range of decreasing values of λ, the iterative algorithm can use efficiently the solution for the previous value of λ as warm starts [129][86].

Another class of related methods are maximum margin matrix factorization (MMMF) methods [157] [177][23][117], which use a factorization model of the matrix $\mathbf{Y}$, as in the formulation (P1) of the WLRA problem, but are also equipped with a regularization term $\lambda \in \mathbb{R}_{+*}$ as in the above Lagrange form of problem (P0):

$$\min_{\mathbf{A} \in \mathbb{R}^{p \times k},\, \mathbf{B} \in \mathbb{R}^{k \times n}} \quad \varphi^*_\lambda(\mathbf{A}, \mathbf{B}) = \frac{1}{2}\|\sqrt{\mathbf{W}} \odot (\mathbf{X} - \mathbf{A}\mathbf{B})\|_F^2 + \frac{\lambda}{2}(\|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2) \,. \qquad \text{(MMMF)}$$

Not surprisingly (e.g., taking into account the equivalence between the original problems (P0) and (P1) stated in Theorem 3.1), there are closed relationships between the set of global minimizers of these Lagrange and regularized formulations of problems (P0) and (P1), see Theorem 3 and Lemma 6 in Mazumder et al. [129] and also Hastie et al. [86] for details. However, the above MMMF criterion is not convex in $(\mathbf{A}, \mathbf{B})$, but only bi-convex as for the original problem (P1), e.g., for a fixed $\mathbf{B}$ matrix, the modified function $\varphi^*_\lambda(.)$ is convex in $\mathbf{A}$, and for a fixed $\mathbf{A}$ matrix, the function $\varphi^*_\lambda(.)$ is convex in $\mathbf{B}$. As the MMMF criterion is not convex, it can have possibly several local minima as the original problem (P1) [72][171][157] and ALS algorithms (see Section 4), which are very often used to solve these MMMF and (P1) problems, get frequently stuck in sub-optimal local minima for a small value of k or a poorly chosen starting point, especially if some elements of the weight matrix $\mathbf{W}$ are equal to zero [72][171]. However, Ban et al. [23] have demonstrated, extending the results of Razenshteyn et al. [167], that it also exists polynomial time algorithms solving this weighted and regularized MMMF formulation of the WLRA problem, with provable guarantees, and also sharper time bounds than those proved in [167].

Some other recent works have proposed to add to $\varphi^*(\mathbf{A}, \mathbf{B})$, or similar regularized cost functions using the bilinear Burer-Monteiro approach, a balancing regularizer of the form

$$R(\mathbf{A}, \mathbf{B}) = \frac{\lambda}{4} \|\mathbf{A}^T\mathbf{A} - \mathbf{B}^T\mathbf{B}\|_F^2 \,,$$

where λ controls the weight for the regularizer as before [156][193][200][201]. $R(\mathbf{A}, \mathbf{B})$ implicitly forces the $\mathbf{A}$ and $\mathbf{B}$ matrices to have the same energy and, thus, helps to remove the scaling ambiguity which inherently affects the cost function $\varphi^*(.)$ and the minimization of $\varphi^*(\mathbf{A}, \mathbf{B})$ in the (P1) formulation of the WLRA problem as discussed in Remark 3.2 above. Moreover, for many cost functions which use the bilinear Burer-Monteiro approach, adding this balancing regularizer does not compromise the quality of the solutions [156][117][201][146].

Many of the proposed recent approaches also recast the WLRA problem as an optimization problem on the Grassmann manifold $\mathrm{Gr}(p, k)$ or on the two Grassmann manifolds $\mathrm{Gr}(p, k)$ and $\mathrm{Gr}(n, k)$ (where $\mathrm{Gr}(p, k)$ is the set of k-dimensional linear subspaces of $\mathbb{R}^p$) and introduce a regularization parameter $\lambda \in \mathbb{R}_{+*}$ as in the above Lagrange forms of problems (P0) and (P1) in order to ensure smoothness of the objective function and hence obtain good convergence at the expense of slight increase of the objective [100][101][47][131][14]. An interesting example in this class of methods, as it is closely related to the formulations (P0) or (P1) of the WLRA problem, is the unconstrained Riemannian optimization methods on a single Grassmann manifold $\mathrm{Gr}(p, k)$ described in Boumal and Absil [13][14] for solving the matrix completion problem, which we now discussed in some details.

To this end, for any weight matrix $\mathbf{W} \in \mathbb{R}_+^{p \times n}$, let us define the set $\bar{\Omega} \subset [p] \times [n]$, be the set of indices of the elements of $\mathbf{W}$ with $\mathbf{W}_{ij} = 0$ (e.g., $\bar{\Omega}$ is the complement of Ω in $[p] \times [n]$) and the seminorms

$$\|\mathbf{Y}\|_\Omega^2 = \sum_{(i,j)\in\Omega} \mathbf{Y}_{ij}^2 \text{ and } \|\mathbf{Y}\|_{\bar{\Omega}}^2 = \sum_{(i,j)\in\bar{\Omega}} \mathbf{Y}_{ij}^2 \,.$$

With these definitions and in our notations, Boumal and Absil [13][14] proposed to solve the following optimization problem

$$\min_{\mathbf{Y}\in\mathbb{R}^{p\times n}_{\leq k}} \quad g(\mathbf{Y}) = \frac{1}{2}\|\sqrt{\mathbf{W}} \odot (\mathbf{X} - \mathbf{Y})\|_\Omega^2 + \frac{\lambda}{2}\|\mathbf{Y}\|_{\bar{\Omega}}^2 \,,$$

where, as before, $\lambda \in \mathbb{R}_{+*}$ is a regularization parameter, which ensures that the solution to this problem exists and the cost function g is smooth. They give the following interpretation for the minimization of the cost function g, which makes sense for the matrix completion problem: "we

are looking for an optimal matrix $\widehat{\mathbf{X}}$ of rank at most k and we have confidence $\sqrt{\mathbf{W}}_{ij}$ that $\widehat{\mathbf{X}}_{ij}$ should equal $\mathbf{X}_{ij}$ for $(i,j) \in \Omega$ and smaller confidence λ that $\widehat{\mathbf{X}}_{ij}$ should equal zero for $(i,j) \in \bar{\Omega}$". They have also illustrated that the solutions of this problem are largely insensitive to the value of λ provided it is much smaller than the strictly positive values $\mathbf{W}_{ij}$. As an illustration, for matrix completion problems in their experiments, they used $\lambda = 10^{-6}$ and $\mathbf{W}_{ij} = 1$ if $(i,j) \in \Omega$. Finally, they describe and apply second-order Riemannian trust-region methods (RTRMC2) and Riemannian conjugate gradient methods (RCGMC) [11] to solve this problem efficiently and accurately, which are still state-of-the-art algorithms on a wide range of problem instances.

Interestingly, we now show that the minimization of the cost function $g(.)$ proposed by Boumal and Absil [13][14] is in fact a simple instance of formulation (P0) of the WLRA problem so that the variable projection framework can also be used to solve this problem as we will illustrate in the following sections. More precisely, if, for any $p \times n$ weight matrix $\mathbf{W}$ with some zero elements and any $\lambda \in \mathbb{R}_{+*}$ (e.g., $\lambda > 0$), we define as above an $p \times n$ weight matrix $\mathbf{W}_\lambda \in \mathbb{R}_{+*}^{p \times n}$ as

$$\big[\mathbf{W}_\lambda\big]_{ij} = \begin{cases} \mathbf{W}_{ij} & \text{if } (i,j) \in \Omega \\ \lambda & \text{if } (i,j) \notin \Omega \end{cases} , \tag{3.16}$$

and we introduce the projection operator associated with an $p \times n$ weight matrix $\mathbf{W}$ by $P_\Omega : \mathbb{R}^{p \times n} \longrightarrow \mathbb{R}^{p \times n}$ with $P_\Omega(\mathbf{X}) = \mathbf{X}_\Omega$ where

$$\big[\mathbf{X}_\Omega\big]_{ij} = \begin{cases} \mathbf{X}_{ij} & \text{if } (i,j) \in \Omega \\ 0 & \text{if } (i,j) \notin \Omega \end{cases} , \tag{3.17}$$

we can rearrange the cost function $g(.)$ introduced by Boumal and Absil [13][14] as

$$g_\lambda(\mathbf{Y}) = \frac{1}{2} \| \sqrt{\mathbf{W}_\lambda} \odot (\mathbf{X}_\Omega - \mathbf{Y}) \|_F^2 , \tag{3.18}$$

and it is readily observed that the minimization of this cost function $g_\lambda(.)$ w.r.t. $\mathbf{Y} \in \mathbb{R}_{\leq k}^{p \times n}$ is equivalent to the form (P0)

$$\min_{\mathbf{Y} \in \mathbb{R}_{\leq k}^{p \times n}} \quad g_\lambda(\mathbf{Y}) = \frac{1}{2} \| \sqrt{\mathbf{W}_\lambda} \odot (\mathbf{X}_\Omega - \mathbf{Y}) \|_F^2 = \frac{1}{2} \| \mathbf{X}_\Omega - \mathbf{Y} \|_{\mathbf{W}_\lambda}^2$$

of a standard WLRA problem in which we use the matrices $\mathbf{X}_\Omega$ and $\mathbf{W}_\lambda$ in place of $\mathbf{X}$ and $\mathbf{W}$, respectively. Furthermore, as all the elements of the weight matrix $\mathbf{W}_\lambda$ are greater than zero for any $\lambda \in \mathbb{R}_{+*}$, $\|\|\|_{\mathbf{W}_\lambda}$ defines a norm on $\mathbb{R}^{p \times k}$ and Theorem 3.3 shows that the set of global minimizers of $g_\lambda(.)$ is nonempty and compact, so that the minimization of this cost function is a well-posed problem. In other words, for any $\lambda \in \mathbb{R}_{+*}$ there exists $\widehat{\mathbf{X}}_\lambda \in \mathbb{R}_{\leq k}^{p \times n}$ such that

$$\widehat{\mathbf{X}}_\lambda = \text{Arg} \min_{\mathbf{Y} \in \mathbb{R}_{\leq k}^{p \times n}} g_\lambda(\mathbf{Y}) .$$

In addition, if we take a regularization parameter λ (also called the Tikhonov parameter, see [70]) sufficiently small, the following theorem shows that the minimization of $g_\lambda(.)$ with a Tikhonov parameter tending to zero is an interesting alternative to the formulations (P0) and (P1) of the WLRA problem, which are not well-posed when some elements of the weight matrix $\mathbf{W}$ are equal to zero as discussed above.

Theorem 3.8. Let $\mathbf{X} \in \mathbb{R}^{p \times n}$, $\mathbf{W} \in \mathbb{R}_+^{p \times n}$ (i.e., $\mathbf{W}_{ij} \geq 0$), $k \in \mathbb{N}_*$ with $k \leq \textit{rank}(\mathbf{X}) \leq \min(p,n)$ and $\lambda \in \mathbb{R}_{+*}$ (i.e., $\lambda > 0$). Furthermore, using definition (3.18) of the cost function, $g_\lambda(.)$, let

$$\widehat{\mathbf{X}}_\lambda = \text{Arg} \min_{\mathbf{Y} \in \mathbb{R}_{\leq k}^{p \times n}} g_\lambda(\mathbf{Y}) \text{ and } f(\lambda) = g_\lambda(\widehat{\mathbf{X}}_\lambda) = \frac{1}{2} \| \mathbf{X}_\Omega - \widehat{\mathbf{X}}_\lambda \|_{\mathbf{W}_\lambda}^2 \text{ for } \lambda \in \mathbb{R}_{+*}$$

then

$$\lim_{\lambda \to 0} f(\lambda) = \bar{\mathbf{c}}_{\varphi} \ ,$$

where $\bar{\mathbf{c}}_{\varphi}$ is the infimum of the cost function $\varphi(.)$ used in the formulation (P0) of the WLRA problem and $\mathbf{X}_{\Omega} = P_{\Omega}(\mathbf{X})$ where P_{Ω} is the projection operator associated with the $p \times n$ weight matrix $\mathbf{W}$.

Proof. We first show that $f(.)$ has a well defined limit, $\bar{\mathbf{c}}_f$, when $f(.)$ tends to zero. To demonstrate this result, we first note that $f(.)$ is an increasing function. For $\alpha \in \mathbb{R}_{+*}$ and $\lambda \in \mathbb{R}_{+*}$ with $\alpha \geq \lambda$, let

$$\widehat{\mathbf{X}}_{\alpha} = \text{Arg} \min_{\mathbf{Y} \in \mathbb{R}^{p \times n}_{\leq k}} g_{\alpha}(\mathbf{Y}) \text{ and } \widehat{\mathbf{X}}_{\lambda} = \text{Arg} \min_{\mathbf{Y} \in \mathbb{R}^{p \times n}_{\leq k}} g_{\lambda}(\mathbf{Y})$$

then we have

$$\|\mathbf{X}_{\Omega} - \widehat{\mathbf{X}}_{\alpha}\|^2_{\mathbf{W}_{\alpha}} \geq \|\mathbf{X}_{\Omega} - \widehat{\mathbf{X}}_{\alpha}\|^2_{\mathbf{W}_{\lambda}} \geq \|\mathbf{X}_{\Omega} - \widehat{\mathbf{X}}_{\lambda}\|^2_{\mathbf{W}_{\lambda}} \ ,$$

which implies that $f(\alpha) \geq f(\lambda)$. Furthermore, for all $\lambda \in \mathbb{R}_{+*}$, we have

$$f(\lambda) = \frac{1}{2}\|\mathbf{X}_{\Omega} - \widehat{\mathbf{X}}_{\lambda}\|^2_{\mathbf{W}_{\lambda}} \geq \frac{1}{2}\|\mathbf{X}_{\Omega} - \widehat{\mathbf{X}}_{\lambda}\|^2_{\mathbf{W}} = \varphi(\widehat{\mathbf{X}}_{\lambda}) \geq \bar{\mathbf{c}}_{\varphi} \ ,$$

which shows that $\lim_{\lambda \to 0} f(\lambda)$ exists and that $\lim_{\lambda \to 0} f(\lambda) = \bar{\mathbf{c}}_f \geq \bar{\mathbf{c}}_{\varphi}$.

It remains to show that $\bar{\mathbf{c}}_{\varphi} \geq \bar{\mathbf{c}}_f$. To this end, suppose that $\bar{\mathbf{c}}_{\varphi} < \bar{\mathbf{c}}_f$, then it exists $\mathbf{Y} \in \mathbb{R}^{p \times n}_{\leq k}$ such that $\bar{\mathbf{c}}_{\varphi} \leq \varphi(\mathbf{Y}) < \bar{\mathbf{c}}_f$, otherwise $\bar{\mathbf{c}}_{\varphi}$ is not the infimum of $\varphi(.)$. As

$$\lim_{\lambda \to 0} \|\mathbf{X}_{\Omega} - \mathbf{Y}\|^2_{\mathbf{W}_{\lambda}} = \|\mathbf{X} - \mathbf{Y}\|^2_{\mathbf{W}} \text{ and } \|\mathbf{X}_{\Omega} - \mathbf{Y}\|^2_{\mathbf{W}_{\lambda}} \geq \|\mathbf{X} - \mathbf{Y}\|^2_{\mathbf{W}} \text{ for all } \lambda \in \mathbb{R}_{+*} \ ,$$

it also exists $\alpha \in \mathbb{R}_{+*}$ such that

$$\varphi(\mathbf{Y}) = \frac{1}{2}\|\mathbf{X} - \mathbf{Y}\|^2_{\mathbf{W}} \leq \frac{1}{2}\|\mathbf{X}_{\Omega} - \mathbf{Y}\|^2_{\mathbf{W}_{\alpha}} < \bar{\mathbf{c}}_f \ .$$

However, we also have

$$\bar{\mathbf{c}}_f \leq f(\alpha) = \frac{1}{2}\|\mathbf{X}_{\Omega} - \widehat{\mathbf{X}}_{\alpha}\|^2_{\mathbf{W}_{\alpha}} \leq \frac{1}{2}\|\mathbf{X}_{\Omega} - \mathbf{Y}\|^2_{\mathbf{W}_{\alpha}} < \bar{\mathbf{c}}_f$$

and we obtain a contradiction.

□

Thus, one way of getting an useful approximate solution to the WLRA problem when missing values are present is to use a continuation Tikhonov method that approximately solves a sequence of regularized WLRA problems for a sequence of decreasing Tikhonov parameter λ. The approximate solution of one regularized WLRA problem with Tikhonov parameter λ_t (e.g., the minimization of $g_{\lambda_t}(.)$) is taken as the starting point for the next regularized WLRA problem with Tikhonov parameter $\lambda_{t+1} < \lambda_t$. This kind of Tikhonov methods has already been proposed in the context of ill-conditioned and uniformly rank-deficient NLLS problems [52][53][54], see Section 6 where such methods are further discussed.

3.4 Variable projection formulation of the WLRA problem

We are now ready to show that the alternative formulation (P1) or its variants (see Remark 3.1) of the WLRA problem

$$\min_{\mathbf{A} \in \mathbb{R}^{p \times k}, \mathbf{B} \in \mathbb{R}^{k \times n}} \varphi^*(\mathbf{A}, \mathbf{B}) = \frac{1}{2}\|\sqrt{\mathbf{W}} \odot (\mathbf{X} - \mathbf{A}\mathbf{B})\|^2_F$$

is a separable NLLS problem as stated in the Definition 2.10 of Subsection 2.4 [63][166]. This means that the minimization of $\varphi^*(\mathbf{A}, \mathbf{B})$ is a mixed linear-nonlinear least-squares problem where the associated residual function $e(\mathbf{A}, \mathbf{B})$ is linear in some variables and nonlinear in others.

In order to demonstrate this result, we first write $\varphi^*(\mathbf{A},\mathbf{B})$ as

$$\varphi^*(\mathbf{A},\mathbf{B}) = \frac{1}{2}\|e(\mathbf{A},\mathbf{B})\|_2^2 = \frac{1}{2}e(\mathbf{A},\mathbf{B})^T e(\mathbf{A},\mathbf{B}) \ ,$$

where the residual vector function $e(\mathbf{A},\mathbf{B}) \in \mathbb{R}^{p.n}$ is defined by

$$e(\mathbf{A},\mathbf{B}) = vec\big(\sqrt{\mathbf{W}} \odot (\mathbf{X} - \mathbf{AB})\big) \ . \tag{3.19}$$

Using equations (2.27) and (2.33), the residual function $e(\mathbf{A},\mathbf{B})$ can be further transformed as

$$\begin{aligned} e(\mathbf{A},\mathbf{B}) &= diag\big(vec(\sqrt{\mathbf{W}})\big)vec(\mathbf{X} - \mathbf{AB}) \\ &= vec(\sqrt{\mathbf{W}} \odot \mathbf{X}) - diag\big(vec(\sqrt{\mathbf{W}})\big)vec(\mathbf{AB}) \\ &= vec(\sqrt{\mathbf{W}} \odot \mathbf{X}) - diag\big(vec(\sqrt{\mathbf{W}})\big)\,(\mathbf{I}_n \otimes \mathbf{A})\,vec(\mathbf{B}) \ , \end{aligned}$$

and $e(\mathbf{A},\mathbf{B})$ is finally equal in explicit matrix form to

$$\begin{bmatrix} \sqrt{\mathbf{W}}_{.1}\mathbf{X}_{.1} \\ \vdots \\ \sqrt{\mathbf{W}}_{.j}\mathbf{X}_{.j} \\ \vdots \\ \sqrt{\mathbf{W}}_{.n}\mathbf{X}_{.n} \end{bmatrix} - \begin{bmatrix} diag(\sqrt{\mathbf{W}}_{.1})\mathbf{A} & 0 & \dots & 0 & 0 \\ 0 & \ddots & 0 & \dots & 0 \\ \vdots & 0 & diag(\sqrt{\mathbf{W}}_{.j})\mathbf{A} & 0 & \vdots \\ 0 & \dots & 0 & \ddots & 0 \\ 0 & 0 & \dots & 0 & diag(\sqrt{\mathbf{W}}_{.n})\mathbf{A} \end{bmatrix} \begin{bmatrix} \mathbf{B}_{.1} \\ \vdots \\ \mathbf{B}_{.j} \\ \vdots \\ \mathbf{B}_{.n} \end{bmatrix} .$$

In this residual function, we first note that all the lines corresponding to a zero weight (e.g., $\mathbf{W}_{ij} = 0$) can be eliminated when evaluating this function in real computations. The same is true for all the equations of the following sections and in a practical computer implementation of the algorithms used to minimize $\varphi^*(.)$. However, for notational simplicity and because we want to consider at the same time both the cases $\mathbf{W} \in \mathbb{R}_+^{p\times n}$ and $\mathbf{W} \in \mathbb{R}_{+*}^{p\times n}$, we do not introduce an incidence matrix in our equations to indicate which rows or columns must be eliminated as was done for example in [147][28][37][66][14]. Then, we may write

$$\varphi^*(\mathbf{A},\mathbf{B}) = \frac{1}{2}\|\mathbf{x} - \mathbf{F}(\mathbf{a})\mathbf{b})\|_2^2 \ ,$$

where $\mathbf{x} = vec(\sqrt{\mathbf{W}}\odot\mathbf{X})$, $\mathbf{a} = vec(\mathbf{A}^T)$, $\mathbf{b} = vec(\mathbf{B})$ and $\mathbf{F}(\mathbf{a})$ is the block diagonal matrix

$$\mathbf{F}(\mathbf{a}) = \bigoplus_{j=1}^{n} \mathbf{F}_j(\mathbf{a}) = \begin{bmatrix} \mathbf{F}_1(\mathbf{a}) & 0 & \dots & 0 & 0 \\ 0 & \ddots & 0 & \dots & 0 \\ \vdots & 0 & \mathbf{F}_j(\mathbf{a}) & 0 & \vdots \\ 0 & \dots & 0 & \ddots & 0 \\ 0 & 0 & \dots & 0 & \mathbf{F}_n(\mathbf{a}) \end{bmatrix} = diag\big(vec(\sqrt{\mathbf{W}})\big)\big(\mathbf{I}_n \otimes \mathbf{A}\big) \ , \tag{3.20}$$

where

$$\mathbf{F}_j(\mathbf{a}) = diag(\sqrt{\mathbf{W}}_{.j})\mathbf{A} = diag(\sqrt{\mathbf{W}}_{.j})\big(mat_{k\times p}(\mathbf{a})\big)^T \ .$$

The reason and interest of defining the vectorized form of $\mathbf{A}$ as

$$\mathbf{a} = vec(\mathbf{A}^T) \ , \tag{3.21}$$

instead of simply $vec(\mathbf{A})$ as usually done, will become clear in the next sections. From this formulation, it is clear that minimizing $\varphi^*(.)$ is a separable NLLS problem, since for a fixed matrix $\mathbf{A}$, we have a linear least-squares problem to determine the optimal vector $\widehat{\mathbf{b}} = vec(\widehat{\mathbf{B}})$, i.e.,

$$\widehat{\mathbf{b}} = \text{Arg}\min_{\mathbf{b}\in\mathbb{R}^{n.k}} \varphi^*(\mathbf{A},\mathbf{B}) = \frac{1}{2}\|\mathbf{x} - \mathbf{F}(\mathbf{a})\mathbf{b})\|_2^2 \ .$$

Moreover, we observe that the residual function $e(\mathbf{A},\mathbf{B})$ is linear in both $\mathbf{A}$ and $\mathbf{B}$, since

$$\begin{aligned}
\mathbf{F}(\mathbf{a})\mathbf{b} &= \Big(\bigoplus_{j=1}^{n} \mathit{diag}(\sqrt{\mathbf{W}}_{.j})\mathbf{A}\Big)\mathbf{b} \\
&= \mathit{diag}\big(\mathit{vec}(\sqrt{\mathbf{W}})\big)\big(\mathbf{I}_n \otimes \mathbf{A}\big)\mathit{vec}(\mathbf{B}) \\
&= \mathit{diag}\big(\mathit{vec}(\sqrt{\mathbf{W}})\big)\mathit{vec}(\mathbf{AB}) \\
&= \mathit{diag}\big(\mathit{vec}(\sqrt{\mathbf{W}})\big)\big(\mathbf{B}^T \otimes \mathbf{I}_p\big)\mathit{vec}(\mathbf{A}) \\
&= \mathit{diag}\big(\mathit{vec}(\sqrt{\mathbf{W}})\big)\big(\mathbf{B}^T \otimes \mathbf{I}_p\big)\mathbf{K}_{(k,p)}\mathbf{K}_{(p,k)}\mathit{vec}(\mathbf{A}) \\
&= \mathit{diag}\big(\mathit{vec}(\sqrt{\mathbf{W}})\big)\big(\mathbf{B}^T \otimes \mathbf{I}_p\big)\mathbf{K}_{(k,p)}\mathit{vec}(\mathbf{A}^T) \\
&= \mathit{diag}\big(\mathit{vec}(\sqrt{\mathbf{W}})\big)\mathbf{K}_{(n,p)}\big(\mathbf{I}_p \otimes \mathbf{B}^T\big)\mathbf{a} \\
&= \mathbf{K}_{(n,p)}\mathit{diag}\big(\mathit{vec}(\sqrt{\mathbf{W}}^T)\big)\big(\mathbf{I}_p \otimes \mathbf{B}^T\big)\mathbf{a} \\
&= \mathbf{K}_{(n,p)}\Big(\bigoplus_{i=1}^{p} \mathit{diag}(\sqrt{\mathbf{W}}_{i.})\mathbf{B}^T\Big)\mathbf{a}\,.
\end{aligned}$$

Defining now

$$\mathbf{G}(\mathbf{b}) = \bigoplus_{i=1}^{p} \mathbf{G}_i(\mathbf{b}) = \begin{bmatrix} \mathbf{G}_1(\mathbf{b}) & 0 & \dots & 0 & 0 \\ 0 & \ddots & 0 & \dots & 0 \\ \vdots & 0 & \mathbf{G}_i(\mathbf{b}) & 0 & \vdots \\ 0 & \dots & 0 & \ddots & 0 \\ 0 & 0 & \dots & 0 & \mathbf{G}_p(\mathbf{b}) \end{bmatrix}, \tag{3.22}$$

where

$$\mathbf{G}_i(\mathbf{b}) = \mathit{diag}(\sqrt{\mathbf{W}}_{i.})\mathbf{B}^T = \mathit{diag}(\sqrt{\mathbf{W}}_{i.})\big(\mathit{mat}_{k\times n}(\mathbf{b})\big)^T\,,$$

we note that the residual function $e(\mathbf{A},\mathbf{B})$ may then be written in the following alternative matrix form

$$\begin{aligned}
e(\mathbf{A},\mathbf{B}) &= \mathbf{x} - \mathbf{F}(\mathbf{a})\mathbf{b} \\
&= \mathbf{x} - \mathbf{K}_{(n,p)}\Big(\bigoplus_{i=1}^{p} \mathit{diag}(\sqrt{\mathbf{W}}_{i.})\mathbf{B}^T\Big)\mathbf{a} \\
&= \mathbf{x} - \mathbf{K}_{(n,p)}\mathbf{G}(\mathbf{b})\mathbf{a} \\
&= \mathbf{K}_{(n,p)}\mathbf{K}_{(p,n)}\mathbf{x} - \mathbf{K}_{(n,p)}\mathbf{G}(\mathbf{b})\mathbf{a} \\
&= \mathbf{K}_{(n,p)}\mathbf{K}_{(p,n)}\mathit{vec}(\sqrt{\mathbf{W}} \odot \mathbf{X}) - \mathbf{K}_{(n,p)}\mathbf{G}(\mathbf{b})\mathbf{a} \\
&= \mathbf{K}_{(n,p)}\mathit{vec}\big((\sqrt{\mathbf{W}} \odot \mathbf{X})^T\big) - \mathbf{K}_{(n,p)}\mathbf{G}(\mathbf{b})\mathbf{a} \\
&= \mathbf{K}_{(n,p)}\big(\mathbf{z} - \mathbf{G}(\mathbf{b})\mathbf{a}\big)\,,
\end{aligned}$$

where $\mathbf{z} = \mathit{vec}\big((\sqrt{\mathbf{W}} \odot \mathbf{X})^T\big)$. This implies that $\varphi^*(\mathbf{A},\mathbf{B})$ may also be expressed as

$$\begin{aligned}
\varphi^*(\mathbf{A},\mathbf{B}) &= \frac{1}{2}\left(\mathbf{z} - \mathbf{G}(\mathbf{b})\mathbf{a}\right)^T \mathbf{K}_{(p,n)}\mathbf{K}_{(n,p)}\left(\mathbf{z} - \mathbf{G}(\mathbf{b})\mathbf{a}\right) \\
&= \frac{1}{2}\left(\mathbf{z} - \mathbf{G}(\mathbf{b})\mathbf{a}\right)^T\left(\mathbf{z} - \mathbf{G}(\mathbf{b})\mathbf{a}\right) \\
&= \frac{1}{2}\|\mathbf{z} - \mathbf{G}(\mathbf{b})\mathbf{a}\|_2^2\,,
\end{aligned}$$

which shows that the roles of $\mathbf{A}$ and $\mathbf{B}$ are interchangeable in $\varphi^*(\mathbf{A},\mathbf{B})$ as already noted in the case of binary weights for example in [147]. As for the choice between the formulations (P1) and (P2) of the WLRA problem (see Remark 3.2), the choice between the formulations

$$\varphi^*(\mathbf{A},\mathbf{B}) = \frac{1}{2}\|\mathbf{x} - \mathbf{F}(\mathbf{a})\mathbf{b})\|_2^2 \quad \text{or} \quad \varphi^*(\mathbf{A},\mathbf{B}) = \frac{1}{2}\|\mathbf{z} - \mathbf{G}(\mathbf{b})\mathbf{a}\|_2^2$$

depends on the values of p and n, and the first one should be preferred if $p < n$ as the number of parameters to estimate (e.g., $\mathbf{A}$) will be smaller once the other matrix variable (e.g., $\mathbf{B}$) has been eliminated as we will show below, and vice-versa if $p > n$. Furthermore, in what follows, we note that the matrices $\mathbf{A}$ and $\mathbf{B}$ can be used in an interchangeable manner with their vectorized forms $\mathbf{a}$ and $\mathbf{b}$, respectively, as the mapping $vec(.)$ is a bijective homeomorphism (see Subsection 2.2 for details).

Thus, the problem of minimizing $\varphi^*(.)$ is separable and this property can be exploited in a least-squares estimation, and a number of special purpose algorithms have been proposed in this context [95][96][63][166][10][149][17]. Moreover, it has been demonstrated that these special algorithms provide greater stability than standard NLLS methods, besides reducing both the dimensionality of the optimization problem and the necessary number of iteration steps [136][65][37][17]. In most cases, the total computational work decreases with separable methods even though the code describing the separable problem is slightly more complicated than in standard NLLS algorithms. We now discuss how to reformulate the problems (P0) and (P1) so that we can exploit the separation property by eliminating one of the matrix variables (e.g., $\mathbf{B}$ if $p < n$) and devise more efficient algorithms to solve these problems.

Assuming that $p < n$, for a fixed $\mathbf{A}$ matrix, we have a linear least-squares problem to determine the optimal vector $\mathbf{b} = vec(\mathbf{B})$ which will minimize the cost function $\varphi^*(.)$ and the solution of this linear least-squares problem is $\widehat{\mathbf{b}} = \mathbf{F}(\mathbf{a})^+\mathbf{x}$ where $\mathbf{x} = vec(\sqrt{\mathbf{W}} \odot \mathbf{X})$ and $\mathbf{F}(\mathbf{a})^+$ is the pseudo-inverse of the $p.n \times k.n$ matrix $\mathbf{F}(\mathbf{a})$, see Subsection 2.1. Inserting now $\widehat{\mathbf{b}}$ in $\varphi^*(.)$, we obtain a new nonlinear functional $\psi : \mathbb{R}^{p.k} \longrightarrow \mathbb{R}$ involving only the vectorized form of the $\mathbf{A}$ matrix

$$\psi(\mathbf{a}) = \frac{1}{2}\| \left(\mathbf{I}_{p.n} - \mathbf{F}(\mathbf{a})\mathbf{F}(\mathbf{a})^+\right)\mathbf{x}\|_2^2 = \frac{1}{2}\|\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\mathbf{x}\|_2^2 = \frac{1}{2}\|\mathbf{r}(\mathbf{a})\|_2^2 \,, \tag{3.23}$$

where $\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}$ is the orthogonal projector onto the orthogonal complement of $ran\big(\mathbf{F}(\mathbf{a})\big)$ and $\mathbf{r}(.)$ is a nonlinear residual function of $\mathbf{a} = vec(\mathbf{A}^T)$ defined by

$$\mathbf{r} : \mathbb{R}^{p.k} \longrightarrow \mathbb{R}^{p.n} : \mathbf{a} \mapsto \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\mathbf{x} = \mathbf{x} - \mathbf{F}(\mathbf{a})\widehat{\mathbf{b}} = \mathbf{K}_{(n,p)}\big(\mathbf{z} - \mathbf{G}(\widehat{\mathbf{b}})\mathbf{a}\big) \,. \tag{3.24}$$

$\mathbf{r}(\mathbf{a})$ is called the variable projection residual of $\mathbf{X}$ at $\mathbf{A}$ (or equivalently of $\mathbf{x}$ at $\mathbf{a}$) and the functional $\psi(.)$ can be termed a variable projection functional since $\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}$ is an orthogonal projector involving only the vectorized form of the $\mathbf{A}$ matrix [63]. Again, if we take into account the block structure of $\mathbf{F}(\mathbf{a})$, we obtain an alternative formulation of $\psi(.)$, which is useful for computational purposes,

$$\psi(\mathbf{a}) = \frac{1}{2}\sum_{j=1}^{n}\psi_j(\mathbf{a}) \,,$$

where $\psi_j(.)$ denotes the j^{th} atomic function, which is defined for all $\mathbf{a} \in \mathbb{R}^{p.k}$, by

$$\begin{aligned}
\psi_j(\mathbf{a}) &= \|\mathbf{P}^{\perp}_{\mathbf{F}_j(\mathbf{a})}\mathbf{x}_j\|_2^2 \\
&= \|\big(\mathbf{I}_p - \mathbf{F}_j(\mathbf{a})\mathbf{F}_j(\mathbf{a})^+\big)\mathbf{x}_j\|_2^2 \\
&= \| \left(\mathbf{I}_p - \left(diag(\sqrt{\mathbf{W}}_{.j})\mathbf{A}\right)\left(diag(\sqrt{\mathbf{W}}_{.j})\mathbf{A}\right)^+\right)(\sqrt{\mathbf{W}}_{.j} \odot \mathbf{X}_{.j})\|_2^2 \,.
\end{aligned} \tag{3.25}$$

Here $\mathbf{P}^{\perp}_{\mathbf{F}_j(\mathbf{a})}$ is the orthogonal projector onto the orthogonal complement of $ran\big(\mathbf{F}_j(\mathbf{a})\big)$ and $\mathbf{x}_j = \sqrt{\mathbf{W}}_{.j} \odot \mathbf{X}_{.j}$.

This new formulation of our NLLS problem, based on the cost function $\psi(.)$, suggests that the minimization of $\varphi^*(.)$ can be separated in two steps. Once a $\mathbf{A}$ matrix has been obtained by minimizing $\psi(\mathbf{a})$, the $\mathbf{B}$ matrix can be obtained by solving a large block diagonal least-squares problem, which is equivalent to solve n independent smaller linear least-squares problems. The rational for employing this separation of variables to minimize $\varphi^*(.)$ is given by the following theorem, which is a

slight variation of a theorem originally proved by Golub and Pereyra in a more general setting (see Theorem 2.1 in [63]).

Theorem 3.9. With the same notations and definitions as in Theorem 3.1, the problem (P1) is equivalent to the problem

$$\min_{\mathbf{A}\in\mathbb{R}^{p\times k}} \psi\big(vec(\mathbf{A}^T)\big) = \psi(\mathbf{a}) = \frac{1}{2}\|\mathbf{P}^{\perp}_{\mathbf{F(a)}}\mathbf{x}\|_2^2 \,, \tag{VP1}$$

where $\mathbf{a} = vec(\mathbf{A}^T) \in \mathbb{R}^{p.k}$ and $\mathbf{x} = vec(\sqrt{\mathbf{W}} \odot \mathbf{X}) \in \mathbb{R}^{p.n}$. In other words, if we consider the range of $\varphi^*(.)$, C_{φ^*}, and the range of $\psi(.)$,

$$\mathrm{C}_{\psi} = \big\{\mathbf{y} \in \mathbb{R} \,/\, \exists \mathbf{A} \in \mathbb{R}^{p\times k} \text{ with } \mathbf{y} = \psi\big(vec(\mathbf{A}^T)\big)\big\} \,,$$

these two subsets of $\mathbb{R}$ have the same infimum and if this infimum is a minimum for one set, the other set also admits a minimum and these two minima are equal.

Proof. As in Theorem 3.1, C_{φ^*} and C_{ψ} are bounded below by zero and, thus, admit an infimum greater or equal to zero, say $\bar{\mathbf{c}}_{\varphi^*}$ and $\bar{\mathbf{c}}_{\psi}$, respectively.
Suppose first that $\bar{\mathbf{c}}_{\psi} < \bar{\mathbf{c}}_{\varphi^*}$. Then, it exists $\mathbf{A} \in \mathbb{R}^{p\times k}$ such that

$$\bar{\mathbf{c}}_{\psi} \leq \psi\big(vec(\mathbf{A}^T)\big) = \psi(a) < \bar{\mathbf{c}}_{\varphi^*} \,,$$

where $\mathbf{a} = vec(\mathbf{A}^T) \in \mathbb{R}^{p.k}$. Now, let $\mathbf{b} = \mathbf{F}(\mathbf{a})^{+}\mathbf{x} \in \mathbb{R}^{k.n}$ and define $\mathbf{B} = mat(\mathbf{b}) \in \mathbb{R}^{k\times n}$, we have

$$\mathbf{P}^{\perp}_{\mathbf{F(a)}}\mathbf{x} = \big(\mathbf{I}_{p.n} - \mathbf{F}(\mathbf{a})\mathbf{F}(\mathbf{a})^{+}\big)\,\mathbf{x} = \mathbf{x} - \mathbf{F}(\mathbf{a})\mathbf{b}$$

and

$$\|\mathbf{P}^{\perp}_{\mathbf{F(a)}}\mathbf{x}\|_2^2 = \|\mathbf{x} - \mathbf{F}(\mathbf{a})\mathbf{b}\|_2^2 = \|\sqrt{\mathbf{W}} \odot (\mathbf{X} - \mathbf{AB})\|_F^2 \,,$$

which implies that $\psi(\mathbf{a}) = \varphi^*(\mathbf{A}, \mathbf{B})$. In other words, we have $\varphi^*(\mathbf{A}, \mathbf{B}) < \bar{\mathbf{c}}_{\varphi^*}$, which contradicts the assertion that $\bar{\mathbf{c}}_{\varphi^*}$ is the infimum of $\varphi^*(.)$. This shows that $\bar{\mathbf{c}}_{\psi} \geq \bar{\mathbf{c}}_{\varphi^*}$.
Suppose now that $\bar{\mathbf{c}}_{\varphi^*} < \bar{\mathbf{c}}_{\psi}$. Then, it exists $(\mathbf{A}, \mathbf{B}) \in \mathbb{R}^{p\times k} \times \mathbb{R}^{k\times n}$ such that $\bar{\mathbf{c}}_{\varphi^*} \leq \varphi^*(\mathbf{A}, \mathbf{B}) < \bar{\mathbf{c}}_{\psi}$, otherwise $\bar{\mathbf{c}}_{\varphi^*}$ is not the infimum of $\varphi^*(.)$. However, if we define $\widehat{\mathbf{b}} = \mathbf{F}(\mathbf{a})^{+}\mathbf{x} \in \mathbb{R}^{k.n}$ and $\widehat{\mathbf{B}} = mat(\widehat{\mathbf{b}})$, we have

$$\|\sqrt{\mathbf{W}} \odot (\mathbf{X} - \mathbf{A}\widehat{\mathbf{B}})\|_F^2 \leq \|\sqrt{\mathbf{W}} \odot (\mathbf{X} - \mathbf{AB})\|_F^2 \,,$$

as, for a fixed $\mathbf{A}$ matrix, $\widehat{\mathbf{b}} = vec(\widehat{\mathbf{B}})$ is the solution of the least-squares problem

$$\min_{\mathbf{b}\in\mathbb{R}^{k.n}} \|\mathbf{x} - \mathbf{F}(\mathbf{a})\mathbf{b}\|_2^2 = \min_{\mathbf{B}\in\mathbb{R}^{k\times n}} \|\sqrt{\mathbf{W}} \odot (\mathbf{X} - \mathbf{AB})\|_F^2$$

This implies that $\psi(\mathbf{a}) = \varphi^*(\mathbf{A}, \widehat{\mathbf{B}}) \leq \varphi^*(\mathbf{A}, \mathbf{B}) < \bar{\mathbf{c}}_{\psi}$, which contradicts the assertion that $\bar{\mathbf{c}}_{\psi}$ is the infimum of $\psi(.)$. This demonstrates that $\bar{\mathbf{c}}_{\varphi^*} \geq \bar{\mathbf{c}}_{\psi}$.
Finally, the inequalities $\bar{\mathbf{c}}_{\varphi^*} \geq \bar{\mathbf{c}}_{\psi}$ and $\bar{\mathbf{c}}_{\varphi^*} \leq \bar{\mathbf{c}}_{\psi}$ imply that $\bar{\mathbf{c}}_{\varphi^*} = \bar{\mathbf{c}}_{\psi}$, which proves the first part of the theorem.

Now assume that $\widehat{\mathbf{A}}$ minimizes $\psi(.)$, e.g., $\psi(vec(\widehat{\mathbf{A}}^T)) = \psi(\widehat{\mathbf{a}}) = \bar{\mathbf{c}}_{\psi}$. If we let $\widehat{\mathbf{b}} = \mathbf{F}(\widehat{\mathbf{a}})^{+}\mathbf{x}$ and $\widehat{\mathbf{B}} = mat(\widehat{\mathbf{b}})$, we have

$$\psi(\widehat{\mathbf{a}}) = \frac{1}{2}\|\mathbf{P}^{\perp}_{\mathbf{F}(\widehat{\mathbf{a}})}\mathbf{x}\|_2^2 = \frac{1}{2}\|\mathbf{x} - \mathbf{F}(\widehat{\mathbf{a}})\widehat{\mathbf{b}}\|_2^2 = \frac{1}{2}\|\sqrt{\mathbf{W}} \odot (\mathbf{X} - \widehat{\mathbf{A}}\widehat{\mathbf{B}})\|_F^2 = \varphi^*(\widehat{\mathbf{A}}, \widehat{\mathbf{B}})$$

and the equalities $\bar{\mathbf{c}}_{\varphi^*} = \bar{\mathbf{c}}_{\psi}$ and $\psi(\widehat{\mathbf{a}}) = \bar{\mathbf{c}}_{\psi}$ show that $\varphi^*(\widehat{\mathbf{A}}, \widehat{\mathbf{B}}) = \bar{\mathbf{c}}_{\varphi^*}$ and we conclude that $(\widehat{\mathbf{A}}, \widehat{\mathbf{B}})$ is a global minimizer of $\varphi^*(.)$.

Reciprocally, assume that $(\widehat{\mathbf{A}}, \widehat{\mathbf{B}})$ minimizes $\varphi^*(.)$, e.g., $\varphi^*(\widehat{\mathbf{A}}, \widehat{\mathbf{B}}) = \bar{\mathbf{c}}_{\varphi^*}$. Let $\widehat{\mathbf{a}} = \mathit{mat}(\widehat{\mathbf{A}}^T)$, $\bar{\mathbf{b}} = \mathbf{F}(\widehat{\mathbf{a}})^+\mathbf{x}$ and $\bar{\mathbf{B}} = \mathit{mat}(\bar{\mathbf{b}})$, we have

$$\bar{\mathbf{c}}_\psi \leq \psi(\widehat{\mathbf{a}}) = \varphi^*(\widehat{\mathbf{A}}, \bar{\mathbf{B}}) \leq \varphi^*(\widehat{\mathbf{A}}, \widehat{\mathbf{B}}) = \bar{\mathbf{c}}_{\varphi^*} = \bar{\mathbf{c}}_\psi \ ,$$

which implies that $\psi(\widehat{\mathbf{a}}) = \bar{\mathbf{c}}_\psi$ and $\widehat{\mathbf{a}}$ is a global minimizer of $\psi(.)$ and we are done.

□

An alternative to the minimization of $\psi(.)$ in the variable projection approach can be introduced with the help of a QRCP (see equation (2.15) in Subsection 2.1) of the matrix $\mathbf{F}(\mathbf{a})$ of rank $r \leq k.n$

$$\mathbf{Q}(\mathbf{a})\mathbf{F}(\mathbf{a})\mathbf{P} = \begin{bmatrix} \mathbf{R} & \mathbf{S} \\ \mathbf{0}^{(p.n-r)\times r} & \mathbf{0}^{(p.n-r)\times(k.n-r)} \end{bmatrix} \ ,$$

where $\mathbf{Q}(\mathbf{a})$ is an $p.n \times p.n$ orthogonal matrix, $\mathbf{P}$ is an $k.n \times k.n$ permutation matrix, $\mathbf{R}$ is an $r \times r$ nonsingular upper triangular matrix and $\mathbf{S}$ an $r \times (k.n - r)$ full matrix. Then, if $\mathbf{Q}(\mathbf{a})$ is partitioned into

$$\mathbf{Q}(\mathbf{a}) = \begin{bmatrix} \mathbf{Q}_1(\mathbf{a}) \\ \mathbf{Q}_2(\mathbf{a}) \end{bmatrix} \ ,$$

where $\mathbf{Q}_1(\mathbf{a})$ and $\mathbf{Q}_2(\mathbf{a})$ are, respectively, $r \times p.n$ and $(p.n - r) \times p.n$ submatrices, using results in Subsection 2.1, we have

$$\mathbf{Q}(\mathbf{a})\mathbf{P}^\perp_{\mathbf{F}(\mathbf{a})} = \begin{bmatrix} \mathbf{0}^{r\times p.n} \\ \mathbf{Q}_2(\mathbf{a}) \end{bmatrix} \ .$$

This implies that, for all $\mathbf{a} \in \mathbb{R}^{k.p}$ and $\mathbf{x} = \mathit{vec}(\sqrt{\mathbf{W}} \odot \mathbf{X}) \in \mathbb{R}^{p.n}$,

$$\|\mathbf{P}^\perp_{\mathbf{F}(\mathbf{a})}\mathbf{x}\|_2^2 = \|\mathbf{Q}(\mathbf{a})\mathbf{P}^\perp_{\mathbf{F}(\mathbf{a})}\mathbf{x}\|_2^2 = \|\mathbf{Q}_2(\mathbf{a})\mathbf{x}\|_2^2 \ ,$$

as first noted by Krogh [95] and Kaufman [96] and later by Shen and Ypma [175]. Thus, instead of minimizing the variable projection functional $\psi(.)$ to solve the (VP1) problem, we can minimize the variable orthogonal functional $\psi^*(.)$ defined by

$$\psi^*(\mathbf{a}) = \|\mathbf{Q}_2(\mathbf{a})\mathbf{x}\|_2^2 \ , \tag{3.26}$$

assuming that the rank of $\mathbf{F}(\mathbf{a})$ stays constant in a neighborhood of a solution $\widehat{\mathbf{a}}$ of the (VP1) problem. Note that this condition is also implicit when using the variable projection functional $\psi(.)$ as this condition is required both for the differentiation of $\mathbf{P}^\perp_{\mathbf{F}(.)}$ and $\mathbf{Q}_2(.)$ in a neighborhood of $\widehat{\mathbf{a}}$ as we will illustrate below. Thus, in the first step, it is mathematically equivalent to minimize $\psi(.)$ or $\psi^*(.)$, even though minimizing $\psi^*(.)$ may involved slightly different numerical algorithms [96][109][175]. Once a minimum of $\psi^*(.)$ has been determined, one can again determine $\widehat{\mathbf{b}}$ by solving the linear least-squares problem

$$\min_{\mathbf{b}\in\mathbb{R}^{k.n}} \|\mathbf{x} - \mathbf{F}(\mathbf{a})\mathbf{b}\|_2^2 \ ,$$

according to Theorem 3.9.

As there is no constraint on the rank of $\mathbf{A} \in \mathbb{R}^{p\times k}$ in the (VP1) problem stated in Theorem 3.9, the search space for minimizing the cost functions $\psi(.)$ or $\psi^*(.)$ is at first sight the linear space $\mathbb{R}^{p\times k}$. However, the following corollaries demonstrate that we can restrict this search space to the submanifold $\mathbb{R}^{p\times k}_k$ or even to $\mathbb{O}^{p\times k}$, the set of $p \times k$ matrices with orthonormal columns, which is called the Stiefel manifold [11]. Moreover, in many practical instances, for example in the matrix completion problem in which we are looking for a matrix $\widehat{\mathbf{X}}$ of specified and fixed rank k, which agrees with the observed entries of the input matrix $\mathbf{X}$, or when we solve the WLRA problem to estimate a consistent factor or principal component model, restricting the search space to the submanifold $\mathbb{R}^{p\times k}_k$ or even to the Stiefel submanifold is fully justified.

Corollary 3.1. With the same notations and definitions as in Theorem 3.9, the problem (P1) is also equivalent to the following alternative formulations of the problem (VP1) in which the search space is restricted to $\mathbb{R}_k^{p\times k}$ and $\mathbb{O}^{p\times k}$, respectively:

$$\min_{\mathbf{A}\in\mathbb{R}_k^{p\times k}} \psi\big(vec(\mathbf{A}^T)\big) = \psi(\mathbf{a}) = \frac{1}{2}\|\mathbf{P}^{\perp}_{\mathbf{F(a)}}\mathbf{x}\|_2^2$$

and

$$\min_{\mathbf{A}\in\mathbb{O}^{p\times k}} \psi\big(vec(\mathbf{A}^T)\big) = \psi(\mathbf{a}) = \frac{1}{2}\|\mathbf{P}^{\perp}_{\mathbf{F(a)}}\mathbf{x}\|_2^2 \,,$$

where $\mathbf{a} = vec(\mathbf{A}^T) \in \mathbb{R}^{p.k}$ and $\mathbf{x} = vec(\sqrt{\mathbf{W}} \odot \mathbf{X}) \in \mathbb{R}^{p.n}$.

Proof. As in Remark 3.1, we note that for all $(\mathbf{A}, \mathbf{B}) \in \mathbb{R}^{p\times k} \times \mathbb{R}^{k\times n}$, the matrix product $\mathbf{AB}$ is of rank at most k and can be factored as $\mathbf{AB} = \mathbf{CD}$ with $(\mathbf{C}, \mathbf{D}) \in \mathbb{R}_k^{p\times k} \times \mathbb{R}^{k\times n}$, which implies that the range of $\varphi^*(.)$, C_{φ^*}, is equal to the set $\big\{y \in \mathbb{R} \,/\, \exists \mathbf{C} \in \mathbb{R}_k^{p\times k}, \exists \mathbf{D} \in \mathbb{R}^{k\times n} \text{ with } y = \varphi^*(\mathbf{C}, \mathbf{D})\big\}$. Taking into account this property, it is easy to verify that a slight modification of the demonstration of Theorem 3.9 leads to the assertion that the problem (P1) is also equivalent to the problem

$$\min_{\mathbf{A}\in\mathbb{R}_k^{p\times k}} \psi\big(vec(\mathbf{A}^T)\big) = \psi(\mathbf{a}) = \frac{1}{2}\|\mathbf{P}^{\perp}_{\mathbf{F(a)}}\mathbf{x}\|_2^2 \,.$$

We omit the details.
Now, since $\mathbb{O}^{p\times k} \subset \mathbb{R}_k^{p\times k}$ and any element $\mathbf{A}$ of $\mathbb{R}_k^{p\times k}$ can also be written as $\mathbf{A} = \mathbf{UR}$ where $\mathbf{U} \in \mathbb{O}^{p\times k}$ and $\mathbf{R} \in \mathbb{R}_k^{k\times k}$, for example by using the QR or SVD decompositions of $\mathbf{A}$, we have

$$\psi(\mathbf{a}) = \varphi^*(\mathbf{A}, \widehat{\mathbf{B}}) = \varphi^*(\mathbf{UR}, \widehat{\mathbf{B}}) = \varphi^*(\mathbf{U}, \mathbf{R}\widehat{\mathbf{B}}) \geq \psi(\mathbf{u}) \,,$$

where $\mathbf{a} = vec(\mathbf{A}^T)$, $\widehat{\mathbf{B}} = mat(\widehat{\mathbf{b}})$ with $\widehat{\mathbf{b}} = \mathbf{F}(\mathbf{a})^+\mathbf{x}$ and $\mathbf{u} = vec(\mathbf{U}^T)$. Reciprocally, if $\mathbf{A} = \mathbf{UR}$ with $\mathbf{U} \in \mathbb{O}^{p\times k}$ and $\mathbf{R} \in \mathbb{R}_k^{k\times k}$, we have

$$\psi(\mathbf{u}) = \varphi^*(\mathbf{U}, \widehat{\mathbf{D}}) = \varphi^*(\mathbf{AR}^{-1}, \widehat{\mathbf{D}}) = \varphi^*(\mathbf{A}, \mathbf{R}^{-1}\widehat{\mathbf{D}}) \geq \psi(\mathbf{a}) \,,$$

where $\widehat{\mathbf{D}} = mat(\widehat{\mathbf{d}})$ with $\widehat{\mathbf{d}} = \mathbf{F}(\mathbf{u})^+\mathbf{x}$. This implies that $\psi(\mathbf{a}) = \psi(\mathbf{u})$ if $\mathbf{A} = \mathbf{UR}$, which demonstrates that the range of $\psi(.)$, C_ψ, is equal to the set

$$\big\{y \in \mathbb{R} \,/\, \exists \mathbf{U} \in \mathbb{O}^{p\times k} \text{ with } y = \psi\big(vec(\mathbf{U}^T)\big)\big\} \,.$$

As a consequence, these two sets have the same infimum and the same minimum, if this minimum exists, and the problems (P1) or (VP1) are also equivalent to

$$\min_{\mathbf{A}\in\mathbb{O}^{p\times k}} \psi\big(vec(\mathbf{A}^T)\big) = \psi(\mathbf{a}) = \frac{1}{2}\|\mathbf{P}^{\perp}_{\mathbf{F(a)}}\mathbf{x}\|_2^2 \,,$$

as claimed in the Corollary.

□

Corollary 3.2. With the same definitions and notations as in Theorem 3.9 and Corollary 3.1, the following two assertions are true:

1) $\psi(vec(\mathbf{A}^T)) = \psi(vec(\mathbf{C}^T))$ if $\mathbf{A} = \mathbf{CD}$ with $\mathbf{A} \in \mathbb{R}_k^{p\times k}, \mathbf{C} \in \mathbb{R}_k^{p\times k}$ and $\mathbf{D} \in \mathbb{R}_k^{k\times k}$,
2) $\psi(vec(\mathbf{A}^T)) = \psi(vec(\mathbf{C}^T))$ if $\mathbf{A} = \mathbf{CD}$ with $\mathbf{A} \in \mathbb{O}^{p\times k}, \mathbf{C} \in \mathbb{O}^{p\times k}$ and $\mathbf{D} \in \mathbb{O}^{k\times k}$.

Proof. The proof is very similar to the one used in Corollary 3.1. Suppose first that $\mathbf{A} = \mathbf{CD}$ with $\mathbf{A} \in \mathbb{R}_k^{p\times k}, \mathbf{C} \in \mathbb{R}_k^{p\times k}$ and $\mathbf{D} \in \mathbb{R}_k^{k\times k}$ and let $\mathbf{a} = vec(\mathbf{A}^T)$ and $\mathbf{c} = vec(\mathbf{C}^T)$. We have

$$\psi(\mathbf{a}) = \varphi^*(\mathbf{A}, \widehat{\mathbf{B}}) = \varphi^*(\mathbf{CD}, \widehat{\mathbf{B}}) = \varphi^*(\mathbf{C}, \mathbf{D}\widehat{\mathbf{B}}) \geq \psi(\mathbf{c}) \,,$$

where $\widehat{\mathbf{B}} = mat(\widehat{\mathbf{b}})$ and $\widehat{\mathbf{b}} = \mathbf{F}(\mathbf{a})^{+}\mathbf{x}$. Reciprocally, we have

$$\psi(\mathbf{c}) = \varphi^*(\mathbf{C}, \widehat{\mathbf{E}}) = \varphi^*(\mathbf{A}\mathbf{D}^{-1}, \widehat{\mathbf{E}}) = \varphi^*(\mathbf{A}, \mathbf{D}^{-1}\widehat{\mathbf{E}}) \geq \psi(\mathbf{a}) \ ,$$

where $\widehat{\mathbf{E}} = mat(\widehat{\mathbf{e}})$ and $\widehat{\mathbf{e}} = \mathbf{F}(\mathbf{c})^{+}\mathbf{x}$. Finally, we obtain $\psi(vec(\mathbf{A}^T)) = \psi(vec(\mathbf{C}^T))$, and the first part of the corollary is demonstrated.

The proof of the second assertion is exactly similar and thus omitted.

□

Remark 3.7. Theorem 3.9 and Corollaries 3.1 and 3.2 also explain why the WLRA problem can be recast as an optimization problem on the Grassmann manifold $\mathrm{Gr}(p,k)$ [47][14], the collection of all linear subspaces of fixed dimension k of the Euclidean space $\mathbb{R}^p$, which is a smooth (quotient) manifold of dimension $k.(p-k)$ [3][11]. As stated in Theorems 3.1 and 3.9, the formulations (P0), (P1) and (VP1) of the WLRA problem are equivalent and Corollary 3.1 shows that the search space for the (VP1) problem can be restricted to $\mathbb{R}_k^{p\times k}$ or $\mathbb{O}^{p\times k}$ without loss of generality. Next, Corollary 3.2 demonstrates that $\psi(vec(\mathbf{A}^T)) = \psi(vec(\mathbf{C}^T))$ as soon as the linear subspaces $ran(\mathbf{A})$ and $ran(\mathbf{C})$ are equal. In other words, the value of $\psi(vec(\mathbf{A}^T))$ for $\mathbf{A} \in \mathbb{R}_k^{p\times k}$ depends only on the linear subspace $ran(\mathbf{A})$ as for any matrix $\mathbf{C} \in \mathbb{R}_k^{p\times k}$ such that the columns of $\mathbf{C}$ is a (orthonormal or not) basis of $ran(\mathbf{A})$, we have $\psi(vec(\mathbf{A}^T)) = \psi(vec(\mathbf{C}^T))$.
As such, an element $\mathcal{U}$ of $\mathrm{Gr}(p,k)$ can be represented by any $p \times k$ matrix $\mathbf{U}$ of full column-rank such that $ran(\mathbf{U}) = \mathcal{U}$, e.g., if the columns of $\mathbf{U}$ form a basis of $\mathcal{U}$. For numerical reasons, elements of $\mathrm{Gr}(p,k)$ are very often represented by elements of $\mathbb{O}^{p\times k}$ [51][125] [28][46][47][14], but any matrix with the same column space can be used, and we will see below that representing $\mathcal{U}$ by elements of $\mathbb{R}_k^{p\times k}$ instead of $\mathbb{O}^{p\times k}$ can also be useful to demonstrate important properties of the cost function $\psi(.)$ used in the (VP1) form of the WLRA problem, especially when the associated weight matrix $\mathbf{W}$ is not strictly positive.
Stated differently, we can say that two $p \times k$ (orthonormal) matrices $\mathbf{U}$ and $\mathbf{V}$ are equivalent if and only if they have the same column space or, equivalently, if it exists some $\mathbf{Q} \in \mathbb{R}_k^{k\times k}$ (or $\mathbf{Q} \in \mathbb{O}^{k\times k}$) such that $\mathbf{U} = \mathbf{V}\mathbf{Q}$. Because of this equivalence relationship in $\mathbb{R}_k^{p\times k}$, $\mathrm{Gr}(p,k)$ can be described as the quotient of $\mathbb{R}_k^{p\times k}$ by the action of $\mathbb{R}_k^{k\times k}$. Alternatively, $\mathrm{Gr}(p,k)$ can also be described as the quotient of $\mathbb{O}^{p\times k}$ by the action of $\mathbb{O}^{k\times k}$, see Subsection 2.4 for details. See [11][24] for a geometrical and comprehensive description of these different approaches of representing elements of $\mathrm{Gr}(p,k)$ with matrices.
Moreover, the fact that the formulation (P1) of the WLRA problem never has a unique or finite set of global minimizers is also related to the preceding discussion. If $\mathbf{Y} = \mathbf{A}\mathbf{B}$ with $\mathbf{A} \in \mathbb{R}_k^{p\times k}$, the elements of the columns of $\mathbf{B} \in \mathbb{R}^{k\times p}$ are the coordinates of the corresponding columns of $\mathbf{Y}$ in the particular basis of $ran(\mathbf{A})$ provided by the columns of $\mathbf{A}$ and, thus, these coordinates depend on the choice of the basis.
Finally, in a similar fashion that the formulation (VP1) and the associated variable projection functionals $\psi(.)$ and $\psi^*(.)$ are derived from the formulation (P1) of the WLRA problem, it is also possible to reformulate the problem (P2) (see Remark 3.2) as a double-minimization problem

$$\min_{\mathbf{N}\in\mathbb{R}_{p-k}^{p\times(p-k)}} \quad \left(\min_{\mathbf{Y}\in\mathbb{R}^{p\times n} \text{ with } \mathbf{N}^T\mathbf{Y}=\mathbf{0}^{(p-k)\times n}} \frac{1}{2}\|\sqrt{\mathbf{W}} \odot (\mathbf{X}-\mathbf{Y})\|_F^2 \right) \ ,$$

which will lead to a dual formulation (VP2) of the WLRA problem and its associated variable projection functional $\psi^{**}(.)$ on the Grassmann manifold $\mathrm{Gr}(p,p-k)$ [125]. More precisely, Manton et al. [125] have demonstrated that the above inner minimization problem has a closed form solution (see Theorem 1 in [125]), which can be calculated analytically and depends only on the range of $\mathbf{N}$ and not on the particular matrix $\mathbf{N}$ when $\mathbf{W} \in \mathbb{R}_{+*}^{p\times n}$ (and these results can be probably extended to the case $\mathbf{W} \in \mathbb{R}_{+}^{p\times n}$). Thus, problem (P2) is also a separable NLLS problem as stated in the Definition 2.10 of Subsection 2.4 despite the variable matrix $\mathbf{Y}$ does not occur linearly in the

residual function associated with problem (P2). This situation is exactly similar to the one described above for the problem (VP1) and its associated variable projection functional $\psi(.)$ where we proceed in two steps, namely, first find the matrix $\widehat{\mathbf{A}}$ such that $\psi(\widehat{\mathbf{a}})$ is minimum and, second, determine $\widehat{\mathbf{B}}$ by solving a large block diagonal least-squares problem. In other words, for $\mathbf{N} \in \mathbb{R}_{p-k}^{p\times(p-k)}$, the dual variable projection functional

$$\psi^{**}(\mathbf{N}) = \min_{\mathbf{Y}\in\mathbb{R}^{p\times n} \text{ with } \mathbf{N}^T\mathbf{Y}=\mathbf{0}^{(p-k)\times n}} \frac{1}{2}\|\sqrt{\mathbf{W}} \odot (\mathbf{X}-\mathbf{Y})\|_F^2 \tag{3.27}$$

is well defined and we can attempt to find a solution $\widehat{\mathbf{N}}$ (more precisely a subspace $\widehat{\mathcal{N}}$, which is represented by $\widehat{\mathbf{N}}$) of the problem

$$\min_{\mathbf{N}\in\mathbb{R}_{p-k}^{p\times(p-k)}} \quad \psi^{**}(\mathbf{N}) \tag{VP2}$$

in a first step by any iterative first- or second-order method working on the Grassmann manifold $\mathrm{Gr}(p, p-k)$ [51][125]. Once a minimum $\widehat{\mathbf{N}}$ has been found, the best rank-k approximation matrix $\widehat{\mathbf{Y}}$ can be determined in a second step by solving the inner minimization problem analytically for the matrix $\widehat{\mathbf{N}}$. As for the (VP1) problem, the search space for the (VP2) problem can be restricted to $\mathbb{O}^{p\times(p-k)}$ without loss of generality for numerical reasons [51][125]. Furthermore, it will be shown in the following sections, that minimizing the cost function $\psi(.)$ associated with the (VP1) problem reduces to one of dimension $k.(p-k)$ instead of $k.p$ as for the minimization of the cost function $\psi^{**}(.)$ associated with the (VP2) problem [51][125], highlighting again the duality between the (VP1) and (VP2) formulations of the WLRA problem. ■

We address now the question of the continuity of the cost function $\psi(.)$, which must be minimized in the (VP1) problem, as was done above for the cost functions $\varphi(.)$ and $\varphi^*(.)$ associated, respectively, with the formulations (P0) and (P1) of the WLRA problem. Taking into account that, for all $\mathbf{a} \in \mathbb{R}^{p.k}$, we have

$$\psi(\mathbf{a}) = \frac{1}{2}\|\mathbf{P}_{\mathbf{F}(\mathbf{a})}^{\perp}\mathbf{x}\|_2^2 \,, \tag{3.28}$$

we see that the continuity of $\psi(.)$ is closely associated to the continuity of the orthogonal projector $\mathbf{P}_{\mathbf{F}(.)}^{\perp}$ (or equivalently $\mathbf{P}_{\mathbf{F}(.)}$) as a function of $\mathbf{a} \in \mathbb{R}^{p.k}$. Furthermore, due to the block diagonal structure of $\mathbf{F}(.)$, we observe that the continuity of $\mathbf{P}_{\mathbf{F}(.)}^{\perp}$ as a function of $\mathbf{a}$ is equivalent to the continuity of the n atomic orthogonal projectors $\mathbf{P}_{\mathbf{F}_j(.)}^{\perp}$, for $j = 1$ to n, since, for all $\mathbf{a} \in \mathbb{R}^{p.k}$, we have

$$\mathbf{P}_{\mathbf{F}(\mathbf{a})}^{\perp} = \bigoplus_{j=1}^{n} \mathbf{P}_{\mathbf{F}_j(\mathbf{a})}^{\perp} \,.$$

The next theorem gives necessary and sufficient conditions for the continuity of a general orthogonal projector $\mathbf{P}_{\Phi(.)}$, which is associated with a $l \times t$ matrix function $\Phi(.)$ of a vector $\mathbf{a} \in \mathbb{R}^m$, but let us first give the following definition:

Definition 3.1. Let $\Phi(.)$ be a matrix function defined as

$$\Phi : \mathbb{R}^m \longrightarrow \mathbb{R}^{l\times t} : \mathbf{a} \mapsto \Phi(\mathbf{a}) \,.$$

We say that the matrix function $\Phi(.)$ has a local constant rank at a point $\mathbf{a}_0 \in \mathbb{R}^m$ if there exists an open neighborhood Υ of $\mathbf{a}_0$ in $\mathbb{R}^m$ such that the matrix $\Phi(\mathbf{a})$ has a constant rank $q \leq \min(l, t)$ for all $\mathbf{a} \in \Upsilon$.

We then restate the following fundamental result about the continuity of the Moore-Penrose inverse and orthogonal projectors for real matrix functions, which can be traced back to the seminal works of Wedin [189] and Stewart [172].

Theorem 3.10. Let $\Phi(.)$ be a matrix function : $\mathbb{R}^m \longrightarrow \mathbb{R}^{l\times t}$, which is continuous at a point $\mathbf{a}_0 \in \mathbb{R}^m$. The following conditions are equivalent.

1) $\Phi(.)$ has a local constant rank at $\mathbf{a}_0$
2) $\Phi(.)^+$ is continuous at $\mathbf{a}_0$
3) $\Phi(.)\Phi(.)^+ = \mathbf{P}_{\Phi(.)}$ is continuous at $\mathbf{a}_0$
4) $\Phi(.)^+\Phi(.) = \mathbf{P}_{\Phi(.)^T}$ is continuous at $\mathbf{a}_0$

In other words, the continuity of the pseudo-inverse of a continuous matrix function $\Phi(.)$ at a point $\mathbf{a}_0 \in \mathbb{R}^m$ is equivalent to the continuity of the orthogonal projectors onto the column or row spaces of this matrix function at $\mathbf{a}_0$ and all these conditions are equivalent to the assertion that $\Phi(.)$ has local constant rank at $\mathbf{a}_0$ if $\Phi(.)$ is itself continuous at $\mathbf{a}_0$.

Proof. See Propositions 8.1 and 8.2 in Chapter 8 of Magnus and Neudecker [124] or Chapter 10 of Campbell and Meyer [29].

□

To ease the notation burden in the rest of this section and the following sections, we define the following linear mapping $h(.)$ and its inverse mapping $h^{-1}(.)$:

$$\begin{aligned} h &: \mathbb{R}^{p.k} \longrightarrow \mathbb{R}^{p\times k}, \mathbf{a} \mapsto [mat_{k\times p}(\mathbf{a})]^T = \mathbf{A} \,, \\ h^{-1} &: \mathbb{R}^{p\times k} \longrightarrow \mathbb{R}^{p.k}, \mathbf{A} \mapsto vec(\mathbf{A}^T) = \mathbf{a} \,, \end{aligned} \tag{3.29}$$

which are homeomorphisms from $\mathbb{R}^{p.k}$ to $\mathbb{R}^{p\times k}$ and from $\mathbb{R}^{p\times k}$ to $\mathbb{R}^{p.k}$, respectively, allowing to identify the elements and the topologies of these two finite dimensional vector spaces. These notations seem cumbersome, but are related to our definition of the vectorized form of the matrix $\mathbf{A}$ as $vec(\mathbf{A}^T)$ instead of $vec(\mathbf{A})$, which will be justified in the next sections. In other words, with these definitions, we have $h^{-1}(\mathbf{A}) = vec(\mathbf{A}^T) = \mathbf{a}$ for all $\mathbf{A} \in \mathbb{R}^{p\times k}$.

Armed with Theorem 3.10, we now consider the continuity of the orthogonal projector $\mathbf{P}^{\perp}_{\mathbf{F}(.)}$ (or equivalently $\mathbf{P}_{\mathbf{F}(.)}$), which is used in the cost function $\psi(.)$ of the (VP1) problem. We first observe that $\mathbf{F}(.)$, as a function of $\mathbf{a}$, is a linear mapping from $\mathbb{R}^{p.k}$ to $\mathbb{R}^{p.n\times n.k}$ since the *mat* and transpose operators are linear mappings, and, the Kronecker and matrix products are bilinear mappings. As $\mathbb{R}^{p.k}$ and $\mathbb{R}^{p.n\times n.k}$ are finite dimensional vector spaces, $\mathbf{F}(.)$ is thus continuous for all $\mathbf{a} \in \mathbb{R}^{p.k}$. Similarly, the n atomic matrix functions $\mathbf{F}_j(.)$ are also continuous linear mappings from $\mathbb{R}^{p.k}$ to $\mathbb{R}^{p\times k}$. In these conditions, Theorem 3.9 shows that the continuity of the orthogonal projectors $\mathbf{P}^{\perp}_{\mathbf{F}(.)}$ and $\mathbf{P}^{\perp}_{\mathbf{F}_j(.)}$ (or equivalently $\mathbf{P}_{\mathbf{F}(.)}$ and $\mathbf{P}_{\mathbf{F}_j(.)}$) at a point $\mathbf{a} \in \mathbb{R}^{p.k}$ is equivalent, respectively, to the propositions that $\mathbf{F}(.)$ and $\mathbf{F}_j(.)$ have a local constant rank at $\mathbf{a}$. Furthermore, the proposition that $\mathbf{F}(.)$ has a local constant rank at $\mathbf{a}$ is equivalent to the proposition that all the n $\mathbf{F}_j(.)$ functions, for $j = 1$ to n, have also a local constant rank at $\mathbf{a}$.

In the special case where the weight matrix $\mathbf{W} \in \mathbb{R}^{p\times n}_{+*}$ (e.g., $\mathbf{W}_{ij} > 0$), we have the following more precise result:

Theorem 3.11. For $\mathbf{X} \in \mathbb{R}^{p\times n}$ and $\mathbf{W} \in \mathbb{R}^{p\times n}_{+*}$, and any fixed integer $k \leq rank(\mathbf{X}) \leq \min(p, n)$, define the matrix function $\mathbf{P}_{\mathbf{F}(.)}$, from $\mathbb{R}^{p.k}$ to $\mathbb{R}^{p.n\times p.n}$, by

$$\mathbf{a} \mapsto \mathbf{P}_{\mathbf{F}(\mathbf{a})} = \mathbf{F}(\mathbf{a})\mathbf{F}(\mathbf{a})^+ \,,$$

where $\mathbf{F}(\mathbf{a})^+$ is the pseudo-inverse of $\mathbf{F}(\mathbf{a})$ and $\mathbf{F}(\mathbf{a})$ is the $p.n \times n.k$ block diagonal matrix

$$\mathbf{F}(\mathbf{a}) = \bigoplus_{j=1}^{n} diag(\sqrt{\mathbf{W}}_{.j})h(\mathbf{a}) = \bigoplus_{j=1}^{n} diag(\sqrt{\mathbf{W}}_{.j})\mathbf{A} \,.$$

$\mathbf{P}_{\mathbf{F}(.)}$ is continuous at all $\mathbf{a} \in h^{-1}(\mathbb{R}^{p\times k}_k)$ and discontinuous at all $\mathbf{a} \in h^{-1}(\mathbb{R}^{p\times k}_{<k})$.

Proof. As noted above, the continuity of $\mathbf{P}_{\mathbf{F}(.)}$ at a point $\mathbf{a}_0 \in \mathbb{R}^{p.k}$ is equivalent to the existence of an open neighborhood of $\mathbf{a}_0$ in $\mathbb{R}^{p.k}$ in which $\mathbf{F}(.)$ has a local constant rank. However, since

$$\begin{aligned}\mathbf{F}(\mathbf{a}) &= \bigoplus_{j=1}^{n} diag(\sqrt{\mathbf{W}}_{.j})\mathbf{A} \\ &= diag\big(vec(\sqrt{\mathbf{W}})\big)\big[\bigoplus_{j=1}^{n}\mathbf{A}\big]\end{aligned}$$

and $rank\big(diag(vec(\sqrt{\mathbf{W}}))\big) = p.n$ if $\mathbf{W}$ is strictly positive, then

$$rank\big(\mathbf{F}(\mathbf{a})\big) = rank(\bigoplus_{j=1}^{n}\mathbf{A}) = n.rank(\mathbf{A})\ .$$

Thus, the rank of $\mathbf{F}(\mathbf{a})$ is entirely determined by the rank of $\mathbf{A} = h(\mathbf{a})$ when the weight matrix is strictly positive. In other words, the continuity of $\mathbf{P}_{\mathbf{F}(.)}$ at $\mathbf{a}_0 \in \mathbb{R}^{p.k}$ is equivalent to the existence of an open neighborhood Υ of $\mathbf{A}_0 = h(\mathbf{a}_0) \in \mathbb{R}^{p\times k}$ such that for all $\mathbf{A} \in \Upsilon$, the rank of $\mathbf{A} = h(\mathbf{a})$ is constant.

Now, we have $\mathbb{R}^{p\times k} = \mathbb{R}_k^{p\times k} \bigcup \mathbb{R}_{<k}^{p\times k}$ and let us consider separately the two cases $\mathbf{A}_0 \in \mathbb{R}_k^{p\times k}$ and $\mathbf{A}_0 \in \mathbb{R}_{<k}^{p\times k}$.

Suppose first that $\mathbf{A}_0 \in \mathbb{R}_k^{p\times k}$. Per definition, the rank of $\mathbf{A}_0$ is constant and equal to k. Note further that $\mathbb{R}_k^{p\times k}$ is an open set of $\mathbb{R}^{p\times k}$ as the preimage of the open set $\mathbb{R}\backslash\{0\}$ under the continuous mapping $\mathbf{A} \mapsto det(\mathbf{A}^T\mathbf{A})$ as stated in Theorem 2.3. In other words, for all $\mathbf{A}_0 \in \mathbb{R}_k^{p\times k}$, there is an open neighborhood Υ of $\mathbf{A}_0$ included in $\mathbb{R}_k^{p\times k}$ and we deduce immediately that $\mathbf{P}_{\mathbf{F}(.)}$ is a continuous mapping for all $\mathbf{a}_0 \in h^{-1}(\mathbb{R}_k^{p\times k})$ using Theorem 3.10.

Suppose now that $\mathbf{A}_0 \in \mathbb{R}_{<k}^{p\times k}$. Since $\mathbb{R}_{<k}^{p\times k}$ is the frontier of the open set $\mathbb{R}_k^{p\times k}$ in $\mathbb{R}^{p\times k}$ according to Theorem 2.3, every open neighborhood Υ of $\mathbf{A}_0$ in $\mathbb{R}_{<k}^{p\times k}$ also contains some points $\mathbf{A} \in \mathbb{R}_k^{p\times k}$ and as such if $\mathbf{A}_0 \in \mathbb{R}_{<k}^{p\times k}$ there is no open neighborhood Υ of $\mathbf{A}_0$ in $\mathbb{R}^{p\times k}$ in which the rank of $\mathbf{A}$ is constant for all $\mathbf{A} \in \Upsilon$. This implies that $\mathbf{P}_{\mathbf{F}(.)}$ (and, thus, also $\mathbf{P}_{\mathbf{F}(.)}^{\perp}$ and $\psi(.)$) is discontinuous at all points $\mathbf{a}_0 \in \mathbb{R}^{p.k}$ such that $h(\mathbf{a}_0) = \mathbf{A}_0 \in \mathbb{R}_{<k}^{p\times k}$.

□

Corollary 3.3. With the same definitions and notations as in Theorem 3.11, the cost function $\psi(.)$ of the (VP1) problem

$$\psi(\mathbf{a}) = \frac{1}{2}\|\mathbf{P}_{\mathbf{F}(\mathbf{a})}^{\perp}\mathbf{x}\|_2^2$$

is continuous at all points $\mathbf{a} \in h^{-1}(\mathbb{R}_k^{p\times k})$, if the weight matrix $\mathbf{W} \in \mathbb{R}_{+*}^{p\times n}$ (e.g., if $\mathbf{W}$ is strictly positive). Furthermore, the sets of global minimizers of $\psi(.)$ over the subsets $\mathbb{R}_k^{p\times k}$ and $\mathbb{O}^{p\times k}$ of $\mathbb{R}^{p\times k}$ are not empty if $\mathbf{W} \in \mathbb{R}_{+*}^{p\times n}$.

Proof. From Theorem 3.11 above, we know that $\mathbf{P}_{\mathbf{F}(.)}^{\perp}$ is a continuous mapping over the set $\mathbb{R}_k^{p\times k}$ if the weight matrix $\mathbf{W}$ is strictly positive. Thus, in the same conditions, the restriction of $\psi(.)$ to $\mathbb{R}_k^{p\times k}$ is the composition of several continuous mappings on their respective domains of definition and, consequently, the restriction of $\psi(.)$ to $\mathbb{R}_k^{p\times k}$ is continuous at all points in $\mathbb{R}_k^{p\times k}$.

The second part of the corollary results immediately from Theorems 3.3, 3.9 and Corollary 3.1, which show, respectively, that the set of solutions of the (WLRA) problem is not empty if the weight matrix is strictly positive and that the (WLRA) and (VP1) problems, or their variants, are equivalent. Alternatively, it results from the first part of the corollary and the fact that $\mathbb{O}^{p\times k}$ is

included in $\mathbb{R}_k^{p\times k}$ and is a compact set for the topology of $\mathbb{R}_k^{p\times k}$ induced by the topology of $\mathbb{R}^{p\times k}$ (as $\mathbb{O}^{p\times k}$ is compact in $\mathbb{R}^{p\times k}$ according to Theorem 2.3). In these conditions, $\psi(.)$ is a continuous mapping over $\mathbb{R}_k^{p\times k}$ and attains its infimum over the compact set $\mathbb{O}^{p\times k}$, which implies directly that the set of global minimizers of the (VP1) problem is not empty and we are done.

□

Consider the preimage of $\mathbb{R}_k^{p\times k}$ by $h(.)$, e.g., $h^{-1}(\mathbb{R}_k^{p\times k})$. As $h(.)$ is continuous and $\mathbb{R}_k^{p\times k}$ is an open set for the topology of $\mathbb{R}^{p\times k}$, $h^{-1}(\mathbb{R}_k^{p\times k})$ is also an open set of $\mathbb{R}^{p.k}$. Similarly, as $\mathbb{O}^{p\times k}$ is a compact set for the topology of $\mathbb{R}^{p\times k}$, $h^{-1}(\mathbb{O}^{p\times k})$ is a compact set of $\mathbb{R}^{p.k}$ (because $h(.)$ is a homeomorphism or more simply because the reciprocal image of the closed set $\mathbb{O}^{p\times k}$ by $h(.)$ is a closed set of $\mathbb{R}^{p.k}$ and $\|h(\mathbf{a})\|_F = \|\mathbf{a}\|_2$ for all $\mathbf{a} \in \mathbb{R}^{p.k}$). Then the preceding results suggest that it is much more convenient to restrict the domain of definition of the cost function $\psi(.)$ to the open set $h^{-1}(\mathbb{R}_k^{p\times k})$ or even to the compact set $h^{-1}(\mathbb{O}^{p\times k})$ when we try to solve the (VP1) problem or its convex variants with a strictly positive weight matrix. In these conditions, the (VP1) problem is a well-posed problem with a non empty set of solutions and the cost function $\psi(.)$ is continuous and even smooth over $h^{-1}(\mathbb{R}_k^{p\times k})$ or $h^{-1}(\mathbb{O}^{p\times k})$. More precisely, the fact that $\mathbf{F}(.)$ is a continuous linear mapping and has the local constant rank property for all neighborhoods included in $h^{-1}(\mathbb{R}_k^{p\times k})$ implies also that $\mathbf{P}_{\mathbf{F}(.)}^{\perp}$ (and thus also $\psi(.)$) is continuously and infinitely differentiable at each point of $h^{-1}(\mathbb{R}_k^{p\times k})$. See Theorem 8.4 in Chapter 8 of Magnus and Neudecker [124] and Subsection 5.2 for details.

We now discuss the continuity of $\mathbf{P}_{\mathbf{F}(.)}$, $\mathbf{P}_{\mathbf{F}(.)}^{\perp}$ and $\psi(.)$ in the case where some elements of the weight matrix $\mathbf{W}$ are equal to zero. We already know that the (VP1) problem is not well-posed in these conditions as the set of global minimizers of $\psi(.)$ over $h^{-1}(\mathbb{R}_k^{p\times k})$ or $h^{-1}(\mathbb{O}^{p\times k})$ can be empty, as already noted for the equivalent forms (P0) or (P1) of the WLRA problem. In this more difficult case, $\mathbf{P}_{\mathbf{F}(.)}$ and $\mathbf{P}_{\mathbf{F}(.)}^{\perp}$ are also generally discontinuous at all $\mathbf{a} \in h^{-1}(\mathbb{R}_{<k}^{p\times k})$, as illustrated by the following theorem.

Theorem 3.12. Let $\mathbf{X} \in \mathbb{R}^{p\times n}$ and $\mathbf{W} \in \mathbb{R}_+^{p\times n}$, and $k \leq \mathit{rank}(\mathbf{X})$. If $\mathbf{W}_{.j} \in \mathbb{R}_{+*}^{p}$ (e.g., if the j^{th} column of the weight matrix has no zero element), then the matrix function $\mathbf{P}_{\mathbf{F}_j(.)}$ from $\mathbb{R}^{p.k}$ to $\mathbb{R}^{p\times p}$ defined by

$$\mathbf{a} \mapsto \mathbf{P}_{\mathbf{F}_j(\mathbf{a})} = \mathbf{F}_j(\mathbf{a})\mathbf{F}_j(\mathbf{a})^+ ,$$

where $\mathbf{F}_j(\mathbf{a})^+$ is the pseudo-inverse of $\mathbf{F}_j(\mathbf{a})$ and $\mathbf{F}_j(\mathbf{a})$ is the $p \times k$ matrix

$$\mathbf{F}_j(\mathbf{a}) = \mathit{diag}(\sqrt{\mathbf{W}}_{.j})h(\mathbf{a}) = \mathit{diag}(\sqrt{\mathbf{W}}_{.j})\mathbf{A} ,$$

is continuous at all $\mathbf{a} \in h^{-1}(\mathbb{R}_k^{p\times k})$ and discontinuous at all $\mathbf{a} \in h^{-1}(\mathbb{R}_{<k}^{p\times k})$. Furthermore, in the same conditions, the matrix function $\mathbf{P}_{\mathbf{F}(.)}$ from $\mathbb{R}^{p.k}$ to $\mathbb{R}^{p.n\times p.n}$ defined by

$$\mathbf{a} \mapsto \mathbf{P}_{\mathbf{F}(\mathbf{a})} = \bigoplus_{j=1}^{n} \mathbf{P}_{\mathbf{F}_j(\mathbf{a})}$$

is also discontinuous at all $\mathbf{a} \in h^{-1}(\mathbb{R}_{<k}^{p\times k})$.

Proof. The proof is similar to the one of Theorem 3.11 and is thus omitted.

□

However, when some weights are equal to zero (e.g., when $\mathbf{W} \in \mathbb{R}_+^{p\times n}$ instead of $\mathbb{R}_{+*}^{p\times n}$), the condition that $h(\mathbf{a}) \in \mathbb{R}_k^{p\times k}$ is still necessary, but is not sufficient to ensure the continuity of the orthogonal projectors $\mathbf{P}_{\mathbf{F}(.)}$ and $\mathbf{P}_{\mathbf{F}(.)}^{\perp}$. More precisely, we will demonstrate now that the set of points of $h^{-1}(\mathbb{R}_k^{p\times k})$ for which $\mathbf{P}_{\mathbf{F}(.)}$ and $\mathbf{P}_{\mathbf{F}(.)}^{\perp}$ are discontinuous, is not always empty, can be even

infinite and grows in size with the number of zero-elements of the weight matrix (see Theorem 3.13 below). Finally, we will show that the j^{th} atomic function $\psi_j(.)$ is also discontinuous at all points of $h^{-1}(\mathbb{R}_k^{p\times k})$ for which $\mathbf{P}^{\perp}_{\mathbf{F}_j(.)}$ is discontinuous (see again Theorem 3.13) and, in these conditions, $\psi(.) = \frac{1}{2}\sum_{j=1}^{n}\psi_j(.)$ can be hardly continuous or differentiable at those points. These results generalize the examples given in Dai et al. [46][47] about the discontinuity of some of the j^{th} atomic functions $\psi_j(.)$ when some entries of the matrix $\mathbf{X}$ are missing in the case of binary weights and provide a systematic characterization of the subset of points of $h^{-1}(\mathbb{R}_k^{p\times k})$ for which some of the atomic functions $\psi_j(.)$ are discontinuous when the weight matrix $\mathbf{W}$ is not strictly positive. This systematic assessment of the discontinuities of $\psi(.)$ is possible because we consider $h^{-1}(\mathbb{R}_k^{p\times k})$ as the domain of definition of $\psi(.)$ instead of $h^{-1}(\mathbb{O}^{p\times k})$ as in Dai et al. [46][47], who used a Grassmann manifold's framework to minimize $\psi(.)$.

In order to identify precisely the points of $h^{-1}(\mathbb{R}_k^{p\times k})$ for which the j^{th} atomic function defined by

$$\psi_j(\mathbf{a}) = \left\|\mathbf{P}^{\perp}_{\mathbf{F}_j(\mathbf{a})}\mathbf{x}_j\right\|_2^2 = \left\|\left(\mathbf{I}_p - \mathbf{F}_j(\mathbf{a})\mathbf{F}_j(\mathbf{a})^{+}\right)\mathbf{x}_j\right\|_2^2, \forall \mathbf{a}\in\mathbb{R}^{p.k}\ ,$$

where $\mathbf{P}^{\perp}_{\mathbf{F}_j(\mathbf{a})}$ is the orthogonal projector onto the orthogonal complement of $ran\big(\mathbf{F}_j(\mathbf{a})\big)$ and $\mathbf{x}_j = \sqrt{\mathbf{W}}_{.j}\odot\mathbf{X}_{.j}$, is discontinuous, let us first define what we call the j^{th} barrier set $\mathcal{B}_j$ associated with this j^{th} atomic function $\psi_j(.)$ and the corresponding $p\times k$ matrix function $\mathbf{F}_j(.)$.

Definition 3.2. Let $\mathbf{W}\in\mathbb{R}_+^{p\times n}$ and define for all $\mathbf{a}\in\mathbb{R}^{p.k}$, the matrix function $\mathbf{F}(.)$

$$\mathbf{F}:\mathbb{R}^{p.k}\longrightarrow\mathbb{R}^{p.n\times k.n}:\mathbf{a}\mapsto\mathbf{F}(\mathbf{a}) = \bigoplus_{j=1}^{n}\mathbf{F}_j(\mathbf{a})\ ,$$

where $\mathbf{F}_j(\mathbf{a}) = diag(\sqrt{\mathbf{W}}_{.j})h(\mathbf{a})\in\mathbb{R}^{p\times k}$ is called the j^{th} matrix function. The j^{th} barrier set $\mathcal{B}_j$ associated with the j^{th} atomic and matrix functions, $\psi_j(.)$ and $\mathbf{F}_j(.)$, is the subset of $\mathbb{R}^{p.k}$ defined by

$$\mathcal{B}_j = \left\{\mathbf{a}\in\mathbb{R}^{p.k}\ /\ \exists\mathbf{A}\in\mathbb{R}_k^{p\times k}\text{ with }\mathbf{A}=h(\mathbf{a})\text{ and }\mathbf{F}_j(\mathbf{a})=\mathbf{0}^{p\times k}\right\}\ ,$$

where $\mathbf{0}^{p\times k}$ is the zero $p\times k$ matrix.

Remark 3.8. This terminology is due to Dai et al. [46][47], who illustrated by a few examples that some points belonging to the intersection of a barrier set $\mathcal{B}_j$ with $h^{-1}(\mathbb{O}^{p\times k})$ act as "barriers", which may prevent gradient descent algorithms used to solve the (VP1) problem from converging to a global minimum or infimum. ■

Remark 3.9. The subset $\mathcal{B}_j$ of $\mathbb{R}^{p.k}$ introduced in Definition 3.2 can also be defined as follows. Consider again the preimage of $\mathbb{R}_k^{p\times k}$ by $h(.)$, e.g., $h^{-1}(\mathbb{R}_k^{p\times k})$, which is an open set of $\mathbb{R}^{p.k}$ and also the continuous linear mapping $\mathbf{F}_j:\mathbb{R}^{p.k}\longrightarrow\mathbb{R}^{p\times k},\mathbf{a}\mapsto\mathbf{F}_j(\mathbf{a})$. The subset of $\mathbb{R}^{p.k}$ such that $\mathbf{F}_j(\mathbf{a}) = \mathbf{0}^{p\times k}$ is the null space of $\mathbf{F}_j(.)$, which is a closed linear subspace of $\mathbb{R}^{p.k}$. With these results, we have

$$\mathcal{B}_j = h^{-1}(\mathbb{R}_k^{p\times k})\cap\mathbf{F}_j^{-1}(\{\mathbf{0}^{p\times k}\}) = h^{-1}(\mathbb{R}_k^{p\times k})\cap null(\mathbf{F}_j)\ ,$$

e.g., $\mathcal{B}_j$ is the intersection of the preimages $h^{-1}(\mathbb{R}_k^{p\times k})$ and $\mathbf{F}_j^{-1}(\{\mathbf{0}^{p\times k}\})$, which is not open and nor closed in $\mathbb{R}^{p.k}$. ■

We first observe that $\mathcal{B}_j$ is empty if all elements of the column vector $\mathbf{W}_{.j}$ are greater than zero as in that case we have $rank\big(diag(\sqrt{\mathbf{W}}_{.j})\big) = p$ and, thus, $rank(\mathbf{F}_j(\mathbf{a})) = rank(\mathbf{A}) = k$ if $h(\mathbf{a}) = \mathbf{A}\in\mathbb{R}_k^{p\times k}$, and in these conditions $\mathbf{F}_j(\mathbf{a})\neq\mathbf{0}^{p\times k}$. On the other hand, if the number of zero

elements of the column vector $\mathbf{W}_{.j}$ is greater or equal to k, $\mathcal{B}_j$ is not empty and is an infinite subset of $h^{-1}(\mathbb{R}_k^{p\times k})$ as demonstrated in Theorem 3.13 below.

To demonstrate this proposition, let us introduce again some notations. Suppose that some elements of the column vector $\mathbf{W}_{.j}$ are equal to zero, which is equivalent to say that the corresponding elements of the column vector $\mathbf{X}_{.j}$ are missing. Then, let p_u be the number of zero elements of $\mathbf{W}_{.j}$ and $p_o = p - p_u$ the number of elements of $\mathbf{W}_{.j}$ which are different from zero, e.g., p_u and p_o are, respectively, the number of "unobserved" and "observed" entries in the column vector $\mathbf{X}_{.j}$. In this case, we can partition the vectors $\mathbf{X}_{.j}$ and $\mathbf{W}_{.j}$ as

$$\begin{bmatrix}\mathbf{X}_{.j}^u\\ \mathbf{X}_{.j}^o\end{bmatrix} = \mathbf{P}_j\mathbf{X}_{.j} \text{ and } \begin{bmatrix}\mathbf{W}_{.j}^u\\ \mathbf{W}_{.j}^o\end{bmatrix} = \begin{bmatrix}\mathbf{0}^{p_u}\\ \mathbf{W}_{.j}^o\end{bmatrix} = \mathbf{P}_j\mathbf{W}_{.j} \ .$$

Here $\mathbf{X}_{.j}^u \in \mathbb{R}^{p_u}$ is the "unobserved" part of the column vector $\mathbf{X}_{.j}$, $\mathbf{X}_{.j}^o \in \mathbb{R}^{p_o}$ is the "observed" part of this vector and $\mathbf{P}_j$ is any $p \times p$ permutation matrix, which reorders the elements of $\mathbf{W}_{.j}$ so that the zero elements of $\mathbf{W}_{.j}$ appear first. Obviously, $\mathbf{P}_j$ is not unique, but we can use any such permutation in what follows. Note also that this permutation matrix $\mathbf{P}_j$ could be different for each pair of column vectors $\mathbf{X}_{.j}$ and $\mathbf{W}_{.j}$, such that some of the elements of $\mathbf{W}_{.j}$ are equal to zero, if the patterns of "unobserved" entries differ among the columns of the matrix $\mathbf{X}$. Similarly, $\mathbf{A} \in \mathbb{R}^{p\times k}$ can be partitioned as

$$\mathbf{P}_j\mathbf{A} = \begin{bmatrix}\mathbf{A}^u\\ \mathbf{A}^o\end{bmatrix} \ ,$$

where $\mathbf{A}^u \in \mathbb{R}^{p_u\times k}$ and $\mathbf{A}^o \in \mathbb{R}^{p_o\times k}$. With these notations, we have the following theorem:

Theorem 3.13. Let $p_u \in \mathbb{N}_*$ and $p_o \in \mathbb{N}_*$ with $p = p_u + p_o$ design, respectively, the numbers of "unobserved" and "observed" entries in the j^{th} column of $\mathbf{X}$. If $p_u \geq k$, the j^{th} barrier set $\mathcal{B}_j$ is equal to the (nonempty) subset of $\mathbb{R}^{p.k}$

$$\mathcal{B}_j^* = \left\{\mathbf{a} \in \mathbb{R}^{p.k} \,/\, \exists \mathbf{A}^u \in \mathbb{R}_k^{p_u\times k} \text{ and } \mathbf{P}_j h(\mathbf{a}) = \begin{bmatrix}\mathbf{A}^u\\ \mathbf{0}^{p_o\times k}\end{bmatrix}\right\} ,$$

where $\mathbf{0}^{p_o\times k}$ is the zero $p_o \times k$ matrix and $\mathbf{P}_j$ is any $p \times p$ permutation matrix, which reorders the elements of $\mathbf{X}_{.j}$ so that the missing elements of $\mathbf{X}_{.j}$ appear first. On the other hand, if $p_u < k$ then $\mathcal{B}_j$ is empty.

Proof. Suppose first that $p_u \geq k$ and let $\mathbf{a} \in \mathcal{B}_j$. Then, it exists $h(\mathbf{a}) = \mathbf{A} \in \mathbb{R}_k^{p\times k}$ and we have the implications

$$\begin{aligned}
\mathbf{F}_j(\mathbf{a}) = \mathbf{0}^{p\times k} &\Rightarrow \mathit{diag}(\sqrt{\mathbf{W}}_{.j})\mathbf{A} = \mathbf{0}^{p\times k}\\
&\Rightarrow \mathit{diag}(\mathbf{P}_j^T\mathbf{P}_j\sqrt{\mathbf{W}}_{.j})\mathbf{P}_j^T\begin{bmatrix}\mathbf{A}^u\\ \mathbf{A}^o\end{bmatrix} = \mathbf{0}^{p\times k}\\
&\Rightarrow \mathbf{P}_j^T\mathit{diag}(\mathbf{P}_j\sqrt{\mathbf{W}}_{.j})\begin{bmatrix}\mathbf{A}^u\\ \mathbf{A}^o\end{bmatrix} = \mathbf{0}^{p\times k}\\
&\Rightarrow \mathit{diag}(\begin{bmatrix}\mathbf{0}^{p_u}\\ \sqrt{\mathbf{W}}_{.j}^o\end{bmatrix})\begin{bmatrix}\mathbf{A}^u\\ \mathbf{A}^o\end{bmatrix} = \mathbf{0}^{p\times k}\\
&\Rightarrow \mathit{diag}(\sqrt{\mathbf{W}}_{.j}^o)\mathbf{A}^o = \mathbf{0}^{p_o\times k}\\
&\Rightarrow \mathbf{A}^o = \mathbf{0}^{p_o\times k} \ .
\end{aligned}$$

The last implication results from the fact that all elements of the vector $\mathbf{W}_{.j}^o$ are different from zero by definition. This implies that $\mathbf{A} = \mathbf{P}_j^T\begin{bmatrix}\mathbf{A}^u\\ \mathbf{0}^{p_o\times k}\end{bmatrix}$. Furthermore, since $\mathbf{P}_j^T$ is a permutation matrix

and, thus, of full rank p, we have $k = \mathit{rank}(\mathbf{A}) = \mathit{rank}(\mathbf{A}^u)$ and $\mathbf{A}^u \in \mathbb{R}_k^{p_u \times k}$, which shows that $\mathbf{a} \in \mathcal{B}_j^*$.

Reciprocally, suppose that $\mathbf{a} \in \mathcal{B}_j^*$. Then, it exists $\mathbf{A} = h(\mathbf{a}) \in \mathbb{R}^{p \times k}$ and $\mathbf{A}^u \in \mathbb{R}_k^{p_u \times k}$ such that $\mathbf{A} = \mathbf{P}_j^T \begin{bmatrix} \mathbf{A}^u \\ \mathbf{0}^{p_o \times k} \end{bmatrix}$ and $\mathbf{A}$ is of rank k as $\mathbf{P}_j^T$ is of full rank p and $\mathbf{A}^u$ is of rank k. Then, we have

$$\begin{aligned}
\mathbf{F}_j(\mathbf{a}) &= \mathit{diag}(\sqrt{\mathbf{W}}_{.j})\mathbf{A} \\
&= \mathit{diag}(\sqrt{\mathbf{W}}_{.j})\mathbf{P}_j^T \begin{bmatrix} \mathbf{A}^u \\ \mathbf{0}^{p_o \times k} \end{bmatrix} \\
&= \mathit{diag}(\mathbf{P}_j^T \mathbf{P}_j \sqrt{\mathbf{W}}_{.j})\mathbf{P}_j^T \begin{bmatrix} \mathbf{A}^u \\ \mathbf{0}^{p_o \times k} \end{bmatrix} \\
&= \mathbf{P}_j^T \mathit{diag}(\mathbf{P}_j \sqrt{\mathbf{W}}_{.j}) \begin{bmatrix} \mathbf{A}^u \\ \mathbf{0}^{p_o \times k} \end{bmatrix} \\
&= \mathbf{P}_j^T \mathit{diag}(\begin{bmatrix} \mathbf{0}^{p_u} \\ \sqrt{\mathbf{W}}^o_{.j} \end{bmatrix}) \begin{bmatrix} \mathbf{A}^u \\ \mathbf{0}^{p_o \times k} \end{bmatrix} \\
&= \mathbf{P}_j^T \mathbf{0}^{p_o \times k} = \mathbf{0}^{p_o \times k} ,
\end{aligned}$$

and $\mathbf{a} \in \mathcal{B}_j$, which demonstrates the first part of the theorem.

Suppose now that $p_u < k$, then if $\mathbf{a} \in \mathbb{R}^{p.k}$ and $h(\mathbf{a}) = \mathbf{A} \in \mathbb{R}^{p \times k}$, we have,

$$\begin{aligned}
\mathbf{F}_j(\mathbf{a}) = \mathbf{0}^{p \times k} &\Rightarrow \mathit{diag}(\sqrt{\mathbf{W}}_{.j})\mathbf{A} = \mathbf{0}^{p \times k} \\
&\Rightarrow \mathit{diag}(\begin{bmatrix} \mathbf{0}^{p_u} \\ \sqrt{\mathbf{W}}^o_{.j} \end{bmatrix}) \begin{bmatrix} \mathbf{A}^u \\ \mathbf{A}^o \end{bmatrix} = \mathbf{0}^{p \times k} \\
&\Rightarrow \mathit{diag}(\sqrt{\mathbf{W}}^o_{.j})\mathbf{A}^o = \mathbf{0}^{p_o \times k} \\
&\Rightarrow \mathbf{A}^o = \mathbf{0}^{p_o \times k} .
\end{aligned}$$

This implies that $\mathbf{A} = \mathbf{P}_j^T \begin{bmatrix} \mathbf{A}^u \\ \mathbf{0}^{p_o \times k} \end{bmatrix}$ with $\mathbf{A}^u \in \mathbb{R}^{p_u \times k}$. But, as $\mathbf{P}_j^T$ is a permutation matrix and, thus, of full rank p, we have $\mathit{rank}(\mathbf{A}) = \mathit{rank}(\mathbf{A}^u) \leq \min(p_u, k) = p_u < k$ and we conclude that $\mathbf{a} \notin \mathcal{B}_j$. In other words, if $p_u < k$, for $\mathbf{a} \in \mathbb{R}^{p.k}$ we cannot have $\mathbf{F}_j(\mathbf{a}) = \mathbf{0}^{p \times k}$ and $h(\mathbf{a}) = \mathbf{A} \in \mathbb{R}_k^{p \times k}$.

□

Remark 3.10. Using Remark 3.9, it is easy to verify that the first part of Theorem 3.13 results from (i) the fact that $null(\mathbf{F}_j)$ is a linear subspace of $\mathbb{R}^{p.k}$ of dimension $p_u.k$ and is equal to

$$null(\mathbf{F}_j) = \left\{ \mathbf{a} \in \mathbb{R}^{p.k} \,/\, \exists \mathbf{A}^u \in \mathbb{R}^{p_u \times k} \text{ such that } \mathbf{P}_j h(\mathbf{a}) = \begin{bmatrix} \mathbf{A}^u \\ \mathbf{0}^{p_o \times k} \end{bmatrix} \right\} ,$$

and (ii) the property: Let $\mathbf{A} \in \mathbb{R}^{p \times k}$ and $\mathbf{A}^u \in \mathbb{R}^{p_u \times k}$ with $\mathbf{P}_j \mathbf{A} = \begin{bmatrix} \mathbf{A}^u \\ \mathbf{0}^{p_o \times k} \end{bmatrix}$. If $p_u \geq k$:

$$\mathit{rank}(\mathbf{A}) = k \Longleftrightarrow \mathit{rank}(\mathbf{A}^u) = k .$$

■

Theorem 3.14. With the same notations as in Theorem 3.13 and above, if $p_u \geq k$ and $p_o \geq 1$ (e.g., if there is at least one observed entry in the column vector $\mathbf{X}_{.j}$), the following two assertions are true:

1) The orthogonal projectors $\mathbf{P}_{\mathbf{F}_j(.)}$ and $\mathbf{P}^{\perp}_{\mathbf{F}_j(.)}$ are discontinuous at all points of $\mathcal{B}_j$,

2) If $\|\mathbf{X}^o_{.j}\|_2 \neq 0,$ then $\psi_j(.)$ is also discontinuous at all points of $\mathcal{B}_j$.

Proof. According to Theorem 3.10, in order to demonstrate the first assertion of the theorem it suffices to show that the matrix function $\mathbf{F}_j(.)$ has no local constant rank for all $\mathbf{a} \in \mathcal{B}_j$. We first note that $\mathbf{F}_j(\mathbf{a}) = \mathbf{0}^{p\times k}$ and, thus, *rank*$(\mathbf{F}_j(\mathbf{a})) = 0$ if $\mathbf{a} \in \mathcal{B}_j$. In these conditions, it suffices to show that, $\forall \alpha \in \mathbb{R}_{+*}$, if we consider the open ball $B_{p.k}(\mathbf{a}, \alpha)$, with center $\mathbf{a}$ and radius α, in $\mathbb{R}^{p\times k}$, it exists $\mathbf{d} \in B_{p.k}(\mathbf{a}, \alpha) \cap h^{-1}(\mathbb{R}_k^{p\times k})$ such that *rank*$(\mathbf{F}_j(\mathbf{d})) \neq 0$.

Since $\mathbf{a} \in \mathcal{B}_j$ and $p_u \geq k$ by hypothesis, according to Theorem 3.13, it exists $\mathbf{A}^u \in \mathbb{R}_k^{p_u\times k}$ such that $\mathbf{P}_j\mathbf{A} = \begin{bmatrix}\mathbf{A}^u \\ \mathbf{0}^{p_o\times k}\end{bmatrix}$ where $\mathbf{A} = h(\mathbf{a}) \in \mathbb{R}_k^{p\times k}$. Now, let $\beta \in \mathbb{R}_{+*}$ such that $\beta < \alpha$ and $\mathbf{D}^o \in \mathbb{R}^{p_o\times k}$ such that $\mathbf{D}^o_{11} = \beta$ and $\mathbf{D}^o_{ij} = 0$ if $i \neq 1$ and $j \neq 1$. If we define $\mathbf{D} = \mathbf{P}_j^T \begin{bmatrix}\mathbf{A}^u \\ \mathbf{D}^o\end{bmatrix}$ and $\mathbf{d} = vec(\mathbf{D}^T)$, we have $\mathbf{d} \in h^{-1}(\mathbb{R}_k^{p\times k})$ as $\mathbf{D} \in \mathbb{R}_k^{p\times k}$ (since $\mathbf{A}^u \in \mathbb{R}_k^{p_u\times k}$) and also

$$\|\mathbf{d} - \mathbf{a}\|_2^2 = \|\mathbf{D} - \mathbf{A}\|_F^2 = \beta^2 \leq \alpha^2 \ .$$

Thus, $\mathbf{d} \in B_{p.k}(\mathbf{a}, \alpha) \cap h^{-1}(\mathbb{R}_k^{p\times k})$. Obviously, $\mathbf{F}_j(\mathbf{d}) \neq \mathbf{0}^{p\times k}$ as $\mathbf{D}^o \neq \mathbf{0}^{p_o\times k}$ and *rank*$(\mathbf{F}_j(\mathbf{d})) = 1$ and we conclude that $\mathbf{F}_j(.)$ has no local constant rank at all points $\mathbf{a}$ of $\mathcal{B}_j$.

For demonstrating the second part of the theorem, we first remark that, for all points $\mathbf{a} \in \mathcal{B}_j$, we have $\mathbf{F}_j(\mathbf{a}) = \mathbf{0}^{p\times k}$ and $\mathbf{F}_j(\mathbf{a})^+ = \mathbf{0}^{k\times p}$ and, thus,

$$\psi_j(\mathbf{a}) = \|\mathbf{P}^{\perp}_{\mathbf{F}_j(\mathbf{a})}\mathbf{x}_j\|_2^2 = \|\mathbf{x}_j\|_2^2 = \|\mathbf{x}_j^u\|_2^2 + \|\mathbf{x}_j^o\|_2^2 = \|\mathbf{x}_j^o\|_2^2 \ ,$$

since $\mathbf{W}^u_{.j} = \mathbf{0}^{p_u}$. By hypothesis, we have $\|\mathbf{X}^o_{.j}\|_2 \neq 0$ and, thus, $\|\mathbf{x}_j^o\|_2^2 \neq 0$, so let us consider the open interval $\left]\frac{1}{2}\|\mathbf{x}_j^o\|_2^2, \frac{3}{2}\|\mathbf{x}_j^o\|_2^2\right[$ of $\mathbb{R}$, e.g., the open ball $B_1\big(\psi_j(\mathbf{a}), \frac{1}{2}\psi_j(\mathbf{a})\big)$ of $\mathbb{R}$. We have to show that, for all $\alpha \in \mathbb{R}_{+*}$, it exists $\mathbf{d} \in \mathbb{R}^{p.k}$ such that $\mathbf{d} \in B_{p.k}(\mathbf{a}, \alpha) \cap h^{-1}(\mathbb{R}_k^{p\times k})$, but $\psi_j(\mathbf{d}) \notin B_1\big(\psi_j(\mathbf{a}), \frac{1}{2}\psi_j(\mathbf{a})\big)$.

Let $\beta \in \mathbb{R}_{+*}$ such that $\beta < \alpha$, and define $\mathbf{D}^o \in \mathbb{R}^{p_o\times k}$ by $\mathbf{D}^o_{.1} = \beta.\frac{\mathbf{X}^o_{.j}}{\|\mathbf{X}^o_{.j}\|_2}$ and $\mathbf{D}^o_{.i} = 0^{p_o}$ for $i = 2$ to k, where 0^{p_o} is the zero vector of dimension p_o. Setting now

$$\mathbf{D} = \mathbf{P}_j^T \begin{bmatrix}\mathbf{A}^u \\ \mathbf{D}^o\end{bmatrix} \in \mathbb{R}^{p\times k}, \mathbf{d} = vec(\mathbf{D}^T) \in \mathbb{R}^{p.k} \text{ and } \mathbf{d}^o = vec\big((\mathbf{D}^o)^T\big) \in \mathbb{R}^{p_o.k} \ ,$$

we have $\|\mathbf{d} - \mathbf{a}\|_2^2 = \|\mathbf{D} - \mathbf{A}\|_F^2 = \beta^2 \leq \alpha^2$ and rank(**D**) = k. Thus, $\mathbf{d} \in B_{p.k}(\mathbf{a}, \alpha) \cap h^{-1}(\mathbb{R}_k^{p\times k})$ and

$$\begin{aligned}
\psi_j(\mathbf{d}) &= \|\mathbf{P}^{\perp}_{\mathbf{F}_j(\mathbf{d})}\mathbf{x}_j\|_2^2 \\
&= \|\mathbf{x}_j - \mathbf{F}_j(\mathbf{d})\mathbf{F}_j(\mathbf{d})^+\mathbf{x}_j\|_2^2 \\
&= \|\mathbf{x}_j^o - \mathbf{F}_j(\mathbf{d}^o)\mathbf{F}_j(\mathbf{d}^o)^+\mathbf{x}_j^o\|_2^2 \\
&= \min_{\mathbf{b}\in\mathbb{R}^k} \|\mathbf{x}_j^o - \mathbf{F}_j(\mathbf{d}^o)\mathbf{b}\|_2^2 \\
&= \min_{\mathbf{b}_1\in\mathbb{R}} \|\mathbf{x}_j^o - (\sqrt{\mathbf{W}}^o_{.j} \odot \mathbf{D}^o_{.1})\mathbf{b}_1\|_2^2 \\
&= \min_{\mathbf{b}_1\in\mathbb{R}} \|\mathbf{x}_j^o - \frac{\beta}{\|\mathbf{X}^o_{.j}\|_2}(\sqrt{\mathbf{W}}^o_{.j} \odot \mathbf{X}^o_{.j})\mathbf{b}_1\|_2^2 \\
&= \min_{\mathbf{b}_1\in\mathbb{R}} \|\mathbf{x}_j^o - \frac{\beta}{\|\mathbf{X}^o_{.j}\|_2}\mathbf{x}_j^o\mathbf{b}_1\|_2^2 \\
&= 0 \text{ with } \mathbf{b}_1 = \frac{\|\mathbf{X}^o_{.j}\|_2}{\beta} \ .
\end{aligned}$$

In other words, $\psi_j(\mathbf{d}) \notin \left]\frac{1}{2}\|\mathbf{x}_j^o\|_2^2, \frac{3}{2}\|\mathbf{x}_j^o\|_2^2\right[= B_1\big(\psi_j(\mathbf{a}), \frac{1}{2}\psi_j(\mathbf{a})\big)$, which demonstrates that $\psi_j(.)$ is discontinuous at all points of $\mathcal{B}_j$ as claimed in the second part of the theorem.

□

Corollary 3.4. With the same definitions and notations as in Theorem 3.14, the orthogonal projectors $\mathbf{P}_{\mathbf{F}(.)}$ and $\mathbf{P}^{\perp}_{\mathbf{F}(.)}$ are discontinuous at all points $\mathbf{a} \in \bigcup_{j=1}^{n} \mathcal{B}_j$.

Proof. This results immediately from Theorem 3.14 and the equality

$$\mathbf{P}_{\mathbf{F}(\mathbf{a})} = \bigoplus_{j=1}^{n} \mathbf{P}_{\mathbf{F}_j(\mathbf{a})},$$

which shows that the continuity of $\mathbf{P}_{\mathbf{F}(.)}$ is equivalent to the continuity of the n atomic orthogonal projectors $\mathbf{P}_{\mathbf{F}_j(.)}$, for $j = 1$ to n.

□

Since $\mathbf{F}_j(.)$ is a continuous linear mapping different from the zero constant function, its null space, $null(\mathbf{F}_j) = \mathbf{F}_j^{-1}(\{\mathbf{0}^{p\times k}\})$, is closed and not dense in $\mathbb{R}^{p.k}$ and its complement $\mathbb{R}^{p.k}/null(\mathbf{F}_j)$ (in $\mathbb{R}^{p.k}$), is nonempty and open. Next, using Theorem 3.14 and Corollary 3.4, we know that $\mathbf{P}^{\perp}_{\mathbf{F}(.)}$ is not continuous at all points $\mathbf{a} \in \bigcup_{j=1}^{n} \mathcal{B}_j$, implying that $\mathbf{P}^{\perp}_{\mathbf{F}(.)}$ will not be differentiable and $\psi(.)$ will also, in general, not be continuous and differentiable at these points. This implies that the "feasible" search set for a solution of the (VP1) problem will be severely restricted in the case of missing values, at least by standard first- and second-order optimization methods, which require that the objective function must be at least differentiable. Moreover, the "size" of this "feasible" search set will also decrease if the number of missing values increases as it is equal to $\bigcap_{j=1}^{n}[h^{-1}(\mathbb{R}_k^{p\times k})/\mathcal{B}_j)]$.

Hence, when the number of missing values in the data matrix $\mathbf{X}$ is very large, one may prefer an alternative formulation of the WLRA problem that will allow for a continuous and differentiable objective function $\psi(.)$ for all points of $h^{-1}(\mathbb{R}_k^{p\times k})$, no barrier sets $\mathcal{B}_j$ and no discontinuities for any of the atomic functions $\psi_j(.)$. In this context, when the weights are equal to one or zero (e.g., the missing value problem), Dai et al. [47] have proposed to replace the traditional Frobenius metric by what they called a "geometric performance metric" to avoid these discontinuities of the atomic functions when solving the matrix completion problem in a Grassmann manifold's setting. More generally, as discussed above and demonstrated in Theorem 3.8, the minimization of the separable form of the cost function $g_\lambda(.)$ defined in equation (3.18) with a small regularization parameter $\lambda \in \mathbb{R}_{+*}$ or continuation Tikhonov methods based on a family of such cost functions $g_\lambda(.)$ in which λ tends to zero during the iterations are also promising alternatives in such difficult situation. Moreover, these alternatives work with nonuniform weight matrices including zero weights and not only for the missing value problem with binary weights.

We are now set to describe the different algorithms which may be used to minimize $\varphi^*(.)$ or $\psi(.)$. We start by a modern description of several ALS regression methods, which all originate from the NIPALS algorithm first introduced by Wold and his collaborators [191][192][93], and alternate between minimization of the two sets of variables, $\mathbf{A}$ and $\mathbf{B}$, for solving the formulation (P1) of the WLRA problem in Section 4. The more complicated separable NLLS algorithms (e.g., first- and second-order variable projection methods), which explicitly eliminate the linear parameters (for example $\mathbf{b} = vec(\mathbf{B})$) obtaining a reduced, but somewhat more complicated, functional $\psi(.)$ that involves only the nonlinear parameters (e.g., $\mathbf{a} = vec(\mathbf{A}^T)$), are described in Section 5.

4 The block alternating least-squares method and its variants

As noted in Subsection 3.4, if we fix $\mathbf{a} = h^{-1}(\mathbf{A}) = vec(\mathbf{A}^T)$, then the problem

$$\min_{\mathbf{b}\in\mathbb{R}^{k.n}} \quad \frac{1}{2}\|\mathbf{x} - \mathbf{F}(\mathbf{a})\mathbf{b}\|_2^2 = \varphi^*(\mathbf{A}, \mathbf{B}) \,,$$

where $\mathbf{F}(\mathbf{a})$ and $\mathbf{x}$ are also defined in Subsection 3.4, is a linear least-squares problem with $k.n$ unknowns $\mathbf{B}_{ij}$. The unique minimum 2-norm solution of this linear least-squares problem for a

fixed $\mathbf{A}$ matrix is $\widehat{\mathbf{b}} = \mathbf{F}(\mathbf{a})^{+}\mathbf{x}$ as stated in Subsection 2.1. More precisely, if we take into account the block structure of $\mathbf{F}(\mathbf{a})$, we observe that the best choice of $\mathbf{b} = vec(\mathbf{B})$ for a given $\mathbf{A}$ matrix is obtained by solving n independent linear least-squares problems, each with k unknowns, and $\widehat{\mathbf{B}}_{.j}$, for $j = 1, \cdots, n$, can be calculated by

$$\widehat{\mathbf{B}}_{.j} = \left(diag(\sqrt{\mathbf{W}}_{.j})\mathbf{A}\right)^{+} (\sqrt{\mathbf{W}}_{.j} \odot \mathbf{X}_{.j}) = \mathbf{F}_j(\mathbf{a})^{+}\mathbf{x}_j \ , \tag{4.1}$$

where $\mathbf{F}_j(\mathbf{a}) = diag(\sqrt{\mathbf{W}}_{.j})\mathbf{A}$ and $\mathbf{x}_j = \sqrt{\mathbf{W}}_{.j} \odot \mathbf{X}_{.j}$. Likewise, if $\mathbf{b} = vec(\mathbf{B})$ is fixed, the minimization problem

$$\min_{\mathbf{a}\in\mathbb{R}^{p.k}} \quad \frac{1}{2}\|\mathbf{z} - \mathbf{G}(\mathbf{b})\mathbf{a}\|_2^2 = \varphi^*(\mathbf{A}, \mathbf{B}) \ ,$$

where $\mathbf{G}(\mathbf{b})$ and $\mathbf{z}$ are again defined in Subsection 3.4, is a linear least-squares problem with $k.p$ unknowns $\mathbf{A}_{ij}$. The unique minimum 2-norm solution of this linear least-squares problem for a fixed $\mathbf{B}$ matrix is then $\widehat{\mathbf{a}} = \mathbf{G}(\mathbf{b})^{+}\mathbf{z}$. Again, by taking into account the block structure of $\mathbf{G}(\mathbf{b})$, we observe that the best choice of $\mathbf{A}$ for a given $\mathbf{B}$ matrix is obtained by solving p independent linear least-squares problems, each with k unknowns, and $\widehat{\mathbf{A}}_{i.}$, for $i = 1, \cdots, p$, can be calculated by

$$\widehat{\mathbf{A}}_{i.} = (\sqrt{\mathbf{W}}_{i.} \odot \mathbf{X}_{i.}) \left(\mathbf{B}diag(\sqrt{\mathbf{W}}_{i.})\right)^{+} = \left(\mathbf{G}_i(\mathbf{b})^{+}\mathbf{z}_i\right)^T \ , \tag{4.2}$$

where $\mathbf{G}_i(\mathbf{b}) = diag(\sqrt{\mathbf{W}}_{i.})\mathbf{B}^T$ and $\mathbf{z}_i = \left(\sqrt{\mathbf{W}}_{i.} \odot \mathbf{X}_{i.}\right)^T$.

The above results suggest that we can minimize the cost function $\varphi^*(.)$ and solve the WLRA problem in its formulation (P1) by taking the block separable least-squares approach (e.g., NIPALS algorithm) of Wold and his collaborators [191][192][93], a method rediscovered and studied many times after, particularly in the context of low-rank matrix completion and optimization problems [188][94][76][118][146] or in the computer vision community [176][15][37]. The idea is to minimize $\varphi^*(.)$ by alternatively improving the $\mathbf{A}$ and $\mathbf{B}$ matrices through a sequence of cyclic linear least-squares optimizations. One starts with some initial guess, say, $\mathbf{A}^0$ and iterates from $\mathbf{A}^0$ to $\mathbf{B}^0$ then from $\mathbf{B}^0$ to $\mathbf{A}^1$, etc ... This method of iterations yields a decreasing sequence of functions values $\{\varphi^*(\mathbf{A}^i, \mathbf{B}^i)\}_{i\in\mathbb{N}}$ as the sandwich inequality

$$\varphi^*(\mathbf{A}^i, \mathbf{B}^i) \geq \varphi^*(\mathbf{A}^i, \mathbf{B}^{i+1}) \geq \varphi^*(\mathbf{A}^{i+1}, \mathbf{B}^{i+1})$$

holds for all $(\mathbf{A}^i, \mathbf{B}^i)$, $i \in \mathbb{N}$. Since the continuous real-valued function $\varphi^*(.)$ is bounded below by zero, the sequence of function values should converge to an infimum. However, the convergence can be quite slow, especially in the presence of missing values [15], and the sequence of points $\{(\mathbf{A}^i, \mathbf{B}^i)\}_{i\in\mathbb{N}}$ may even cycle, stagnate and not converge to a stationary point of the WLRA problem [69] as this block separable least-squares approach is a simple instance of a block coordinate descent method (also known as the block-nonlinear Gauss-Seidel method, see [148][139]) for minimizing $\varphi^*(.)$ and the cost function is not convex [152]. Importantly, subsequence convergence to a stationary point can still be obtained for this block coordinate descent algorithm applied to a non-convex function (as $\varphi^*(.)$) for special cases such as the existence of an unique minimizer per block of variables [108][7] or in the case of two block of variables with a nonempty set of minimizers per block, but without the unicity condition [68], as stated in the following theorem and corollary:

Theorem 4.1. Suppose that $f(.)$ is a continuously differentiable function over a set $\Omega \subset \mathbb{R}^l$, which is a Cartesian product of closed convex sets $\Omega_1, \Omega_2, ..., \Omega_m$, where $\Omega_i \subset \mathbb{R}^{n_i}$ for $i = 1, ..., m$ and $\sum_{i=1}^m n_i = l$. Suppose that the variable $\mathbf{x}$ is also partitioned accordingly as $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_m)$ where $\mathbf{x}_i \in \Omega_i$. Furthermore, suppose that for each i and $\mathbf{x} \in \Omega$, the solution of

$$\min_{\zeta\in\Omega_i} \quad f(\mathbf{x}_1, ..., \mathbf{x}_{i-1}, \zeta, \mathbf{x}_{i+1}, ..., \mathbf{x}_m)$$

is uniquely attained. If $f(.)$ is minimized by a block coordinate descent algorithm, in which a single block of variables, $\mathbf{x}_i$, is optimized while the remaining variables are held fixed at each iteration,

then any accumulation point of the sequence of points, $\{\mathbf{x}^k\}_{k\in\mathbb{N}}$ generated by this block coordinate descent algorithm is also a first-order stationary point of $f(.)$.

Proof. Omitted. See p.195 in [152] or Proposition 2.7.1 in [7] for details. □

Corollary 4.1. In the same conditions as in Theorem 4.1, if $f(.)$ is defined only over a Cartesian product of two closed convex sets, Ω_1 and Ω_2, and the global minimization of $f(.)$ with respect to each component is well defined, but not necessarily unique, then any accumulation point of the sequence of points, $\{(\mathbf{x}^k, \mathbf{y}^k)\}_{k\in\mathbb{N}}$ generated by this two-block coordinate descent algorithm is also a first-order stationary point of $f(.)$.

Proof. Omitted. See Theorem 6.3 in [68]. □

As $\mathbb{R}^{p\times k}$ and $\mathbb{R}^{k\times n}$ are closed convex sets and $\varphi^*(.)$ is continuous as stated in Theorem 3.2 and also continuously differentiable (as it is a polynomial in $(p\times k)+(k\times n)$ variables), Corollary 4.1 can be applied to the two-block separable least-squares approach described above. Note, on the other hand, that Theorem 4.1 cannot be used here because we cannot assume that all the regression problems for computing $\mathbf{A}$ and $\mathbf{B}$ cyclically in the two-block separable least-squares method to minimize $\varphi^*(.)$ can be solved uniquely in general. For example, this will not be the case if some rows or columns of $\mathbf{X}$ have less than k "observed" values. However, using Corollary 4.1, we still obtain that any accumulation point of the sequence $\{(\mathbf{A}^i, \mathbf{B}^i)\}_{i\in\mathbb{N}}$, say $(\widehat{\mathbf{A}}, \widehat{\mathbf{B}})$, is a first-order stationary point of $\varphi^*(.)$ and, thus, satisfies

$$\frac{\partial\varphi^*(\widehat{\mathbf{A}}, \widehat{\mathbf{B}})}{\partial\mathbf{a}} = \nabla\varphi^*_{\mathbf{a}}(\widehat{\mathbf{A}}, \widehat{\mathbf{B}}) = \mathbf{0}^{k.p} \quad \text{and} \quad \frac{\partial\varphi^*(\widehat{\mathbf{A}}, \widehat{\mathbf{B}})}{\partial\mathbf{b}} = \nabla\varphi^*_{\mathbf{b}}(\widehat{\mathbf{A}}, \widehat{\mathbf{B}}) = \mathbf{0}^{k.n} ,$$

where the partial functions $\varphi^*_{\mathbf{a}}(.)$ and $\varphi^*_{\mathbf{b}}(.)$ are defined by

$$\varphi^*_{\mathbf{a}} : \mathbb{R}^{p.k} \longrightarrow \mathbb{R} : \mathbf{c} \mapsto \varphi^*(mat_{k\times p}(\mathbf{c})^T, \mathbf{B}) = \frac{1}{2}\|\sqrt{\mathbf{W}} \odot (\mathbf{X} - \mathbf{CB})\|^2_F ,$$
$$\varphi^*_{\mathbf{b}} : \mathbb{R}^{k.n} \longrightarrow \mathbb{R} : \mathbf{d} \mapsto \varphi^*(\mathbf{A}, mat_{k\times n}(\mathbf{d})) = \frac{1}{2}\|\sqrt{\mathbf{W}} \odot (\mathbf{X} - \mathbf{AD})\|^2_F ,$$

see Theorem 4.3 below for details. Note that this condition is however not sufficient to ensure that $\widehat{\mathbf{Y}} = \widehat{\mathbf{A}}\widehat{\mathbf{B}} \in \mathbb{R}^{p\times n}_{\leq k}$ is a Frechet first-order stationary point of $\varphi(.)$ according to Theorem 3.7 if $\widehat{\mathbf{Y}} \in \mathbb{R}^{p\times n}_{<k}$. However, from Theorem 4.3 below, we also get that the (partial) Hessian matrices of the vectorized form of $\varphi^*(.)$ are equal to

$$\frac{\partial^2\varphi^*(\mathbf{A}, \mathbf{B})}{\partial^2\mathbf{a}} = \nabla^2\varphi^*_{\mathbf{a}}(\mathbf{A}, \mathbf{B}) = \mathbf{G}(\mathbf{b})^T\mathbf{G}(\mathbf{b}) ,$$
$$\frac{\partial^2\varphi^*(\mathbf{A}, \mathbf{B})}{\partial^2\mathbf{b}} = \nabla^2\varphi^*_{\mathbf{b}}(\mathbf{A}, \mathbf{B}) = \mathbf{F}(\mathbf{a})^T\mathbf{F}(\mathbf{a}) ,$$

and are, thus, positive semi-definite for all $\mathbf{A} \in \mathbb{R}^{p\times k}$ and $\mathbf{B} \in \mathbb{R}^{k\times n}$, which implies that $\varphi^*(.)$ is bi-convex in its whole domain. In addition, if the block matrices $\mathbf{F}(\widehat{\mathbf{a}})$ and $\mathbf{G}(\widehat{\mathbf{b}})$ are of full column-rank, which will be the rule rather than the exception if there are at least k "observed" values in each column and row of the incomplete data matrix $\mathbf{X}$, the (partial) Hessian matrices $\frac{\partial^2\varphi^*(\widehat{\mathbf{A}},\widehat{\mathbf{B}})}{\partial^2\mathbf{a}}$ and $\frac{\partial^2\varphi^*(\widehat{\mathbf{A}},\widehat{\mathbf{B}})}{\partial^2\mathbf{b}}$ will be further positive definite and this implies that

$$\widehat{\mathbf{A}} = \text{Arg}\min_{\mathbf{A}\in\mathbb{R}^{p\times k}} \varphi^*(\mathbf{A}, \widehat{\mathbf{B}}) \quad , \quad \widehat{\mathbf{B}} = \text{Arg}\min_{\mathbf{B}\in\mathbb{R}^{k\times n}} \varphi^*(\widehat{\mathbf{A}}, \mathbf{B}) ,$$

are strict local minima for the partial functions $\varphi^*(., \widehat{\mathbf{B}})$ and $\varphi^*(\widehat{\mathbf{A}}, .)$, respectively, and

$$\varphi^*(\widehat{\mathbf{A}}, \widehat{\mathbf{B}}) = \min_{\mathbf{A}\in\mathbb{R}^{p\times k}} \varphi^*(\mathbf{A}, \widehat{\mathbf{B}}) = \min_{\mathbf{B}\in\mathbb{R}^{k\times n}} \varphi^*(\widehat{\mathbf{A}}, \mathbf{B}) ,$$

which is a necessary, though not sufficient condition, for the pair $(\widehat{\mathbf{A}}, \widehat{\mathbf{B}})$'s being a minimum point of $\varphi^*(.)$. Finally, the resulting algorithm is globally convergent with a sublinear or linear convergence rate at best [166][15][21]. However, little can be said in general about the convergence behaviour of the sequence $\{(\mathbf{A}^i, \mathbf{B}^i)\}_{i\in\mathbb{N}}$ without additional assumptions or modifications (e.g., regularizations) of the cost function of the WLRA problem as we will discuss now in some details.

If we use vectorized matrix variables, e.g., $\mathbf{a} = vec(\mathbf{A}^T)$ and $\mathbf{b} = vec(\mathbf{B})$, then the iterations in the block ALS algorithm (e.g., NIPALS) take the following form

$$\begin{aligned}\mathbf{a}^{i+1} &= \mathbf{G}(\mathbf{b}^i)^+\mathbf{z} = \omega(\mathbf{b}^i)\ ,\\ \mathbf{b}^{i+1} &= \mathbf{F}(\mathbf{a}^{i+1})^+\mathbf{x} = \upsilon(\mathbf{a}^{i+1})\ ,\end{aligned}$$

for $i = 0, 1, 2, \ldots$, where $\upsilon(.)$ and $\omega(.)$ are two real-vector functions from $\mathbb{R}^{p.k}$ to $\mathbb{R}^{k.n}$ and from $\mathbb{R}^{k.n}$ to $\mathbb{R}^{p.k}$, respectively, defined by

$$\upsilon(\mathbf{a}) = \begin{cases}\text{Arg}\min_{\mathbf{b}\in\mathbb{R}^{k.n}}\ \frac{1}{2}\|\mathbf{x} - \mathbf{F}(\mathbf{a})\mathbf{b}\|_2^2 = \varphi^*(\mathbf{A}, \mathbf{B})\\ \text{s.t. Arg}\min_{\mathbf{b}\in\mathbb{R}^{k.n}}\ \|\mathbf{b}\|_2\end{cases}$$

$$\omega(\mathbf{b}) = \begin{cases}\text{Arg}\min_{\mathbf{a}\in\mathbb{R}^{p.k}}\ \frac{1}{2}\|\mathbf{z} - \mathbf{G}(\mathbf{b})\mathbf{a}\|_2^2 = \varphi^*(\mathbf{A}, \mathbf{B})\\ \text{s.t. Arg}\min_{\mathbf{a}\in\mathbb{R}^{p.k}}\ \|\mathbf{a}\|_2\end{cases}$$

That is, either subproblem has an unique minimum 2-norm minimizer (see Subsection 2.1) and the functions $\upsilon(.)$ and $\omega(.)$ are thus well-defined. In these conditions, the composition map $\chi(.) = \omega(.) \circ \upsilon(.)$ is also a well-defined function from $\mathbb{R}^{p.k}$ to $\mathbb{R}^{p.k}$ and the ALS algorithm takes the form of a standard fixed point iteration [148]

$$\mathbf{a}^{i+1} = \chi(\mathbf{a}^i) = \chi^i(\mathbf{a}^0) \text{ for } i = 0, 1, 2, \ldots$$

Clearly, if $\upsilon(.)$ and $\omega(.)$ are continuous, $\chi(.)$ is also continuous and if, in addition,

$$\lim_{i\to\infty} \mathbf{a}^i = \widehat{\mathbf{a}}\ ,$$

then $\widehat{\mathbf{a}}$ solves the system $\mathbf{a} = \chi(\mathbf{a})$, (e.g., $\widehat{\mathbf{a}}$ is a fixed point of $\chi(.)$) and

$$\lim_{i\to\infty} \left(\mathbf{a}^i, \upsilon(\mathbf{a}^i)\right) = \lim_{i\to\infty} \left(\mathbf{a}^i, \mathbf{b}^i\right) = (\widehat{\mathbf{a}}, \widehat{\mathbf{b}}) \text{ with } \widehat{\mathbf{b}} = \upsilon(\widehat{\mathbf{a}}) \text{ and } \widehat{\mathbf{a}} = \omega(\widehat{\mathbf{b}})\ .$$

Then, under these hypotheses, the ALS iterations converge to $(\widehat{\mathbf{a}}, \widehat{\mathbf{b}})$ and we have

$$\widehat{\mathbf{a}} = \text{Arg}\min_{\mathbf{a}\in\mathbb{R}^{p.k}} \frac{1}{2}\|\mathbf{z} - \mathbf{G}(\widehat{\mathbf{b}})\mathbf{a}\|_2^2 \quad \text{and} \quad \widehat{\mathbf{b}} = \text{Arg}\min_{\mathbf{b}\in\mathbb{R}^{k.n}} \frac{1}{2}\|\mathbf{x} - \mathbf{F}(\widehat{\mathbf{a}})\mathbf{b}\|_2^2\ ,$$

which implies

$$\nabla\varphi^*_{\mathbf{a}}(\widehat{\mathbf{A}}, \widehat{\mathbf{B}}) = \frac{\partial\varphi^*(\widehat{\mathbf{A}}, \widehat{\mathbf{B}})}{\partial\mathbf{a}} = \mathbf{0}^{k.p} \quad , \quad \nabla\varphi^*_{\mathbf{b}}(\widehat{\mathbf{A}}, \widehat{\mathbf{B}}) = \frac{\partial\varphi^*(\widehat{\mathbf{A}}, \widehat{\mathbf{B}})}{\partial\mathbf{b}} = \mathbf{0}^{k.n}\ ,$$

and $(\widehat{\mathbf{A}}, \widehat{\mathbf{B}})$ is a first-order stationary point of $\varphi^*(.)$. Thus, in these conditions, we have a one-to-one correspondence between the fixed points of $\chi(.)$ and the first-order stationary points of $\varphi^*(.)$. However, the hypotheses that $\upsilon(.)$ and $\omega(.)$ are continuous cannot be proved here as the generalized inverse functions $\mathbf{F}(.)^+$ and $\mathbf{G}(.)^+$ are clearly not continuous on all points of $\mathbb{R}^{p.k}$ and $\mathbb{R}^{k.n}$, respectively, according to Theorems 3.10, 3.11 and 3.12. Consequently, this approach cannot be used to establish the general convergence of the whole sequence $\{(\mathbf{A}^i, \mathbf{B}^i)\}_{i\in\mathbb{N}}$. Similarly, $\chi(.)$ is not a contraction in any open ball $B_{p.k}(\mathbf{a}^0, r)$ of radius r around the starting point $\mathbf{a}^0$ as otherwise the Contraction Mapping Theorem [148] will imply that the equation $\mathbf{a} = \chi(\mathbf{a})$ has an unique solution $\widehat{\mathbf{a}}$ in the closed ball $\bar{B}_{p.k}(\mathbf{a}^0, r)$, which is false according to Remark 3.4 and the

over-parameterization of the formulation (P1) of the WLRA problem. In other words, the convergence of the whole sequence $\{(\mathbf{A}^i, \mathbf{B}^i)\}_{i\in\mathbb{N}}$ cannot be proved either with the help of the Contraction Mapping Theorem.

First, we observe that more precise and stronger results can be derived when all the weights are strictly positive, e.g., when $\mathbf{W} \in \mathbb{R}^{p\times n}_{+*}$, because the cost function $\varphi(.)$ is λ-smooth in that case, which means that the gradient mapping $\nabla\varphi(.)$ from $\mathbb{R}^{p\times n}$ to $\mathbb{R}^{p\times n}$ is Lipschitz continuous with a Lipschitz constant $\lambda > 0$, e.g.,

$$\|\nabla\varphi(\mathbf{Y}) - \nabla\varphi(\mathbf{Z})\|_F \leq \lambda\|\mathbf{Y} - \mathbf{Z}\|_F \ , \ \forall \mathbf{Y}, \mathbf{Z} \in \mathbb{R}^{p\times n} \ .$$

Using equation (3.3) in Subsection 3.2, we get immediately

$$\|\nabla\varphi(\mathbf{Y}) - \nabla\varphi(\mathbf{Z})\|_F = \|\mathbf{W} \odot (\mathbf{Y} - \mathbf{Z})\|_F \leq \lambda\|\mathbf{Y} - \mathbf{Z}\|_F \ , \ \forall \mathbf{Y}, \mathbf{Z} \in \mathbb{R}^{p\times n} \ ,$$

with $\lambda = \max_{(i,j)\in[p]\times[n]} \mathbf{W}_{ij}$, implying that $\varphi(.)$ is effectively λ-smooth when $\mathbf{W} \in \mathbb{R}^{p\times n}_{+*}$. As $\varphi(.)$ is also bounded from below, e.g., $\forall \mathbf{Y} \in \mathbb{R}^{p\times n}, \varphi(\mathbf{Y}) \geq 0$, we have the following result, which is a direct application of Corollary 3.9 in Olikier et al. [146].

Theorem 4.2. Let $\mathbf{Y}^i = \mathbf{A}^i\mathbf{B}^i \in \mathbb{R}^{p\times n}_{\leq k}$, $\forall i \in \mathbb{N}$, where the sequence $\{(\mathbf{A}^i, \mathbf{B}^i)\}_{i\in\mathbb{N}}$ is the iterates of the block ALS algorithm defined by equations (4.1) and (4.2).

Then, the generated sequence $\{\varphi(\mathbf{Y}^i)\}_{i\in\mathbb{N}} = \{\varphi^*(\mathbf{A}^i, \mathbf{B}^i)\}_{i\in\mathbb{N}}$ of cost function values is monotonically decreasing and converges to some value $\varphi_* \geq \bar{\mathbf{c}}_{\varphi^*} = \bar{\mathbf{c}}_{\varphi}$, where $\bar{\mathbf{c}}_{\varphi^*} = \bar{\mathbf{c}}_{\varphi}$ is the infimum of $\varphi(.)$ on $\mathbb{R}^{p\times n}_{\leq k}$, which is equal to the infimum of $\varphi^*(.)$ on $\mathbb{R}^{p\times k} \times \mathbb{R}^{k\times n}$ (see Theorem 3.1 for details). Moreover, the Riemannian gradient of $\varphi(.)$ at $\mathbf{Y}^i$ tends to zero, e.g.,

$$\lim_{i\to\infty} \nabla_R\varphi(\mathbf{Y}^i) = \mathbf{P}_{\mathcal{T}_{\mathbf{Y}^i}\mathbb{R}^{p\times n}_{rank(\mathbf{Y}^i)}}\big(\nabla\varphi(\mathbf{Y}^i)\big) = \mathbf{0}^{p\times n}$$

and every point of accumulation $\widehat{\mathbf{Y}}$ of the sequence $\{\mathbf{Y}^i\}_{i\in\mathbb{N}}$ satisfies $\varphi(\widehat{\mathbf{Y}}) = \varphi_*$ and $\nabla_R\varphi(\widehat{\mathbf{Y}}) = \mathbf{0}^{p\times n}$, which means that $\widehat{\mathbf{Y}}$ is a Riemannian first-order stationarity point of $\varphi(.)$ on the smooth manifold $\mathbb{R}^{p\times n}_{k'}$ where $k' = rank(\widehat{\mathbf{Y}}) \leq k$. In particular, if $rank(\widehat{\mathbf{Y}}) = k$ then $\widehat{\mathbf{Y}}$ is also a Frechet first-order stationarity point of $\varphi(.)$ on $\mathbb{R}^{p\times n}_{\leq k}$ in the sense of Theorem 3.5 and the pair $(\widehat{\mathbf{A}}, \widehat{\mathbf{B}})$ is a first-order critical point of $\varphi^*(.)$ on $\mathbb{R}^{p\times k} \times \mathbb{R}^{k\times n}$ according to Theorem 3.7.

Furthermore, $\forall j \in \mathbb{N}$, it holds that

$$\min_{0\leq i\leq j} \|\nabla_R\varphi(\mathbf{Y}^i)\|_F \leq \Big(2.\lambda.\frac{\varphi(\mathbf{Y}^0) - \varphi_*}{2.j+1}\Big)^{\frac{1}{2}} \ .$$

In particular, given $\varepsilon > 0$, the algorithm returns a matrix satisfying $\|\nabla_R\varphi(\mathbf{Y}^i)\|_F \leq \varepsilon$ after at most $\left\lceil \lambda.\frac{\varphi(\mathbf{Y}^0)-\varphi_*}{\varepsilon^2} - \frac{1}{2} \right\rceil$ iterations.

Proof. Omitted. See Corollary 3.9 of Olikier et al. [146] for details. □

Interestingly, this theorem also illustrated the impact of a "good" initialization of the block ALS algorithm on the required number iterations for the convergence of the sequence in terms of the norm of the Riemannian gradient of $\varphi(.)$. However, even in the case where $\mathbf{W} \in \mathbb{R}^{p\times n}_{+*}$, we cannot ensure that $\mathbf{Y}^{i+1} - \mathbf{Y}^i \longrightarrow \mathbf{0}^{p\times n}$ and this result requires additional modifications of the algorithm or hypotheses.

In fact, many of the past works study when the ALS minimization algorithm converges to its infimum for the matrix completion problem in polynomial time under the additional assumptions that (i) there is a solution $\widehat{\mathbf{X}} = \widehat{\mathbf{A}}\widehat{\mathbf{B}}$, which is incoherent (e.g., the squared row norms of $\widehat{\mathbf{A}}$ and squared column norms of $\widehat{\mathbf{B}}$ are not small) and (ii) the non-missing entries of $\mathbf{X}$ are selected uniformly at random or have pseudorandom properties [94][76]. More precisely, these two studies

have shown that with an appropriate SVD-based initialization, the ALS algorithm (with a few modifications) recovers the ground-truth in the case of random binary weights and under a resampling scheme. Convergence results with a relaxation of the random sampling hypothesis can be found in [16][114][170]. However, all these past studies concern mainly the matrix completion problem with a binary weight matrix [76][94][16][170] or assume that there are no zero weights and that $\mathbf{W}$ is spectrally closed to the all one matrix in the case of a nonuniform weight matrix [114]. Finally, there is some ongoing debate as to whether these different assumptions are valid for real-world datasets [174]. Interestingly, the incoherency hypothesis of the solution pair $(\widehat{\mathbf{A}}, \widehat{\mathbf{B}})$ stated above means that $\widehat{\mathbf{A}}$ is far away from any of the barrier sets $\mathcal{B}_j$, defined in the previous subsection (see Definition 3.2), illustrating how the variable projection framework shed also some lights on the solvability of the WLRA problem by other methods such as the block ALS algorithm described above.

The block ALS method can also be adapted to solve the MMMF formulation of the WLRA problem equipped with a regularization parameter $\lambda \in \mathbb{R}_{+*}$ already discussed in Subsection 3.3 (see equation (MMMF)), since

$$\begin{aligned}\min_{\mathbf{A}\in\mathbb{R}^{p\times k},\, \mathbf{B}\in\mathbb{R}^{k\times n}} \quad \varphi^*_\lambda(\mathbf{A},\mathbf{B}) &= \frac{1}{2}\|\sqrt{\mathbf{W}} \odot (\mathbf{X}-\mathbf{A}\mathbf{B})\|_F^2 + \frac{\lambda}{2}(\|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2)\\ &= \frac{1}{2}\big\| \left[\mathbf{z} - \mathbf{G}(\mathbf{b})\mathbf{a}\right] \big\|_2^2 + \frac{\lambda}{2}\|\mathbf{a}\|_2^2 + \frac{\lambda}{2}\|\mathbf{b}\|_2^2\\ &= \frac{1}{2}\big\| \left[\mathbf{x} - \mathbf{F}(\mathbf{a})\mathbf{b}\right] \big\|_2^2 + \frac{\lambda}{2}\|\mathbf{a}\|_2^2 + \frac{\lambda}{2}\|\mathbf{b}\|_2^2\,,\end{aligned}$$

where $\mathbf{a} = vec(\mathbf{A}^T)$, $\mathbf{b} = vec(\mathbf{B})$, $\mathbf{x}, \mathbf{z}$, $\mathbf{F}(\mathbf{a})$ and $\mathbf{G}(\mathbf{b})$ are defined as above.

In this case, the block ALS algorithm computes alternatively the solutions of the two regularized least-squares problems

$$\text{Arg}\min_{\mathbf{a}\in\mathbb{R}^{p.k}} \frac{1}{2}\big\|\mathbf{z}-\mathbf{G}(\mathbf{b})\mathbf{a}\big\|_2^2 + \frac{\lambda}{2}\big\|\mathbf{a}\big\|_2^2 = \frac{1}{2}\left\| \begin{bmatrix}\mathbf{z}-\mathbf{G}(\mathbf{b})\mathbf{a}\\ \sqrt{\lambda}.\mathbf{a}\end{bmatrix}\right\|_2^2 = \frac{1}{2}\left\| \begin{bmatrix}\mathbf{z}\\ \mathbf{0}^{k.p}\end{bmatrix} - \begin{bmatrix}\mathbf{G}(\mathbf{b})\\ \sqrt{\lambda}.\mathbf{I}_{p.k}\end{bmatrix}\mathbf{a}\right\|_2^2,$$

and

$$\text{Arg}\min_{\mathbf{b}\in\mathbb{R}^{k.n}} \frac{1}{2}\big\|\mathbf{x}-\mathbf{F}(\mathbf{a})\mathbf{b}\big\|_2^2 + \frac{\lambda}{2}\big\|\mathbf{b}\big\|_2^2 = \frac{1}{2}\left\| \begin{bmatrix}\mathbf{x}-\mathbf{F}(\mathbf{a})\mathbf{b}\\ \sqrt{\lambda}.\mathbf{b}\end{bmatrix}\right\|_2^2 = \frac{1}{2}\left\| \begin{bmatrix}\mathbf{x}\\ \mathbf{0}^{k.n}\end{bmatrix} - \begin{bmatrix}\mathbf{F}(\mathbf{a})\\ \sqrt{\lambda}.\mathbf{I}_{k.n}\end{bmatrix}\mathbf{b}\right\|_2^2.$$

In other words, the MMMF ALS algorithm updates $\mathbf{a}$ and $\mathbf{b}$ at the $i+1$ iteration according to the rules

$$\mathbf{a}^{i+1} = \left(\mathbf{G}(\mathbf{b}^i)^T\mathbf{G}(\mathbf{b}^i) + \lambda.\mathbf{I}_{p.k}\right)^{-1}\mathbf{G}(\mathbf{b}^i)^T\mathbf{z}$$

and

$$\mathbf{b}^{i+1} = \left(\mathbf{F}(\mathbf{a}^{i+1})^T\mathbf{F}(\mathbf{a}^{i+1}) + \lambda.\mathbf{I}_{k.n}\right)^{-1}\mathbf{F}(\mathbf{a}^{i+1})^T\mathbf{x}\,.$$

Furthermore, Theorem 4.1 can now be applied directly to this regularized ALS algorithm in order to show that any accumulation point of the sequence $\{(\mathbf{A}^i, \mathbf{B}^i)\}_{i\in\mathbb{N}}$, say $(\widehat{\mathbf{A}}, \widehat{\mathbf{B}})$, is a stationary point of $\varphi^*_\lambda(.)$ as we are now sure that all the regression subproblems for computing $\mathbf{A}^i$ and $\mathbf{B}^i$ can be solved uniquely because of the presence of the regularization terms in $\varphi^*_\lambda(.)$.

Next, from Theorem 4.3 and its corollary (see below), we deduce that the partial Hessian matrices $\frac{\partial^2\varphi^*_\lambda(\mathbf{A},\mathbf{B})}{\partial^2\mathbf{a}} = \nabla^2(\varphi^*_\lambda)_{\mathbf{a}}(\mathbf{A},\mathbf{B})$ and $\frac{\partial^2\varphi^*_\lambda(\mathbf{A},\mathbf{B})}{\partial^2\mathbf{b}} = \nabla^2(\varphi^*_\lambda)_{\mathbf{b}}(\mathbf{A},\mathbf{B})$ are positive definite for all $\mathbf{A} \in \mathbb{R}^{p\times k}$ and $\mathbf{B} \in \mathbb{R}^{k\times n}$ as soon as $\lambda > 0$, which implies that $\varphi^*_\lambda(.)$ is now strongly bi-convex in its whole domain instead of only bi-convex as $\varphi^*(.)$. Using the facts that $\varphi^*_\lambda(.)$ is also a coercive (thanks to the inclusion of the regularization term $\frac{\lambda}{2}\|\mathbf{a}\|_2^2 + \frac{\lambda}{2}\|\mathbf{b}\|_2^2$) and real-analytic (as it is a polynomial in $(p\times k)+(k\times n)$ variables) function, it can be demonstrated that this strongly bi-convex cost function also verifies the so-called Kurdyka-Lojasiewicz inequality, the sequence $(\mathbf{A}^i, \mathbf{B}^i)$ is bounded and that the whole sequence $(\mathbf{A}^i, \mathbf{B}^i)$ generated by the MMMF ALS algorithm converges to a first-order

stationary point of $\varphi_\lambda^*(.)$, say $(\widehat{\mathbf{A}}, \widehat{\mathbf{B}})$ [194], which is a much stronger result than the one delivered by Theorem 4.1 and its corollary.

Finally, Li et al. [118], using results from [4][194], were able to demonstrate recently that the sequence $(\mathbf{A}^i, \mathbf{B}^i)$ generated by the following proximal version of the MMMF ALS algorithm

$$\begin{aligned}
\mathbf{a}^{i+1} &= \text{Arg}\min_{\mathbf{a}\in\mathbb{R}^{p.k}} \frac{1}{2}\|\mathbf{z} - \mathbf{G}(\mathbf{b}^i)\mathbf{a}\|_2^2 + \frac{\lambda}{2}\|\mathbf{a}\|_2^2 + \frac{\beta}{2}\|\mathbf{a}^i - \mathbf{a}\|_2^2 \\
&= \text{Arg}\min_{\mathbf{a}\in\mathbb{R}^{p.k}} \frac{1}{2}\left\| \begin{bmatrix} \mathbf{z} - \mathbf{G}(\mathbf{b}^i)\mathbf{a} \\ \sqrt{\lambda}.\mathbf{a} \\ \sqrt{\beta}(\mathbf{a}^i - \mathbf{a}) \end{bmatrix} \right\|_2^2 \\
&= \text{Arg}\min_{\mathbf{a}\in\mathbb{R}^{p.k}} \frac{1}{2}\left\| \begin{bmatrix} \mathbf{z} \\ \mathbf{0}^{p.k} \\ \sqrt{\beta}.\mathbf{a}^i \end{bmatrix} - \begin{bmatrix} \mathbf{G}(\mathbf{b}^i) \\ \sqrt{\lambda}.\mathbf{I}_{p.k} \\ \sqrt{\beta}.\mathbf{I}_{p.k} \end{bmatrix} \mathbf{a} \right\|_2^2 \\
&= \left(\mathbf{G}(\mathbf{b}^i)^T\mathbf{G}(\mathbf{b}^i) + (\lambda+\beta)\mathbf{I}_{p.k}\right)^{-1}\left(\mathbf{G}(\mathbf{b}^i)^T\mathbf{z} + \beta.\mathbf{a}^i\right)
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{b}^{i+1} &= \text{Arg}\min_{\mathbf{b}\in\mathbb{R}^{k.n}} \frac{1}{2}\|\mathbf{x} - \mathbf{F}(\mathbf{a}^{i+1})\mathbf{b}\|_2^2 + \frac{\lambda}{2}\|\mathbf{b}\|_2^2 + \frac{\beta}{2}\|\mathbf{b}^i - \mathbf{b}\|_2^2 \\
&= \text{Arg}\min_{\mathbf{b}\in\mathbb{R}^{k.n}} \frac{1}{2}\left\| \begin{bmatrix} \mathbf{x} - \mathbf{F}(\mathbf{a}^{i+1})\mathbf{b} \\ \sqrt{\lambda}.\mathbf{b} \\ \sqrt{\beta}(\mathbf{b}^i - \mathbf{b}) \end{bmatrix} \right\|_2^2 \\
&= \text{Arg}\min_{\mathbf{b}\in\mathbb{R}^{k.n}} \frac{1}{2}\left\| \begin{bmatrix} \mathbf{x} \\ \mathbf{0}^{k.n} \\ \sqrt{\beta}.\mathbf{b}^i \end{bmatrix} - \begin{bmatrix} \mathbf{F}(\mathbf{a}^{i+1}) \\ \sqrt{\lambda}.\mathbf{I}_{k.n} \\ \sqrt{\beta}.\mathbf{I}_{k.n} \end{bmatrix} \mathbf{b} \right\|_2^2 \\
&= \left(\mathbf{F}(\mathbf{a}^{i+1})^T\mathbf{F}(\mathbf{a}^{i+1}) + (\lambda+\beta)\mathbf{I}_{k.n}\right)^{-1}\left(\mathbf{F}(\mathbf{a}^{i+1})^T\mathbf{x} + \beta.\mathbf{b}^i\right),
\end{aligned}$$

where

$$\beta > 8.\|\mathbf{W}\|_S^2\varphi_\lambda^*(\mathbf{A}^0, \mathbf{B}^0)/\lambda + 4.\|\mathbf{W}\|_S\sqrt{\varphi_\lambda^*(\mathbf{A}^0, \mathbf{B}^0)} + \lambda,$$

converges not only to a first-order stationary point, but in fact to a second-order stationary point of $\varphi_\lambda^*(.)$ (see Proposition 4 and example 3 in Section 4.3 of [118]), e.g., to a point $(\widehat{\mathbf{A}}, \widehat{\mathbf{B}})$ which verifies

$$\nabla\varphi_\lambda^*(\widehat{\mathbf{A}}, \widehat{\mathbf{B}}) = (\mathbf{0}^{p\times k}, \mathbf{0}^{k\times n}) \text{ and } \left(\nabla^2\varphi_\lambda^*(\widehat{\mathbf{A}}, \widehat{\mathbf{B}})\right)\left((\mathbf{C}, \mathbf{D}), (\mathbf{C}, \mathbf{D})\right) \geq 0,$$

$\forall(\mathbf{C}, \mathbf{D}) \in \mathbb{R}^{p\times k} \times \mathbb{R}^{k\times n}$, e.g., $\nabla^2\varphi_\lambda^*(\widehat{\mathbf{A}}, \widehat{\mathbf{B}})$ is a positive semi-definite (symmetric) matrix. Importantly, if $\varphi_\lambda^*(.)$ is well-conditioned (e.g., depending on the form of the weight matrix $\mathbf{W}$), these second-order stationary points may correspond to a local or even global optimal solution, see [200] and Theorem 3.10 of [146] for more information.

Remark 4.1. An interesting and open question is to determine if these strong first- and second-order convergence properties of the ALS method for solving the MMMF formulation of the WLRA problem may also extend to the cost function $g_\lambda(.)$ proposed by Boumal and Absil [13][14] and discussed in Subsection 3.3 (see equation (3.18)). ■

The block ALS algorithm and its MMMF variant have also been incorporated as a building block in various Expectation-Maximization or other first-order methods to increase their efficiency for large datasets by avoiding costly SVD computations in high dimensions [93][86][181].

Interestingly, we note that Szlam et al. [178] have recently demonstrated that only a few iterations of such ALS are sufficient to produce nearly optimal spectral- and Frobenius-norm accuracies of low-rank approximations to a matrix when all the weights $\mathbf{W}_{ij}$ are equal to one, provided that $\mathbf{A}_0$ is one of the random matrices used by [85] (for example, the entries of $\mathbf{A}_0$ can be independent

and identically distributed standard normal variates) and that iterating until convergence is unnecessary. Extending their demonstration to the case when the weights $\mathbf{W}_{ij}$ are unequal (and eventually with some equal to zero) is an interesting issue already discussed in [167][23], but is outside the scope of this paper. However, we highlight again that proper initialization of the ALS or variable projection methods described here is obviously an important topic, which also needs a careful attention [72][94][76][170][171]. As an illustration, [94][76][170][167] showed that given a good enough initialization, many simple local search algorithms, like ALS, succeed, a result which is consistent with Theorem 4.2 above.

Now, let us consider how to compute efficiently the first- and second-order derivatives of the vectorized form of $\varphi^*(.)$ (and $\varphi^*_\lambda(.)$) in order to obtain meaningful tests of convergence of these ALS methods to a (local) minimizer of this cost function. We already know from Subsection 3.1 that the objective function $\varphi^*(.)$ used in the (P1) formulation of the WLRA problem,

$$\varphi^* : \mathbb{R}^{p\times k} \times \mathbb{R}^{k\times n} \longrightarrow \mathbb{R} : (\mathbf{A}, \mathbf{B}) \mapsto \varphi^*(\mathbf{A}, \mathbf{B}) = \frac{1}{2}\|\sqrt{\mathbf{W}} \odot (\mathbf{X} - \mathbf{AB})\|_F^2 \,,$$

is C^∞ differentiable over its domain of definition. Furthermore, we have also already derived the first- and second-order derivatives of $\varphi^*(.)$ in equations (3.12) and (3.15), respectively. As the vectorized form of $\varphi^*(.)$ is defined by the composition of $\varphi^*(.)$ with the linear mapping

$$\mathbb{R}^{p.k} \times \mathbb{R}^{k.n} \longrightarrow \mathbb{R}^{p\times k} \times \mathbb{R}^{k\times n} : (\mathbf{a}, \mathbf{b}) \mapsto \big(\mathit{mat}_{k\times p}(\mathbf{a})^T, \mathit{mat}_{k\times n}(\mathbf{b})\big) = (\mathbf{A}, \mathbf{B}) \,,$$

it is also C^∞ differentiable over its domain of definition, $\mathbb{R}^{p.k} \times \mathbb{R}^{k.n}$, and we have the following results concerning the vectorized forms of the first- and second-order derivatives of $\varphi^*(.)$, which offer more convenient expressions for checking the first- and second-KKT conditions of $\varphi^*(.)$ at a given pair of $\mathbb{R}^{p\times k} \times \mathbb{R}^{k\times n}$ then equations (3.12) and (3.15).

Theorem 4.3. For $\mathbf{X} \in \mathbb{R}^{p\times n}$, $\sqrt{\mathbf{W}} \in \mathbb{R}^{p\times n}_+$ and any fixed integer $k \leq \mathit{rank}(\mathbf{X}) \leq \min(p, n)$, the vectorized partial first-derivatives of $\varphi^*(.)$ with respect to $\mathbf{a} = \mathit{vec}(\mathbf{A}^T)$ and $\mathbf{b} = \mathit{vec}(\mathbf{B})$ are equal, respectively, to

$$\frac{\partial\varphi^*(\mathbf{A}, \mathbf{B})}{\partial \mathbf{a}} = \nabla\varphi^*_{\mathbf{a}}(\mathbf{A}, \mathbf{B}) = \mathbf{G}(\mathbf{b})^T\mathbf{G}(\mathbf{b})\mathbf{a} - \mathbf{G}(\mathbf{b})^T\mathbf{z} \tag{4.3}$$

and

$$\frac{\partial\varphi^*(\mathbf{A}, \mathbf{B})}{\partial \mathbf{b}} = \nabla\varphi^*_{\mathbf{b}}(\mathbf{A}, \mathbf{B}) = \mathbf{F}(\mathbf{a})^T\mathbf{F}(\mathbf{a})\mathbf{b} - \mathbf{F}(\mathbf{a})^T\mathbf{x} \,, \tag{4.4}$$

where

$$\begin{aligned}
\mathbf{F}(\mathbf{a}) &= \bigoplus_{j=1}^{n} \mathbf{F}_j(\mathbf{a}) = \bigoplus_{j=1}^{n} \mathit{diag}(\sqrt{\mathbf{W}}_{.j})\big(\mathit{mat}_{k\times p}(\mathbf{a})\big)^T \,, \\
\mathbf{G}(\mathbf{b}) &= \bigoplus_{i=1}^{p} \mathbf{G}_i(\mathbf{b}) = \bigoplus_{i=1}^{p} \mathit{diag}(\sqrt{\mathbf{W}}_{i.})\big(\mathit{mat}_{k\times n}(\mathbf{b})\big)^T \,, \\
\mathbf{x} &= \mathit{vec}(\sqrt{\mathbf{W}} \odot \mathbf{X}) \text{ and } \mathbf{z} = \mathit{vec}\big((\sqrt{\mathbf{W}} \odot \mathbf{X})^T\big) \,.
\end{aligned}$$

Moreover, we have

$$\begin{aligned}
\nabla\varphi^*_{\mathbf{a}}(\mathbf{A}, \mathbf{B}) &= \mathit{vec}\big(\nabla\varphi^*_{\mathbf{A}}(\mathbf{A}, \mathbf{B})\big) \,, \\
\nabla\varphi^*_{\mathbf{b}}(\mathbf{A}, \mathbf{B}) &= \mathit{vec}\big(\nabla\varphi^*_{\mathbf{B}}(\mathbf{A}, \mathbf{B})\big) \,,
\end{aligned} \tag{4.5}$$

where $\nabla\varphi^*_{\mathbf{A}}(\mathbf{A}, \mathbf{B})$ and $\nabla\varphi^*_{\mathbf{B}}(\mathbf{A}, \mathbf{B})$ are defined in equation (3.11).

The vectorized second-derivative (symmetric) matrix form of $\varphi^*(.)$ is given by

$$\begin{aligned}
\big[\nabla^2\varphi^*(\mathbf{A}, \mathbf{B})\big] &= \begin{bmatrix} \frac{\partial^2\varphi^*(\mathbf{A},\mathbf{B})}{\partial^2\mathbf{a}} & \frac{\partial^2\varphi^*(\mathbf{A},\mathbf{B})}{\partial\mathbf{a}\partial\mathbf{b}} \\ \frac{\partial^2\varphi^*(\mathbf{A},\mathbf{B})}{\partial\mathbf{b}\partial\mathbf{a}} & \frac{\partial^2\varphi^*(\mathbf{A},\mathbf{B})}{\partial^2\mathbf{b}} \end{bmatrix} \\
&= \begin{bmatrix} \nabla^2\varphi^*_{\mathbf{a}}(\mathbf{A}, \mathbf{B}) & \nabla^2\varphi^*_{\mathbf{a},\mathbf{b}}(\mathbf{A}, \mathbf{B}) \\ \nabla^2\varphi^*_{\mathbf{b},\mathbf{a}}(\mathbf{A}, \mathbf{B}) & \nabla^2\varphi^*_{\mathbf{b}}(\mathbf{A}, \mathbf{B}) \end{bmatrix} \,,
\end{aligned} \tag{4.6}$$

where

$$\nabla^2\varphi^*_{\mathbf{a}}(\mathbf{A},\mathbf{B}) = \frac{\partial^2\varphi^*(\mathbf{A},\mathbf{B})}{\partial^2\mathbf{a}} = \mathbf{G}(\mathbf{b})^T\mathbf{G}(\mathbf{b}) \,,$$
$$\nabla^2\varphi^*_{\mathbf{b}}(\mathbf{A},\mathbf{B}) = \frac{\partial^2\varphi^*(\mathbf{A},\mathbf{B})}{\partial^2\mathbf{b}} = \mathbf{F}(\mathbf{a})^T\mathbf{F}(\mathbf{a}) \,,$$
$$\nabla^2\varphi^*_{\mathbf{b},\mathbf{a}}(\mathbf{A},\mathbf{B}) = \frac{\partial^2\varphi^*(\mathbf{A},\mathbf{B})}{\partial\mathbf{a}\partial\mathbf{b}} = \Big(\big(\mathbf{W}\odot(\mathbf{AB}-\mathbf{X})\big)^T\otimes\mathbf{I}_k\Big) + \mathbf{F}(\mathbf{a})^T\mathbf{K}_{(n,p)}\mathbf{G}(\mathbf{b}) \,,$$
$$\nabla^2\varphi^*_{\mathbf{a},\mathbf{b}}(\mathbf{A},\mathbf{B}) = \big[\nabla^2\varphi^*_{\mathbf{b},\mathbf{a}}(\mathbf{A},\mathbf{B})\big]^T \,.$$

Finally, we have the following equalities, which precise the relationships between the quadratic forms $\big(\nabla^2\varphi^*(\mathbf{A},\mathbf{B})\big)$ and $\big(\nabla^2\varphi(\mathbf{AB})\big)$ in complement of equations (3.14) and (3.15)

$$\mathbf{c}^T\nabla^2\varphi^*_{\mathbf{a}}(\mathbf{A},\mathbf{B})\mathbf{c} = \big(\nabla^2\varphi^*_{\mathbf{A}}(\mathbf{A},\mathbf{B})\big)(\mathbf{C},\mathbf{C}) = \big(\nabla^2\varphi(\mathbf{AB})\big)(\mathbf{CB},\mathbf{CB}) \,,$$
$$\mathbf{d}^T\nabla^2\varphi^*_{\mathbf{b}}(\mathbf{A},\mathbf{B})\mathbf{d} = \big(\nabla^2\varphi^*_{\mathbf{B}}(\mathbf{A},\mathbf{B})\big)(\mathbf{D},\mathbf{D}) = \big(\nabla^2\varphi(\mathbf{AB})\big)(\mathbf{AD},\mathbf{AD}) \,,$$
$$\mathbf{d}^T\nabla^2\varphi^*_{\mathbf{b},\mathbf{a}}(\mathbf{A},\mathbf{B})\mathbf{c} = \big\langle\nabla\varphi(\mathbf{AB}),\mathbf{CD}\big\rangle_F + \big(\nabla^2\varphi(\mathbf{AB})\big)(\mathbf{AD},\mathbf{CB}) \,,$$

$\forall\mathbf{C}\in\mathbb{R}^{p\times k}$ with $\mathbf{c} = vec(\mathbf{C}^T)$ and $\forall\mathbf{D}\in\mathbb{R}^{k\times n}$ with $\mathbf{d} = vec(\mathbf{D})$.

Proof. First, we observe that the matrix of first-derivatives of the vectorized residual function $\mathbf{e}(\mathbf{a},\mathbf{b}) = \mathbf{e}(\mathbf{A},\mathbf{B}) = \mathbf{x} - \mathbf{F}(\mathbf{a})\mathbf{b}$ with respect to $\mathbf{b}$ ($\mathbf{e}(\mathbf{A},\mathbf{B})$ is first defined in equation (3.19)) is simply

$$\frac{\partial\mathbf{e}(\mathbf{A},\mathbf{B})}{\partial\mathbf{b}} = -\mathbf{F}(\mathbf{a})$$

and is very sparse with only k non-zero elements in each row as $\mathbf{F}(\mathbf{a})$ is a block diagonal matrix (see equation (3.20)). Since

$$\varphi^*(\mathbf{A},\mathbf{B}) = \frac{1}{2}\mathbf{e}(\mathbf{A},\mathbf{B})^T\mathbf{e}(\mathbf{A},\mathbf{B}) = \frac{1}{2}\|\mathbf{e}(\mathbf{a},\mathbf{b})\|_2^2 \,,$$

the derivative of $\varphi^*(\mathbf{A},\mathbf{B})$ with respect to $\mathbf{b}$ is then easy to compute, using a standard differential rule for a mapping of the form $\mathbb{R}^{k.n}\longrightarrow\mathbb{R}:\mathbf{d}\mapsto\frac{1}{2}\|g(\mathbf{d})\|_2^2$, where $g(.)$ is a differentiable mapping from $\mathbb{R}^{k.n}$ to $\mathbb{R}^{p.n}$ [26],

$$\frac{\partial\varphi^*(\mathbf{A},\mathbf{B})}{\partial\mathbf{b}} = -\mathbf{F}(\mathbf{a})^T\big(\mathbf{x}-\mathbf{F}(\mathbf{a})\mathbf{b}\big) = \mathbf{F}(\mathbf{a})^T\mathbf{F}(\mathbf{a})\mathbf{b} - \mathbf{F}(\mathbf{a})^T\mathbf{x} \,.$$

For computing the derivative of $\varphi^*(\mathbf{A},\mathbf{B})$ with respect to $\mathbf{a}$, we first recall that the vectorized residual function $\mathbf{e}(\mathbf{a},\mathbf{b})$ may also be expressed in the alternative form

$$\mathbf{e}(\mathbf{A},\mathbf{B}) = \mathbf{x} - \mathbf{K}_{(n,p)}\mathbf{G}(\mathbf{b})\mathbf{a} = \mathbf{K}_{(n,p)}\big(\mathbf{z}-\mathbf{G}(\mathbf{b})\mathbf{a}\big) \,,$$

see the paragraph after equation (3.22) in Subsection 3.4 for details. Hence

$$\frac{\partial\mathbf{e}(\mathbf{A},\mathbf{B})}{\partial\mathbf{a}} = -\mathbf{K}_{(n,p)}\mathbf{G}(\mathbf{b})$$

and this matrix of derivatives with respect to $\mathbf{a}$ is also very sparse with only k non-zero elements in each row. Now the derivative of $\varphi^*(\mathbf{A},\mathbf{B})$ with respect to $\mathbf{a}$, is also very simple to obtain, using the same differentiation rule as above and properties of the commutation matrix given in Subsection 2.2,

$$\begin{aligned}\frac{\partial\varphi^*(\mathbf{A},\mathbf{B})}{\partial\mathbf{a}} &= -\big(\mathbf{K}_{(n,p)}\mathbf{G}(\mathbf{b})\big)^T\mathbf{K}_{(n,p)}\big(\mathbf{z}-\mathbf{G}(\mathbf{b})\mathbf{a}\big)\\ &= -\mathbf{G}(\mathbf{b})^T\mathbf{K}_{(p,n)}\mathbf{K}_{(n,p)}\big(\mathbf{z}-\mathbf{G}(\mathbf{b})\mathbf{a}\big)\\ &= \mathbf{G}(\mathbf{b})^T\mathbf{G}(\mathbf{b})\mathbf{a} - \mathbf{G}(\mathbf{b})^T\mathbf{z} \,.\end{aligned}$$

Next, to demonstrate that $\nabla\varphi^*_{\mathbf{b}}(\mathbf{A},\mathbf{B}) = vec\big(\nabla\varphi^*_{\mathbf{B}}(\mathbf{A},\mathbf{B})\big)$, we observe that, by definition, we have

$$\mathbf{F}(\mathbf{a})\mathbf{b} = diag\big(vec(\sqrt{\mathbf{W}})\big)(\mathbf{I}_n \otimes \mathbf{A})vec(\mathbf{B}) = diag\big(vec(\sqrt{\mathbf{W}})\big)vec(\mathbf{AB})$$

and, thus,

$$\begin{aligned}\mathbf{F}(\mathbf{a})\mathbf{b} - \mathbf{x} &= diag\big(vec(\sqrt{\mathbf{W}})\big)vec(\mathbf{AB}) - vec(\sqrt{\mathbf{W}} \odot \mathbf{X}) \\ &= diag\big(vec(\sqrt{\mathbf{W}})\big)\big(vec(\mathbf{AB}) - vec(\mathbf{X})\big) \\ &= diag\big(vec(\sqrt{\mathbf{W}})\big)vec(\mathbf{AB} - \mathbf{X}) \ ,\end{aligned}$$

which implies that

$$\begin{aligned}\nabla\varphi^*_{\mathbf{b}}(\mathbf{A},\mathbf{B}) &= \mathbf{F}(\mathbf{a})^T\big(\mathbf{F}(\mathbf{a})\mathbf{b} - \mathbf{x}\big) \\ &= (\mathbf{I}_n \otimes \mathbf{A})^T diag\big(vec(\mathbf{W})\big)vec(\mathbf{AB} - \mathbf{X}) \\ &= (\mathbf{I}_n \otimes \mathbf{A}^T) diag\big(vec(\mathbf{W})\big)vec(\mathbf{AB} - \mathbf{X}) \\ &= (\mathbf{I}_n \otimes \mathbf{A}^T)vec\big(\mathbf{W} \odot (\mathbf{AB} - \mathbf{X})\big) \\ &= vec\Big(\mathbf{A}^T\big(\mathbf{W} \odot (\mathbf{AB} - \mathbf{X})\big)\Big),\end{aligned}$$

and, using equation (3.11), we conclude that

$$\nabla\varphi^*_{\mathbf{b}}(\mathbf{A},\mathbf{B}) = vec\big(\nabla\varphi^*_{\mathbf{B}}(\mathbf{A},\mathbf{B})\big) \ .$$

Similarly, for demonstrating that $\nabla\varphi^*_{\mathbf{a}}(\mathbf{A},\mathbf{B}) = vec\big(\nabla\varphi^*_{\mathbf{A}}(\mathbf{A},\mathbf{B})\big)$, we observe that

$$\mathbf{G}(\mathbf{b})\mathbf{a} = diag\big(vec(\sqrt{\mathbf{W}^T})\big)(\mathbf{I}_p \otimes \mathbf{B}^T)vec(\mathbf{A}^T) = diag\big(vec(\sqrt{\mathbf{W}^T})\big)vec\big((\mathbf{AB})^T\big)$$

and, thus,

$$\begin{aligned}\mathbf{G}(\mathbf{b})\mathbf{a} - \mathbf{z} &= diag\big(vec(\sqrt{\mathbf{W}^T})\big)vec\big((\mathbf{AB})^T\big) - vec\big((\sqrt{\mathbf{W}} \odot \mathbf{X})^T\big) \\ &= diag\big(vec(\sqrt{\mathbf{W}^T})\big)\Big(vec\big((\mathbf{AB})^T\big) - vec(\mathbf{X}^T)\Big) \\ &= diag\big(vec(\sqrt{\mathbf{W}^T})\big)vec\big((\mathbf{AB} - \mathbf{X})^T\big) \ ,\end{aligned}$$

which implies that

$$\begin{aligned}\nabla\varphi^*_{\mathbf{a}}(\mathbf{A},\mathbf{B}) &= \mathbf{G}(\mathbf{b})^T\big(\mathbf{G}(\mathbf{b})\mathbf{a} - \mathbf{z}\big) \\ &= (\mathbf{I}_p \otimes \mathbf{B}) diag\big(vec(\mathbf{W}^T)\big)vec\big((\mathbf{AB} - \mathbf{X})^T\big) \\ &= (\mathbf{I}_p \otimes \mathbf{B})vec\big(\mathbf{W}^T \odot (\mathbf{AB} - \mathbf{X})^T\big) \\ &= vec\Big(\mathbf{B}\big(\mathbf{W} \odot (\mathbf{AB} - \mathbf{X})\big)^T\Big) \\ &= vec\Big(\big(\mathbf{W} \odot (\mathbf{AB} - \mathbf{X})\big)\mathbf{B}^T\big)^T\Big) \ ,\end{aligned}$$

and, using again equation (3.11), we conclude that

$$\nabla\varphi^*_{\mathbf{a}}(\mathbf{A},\mathbf{B}) = vec\big(\nabla\varphi^*_{\mathbf{A}}(\mathbf{A},\mathbf{B})^T\big) \ .$$

Next, we immediately get that the (partial) Hessian matrices of the vectorized form of $\varphi^*(.)$ are equal to

$$\begin{aligned}\nabla^2\varphi^*_{\mathbf{a}}(\mathbf{A},\mathbf{B}) &= \mathbf{G}(\mathbf{b})^T\mathbf{G}(\mathbf{b}) \ , \\ \nabla^2\varphi^*_{\mathbf{b}}(\mathbf{A},\mathbf{B}) &= \mathbf{F}(\mathbf{a})^T\mathbf{F}(\mathbf{a}) \ ,\end{aligned}$$

since the specific forms of $\frac{\partial\varphi^*(\mathbf{A},\mathbf{B})}{\partial\mathbf{a}}$ and $\frac{\partial\varphi^*(\mathbf{A},\mathbf{B})}{\partial\mathbf{b}}$ derived above can be both interpreted as the sum of a linear mapping and a constant term, when they are considered as a function of $\mathbf{a}$ and $\mathbf{b}$, respectively.

To derive an explicit formula for $\nabla^2\varphi^*_{\mathbf{b},\mathbf{a}}(\mathbf{A},\mathbf{B})$, we start from the equation

$$\frac{\partial\varphi^*(\mathbf{A},\mathbf{B})}{\partial\mathbf{b}} = \mathbf{F}(\mathbf{a})^T\big(\mathbf{F}(\mathbf{a})\mathbf{b}-\mathbf{x}\big) = \big(\mathbf{I}_n\otimes\mathbf{A}^T\big)vec\big(\mathbf{W}\odot(\mathbf{AB}-\mathbf{X})\big)\ ,$$

and apply the differential rule for a matrix product [26] to get

$$\nabla^2\varphi^*_{\mathbf{b},\mathbf{a}}(\mathbf{A},\mathbf{B})\mathbf{c} = (\mathbf{I}_n\otimes\mathbf{C}^T)vec\big(\mathbf{W}\odot(\mathbf{AB}-\mathbf{X})\big) + (\mathbf{I}_n\otimes\mathbf{A}^T)vec(\mathbf{W}\odot\mathbf{CB})\ ,$$

$\forall\mathbf{C}\in\mathbb{R}^{p\times k}$ with $\mathbf{c}=vec(\mathbf{C}^T)$. On one hand, using equation (2.33), we have

$$\begin{aligned}(\mathbf{I}_n\otimes\mathbf{C}^T)vec\big(\mathbf{W}\odot(\mathbf{AB}-\mathbf{X})\big) &= vec\Big(\mathbf{C}^T\big(\mathbf{W}\odot(\mathbf{AB}-\mathbf{X})\big)\Big)\\ &= \Big(\big(\mathbf{W}\odot(\mathbf{AB}-\mathbf{X})\big)^T\otimes\mathbf{I}_k\Big)vec(\mathbf{C}^T)\\ &= \Big(\big(\mathbf{W}\odot(\mathbf{AB}-\mathbf{X})\big)^T\otimes\mathbf{I}_k\Big)\mathbf{c}\ ,\end{aligned}$$

and, on the other hand, using equations (2.33), (2.34), (2.35) and Lemma 2.2, we get

$$\begin{aligned}(\mathbf{I}_n\otimes\mathbf{A}^T)vec(\mathbf{W}\odot\mathbf{CB}) &= (\mathbf{I}_n\otimes\mathbf{A}^T)diag\big(vec(\mathbf{W})\big)vec(\mathbf{CB})\\ &= \Big(diag\big(vec(\sqrt{\mathbf{W}})\big)\mathbf{F}(\mathbf{a})\Big)^T(\mathbf{B}^T\otimes\mathbf{I}_p)vec(\mathbf{C})\\ &= \Big(diag\big(vec(\sqrt{\mathbf{W}})\big)\mathbf{F}(\mathbf{a})\Big)^T(\mathbf{B}^T\otimes\mathbf{I}_p)\mathbf{K}_{(k,p)}\mathbf{K}_{(p,k)}vec(\mathbf{C})\\ &= \mathbf{F}(\mathbf{a})^T diag\big(vec(\sqrt{\mathbf{W}})\big)(\mathbf{B}^T\otimes\mathbf{I}_p)\mathbf{K}_{(k,p)}vec(\mathbf{C}^T)\\ &= \mathbf{F}(\mathbf{a})^T diag\big(vec(\sqrt{\mathbf{W}})\big)\mathbf{K}_{(n,p)}(\mathbf{I}_p\otimes\mathbf{B}^T)vec(\mathbf{C}^T)\\ &= \mathbf{F}(\mathbf{a})^T\mathbf{K}_{(n,p)}diag\big(vec(\sqrt{\mathbf{W}}^T)\big)(\mathbf{I}_p\otimes\mathbf{B}^T)vec(\mathbf{C}^T)\\ &= \mathbf{F}(\mathbf{a})^T\mathbf{K}_{(n,p)}\mathbf{G}(\mathbf{b})\mathbf{c}\ .\end{aligned}$$

Together, these equalities imply, finally, that

$$\nabla^2\varphi^*_{\mathbf{b},\mathbf{a}}(\mathbf{A},\mathbf{B}) = \Big(\big(\mathbf{W}\odot(\mathbf{AB}-\mathbf{X})\big)^T\otimes\mathbf{I}_k\Big) + \mathbf{F}(\mathbf{a})^T\mathbf{K}_{(n,p)}\mathbf{G}(\mathbf{b})\ ,$$

as claimed in the theorem.

Finally, the equality $\nabla^2\varphi^*_{\mathbf{a},\mathbf{b}}(\mathbf{A},\mathbf{B}) = \big[\nabla^2\varphi^*_{\mathbf{b},\mathbf{a}}(\mathbf{A},\mathbf{B})\big]^T$ is a direct consequence of the fact that the Hessian $\nabla^2\varphi^*(\mathbf{A},\mathbf{B})$ is a $(p.k+k.n)\times(p.k+k.n)$ symmetric matrix according to the Schwarz's theorem [26], see Subsection 2.4 and Remark 4.3 below for details.

It remains to establish the equalities between the quadratic forms $\big(\nabla^2\varphi^*(\mathbf{A},\mathbf{B})\big)$ and $\big(\nabla^2\varphi(\mathbf{AB})\big)$. First, note that

$$\begin{aligned}\nabla^2\varphi^*_{\mathbf{a}}(\mathbf{A},\mathbf{B})\mathbf{c} &= vec\Big(\big((\mathbf{W}\odot\mathbf{CB})\mathbf{B}^T\big)^T\Big)\ ,\\ \nabla^2\varphi^*_{\mathbf{b}}(\mathbf{A},\mathbf{B})\mathbf{d} &= vec\big(\mathbf{A}^T(\mathbf{W}\odot\mathbf{AD})\big)\ ,\end{aligned}$$

$\forall\mathbf{C}\in\mathbb{R}^{p\times k}$ with $\mathbf{c}=vec(\mathbf{C}^T)$ and $\forall\mathbf{D}\in\mathbb{R}^{k\times n}$ with $\mathbf{d}=vec(\mathbf{D})$.

Using these equalities, we deduce

$$\begin{aligned}
\mathbf{c}^T\nabla^2\varphi^*_{\mathbf{a}}(\mathbf{A},\mathbf{B})\mathbf{c} &= \big\langle \nabla^2\varphi^*_{\mathbf{a}}(\mathbf{A},\mathbf{B})\mathbf{c},\mathbf{c}\big\rangle_2 \\
&= \Big\langle vec\Big(\big((\mathbf{W}\odot\mathbf{CB})\mathbf{B}^T\big)^T\Big), vec(\mathbf{C}^T)\Big\rangle_2 \\
&= \big\langle vec\big((\mathbf{W}\odot\mathbf{CB})\mathbf{B}^T\big), vec(\mathbf{C})\big\rangle_2 \\
&= \big\langle (\mathbf{W}\odot\mathbf{CB})\mathbf{B}^T,\mathbf{C}\big\rangle_F \\
&= \big\langle [\nabla^2\varphi^*_{\mathbf{A}}(\mathbf{A},\mathbf{B})](\mathbf{C}),\mathbf{C}\big\rangle_F \\
&= \big(\nabla^2\varphi^*_{\mathbf{A}}(\mathbf{A},\mathbf{B})\big)(\mathbf{C},\mathbf{C}) \\
&= \big(\nabla^2\varphi(\mathbf{AB})\big)(\mathbf{CB},\mathbf{CB}) \ ,
\end{aligned}$$

and also

$$\begin{aligned}
\mathbf{d}^T\nabla^2\varphi^*_{\mathbf{b}}(\mathbf{A},\mathbf{B})\mathbf{d} &= \big\langle \nabla^2\varphi^*_{\mathbf{b}}(\mathbf{A},\mathbf{B})\mathbf{d},\mathbf{d}\big\rangle_2 \\
&= \big\langle vec\big(\mathbf{A}^T(\mathbf{W}\odot\mathbf{AD})\big), vec(\mathbf{D})\big\rangle_2 \\
&= \big\langle \mathbf{A}^T(\mathbf{W}\odot\mathbf{AD}),\mathbf{D}\big\rangle_F \\
&= \big\langle [\nabla^2\varphi^*_{\mathbf{B}}(\mathbf{A},\mathbf{B})](\mathbf{D}),\mathbf{D}\big\rangle_F \\
&= \big(\nabla^2\varphi^*_{\mathbf{B}}(\mathbf{A},\mathbf{B})\big)(\mathbf{D},\mathbf{D}) \\
&= \big(\nabla^2\varphi(\mathbf{AB})\big)(\mathbf{AD},\mathbf{AD}) \ ,
\end{aligned}$$

where, in both cases, the last equality results from equation (3.14). The last equality in the theorem,

$$\mathbf{d}^T\nabla^2\varphi^*_{\mathbf{b},\mathbf{a}}(\mathbf{A},\mathbf{B})\mathbf{c} = \big\langle \nabla\varphi(\mathbf{AB}),\mathbf{CD}\big\rangle_F + \big(\nabla^2\varphi(\mathbf{AB})\big)(\mathbf{AD},\mathbf{CB}) \ ,$$

can be derived in a similar way, by a lengthy, but direct, computation.

□

Corollary 4.2. For $\mathbf{X}\in\mathbb{R}^{p\times n}$, $\sqrt{\mathbf{W}}\in\mathbb{R}^{p\times n}_+$, $\lambda\in\mathbb{R}_{+*}$ and any fixed integer $k\leq$ *rank*$(\mathbf{X})\leq \min(p,n)$, the objective function $\varphi^*_\lambda(.)$ used in the (MMMF) formulation of the WLRA problem

$$\varphi^*_\lambda : \mathbb{R}^{p\times k}\times\mathbb{R}^{k\times n}\longrightarrow\mathbb{R} : (\mathbf{A},\mathbf{B})\mapsto\varphi^*_\lambda(\mathbf{A},\mathbf{B}) = \frac{1}{2}\|\sqrt{\mathbf{W}}\odot(\mathbf{X}-\mathbf{AB})\|^2_F + \frac{\lambda}{2}(\|\mathbf{A}\|^2_F+\|\mathbf{B}\|^2_F)$$

is C^∞ differentiable over its domain of definition and the partial first-order derivatives of $\varphi^*_\lambda(.)$ with respect to $\mathbf{a}=vec(\mathbf{A}^T)$ and $\mathbf{b}=vec(\mathbf{B})$ are equal, respectively, to

$$\frac{\partial\varphi^*_\lambda(\mathbf{A},\mathbf{B})}{\partial\mathbf{a}} = \nabla(\varphi^*_\lambda)_{\mathbf{a}}(\mathbf{A},\mathbf{B}) = \mathbf{G}(\mathbf{b})^T\mathbf{G}(\mathbf{b})\mathbf{a} - \mathbf{G}(\mathbf{b})^T\mathbf{z} + \lambda\mathbf{a}$$

and

$$\frac{\partial\varphi^*_\lambda(\mathbf{A},\mathbf{B})}{\partial\mathbf{b}} = \nabla(\varphi^*_\lambda)_{\mathbf{b}}(\mathbf{A},\mathbf{B}) = \mathbf{F}(\mathbf{a})^T\mathbf{F}(\mathbf{a})\mathbf{b} - \mathbf{F}(\mathbf{a})^T\mathbf{x} + \lambda\mathbf{b} \ .$$

Furthermore, the partial second-order derivatives of $\varphi^*_\lambda(.)$ with respect to $\mathbf{a}=vec(\mathbf{A}^T)$ and $\mathbf{b}=vec(\mathbf{B})$ are given by

$$\frac{\partial^2\varphi^*_\lambda(\mathbf{A},\mathbf{B})}{\partial^2\mathbf{a}} = \nabla^2(\varphi^*_\lambda)_{\mathbf{a}}(\mathbf{A},\mathbf{B}) = \mathbf{G}(\mathbf{b})^T\mathbf{G}(\mathbf{b}) + \lambda.\mathbf{I}_{p.k}$$

and

$$\frac{\partial^2\varphi^*_\lambda(\mathbf{A},\mathbf{B})}{\partial^2\mathbf{b}} = \nabla^2(\varphi^*_\lambda)_{\mathbf{b}}(\mathbf{A},\mathbf{B}) = \mathbf{F}(\mathbf{a})^T\mathbf{F}(\mathbf{a}) + \lambda.\mathbf{I}_{k.n} \ .$$

Proof. $\varphi^*_\lambda(.)$ is the sum of three C^∞ differentiable functions, e.g., $\varphi^*(.)$ and the mappings $\frac{\lambda}{2}\|\mathbf{A}\|^2_F = \frac{\lambda}{2}\|\mathbf{a}\|^2_2$ and $\frac{\lambda}{2}\|\mathbf{B}\|^2_F = \frac{\lambda}{2}\|\mathbf{b}\|^2_2$ and is thus C^∞ differentiable. The formulas for $\frac{\partial\varphi^*_\lambda(\mathbf{A},\mathbf{B})}{\partial\mathbf{a}}$ and

$\frac{\partial \varphi_\lambda^*(\mathbf{A},\mathbf{B})}{\partial \mathbf{b}}$ follow immediately from Theorem 4.3, standard differentiation rules and the differential rule for a mapping of the form $\mathbb{R}^m \longrightarrow \mathbb{R} : \mathbf{x} \mapsto \frac{1}{2}\|\mathbf{x}\|_2^2$.

The form of the partial second-order derivatives of $\varphi_\lambda^*(.)$ given in the theorem is a direct consequence of Theorem 4.3 and the fact that $\nabla(\varphi_\lambda^*)_\mathbf{a}(\mathbf{A},\mathbf{B})$ and $\nabla(\varphi_\lambda^*)_\mathbf{b}(\mathbf{A},\mathbf{B})$ are both the sum of two linear mappings and of a constant term when they are considered as a function of $\mathbf{a}$ and $\mathbf{b}$, respectively.

□

Remark 4.2. The equations

$$\frac{\partial \mathbf{e}(\mathbf{A},\mathbf{B})}{\partial \mathbf{b}} = -\mathbf{F}(\mathbf{a}) \quad \text{and} \quad \frac{\partial \mathbf{e}(\mathbf{A},\mathbf{B})}{\partial \mathbf{a}} = -\mathbf{K}_{(n,p)}\mathbf{G}(\mathbf{b})$$

derived in the proof of Theorem 4.3 show that residual function $\mathbf{e}(.)$ is not a nonlinear function of its arguments as defined in Subsection 2.4. However, despite of this, the cost function $\varphi^*(.)$ is still nonlinear as the partial derivatives of $\varphi^*(.)$ with respect to $\mathbf{a}$ and $\mathbf{b}$ are functions of $\mathbf{b}$ and $\mathbf{a}$, respectively, as demonstrated in Theorem 4.3. Furthermore, as the minimization of the cost function $\varphi^*(.)$ has no closed form solution in general, we can still consider $\varphi^*(.)$ as a NLLS functional as defined in Subsection 2.4. ■

Due to the block diagonal structures of both $\mathbf{F}(\mathbf{a})$ and $\mathbf{G}(\mathbf{b})$, the evaluation of the partial derivatives of $\varphi^*(.)$ is fast, easy to implement and may be parallelized. Moreover, we already know that

$$\frac{\partial \varphi^*(\mathbf{A},\mathbf{B})}{\partial \mathbf{a}} = \mathbf{0}^{k.p} \quad \text{or} \quad \frac{\partial \varphi^*(\mathbf{A},\mathbf{B})}{\partial \mathbf{b}} = \mathbf{0}^{k.n}$$

if the ALS algorithm is used to minimize $\varphi^*(\mathbf{A},\mathbf{B})$ and the iterations are stopped after computing $\mathbf{A}$ or $\mathbf{B}$, respectively. Similar remarks apply to the partial derivatives of $\varphi_\lambda^*(.)$.

The main payoff of the two-block ALS method is its simplicity since it involves solving mainly two sequences of small (eventually regularized) linear least-squares problems. Moreover, compared to gradient-type algorithms, it has the advantage that there is no need to tune optimization parameters like step sizes [146]. However, practical experience with this algorithm shows that, in many cases, the "NIPALS" iterates do not converge to the closest fit (e.g., the infimum or minimum of $\varphi^*(\mathbf{A},\mathbf{B})$ or $\varphi_\lambda^*(\mathbf{A},\mathbf{B})$) and get frequently stuck in sub-optimal local minima for a small value of k or a poorly chosen starting point [72][157]. This is especially true when some weights are equal to 0 (i.e., when missing values are present in $\mathbf{X}$), even with the initialization procedure proposed by Gabriel and Zamir [72]. Moreover this initialization procedure is only applicable if $k = 1$ and if there is one and only one missing cell ($\mathbf{W}_{ij} = 0$) for the matrix entries in the i^{th} row and j^{th} column of $\mathbf{X}$ for all i and j (see Gabriel and Zamir [72] for more details). Furthermore, it is known that the two-block ALS algorithm is vulnerable to flatlining [15] and inherits in many cases of the very slow convergence of the block coordinate descent method [139]. To overcome these difficulties, we describe in the next section, various first- and second-order separable NLLS algorithms for minimizing $\psi(.)$ instead of $\varphi^*(.)$.

Remark 4.3. If we concatenate the vectors $\mathbf{a}$ and $\mathbf{b}$ in $\mathbf{c} = \begin{bmatrix}\mathbf{a}\\ \mathbf{b}\end{bmatrix} \in \mathbb{R}^{k.(p+n)}$, we may define the following residual and objective functions:

$$r(\mathbf{c}) = \mathbf{x} - \mathbf{F}(\mathbf{a})\mathbf{b} = \mathbf{K}_{(n,p)}\big(\mathbf{z} - \mathbf{G}(\mathbf{b})\mathbf{a}\big) = \mathbf{e}(\mathbf{A},\mathbf{B})$$

and

$$\phi(\mathbf{c}) = \frac{1}{2}r(\mathbf{c})^T r(\mathbf{c}) = \frac{1}{2}\|r(\mathbf{c})\|_2^2 = \varphi^*(\mathbf{A},\mathbf{B}).$$

According to equation (2.65), the gradient of $\phi(.)$ is then equal to

$$\nabla\phi(\mathbf{c}) = J(\mathbf{r}(\mathbf{c}))^T\mathbf{r}(\mathbf{c}) \ ,$$

with the Jacobian matrix $J(\mathbf{r}(\mathbf{c})) \in \mathbb{R}^{p.n \times k.(p+n)}$ having the form

$$J(\mathbf{r}(\mathbf{c})) = \begin{bmatrix} \frac{\partial \mathbf{r}(\mathbf{c})}{\partial \mathbf{a}} & \frac{\partial \mathbf{r}(\mathbf{c})}{\partial \mathbf{b}} \end{bmatrix} = \begin{bmatrix} \frac{\partial \mathbf{e}(\mathbf{A},\mathbf{B})}{\partial \mathbf{a}} & \frac{\partial \mathbf{e}(\mathbf{A},\mathbf{B})}{\partial \mathbf{b}} \end{bmatrix} .$$

Now, using Remark 4.2 and Theorem 4.3, we have:

$$J(\mathbf{r}(\mathbf{c})) = -\begin{bmatrix} \mathbf{K}_{(n,p)}\mathbf{G}(\mathbf{b}) & \mathbf{F}(\mathbf{a}) \end{bmatrix}$$

and

$$\nabla\phi(\mathbf{c}) = \begin{bmatrix} \nabla\varphi^*_{\mathbf{a}}(\mathbf{A},\mathbf{B}) \\ \nabla\varphi^*_{\mathbf{b}}(\mathbf{A},\mathbf{B}) \end{bmatrix} = \begin{bmatrix} \mathbf{G}(\mathbf{b})^T\mathbf{G}(\mathbf{b})\mathbf{a} - \mathbf{G}(\mathbf{b})^T\mathbf{z} \\ \mathbf{F}(\mathbf{a})^T\mathbf{F}(\mathbf{a})\mathbf{b} - \mathbf{F}(\mathbf{a})^T\mathbf{x} \end{bmatrix} .$$

Finally, if we differentiate again $\nabla\phi(\mathbf{c})$ with respect to $\mathbf{c}$, an analytic formulae for the Hessian matrix $\nabla^2\phi(\mathbf{c})$ can be obtained, which is essentially equivalent to the results given in Theorem 4.3, see [15][82] for a derivation of this Hessian matrix. Equipped with these exact formulas for $\nabla\phi(\mathbf{c})$, $J(\mathbf{r}(\mathbf{c}))$ and $\nabla^2\phi(\mathbf{c})$, standard first- and second-order NLLS methods such as the steepest gradient, Gauss-Newton, Levenberg-Marquardt and Newton algorithms [45][139][123] can also be used (and have been used) to minimize directly $\phi(\mathbf{c}) = \varphi^*(\mathbf{A},\mathbf{B})$ or $\phi_\lambda(\mathbf{c}) = \varphi^*_\lambda(\mathbf{A},\mathbf{B})$, and to solve the WLRA problem and its MMMF variant [15][37][81]. However, as it is arguably preferable to keep the dimension of the search space as much low as possible and because the joint optimization strategy of minimizing directly $\phi(.)$ has been found to be much less efficient and less robust than the variable projection framework (based on the minimization of $\psi(.)$) detailed in the next section [37][150][14][81][17], we don't focus here anymore on the direct minimization of $\phi(.)$ or $\varphi^*(.)$ (or alternatively $\phi_\lambda(.)$ or $\varphi^*_\lambda(.)$) for solving the WLRA problem. ■

5 The variable projection framework

We now explain how to minimize the cost function $\psi(.)$, which is used in the (VP1) formulation of the WLRA problem. In addition to the equivalence of the (P1) and (VP1) formulations of the WLRA problem stated in Theorem 3.9, the variable projection approach is further justified by a theorem originally proved by Golub and Pereyra in [63], which shows, under some differentiability conditions, that if $\widehat{\mathbf{a}} = vec(\widehat{\mathbf{A}}^T)$ is a critical point of $\psi(.)$ and $\widehat{\mathbf{B}}$ is calculated by equation (4.1), e.g., by solving n independent linear least-squares problems as described in the beginning of Section 4, then $(\widehat{\mathbf{A}},\widehat{\mathbf{B}})$ is also a first-order critical point of $\varphi^*(.)$. We will give a demonstration of this result later in Theorem 5.7 (see Subsection 5.3) for completeness.

General optimization methods used to minimize a functional like $\psi(.)$ are termed variable projection algorithms and are described in [63][166][95][96][10][65][149]. Their advantages are that they usually solve mixed linear-nonlinear least-squares problems like $\varphi^*(.)$ in less time, fewer function evaluations and better global convergence than standard NLLS codes, and that no starting estimate of the linear variable (e.g., $\mathbf{B}$) is required [136]. In the context of the WLRA or matrix completion problems, they offer also other advantages as shown in [147][150][37][14][81][88] and as we will illustrate in the next sections. However, many of them have also a major drawback as they expand considerably the dimensionality of the WLRA problem (see Subsection 3.4 for details). This limits severely their use for medium and large datasets, which are currently found now in many applications, beyond variations of the variable projection steepest (e.g., gradient) descent method or similar first-order methods [171][46][14][17][146]. In our WLRA context, the simplest variable projection steepest descent method can be written as

$$\mathbf{a}_{i+1} = \mathbf{a}_i - \alpha_i \nabla\psi(\mathbf{a}_i).$$

In words, with this basic method, we move by making a correction step that is proportional to the negative of the gradient of $\psi(.)$ and the positive scalar α_i can be used to control the size of the step without changing its direction [123][146]. This basic method works fine for simple models, but is often too simplistic when there are many parameters to estimate like in our WLRA problem.

Furthermore, its convergence can be very slow without cleaver strategies to control α_i or the use of second-order information, especially in the final stage [139][17][146]. Near a local minimizer, the steepest descent method converges at a linear rate depending on the condition number in a neighborhood of this minimizer. However, this convergence rate deteriorates dramatically when the Hessian of the cost function is ill-conditioned and we will demonstrate later, in Subsection 5.3, that this always the case for the cost function $\psi(.)$. As another illustration, during the iterations, the curvature of $\psi(.)$ is usually not the same in all directions. If there is a long and narrow valley in the values of $\psi(.)$, which is not unusual when the weights are not uniform [171][200], the component of the gradient in the direction that points along the bottom of the valley can be very small while the component perpendicular to the walls of the valley can be quite large even though we have to move a long distance along the base and a small distance perpendicular to the walls to move in the right direction. This is the so-called "error valley" problem, which can be alleviated only if we use some information about the curvature as well as the gradient of $\psi(.)$ in the design of the method [139]. However, second-order derivatives of the cost function are very often prohibitively expensive to compute and we need to find a good compromise between accuracy and speed when the dimensions and the number of variables of the problem are large [123].

Thus, since the convergence of the steepest descent method or its variants, like conjugate gradient methods, can be very slow and second derivatives are expensive to evaluate, we concentrate our attention on (pseudo) second-order or quasi-Newton methods well adapted to NLLS problems [45][139][123][87]. These methods aim to avoid the drawbacks of Newton's methods while maintaining the benefits of using second-order information and introduce also some suitable regularization to cup with the singularity of the Hessian. After a brief description of the Newton, Gauss-Newton, augmented Gauss-Newton and Levenberg-Marquardt algorithms in Subsection 5.1, we give in the next sections a detailed study of the Jacobian matrix $J(\mathbf{r}(\mathbf{a}))$, gradient vector $\nabla\psi(\mathbf{a})$ and Hessian matrix $\nabla^2\psi(\mathbf{a})$, which are pivotal in these variable projection quasi-Newton algorithms and whose specific properties in the context of the WLRA problem have not always been well appreciated in past studies, except in [158][147][150].

5.1 Second-order NLLS optimization methods

As discussed in Subsection 3.4, the minimization of $\psi(.)$ is equivalent to the standard NLLS problem

$$\min_{\mathbf{a}\in\mathbb{R}^{p.k}} \psi(\mathbf{a}) = \frac{1}{2}\|\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\mathbf{x}\|_2^2 = \frac{1}{2}\|\mathbf{r}(\mathbf{a})\|_2^2 = \frac{1}{2}\mathbf{r}(\mathbf{a})^T\mathbf{r}(\mathbf{a}) \ ,$$

where $\mathbf{r}(\mathbf{a}) = \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\mathbf{x}$. Numerous first- and second-order iterative methods are available for minimizing a sum of squares of nonlinear functions such as $\psi(.)$ [45][139][123][87]. However, for finding a solution of our (VP1) problem with these methods, we first note that a certain degree of smoothness of the objective function $\psi(.)$ is required, meaning that $\psi(.)$ must possess one or better two continuous derivatives and the results of Subsection 3.4 show that these smoothness conditions are not systematically verified if $\mathbf{W}$ has some zero elements as the orthogonal projector $\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}$ can be a discontinuous function of $\mathbf{a}$ even if $\mathbf{A} = \mathit{mat}_{k\times p}(\mathbf{a})^T = h(\mathbf{a})$ is of full column rank (see Theorem 3.14 and Corollary 3.4). The degree of smoothness of $\psi(.)$ and $\mathbf{P}^{\perp}_{\mathbf{F}(.)}$ will be further studied in Subsection 5.2, but we note that, despite these caveats, some standard iterative NLLS algorithms have been used very successfully to solve the (VP1) problem even without proper regularization of $\psi(.)$ to insure its smoothness when missing values are present [28][150][66][81][88].

The recommended standard methods are the Gauss-Newton, Levenberg-Marquardt, trust-region Gauss-Newton and augmented Gauss-Newton algorithms if second-order derivatives are difficult or cumbersome to evaluate [45][139][123][87]. All these methods attempt to minimize $\psi(.)$ by finding a zero of $\nabla\psi(.)$, i.e., a point $\mathbf{a} = \mathit{vec}(\mathbf{A}^T)$ such that

$$\nabla\psi(\mathbf{a}) = J\big(\mathbf{r}(\mathbf{a})\big)^T\mathbf{r}(\mathbf{a}) = \mathbf{0}^{k.p} \ .$$

Moreover, all four methods may be interpreted as variations of Newton's method to find a zero of $\nabla\psi(.)$ [45][139][123]. In Newton's method, the correction vector $d\mathbf{a}_n$ for improving an approximate initial solution vector $\mathbf{a}$ of the equation $\nabla\psi(\mathbf{a}) = \mathbf{0}^{k.p}$ is found as the solution to the linear system

$$\nabla^2\psi(\mathbf{a})d\mathbf{a}_n = -J\big(\mathbf{r}(\mathbf{a})\big)^T\mathbf{r}(\mathbf{a}) \;, \tag{5.1}$$

where $\nabla^2\psi(\mathbf{a})$ is the Hessian of $\psi(.)$ at $\mathbf{a}$ given by

$$\nabla^2\psi(\mathbf{a}) = J\big(\mathbf{r}(\mathbf{a})\big)^T J\big(\mathbf{r}(\mathbf{a})\big) + \sum_{l=1}^{n.p}\mathbf{r}_l(\mathbf{a})\nabla^2\mathbf{r}_l(\mathbf{a}) \;. \tag{5.2}$$

In this last equation, $\nabla^2\mathbf{r}_l(\mathbf{a})$ is the Hessian matrix of the l^{th} component of the residual functional $\mathbf{r}(\mathbf{a})$ (i.e., $\mathbf{r}_l(\mathbf{a})$), which is a $p.k \times p.k$ symmetric matrix. The Newton method is based on the second-order Taylor expansion of $\psi(.)$ in a neighborhood of the current iterate $\mathbf{a}$ (see equation (2.43) in Subsection 2.4), e.g.,

$$\psi(\mathbf{a} + d\mathbf{a}) \approx N(d\mathbf{a}) = \psi(\mathbf{a}) + d\mathbf{a}^T\nabla\psi(\mathbf{a}) + \frac{1}{2}d\mathbf{a}^T\nabla^2\psi(\mathbf{a})d\mathbf{a} \;.$$

More precisely, the Newton method attempts to minimize $\psi(.)$ at each iteration by finding a first-order stationary point $d\mathbf{a}_n$ of this quadratic model $N(.)$. Setting the gradient of $N(.)$ to zero, e.g., $\nabla N(d\mathbf{a}_n) = \mathbf{0}^{k.p}$, we obtain the following equation

$$\nabla\psi(\mathbf{a}) + \nabla^2\psi(\mathbf{a})d\mathbf{a}_n = \mathbf{0}^{k.p} \;,$$

from which we derived immediately equation (5.1) defining the Newton iteration. Moreover, if the Hessian matrix $\nabla^2\psi(\mathbf{a})$, which is also equal to $\nabla^2 N(d\mathbf{a}_n)$, is positive definite then $d\mathbf{a}_n$ is a strict global minimizer of $N(.)$ and in a descent direction for $\psi(.)$. In other words, the Newton iteration is well defined as soon as $\nabla^2\psi(\mathbf{a})$ is positive definite, but runs into troubles when it is not, for example in regions of mixed curvature of $\psi(.)$. It may even happen during the iterations that $\nabla^2\psi(\mathbf{a})$ becomes definite negative in which case $d\mathbf{a}_n$ will be a strict global maximizer of $N(.)$ instead of a minimizer, which is a major drawback of the basic Newton method and explains why it lacks global convergence [123][87]. Moreover, since Newton's method requires the computation of second-order derivatives, which can be cumbersome for large-scale problems (see equation (5.2)), it is rarely used in practice despite its quadratic convergence in a neighborhood of a first-order critical point of $\psi(.)$ [45][139][123].

Importantly, the smallest eigenvalue of the positive (semi-definite) matrix $J(\mathbf{r}(\mathbf{a}))^T J(\mathbf{r}(\mathbf{a}))$ can be used to assess the relative importance of the two terms in $\nabla^2\psi(\mathbf{a})$ [87]. More precisely, if for all $\mathbf{a}$ in a neighborhood of a minimizer of $\psi(.)$, the quantities $|\mathbf{r}_l(\mathbf{a})|\|\nabla^2\mathbf{r}_l(\mathbf{a})\|_2$ for $l = 1, \cdots, n.p$ are small relative to this eigenvalue, the term $J(\mathbf{r}(\mathbf{a}))^T J(\mathbf{r}(\mathbf{a}))$ will dominate the Hessian matrix [87]. Now, depending on the relative importance of these two terms in $\nabla^2\psi(\mathbf{a})$, the recommended methods are the Gauss-Newton, Levenberg-Marquardt, trust-region Gauss-Newton and augmented Gauss-Newton algorithms, which involve different approximations of the second term in the Hessian of $\psi(.)$.

The Gauss-Newton method approximates $\nabla^2\psi(\mathbf{a})$ with $J(\mathbf{r}(\mathbf{a}))^T J(\mathbf{r}(\mathbf{a}))$, i.e., drops the second term of the Hessian of $\psi(.)$, which contains products of the $\mathbf{r}_l(\mathbf{a})$ functions and their second-order derivatives. This approximation is exact only if the residual function $\mathbf{r}(.)$ is linear in $\mathbf{a}$, which is usually valid only in a neighborhood of a minimum of $\psi(.)$. The Gauss-Newton method is intended for problems in which the second term of the Hessian matrix is small relative to the first term. Thus, this Gauss-Newton approximation is based on the assumptions that the functions $\mathbf{r}_l(\mathbf{a})$ have small curvatures or that near the solution the magnitudes of the $\mathbf{r}_l(\mathbf{a})$ functions are small. If these conditions are satisfied the Gauss-Newton method will ultimately converge at the same rate as Newton's method despite full second-order derivatives are not used. In Gauss-Newton's method, the correction vector $d\mathbf{a}_{gn}$ for improving an approximate solution is then found as the solution to the linear system of equations

$$J\big(\mathbf{r}(\mathbf{a})\big)^T J\big(\mathbf{r}(\mathbf{a})\big)d\mathbf{a}_{gn} = -J\big(\mathbf{r}(\mathbf{a})\big)^T\mathbf{r}(\mathbf{a}) \;. \tag{5.3}$$

The Gauss-Newton method can be also introduced by a linearization argument. If, given $\mathbf{a} \in \mathbb{R}^{p.k}$, we could solve the problem

$$\min_{d\mathbf{a}\in\mathbb{R}^{p.k}} \psi(\mathbf{a}+d\mathbf{a}) = \frac{1}{2}\|\mathbf{r}(\mathbf{a}+d\mathbf{a})\|_2^2 = \frac{1}{2}\mathbf{r}(\mathbf{a}+d\mathbf{a})^T\mathbf{r}(\mathbf{a}+d\mathbf{a}) \;,$$

then $\mathbf{a}+d\mathbf{a}$ is a minimizer of $\psi(.)$. Since $\mathbf{r}(.)$ is a nonlinear residual function, we must seek an approximate solution that can be improved iteratively. A natural way to find an approximate solution is to linearize the residual function $\mathbf{r}(.)$ around $\mathbf{a}$. If we assume that $\mathbf{r}(.)$ is twice continuously differentiable at $\mathbf{a} \in \mathbb{R}^{p.k}$, we have the first-order Taylor expansion

$$\mathbf{r}(\mathbf{a}+d\mathbf{a}) = \mathbf{r}(\mathbf{a}) + J(\mathbf{r}(\mathbf{a}))d\mathbf{a} + \mathcal{O}(\|d\mathbf{a}\|_2^2)$$

and if we substitute this Taylor approximation for $\mathbf{r}(\mathbf{a}+d\mathbf{a})$ in the definition of $\psi(\mathbf{a}+d\mathbf{a})$, this leads to the quadratic function approximation

$$\psi(\mathbf{a}+d\mathbf{a}) \approx G(d\mathbf{a}) = \psi(\mathbf{a}) + d\mathbf{a}^T J\big(\mathbf{r}(\mathbf{a})\big)^T\mathbf{r}(\mathbf{a}) + \frac{1}{2}d\mathbf{a}^T J\big(\mathbf{r}(\mathbf{a})\big)^T J\big(\mathbf{r}(\mathbf{a})\big)d\mathbf{a} \;,$$

which must be minimized at each iteration. As a model for the change of the cost function $\psi(.)$, the quadratic function $G(.)$ has two important advantages compared to the Newton quadratic model $N(.)$, first, it involves only first derivatives of the residual function $\mathbf{r}(.)$ and, second, the symmetric matrix $J(\mathbf{r}(\mathbf{a}))^T J(\mathbf{r}(\mathbf{a}))$ is always positive semi-definite and is positive definite if $J(\mathbf{r}(\mathbf{a}))$ is of full column rank. Of course, the drawback is a lost of accuracy as full second-order information from the Hessian matrix is not taken into account. The gradient of this quadratic function is equal to

$$\nabla G(d\mathbf{a}) = J\big(\mathbf{r}(\mathbf{a})\big)^T\mathbf{r}(\mathbf{a}) + J\big(\mathbf{r}(\mathbf{a})\big)^T J\big(\mathbf{r}(\mathbf{a})\big)d\mathbf{a} \;,$$

and setting it to zero leads to the linear system (5.3), which is also the normal equations of the linear least-squares problem

$$d\mathbf{a}_{gn} = \text{Arg}\min_{d\mathbf{a}\in\mathbb{R}^{p.k}} \frac{1}{2}\|\mathbf{r}(\mathbf{a}) + J\big(\mathbf{r}(\mathbf{a})\big)d\mathbf{a}\|_2^2 \;, \tag{5.4}$$

whose unique solution is

$$d\mathbf{a}_{gn} = -\Big(J\big(\mathbf{r}(\mathbf{a})\big)^T J\big(\mathbf{r}(\mathbf{a})\big)\Big)^{-1} J\big(\mathbf{r}(\mathbf{a})\big)^T\mathbf{r}(\mathbf{a}) \;,$$

if $J(\mathbf{r}(\mathbf{a}))$ has full column rank or, if this Jacobian matrix is rank-deficient or ill-conditioned, whose unique minimum 2-norm solution is

$$d\mathbf{a}_{gn} = -J\big(\mathbf{r}(\mathbf{a})\big)^+\mathbf{r}(\mathbf{a}) \;, \tag{5.5}$$

where $J(\mathbf{r}(\mathbf{a}))^+$ is the pseudo-inverse of the Jacobian matrix of the residual function $\mathbf{r}(.)$ at $\mathbf{a}$. In the rest of this monograph, we will mostly use the pseudo-inverse notation $J(\mathbf{r}(\mathbf{a}))^+$ to indicate that the normal equations shall not be used to compute $d\mathbf{a}_{gn}$ if $J(\mathbf{r}(\mathbf{a}))$ is ill-conditioned or singular.

The linear least-squares problem (5.4) can be solved by stable orthogonalization methods or the SVD decomposition of the Jacobian matrix $J(\mathbf{r}(\mathbf{a}))$, see Subsection 2.1 and [71][87] for details. Thus, the last equation becomes the iteration formula

$$\mathbf{a}_{i+1} = \mathbf{a}_i - J\big(\mathbf{r}(\mathbf{a}_i)\big)^+\mathbf{r}(\mathbf{a}_i) \;, \tag{5.6}$$

which is known as the Gauss-Newton algorithm. Given an initial estimate $\mathbf{a}_0$, the linear least-squares problem (5.4) associated with the Taylorized equations are solved to yield a correction to this vector $\mathbf{a}_0$. This process is repeated and stops if and when the vectors $\mathbf{a}_i$ (or the values $\psi(\mathbf{a}_i)$) converge or the norm of $\nabla\psi(\mathbf{a}_i)$ is sufficiently small to assume that we have reached a stationary point of $\psi(.)$.

Of course, the linearization argument used to derive the Gauss-Newton iteration is only valid in a neighborhood of $\mathbf{a}_i$ and it may happens that $\psi(\mathbf{a}_{i+1}) > \psi(\mathbf{a}_i)$ meaning that the Gauss-Newton algorithm may compute bad corrections by taking steps that are too long, reaching points outside the region of validity of the affine model used to approximate $\mathbf{r}(.)$ around $\mathbf{a}_i$. Several cleaver variants have been proposed to overcome this problem in practice.

The first one is the damped Gauss-Newton algorithm which is defined as

$$\mathbf{a}_{i+1} = \mathbf{a}_i - \alpha_i J\big(\mathbf{r}(\mathbf{a}_i)\big)^+ \mathbf{r}(\mathbf{a}_i) \; .$$

In this equation, α_i is a damping parameter which is chosen at each iteration to make the algorithm a descent method (i.e, such that $\psi(\mathbf{a}_{i+1}) < \psi(\mathbf{a}_i)$). The Gauss-Newton approximation of the Hessian is always positive semi-definite and it is positive definite, if and only if, the Jacobian matrix has full column rank, and, in this case, $d\mathbf{a}_{gn}$ is the unique solution of the above linear least-squares problem and is also in a descent direction for $\psi(.)$ if $d\mathbf{a}_{gn} \neq \mathbf{0}^{k.p}$ or, equivalently, if $\nabla\psi(\mathbf{a}) = J(\mathbf{r}(\mathbf{a}))^T\mathbf{r}(\mathbf{a}) \neq \mathbf{0}^{k.p}$ since in these conditions

$$0 < d\mathbf{a}_{gn}^T J\big(\mathbf{r}(\mathbf{a})\big)^T J\big(\mathbf{r}(\mathbf{a})\big) d\mathbf{a}_{gn} = -d\mathbf{a}_{gn}^T J\big(\mathbf{r}(\mathbf{a})\big)^T \mathbf{r}(\mathbf{a}) = -d\mathbf{a}_{gn}^T \nabla\psi(\mathbf{a}) \; .$$

This shows that, when the Jacobian matrix has full column rank, the Gauss-Newton method can always be complemented with a line search in order to enforce the descending condition $\psi(\mathbf{a}_{i+1}) < \psi(\mathbf{a}_i)$ during the iterations [139][123][87]. Here, $\mathbf{a}_{i+1} = \mathbf{a}_i + \widehat{\alpha} d\mathbf{a}_{gn}$ and $\widehat{\alpha}$ is found as a (approximate) solution to the problem

$$\widehat{\alpha} \approx \text{Arg}\min_{\alpha>0} \; \psi(\mathbf{a}_i + \alpha d\mathbf{a}_{gn}) \; .$$

Many strategies have been proposed to choose the damping parameter $\widehat{\alpha}$ [45][139]. The Gauss-Newton method with a line search can be shown to have guaranteed convergence, provided that the level set $\big\{\mathbf{a} \in \mathbb{R}^{p.k} \mid \psi(\mathbf{a}) \leq \psi(\mathbf{a}_0)\big\}$ is bounded, and the Jacobian matrix $J(\mathbf{r}(\mathbf{a}_i))$ has full rank in all iterations [45][139]. Practical experience shows that the Gauss-Newton method may fail with or without a line search and that it usually has only linear convergence as opposed to the Newton's method, which exhibits quadratic convergence near a solution vector $\widehat{\mathbf{a}}$. However, if, at a solution $\widehat{\mathbf{a}}$, we have $\mathbf{r}(\widehat{\mathbf{a}}) = \mathbf{0}^{k.p}$, then we have the equality

$$\nabla^2\psi(\widehat{\mathbf{a}}) = J\big(\mathbf{r}(\widehat{\mathbf{a}})\big)^T J\big(\mathbf{r}(\widehat{\mathbf{a}})\big)$$

and we can also get quadratic convergence with the Gauss-Newton method. Similarly, if the component residual functions $\mathbf{r}_l(.)$ have small curvatures or if the $|\mathbf{r}_l(\widehat{\mathbf{a}})|$ are small, we can also get superlinear convergence. For example, this will be the case for the matrix completion problem. This can also be observed if the values of the residual matrix $\mathbf{X} - \widehat{\mathbf{A}}\widehat{\mathbf{B}}$ behave like white noise, as in this case we can expect partial canceling in the sum

$$\sum_{l=1}^{n.p} \mathbf{r}_l(\widehat{\mathbf{a}}) \nabla^2 \mathbf{r}_l(\widehat{\mathbf{a}}) \; ,$$

in which case, we also get

$$\nabla^2\psi(\widehat{\mathbf{a}}) \approx J\big(\mathbf{r}(\widehat{\mathbf{a}})\big)^T J\big(\mathbf{r}(\widehat{\mathbf{a}})\big) \; .$$

This situation also occurs in many applications, especially in climate science.

When the Hessian matrix $\nabla^2\psi(\mathbf{a})$ is positive definite, the full Newton direction $d\mathbf{a}_n$ is also a descent direction for $\psi(.)$ and, in this case, the full Newton method can also be complemented by a line search to enforce the descending condition $\psi(\mathbf{a}_{i+1}) < \psi(\mathbf{a}_i)$ [45][139][123][87]. However, contrary to the Gauss-Newton approximation of the Hessian, which is always positive semi-definite, the full Hessian matrix can be indefinite in some regions of mixed curvature of the search space or

even negative definite, in which cases, further regularization of the Hessian matrix, such that inflating its diagonal elements, is required to transform it in a positive definite matrix before applying a line search (see [139][123] and Subsection 6.3 for more details).

The second modification of the Gauss-Newton algorithm used in practice is the Levenberg-Marquardt method. This method approximates the second term in the Hessian of $\psi(.)$ with $\lambda.\mathbf{D}^T\mathbf{D}$ where $\mathbf{D}$ is a full rank matrix and λ a strictly positive real scalar (the Marquardt damping parameter). The standard choice for $\mathbf{D}$ is the identity matrix or a diagonal matrix $\mathbf{D} = diag(\mathbf{d})$ with appropriately chosen components $\mathbf{d}_j > 0$, which take into account the scaling of the problem and can be kept fixed or changed during the iterations [122][45][139][123]. In all cases, this implies that the approximate Hessian matrix

$$J\big(\mathbf{r}(\mathbf{a})\big)^T J\big(\mathbf{r}(\mathbf{a})\big) + \lambda.\mathbf{D}^T\mathbf{D}$$

is positive definite if $\lambda > 0$. Thus, the Levenberg-Marquardt's method is based on the following quadratic approximation model

$$\psi(\mathbf{a} + d\mathbf{a}) \approx L_\lambda(d\mathbf{a}) = \psi(\mathbf{a}) + d\mathbf{a}^T J\big(\mathbf{r}(\mathbf{a})\big)^T \mathbf{r}(\mathbf{a}) + \frac{1}{2} d\mathbf{a}^T \Big(J\big(\mathbf{r}(\mathbf{a})\big)^T J\big(\mathbf{r}(\mathbf{a})\big) + \lambda.\mathbf{D}^T\mathbf{D} \Big) d\mathbf{a}$$

and the correction vector $d\mathbf{a}_{lm}$ for improving an approximate solution $\mathbf{a}$ is found as the solution of the regularized normal system

$$\Big(J\big(\mathbf{r}(\mathbf{a})\big)^T J\big(\mathbf{r}(\mathbf{a})\big) + \lambda.\mathbf{D}^T\mathbf{D} \Big) d\mathbf{a}_{lm} = -J\big(\mathbf{r}(\mathbf{a})\big)^T \mathbf{r}(\mathbf{a}) \tag{5.7}$$

and is always in a descent direction for $\psi(.)$, even when $J(\mathbf{r}(\mathbf{a}))$ is not of full column rank, if $\lambda > 0$. Rather than dividing the steps when $\psi(\mathbf{a}_{i+1}) > \psi(\mathbf{a}_i)$ as in the damped Gauss-Newton method, the Levenberg-Marquardt algorithm deflates the steps by inflating the diagonals of the cross-product Jacobian matrix (which is equivalent to shift positively its spectrum) before inverting it to solve for the correction vector. It may be demonstrated that a sufficiently large λ always exists such that $\psi(\mathbf{a}_{i+1}) < \psi(\mathbf{a}_i)$ will be satisfied unless $\mathbf{a}_i$ is already a stationary point of $\psi(.)$ [45][139][123][87].

In other words, the Marquardt damping parameter λ controls the nature of the iterations and limits the size of $d\mathbf{a}_{lm}$ at the same time. If we assume that $\mathbf{D}$ is the identity matrix and λ is very large, then

$$d\mathbf{a}_{lm} \approx -\frac{1}{\lambda}.J\big(\mathbf{r}(\mathbf{a})\big)^T \mathbf{r}(\mathbf{a}) = -\frac{1}{\lambda}.\nabla\psi(\mathbf{a})$$

is a short step in a direction very close to the steepest descent direction. If, on the other hand, λ is very small, then $L_\lambda(\mathbf{a}) \approx G(\mathbf{a})$ and $d\mathbf{a}_{lm}$ is close to the Gauss-Newton step $d\mathbf{a}_{gn}$ described above. In other words, we can think of the Levenberg-Marquardt method as a hybrid method between the steepest descent and Gauss-Newton methods with the good performance of the steepest descent method in the initial stage and the faster convergence of the Gauss-Newton method at the final stage of the iterative process, assuming that the value of the Marquardt damping parameter decreases during the iterative process. Taking $\mathbf{D}$ as the identity matrix corresponds to the algorithm originally proposed by Levenberg [107]. Later, Marquardt [120] improved the method by choosing the diagonals of $\mathbf{D}$ to match the 2-norms of the columns of the Jacobian matrix $J(\mathbf{r}(\mathbf{a}))$. This makes the algorithm invariant under diagonal scaling of the elements of the vector $\mathbf{a}$ [122][139]. This also allows to include local curvature information, even when λ is large and we are essentially moving in the (negative) steepest gradient direction. This is, for example, useful to alleviate the "error valley" problem affecting the steepest gradient method discussed at the beginning of this section since, in that case, we are moving further in the directions in which the gradient is smaller. Later, many other choices for $\mathbf{D}$ have been proposed and tested [122][49][45].

The above equations defining the Levenberg-Marquardt's correction vector $d\mathbf{a}_{lm}$ are the normal equations for the regularized linear least-squares problem

$$\min_{d\mathbf{a}\in\mathbb{R}^{p.k}} \frac{1}{2} \left\| \begin{bmatrix} \mathbf{r}(\mathbf{a}) \\ \mathbf{0}^{p.k} \end{bmatrix} + \begin{bmatrix} J\big(\mathbf{r}(\mathbf{a})\big) \\ \sqrt{\lambda}.\mathbf{D} \end{bmatrix} d\mathbf{a} \right\|_2^2 = \frac{1}{2} \left\| \mathbf{r}(\mathbf{a}) + J\big(\mathbf{r}(\mathbf{a})\big) d\mathbf{a} \right\|_2^2 + \frac{\lambda}{2} \left\| \mathbf{D} d\mathbf{a} \right\|_2^2, \tag{5.8}$$

which can also be solved accurately by stable methods as for the Gauss-Newton correction and there is no need to form nor to invert the symmetric matrix $J(\mathbf{r}(\mathbf{a}))^T J(\mathbf{r}(\mathbf{a})) + \lambda.\mathbf{D}^T\mathbf{D}$ [87][139]. Moreover, this linear least-squares problem has always a unique solution if $\lambda > 0$.

The Levenberg-Marquardt algorithm is often considered superior to the (damped) Gauss-Newton algorithm since it is well defined even when the Jacobian matrix is rank deficient. Another advantage is that the Levenberg-Marquardt correction assures an optimal interpolation between a Gauss-Newton step and the steepest descent direction (e.g., negative gradient direction) when the Gauss-Newton step is much too long.

Similarly, we can define a Levenberg-Marquardt variant of the Newton method by computing the correction vector $d\mathbf{a}_n$ as

$$\big(\nabla^2\psi(\mathbf{a}) + \lambda.\mathbf{I}_{k.p}\big)\, d\mathbf{a}_n = -J\big(\mathbf{r}(\mathbf{a})\big)^T\mathbf{r}(\mathbf{a})\ , \tag{5.9}$$

where the term $\lambda.\mathbf{I}_{k.p}$ is included when $\nabla^2\psi(\mathbf{a})$ is not positive definite and hence the Newton direction may not be a descent direction. In such conditions, it is always possible to choose λ sufficiently large such that, first, the matrix $\nabla^2\psi(\mathbf{a}) + \lambda.\mathbf{I}_{k.p}$ becomes positive definite and, second, $\psi(\mathbf{a}+d\mathbf{a}_n) < \psi(\mathbf{a})$ [139][123]. This strategy is based on the quadratic approximation model

$$\psi(\mathbf{a}+d\mathbf{a}) \approx N_\lambda(d\mathbf{a}) = \psi(\mathbf{a}) + d\mathbf{a}^T J\big(\mathbf{r}(\mathbf{a})\big)^T\mathbf{r}(\mathbf{a}) + \frac{1}{2}d\mathbf{a}^T\left(\nabla^2\psi(\mathbf{a}) + \lambda.\mathbf{I}_{k.p}\right)d\mathbf{a}\ .$$

As in the Levenberg-Marquardt algorithm, the damping parameter λ can be used to control both the size and direction of the correction vector $d\mathbf{a}_n$ and, in this case, we can avoid the use of a line search to control the step size in order to get reasonable convergence in the Newton method. Thus, we can also think of this Levenberg-Marquardt variant of the Newton method as an hybrid between the steepest descent and Newton methods with the good performance of the steepest descent method in the initial stage, but the quadratic convergence of the Newton method at the final stage [139][123]. See the variable projection Newton algorithms (5), (6) and (7) described in Subsection 6.3, which all integrate a damping term $\lambda.\mathbf{I}_{k.p}$ for some illustrations of this simple strategy in the context of the Newton method applied to the WLRA problem.

A variation of the Levenberg-Marquardt method is the trust-region Gauss-Newton algorithm where the correction vector $d\mathbf{a}_{t-gn}$ is defined as the solution of the constrained linear least-squares problem

$$d\mathbf{a}_{t-gn} = \text{Arg}\min_{d\mathbf{a}\in\mathbb{R}^{p.k}} \frac{1}{2}\big\|\mathbf{r}(\mathbf{a}) + J\big(\mathbf{r}(\mathbf{a})\big)d\mathbf{a}\big\|_2^2 \text{ subject to } \big\|\mathbf{D}d\mathbf{a}\big\|_2 \leq \delta\ .$$

Here, the set of feasible correction vectors $d\mathbf{a}$ is restricted to the ellipsoid $\{d\mathbf{a} \in \mathbb{R}^{p.k} / \|\mathbf{D}d\mathbf{a}\|_2 \leq \delta\}$ which is called the trust region. $\delta > 0$ is the trust region radius, which controls the size of the trust region and is updated recursively during the iterative process [45][139]. In this class of methods, the scaling matrix $\mathbf{D}$ generates the elliptic norm $\|d\mathbf{a}\|_\mathbf{D} = \|\mathbf{D}d\mathbf{a}\|_2$ in which the correction vector is measured [139]. The trust region can then be thought of as a region of trust for the linear model

$$\mathbf{r}(\mathbf{a}+d\mathbf{a}) \approx \mathbf{r}(\mathbf{a}) + J\big(\mathbf{r}(\mathbf{a})\big)d\mathbf{a}$$

and the idea in the trust-region Gauss-Newton method is to avoid using this linear model outside its range of validity. Note that the Gauss-Newton step $d\mathbf{a}_{gn}$ solves this constrained problem if $\|\mathbf{D}d\mathbf{a}_{gn}\|_2 \leq \delta$. Otherwise, it can be shown that the trust-region Gauss-Newton correction vector is the unique solution $d\mathbf{a}(\lambda)$ of the unconstrained regularized linear least-squares problem

$$d\mathbf{a}(\lambda) = \text{Arg}\min_{d\mathbf{a}\in\mathbb{R}^{p.k}} \frac{1}{2}\big\|\mathbf{r}(\mathbf{a}) + J\big(\mathbf{r}(\mathbf{a})\big)d\mathbf{a}\big\|_2^2 + \frac{\lambda}{2}\big\|\mathbf{D}d\mathbf{a}\ ,\big\|_2^2\ ,$$

where $\lambda > 0$ is determined from the scalar equation $\|\mathbf{D}d\mathbf{a}(\lambda)\|_2 = \delta$ which is nonlinear in λ. In other words, when the correction vector $d\mathbf{a}$ is directly controlled by the Marquardt damping parameter λ and not by δ, we obtain the Levenberg-Marquardt algorithm, otherwise we have a trust

region Gauss-Newton algorithm [139]. We also observe that if $\mathbf{D}$ is nonsingular then a change of variables yields an equivalent linear least-squares problem with $\mathbf{D} = \mathbf{I}$ for computing both the Levenberg-Marquardt and trust-region Gauss-Newton corrections.

Finally, the augmented Gauss-Newton method partly takes second-order derivatives into account by approximating the second term of the Hessian of $\psi(.)$ by either finite differencing or a quasi-Newton update in order to improve the above NLLS methods in the large residuals case [49][45][139][123]. The variable projection quasi-Newton algorithms discussed in Subsection 6.3 belong to this class of methods.

What has been described so far is well-known. In the following subsections, let us quantify the smoothness of $\psi(.)$ in more details and study the specific properties of the Jacobian matrix $J(\mathbf{r}(\mathbf{a}))$, Hessian matrix $\nabla^2\psi(\mathbf{a})$ and vector gradient $\nabla\psi(\mathbf{a})$, which need to be evaluated in the above second-order or pseudo second-order NLLS algorithms. Implementation details of these variable projection NLLS algorithms will be presented in Section 6 after their main properties have been derived in the rest of this section. For small or medium sized NLLS or WLRA problems, the above methods will be much faster than variants of the steepest gradient method. However, for larger problems, the cost of solving a linear least-squares problem or a linear system with a huge coefficient matrix at each iteration is a major drawback compared to the (steepest) gradient methods as the time spent in each iteration scales as $\mathcal{O}((k.p)^3)$ and, thus, increases considerably for large and square data matrices and a large value of the k parameter. Taking these difficulties in consideration, we propose also some parallel implementations of all our variable projection NLLS algorithms for the WLRA problem in Section 6 so that they can also be used for larger sized problems also found now in many practical applications.

5.2 Computation and properties of the Jacobian matrix

In order to use a (damped or trust-region) Gauss-Newton or Levenberg-Marquardt algorithm for minimizing $\psi(.)$ (and solve the WLRA problem), we must compute the Jacobian of the residual function

$$\mathbf{r}(\mathbf{a}) = \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\mathbf{x} = (\mathbf{I}_{n.p} - \mathbf{P}_{\mathbf{F}(\mathbf{a})})\mathbf{x}\,,$$

defined in equation (3.24). This requires computing the derivative of the orthogonal projector $\mathbf{P}_{\mathbf{F}(.)}$ with respect to $\mathbf{a} \in h^{-1}(\mathbb{R}^{p\times k}_k)$ as shown in Subsection 3.4. If $\mathbf{a} \in h^{-1}(\mathbb{R}^{p\times k}_{<k})$, keep in mind that $\mathbf{P}_{\mathbf{F}(.)}$ is not even continuous at $\mathbf{a}$ (see Theorems 3.11 and 3.12) and cannot be differentiable either at this point.

A close formula for the derivative of orthogonal projectors has been derived first by Golub and Pereyra [63] and Decell [38] under the assumption that $\mathbf{F}(.)$ is of local constant rank at any point $\mathbf{a}$ (this means that $\mathbf{F}(.)$ is of constant rank in a neighborhood of $\mathbf{a}$, but not necessarily of full column-rank, see Definition 3.1 for details) in which differentiation is to be performed as stated in the following theorem, which extends the results about the continuity of $\mathbf{P}_{\mathbf{F}(.)}$ given in Theorem 3.10:

Theorem 5.1. Let $\Phi(.)$ be a matrix function : $\mathbb{R}^m \longrightarrow \mathbb{R}^{l\times t}$, which is q times continuously differentiable at a point $\mathbf{a} \in \mathbb{R}^m$. The following conditions are equivalent:

1) $\Phi(.)$ has a local constant rank at $\mathbf{a}$,
2) $\Phi(.)^+$ is q times continuously differentiable at $\mathbf{a}$,
3) $\Phi(.)\Phi(.)^+ = \mathbf{P}_{\Phi(.)}$ is q times continuously differentiable at $\mathbf{a}$,
4) $\Phi(.)^+\Phi(.)$ is q times continuously differentiable at $\mathbf{a}$.

In other words, the differentiability of the pseudo-inverse of a matrix function $\Phi(.)$ at a point $\mathbf{a} \in \mathbb{R}^m$ is equivalent to the differentiability of the orthogonal projectors onto the column or row spaces of this matrix function at $\mathbf{a}$ and all these conditions are equivalent to the assertion that this matrix

function has local constant rank at $\mathbf{a}$ if $\Phi(.)$ is itself differentiable at $\mathbf{a}$. Furthermore, in these conditions, we have for any point $\mathbf{a} \in \mathbb{R}^m$ for which $\Phi(.)$ is differentiable

$$D(\mathbf{P}_{\Phi(\mathbf{a})}) = \mathbf{P}^{\perp}_{\Phi(\mathbf{a})} D(\Phi(\mathbf{a}))\Phi(\mathbf{a})^+ + \big(\mathbf{P}^{\perp}_{\Phi(\mathbf{a})} D(\Phi(\mathbf{a}))\Phi(\mathbf{a})^+\big)^T \tag{5.10}$$

and

$$\begin{aligned} D(\Phi(\mathbf{a})^+) = &- \Phi(\mathbf{a})^+ D(\Phi(\mathbf{a}))\Phi(\mathbf{a})^+ + \Phi(\mathbf{a})^+(\Phi(\mathbf{a})^+)^T D(\Phi(\mathbf{a})^T)\mathbf{P}^{\perp}_{\Phi(\mathbf{a})} \\ &+ \big(\mathbf{I}_m - \Phi(\mathbf{a})^+\Phi(\mathbf{a})\big) D(\Phi(\mathbf{a})^T)(\Phi(\mathbf{a})^+)^T \Phi(\mathbf{a})^+ \, . \end{aligned} \tag{5.11}$$

Finally, note that, in the above equation defining the differential of the orthogonal projector $\mathbf{P}_{\Phi(.)}$, we can substitute in place of the pseudo-inverse $\Phi(\mathbf{a})^+$ any symmetric generalized inverse $\Phi(\mathbf{a})^-$ as defined in equations (2.10) or (2.19) of Subsection 2.1.

Proof. See Theorems 8.4 and 8.5 in Chapter 8 of [124] and also [63][64][38][29].

□

As noted already in Subsection 3.4, $\mathbf{F}(.)$ is a continuous linear mapping from $\mathbb{R}^{p.k}$ into $\mathbb{R}^{n.p \times n.k}$ (since the *mat* and transpose operators are linear mappings and the Kronecker product is a bilinear operator) and is, thus, continuously and infinitely differentiable at any point $\mathbf{a} \in \mathbb{R}^{p.k}$ [26]. Collecting the results from Theorems 3.10 and 5.1, we then deduce that the proposition that $\mathbf{P}_{\mathbf{F}(.)}$ is infinitely differentiable (e.g., of class C^∞) at a point $\mathbf{a} \in h^{-1}(\mathbb{R}^{p \times k}_k)$ is equivalent to its continuity at this point and to the proposition that $\mathbf{F}(.)$ is of constant rank in a neighborhood of $\mathbf{a}$. Next, using Theorem 3.11, we obtain the following corollary in the case where $\mathbf{W} \in \mathbb{R}^{p \times n}_{+*}$:

Corollary 5.1. For $\mathbf{X} \in \mathbb{R}^{p \times n}$ and $\mathbf{W} \in \mathbb{R}^{p \times n}_{+*}$, and any fixed integer $k \leq \mathit{rank}(\mathbf{X})$, the matrix function $\mathbf{P}_{\mathbf{F}(.)}$ from $\mathbb{R}^{p.k}$ to $\mathbb{R}^{p.n \times p.n}$ defined by

$$\mathbf{a} \mapsto \mathbf{P}_{\mathbf{F}(\mathbf{a})} = \mathbf{F}(\mathbf{a})\mathbf{F}(\mathbf{a})^+ \, ,$$

where $\mathbf{F}(\mathbf{a})^+$ is the pseudo-inverse of $\mathbf{F}(\mathbf{a})$ and $\mathbf{F}(\mathbf{a})$ is the $p.n \times n.k$ block diagonal matrix

$$\mathbf{F}(\mathbf{a}) = \bigoplus_{j=1}^{n} \mathit{diag}(\sqrt{\mathbf{W}}_{.j}) h(\mathbf{a}) = \bigoplus_{j=1}^{n} \mathit{diag}(\sqrt{\mathbf{W}}_{.j}) \mathbf{A} \, ,$$

is of class C^∞ (e.g., infinitely differentiable) at all points $\mathbf{a} \in h^{-1}(\mathbb{R}^{p \times k}_k)$. Furthermore, for all points $\mathbf{a} \in h^{-1}(\mathbb{R}^{p \times k}_k)$, we have

$$D(\mathbf{P}_{\mathbf{F}(\mathbf{a})}) = \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})} D\big(\mathbf{F}(\mathbf{a})\big)\mathbf{F}(\mathbf{a})^+ + \big(\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})} D\big(\mathbf{F}(\mathbf{a})\big)\mathbf{F}(\mathbf{a})^+\big)^T \, . \tag{5.12}$$

Here, as in equation (5.10) of Theorem 5.1, we can substitute in place of $\mathbf{F}(\mathbf{a})^+$ any symmetric generalized inverse $\mathbf{F}(\mathbf{a})^-$ as defined in equations (2.10) or (2.19) of Subsection 2.1.

As expected from Corollary 3.4, the situation is much less favourable when $\mathbf{W}$ has some zero elements, as the condition that $\mathbf{a} \in h^{-1}(\mathbb{R}^{p \times k}_k)$ is not sufficient to ensure that $\mathbf{F}(.)$ is of constant rank in a neighborhood of $\mathbf{a}$ and, thus, that $\mathbf{P}^{\perp}_{\mathbf{F}(.)}$ is differentiable at $\mathbf{a}$ in such situation:

Corollary 5.2. For $\mathbf{X} \in \mathbb{R}^{p \times n}$ and $\mathbf{W} \in \mathbb{R}^{p \times n}_{+}$, and any fixed integer $k \leq \mathit{rank}(\mathbf{X})$, the matrix function $\mathbf{P}_{\mathbf{F}(.)}$ from $\mathbb{R}^{p.k}$ into $\mathbb{R}^{p.n \times p.n}$ defined by

$$\mathbf{a} \mapsto \mathbf{P}_{\mathbf{F}(\mathbf{a})} = \mathbf{F}(\mathbf{a})\mathbf{F}(\mathbf{a})^+$$

is not differentiable at all points $\mathbf{a} \in \bigcup_{j=1}^{n} \mathcal{B}_j$, where $\mathcal{B}_j$ is the j^{th} barrier set associated with the j^{th} atomic and matrix functions, $\psi_j(.)$ and $\mathbf{F}_j(.)$, as defined, respectively, in equation (3.25) and Definition 3.2.

Despite the caveats stated in Corollary 5.2 when some elements of $\mathbf{W}$ are equal to zero, it is important to keep in mind that the general differential formula (5.12) is still valid in that case as soon as $\mathbf{F}(.)$ has a local constant rank at $\mathbf{a} \in h^{-1}(\mathbb{R}_k^{p\times k})$. Furthermore, previous comparative studies have also demonstrated that first- and second-order variable projection methods used for minimizing $\psi(.)$ generally outperform other concurrent methods even for a large number of missing values in the case of binary weights and without any form of regularization to ensure the smoothness of $\psi(.)$ despite the non differentiability of $\mathbf{P}_{\mathbf{F}(.)}$ in some regions of the search space $h^{-1}(\mathbb{R}_k^{p\times k})$ [28][150][81][88].

Here, $D(\mathbf{F}(\mathbf{a}))$ and $D(\mathbf{P}_{\mathbf{F}(\mathbf{a})})$ are, for $\mathbf{a} \in h^{-1}(\mathbb{R}_k^{p\times k})$, elements of $\pounds(\mathbb{R}^{p.k}, \pounds(\mathbb{R}^{n.k}, \mathbb{R}^{n.p}))$ and $\pounds(\mathbb{R}^{p.k}, \pounds(\mathbb{R}^{n.p}, \mathbb{R}^{n.p}))$, respectively, and could be interpreted as tridimensional tensors (see equation (2.38) in Subsection 2.4). Now, since $\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})} = \mathbf{I}_{n.p} - \mathbf{P}_{\mathbf{F}(\mathbf{a})}$, we then have

$$D(\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}) = D(\mathbf{I}_{n.p} - \mathbf{P}_{\mathbf{F}(\mathbf{a})}) = -D(\mathbf{P}_{\mathbf{F}(\mathbf{a})})$$

and we deduce by the product differentiation rule [26] that

$$J(\mathbf{r}(\mathbf{a})) = J(\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\mathbf{x}) = D(\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})})\mathbf{x} + \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}J(\mathbf{x}) = -D(\mathbf{P}_{\mathbf{F}(\mathbf{a})})\mathbf{x}\ . \tag{5.13}$$

Substituting now for $D(\mathbf{P}_{\mathbf{F}(\mathbf{a})})$ yields

$$\begin{aligned} J(\mathbf{r}(\mathbf{a})) &= -\big(\ \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}D(\mathbf{F}(\mathbf{a}))\mathbf{F}(\mathbf{a})^{+}\mathbf{x} + (\mathbf{F}(\mathbf{a})^{+})^{T}D(\mathbf{F}(\mathbf{a}))^{T}(\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})})^{T}\mathbf{x}\ \ \big) \\ &= -\big(\ \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}D(\mathbf{F}(\mathbf{a}))\widehat{\mathbf{b}} \qquad\quad + (\mathbf{F}(\mathbf{a})^{+})^{T}D(\mathbf{F}(\mathbf{a}))^{T}\mathbf{r}(\mathbf{a}) \qquad\ \big)\ , \end{aligned}$$

where we have used the fact that $\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}$ is a symmetric matrix (see Subsection 2.1). In these equations, $\widehat{\mathbf{b}} = \mathbf{F}(\mathbf{a})^{+}\mathbf{x}$ and $\mathbf{r}(\mathbf{a}) = \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\mathbf{x}$ are, respectively, the minimum Euclidean norm solution and residual vector of the following linear least-squares problem already encountered when describing the block ALS method in Section 4

$$\min_{\mathbf{b}\in\mathbb{R}^{n.k}} \quad \frac{1}{2}\|\mathbf{x} - \mathbf{F}(\mathbf{a})\mathbf{b}\|_2^2 = \varphi^{*}(\mathbf{A}, \mathbf{B})\ ,$$

where $\mathbf{B} = \mathit{mat}(\mathbf{b})$. Note that we can also use $\widehat{\mathbf{b}} = \mathbf{F}(\mathbf{a})^{-}\mathbf{x}$, which is cheaper to evaluate, in the above equations. Moreover, we recall that the linear mappings $D(\mathbf{F}(\mathbf{a}))\widehat{\mathbf{b}}$ and $D(\mathbf{F}(\mathbf{a}))^{T}\mathbf{r}(\mathbf{a})$ are elements of $\pounds(\mathbb{R}^{p.k}, \mathbb{R}^{n.p})$ and $\pounds(\mathbb{R}^{p.k}, \mathbb{R}^{n.k})$, respectively, since transposition in the tensor $D(\mathbf{F}(\mathbf{a}))$ is performed on each slab $\partial\mathbf{F}(\mathbf{a})/\partial\mathbf{a}_i$. See equation (2.38) in Subsection 2.4 for details. Thus, these two factors correspond to $n.p \times p.k$ and $n.k \times p.k$ matrices, respectively.

We now derive an explicit formulation for the $n.p \times p.k$ matrix $J(\mathbf{r}(\mathbf{a}))$, which is independent of the differentiability of the residual function $\mathbf{r}(.)$ and the existence of the "true" Jacobian matrix of this residual function. We first consider the first term in $J(\mathbf{r}(\mathbf{a}))$, i.e.,

$$\mathbf{M}(\mathbf{a}) = \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}D\big(\mathbf{F}(\mathbf{a})\big)\widehat{\mathbf{b}}\ ,$$

which is also a $n.p \times p.k$ matrix. As derived in equation (3.20) of Subsection 3.4, $\mathbf{F}(\mathbf{a})$ may be expressed in the form

$$\mathbf{F}(\mathbf{a}) = \mathit{diag}\big(\mathit{vec}(\sqrt{\mathbf{W}})\big)(\mathbf{I}_n \otimes \mathbf{A}) = \mathit{diag}\big(\mathit{vec}(\sqrt{\mathbf{W}})\big)\big(\mathbf{I}_n \otimes \mathit{mat}_{k\times p}(\mathbf{a})^{T}\big)$$

and it is clear that $\mathbf{F}(.)$ is a continuous linear mapping from $\mathbb{R}^{p.k}$ into $\mathbb{R}^{n.p\times n.k}$ since the *mat* and transpose operators are linear mappings and the Kronecker and matrix products are bilinear operators. Hence, $\forall\ \mathbf{a}, \triangle\mathbf{a} \in \mathbb{R}^{p.k}$ and $\triangle\mathbf{A} = h(\triangle\mathbf{a}) = \mathit{mat}_{k\times p}(\triangle\mathbf{a})^{T}$, we have

$$D\big(\mathbf{F}(\mathbf{a})\big)(\triangle\mathbf{a}) = \mathbf{F}(\triangle\mathbf{a}) = \mathit{diag}\big(\mathit{vec}(\sqrt{\mathbf{W}})\big)(\mathbf{I}_n \otimes \triangle\mathbf{A})\ .$$

Noting that (see equation (2.33) in Subsection 2.2)

$$\mathbf{F}(\triangle\mathbf{a})\widehat{\mathbf{b}} = \mathit{diag}\big(\mathit{vec}(\sqrt{\mathbf{W}})\big)(\mathbf{I}_n \otimes \triangle\mathbf{A})\mathit{vec}(\widehat{\mathbf{B}}) = \mathit{diag}\big(\mathit{vec}(\sqrt{\mathbf{W}})\big)(\widehat{\mathbf{B}}^{T} \otimes \mathbf{I}_p)\mathit{vec}(\triangle\mathbf{A})\ ,$$

where $\widehat{\mathbf{B}} = mat_{k\times n}(\widehat{\mathbf{b}})$ and using the $p.k \times p.k$ commutation matrix $\mathbf{K}_{(k,p)}$ (see equation (2.34) in Subsection 2.2), we deduce that

$$(D\big(\mathbf{F}(\mathbf{a})\big)(\triangle\mathbf{a}))\widehat{\mathbf{b}} = diag\big(vec(\sqrt{\mathbf{W}})\big)(\widehat{\mathbf{B}}^T \otimes \mathbf{I}_p)\mathbf{K}_{(k,p)}\triangle\mathbf{a} \;, \tag{5.14}$$

since

$$vec(\triangle\mathbf{A}) = \mathbf{K}_{(k,p)}vec(\triangle\mathbf{A}^T) = \mathbf{K}_{(k,p)}\triangle\mathbf{a} \;,$$

following our conventions for the vectorized form of the $\mathbf{A}$ matrix defined in equation (3.21) of Subsection 3.4. In view of this, we finally obtain the following explicit formulation for the $n.p \times p.k$ matrix $\mathbf{M}(\mathbf{a})$

$$\mathbf{M}(\mathbf{a}) = \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}diag\big(vec(\sqrt{\mathbf{W}})\big)(\widehat{\mathbf{B}}^T \otimes \mathbf{I}_p)\mathbf{K}_{(k,p)} \;. \tag{5.15}$$

An alternative useful formulation of the $\mathbf{M}(\mathbf{a})$ matrix may be derived by noting that (see equation (2.36) and Lemma 2.2 in Subsection 2.2)

$$\begin{aligned} diag\big(vec(\sqrt{\mathbf{W}})\big)(\widehat{\mathbf{B}}^T \otimes \mathbf{I}_p)\mathbf{K}_{(k,p)} &= diag\big(vec(\sqrt{\mathbf{W}})\big)\mathbf{K}_{(n,p)}(\mathbf{I}_p \otimes \widehat{\mathbf{B}}^T) \\ &= \mathbf{K}_{(n,p)}diag\big(vec(\sqrt{\mathbf{W}}^T)\big)(\mathbf{I}_p \otimes \widehat{\mathbf{B}}^T) \\ &= \mathbf{K}_{(n,p)}\mathbf{G}(\widehat{\mathbf{b}}) \;, \end{aligned}$$

where $\mathbf{G}(\widehat{\mathbf{b}})$ is defined in equation (3.22) of Subsection 3.4. Thus,

$$\mathbf{M}(\mathbf{a}) = \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\mathbf{K}_{(n,p)}\mathbf{G}(\widehat{\mathbf{b}}) \;, \tag{5.16}$$

which will be used later, in particular in Theorem 5.3 and for computing $\nabla\psi(\mathbf{a})$ in Subsection 5.3 (see Theorem 5.7). As (see the demonstration of Theorem 4.3 for details)

$$e(\mathbf{A}, \widehat{\mathbf{B}}) = \mathbf{x} - \mathbf{F}(\mathbf{a})\widehat{\mathbf{b}} = \mathbf{K}_{(n,p)}\big(\mathbf{z} - \mathbf{G}(\widehat{\mathbf{b}})\mathbf{a}\big) \;,$$

where $\mathbf{z} = vec\big((\sqrt{\mathbf{W}} \odot \mathbf{X})^T\big)$, we can also write

$$\mathbf{M}(\mathbf{a}) = -\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\frac{\partial e(\mathbf{A}, \widehat{\mathbf{B}})}{\partial \mathbf{a}} \;. \tag{5.17}$$

In order to evaluate the second term in $J(\mathbf{r}(\mathbf{a}))$, i.e.,

$$\mathbf{L}(\mathbf{a}) = \big(\mathbf{F}(\mathbf{a})^+\big)^T D\big(\mathbf{F}(\mathbf{a})\big)^T\mathbf{r}(\mathbf{a}) \;,$$

which corresponds also to a $n.p \times p.k$ matrix, we first remark that, $\forall\ \mathbf{a}, \triangle\mathbf{a} \in \mathbb{R}^{p.k}$ and $\triangle\mathbf{A} = h(\triangle\mathbf{a}) = mat_{k\times p}(\triangle\mathbf{a})^T$, we have

$$\begin{aligned} \Big(D\big(\mathbf{F}(\mathbf{a})\big)(\triangle\mathbf{a})\Big)^T &= \mathbf{F}(\triangle\mathbf{a})^T \\ &= \Big(diag\big(vec(\sqrt{\mathbf{W}})\big)(\mathbf{I}_n \otimes \triangle\mathbf{A})\Big)^T \\ &= (\mathbf{I}_n \otimes \triangle\mathbf{A}^T)diag\big(vec(\sqrt{\mathbf{W}})\big) \;, \end{aligned}$$

since $\mathbf{F}(.)$ is a linear mapping and the transpose operator distributes over the Kronecker product. Now, $\forall\ \mathbf{Z} \in \mathbb{R}^{p\times n}$ and $\forall\ \mathbf{a}, \triangle\mathbf{a} \in \mathbb{R}^{p.k}$, using equation (2.33), we have

$$\begin{aligned} \big(D(\mathbf{F}(\mathbf{a}))(\triangle\mathbf{a})\big)^T vec(\mathbf{Z}) &= (\mathbf{I}_n \otimes \triangle\mathbf{A}^T)diag\big(vec(\sqrt{\mathbf{W}})\big)vec(\mathbf{Z}) \\ &= (\mathbf{I}_n \otimes \triangle\mathbf{A}^T)vec(\sqrt{\mathbf{W}} \odot \mathbf{Z}) \\ &= vec\big(\triangle\mathbf{A}^T(\sqrt{\mathbf{W}} \odot \mathbf{Z})\big) \\ &= \big((\sqrt{\mathbf{W}} \odot \mathbf{Z})^T \otimes \mathbf{I}_k\big)vec(\triangle\mathbf{A}^T) \\ &= \big((\sqrt{\mathbf{W}} \odot \mathbf{Z})^T \otimes \mathbf{I}_k\big)\triangle\mathbf{a} \;, \end{aligned}$$

and, thus, the $n.k \times p.k$ matrix representing the linear mapping $D(\mathbf{F}(\mathbf{a})(.))^T vec(\mathbf{Z})$ is

$$(\sqrt{\mathbf{W}} \odot \mathbf{Z})^T \otimes \mathbf{I}_k \ .$$

Now, using the projection operator $P_\Omega(.)$ associated with the $p \times n$ weight matrix $\mathbf{W}$ defined in equation (3.17), we have

$$\big[P_\Omega(\mathbf{X} - \mathbf{A}\widehat{\mathbf{B}})\big]_{ij} = \begin{cases} \mathbf{X}_{ij} - \sum_{l=1}^{k} \mathbf{A}_{il}\widehat{\mathbf{B}}_{lj} & \text{if } \mathbf{W}_{ij} \neq 0 \\ 0 & \text{if } \mathbf{W}_{ij} = 0 \end{cases} ,$$

and the variable projection residual vector of $\mathbf{x}$ at $\mathbf{A}$ can be written as

$$\mathbf{r}(\mathbf{a}) = vec\big(\sqrt{\mathbf{W}} \odot P_\Omega(\mathbf{X} - \mathbf{A}\widehat{\mathbf{B}})\big)$$

and it follows that, $\forall\ \mathbf{a}, \triangle\mathbf{a} \in \mathbb{R}^{p.k}$,

$$\Big(\big(D\big(\mathbf{F}(\mathbf{a})\big)(\triangle\mathbf{a})\Big)^T \mathbf{r}(\mathbf{a}) = \Big(\big(\sqrt{\mathbf{W}} \odot \sqrt{\mathbf{W}} \odot P_\Omega(\mathbf{X} - \mathbf{A}\widehat{\mathbf{B}})\big)^T \otimes \mathbf{I}_k\Big)\, \triangle\mathbf{a} \ , \tag{5.18}$$

hence

$$\mathbf{L}(\mathbf{a}) = \big(\mathbf{F}(\mathbf{a})^+\big)^T \Big(\big(\mathbf{W} \odot P_\Omega(\mathbf{X} - \mathbf{A}\widehat{\mathbf{B}})\big)^T \otimes \mathbf{I}_k\Big) \ . \tag{5.19}$$

At this point, we will introduce two new intermediate quantities to simplify the notation going forward, especially in the computation of the Hessian matrix in the next section:

$$\mathbf{U}(\mathbf{a}) = diag\big(vec(\sqrt{\mathbf{W}})\big)(\widehat{\mathbf{B}}^T \otimes \mathbf{I}_p)\mathbf{K}_{(k,p)} = \mathbf{K}_{(n,p)}\mathbf{G}(\widehat{\mathbf{b}}) \tag{5.20}$$

and

$$\mathbf{V}(\mathbf{a}) = \big(\mathbf{W} \odot P_\Omega(\mathbf{X} - \mathbf{A}\widehat{\mathbf{B}})\big)^T \otimes \mathbf{I}_k \ . \tag{5.21}$$

With these definitions, we have finally,

$$J\big(\mathbf{r}(\mathbf{a})\big) = -\big(\mathbf{M}(\mathbf{a}) + \mathbf{L}(\mathbf{a})\big) = -\Big(\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\mathbf{U}(\mathbf{a}) + \big(\mathbf{F}(\mathbf{a})^+\big)^T\mathbf{V}(\mathbf{a})\Big) \ . \tag{5.22}$$

We now demonstrate several important results concerning the ranges and null spaces associated with the $\mathbf{M}(\mathbf{a})$, $\mathbf{L}(\mathbf{a})$ and $J(\mathbf{r}(\mathbf{a}))$ matrices, which result directly from the use of the variable projection method.

First, we have $\mathbf{M}(\mathbf{a}) = \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\mathbf{U}(\mathbf{a})$ and this leads to $ran(\mathbf{M}(\mathbf{a})) \subset ran(\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}) = ran(\mathbf{F}(\mathbf{a}))^{\perp}$. Using the properties of the Moore-Penrose inverse (see equation (2.9) or more generally of any symmetric generalized inverse of the form (2.19) defined in Subsection 2.1), we also have

$$(\mathbf{F}(\mathbf{a})^+)^T = \big(\mathbf{F}(\mathbf{a})^+\mathbf{F}(\mathbf{a})\mathbf{F}(\mathbf{a})^+\big)^T = \big(\mathbf{F}(\mathbf{a})\mathbf{F}(\mathbf{a})^+\big)^T\big(\mathbf{F}(\mathbf{a})^+\big)^T = \mathbf{P}_{\mathbf{F}(\mathbf{a})}\big(\mathbf{F}(\mathbf{a})^+\big)^T$$

and we deduce that

$$\mathbf{L}(\mathbf{a}) = \mathbf{P}_{\mathbf{F}(\mathbf{a})}\big(\mathbf{F}(\mathbf{a})^+\big)^T\mathbf{V}(\mathbf{a}) \tag{5.23}$$

and $ran(\mathbf{L}(\mathbf{a})) \subset ran(\mathbf{P}_{\mathbf{F}(\mathbf{a})}) = ran(\mathbf{F}(\mathbf{a}))$. Hence the subspaces $ran(\mathbf{M}(\mathbf{a}))$ and $ran(\mathbf{L}(\mathbf{a}))$ of $\mathbb{R}^{p.n}$ are orthogonal and $ran(\mathbf{M}(\mathbf{a})) \cap ran(\mathbf{L}(\mathbf{a})) = \{\mathbf{0}^{p.n}\}$. Now, since $J(\mathbf{r}(\mathbf{a})) = -(\mathbf{M}(\mathbf{a}) + \mathbf{L}(\mathbf{a}))$, any element of $ran(J(\mathbf{r}(\mathbf{a})))$ may be written uniquely as a sum of an element of $ran(\mathbf{M}(\mathbf{a}))$ and an element of $ran(\mathbf{L}(\mathbf{a}))$ and it follows that

$$ran\,\big(J\big(\mathbf{r}(\mathbf{a})\big)\big) \subset ran\big(\mathbf{M}(\mathbf{a})\big) \oplus ran\big(\mathbf{L}(\mathbf{a})\big)$$

where $\oplus$ stands for the direct sum. From these results, it is then easy to show that

$$null\,\big(J\big(\mathbf{r}(\mathbf{a})\big)\big) = null\big(\mathbf{M}(\mathbf{a})\big) \cap null\big(\mathbf{L}(\mathbf{a})\big) \ .$$

Since $J(\mathbf{r}(\mathbf{a})) = -(\mathbf{M}(\mathbf{a}) + \mathbf{L}(\mathbf{a}))$, we have, by definition,

$$\textit{null}(\mathbf{M}) \cap \textit{null}(\mathbf{L}) \subset \textit{null}\left(J\big(\mathbf{r}(\mathbf{a})\big)\right) ,$$

and, reciprocally,

$$\begin{aligned}
\mathbf{c} \in \textit{null}\left(J\big(\mathbf{r}(\mathbf{a})\big)\right) &\Rightarrow \mathbf{M}(\mathbf{a})\mathbf{c} + \mathbf{L}(\mathbf{a})\mathbf{c} = \mathbf{0}^{p.n} \\
&\Rightarrow \mathbf{M}(\mathbf{a})\mathbf{c} = -\mathbf{L}(\mathbf{a})\mathbf{c} \\
&\Rightarrow \mathbf{M}(\mathbf{a})\mathbf{c} \in \textit{ran}(\mathbf{M}(\mathbf{a})) \cap \textit{ran}(\mathbf{L}(\mathbf{a})) \\
&\Rightarrow \mathbf{M}(\mathbf{a})\mathbf{c} = \mathbf{L}(\mathbf{a})\mathbf{c} = \mathbf{0}^{p.n} \\
&\Rightarrow \mathbf{c} \in \textit{null}\big(\mathbf{M}(\mathbf{a})\big) \cap \textit{null}\big(\mathbf{L}(\mathbf{a})\big) .
\end{aligned}$$

Now, we demonstrate that the matrices $\mathbf{M}(\mathbf{a})$, $\mathbf{L}(\mathbf{a})$ and $J(\mathbf{r}(\mathbf{a}))$ are rank deficient $\forall\ \mathbf{a} \in \mathbb{R}^{p.k}$. This result for $\mathbf{M}(\mathbf{a})$ was first noted by Ruhe [158] for the case $k = 1$ and $\mathbf{W}_{ij} \in \{0, 1\}$. It was proved later for general k, again only for $\mathbf{M}(\mathbf{a})$ and $\mathbf{W}_{ij} \in \{0, 1\}$, by Okatani and Deguchi [147], but under the restrictive hypotheses that $\mathbf{A}$, $\mathbf{B}$, $\mathbf{F}(\mathbf{a})$ and $\mathbf{G}(\mathbf{b})$ are of full rank. See also Okatani et al. [150], where these results are further developed. The next theorem and corollary extend this result for general k and any nonnegative real matrix $\mathbf{W}$ and to $\mathbf{M}(\mathbf{a})$, $\mathbf{L}(\mathbf{a})$ and $J(\mathbf{r}(\mathbf{a}))$ matrices without any restrictive assumptions.

Theorem 5.2. Let $k_{\mathbf{A}} = \textit{rank}(\mathbf{A})$. If,

$$\begin{aligned}
\mathbf{M}(\mathbf{a}) &= \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\mathbf{U}(\mathbf{a}) = \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\textit{diag}\left(\textit{vec}\big(\sqrt{\mathbf{W}}\big)\right)(\widehat{\mathbf{B}}^T \otimes \mathbf{I}_p)\mathbf{K}_{(k,p)} , \\
\mathbf{L}(\mathbf{a}) &= \big(\mathbf{F}(\mathbf{a})^+\big)^T\mathbf{V}(\mathbf{a}) = \big(\mathbf{F}(\mathbf{a})^+\big)^T\big((\mathbf{W} \odot P_{\Omega}(\mathbf{X} - \mathbf{A}\widehat{\mathbf{B}}))^T \otimes \mathbf{I}_k\big) , \\
J(\mathbf{r}(\mathbf{a})) &= -\big(\mathbf{M}(\mathbf{a}) + \mathbf{L}(\mathbf{a})\big) ,
\end{aligned}$$

where all the matrices and vectors have the same definitions as above, then the following relationships hold

$$\begin{aligned}
\textit{dim}\left(\textit{null}\big(\mathbf{M}(\mathbf{a})\big)\right) &\geqslant k_{\mathbf{A}}.k , \\
\textit{dim}\left(\textit{null}\big(\mathbf{L}(\mathbf{a})\big)\right) &\geqslant k_{\mathbf{A}}.k , \\
\textit{dim}\left(\textit{null}\big(J(\mathbf{r}(\mathbf{a}))\big)\right) &\geqslant k_{\mathbf{A}}.k .
\end{aligned}$$

Proof. Consider first the matrix $\mathbf{N}$ defined by

$$\mathbf{N} = \mathbf{K}_{(p,k)}(\mathbf{I}_k \otimes \mathbf{A}) .$$

Since $\mathbf{A}$ is of rank $k_{\mathbf{A}}$, $\mathbf{I}_k \otimes \mathbf{A}$ is of rank $k.k_{\mathbf{A}}$ (see equation (2.30)) and $\mathbf{N}$ is also of rank $k.k_{\mathbf{A}}$ because $\mathbf{K}_{(p,k)}$ is a permutation matrix and the rank of a matrix is unaltered by multiplication with a nonsingular square matrix.

Now, we first demonstrate that the space spanned by the columns of $\mathbf{N}$, which is of dimension $k.k_{\mathbf{A}}$, is included in $\textit{null}(\mathbf{M}(\mathbf{a}))$ and so $\textit{dim}(\textit{null}(\mathbf{M}(\mathbf{a}))) \geqslant k_{\mathbf{A}}.k$.

Let $\mathbf{t} \in \textit{ran}(\mathbf{N})$, then $\exists\mathbf{Z} \in \mathbb{R}^{k\times k}$ such that

$$\mathbf{t} = \mathbf{N}\textit{vec}(\mathbf{Z}) = \mathbf{K}_{(p,k)}(\mathbf{I}_k \otimes \mathbf{A})\textit{vec}(\mathbf{Z}) = \mathbf{K}_{(p,k)}\textit{vec}(\mathbf{A}\mathbf{Z}) .$$

From this equality, we deduce that

$$\begin{aligned}
\mathbf{M}(\mathbf{a})\mathbf{t} &= \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\textit{diag}(\textit{vec}(\sqrt{\mathbf{W}}))(\widehat{\mathbf{B}}^T \otimes \mathbf{I}_p)\mathbf{K}_{(k,p)}\mathbf{K}_{(p,k)}\textit{vec}(\mathbf{A}\mathbf{Z}) \\
&= \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\textit{diag}(\textit{vec}(\sqrt{\mathbf{W}}))(\widehat{\mathbf{B}}^T \otimes \mathbf{I}_p)\textit{vec}(\mathbf{A}\mathbf{Z}) \\
&= \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\textit{diag}(\textit{vec}(\sqrt{\mathbf{W}}))\textit{vec}(\mathbf{A}\mathbf{Z}\widehat{\mathbf{B}}) \\
&= \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\textit{diag}(\textit{vec}(\sqrt{\mathbf{W}}))(\mathbf{I}_n \otimes \mathbf{A})\textit{vec}(\mathbf{Z}\widehat{\mathbf{B}}) \\
&= \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\mathbf{F}(\mathbf{a})\textit{vec}(\mathbf{Z}\widehat{\mathbf{B}}) \\
&= \mathbf{0}^{p.n} ,
\end{aligned}$$

since $\mathbf{P}^{\perp}_{\mathbf{F(a)}}$ is the orthogonal projector onto $ran(\mathbf{F}(\mathbf{a}))^{\perp}$ and, finally, $\mathbf{t} \in null(\mathbf{M}(\mathbf{a}))$. In other words, $ran(\mathbf{N}) \subset null(\mathbf{M}(\mathbf{a}))$ and, hence, $dim(null(\mathbf{M}(\mathbf{a}))) \geqslant dim(ran(\mathbf{N})) = k_{\mathbf{A}}.k$.

We now demonstrate that the relation $ran(\mathbf{N}) \subset null(\mathbf{L}(\mathbf{a}))$ also holds. If $\mathbf{t} \in ran(\mathbf{N})$ and $\mathbf{Z} \in \mathbb{R}^{k\times k}$ with $\mathbf{t} = \mathbf{N}vec(\mathbf{Z})$, using equation (2.33) and Lemma 2.2, we have

$$\begin{aligned}
\mathbf{L}(\mathbf{a})\mathbf{t} &= (\mathbf{F}(\mathbf{a})^{+})^{T}((\mathbf{W} \odot P_{\Omega}(\mathbf{X} - \mathbf{A}\widehat{\mathbf{B}}))^{T} \otimes \mathbf{I}_k)\mathbf{K}_{(p,k)}vec(\mathbf{A}\mathbf{Z}) \\
&= (\mathbf{F}(\mathbf{a})^{+})^{T}((\mathbf{W} \odot P_{\Omega}(\mathbf{X} - \mathbf{A}\widehat{\mathbf{B}}))^{T} \otimes \mathbf{I}_k)vec(\mathbf{Z}^{T}\mathbf{A}^{T}) \\
&= (\mathbf{F}(\mathbf{a})^{+})^{T}vec(\mathbf{Z}^{T}\mathbf{A}^{T}(\mathbf{W} \odot P_{\Omega}(\mathbf{X} - \mathbf{A}\widehat{\mathbf{B}}))) \\
&= (\mathbf{F}(\mathbf{a})^{+})^{T}(\mathbf{I}_n \otimes \mathbf{Z}^{T})vec(\mathbf{A}^{T}(\mathbf{W} \odot P_{\Omega}(\mathbf{X} - \mathbf{A}\widehat{\mathbf{B}}))) \\
&= (\mathbf{F}(\mathbf{a})^{+})^{T}(\mathbf{I}_n \otimes \mathbf{Z}^{T})(\mathbf{I}_n \otimes \mathbf{A}^{T})vec(\mathbf{W} \odot P_{\Omega}(\mathbf{X} - \mathbf{A}\widehat{\mathbf{B}})) \\
&= (\mathbf{F}(\mathbf{a})^{+})^{T}(\mathbf{I}_n \otimes \mathbf{Z}^{T})(\mathbf{I}_n \otimes \mathbf{A}^{T})diag(vec(\sqrt{\mathbf{W}}))vec(\sqrt{\mathbf{W}} \odot P_{\Omega}(\mathbf{X} - \mathbf{A}\widehat{\mathbf{B}})) \\
&= (\mathbf{F}(\mathbf{a})^{+})^{T}(\mathbf{I}_n \otimes \mathbf{Z}^{T})\mathbf{F}(\mathbf{a})^{T}\mathbf{r}(\mathbf{a}) \\
&= \mathbf{0}^{p.n} ,
\end{aligned}$$

since $\mathbf{F}(\mathbf{a})^{T}\mathbf{r}(\mathbf{a}) = \mathbf{0}^{k.n}$. Thus, $ran(\mathbf{N}) \subset null(\mathbf{L}(\mathbf{a}))$ and, hence,

$$dim\big(null\big(\mathbf{L}(\mathbf{a})\big)\big) \geqslant dim\big(ran(\mathbf{N})\big) = k_{\mathbf{A}}.k .$$

Finally, we have $ran(\mathbf{N}) \subset null(\mathbf{M}(\mathbf{a})) \cap null(\mathbf{L}(\mathbf{a})) = null(J(\mathbf{r}(\mathbf{a})))$ and, so,

$$dim\big(null\big(J(\mathbf{r}(\mathbf{a}))\big)\big) \geqslant k_{\mathbf{A}}.k .$$

□

Corollary 5.3. With the same notations as in Theorem 5.2, we have

$$\begin{aligned}
rank\big(\mathbf{M}(\mathbf{a})\big) &\leqslant (p - k_{\mathbf{A}}).k \\
rank\big(\mathbf{L}(\mathbf{a})\big) &\leqslant (p - k_{\mathbf{A}}).k \\
rank\big(J(\mathbf{r}(\mathbf{a}))\big) &\leqslant (p - k_{\mathbf{A}}).k .
\end{aligned}$$

Proof. These inequalities follow directly from Theorem 5.2 and the rank-nullity theorem (see equation (2.1) in Subsection 2.1):

$$\begin{aligned}
rank\big(\mathbf{M}(\mathbf{a})\big) &= k.p - dim\Big(null\big(\mathbf{M}(\mathbf{a})\big)\Big) \\
rank\big(\mathbf{L}(\mathbf{a})\big) &= k.p - dim\Big(null\big(\mathbf{L}(\mathbf{a})\big)\Big) \\
rank\big(J(\mathbf{r}(\mathbf{a}))\big) &= k.p - dim\Big(null\big(J(\mathbf{r}(\mathbf{a}))\big)\Big) .
\end{aligned}$$

□

Remark 5.1. Theorem 5.2 and Corollary 5.3 are obviously connected to the fact that the matrix factorization $\mathbf{Y} = \mathbf{AB}$ used in the (P1) formulation of the WLRA problem is overparameterized, that the minimization of $\psi(.)$ is an optimization problem on $\mathrm{Gr}(p, k)$, the set of linear subspaces of fixed dimension k of the Euclidean space $\mathbb{R}^p$, and that $\mathrm{Gr}(p, k)$ is a smooth manifold of dimension $p.k - k.k$ (see Remarks 3.4 and 3.7 for details). Assuming that $\mathbf{A} \in \mathbb{R}^{p\times k}_{k}$, the minimization of the cost function $\psi(.)$ is at first sight a $k.p$ dimensional problem. However, $\psi(\mathbf{a})$ depends only on the column space of $\mathbf{A}$ and not on its individual elements [51][125][28][14]. As an illustration, if $\mathbf{A} \in \mathbb{R}^{p\times k}_{k}$ and $d\mathbf{A} \in \mathbb{R}^{p\times k}$ is a perturbation matrix, for certain matrices $d\mathbf{A}$, we will have $ran(\mathbf{A} + d\mathbf{A}) = ran(\mathbf{A})$, which will imply that $\psi(\mathbf{A} + d\mathbf{A}) = \psi(\mathbf{A})$. This demonstrates that it is

not useful to consider all $k.p$ search directions for minimizing $\psi(.)$ from a previous matrix estimate $\mathbf{A}$. As demonstrated in [51][125][28][14], this symmetry can be exploited to reduce the dimension of the problem to $k.(p-k)$ parameters instead of $k.p$ in both the (VP1) and (VP2) formulations of the WLRA problem. Thus, in that sense, the column space of $\mathbf{A}$ has only $p.k - k.k$ degrees of freedom, which is consistent to the facts that the rank of $J(\mathbf{r}(\mathbf{a}))$ is at most $p.k-k.k$ if $\mathit{rank}(\mathbf{A}) = k$ and that the the dimension of $\mathrm{Gr}(p,k)$ is also $p.k - k.k$. ■

Theorem 5.2 demonstrates that the Jacobian matrix $J(\mathbf{r}(\mathbf{a}))$ is always rank-deficient. This implies that the linear least-squares problem

$$\min_{d\mathbf{a}\in\mathbb{R}^{p.k}} \frac{1}{2}\|\mathbf{r}(\mathbf{a}) + J(\mathbf{r}(\mathbf{a}))d\mathbf{a}\|_2^2 ,$$

which must be solved at each iteration of a Gauss-Newton type algorithm (see Subsection 5.1 for details) has an infinite set of solutions [71][87][8] and we must remove this ambiguity in any practical implementation of the Gauss-Newton algorithm in a such way that the direction vector $d\mathbf{a}_{gn}$ can be determined uniquely at each iteration. The general solution $d\widehat{\mathbf{a}} \in \mathbb{R}^{k.p}$ of the above rank-deficient linear least-squares problem can be written as

$$d\widehat{\mathbf{a}} = -J(\mathbf{r}(\mathbf{a}))^+\mathbf{r}(\mathbf{a}) + \mathbf{c} = d\mathbf{a}_{min} + \mathbf{c} ,$$

where, as before, $J(\mathbf{r}(\mathbf{a}))^+$ is the pseudo-inverse of $J(\mathbf{r}(\mathbf{a}))$, $d\mathbf{a}_{min}$ is the (unique) minimum 2-norm solution of the above linear least-squares problem (see Subsection 2.1 and [71][87][8]), and $\mathbf{c}$ is an arbitrary $k.p$ dimensional vector belonging to $\mathit{null}(J(\mathbf{r}(\mathbf{a})))$.

First, the pseudo-inverse solution $d\mathbf{a}_{min}$ is characterized uniquely by the two conditions

$$J(\mathbf{r}(\mathbf{a}))^T J(\mathbf{r}(\mathbf{a}))d\mathbf{a}_{min} = -J(\mathbf{r}(\mathbf{a}))^T\mathbf{r}(\mathbf{a}) \text{ and } d\mathbf{a}_{min} \in \mathit{null}(J(\mathbf{r}(\mathbf{a})))^\perp .$$

The first condition states simply that $d\mathbf{a}_{min}$ is a solution of the normal equations of the linear-least-squares problem. Note, further, that

$$d\mathbf{a}_{min} = -J(\mathbf{r}(\mathbf{a}))^+\mathbf{r}(\mathbf{a}) = -J(\mathbf{r}(\mathbf{a}))^+J(\mathbf{r}(\mathbf{a}))J(\mathbf{r}(\mathbf{a}))^+\mathbf{r}(\mathbf{a}) = \mathbf{P}_{J(\mathbf{r}(\mathbf{a}))^T}d\mathbf{a}_{min} ,$$

where $\mathbf{P}_{J(\mathbf{r}(\mathbf{a}))^T}$ is the orthogonal projector onto the row space of $J(\mathbf{r}(\mathbf{a}))$ (e.g., $\mathit{ran}(J(\mathbf{r}(\mathbf{a}))^T)$, see Subsection 2.1 for details. Since $\mathit{ran}(J(\mathbf{r}(\mathbf{a}))^T) = \mathit{null}(J(\mathbf{r}(\mathbf{a})))^\perp$, we deduce immediately that $d\mathbf{a}_{min} \in \mathit{null}(J(\mathbf{r}(\mathbf{a})))^\perp$ as stated in the second condition.

Now, if $d\widehat{\mathbf{a}} = d\mathbf{a}_{min} + \mathbf{c}$, we have

$$J(\mathbf{r}(\mathbf{a}))d\widehat{\mathbf{a}} = J(\mathbf{r}(\mathbf{a}))(d\mathbf{a}_{min} + \mathbf{c}) = J(\mathbf{r}(\mathbf{a}))d\mathbf{a}_{min} ,$$

and, thus, $\|\mathbf{r}(\mathbf{a}) + J(\mathbf{r}(\mathbf{a}))d\widehat{\mathbf{a}}\|_2 = \|\mathbf{r}(\mathbf{a}) + J(\mathbf{r}(\mathbf{a}))d\mathbf{a}_{min}\|_2$, which implies that $d\widehat{\mathbf{a}}$ is also a solution of the above linear least-squares problem. Reciprocally, if $d\widehat{\mathbf{a}}$ is an arbitrary solution, we have $\|\mathbf{r}(\mathbf{a}) + J(\mathbf{r}(\mathbf{a}))d\widehat{\mathbf{a}}\|_2 = \|\mathbf{r}(\mathbf{a}) + J(\mathbf{r}(\mathbf{a}))d\mathbf{a}_{min}\|_2$, which implies that

$$J(\mathbf{r}(\mathbf{a}))d\widehat{\mathbf{a}} = -\mathbf{P}_{J(\mathbf{r}(\mathbf{a}))}\mathbf{r}(\mathbf{a}) = J(\mathbf{r}(\mathbf{a}))d\mathbf{a}_{min} ,$$

as $-\mathbf{P}_{J(\mathbf{r}(\mathbf{a}))}\mathbf{r}(\mathbf{a})$ is the unique closest point to $\mathbf{r}(\mathbf{a})$ in $\mathit{ran}(J(\mathbf{r}(\mathbf{a})))$, see equation (2.14) of Subsection 2.1 for details. Thus, $J(\mathbf{r}(\mathbf{a}))(d\widehat{\mathbf{a}} - d\mathbf{a}_{min}) = \mathbf{0}^{p.n}$ and we can write $d\widehat{\mathbf{a}}$ as

$$d\widehat{\mathbf{a}} = d\mathbf{a}_{min} + (d\widehat{\mathbf{a}} - d\mathbf{a}_{min}) = d\mathbf{a}_{min} + \mathbf{c} ,$$

with $\mathbf{c} = d\widehat{\mathbf{a}} - d\mathbf{a}_{min} \in \mathit{null}(J(\mathbf{r}(\mathbf{a})))$.

In other words, all solution vectors $d\widehat{\mathbf{a}}$ can be written uniquely as the sum of $d\mathbf{a}_{min} \in \mathit{null}(J(\mathbf{r}(\mathbf{a})))^\perp$ and a vector $\mathbf{c} \in \mathit{null}(J(\mathbf{r}(\mathbf{a})))$ and finding all the solutions of the above rank-deficient linear least-squares problem requires computing both a generalized inverse and a basis of the null space of $J(\mathbf{r}(\mathbf{a}))$. Obviously, this also implies to determine accurately the rank of $J(\mathbf{r}(\mathbf{a}))$ or, equivalently,

the rank of its null space. More generally, proceeding in a similar manner, it is also easy to establish an one to one mapping between the elements of $null(J(\mathbf{r}(\mathbf{a})))^{\perp}$ and those of $ran(J(\mathbf{r}(\mathbf{a})))$.

Now, the most natural choice is to select $d\mathbf{a}_{gn} = d\mathbf{a}_{min}$ as the solution of our linear least-squares problem since, with such minimum Euclidean norm solution, the first order Taylor's expansion

$$\mathbf{r}(\mathbf{a} + d\mathbf{a}_{gn}) = \mathbf{r}(\mathbf{a}) + J(\mathbf{r}(\mathbf{a}))d\mathbf{a}_{gn} + \mathcal{O}(\|d\mathbf{a}_{gn}\|_2^2) \ ,$$

which is at the base of the Gauss-Newton algorithm is the most accurate. Selecting $d\mathbf{a}_{gn} = d\mathbf{a}_{min}$ has also a strong theoretical justification as, with this choice, a variable projection Gauss-Newton algorithm used to minimize $\psi(.)$ is equivalent to a Riemannian optimization method operating directly on the Grassmann manifold $\mathrm{Gr}(p, k)$ [3][11] as we will explain later in this subsection.

These considerations related to the uniform rank degeneracy of the Jacobian matrix $J(\mathbf{r}(\mathbf{a}))$ apply also to the computation of the correction vector $d\mathbf{a}_{lm}$ in the Levenberg-Marquardt method as soon as the Marquardt damping parameter λ approaches zero, as it is expected after some iterations of the Levenberg-Marquardt algorithm. Moreover, if the Marquardt parameter λ is controlled so that it does not approach to zero in order to remove the uniform singularity of the Jacobian matrix $J(\mathbf{r}(\mathbf{a}))$, this may severely deteriorate the global convergence as well as the local convergence of the method in a neighborhood of a critical point. In other words, adding the additional constraint that $\|d\mathbf{a}_{lm}\|_2$ is minimum when λ approaches zero, is also important for the robustness and efficiency of the Levenberg-Marquardt or similar regularized methods described in Subsection 5.1 when they are used to solve NNLS problems with an uniformly deficient Jacobian matrix, like the WLRA problem.

We now give sufficient conditions for the equalities:

$$\begin{aligned} dim\Big(null\big(\mathbf{M}(\mathbf{a})\big)\Big) &= k_{\mathbf{A}}.k \\ dim\Big(null\big(J(\mathbf{r}(\mathbf{a}))\big)\Big) &= k_{\mathbf{A}}.k \ , \end{aligned}$$

which will be helpful to remove these ambiguities in determining uniquely and efficiently the correction vectors $d\mathbf{a}_{gn}$ and $d\mathbf{a}_{lm}$ in many practical applications.

Let us first introduce some definitions and notations. For any nonnegative real $p \times n$ matrix $\mathbf{W}$, we define the finite subset of $\mathbb{N}$

$$\Theta(\mathbf{W}) = \big\{j \in \{1, 2, \cdots, n\} \ / \ \mathbf{W}_{.j} \in \mathbb{R}^p_{+*}\big\} \ .$$

$\Theta(\mathbf{W})$ is the set of the column-vector indices of $\mathbf{W}$ such that 0 is not an element of such column-vector of $\mathbf{W}$. Furthermore, let $card(\Theta(\mathbf{W}))$ be the number of elements of $\Theta(\mathbf{W})$ and, for any $s \times n$ matrix $\mathbf{C}$, define the $s \times card(\Theta(\mathbf{W}))$ real submatrix $\mathbf{C}^{'}$ obtained from $\mathbf{C}$ by deleting the columns of $\mathbf{C}$ whose indices do not belong to $\Theta(\mathbf{W})$. We then have the following result, which is new as far as we know.

Theorem 5.3. With these definitions and the same notations as in Theorem 5.2, if $card(\Theta(\mathbf{W})) = n^{'} \geqslant k$ and $rank(\widehat{\mathbf{B}}^{'}) = k$ then the following equalities hold:

$$\begin{aligned} null\big(J(\mathbf{r}(\mathbf{a}))\big) &= null\big(\mathbf{M}(\mathbf{a})\big) \ , \\ dim\big(null(J(\mathbf{r}(\mathbf{a})))\big) &= k_{\mathbf{A}}.k \ . \end{aligned}$$

Proof. First, consider the second formulation of the $\mathbf{M}(\mathbf{a})$ matrix (see equation (5.16)), e.g.,

$$\mathbf{M}(\mathbf{a}) = \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\mathbf{K}_{(n,p)}\mathbf{G}(\widehat{\mathbf{b}}) \ ,$$

where $\widehat{\mathbf{b}} = vec(\widehat{\mathbf{B}})$ and $\mathbf{G}(\widehat{\mathbf{b}}) = diag(vec(\sqrt{\mathbf{W}}^T))(\mathbf{I}_p \otimes \widehat{\mathbf{B}}^T)$. Using the two hypotheses $card(\Theta(\mathbf{W})) = n' \geqslant k$ and $rank(\widehat{\mathbf{B}}') = k$, we first deduce that

$$\begin{aligned} rank\Big(diag\big(vec(\sqrt{\mathbf{W}'}^T)\big)(\mathbf{I}_p \otimes \widehat{\mathbf{B}}'^T)\Big) &= rank\big(\mathbf{I}_p \otimes \widehat{\mathbf{B}}'^T\big) \ , \\ &= rank(\mathbf{I}_p).rank(\widehat{\mathbf{B}}'^T) \\ &= p.k \ , \end{aligned}$$

since $diag(vec(\sqrt{\mathbf{W}'}^T))$ is a nonsingular diagonal matrix. Now, using this equality, we have also

$$rank\big(\mathbf{K}_{(n,p)}\mathbf{G}(\widehat{\mathbf{b}})\big) = rank(\mathbf{G}(\widehat{\mathbf{b}})) = k.p \ ,$$

as $\mathbf{K}_{(n,p)}$ is a (nonsingular) permutation matrix and $diag(vec(\sqrt{\mathbf{W}'}^T))(\mathbf{I}_p \otimes \widehat{\mathbf{B}}'^T)$ is a submatrix of $\mathbf{G}(\widehat{\mathbf{b}})$ formed simply by eliminating some rows of $\mathbf{G}(\widehat{\mathbf{b}})$.

Now, for any matrix $\mathbf{C}$ with s columns, we have the basic rank-nullity relation (see equation (2.1))

$$s = rank(\mathbf{C}) + dim\big(null(\mathbf{C})\big) \ .$$

Furthermore, for any matrix $\mathbf{D}$ with s rows, we also assume the equality

$$rank(\mathbf{D}) = rank(\mathbf{CD}) + dim\big(null(\mathbf{C}) \cap ran(\mathbf{D})\big) \ ,$$

see Marsaglia and Styan [126] for a proof.

Using these two relations, we deduce

$$rank\big(\mathbf{M}(\mathbf{a})\big) + dim\big(null(\mathbf{M}(\mathbf{a}))\big) = k.p$$

and

$$k.p = rank\big(\mathbf{M}(\mathbf{a})\big) + dim\big(null(\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}) \cap ran(\mathbf{K}_{(n,p)}\mathbf{G}(\widehat{\mathbf{b}}))\big) \ ,$$

and so

$$\begin{aligned} dim\big(null(\mathbf{M}(\mathbf{a}))\big) &= dim\big(null(\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\big) \cap ran\big(\mathbf{K}_{(n,p)}\mathbf{G}(\widehat{\mathbf{b}})\big)\big) \\ &= dim\big(ran(\mathbf{F}(\mathbf{a})\big) \cap ran\big(\mathbf{K}_{(n,p)}\mathbf{G}(\widehat{\mathbf{b}})\big)\big) \ . \end{aligned}$$

Next, we consider the matrix $\mathbf{H}$ defined by

$$\mathbf{H} = diag(vec(\sqrt{\mathbf{W}}))(\widehat{\mathbf{B}}^T \otimes \mathbf{A}) \ .$$

We have

$$rank(\mathbf{H}) \leqslant rank(\widehat{\mathbf{B}}^T \otimes \mathbf{A}) = rank(\widehat{\mathbf{B}}).rank(\mathbf{A}) = k.k_{\mathbf{A}} \ ,$$

since the hypothesis $rank(\widehat{\mathbf{B}}') = k$ implies $rank(\widehat{\mathbf{B}}) = k$. We now demonstrate the inclusion

$$ran(\mathbf{F}(\mathbf{a})) \cap ran(\mathbf{K}_{(n,p)}\mathbf{G}(\widehat{\mathbf{b}})) \subset ran(\mathbf{H}) \ .$$

Let $\mathbf{c} \in ran(\mathbf{F}(\mathbf{a})) \cap ran(\mathbf{K}_{(n,p)}\mathbf{G}(\widehat{\mathbf{b}}))$, then $\exists\ \mathbf{S} \in \mathbb{R}^{k \times n}$ and $\mathbf{Z} \in \mathbb{R}^{p \times k}$ such that

$$\mathbf{c} = \mathbf{F}(\mathbf{a})vec(\mathbf{S}) = \mathbf{K}_{(n,p)}\mathbf{G}(\widehat{\mathbf{b}})vec(\mathbf{Z}^T)$$

and we want to show that $\exists \mathbf{T} \in \mathbb{R}^{k \times k}$ such that $\mathbf{c} = \mathbf{H}vec(\mathbf{T})$. But, $\forall \mathbf{T} \in \mathbb{R}^{k \times k}$, we have

$$\begin{aligned} \mathbf{H}vec(\mathbf{T}) &= diag(vec(\sqrt{\mathbf{W}}))(\widehat{\mathbf{B}}^T \otimes \mathbf{A})vec(\mathbf{T}) \\ &= diag(vec(\sqrt{\mathbf{W}}))(\widehat{\mathbf{B}}^T \otimes \mathbf{I}_p)(\mathbf{I}_k \otimes \mathbf{A})vec(\mathbf{T}) \\ &= diag(vec(\sqrt{\mathbf{W}}))(\widehat{\mathbf{B}}^T \otimes \mathbf{I}_p)\mathbf{K}_{(k,p)}\mathbf{K}_{(p,k)}vec(\mathbf{AT}) \\ &= \mathbf{K}_{(n,p)}\mathbf{G}(\widehat{\mathbf{b}})\mathbf{K}_{(p,k)}vec(\mathbf{AT}) \ , \end{aligned}$$

and so, using the facts that $\mathbf{K}_{(n,p)}$ is a (nonsingular) permutation matrix and $\mathbf{G}(\widehat{\mathbf{b}})$ has full column rank demonstrated above, we have the equivalences

$$\begin{aligned}\mathbf{c} = \mathbf{H}vec(\mathbf{T}) &\Leftrightarrow \mathbf{K}_{(n,p)}\mathbf{G}(\widehat{\mathbf{b}})vec(\mathbf{Z}^T) = \mathbf{K}_{(n,p)}\mathbf{G}(\widehat{\mathbf{b}})\mathbf{K}_{(p,k)}vec(\mathbf{AT})\\ &\Leftrightarrow \mathbf{G}(\widehat{\mathbf{b}})vec(\mathbf{Z}^T) = \mathbf{G}(\widehat{\mathbf{b}})\mathbf{K}_{(p,k)}vec(\mathbf{AT})\\ &\Leftrightarrow vec(\mathbf{Z}^T) = \mathbf{K}_{(p,k)}vec(\mathbf{AT})\\ &\Leftrightarrow \mathbf{K}_{(k,p)}vec(\mathbf{Z}^T) = vec(\mathbf{AT})\\ &\Leftrightarrow vec(\mathbf{Z}) = vec(\mathbf{AT})\\ &\Leftrightarrow \mathbf{Z} = \mathbf{AT}\ .\end{aligned}$$

Thus, to demonstrate that $\exists \mathbf{T} \in \mathbb{R}^{k\times k}$ such that $\mathbf{c} = \mathbf{H}vec(\mathbf{T})$, it suffices to show that $\exists \mathbf{T} \in \mathbb{R}^{k\times k}$ such that $\mathbf{Z} = \mathbf{AT}$. But,

$$\begin{aligned}\mathbf{c} &= \mathbf{F}(\mathbf{a})vec(\mathbf{S})\\ &= diag(vec(\sqrt{\mathbf{W}}))(\mathbf{I}_n \otimes \mathbf{A})vec(\mathbf{S})\\ &= diag(vec(\sqrt{\mathbf{W}}))vec(\mathbf{AS})\end{aligned}$$

and also

$$\begin{aligned}\mathbf{c} &= \mathbf{K}_{(n,p)}\mathbf{G}(\widehat{\mathbf{b}})vec(\mathbf{Z}^T)\\ &= \mathbf{K}_{(n,p)}diag(vec(\sqrt{\mathbf{W}}^T))(\mathbf{I}_p \otimes \widehat{\mathbf{B}}^T)vec(\mathbf{Z}^T)\\ &= diag(vec(\sqrt{\mathbf{W}}))(\mathbf{I}_n \otimes \mathbf{Z})vec(\widehat{\mathbf{B}})\\ &= diag(vec(\sqrt{\mathbf{W}}))vec(\mathbf{Z}\widehat{\mathbf{B}})\ .\end{aligned}$$

From these equalities, we then deduce that

$$diag(vec(\sqrt{\mathbf{W}'}))vec(\mathbf{AS}') = diag(vec(\sqrt{\mathbf{W}'}))vec(\mathbf{Z}\widehat{\mathbf{B}}')\ ,$$

and since $diag(vec(\sqrt{\mathbf{W}'}))$ is a nonsingular diagonal matrix, we obtain

$$vec(\mathbf{AS}') = vec(\mathbf{Z}\widehat{\mathbf{B}}') \text{ and, finally, } \mathbf{AS}' = \mathbf{Z}\widehat{\mathbf{B}}'\ .$$

Now, $\widehat{\mathbf{B}}'$ has full row-rank by hypothesis and, consequently, admits a right inverse $\mathbf{R}$ such that $\widehat{\mathbf{B}}'\mathbf{R} = \mathbf{I}_k$ (see [126] for details) and so $\mathbf{AS}'\mathbf{R} = \mathbf{Z}$ and we can take $\mathbf{T} = \mathbf{S}'\mathbf{R}$. Consequently, we have proved

$$ran(\mathbf{F}(\mathbf{a})) \cap ran(\mathbf{K}_{(n,p)}\mathbf{G}(\widehat{\mathbf{b}})) \subset ran(\mathbf{H})\ ,$$

which implies

$$dim(null(\mathbf{M}(\mathbf{a}))) = dim(ran(\mathbf{F}(\mathbf{a})) \cap ran(\mathbf{K}_{(n,p)}\mathbf{G}(\widehat{\mathbf{b}}))) \leqslant rank(\mathbf{H}) \leqslant k_{\mathbf{A}}.k\ .$$

But, from Theorem 5.2, we already know that

$$k_{\mathbf{A}}.k \leqslant dim(null(\mathbf{M}(\mathbf{a})))$$

and we obtain $dim(null(\mathbf{M}(\mathbf{a}))) = k_{\mathbf{A}}.k$.

Again, from Theorem 5.2, we also know that

$$k_{\mathbf{A}}.k \leqslant dim(null(J(\mathbf{r}(\mathbf{a})))) = dim(null(\mathbf{M}(\mathbf{a})) \cap null(\mathbf{L}(\mathbf{a})))$$

and since

$$dim(null(\mathbf{M}(\mathbf{a})) \cap null(\mathbf{L}(\mathbf{a}))) \leqslant dim(null(\mathbf{M}(\mathbf{a}))) = k_{\mathbf{A}}.k\ ,$$

we also conclude that $dim(null(J(\mathbf{r}(\mathbf{a})))) = k_{\mathbf{A}}.k$.

Finally, from the propositions,

$$dim(null(J(\mathbf{r}(\mathbf{a})))) = dim(null(\mathbf{M}(\mathbf{a}))) \text{ and } null(J(\mathbf{r}(\mathbf{a}))) \subset null(\mathbf{M}(\mathbf{a}))\ ,$$

we deduce that $null(J(\mathbf{r}(\mathbf{a}))) = null(\mathbf{M}(\mathbf{a}))$ as claimed in the theorem.

□

Before stating some consequences of Theorem 5.3, it is important to highlight that the two hypotheses of this theorem are not very stringent and are easily checked in practice. Moreover, in most practical cases, these two conditions will be meet as long as $\mathbf{W}$ has k column-vectors without zero elements and each other column-vector of $\mathbf{W}$ has at least k nonzero elements in it. The following corollary is then obvious and is stated without proof:

Corollary 5.4. With the same notations and hypotheses as in Theorem 5.3, we also have the relations

$$\begin{aligned} \mathit{rank}\big(\mathbf{M}(\mathbf{a})\big) &= (p - k_{\mathbf{A}}).k \ , \\ \mathit{rank}\big(J(\mathbf{r}(\mathbf{a}))\big) &= (p - k_{\mathbf{A}}).k \ . \end{aligned}$$

□

Furthermore, the next corollary shows that the sufficient conditions stated in Theorem 5.3 and used in Corollary 5.4 can be simplified when $\mathbf{W}$ is a strictly positive $p \times n$ real matrix. A similar result has been obtained by Chen [27] for $\mathbf{M}(\mathbf{a})$ only, but with a different proof and the additional hypothesis that $\mathbf{A}$ is of full column-rank.

Corollary 5.5. With the same notations as in Theorem 5.2, if $\mathbf{W} \in \mathbb{R}_{+*}^{p \times n}$ and $\mathit{rank}(\widehat{\mathbf{B}}) = k$ then the following equalities hold:

$$\begin{aligned} \mathit{dim}\big(\mathit{null}(\mathbf{M}(\mathbf{a}))\big) &= k_{\mathbf{A}}.k \ , \\ \mathit{dim}\big(\mathit{null}(J(\mathbf{r}(\mathbf{a})))\big) &= k_{\mathbf{A}}.k \ , \end{aligned}$$

and also

$$\begin{aligned} \mathit{rank}\big(\mathbf{M}(\mathbf{a})\big) &= (p - k_{\mathbf{A}}).k \ , \\ \mathit{rank}\big(J(\mathbf{r}(\mathbf{a}))\big) &= (p - k_{\mathbf{A}}).k \ . \end{aligned}$$

Proof. It suffices to note that $\mathbf{W} \in \mathbb{R}_{+*}^{p \times n}$ and $\mathit{rank}(\widehat{\mathbf{B}}) = k$ lead to

$$\mathit{card}(\Theta(\mathbf{W})) = n \geqslant k \text{ and } \widehat{\mathbf{B}}' = \widehat{\mathbf{B}}$$

and the results follow immediately from Theorem 5.3 and Corollary 5.4.

□

Remark 5.2. More generally, if $\mathbf{W} \in \mathbb{R}_{+*}^{p \times n}$ and $\mathit{rank}(\widehat{\mathbf{B}}) = k_{\widehat{\mathbf{B}}} < k$ it is possible to demonstrate

$$\begin{aligned} \mathit{dim}\big(\mathit{null}(\mathbf{M}(\mathbf{a}))\big) &= p.(k - k_{\widehat{\mathbf{B}}}) + k_{\mathbf{A}}.k_{\widehat{\mathbf{B}}} \ , \\ \mathit{rank}\big(\mathbf{M}(\mathbf{a})\big) &= k_{\widehat{\mathbf{B}}}.(p - k_{\mathbf{A}}) \ , \end{aligned}$$

and also

$$\begin{aligned} k_{\mathbf{A}}.k \leqslant \mathit{dim}\big(\mathit{null}(J(\mathbf{r}(\mathbf{a})))\big) &\leqslant p.(k - k_{\widehat{\mathbf{B}}}) + k_{\mathbf{A}}.k_{\widehat{\mathbf{B}}} \ , \\ k_{\widehat{\mathbf{B}}}.(p - k_{\mathbf{A}}) \leqslant \quad \mathit{rank}\big(J(\mathbf{r}(\mathbf{a}))\big) \quad &\leqslant k.(p - k_{\mathbf{A}}) \ , \end{aligned}$$

but we omit the details since these results are not very useful in practical applications in which $\mathit{rank}(\mathbf{A}) = \mathit{rank}(\widehat{\mathbf{B}}) = k$ is the rule. ■

If the hypotheses of Theorem 5.3 and Corollaries 5.4 and 5.5 are satisfied, we know precisely the dimensions of $\mathit{null}(J(\mathbf{r}(\mathbf{a}))) = \mathit{null}(\mathbf{M}(\mathbf{a}))$ and of its orthogonal complement in $\mathbb{R}^{k.p}$, e.g.,

$$\mathit{dim}\big(\mathit{null}(J(\mathbf{r}(\mathbf{a})))\big) = k_{\mathbf{A}}.k \text{ and } \mathit{dim}\big(\mathit{null}(J(\mathbf{r}(\mathbf{a})))^{\perp}\big) = (p - k_{\mathbf{A}}).k \; .$$

Furthermore, while Theorem 5.3, Corollaries 5.4 and 5.5 are valid for any $\mathbf{A} \in \mathbb{R}^{p \times k}$, we recall that the condition $\mathit{rank}(\mathbf{A}) = k$ is required for $\psi(.)$ to be differentiable, which implies that for all practical WLRA applications, we will have

$$\mathit{dim}\big(\mathit{null}(J(\mathbf{r}(\mathbf{a})))\big) = k.k \text{ and } \mathit{dim}\big(\mathit{null}(J(\mathbf{r}(\mathbf{a})))^{\perp}\big) = (p - k).k \; .$$

Then, if (orthonormal or not) bases of $\mathit{null}(J(\mathbf{r}(\mathbf{a}))) = \mathit{null}(\mathbf{M}(\mathbf{a}))$ and of its orthogonal complement are available, it is easy to obtain the minimal 2-norm solution and also all solutions of the linear least-squares problem

$$\min_{d\mathbf{a} \in \mathbb{R}^{p.k}} \frac{1}{2} \| \mathbf{r}(\mathbf{a}) + J\big(\mathbf{r}(\mathbf{a})\big) d\mathbf{a} \|_2^2 \; .$$

More, precisely, if the columns of $\mathbf{N} \in \mathbb{R}^{k.p \times k_{\mathbf{A}}.k}$ and $\mathbf{N}^{\perp} \in \mathbb{R}^{k.p \times (p - k_{\mathbf{A}}).k}$ form, respectively, (orthonormal) bases of $\mathit{null}(J(\mathbf{r}(\mathbf{a})))$ and $\mathit{null}(J(\mathbf{r}(\mathbf{a})))^{\perp}$, then $d\mathbf{a}_{gn} = d\mathbf{a}_{min}$ can be computed in a two-step procedure. In the first step, we need to solve the following reduced (and nonsingular) linear least-squares problem

$$\min_{d\mathbf{a} \in \mathbb{R}^{(p - k_{\mathbf{A}}).k}} \frac{1}{2} \| \mathbf{r}(\mathbf{a}) + J\big(\mathbf{r}(\mathbf{a})\big) \mathbf{N}^{\perp} d\mathbf{a} \|_2^2 \; ,$$

which has an unique solution as demonstrated above, say $d\bar{\mathbf{a}}_{gn} \in \mathbb{R}^{(p - k_{\mathbf{A}}).k}$. Next, in a second step, we obtain $d\mathbf{a}_{gn} = d\mathbf{a}_{min}$ as the matrix-vector product

$$d\mathbf{a}_{gn} = \mathbf{N}^{\perp} d\bar{\mathbf{a}}_{gn} \; .$$

Then, the general solution $d\widehat{\mathbf{a}}$ of the full linear least-squares problem can be also computed in a third step as

$$d\widehat{\mathbf{a}} = d\mathbf{a}_{min} + \mathbf{N}\mathbf{c} = \mathbf{N}^{\perp} d\bar{\mathbf{a}}_{gn} + \mathbf{N}\mathbf{c} \; ,$$

with $\mathbf{c} \in \mathbb{R}^{k_{\mathbf{A}}.k}$ is a vector of $k_{\mathbf{A}}.k$ arbitrary constants. Note that if we use the matrix $-\mathbf{M}(\mathbf{a})$ as an approximate Jacobian instead, we can also find all the solutions of the linear least-squares problem

$$\min_{d\mathbf{a} \in \mathbb{R}^{p.k}} \frac{1}{2} \| \mathbf{r}(\mathbf{a}) - \mathbf{M}(\mathbf{a}) d\mathbf{a} \|_2^2$$

in three steps using again the bases $\mathbf{N}$ and $\mathbf{N}^{\perp}$, but in the first step, we need to solve the reduced and nonsingular linear least-squares problem

$$\min_{d\mathbf{a} \in \mathbb{R}^{(p - k_{\mathbf{A}}).k}} \frac{1}{2} \| \mathbf{r}(\mathbf{a}) - \mathbf{M}(\mathbf{a}) \mathbf{N}^{\perp} d\mathbf{a} \|_2^2 \; ,$$

instead of the previous one involving the rank-deficient matrix $J(\mathbf{r}(\mathbf{a}))$.

Alternatively, it is also possible to obtain $d\mathbf{a}_{gn}$ and all solutions of the above Gauss-Newton linear least-squares problems involving the Jacobian matrix $J(\mathbf{r}(\mathbf{a}))$ or its approximation $-\mathbf{M}(\mathbf{a})$ using only a basis $\mathbf{N}$ of $\mathit{null}(J(\mathbf{r}(\mathbf{a}))) = \mathit{null}(\mathbf{M}(\mathbf{a}))$ as first noted by Okatani et al. [150]. Using the fact, demonstrated above, that $d\mathbf{a}_{gn} = d\mathbf{a}_{min}$ is the unique solution of these linear least-squares problems, which belongs to $\mathit{null}(J(\mathbf{r}(\mathbf{a})))^{\perp}$, it is not difficult to see that $d\mathbf{a}_{gn}$ is also the unique solution among the infinite set of solutions $d\widehat{\mathbf{a}}$ of these linear least-squares problems, which verifies the equality

$$\mathbf{N}^T d\widehat{\mathbf{a}} = \mathbf{0}^{k_{\mathbf{A}}.k} \; ,$$

if the columns of $\mathbf{N}$ form a basis of $null(J(\mathbf{r}(\mathbf{a}))) = null(\mathbf{M}(\mathbf{a}))$.

In addition, using results in the demonstration of Theorem 5.2 and assuming for simplicity that $k_{\mathbf{A}} = k$, e.g., $\mathbf{A}$ is of full column-rank, we observe that the matrix defined by

$$\mathbf{N} = \mathbf{K}_{(p,k)}(\mathbf{I}_k \otimes \mathbf{A})$$

is also of full column-rank (e.g., $rank(\mathbf{N}) = k.k$), and that

$$null\Big(J\big(\mathbf{r}(\mathbf{a})\big)\Big) = null\big(\mathbf{M}(\mathbf{a})\big) = ran(\mathbf{N}) \ ,$$

if the hypotheses of Theorem 5.3 are satisfied. In other words, it is very easy to compute a basis $\mathbf{N}$ of $null\big(J(\mathbf{r}(\mathbf{a}))\big) = null(\mathbf{M}(\mathbf{a}))$ in practical applications or even an orthonormal basis of this linear subspace of $\mathbb{R}^{p.k}$ with the help of Corollary 5.6 demonstrated below.

Furthermore, it is also very easy to introduce the linear constraint $\mathbf{N}^T d\mathbf{a} = \mathbf{0}^{k_{\mathbf{A}}.k}$ in the linear least-squares problems, which must be solved for computing the correction vector $d\mathbf{a}_{gn}$ at each iteration of the Gauss-Newton algorithm, as follows

$$d\mathbf{a}_{gn} = \text{Arg} \min_{d\mathbf{a} \in \mathbb{R}^{p.k}} \frac{1}{2}\|\mathbf{r}(\mathbf{a}) + J(\mathbf{r}(\mathbf{a}))d\mathbf{a}\|_2^2 + \frac{1}{2}\|\mathbf{N}^T d\mathbf{a}\|_2^2 \ ,$$

or

$$d\mathbf{a}_{gn} = \text{Arg} \min_{d\mathbf{a} \in \mathbb{R}^{p.k}} \frac{1}{2}\|\mathbf{r}(\mathbf{a}) - \mathbf{M}(\mathbf{a})d\mathbf{a}\|_2^2 + \frac{1}{2}\|\mathbf{N}^T d\mathbf{a}\|_2^2 \ ,$$

if we use the approximate Jacobian matrix $-\mathbf{M}(\mathbf{a})$. These two "constrained" linear least-squares problems are also, respectively, equivalent to the following standard linear least-squares problems

$$d\mathbf{a}_{gn} = \text{Arg} \min_{d\mathbf{a} \in \mathbb{R}^{p.k}} \frac{1}{2}\Big\| \begin{bmatrix} \mathbf{r}(\mathbf{a}) \\ \mathbf{0}^{k_{\mathbf{A}}.k} \end{bmatrix} + \begin{bmatrix} J(\mathbf{r}(\mathbf{a})) \ , \\ \mathbf{N}^T \end{bmatrix} d\mathbf{a} \Big\|_2^2$$

and

$$d\mathbf{a}_{gn} = \text{Arg} \min_{d\mathbf{a} \in \mathbb{R}^{p.k}} \frac{1}{2}\Big\| \begin{bmatrix} \mathbf{r}(\mathbf{a}) \\ \mathbf{0}^{k_{\mathbf{A}}.k} \end{bmatrix} - \begin{bmatrix} \mathbf{M}(\mathbf{a}) \\ \mathbf{N}^T \end{bmatrix} d\mathbf{a} \Big\|_2^2 \ ,$$

which are both easily solved and have an unique solution as the associated coefficient matrices are nonsingular if the columns of $\mathbf{N}$ form a basis of $null(J(\mathbf{r}(\mathbf{a}))) = null(\mathbf{M}(\mathbf{a}))$.

To demonstrate this result, it suffices to show that the null space of these matrices is reduced to the zero vector. To this end, we first observe that

$$null\Big(\begin{bmatrix} J(\mathbf{r}(\mathbf{a})) \\ \mathbf{N}^T \end{bmatrix} \Big) = null\big(J(\mathbf{r}(\mathbf{a}))\big) \cap null(\mathbf{N}^T) = null\Big(\begin{bmatrix} \mathbf{M}(\mathbf{a}) \\ \mathbf{N}^T \end{bmatrix} \Big) \ ,$$

if the hypotheses of Theorem 5.3 are satisfied. Moreover, the following relationships hold (see equation (2.1) in Subsection 2.1)

$$null(\mathbf{N}^T)^{\perp} = ran(\mathbf{N}) = null\big(J(\mathbf{r}(\mathbf{a}))\big) = null(\mathbf{M}(\mathbf{a})) \ ,$$

and this implies, finally, that

$$null(\begin{bmatrix} J(\mathbf{r}(\mathbf{a})) \\ \mathbf{N}^T \end{bmatrix}) = null(\begin{bmatrix} \mathbf{M}(\mathbf{a}) \\ \mathbf{N}^T \end{bmatrix}) = null(\mathbf{N}^T)^{\perp} \cap null(\mathbf{N}^T) = \{\mathbf{0}^{p.k}\} \ , \tag{5.24}$$

which demonstrates that the corresponding matrices are effectively nonsingular if the columns of $\mathbf{N}$ form a basis of $null(J(\mathbf{r}(\mathbf{a}))) = null(\mathbf{M}(\mathbf{a}))$.

The above results can also be used to improve the Levenberg-Marquardt algorithm. For example, using an exact Jacobian matrix, $J(\mathbf{r}(\mathbf{a}))$, an accurate Levenberg-Marquardt's correction vector $d\mathbf{a}_{lm}$

can also be obtained in two steps. First, by solving the following reduced and regularized linear least-squares problem

$$\min_{d\mathbf{a}\in\mathbb{R}^{(p-k_{\mathbf{A}}).k}} \frac{1}{2}\left\| \begin{bmatrix} \mathbf{r}(\mathbf{a}) \\ \mathbf{0}^{(p-k_{\mathbf{A}}).k} \end{bmatrix} + \begin{bmatrix} J(\mathbf{r}(\mathbf{a}))\mathbf{N}^{\perp} \\ \sqrt{\lambda}\mathbf{D} \end{bmatrix} d\mathbf{a} \right\|_2^2 = \frac{1}{2}\left\|\mathbf{r}(\mathbf{a}) + J(\mathbf{r}(\mathbf{a}))\mathbf{N}^{\perp}d\mathbf{a}\right\|_2^2 + \frac{\lambda}{2}\left\|\mathbf{D}d\mathbf{a}\right\|_2^2 ,$$

where λ is the damping Marquardt parameter and $\mathbf{D}$ is a diagonal scaling matrix of dimension $(p-k_{\mathbf{A}}).k$. This damped linear least-squares problem has always an unique solution, $d\bar{\mathbf{a}}_{lm}$, even when λ tends to zero if the hypotheses of Theorem 5.3 are satisfied. Once $d\bar{\mathbf{a}}_{lm}$ has been found, $d\mathbf{a}_{lm}$ can be computed by the matrix-vector product

$$d\mathbf{a}_{lm} = \mathbf{N}^{\perp}d\bar{\mathbf{a}}_{lm} ,$$

as for the correction vector $d\mathbf{a}_{gn}$ in the Gauss-Newton algorithm.

Alternatively, the linear constraint

$$\mathbf{N}^T d\mathbf{a} = \mathbf{0}^{k_{\mathbf{A}}.k}$$

can also be introduced in the linear least-squares problem, which must be solved for computing the correction vector at each iteration of the Levenberg-Marquardt algorithm. As an illustration, if we use an exact Jacobian matrix, a correction vector can be computed as

$$d\mathbf{a}_{lm}^{'} = \text{Arg}\min_{d\mathbf{a}\in\mathbb{R}^{p.k}} \frac{1}{2}\|\mathbf{r}(\mathbf{a}) + J(\mathbf{r}(\mathbf{a}))d\mathbf{a}\|_2^2 + \frac{1}{2}\|\mathbf{N}^T d\mathbf{a}\|_2^2 + \frac{\lambda}{2}\left\|\mathbf{D}d\mathbf{a}\right\|_2^2 ,$$

where λ is the damping Marquardt parameter and $\mathbf{D}$ is now a diagonal scaling matrix of dimension $k.p$. This problem is also equivalent to the standard linear least-squares problem

$$d\mathbf{a}_{lm}^{'} = \text{Arg}\min_{d\mathbf{a}\in\mathbb{R}^{p.k}} \frac{1}{2}\left\| \begin{bmatrix} \mathbf{r}(\mathbf{a}) \\ \mathbf{0}^{k_{\mathbf{A}}.k} \\ \mathbf{0}^{k.p} \end{bmatrix} + \begin{bmatrix} J(\mathbf{r}(\mathbf{a})) \\ \mathbf{N}^T \\ \sqrt{\lambda}\mathbf{D} \end{bmatrix} d\mathbf{a} \right\|_2^2 .$$

However, contrary to the case of the Gauss-Newton method, the correction vectors, $d\mathbf{a}_{lm}$ and $d\mathbf{a}_{lm}^{'}$, obtained by these two alternative formulations of the Levenberg-Marquardt algorithm will differ in general. More precisely, if $\lambda \neq 0$, we cannot assume that it always exists $\mathbf{c} \in \mathbb{R}^{(p-k_{\mathbf{A}}).k}$ such that

$$d\mathbf{a}_{lm}^{'} = \mathbf{N}^{\perp}\mathbf{c} ,$$

as we only have $\mathbf{N}^T d\mathbf{a}_{lm}^{'} \approx \mathbf{0}^{k_{\mathbf{A}}.k}$, but not exactly $\mathbf{N}^T d\mathbf{a}_{lm}^{'} = \mathbf{0}^{k_{\mathbf{A}}.k}$ as for $d\mathbf{a}_{lm}$. Thus, in these conditions, $d\mathbf{a}_{lm}^{'} \notin \mathit{null}(J(\mathbf{r}(\mathbf{a})))^{\perp}$, while $d\mathbf{a}_{lm} \in \mathit{null}(J(\mathbf{r}(\mathbf{a})))^{\perp}$, and these two correction vectors will not be equal in general. Moreover, the fact that $d\mathbf{a}_{lm} \in \mathit{null}(J(\mathbf{r}(\mathbf{a})))^{\perp}$ implies that the version of the Levenberg-Marquardt algorithm using this correction vector can also be considered as a Riemannian optimization algorithm operating directly on the Grassmann manifold $\text{Gr}(p,k)$ [3][11] in the same way as the Gauss-Newton algorithm discussed above (see below for details), while the version using $d\mathbf{a}_{lm}^{'}$ as a correction step does not enjoy this theoretical property.

At first sight, the approach using only a basis $\mathbf{N}$ of $\mathit{null}\big(J(\mathbf{r}(\mathbf{a}))\big) = \mathit{null}(\mathbf{M}(\mathbf{a}))$ and a linear constraint for computing the Gauss-Newton and Levenberg-Marquardt directions at each iteration, seems to be much cheaper than the first approach, which needs to compute a nonsingular matrix $\mathbf{N}^{\perp} \in \mathbb{R}^{k.p\times(p-k_{\mathbf{A}}).k}$ whose columns form a basis of $\mathit{null}(J(\mathbf{r}(\mathbf{a})))^{\perp}$, to multiply the huge Jacobian matrix $J(\mathbf{r}(\mathbf{a})$ (or its approximation) by this matrix $\mathbf{N}^{\perp}$ and, finally, to compute the matrix-vector products $d\mathbf{a}_{gn} = \mathbf{N}^{\perp}d\bar{\mathbf{a}}_{gn}$ or $d\mathbf{a}_{lm} = \mathbf{N}^{\perp}d\bar{\mathbf{a}}_{lm}$ in the case of the Levenberg-Marquardt algorithm [150].

However, the next corollary shows that the overhead cost incurred by the first approach can be drastically reduced since it is easy to obtain orthonormal bases of $\mathit{null}\big(J(\mathbf{r}(\mathbf{a}))\big)$ and its orthogonal complement in $\mathbb{R}^{k.p}$ if the conditions of Theorem 5.3 are fulfilled. Furthermore, thanks to the particular form of these orthonormal bases, the above matrix product between the Jacobian matrix (or

its approximation) and a basis of $\mathit{null}\big(J(\mathbf{r}(\mathbf{a}))\big)^{\perp}$ can be computed very efficiently with almost the same cost as evaluating the Jacobian matrix or its approximation itself as also demonstrated in this corollary.

Corollary 5.6. With the same notations and hypotheses as in Theorem 5.3, let $\mathbf{O} \in \mathbb{O}^{p\times k_{\mathbf{A}}}$ be an orthonormal basis of $\mathit{ran}(\mathbf{A})$ and $\mathbf{O}^{\perp} \in \mathbb{O}^{p\times(p-k_{\mathbf{A}})}$ be an orthonormal basis of $\mathit{ran}(\mathbf{A})^{\perp}$, then

$$\bar{\mathbf{O}} = \mathbf{K}_{(p,k)}(\mathbf{I}_k \otimes \mathbf{O}) \text{ is an orthonormal basis of } \mathit{null}\big(J(\mathbf{r}(\mathbf{a}))\big) = \mathit{null}(\mathbf{M}(\mathbf{a}))\ ,$$
$$\bar{\mathbf{O}}^{\perp} = \mathbf{K}_{(p,k)}(\mathbf{I}_k \otimes \mathbf{O}^{\perp}) \text{ is an orthonormal basis of } \mathit{null}\big(J(\mathbf{r}(\mathbf{a}))\big)^{\perp} = \mathit{null}(\mathbf{M}(\mathbf{a}))^{\perp}\ .$$

Furthermore, we have

$$\mathbf{M}(\mathbf{a})\bar{\mathbf{O}}^{\perp} = \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\mathbf{U}(\mathbf{a})\bar{\mathbf{O}}^{\perp} = \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\mathit{diag}\big(\mathit{vec}(\sqrt{\mathbf{W}})\big)\big(\widehat{\mathbf{B}}^{T}\otimes\mathbf{O}^{\perp}\big)\ ,$$
$$\mathbf{L}(\mathbf{a})\bar{\mathbf{O}}^{\perp} = \big(\mathbf{F}(\mathbf{a})^{+}\big)^{T}\mathbf{V}(\mathbf{a})\bar{\mathbf{O}}^{\perp} = \big(\mathbf{F}(\mathbf{a})^{+}\big)^{T}\Big(\big(\mathbf{W}\odot P_{\Omega}(\mathbf{X}-\mathbf{A}\widehat{\mathbf{B}})\big)^{T}\mathbf{O}^{\perp}\otimes\mathbf{I}_k\Big)\mathbf{K}_{(p-k,p)}\ .$$

Proof. Using the results of Theorem 5.2, we first observe that

$$\mathit{null}\big(J(\mathbf{r}(\mathbf{a}))\big) = \mathit{null}\big(\mathbf{M}(\mathbf{a})\big) = \mathit{ran}(\mathbf{N})\ ,$$

where $\mathbf{N} = \mathbf{K}_{(p,k)}(\mathbf{I}_k \otimes \mathbf{A})$.

Now, since $\mathbf{O}$ is an orthonormal basis of the column space of $\mathbf{A}$, then $\exists\mathbf{C} \in \mathbb{R}^{k_{\mathbf{A}}\times k}$ such that $\mathbf{A} = \mathbf{OC}$, $\mathit{rank}(\mathbf{C}) = k_{\mathbf{A}}$ and $\mathbf{C}$ is uniquely determined (see Theorem 1 of Marsaglia and Styan [126]). Moreover, since $\mathbf{C}$ has full row-rank, $\mathbf{C}$ admits a right-inverse $\mathbf{R}$ such that $\mathbf{CR} = \mathbf{I}_{k_{\mathbf{A}}}$. From this equality, we deduce that

$$\mathbf{AR} = \mathbf{OCR} = \mathbf{O}\ .$$

Using these properties and equation (2.33), we have, $\forall\mathbf{Z} \in \mathbb{R}^{k\times k}$,

$$\begin{aligned}(\mathbf{I}_k \otimes \mathbf{A})\mathit{vec}(\mathbf{Z}) &= (\mathbf{I}_k \otimes \mathbf{OC})\mathit{vec}(\mathbf{Z})\\ &= \mathit{vec}(\mathbf{OCZ})\\ &= (\mathbf{I}_k \otimes \mathbf{O})\mathit{vec}(\mathbf{CZ})\ ,\end{aligned}$$

and, $\forall\mathbf{T} \in \mathbb{R}^{k_{\mathbf{A}}\times k_{\mathbf{A}}}$,

$$\begin{aligned}(\mathbf{I}_k \otimes \mathbf{O})\mathit{vec}(\mathbf{T}) &= (\mathbf{I}_k \otimes \mathbf{AR})\mathit{vec}(\mathbf{T})\\ &= \mathit{vec}(\mathbf{ART})\\ &= (\mathbf{I}_k \otimes \mathbf{A})\mathit{vec}(\mathbf{RT})\ .\end{aligned}$$

From these equalities, it can be easily proved that

$$\mathit{ran}(\mathbf{N}) = \mathit{ran}\big(\mathbf{K}_{(p,k)}(\mathbf{I}_k \otimes \mathbf{O})\big) = \mathit{ran}(\bar{\mathbf{O}})\ .$$

Moreover, $\bar{\mathbf{O}} = \mathbf{K}_{(p,k)}(\mathbf{I}_k \otimes \mathbf{O})$ is a matrix with orthonormal columns since:

$$\begin{aligned}\big(\mathbf{K}_{(p,k)}(\mathbf{I}_k \otimes \mathbf{O})\big)^{T}\mathbf{K}_{(p,k)}(\mathbf{I}_k \otimes \mathbf{O}) &= (\mathbf{I}_k \otimes \mathbf{O}^{T})(\mathbf{I}_k \otimes \mathbf{O})\\ &= (\mathbf{I}_k \otimes \mathbf{O}^{T}\mathbf{O})\\ &= \mathbf{I}_k \otimes \mathbf{I}_{k_{\mathbf{A}}}\\ &= \mathbf{I}_{k.k_{\mathbf{A}}}\ .\end{aligned}$$

A similar argument applies to $\bar{\mathbf{O}}^{\perp} = \mathbf{K}_{(p,k)}(\mathbf{I}_k \otimes \mathbf{O}^{\perp})$. Finally, we have

$$\begin{aligned}\big(\mathbf{K}_{(p,k)}(\mathbf{I}_k \otimes \mathbf{O})\big)^{T}\mathbf{K}_{(p,k)}(\mathbf{I}_k \otimes \mathbf{O}^{\perp}) &= (\mathbf{I}_k \otimes \mathbf{O}^{T})(\mathbf{I}_k \otimes \mathbf{O}^{\perp})\\ &= (\mathbf{I}_k \otimes \mathbf{O}^{T}\mathbf{O}^{\perp})\\ &= \mathbf{I}_k \otimes \mathbf{0}^{k_A\times(p-k_A)}\\ &= \mathbf{0}^{k.k_A\times k.(p-k_A)}\ ,\end{aligned}$$

and it follows that $\bar{\mathbf{O}}$ is an orthonormal basis of $\mathit{null}\big(J(\mathbf{r}(\mathbf{a}))\big)$ and $\bar{\mathbf{O}}^{\perp}$ is an orthonormal basis of $\mathit{null}\big(J(\mathbf{r}(\mathbf{a}))\big)^{\perp}$ as stated in the corollary. Furthermore, the columns of the matrix $\begin{bmatrix}\bar{\mathbf{O}} & \bar{\mathbf{O}}^{\perp}\end{bmatrix}$ form an orthonormal basis of $\mathbb{R}^{k.p}$.

Finally, to demonstrate the second part of the corollary, let us evaluate compactly the matrix products $\mathbf{U}\bar{\mathbf{O}}^{\perp}$ and $\mathbf{V}\bar{\mathbf{O}}^{\perp}$ using the properties of the Kronecker product, commutation matrix and *vec* operator stated in Subsection 2.2. We have

$$\begin{aligned}\mathbf{U}\bar{\mathbf{O}}^{\perp} &= \mathit{diag}\big(\mathit{vec}(\sqrt{\mathbf{W}})\big)(\widehat{\mathbf{B}}^T \otimes \mathbf{I}_p)\mathbf{K}_{(k,p)}\mathbf{K}_{(p,k)}(\mathbf{I}_k \otimes \mathbf{O}^{\perp}) \\ &= \mathit{diag}\big(\mathit{vec}(\sqrt{\mathbf{W}})\big)(\widehat{\mathbf{B}}^T \otimes \mathbf{I}_p)(\mathbf{I}_k \otimes \mathbf{O}^{\perp}) \\ &= \mathit{diag}\big(\mathit{vec}(\sqrt{\mathbf{W}})\big)(\widehat{\mathbf{B}}^T \otimes \mathbf{O}^{\perp}) \ ,\end{aligned}$$

and also

$$\begin{aligned}\mathbf{V}\bar{\mathbf{O}}^{\perp} &= \big((\mathbf{W} \odot P_{\Omega}(\mathbf{X} - \mathbf{A}\widehat{\mathbf{B}}))^T \otimes \mathbf{I}_k\big)\mathbf{K}_{(p,k)}(\mathbf{I}_k \otimes \mathbf{O}^{\perp}) \\ &= \big((\mathbf{W} \odot P_{\Omega}(\mathbf{X} - \mathbf{A}\widehat{\mathbf{B}}))^T \otimes \mathbf{I}_k\big)(\mathbf{O}^{\perp} \otimes \mathbf{I}_k)\mathbf{K}_{(p-k,k)} \\ &= \big((\mathbf{W} \odot P_{\Omega}(\mathbf{X} - \mathbf{A}\widehat{\mathbf{B}}))\big)^T\mathbf{O}^{\perp} \otimes \mathbf{I}_k\big)\mathbf{K}_{(p-k,p)} \ ,\end{aligned}$$

which concludes the demonstration of the corollary.

□

Using Corollary 5.6 and the preceding results, we can write the correction vectors $d\mathbf{a}_{gn}$ and $d\mathbf{a}_{lm}$ of the Gauss-Newton and Levenberg-Marquardt algorithms as the matrix-vector products

$$d\mathbf{a}_{gn} = \bar{\mathbf{O}}^{\perp} d\bar{\mathbf{a}}_{gn} \text{ and } d\mathbf{a}_{lm} = \bar{\mathbf{O}}^{\perp} d\bar{\mathbf{a}}_{lm} \ ,$$

where $d\bar{\mathbf{a}}_{gn}$ and $d\bar{\mathbf{a}}_{lm}$ are, respectively, the solutions of the problems

$$d\bar{\mathbf{a}}_{gn} = \text{Arg} \min_{d\mathbf{a} \in \mathbb{R}^{(p-k_{\mathbf{A}}).k}} \frac{1}{2}\big\|\mathbf{r}(\mathbf{a}) + J\big(\mathbf{r}(\mathbf{a})\big)\bar{\mathbf{O}}^{\perp} d\mathbf{a}\big\|_2^2$$

and

$$d\bar{\mathbf{a}}_{lm} = \text{Arg} \min_{d\mathbf{a} \in \mathbb{R}^{(p-k_{\mathbf{A}}).k}} \frac{1}{2}\big\|\mathbf{r}(\mathbf{a}) + J\big(\mathbf{r}(\mathbf{a})\big)\bar{\mathbf{O}}^{\perp} d\mathbf{a}\big\|_2^2 + \frac{\lambda}{2}\big\|\mathbf{D} d\mathbf{a}\big\|_2^2 \ ,$$

or of similar linear least-squares problems involving the approximate Jacobian matrix $-\mathbf{M}(\mathbf{a})$ instead of $J(\mathbf{r}(\mathbf{a}))$. Defining $d\mathbf{A}_{gn} \in \mathbb{R}^{p \times k}$ and $d\bar{\mathbf{A}}_{gn} \in \mathbb{R}^{(p-k_{\mathbf{A}}) \times k}$ such that $d\mathbf{a}_{gn} = \mathit{vec}(d\mathbf{A}_{gn}^T)$ and $d\bar{\mathbf{a}}_{gn} = \mathit{vec}(d\bar{\mathbf{A}}_{gn})$, we have (using equation (2.33) and Lemma 2.2 in Subsection 2.2)

$$\mathit{vec}(d\mathbf{A}_{gn}^T) = \mathbf{K}_{(p,k)}(\mathbf{I}_k \otimes \mathbf{O}^{\perp})\mathit{vec}(d\bar{\mathbf{A}}_{gn}) = \mathbf{K}_{(p,k)}\mathit{vec}(\mathbf{O}^{\perp} d\bar{\mathbf{A}}_{gn}) = \mathit{vec}\big((\mathbf{O}^{\perp} d\bar{\mathbf{A}}_{gn})^T\big) \ ,$$

which implies that $d\mathbf{A}_{gn} = \mathbf{O}^{\perp} d\bar{\mathbf{A}}_{gn}$. Obviously, the equality $d\mathbf{A}_{lm} = \mathbf{O}^{\perp} d\bar{\mathbf{A}}_{lm}$ can be derived in a similar fashion.

Thus, the columns of the perturbation matrices $d\mathbf{A}_{gn}$ and $d\mathbf{A}_{lm}$ belong to $\mathit{ran}(\mathbf{O}^{\perp}) = \mathit{ran}(\mathbf{A})^{\perp}$. In other words, these variations of the variable projection Gauss-Newton and Levenberg-Marquardt methods described above to deal with the singularity of the Jacobian matrix $J(\mathbf{r}(\mathbf{a}))$, or of its approximation $-\mathbf{M}(\mathbf{a})$, consider only search directions of the form $d\mathbf{A} = \mathbf{O}^{\perp}\mathbf{C}$ where $\mathbf{C} \in \mathbb{R}^{(p-k_{\mathbf{A}}) \times k}$. This is consistent with Remark 5.1 and the fact that we have only $k.(p-k)$ degrees of freedom to update $\mathbf{A}$ if $\mathit{rank}(\mathbf{A}) = k$ at each iteration of the Gauss-Newton or Levenberg-Marquardt algorithms.

In addition, if $\mathbf{W} \in \mathbb{R}_{+*}^{p \times n}$, the unvectorized form of the cost function $\psi(.)$ (e.g., $\psi \circ h^{-1}(.)$ where $h^{-1}(\mathbf{A}) = \mathit{vec}(\mathbf{A}^T) = \mathbf{a}, \forall \mathbf{A} \in \mathbb{R}^{p \times k}$, with $h(.)$ and $h^{-1}(.)$ defined in equation (3.29) of Subsection 3.4), which is used in the (VP1) formulation of the WLRA problem, is smooth (e.g., of class C^{∞}) over the subdomains $\mathbb{R}_k^{p \times k}$ or $\mathbb{O}^{p \times k}$ according to Corollaries 3.3 and 5.1. Note that this

(unvectorized) cost function $\psi(.)$ can be an instance of the (VP1) formulation of the cost function $g_\lambda(.)$ introduced by Boumal and Absil [13][14], defined in equation (3.18), and already discussed in Subsection 3.3 since its associated weight matrix $\mathbf{W}_\lambda \in \mathbb{R}^{p\times n}_{+*}$ if $\lambda > 0$. In the same conditions, if $\mathbf{A} \in \mathbb{R}^{p\times k}_k$ or $\mathbf{A} \in \mathbb{O}^{p\times k}$, we have $\mathit{rank}(\mathbf{M}(\mathbf{a})) = \mathit{rank}(J(\mathbf{r}(\mathbf{a}))) = k.(p-k)$ according to Corollary 5.5. Using these different results, we can recast the variable projection formulation of the WLRA problem as an optimization problem on the Grassmann manifold $\mathrm{Gr}(p,k)$ [14][11] and the above variable projection Gauss-Newton and Levenberg-Marquardt methods to solve the WLRA problem as Riemannian optimization algorithms operating on this Grassmann manifold [3][11] as their numerical behavior only depends on $\mathring{\mathbf{A}} = \mathit{ran}(\mathbf{A}) \in \mathrm{Gr}(p,k)$, for $\mathbf{A} \in \mathbb{R}^{p\times k}_k$ or $\mathbf{A} \in \mathbb{O}^{p\times k}$, and not on the arbitrarily chosen matrix $\mathbf{A}$ to represent $\mathit{ran}(\mathbf{A})$ according to Corollaries 3.1, 3.2 and Remark 3.7.

More precisely, the smooth function $\psi \circ h^{-1}(.)$ defined on the smooth submanifold $\mathbb{R}^{p\times k}_k$ embedded in $\mathbb{R}^{p\times k}$ is invariant on the equivalence classes of the equivalence relation $\sim$ defined on $\mathbb{R}^{p\times k}_k$ by, $\forall \mathbf{A}, \mathbf{C} \in \mathbb{R}^{p\times k}_k$,

$$\mathbf{A} \sim \mathbf{C} \text{ if and only if it exists } \mathbf{D} \in \mathbb{R}^{k\times k}_k \text{ such that } \mathbf{A} = \mathbf{C}\mathbf{D} \ ,$$

according to Corollary 3.2. In this setting, we can say that $\mathrm{Gr}(p,k)$ is the quotient of $\mathbb{R}^{p\times k}_k$ by the action of the group $\mathbb{R}^{k\times k}_k$ following the terminology introduced in Subsection 2.4. Alternatively, if we prefer to work with orthogonal matrices (e.g., with the Stiefel manifold $\mathbb{O}^{p\times k}$), we can consider the restriction of $\psi \circ h^{-1}(.)$ to $\mathbb{O}^{p\times k}$ and the equivalence relation $\sim$ defined on $\mathbb{O}^{p\times k}$ by, $\forall \mathbf{A}, \mathbf{C} \in \mathbb{O}^{p\times k}$,

$$\mathbf{A} \sim \mathbf{C} \text{ if and only if it exists } \mathbf{D} \in \mathbb{O}^{k\times k} \text{ such that } \mathbf{A} = \mathbf{C}\mathbf{D} \ ,$$

and, similarly, $\psi \circ h^{-1}(.)$ is invariant on the equivalence classes of this equivalence relation according to Corollary 3.2 and $\mathrm{Gr}(p,k)$ is defined now as the quotient of the Stiefel manifold $\mathbb{O}^{p\times k}$ by the action of the orthogonal group $\mathbb{O}^{k\times k}$.

Moreover, as a quotient manifold (see Subsection 2.4 and Chapter 9 of [11]), the Grassmannian admits a tangent space at $\mathring{\mathbf{A}} = \mathit{ran}(\mathbf{A}) \in \mathrm{Gr}(p,k)$, $\forall \mathbf{A} \in \mathbb{R}^{p\times k}_k$ or $\forall \mathbf{A} \in \mathbb{O}^{p\times k}$, designed by $\mathcal{T}_{\mathring{\mathbf{A}}}\mathrm{Gr}(p,k)$ (in the terminology of Subsection 2.4), which can be identified uniquely with the linear subspace of $\mathbb{R}^{p\times k}$ of dimension $k.(p-k)$ defined by

$$\mathcal{T}_{\mathbf{A}}\mathrm{Gr}(p,k) = \left\{ \mathbf{D} \in \mathbb{R}^{p\times k} \ / \ \mathbf{A}^T\mathbf{D} = \mathbf{0}^{k\times k} \right\} .$$

With this identification, $\mathcal{T}_{\mathring{\mathbf{A}}}\mathrm{Gr}(p,k)$ is nothing else than the horizontal space, $\mathcal{H}_{\mathbf{A}}\mathbb{R}^{p\times k}_k$, of $\mathbb{R}^{p\times k}_k$ at $\mathbf{A} \in \mathbb{R}^{p\times k}_k$ or, alternatively, the horizontal space, $\mathcal{H}_{\mathbf{A}}\mathbb{O}^{p\times k}$, of $\mathbb{O}^{p\times k}$ at $\mathbf{A} \in \mathbb{O}^{p\times k}$; see Subsection 2.4 for details. Furthermore, the orthogonal projector onto $\mathcal{T}_{\mathbf{A}}\mathrm{Gr}(p,k)$ with respect to the Frobenius inner product in $\mathbb{R}^{p\times k}$ is given by

$$\mathbf{P}_{\mathcal{T}_{\mathbf{A}}\mathrm{Gr}(p,k)} : \mathbb{R}^{p\times k} \longrightarrow \mathcal{T}_{\mathbf{A}}\mathrm{Gr}(p,k), \mathbf{D} \mapsto \mathbf{P}_{\mathcal{H}_{\mathbf{A}}\mathbb{R}^{p\times k}_k}(\mathbf{D}) = (\mathbf{I}_p - \mathbf{A}\mathbf{A}^+)\mathbf{D} \ ,$$

or, equivalently, if we prefer to work with the Stiefel submanifold, by

$$\mathbf{P}_{\mathcal{T}_{\mathbf{A}}\mathrm{Gr}(p,k)}(\mathbf{D}) = \mathbf{P}_{\mathcal{H}_{\mathbf{A}}\mathbb{O}^{p\times k}}(\mathbf{D}) = (\mathbf{I}_p - \mathbf{O}\mathbf{O}^T)\mathbf{D} = \mathbf{O}^\perp(\mathbf{O}^\perp)^T\mathbf{D} \ ,$$

where the columns of $\mathbf{O}$ and $\mathbf{O}^\perp$ form, respectively, orthogonal bases of $\mathit{ran}(\mathbf{A})$ and $\mathit{ran}(\mathbf{A})^\perp$, and $\mathbf{P}_{\mathcal{H}_{\mathbf{A}}\mathbb{R}^{p\times k}_k}$ and $\mathbf{P}_{\mathcal{H}_{\mathbf{A}}\mathbb{O}^{p\times k}}$ design, respectively, the orthogonal projectors onto the horizontal subspaces of the tangent spaces $\mathcal{T}_{\mathbf{A}}\mathbb{R}^{p\times k}_k$ and $\mathcal{T}_{\mathbf{A}}\mathbb{O}^{p\times k}$. See Subsection 2.4 and [3][11] for more details on the geometry of smooth manifolds, including the (quotient) Grassmann manifold.

Thus, in this Grassmann manifold framework, we have $d\mathbf{A}_{gn}, d\mathbf{A}_{lm} \in \mathcal{T}_{\mathbf{A}}\mathrm{Gr}(p,k)$, since $d\mathbf{A}_{gn} = \mathbf{O}^\perp d\bar{\mathbf{A}}_{gn}$ and $d\mathbf{A}_{lm} = \mathbf{O}^\perp d\bar{\mathbf{A}}_{lm}$, and this implies that the above Gauss-Newton and Levenberg-Marquardt algorithms can be interpreted exactly as Riemannian optimization methods operating on

the Grassmann manifold $\mathrm{Gr}(p,k)$: at the $(i+1)^{th}$ iteration, these algorithms move on the Grassmann manifold from $\mathbf{A}^i \in \mathbb{R}_k^{p\times k}$ along some direction prescribed by the tangent vectors $d\mathbf{A}_{gn}^i$ or $d\mathbf{A}_{lm}^i$ to

$$\mathbf{A}^{i+1} = \mathbf{A}^i + d\mathbf{A}_{gn}^i \text{ or } \mathbf{A}^{i+1} = \mathbf{A}^i + d\mathbf{A}_{lm}^i \,.$$

As at each iteration, $d\mathbf{A}_{gn}^i$ and $d\mathbf{A}_{lm}^i$ belong to $\mathcal{T}_{\mathbf{A}^i}\mathrm{Gr}(p,k)$, we have

$$(\mathbf{A}^i)^T d\mathbf{A}_{gn}^i = \mathbf{0}^{k\times k} \text{ and } (\mathbf{A}^i)^T d\mathbf{A}_{lm}^i = \mathbf{0}^{k\times k} \,,$$

and, in these conditions, $\mathbf{A}^{i+1}$ is always of full column-rank and so $\mathbf{A}^{i+1} \in \mathbb{R}_k^{p\times k}$, e.g., $\mathring{\mathbf{A}}^{i+1} = \mathit{ran}(\mathbf{A}^{i+1})$ belongs to the Grassmann manifold $\mathrm{Gr}(p,k)$, validating our claim about the nature of these Gauss-Newton and Levenberg-Marquardt algorithms.

Alternatively, if we require that each element of $\mathrm{Gr}(p,k)$ must be represented by an element of the Stiefel manifold $\mathrm{St}(p,k) = \mathbb{O}^{p\times k}$ as in [13][14], it is necessary to perform an additional retraction step to the correct (Stiefel) manifold at each iteration of the above variable projection Gauss-Newton and Levenberg-Marquardt algorithms in order to consider these algorithms as Riemannian optimization methods operating on the Grassmann manifold $\mathrm{Gr}(p,k)$ [3][14][11]. In general terms, a retraction on a manifold can be interpreted as a tool that transforms a tangent update vector at a point of this manifold into a new iterate on this manifold. In other words, at the $(i+1)^{th}$ iteration, in order to move from $\mathbf{O}^i \in \mathrm{St}(p,k) = \mathbb{O}^{p\times k}$ along the tangent vectors $d\mathbf{O}_{gn}^i$ or $d\mathbf{O}_{lm}^i \in \mathcal{T}_{\mathbf{O}^i}\mathrm{Gr}(p,k)$ while remaining on the Stiefel manifold, after computing

$$\mathbf{A}^{i+1} = \mathbf{O}^i + d\mathbf{O}_{gn}^i \text{ or } \mathbf{A}^{i+1} = \mathbf{O}^i + d\mathbf{O}_{lm}^i \,,$$

we need to perform the retraction

$$\mathrm{Retraction}_{\mathbf{O}^i}(d\mathbf{O}_{gn}^i) = \mathrm{Ortho}(\mathbf{O}^i + d\mathbf{O}_{gn}^i) \text{ or } \mathrm{Retraction}_{\mathbf{O}^i}(d\mathbf{O}_{lm}^i) = \mathrm{Ortho}(\mathbf{O}^i + d\mathbf{O}_{lm}^i) \,, \quad (5.25)$$

where $\mathrm{Ortho}(\mathbf{H}) \in \mathrm{St}(p,k) = \mathbb{O}^{p\times k}$ designates the $p \times k$ orthonormal factor of a thin QR or polar decomposition of $\mathbf{H} \in \mathbb{R}^{p\times k}$; see [3][13][1][14][11] for more details on the concept of retraction on manifolds and how these retractions can be computed and used as cheap ways of moving on a specific manifold in Riemannian optimization algorithms.

Interestingly, these results are also very similar to those concerning the algorithms derived in Edelman et al. [51] and Manton et al. [125] for minimizing the cost function $\psi^{**}(.)$ on the Grassmann manifold $\mathrm{Gr}(p,p-k)$, defined in equation (3.27) and discussed in Remark 3.7.

More generally, for $\mathbf{W} \in \mathbb{R}_+^{p\times n}$, $\mathbf{A} \in \mathbb{R}_k^{p\times k}$ and any iterative NLLS algorithms use to solve the (VP1) or (VP2) problems, it is possible to demonstrate that there is almost no loss of generality to restrict the search directions during the iterations to the subspace $\mathit{ran}(\mathbf{A})^\perp = \mathit{ran}(\mathbf{O}^\perp)$ when minimizing the cost function $\psi(.)$, or to $\mathit{ran}(\mathbf{A}) = \mathit{ran}(\mathbf{O})$ when minimizing the cost function $\psi^{**}(.)$ as noted, respectively, in [28] and [51][125]. As an illustration, consider the minimization of $\psi(.)$ in the (VP1) problem and take an arbitrary matrix $\mathbf{A} \in \mathbb{R}_k^{p\times k}$ and an arbitrary perturbation matrix $d\mathbf{A} \in \mathbb{R}^{p\times k}$. With the orthonormal basis $\begin{bmatrix}\mathbf{O} & \mathbf{O}^\perp\end{bmatrix}$ of $\mathbb{R}^p$ derived from $\mathbf{A}$ in Corollary 5.6, where $\mathbf{O} \in \mathbb{O}^{p\times k}$ and $\mathbf{O}^\perp \in \mathbb{O}^{p\times(p-k)}$, the perturbation matrix $d\mathbf{A}$ can be decomposed uniquely as

$$d\mathbf{A} = \mathbf{O}\mathbf{K} + \mathbf{O}^\perp\mathbf{K}^\perp \,,$$

where $\mathbf{K} \in \mathbb{R}^{k\times k}$ and $\mathbf{K}^\perp \in \mathbb{R}^{(p-k)\times k}$, and we have the equalities

$$\mathbf{A} + d\mathbf{A} = (\mathbf{A} + \mathbf{O}\mathbf{K}) + \mathbf{O}^\perp\mathbf{K}^\perp = \mathbf{A}\mathbf{Z} + \mathbf{O}^\perp\mathbf{K}^\perp \,,$$

since $\mathbf{A} + \mathbf{O}\mathbf{K} \in \mathit{ran}(\mathbf{A}) = \mathit{ran}(\mathbf{O})$ implies that it exists $\mathbf{Z} \in \mathbb{R}^{k\times k}$ such that $\mathbf{A} + \mathbf{O}\mathbf{K} = \mathbf{A}\mathbf{Z}$. Assuming further that $\mathbf{Z}$ is non-singular, which will be the rule in most practical applications, we have the equalities

$$\psi(\mathbf{A} + d\mathbf{A}) = \psi(\mathbf{A}\mathbf{Z} + \mathbf{O}^\perp\mathbf{K}^\perp) = \psi\big((\mathbf{A}\mathbf{Z} + \mathbf{O}^\perp\mathbf{K}^\perp)\mathbf{Z}^{-1}\big) = \psi\big(\mathbf{A} + \mathbf{O}^\perp\mathbf{K}^\perp\mathbf{Z}^{-1}\big) \,,$$

where the second equality results from Corollary 3.2. In other words, as noted by Chen [28], for most perturbation matrices $d\mathbf{A}$ around $\mathbf{A}$, there exists a perturbation matrix $d\mathbf{A}' = \mathbf{O}^{\perp}\mathbf{K}^{\perp}\mathbf{Z}^{-1}$ whose range is included in $ran(\mathbf{A})^{\perp}$ and which has exactly the same effect as $d\mathbf{A}$ since $\psi(\mathbf{A} + d\mathbf{A}) = \psi(\mathbf{A} + d\mathbf{A}')$. Thus, for any iterative NLLS algorithm used to minimize $\psi(.)$, at each iteration, it suffices generally to search for a perturbation matrix $d\mathbf{A}'$ such that $ran(d\mathbf{A}') \subset ran(\mathbf{A})^{\perp}$.

Since, in these conditions, $\exists \mathbf{T} \in \mathbb{R}^{(p-k)\times k}$ such that $d\mathbf{A}' = \mathbf{O}^{\perp}\mathbf{T}$, this reduces the number of degrees of freedom, or equivalently the number of parameters, to estimate from $k.p$ to $k.(p-k)$. Moreover, as $\mathbf{A}^T d\mathbf{A}' = \mathbf{0}^{k\times k}$, we are also sure that $\mathbf{A} + d\mathbf{A}'$ is always of full column-rank (e.g., $rank(\mathbf{A} + d\mathbf{A}') = k$) across the iterations, which is required for the validity of the algorithms as otherwise the associated orthogonal projector $\mathbf{P}^{\perp}_{\mathbf{F}(.)}$ is not differentiable at $\mathbf{a}$. This is also a useful benefit of restricting the domain of the perturbation matrices $d\mathbf{A}'$ to $ran(\mathbf{A})^{\perp}$ (e.g., such that $ran(d\mathbf{A}') \subset ran(\mathbf{A})^{\perp}$).

These different properties further justify all the variations of the Gauss-Newton and Levenberg-Marquardt algorithms already described in this subsection, which restrict the search directions to the subspace $ran(\mathbf{A})^{\perp}$. Proceeding with similar arguments, Edelman et al. [51] and Manton et al. [125] have also demonstrated that it suffices to consider perturbation matrices whose range is included in $ran(\mathbf{O})$ at each iteration of any NLLS algorithm used to minimize the cost function $\psi^{**}(.)$ in the (VP2) problem. More precisely, in our notations, these algorithms try to minimize the functional $\psi^{**}(\mathbf{O}^{\perp})$ for $\mathbf{O}^{\perp} \in \mathbb{O}^{p\times(p-k)}$ and at each iteration of these algorithms, the search directions for updating $\mathbf{O}^{\perp}$ are restricted to perturbation matrices of the form $d\mathbf{O}^{\perp} = \mathbf{OT}$ where $\mathbf{O} \in \mathbb{O}^{p\times k}$ with $\mathbf{O}^T\mathbf{O}^{\perp} = \mathbf{0}^{k\times(p-k)}$ and $\mathbf{T} \in \mathbb{R}^{k\times(p-k)}$. Obviously, as for the minimization of $\psi(.)$ in the (VP1) problem, this reduces the dimension of the problem from $p.(p-k)$ to $k.(p-k)$ since $\mathbf{T} \in \mathbb{R}^{k\times(p-k)}$. This confirms that the (VP1) and (VP2) formulations of the WLRA problem are dual of each other and should have a similar performance, but not necessarily the same cost depending on the values of p, n and k.

In the previous linear algebra theorems and corollaries, the matrix $\mathbf{A}$ is never assumed to have full column-rank; in other words, the rank of $\mathbf{A}$, $k_{\mathbf{A}}$, is a free parameter and, consequently, it can be less than k. However, we also recall that the case $k_{\mathbf{A}} < k$ is an anomaly in the framework of the WLRA problem because the condition $rank(\mathbf{A}) = k$ is a necessary condition for the continuity and differentiability of the orthogonal projector $\mathbf{P}_{\mathbf{F}(.)}$ as demonstrated in Theorems 3.11, 3.12 and Corollaries 5.1, 5.2 at the beginning of this subsection. Thus, for the following theorem we will make the natural assumption that $\mathbf{A}$ has full column-rank in order to derive stronger results. This theorem demonstrates that it is easy to localize the $k.k$ linearly dependent columns in $\mathbf{M}(\mathbf{a})$, $\mathbf{L}(\mathbf{a})$ and $J(\mathbf{r}(\mathbf{a}))$ in almost all practical cases if $k_{\mathbf{A}} = k$. This result is also new as far we know.

Theorem 5.4. With the same notations as in Theorem 5.2, if $\mathbf{A}$, $\mathbf{M}(\mathbf{a})$, $\mathbf{L}(\mathbf{a})$ and $J(\mathbf{r}(\mathbf{a}))$ are partitioned, respectively, as

$$\begin{array}{lll}
\mathbf{A} = \begin{bmatrix}\mathbf{A}_1\\ \mathbf{A}_2\end{bmatrix} & \text{with } \mathbf{A}_1 \in \mathbb{R}^{(p-k)\times k} & \text{and } \mathbf{A}_2 \in \mathbb{R}^{k\times k}\,,\\
\mathbf{M}(\mathbf{a}) = \begin{bmatrix}\mathbf{M}(\mathbf{a})_1 & \mathbf{M}(\mathbf{a})_2\end{bmatrix} & \text{with } \mathbf{M}(\mathbf{a})_1 \in \mathbb{R}^{n.p\times k.(p-k)} & \text{and } \mathbf{M}(\mathbf{a})_2 \in \mathbb{R}^{n.p\times k.k}\,,\\
\mathbf{L}(\mathbf{a}) = \begin{bmatrix}\mathbf{L}(\mathbf{a})_1 & \mathbf{L}(\mathbf{a})_2\end{bmatrix} & \text{with } \mathbf{L}(\mathbf{a})_1 \in \mathbb{R}^{n.p\times k.(p-k)} & \text{and } \mathbf{L}(\mathbf{a})_2 \in \mathbb{R}^{n.p\times k.k}\,,\\
J(\mathbf{r}(\mathbf{a})) = \begin{bmatrix}J(\mathbf{r}(\mathbf{a}))_1 & J(\mathbf{r}(\mathbf{a}))_2\end{bmatrix} & \text{with } J(\mathbf{r}(\mathbf{a}))_1 \in \mathbb{R}^{n.p\times k.(p-k)} & \text{and } J(\mathbf{r}(\mathbf{a}))_2 \in \mathbb{R}^{n.p\times k.k}\,,
\end{array}$$

and if $rank(\mathbf{A}_2) = k$, then $\exists \mathbf{Z} \in \mathbb{R}^{k.(p-k)\times k.k}$ such that

$$\mathbf{M}(\mathbf{a})_2 = \mathbf{M}(\mathbf{a})_1\mathbf{Z}\,,\ \mathbf{L}(\mathbf{a})_2 = \mathbf{L}(\mathbf{a})_1\mathbf{Z} \text{ and } J(\mathbf{r}(\mathbf{a}))_2 = J(\mathbf{r}(\mathbf{a}))_1\mathbf{Z}\,.$$

In other words, if $rank(\mathbf{A}_2) = k$, the last $k.k$ columns of $\mathbf{M}(\mathbf{a})$, $\mathbf{L}(\mathbf{a})$ and $J(\mathbf{r}(\mathbf{a}))$ are linearly dependent upon the first $k.(p-k)$ columns of $\mathbf{M}(\mathbf{a})$, $\mathbf{L}(\mathbf{a})$ and $J(\mathbf{r}(\mathbf{a}))$, respectively.

Proof. We will first demonstrate that $\mathbf{M}(\mathbf{a})_2 = \mathbf{M}(\mathbf{a})_1\mathbf{Z}$ for some $\mathbf{Z} \in \mathbb{R}^{k.(p-k)\times k.k}$. To this end, let us derive an explicit expression for the $\mathbf{M}(\mathbf{a})_1$ and $\mathbf{M}(\mathbf{a})_2$ submatrices using the formulation of the $\mathbf{M}(\mathbf{a})$ matrix given by equation (5.15), e.g.,

$$\mathbf{M}(\mathbf{a}) = \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})} diag(vec(\sqrt{\mathbf{W}}))(\widehat{\mathbf{B}}^T \otimes \mathbf{I}_p)\mathbf{K}_{(k,p)} \ .$$

Using equation (2.36), we have

$$(\widehat{\mathbf{B}}^T \otimes \mathbf{I}_p)\mathbf{K}_{(k,p)} = \mathbf{K}_{(n,p)}(\mathbf{I}_p \otimes \widehat{\mathbf{B}}^T) = \mathbf{K}_{(n,p)} \begin{bmatrix} \mathbf{I}_{p-k} \otimes \widehat{\mathbf{B}}^T & \mathbf{0}^{n.(p-k)\times k.k} \\ \mathbf{0}^{k.n\times n.(p-k)} & \mathbf{I}_k \otimes \widehat{\mathbf{B}}^T \end{bmatrix} ,$$

hence

$$\mathbf{M}(\mathbf{a}) = \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})} diag(vec(\sqrt{\mathbf{W}}))\mathbf{K}_{(n,p)} \begin{bmatrix} \mathbf{I}_{p-k} \otimes \widehat{\mathbf{B}}^T & \mathbf{0}^{n.(p-k)\times k.k} \\ \mathbf{0}^{k.n\times n.(p-k)} & \mathbf{I}_k \otimes \widehat{\mathbf{B}}^T \end{bmatrix} ,$$

and so

$$\mathbf{M}(\mathbf{a})_1 = \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})} diag(vec(\sqrt{\mathbf{W}}))\mathbf{K}_{(n,p)} \begin{bmatrix} \mathbf{I}_{p-k} \otimes \widehat{\mathbf{B}}^T \\ \mathbf{0}^{k.n\times n.(p-k)} \end{bmatrix} ,$$

$$\mathbf{M}(\mathbf{a})_2 = \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})} diag(vec(\sqrt{\mathbf{W}}))\mathbf{K}_{(n,p)} \begin{bmatrix} \mathbf{0}^{n.(p-k)\times k.k} \\ \mathbf{I}_k \otimes \widehat{\mathbf{B}}^T \end{bmatrix} .$$

Now, we want to show that, $\forall j \in \{1, 2, \cdots, k.k\}$, there is $\mathbf{z}_j \in \mathbb{R}^{k.(p-k)}$ such that $\mathbf{M}(\mathbf{a})_2\mathbf{e}_j = \mathbf{M}(\mathbf{a})_1\mathbf{z}_j$, where $\mathbf{e}_j$ is the j^{th} column unit vector of order $k.k$ and $\mathbf{z}_j$ is the j^{th} column of the matrix $\mathbf{Z}$ we are looking for. Using the preceding expressions of the $\mathbf{M}(\mathbf{a})_1$ and $\mathbf{M}(\mathbf{a})_2$ submatrices, $\forall j \in \{1, 2, \cdots, k.k\}$, we have the equivalences

$$\begin{aligned}
&\mathbf{M}(\mathbf{a})_2\mathbf{e}_j = \mathbf{M}(\mathbf{a})_1\mathbf{z}_j \\
&\Leftrightarrow \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})} diag(vec(\sqrt{\mathbf{W}}))\mathbf{K}_{(n,p)} \left(\begin{bmatrix} \mathbf{0}^{n.(p-k)\times k.k} \\ \mathbf{I}_k \otimes \widehat{\mathbf{B}}^T \end{bmatrix} \mathbf{e}_j - \begin{bmatrix} \mathbf{I}_{p-k} \otimes \widehat{\mathbf{B}}^T \\ \mathbf{0}^{k.n\times n.(p-k)} \end{bmatrix} \mathbf{z}_j \right) = \mathbf{0}^{p.n} \\
&\Leftrightarrow \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})} diag(vec(\sqrt{\mathbf{W}}))\mathbf{K}_{(n,p)} \begin{bmatrix} -(\mathbf{I}_{p-k} \otimes \widehat{\mathbf{B}}^T)\mathbf{z}_j \\ (\mathbf{I}_k \otimes \widehat{\mathbf{B}}^T)\mathbf{e}_j \end{bmatrix} = \mathbf{0}^{p.n} \\
&\Leftrightarrow diag(vec(\sqrt{\mathbf{W}}))\mathbf{K}_{(n,p)} \begin{bmatrix} -(\mathbf{I}_{p-k} \otimes \widehat{\mathbf{B}}^T)\mathbf{z}_j \\ (\mathbf{I}_k \otimes \widehat{\mathbf{B}}^T)\mathbf{e}_j \end{bmatrix} \in ran(\mathbf{F}(\mathbf{a})) \\
&\Leftrightarrow \exists \mathbf{t}_j \in \mathbb{R}^{n.k} \ / \ diag(vec(\sqrt{\mathbf{W}}))\mathbf{K}_{(n,p)} \begin{bmatrix} -(\mathbf{I}_{p-k} \otimes \widehat{\mathbf{B}}^T)\mathbf{z}_j \\ (\mathbf{I}_k \otimes \widehat{\mathbf{B}}^T)\mathbf{e}_j \end{bmatrix} = \mathbf{F}(\mathbf{a})\mathbf{t}_j \ .
\end{aligned}$$

For all $j \in \{1, 2, \cdots, k.k\}$, we will look for a vector $\mathbf{t}_j$ such that $\mathbf{t}_j = vec(\mathbf{C}_j\widehat{\mathbf{B}})$ where $\mathbf{C}_j \in \mathbb{R}^{k\times k}$. But, using equations (2.33), (2.36) and (2.34), $\forall \mathbf{C} \in \mathbb{R}^{k\times k}$, we have

$$\begin{aligned}
\mathbf{F}(\mathbf{a})vec(\mathbf{C}\widehat{\mathbf{B}}) &= diag(vec(\sqrt{\mathbf{W}}))(\mathbf{I}_n \otimes \mathbf{A})vec(\mathbf{C}\widehat{\mathbf{B}}) \\
&= diag(vec(\sqrt{\mathbf{W}}))vec(\mathbf{AC}\widehat{\mathbf{B}}) \\
&= diag(vec(\sqrt{\mathbf{W}}))(\widehat{\mathbf{B}}^T \otimes \mathbf{I}_p)vec(\mathbf{AC}) \\
&= diag(vec(\sqrt{\mathbf{W}}))\mathbf{K}_{(n,p)}(\mathbf{I}_p \otimes \widehat{\mathbf{B}}^T)\mathbf{K}_{(p,k)}vec(\mathbf{AC}) \\
&= diag(vec(\sqrt{\mathbf{W}}))\mathbf{K}_{(n,p)}(\mathbf{I}_p \otimes \widehat{\mathbf{B}}^T)vec(\mathbf{C}^T\mathbf{A}^T) \\
&= diag(vec(\sqrt{\mathbf{W}}))\mathbf{K}_{(n,p)} \begin{bmatrix} \mathbf{I}_{p-k} \otimes \widehat{\mathbf{B}}^T & \mathbf{0}^{n(p-k)\times k.k} \\ \mathbf{0}^{k.n\times n(p-k)} & \mathbf{I}_k \otimes \widehat{\mathbf{B}}^T \end{bmatrix} vec(\mathbf{C}^T\mathbf{A}^T) \\
&= diag(vec(\sqrt{\mathbf{W}}))\mathbf{K}_{(n,p)} \begin{bmatrix} (\mathbf{I}_{p-k} \otimes \widehat{\mathbf{B}}^T)vec(\mathbf{C}^T\mathbf{A}_1^T) \\ (\mathbf{I}_k \otimes \widehat{\mathbf{B}}^T)vec(\mathbf{C}^T\mathbf{A}_2^T) \end{bmatrix} ,
\end{aligned}$$

and it is therefore sufficient to demonstrate that, $\forall j \in \{1, 2, \cdots, k.k\}$, there are $\mathbf{z}_j \in \mathbb{R}^{k.(p-k)}$ and $\mathbf{C}_j \in \mathbb{R}^{k \times k}$ such that

$$\begin{bmatrix} -(\mathbf{I}_{p-k} \otimes \widehat{\mathbf{B}}^T)\mathbf{z}_j \\ (\mathbf{I}_k \otimes \widehat{\mathbf{B}}^T)\mathbf{e}_j \end{bmatrix} = \begin{bmatrix} (\mathbf{I}_{p-k} \otimes \widehat{\mathbf{B}}^T)vec(\mathbf{C}_j^T \mathbf{A}_1^T) \\ (\mathbf{I}_k \otimes \widehat{\mathbf{B}}^T)vec(\mathbf{C}_j^T \mathbf{A}_2^T) \end{bmatrix}$$

in order to have $\mathbf{M}(\mathbf{a})_2\mathbf{e}_j = \mathbf{M}(\mathbf{a})_1\mathbf{z}_j$. Furthermore, if $\mathbf{z}_j$ is set to $-vec(\mathbf{C}_j^T \mathbf{A}_1^T)$ in the above equation, then it suffices to show that there is $\mathbf{C}_j \in \mathbb{R}^{k \times k}$ such that

$$\mathbf{e}_j = vec(\mathbf{C}_j^T \mathbf{A}_2^T)$$

in order to obtain the desired result. But, $\forall j \in \{1, 2, \cdots, k.k\}$, it is easily verified that there are two integers $t(j)$ and $u(j)$ such that

$$\mathbf{e}_j = vec(\mathbf{i}_{t(j)} \mathbf{i}_{u(j)}^T) \, ,$$

where $\mathbf{i}_t$ is the t^{th} column unit vector of order k, and

$$\begin{aligned} \mathbf{e}_j = vec(\mathbf{C}_j^T \mathbf{A}_2^T) &\Leftrightarrow \mathbf{i}_{u(j)} \mathbf{i}_{t(j)}^T = \mathbf{A}_2 \mathbf{C}_j \\ &\Leftrightarrow \mathbf{C}_j = \mathbf{A}_2^{-1} \mathbf{i}_{u(j)} \mathbf{i}_{t(j)}^T \, , \end{aligned}$$

since $\mathbf{A}_2$ is nonsingular by hypothesis. Consequently, $\forall j \in \{1, 2, \cdots, k.k\}$, we have $\mathbf{M}(\mathbf{a})_2\mathbf{e}_j = \mathbf{M}(\mathbf{a})_1\mathbf{z}_j$ with

$$\mathbf{z}_j = -vec(\mathbf{C}_j^T \mathbf{A}_1^T) = -vec(\mathbf{i}_{t(j)} \mathbf{i}_{u(j)}^T (\mathbf{A}_2^{-1})^T \mathbf{A}_1^T) \, ,$$

which is the desired result.

We now demonstrate that, $\forall j \in \{1, 2, \cdots, k.k\}$, we also have $\mathbf{L}(\mathbf{a})_2\mathbf{e}_j = \mathbf{L}(\mathbf{a})_1\mathbf{z}_j$ with $\mathbf{z}_j$ defined as above. To this end, we first observe that if $\mathbf{W}$ and $P_\Omega(\mathbf{X} - \mathbf{A}\widehat{\mathbf{B}})$ are partitioned as

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \end{bmatrix} \text{ and } P_\Omega(\mathbf{X} - \mathbf{A}\widehat{\mathbf{B}}) = \begin{bmatrix} \mathbf{P}_1 \\ \mathbf{P}_2 \end{bmatrix}$$

, with

$$\mathbf{W}_1, \mathbf{P}_1 \in \mathbb{R}^{(p-k) \times n} \quad \text{and} \quad \mathbf{W}_2, \mathbf{P}_2 \in \mathbb{R}^{k \times n} \, ,$$

then, using equation (5.19) in this subsection and equation (2.31) in Subsection 2.2,

$$\begin{aligned} \mathbf{L}(\mathbf{a}) &= (\mathbf{F}(\mathbf{a})^+)^T \big((\mathbf{W} \odot P_\Omega(\mathbf{X} - \mathbf{A}\widehat{\mathbf{B}}))^T \otimes \mathbf{I}_k \big) \\ &= (\mathbf{F}(\mathbf{a})^+)^T \big[(\mathbf{W}_1 \odot \mathbf{P}_1)^T \otimes \mathbf{I}_k, (\mathbf{W}_2 \odot \mathbf{P}_2)^T \otimes \mathbf{I}_k \big] \, ; \end{aligned}$$

hence $\mathbf{L}(\mathbf{a})_1$ and $\mathbf{L}(\mathbf{a})_2$ are given by

$$\mathbf{L}(\mathbf{a})_1 = (\mathbf{F}(\mathbf{a})^+)^T \big((\mathbf{W}_1 \odot \mathbf{P}_1)^T \otimes \mathbf{I}_k \big) \text{ and } \mathbf{L}(\mathbf{a})_2 = (\mathbf{F}(\mathbf{a})^+)^T \big((\mathbf{W}_2 \odot \mathbf{P}_2)^T \otimes \mathbf{I}_k \big) \, ,$$

$\forall j \in \{1, 2, \cdots, k.k\}$, we then have the implication

$$\big((\mathbf{W}_2 \odot \mathbf{P}_2)^T \otimes \mathbf{I}_k \big) \mathbf{e}_j = \big((\mathbf{W}_1 \odot \mathbf{P}_1)^T \otimes \mathbf{I}_k \big) \mathbf{z}_j \Rightarrow \mathbf{L}(\mathbf{a})_2 \mathbf{e}_j = \mathbf{L}(\mathbf{a})_1 \mathbf{z}_j \, ,$$

and it suffices to show that

$$\big((\mathbf{W}_2 \odot \mathbf{P}_2)^T \otimes \mathbf{I}_k \big) \mathbf{e}_j - \big((\mathbf{W}_1 \odot \mathbf{P}_1)^T \otimes \mathbf{I}_k \big) \mathbf{z}_j = \mathbf{0}^{p.k}$$

in order to obtain $\mathbf{L}(\mathbf{a})_2\mathbf{e}_j = \mathbf{L}(\mathbf{a})_1\mathbf{z}_j$. Defining, as above,

$$\mathbf{C}_j = \mathbf{A}_2^{-1} \mathbf{i}_{u(j)} \mathbf{i}_{t(j)}^T$$

and remembering that

$$\mathbf{z}_j = -vec(\mathbf{C}_j^T \mathbf{A}_1^T) \text{ and } \mathbf{e}_j = vec(\mathbf{C}_j^T \mathbf{A}_2^T) \, ,$$

we have (using equation (2.33) in Subsection 2.2)

$$\begin{aligned}\big((\mathbf{W}_1 \odot \mathbf{P}_1)^T \otimes \mathbf{I}_k\big)\mathbf{z}_j &= -\big((\mathbf{W}_1 \odot \mathbf{P}_1)^T \otimes \mathbf{I}_k\big)vec(\mathbf{C}_j^T\mathbf{A}_1^T) \\ &= -vec\big(\mathbf{C}_j^T\mathbf{A}_1^T(\mathbf{W}_1 \odot \mathbf{P}_1)\big) \\ &= -\Big(\big((\mathbf{W}_1 \odot \mathbf{P}_1)^T\mathbf{A}_1\big) \otimes \mathbf{I}_k\Big)vec(\mathbf{C}_j^T)\ ,\end{aligned}$$

and also

$$\begin{aligned}\big((\mathbf{W}_2 \odot \mathbf{P}_2)^T \otimes \mathbf{I}_k\big)\mathbf{e}_j &= \big((\mathbf{W}_2 \odot \mathbf{P}_2)^T \otimes \mathbf{I}_k\big)vec(\mathbf{C}_j^T\mathbf{A}_2^T) \\ &= vec\big(\mathbf{C}_j^T\mathbf{A}_2^T(\mathbf{W}_2 \odot \mathbf{P}_2)\big) \\ &= \Big(\big((\mathbf{W}_2 \odot \mathbf{P}_2)^T\mathbf{A}_2\big) \otimes \mathbf{I}_k\Big)vec(\mathbf{C}_j^T)\ .\end{aligned}$$

From these equalities, using equations (2.31), (2.32) and (2.33) in Subsection 2.2, we deduce that

$$\begin{aligned}&\big((\mathbf{W}_2 \odot \mathbf{P}_2)^T \otimes \mathbf{I}_k\big)\mathbf{e}_j - \big((\mathbf{W}_1 \odot \mathbf{P}_1)^T \otimes \mathbf{I}_k\big)\mathbf{z}_j \\ &= \Big(\big((\mathbf{W}_2 \odot \mathbf{P}_2)^T\mathbf{A}_2\big) \otimes \mathbf{I}_k + \big((\mathbf{W}_1 \odot \mathbf{P}_1)^T\mathbf{A}_1\big) \otimes \mathbf{I}_k\Big)vec(\mathbf{C}_j^T) \\ &= \Big(\big((\mathbf{W}_2 \odot \mathbf{P}_2)^T\mathbf{A}_2 + (\mathbf{W}_1 \odot \mathbf{P}_1)^T\mathbf{A}_1\big) \otimes \mathbf{I}_k\Big)vec(\mathbf{C}_j^T) \\ &= \Big(\big((\mathbf{W} \odot P_\Omega(\mathbf{X} - \mathbf{A}\widehat{\mathbf{B}}))^T\mathbf{A}\big) \otimes \mathbf{I}_k\Big)vec(\mathbf{C}_j^T) \\ &= vec\big(\mathbf{C}_j^T\mathbf{A}^T(\mathbf{W} \odot P_\Omega(\mathbf{X} - \mathbf{A}\widehat{\mathbf{B}}))\big) \\ &= (\mathbf{I}_n \otimes \mathbf{C}_j^T)(\mathbf{I}_n \otimes \mathbf{A}^T)vec(\mathbf{W} \odot P_\Omega(\mathbf{X} - \mathbf{A}\widehat{\mathbf{B}})) \\ &= (\mathbf{I}_n \otimes \mathbf{C}_j^T)(\mathbf{I}_n \otimes \mathbf{A}^T)diag(vec(\sqrt{\mathbf{W}}))vec(\sqrt{\mathbf{W}} \odot P_\Omega(\mathbf{X} - \mathbf{A}\widehat{\mathbf{B}})) \\ &= (\mathbf{I}_n \otimes \mathbf{C}_j^T)\mathbf{F}(\mathbf{a})^T\mathbf{r}(\mathbf{a}) \\ &= \mathbf{0}^{p.k}\ ,\end{aligned}$$

since $\mathbf{F}(\mathbf{a})^T\mathbf{r}(\mathbf{a}) = \mathbf{0}^{n.k}$. This implies that, $\forall j \in \{1, 2, \cdots, k.k\}$, we also have $\mathbf{L}(\mathbf{a})_2\mathbf{e}_j = \mathbf{L}(\mathbf{a})_1\mathbf{z}_j$ as claimed in the theorem.

Finally, since $J(\mathbf{r}(\mathbf{a})) = -(\mathbf{M}(\mathbf{a})+\mathbf{L}(\mathbf{a}))$, the preceding results also imply that, $\forall j \in \{1, 2, \cdots, k.k\}$, we have

$$J\big(\mathbf{r}(\mathbf{a})\big)_2\mathbf{e}_j = J\big(\mathbf{r}(\mathbf{a})\big)_1\mathbf{z}_j\ .$$

□

Theorems 5.4 shows that it is possible to compute $d\mathbf{a}_{gn}$ efficiently, and without using the linear constraint $\mathbf{N}^T d\mathbf{a} = \mathbf{0}^{k.k}$ or an orthonormal basis $\mathbf{O}^\perp$ of $ran(\mathbf{A})^\perp$, by using a simplified stable COD of $J(\mathbf{r}(\mathbf{a}))$ (see equation (2.20) in Subsection 2.1 for details). When $rank(J(\mathbf{r}(\mathbf{a}))) = k.(p-k)$, the first $k.(p-k)$ columns of $J(\mathbf{r}(\mathbf{a}))$ are linearly independent and the last $k.k$ columns of this matrix are linearly dependent upon these first $k.(p-k)$ columns as soon as $rank(\mathbf{A}_2) = k$. This property is also verified for the matrix $-\mathbf{M}(\mathbf{a})$, which can be used as an approximation of $J(\mathbf{r}(\mathbf{a}))$ in the optimization algorithms as we will discuss in Section 6. In these conditions, a simplified COD of $J(\mathbf{r}(\mathbf{a}))$ (or alternatively of $-\mathbf{M}(\mathbf{a})$) is given by

$$J(\mathbf{r}(\mathbf{a})) = \mathbf{Q}\mathbf{T}\mathbf{Z}^T = \mathbf{Q}\begin{bmatrix}\mathbf{T}_{11} & \mathbf{0}^{k(p-k)\times k.k}\end{bmatrix}\mathbf{Z}^T\ ,$$

where $\mathbf{Q} \in \mathbb{O}^{n.p\times k.(p-k)}$, $\mathbf{Z} \in \mathbb{O}^{k.p\times k.p}$, $\mathbf{T} \in \mathbb{R}^{k.(p-k)\times k.p}$ and $\mathbf{T}_{11} \in \mathbb{R}^{k.(p-k)\times k.(p-k)}$ is a full rank upper triangular matrix. The advantage of this COD formulation for rank deficient matrices, such as $J(\mathbf{r}(\mathbf{a}))$ or $-\mathbf{M}(\mathbf{a})$, is the ability to compute directly the unique minimum 2-norm solution of the problem

$$\min_{d\mathbf{a}\in\mathbb{R}^{p.k}} \frac{1}{2}\|\mathbf{r}(\mathbf{a}) + J(\mathbf{r}(\mathbf{a}))d\mathbf{a}\|_2^2$$

as follows

$$d\mathbf{a}_{gn} = -\mathbf{Z}\begin{bmatrix}\mathbf{T}_{11}^{-1}\mathbf{Q}^T\mathbf{r}(\mathbf{a})\\ \mathbf{0}^{k.k\times k(p-k)}\end{bmatrix} ,$$

without computing the SVD or the pseudo-inverse of $J(\mathbf{r}(\mathbf{a}))$, but only this simplified COD of $J(\mathbf{r}(\mathbf{a}))$, which is based on a simple QR factorization without column pivoting of the first $k.(p-k)$ columns of $J(\mathbf{r}(\mathbf{a}))$ as a first step. Furthermore, as $J(\mathbf{r}(\mathbf{a}))$ is a tall and skinny matrix, this QR factorization can be parallelized very efficiently to reduce the computational time and the memory requirements as we will illustrate in Section 6, see also [48] for details. We also note that this COD can also be used to derive another variation of the Levenberg-Marquardt algorithm, which will also take care of the singularity of the Jacobian matrix or its approximation without using the linear constraint $\mathbf{N}^T d\mathbf{a} = \mathbf{0}^{k.k}$ or an orthonormal basis, $\mathbf{O}^{\perp}$, of $ran(\mathbf{A})^{\perp}$ as in the formulations derived above; see again Section 6 for details.

Theorems 5.2 and 5.4 are valid for an arbitrary weight matrix $\mathbf{W}$. However, the ranks of $\mathbf{M}(\mathbf{a})$, $\mathbf{L}(\mathbf{a})$ and $J(\mathbf{r}(\mathbf{a}))$ may also be altered by the choice of a particular weight matrix, as already illustrated by Theorem 5.3. The two following theorems further illustrate the strong dependency of the ranks of $\mathbf{M}(\mathbf{a})$, $\mathbf{L}(\mathbf{a})$ and $J(\mathbf{r}(\mathbf{a}))$ to the number and distribution of zero elements in $\mathbf{W}$, respectively.

Let us again give some definitions before stating the two theorems. First, we introduce the $p \times n$ incidence matrix, $\boldsymbol{\delta}$, associated with the matrix $\mathbf{W}$, defined by

$$\boldsymbol{\delta}_{ij} = \begin{cases} 1 & \text{if } \mathbf{W}_{ij} \neq 0 \\ 0 & \text{if } \mathbf{W}_{ij} = 0 \end{cases} . \tag{5.26}$$

The number of "nonmissing" elements in $\mathbf{X}$ (or equivalently the number of non-zero elements in $\mathbf{W}$) is then equal to $\sum_{ij} \boldsymbol{\delta}_{ij} = nobs$. Since

$$\mathbf{F}(\mathbf{a}) = \bigoplus_{j=1}^{n} \Big(diag(vec(\sqrt{\mathbf{W}}_{.j}))\mathbf{A} \Big) = \bigoplus_{j=1}^{n} \mathbf{F}_j(\mathbf{a}) ,$$

we first observe that

$$rank\big(\mathbf{F}(\mathbf{a})\big) = \sum_{j=1}^{n} r_j \quad \text{where } r_j = rank(\mathbf{F}_j(\mathbf{a})), \text{ for } j = 1, \cdots, n .$$

Furthermore, using equation (2.2) in Subsection 2.1, we have

$$r_j \leqslant min\left(\sum_{i=1}^{p} \boldsymbol{\delta}_{ij}, k \right), \text{ for } j = 1, \cdots, n, \quad \text{and} \quad rank\big(\mathbf{F}(\mathbf{a})\big) \leqslant min\left(nobs, k.n\right) .$$

We now state the following simple bounds for the ranks of $\mathbf{M}(\mathbf{a})$, $\mathbf{L}(\mathbf{a})$ and $J(\mathbf{r}(\mathbf{a}))$:

Theorem 5.5. With these definitions and the same notations as in Theorem 5.2, then the following inequalities hold:

$$\begin{aligned} rank(\mathbf{M}(\mathbf{a})) &\leqslant min\left(nobs - rank\big(\mathbf{F}(\mathbf{a})\big), k.(p - k_{\mathbf{A}})\right) \\ rank(\mathbf{L}(\mathbf{a})) &\leqslant min\left(nobs, k.min(n, p - k_{\mathbf{A}})\right) \\ rank(J(\mathbf{r}(\mathbf{a}))) &\leqslant min\left(nobs, k.(p - k_{\mathbf{A}})\right) . \end{aligned}$$

Proof. Omitted.

□

Let us further define for any integer $i \in \{1, 2, \cdots, p\}$ and any $p \times k$ matrix $\mathbf{A}$, the finite subset of $\mathbb{N}$

$$\xi(i, \mathbf{A}) = \left\{ j \in \{1, 2, \cdots, n\} \ / \ \boldsymbol{\delta}_{ij} \neq 0 \text{ and } \sum_{l=1}^{p} \boldsymbol{\delta}_{lj} > r_j \right\},$$

where $r_j = \textit{rank}(\mathbf{F}_j(\mathbf{a}))$ for $j = 1, \cdots, n$ as above. If now, we define the finite subset of $\mathbb{N}$

$$\Omega(\mathbf{A}) = \left\{ i \in \{1, 2, \cdots, p\} \ / \ \xi(i, \mathbf{A}) = \varnothing \right\},$$

the following theorem is valid.

Theorem 5.6. With these definitions, let $\textit{card}\big(\Omega(\mathbf{A})\big)$ be the number of elements of $\Omega(\mathbf{A})$. If $\Omega(\mathbf{A})$ is not empty then the matrices $\mathbf{M}(\mathbf{a})$, $\mathbf{L}(\mathbf{a})$ and $J(\mathbf{r}(\mathbf{a}))$ have $k.\textit{card}\big(\Omega(\mathbf{A})\big)$ columns equal to zero, moreover the indices of these columns in these matrices correspond.

Proof. We first consider the matrix $\mathbf{M}(\mathbf{a})$ and recall that this matrix has the following form

$$\mathbf{M}(\mathbf{a}) = \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})} \textit{diag}(\textit{vec}(\sqrt{\mathbf{W}}))(\widehat{\mathbf{B}}^T \otimes \mathbf{I}_p)\mathbf{K}_{(k,p)} \ .$$

Then, we also recall that $\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}$ is a block-diagonal matrix of the form

$$\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})} = \begin{bmatrix} \mathbf{P}^{\perp}_{\mathbf{F}_1(\mathbf{a})} & 0 & \dots & 0 & 0 \\ 0 & \ddots & 0 & \dots & 0 \\ \vdots & 0 & \mathbf{P}^{\perp}_{\mathbf{F}_j(\mathbf{a})} & 0 & \vdots \\ 0 & \dots & 0 & \ddots & 0 \\ 0 & 0 & \dots & 0 & \mathbf{P}^{\perp}_{\mathbf{F}_n(\mathbf{a})} \end{bmatrix} = \bigoplus_{j=1}^{n} \mathbf{P}^{\perp}_{\mathbf{F}_j(\mathbf{a})} \ ,$$

with

$$\mathbf{P}^{\perp}_{\mathbf{F}_j(\mathbf{a})} = \mathbf{I}_p - \mathbf{F}_j(\mathbf{a})\mathbf{F}_j(\mathbf{a})^{+} = \mathbf{I}_p - \big(\textit{diag}(\textit{vec}(\sqrt{\mathbf{W}}_{.j}))\mathbf{A}\big)\big(\textit{diag}(\textit{vec}(\sqrt{\mathbf{W}}_{.j}))\mathbf{A}\big)^{+} \ ,$$

for $j \in \{1, 2, \cdots, n\}$. Hence,

$$\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})} \textit{diag}(\textit{vec}(\sqrt{\mathbf{W}})) = \bigoplus_{j=1}^{n} \Big(\mathbf{P}^{\perp}_{\mathbf{F}_j(\mathbf{a})} \textit{diag}(\textit{vec}(\sqrt{\mathbf{W}}_{.j})) \Big) \ .$$

Consider now the following uniform blocking of the $np \times pk$ matrix $(\widehat{\mathbf{B}}^T \otimes \mathbf{I}_p)\mathbf{K}_{(k,p)}$ into $n.p$ submatrices $\mathbf{O}_{ji} \in \mathbb{R}^{p \times k}$, for $j \in \{1, 2, \cdots, n\}$ and $i \in \{1, 2, \cdots, p\}$,

$$(\widehat{\mathbf{B}}^T \otimes \mathbf{I}_p)\mathbf{K}_{(k,p)} = \begin{bmatrix} \mathbf{O}_{11} & \mathbf{O}_{12} & \dots & \mathbf{O}_{1p} \\ \mathbf{O}_{21} & \mathbf{O}_{22} & \dots & \mathbf{O}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{O}_{n1} & \mathbf{O}_{n2} & \dots & \mathbf{O}_{np} \end{bmatrix}, \mathbf{O}_{ji} \in \mathbb{R}^{p \times k} \ .$$

Since $(\widehat{\mathbf{B}}^T \otimes \mathbf{I}_p)\mathbf{K}_{(k,p)} = \mathbf{K}_{(n,p)}(\mathbf{I}_p \otimes \widehat{\mathbf{B}}^T)$, $(\widehat{\mathbf{B}}^T \otimes \mathbf{I}_p)\mathbf{K}_{(k,p)}$ is the matrix having as rows, every n^{th} row of the matrix $\mathbf{I}_p \otimes \widehat{\mathbf{B}}^T$ of order $n.p \times p.k$, starting with the first, then every n^{th} row starting with the second, and so on. Thus, keeping in mind that $\mathbf{I}_p \otimes \widehat{\mathbf{B}}^T$ is a block-diagonal matrix, it is not hard to see that the $\mathbf{O}_{ji}$ submatrices are very sparse with the following simple structure

$$\mathbf{O}_{ji} = \begin{bmatrix} \mathbf{0}^{(i-1)\times k} \\ [\widehat{\mathbf{B}}^T]_{j.} \\ \mathbf{0}^{(p-i)\times k} \end{bmatrix} \begin{matrix} \}(i-1) \text{ rows} \\ \}1 \text{ row} \\ \}(p-i) \text{ rows} \end{matrix} \ , \text{ for } j = 1, \dots, n \text{ and } i = 1, \dots, p \ \ .$$

Now, let $i \in \Omega(\mathbf{A})$ and consider the submatrix $\mathbf{M}_i$ defined by the columns $(i-1).k+1$ to $i.k$ of $\mathbf{M}(\mathbf{a})$. $\mathbf{M}_i$ is equal to

$$\mathbf{M}_i = \bigoplus_{j=1}^{n} \left(\mathbf{P}^{\perp}_{\mathbf{F}_j(\mathbf{a})} diag(\sqrt{\mathbf{W}}_{.j}) \right) \begin{bmatrix} \mathbf{O}_{1i} \\ \vdots \\ \mathbf{O}_{ni} \end{bmatrix} = \begin{bmatrix} \mathbf{P}^{\perp}_{\mathbf{F}_1(\mathbf{a})} diag(\sqrt{\mathbf{W}}_{.1}) \mathbf{O}_{1i} \\ \vdots \\ \mathbf{P}^{\perp}_{\mathbf{F}_n(\mathbf{a})} diag(\sqrt{\mathbf{W}}_{.n}) \mathbf{O}_{ni} \end{bmatrix}.$$

Now, for $j = 1, \ldots, n$, we have $\mathbf{W}_{ij} = 0$ or $\mathbf{W}_{ij} \neq 0$:

- If $\mathbf{W}_{ij} = 0$, then $diag(\sqrt{\mathbf{W}}_{.j})\mathbf{O}_{ji} = 0^{p \times k}$ and the rows $(j-1).p+1$ to $j.p$ of $\mathbf{M}_i$ are equal to zero.

- If $\mathbf{W}_{ij} \neq 0$, we have $\boldsymbol{\delta}_{ij} = 1$ and $\sum_{l=1}^{p} \boldsymbol{\delta}_{lj} = r_j$ since $i \in \Omega(\mathbf{A})$, and it follows that

$$\mathbf{P}^{\perp}_{\mathbf{F}_j(\mathbf{a})} = \mathbf{I}_p - diag(\boldsymbol{\delta}_{.j})$$

and

$$\begin{aligned} \mathbf{P}^{\perp}_{\mathbf{F}_j(\mathbf{a})} diag(\sqrt{\mathbf{W}}_{.j}) &= \big(\mathbf{I}_p - diag(\boldsymbol{\delta}_{.j})\big) diag(\sqrt{\mathbf{W}}_{.j}) \\ &= diag(\sqrt{\mathbf{W}}_{.j}) - diag(\sqrt{\mathbf{W}}_{.j}) \\ &= \mathbf{0}^p, \end{aligned}$$

and the rows $(j-1).p+1$ to $j.p$ of $\mathbf{M}_i$ are also equal to zero if $\mathbf{W}_{ij} \neq 0$.

Hence, we finally obtain $\mathbf{M}_i = 0^{p.n \times k}$ if $i \in \Omega(\mathbf{A})$ and we conclude that $\mathbf{M}(\mathbf{a})$ has $k.card\big(\Omega(\mathbf{A})\big)$ columns equal to zero, as claimed in the theorem.

Turning now our attention to $\mathbf{L}(\mathbf{a})$, which is equal to

$$\mathbf{L}(\mathbf{a}) = (\mathbf{F}(\mathbf{a})^{+})^T \big((\mathbf{W} \odot P_{\Omega}(\mathbf{X} - \mathbf{A}\widehat{\mathbf{B}}))^T \otimes \mathbf{I}_k \big),$$

we observe that it is sufficient to show that the i^{th} row of $\mathbf{W} \odot P_{\Omega}(\mathbf{X} - \mathbf{A}\widehat{\mathbf{B}})$ is equal to zero to establish that the columns $(i-1).k+1$ to $i.k$ of $\mathbf{L}(\mathbf{a})$ are equal to zero if $i \in \Omega(\mathbf{A})$.

Now, for $j = 1, \ldots, n$, we have $\mathbf{W}_{ij} = 0$ or $\mathbf{W}_{ij} \neq 0$:

- If $\mathbf{W}_{ij} = 0$, then $\Big[\mathbf{W} \odot P_{\Omega}(\mathbf{X} - \mathbf{A}\widehat{\mathbf{B}})\Big]_{ij} = 0$ for any $p \times n$ matrix $P_{\Omega}(\mathbf{X} - \mathbf{A}\widehat{\mathbf{B}})$.

- If $\mathbf{W}_{ij} \neq 0$, as above, we have $\boldsymbol{\delta}_{ij} = 1$ and $\sum_{l=1}^{p} \boldsymbol{\delta}_{lj} = r_j$ since $i \in \Omega(\mathbf{A})$, implying that

$$\mathbf{P}^{\perp}_{\mathbf{F}_j(\mathbf{a})} = \mathbf{I}_p - diag(\boldsymbol{\delta}_{.j})$$

and it follows that

$$\begin{aligned} \Big[\mathbf{W} \odot P_{\Omega}(\mathbf{X} - \mathbf{A}\widehat{\mathbf{B}})\Big]_{.j} &= \sqrt{\mathbf{W}}_{.j} \odot \Big(\sqrt{\mathbf{W}}_{.j} \odot P_{\Omega}(\mathbf{X} - \mathbf{A}\widehat{\mathbf{B}})_{.j} \Big) \\ &= \sqrt{\mathbf{W}}_{.j} \odot \mathbf{P}^{\perp}_{\mathbf{F}_j(\mathbf{a})} (\sqrt{\mathbf{W}}_{.j} \odot \mathbf{X}_{.j}) \\ &= \sqrt{\mathbf{W}}_{.j} \odot \big(\mathbf{I}_p - diag(\boldsymbol{\delta}_{.j})\big)(\sqrt{\mathbf{W}}_{.j} \odot \mathbf{X}_{.j}) \\ &= \sqrt{\mathbf{W}}_{.j} \odot \big(\sqrt{\mathbf{W}}_{.j} \odot \mathbf{X}_{.j} - \sqrt{\mathbf{W}}_{.j} \odot \mathbf{X}_{.j}\big) \\ &= \mathbf{0}^p, \end{aligned}$$

and $\Big[\mathbf{W} \odot P_{\Omega}(\mathbf{X} - \mathbf{A}\widehat{\mathbf{B}})\Big]_{ij} = 0$ also in the case $\mathbf{W}_{ij} \neq 0$.

Hence, we conclude that the columns $(i-1).k+1$ to $i.k$ of $\mathbf{L}(\mathbf{a})$ are equal to zero if $i \in \Omega(\mathbf{A})$ and $\mathbf{L}(\mathbf{a})$ has also $k.card\big(\Omega(\mathbf{A})\big)$ columns equal to zero, as claimed in the theorem. Finally, the same result holds for $J(\mathbf{r}(\mathbf{a}))$ since $J(\mathbf{r}(\mathbf{a})) = -(\mathbf{M}(\mathbf{a}) + \mathbf{L}(\mathbf{a}))$.

□

Clearly, any variable projection Gauss-Newton or Levenberg-Marquardt algorithm using the Jacobian matrix $J(\mathbf{r}(\mathbf{a}))$, or approximating this Jacobian matrix with $-\mathbf{M}(\mathbf{a})$, will be incorrect if $\Omega(\mathbf{A})$ is not empty as demonstrated in Theorem 5.6. The fact that $\Omega(\mathbf{A})$ is not empty, is symptomatic of the situation where we try to fit the matrix $\mathbf{X}$ by a model with too many components with respect to the number of missing elements in this matrix. However, if we restrict the set of WLRA problems by imposing the condition $\sum_{l=1}^{p} \boldsymbol{\delta}_{lj} > k$ for all $j = 1, \cdots, n$, we are sure that $\Omega(\mathbf{A})$ will be empty and the events described in Theorem 5.6 will not occurred.

Finally, if $card\big(\Omega(\mathbf{A})\big)$ is not equal to zero, a much better alternative to find a solution of this particular WLRA problem, is to minimize the regularized cost function $g_\lambda(.)$ defined in equation (3.18) of Subsection 3.3 with a variable projection algorithm, as this regularized minimization problem is always well-posed (see Subsection 3.3 for details).

5.3 Computations and properties of the gradient vector and Hessian matrix

Using the properties of the Jacobian matrix $J(\mathbf{r}(\mathbf{a}))$ demonstrated in Subsection 5.2, we now derive a simple expression for the gradient of $\psi(.)$ and we give a detailed study of the Hessian matrix $\nabla^2\psi(\mathbf{a})$. These results may be used to formulate steepest descent or Newton algorithms for minimizing the variable projection functional $\psi(.)$ [171][28][169][46][13][14][17]. We also again illustrate the tight relationships between these Euclidean gradient and Hessian operators and the corresponding Riemannian gradient and Hessian operators when we consider the cost function $\psi(.)$ as operating on the Grassmann manifold $\mathrm{Gr}(p,k)$ [3][14][11] extending the investigations of [82] on this topic.

Since $J(\mathbf{r}(\mathbf{a}))$ is rank-deficient everywhere according to Theorem 5.2, the smallest eigenvalue of $J(\mathbf{r}(\mathbf{a}))^T J(\mathbf{r}(\mathbf{a}))$ is always equal to zero and, in these conditions, we cannot expect that the Gauss-Newton term $J(\mathbf{r}(\mathbf{a}))^T J(\mathbf{r}(\mathbf{a}))$ will dominate the second term in the expression of $\nabla^2\psi(\mathbf{a})$ (see Subsection 5.1). This suggests that a full-Newton approach can perform much better than the Gauss-Newton or Levenberg-Marquardt methods for solving the (VP1) problem. However, the Gauss-Newton or Levenberg-Marquardt algorithms perform surprisingly better than different versions of the full Newton algorithm for solving the (VP1) problem in the comparative studies of Okatani et al. [150] or Hong et al. [81].These conclusions are thus very counterintuitive taking into account the excellent properties (e.g., fast convergence for NLLS problems with large residuals and quadratic local convergence in a neighborhood of a stationary point) of the full Newton approach compared to the Gauss-Newton or Levenberg-Marquardt methods for general or variable projection NLLS problems [45][123][10][87].

The results of this Subsection will try to elucidate these contradictions and also will reveal the power of the variable projection framework in understanding the intrinsic difficulties associated with the WLRA problem. As an illustration, we will demonstrate below, that the Hessian matrix $\nabla^2\psi(\widehat{\mathbf{a}})$ is deficient at all first-order stationary points $\widehat{\mathbf{a}}$ of $\psi(.)$ (see Theorem 5.9 below) and, consequently, this Hessian matrix is expected to be nearly singular and ill-conditioned in a "small" neighborhood of a first-order stationary point $\widehat{\mathbf{a}}$ of $\psi(.)$. This result is consistent with the fact that (local) minimizers of $\varphi^*(.)$ and $\psi(.)$ are never isolated, as already discussed in Subsection 3.1, and that any neighborhood of a (local) minimizer of these cost functions contains also an infinite number of other minimizers, which attain the same minimum of $\psi(.)$, implying that the Hessian matrix $\nabla^2\psi(\widehat{\mathbf{a}})$ at a local minimizer $\widehat{\mathbf{a}}$ is at best positive semi-definite, but never positive definite. When the weight matrix $\mathbf{W} \in \mathbb{R}_{*+}^{p\times n}$ and we consider the cost function $\psi(.)$ as defined on the (quotient) Grassmann

manifold $\mathrm{Gr}(n,k)$, because $\psi(.)$ is invariant on the equivalence classes of this quotient, the fact that the Hessian cannot possibly be positive definite at first-order stationary points is already a known result, see Chapter 9 and Lemma 9.41 in [11]. Our Theorem 5.9 thus provides an extension of this result as it also applies to the case where $\mathbf{W} \in \mathbb{R}_{+}^{p \times n}$. These results also imply that the Hessian matrices at points arbitrarily closed to minima have vanishingly small, possibly negative eigenvalues, leading to ill-conditioned and indefinite linear systems with a severe loss of accuracy in Newton and trust-regions methods when the iterates reach a neighborhood of a minimum [162].

In these conditions, many of the excellent properties of full-Newton and trust-regions approaches are lost, which may explain the degraded performance of these methods for minimizing $\psi(.)$ in the comparative studies of Okatani et al. [150] and Hong et al. [81]. This poor performance concerns especially the second-order Riemannian trust-region method (RTRMC2) operating on the Grassmann manifold developed initially in Boumal and Absil [13] and already discussed in Subsection 3.3. In their Riemannian Newton approach, Boumal and Absil [13][14] have derived a compact directional derivative formulae for the Riemannian Hessian of the cost function of a regularized form of the WLRA problem (see the regularized cost function $g_{\lambda}(.)$ defined in equation (3.18) of Subsection 3.3), which is then used in an inexact subproblem solver based on a truncated conjugate gradient method to compute an approximate (Riemannian) Newton step at each iteration of their RTRMC2 method. However, in order to ensure convergence to (local) minima, the truncated conjugate gradient theory assumes a positive definite system at these minima of the cost function, an hypothesis which is not verified here near first-order critical points, where the RTRMC2 method may be applied to ill-conditioned, indefinite systems, leading to an erratic behaviour for some WLRA problems as the truncated conjugate gradient subproblem solver may be highly sensitive to negative eigenvalues, even of small magnitude. This challenging question of the convergence of trust-regions methods for non-isolated minima of smooth functions defined on a (arbitrary) manifold has been revisited recently in [162] in which the authors were still able to derive convergence results for an inexact solver based on a truncated conjugate gradient method under some additional hypotheses and with a carefully designed inexact subproblem solver with had hoc stopping criteria for the truncated conjugate gradient iterations so that the subproblem solver is not affected by the small negative eigenvalues of the Hessian matrix.

These convergence problems of trust-region methods near (local) minima may explain why Boumal and Absil [14] have subsequently introduced a pre-conditioner for the Hessian matrix in their RTRMC2 method originally proposed in [13]. However, we are not aware of any new comparison studies, which evaluate the performance of the updated and preconditioned RTRMC2 algorithm proposed in [14] with the Gauss-Newton or Levenberg-Marquardt approaches described in [147][150][81][88] to verify if this preconditioned version of RTRMC2 performs now better than the variable projection Gauss-Newton or Levenberg-Marquardt methods for solving WLRA problems. Furthermore, almost all these previous studies deal only of WLRA problems with binary weights (e.g., the missing value problem) excepted of the work of Boumal and Absil [14]. This clearly shows the need of new extensive comparison studies to clarify the respective performance of the most recent various first- or second-order methods proposed in the literature for solving the general WLRA problem, but such ambitious task is outside of the scope of this paper, which is mainly devoted to a better understanding of the theoretical properties of variable projections methods for solving the WLRA problem.

To start with, we give convenient different expressions for the Euclidean gradient of the variable projection functional $\psi(.)$. This proposition is mainly a reformulation of Theorem 2.1 in Golub and Pereyra [63]; however, we give a direct proof using only linear algebra in order to be self-contained and because this formula is not well-known in the literature related to the WLRA problem.

Theorem 5.7. Let $\varphi^{*}(\mathbf{A}, \widehat{\mathbf{B}})$ and $\psi(\mathbf{a})$ be defined, respectively, as in equations (P1) and (3.23) of Section 3 with $\mathbf{A} \in \mathbb{R}_{k}^{p \times k}$, $\mathbf{a} = vec(\mathbf{A}^{T}) \in \mathbb{R}^{p.k}$, $\widehat{\mathbf{B}} \in \mathbb{R}^{k \times n}$ is such that $\widehat{\mathbf{b}} = vec(\widehat{\mathbf{B}}) = \mathbf{F}(\mathbf{a})^{+}\mathbf{x} \in \mathbb{R}^{k.n}$ and $\mathbf{x} = vec(\sqrt{\mathbf{W}} \odot \mathbf{X}) \in \mathbb{R}^{p.n}$, where $\mathbf{X} \in \mathbb{R}^{p \times n}$ and $\mathbf{W} \in \mathbb{R}_{+}^{p \times n}$ are, respectively, the data

and weight matrices of the WLRA problems (P0) or (P1).

We further assume that $\mathbf{a}$ belongs to an open set $\Omega \subset \mathbb{R}^{p.k}$ in which $\mathbf{F}(.)$ has a constant rank, so that the Jacobian matrix $J(\mathbf{r}(\mathbf{a}))$ derived in Subsection 5.2 and the Euclidean gradient $\nabla\psi(\mathbf{a})$ are both well-defined. Then,

$$\nabla\psi(\mathbf{a}) = -\mathbf{M}(\mathbf{a})^T\mathbf{r}(\mathbf{a}) = -\mathbf{M}(\mathbf{a})^T\big(\mathbf{x} - \mathbf{F}(\mathbf{a})\widehat{\mathbf{b}}\big) = \frac{\partial\varphi^*(\mathbf{A},\widehat{\mathbf{B}})}{\partial\mathbf{a}}$$

and

$$\|\nabla\psi(\mathbf{a})\|_2 = \big\|\nabla\varphi^*_{\mathbf{a}}(\mathbf{A},\widehat{\mathbf{B}})\big\|_2 = \big\|\nabla\varphi^*_{\mathbf{A}}(\mathbf{A},\widehat{\mathbf{B}})\big\|_F \,,$$

where $\frac{\partial\varphi^*(\mathbf{A},\widehat{\mathbf{B}})}{\partial\mathbf{a}}$, $\nabla\varphi^*_{\mathbf{a}}(\mathbf{A},\widehat{\mathbf{B}})$ and $\nabla\varphi^*_{\mathbf{A}}(\mathbf{A},\widehat{\mathbf{B}})$ are defined, respectively, by equations (4.3) and (4.5) in Theorem 4.3 and equation (3.11) of Subsection 3.2.

In addition, the theorem remains valid if $\widehat{\mathbf{b}} = \mathbf{F}(\mathbf{a})^-\mathbf{x}$ where $\mathbf{F}(\mathbf{a})^-$ is a symmetric generalized inverse of $\mathbf{F}(\mathbf{a})$ as defined in equations (2.10) or (2.19) of Subsection 2.1.

Proof. Using the notations and results in Subsection 5.2, we have

$$\begin{aligned}
\nabla\psi(\mathbf{a}) &= J\big(\mathbf{r}(\mathbf{a})\big)^T\mathbf{r}(\mathbf{a}) \\
&= -\big(\mathbf{M}(\mathbf{a}) + \mathbf{L}(\mathbf{a})\big)^T\mathbf{r}(\mathbf{a}) \\
&= -\mathbf{M}(\mathbf{a})^T\mathbf{r}(\mathbf{a}) - \mathbf{L}(\mathbf{a})^T\mathbf{r}(\mathbf{a}) \\
&= -\mathbf{M}(\mathbf{a})^T\mathbf{r}(\mathbf{a}) \,,
\end{aligned}$$

since $\mathbf{r}(\mathbf{a}) \in \mathit{ran}(\mathbf{F}(\mathbf{a}))^\perp$ (see equation (3.24)), $\mathit{ran}(\mathbf{F}(\mathbf{a}))^\perp \subset \mathit{ran}(\mathbf{L}(\mathbf{a}))^\perp$ (see equation (5.23)) and $\mathit{ran}(\mathbf{L}(\mathbf{a}))^\perp = \mathit{null}(\mathbf{L}(\mathbf{a})^T)$. The last equality resulting from equation (2.4) in Subsection 2.1.

Hence, $\mathbf{L}(\mathbf{a})$, the second term of $J(\mathbf{r}(\mathbf{a}))$, does not contribute to the gradient $\nabla\psi(\mathbf{a})$. Now, using the second formulation of $\mathbf{M}(\mathbf{a})$ given in equation (5.16) of Subsection 5.2, we deduce

$$\begin{aligned}
\nabla\psi(\mathbf{a}) &= -\big(\mathbf{P}^\perp_{\mathbf{F}(\mathbf{a})}\mathbf{K}_{(n,p)}\mathbf{G}(\widehat{\mathbf{b}})\big)^T\mathbf{r}(\mathbf{a}) \\
&= -\mathbf{G}(\widehat{\mathbf{b}})^T\mathbf{K}_{(p,n)}\mathbf{P}^\perp_{\mathbf{F}(\mathbf{a})}\mathbf{r}(\mathbf{a}) \\
&= -\mathbf{G}(\widehat{\mathbf{b}})^T\mathbf{K}_{(p,n)}\mathbf{r}(\mathbf{a}) \,,
\end{aligned}$$

since $\mathbf{r}(\mathbf{a}) \in \mathit{ran}(\mathbf{F}(\mathbf{a}))^\perp = \mathit{ran}(\mathbf{P}^\perp_{\mathbf{F}(\mathbf{a})})$. Noting that

$$\mathbf{r}(\mathbf{a}) = \mathbf{x} - \mathbf{F}(\mathbf{a})\widehat{\mathbf{b}} = \mathbf{K}_{(n,p)}\big(\mathbf{z} - \mathbf{G}(\widehat{\mathbf{b}})\mathbf{a}\big) \,,$$

where $\mathbf{z} = \mathit{vec}\big((\sqrt{\mathbf{W}} \odot \mathbf{X})^T\big)$ (see equation (3.24) in Section 3 for details), we finally obtain

$$\begin{aligned}
\nabla\psi(\mathbf{a}) &= -\mathbf{G}(\widehat{\mathbf{b}})^T\mathbf{K}_{(p,n)}\mathbf{K}_{(n,p)}\big(\mathbf{z} - \mathbf{G}(\widehat{\mathbf{b}})\mathbf{a}\big) \\
&= -\mathbf{G}(\widehat{\mathbf{b}})^T\big(\mathbf{z} - \mathbf{G}(\widehat{\mathbf{b}})\mathbf{a}\big) \\
&= \mathbf{G}(\widehat{\mathbf{b}})^T\mathbf{G}(\widehat{\mathbf{b}})\mathbf{a} - \mathbf{G}(\widehat{\mathbf{b}})^T\mathbf{z} \\
&= \frac{\partial\varphi^*(\mathbf{A},\widehat{\mathbf{B}})}{\partial\mathbf{a}} \,,
\end{aligned}$$

where the last equality results from equation (4.3) in Theorem 4.3. Finally, using the fact that

$$\widehat{\mathbf{B}} = \mathrm{Arg}\min_{\mathbf{B}\in\mathbb{R}^{k\times n}} \varphi^*(\mathbf{A},\mathbf{B}) = \frac{1}{2}\|\mathbf{x} - \mathbf{F}(\mathbf{a})\mathbf{b}\|_2^2 \,,$$

we have

$$\frac{\partial\varphi^*(\mathbf{A},\widehat{\mathbf{B}})}{\partial\mathbf{b}} = \mathbf{0}^{k.n} \,,$$

and so the vectorized form of the Euclidean gradient of $\varphi^*(.)$ at $(\mathbf{a}, \widehat{\mathbf{b}})$ is given by

$$\nabla\varphi^*(\mathbf{A}, \widehat{\mathbf{B}}) = \begin{bmatrix} \frac{\partial\varphi^*(\mathbf{A},\widehat{\mathbf{B}})}{\partial\mathbf{a}} & \mathbf{0}^{k.n} \end{bmatrix} ,$$

which implies that

$$\|\nabla\psi(\mathbf{a})\|_2 = \left\|\nabla\varphi^*_{\mathbf{a}}(\mathbf{A}, \widehat{\mathbf{B}})\right\|_2 .$$

Next, the equality

$$\|\nabla\psi(\mathbf{a})\|_2 = \left\|\nabla\varphi^*_{\mathbf{A}}(\mathbf{A}, \widehat{\mathbf{B}})\right\|_F$$

is a direct consequence of equation (4.5) in Theorem 4.3.

Finally, the above demonstration remains valid if $\widehat{\mathbf{b}} = \mathbf{F}(\mathbf{a})^{-}\mathbf{x}$ because the differential formula (5.12) for an orthogonal projector is unchanged if a symmetric generalized inverse is used in place of the pseudo-inverse and the residual vector $\mathbf{r}(\mathbf{a})$ is also identical since $\mathbf{r}(\mathbf{a}) = \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\mathbf{x}$. This concludes the proof of Theorem 5.7.

□

Remembering that $\mathbf{G}(\widehat{\mathbf{b}})$ is a block-diagonal matrix, we see that the computation of $\nabla\psi(\mathbf{a})$ is easy, fast and may be efficiently parallelized using Theorem 5.7, since $\mathbf{G}(\widehat{\mathbf{b}})^T\mathbf{G}(\widehat{\mathbf{b}})$ is also a block-diagonal matrix. We further highlight that this formulation of $\nabla\psi(\mathbf{a})$ is much more efficient than the formulae given in Chen [28] (see its equations 24 and 27), which does not exploit the fact that the residual vector $\mathbf{r}(\mathbf{a})$ is linear in both $\mathbf{a}$ and $\mathbf{b}$, and involved multiplication of matrices with many zeros.

With the help of Theorem 5.7, it is also easy to demonstrate that the Gauss-Newton directions defined in Subsection 5.2 are in a descent direction for $\psi(.)$ despite the systematic rank deficiency of the Jacobian matrix $J(\mathbf{r}(\mathbf{a}))$ or of its approximation $-\mathbf{M}(\mathbf{a})$ proved in Theorem 5.2.

Corollary 5.7. Let $\mathbf{A} \in \mathbb{R}^{p\times k}_k$ and $\mathbf{a} = vec(\mathbf{A}^T) \in \mathbb{R}^{k.p}$.

If $\nabla\psi(\mathbf{a}) \neq \mathbf{0}^{p.k}$, e.g., if $\mathbf{a}$ is not a first-order stationary point of $\psi(.)$, the Gauss-Newton directions defined by

$$\begin{aligned} d\mathbf{a}_{gn} &= -J\big(\mathbf{r}(\mathbf{a})\big)^{+}\mathbf{r}(\mathbf{a}) = \big(\mathbf{M}(\mathbf{a}) + \mathbf{L}(\mathbf{a})\big)^{+}\mathbf{r}(\mathbf{a}) , \\ d\mathbf{a}_{gn} &= \mathbf{M}(\mathbf{a})^{+}\mathbf{r}(\mathbf{a}) \end{aligned}$$

are in a descent direction for $\psi(.)$.

Proof. To prove the assertion if the Gauss-Newton direction $d\mathbf{a}_{gn}$ is defined from the pseudo-inverse of the full Jacobian matrix $J(\mathbf{r}(\mathbf{a}))$, note that

$$\begin{aligned} d\mathbf{a}^T_{gn}\nabla\psi(\mathbf{a}) &= -\mathbf{r}(\mathbf{a})^T J\big(\mathbf{r}(\mathbf{a})\big)^{+T} J\big(\mathbf{r}(\mathbf{a})\big)^T\mathbf{r}(\mathbf{a}) \\ &= -\mathbf{r}(\mathbf{a})^T J\big(\mathbf{r}(\mathbf{a})\big) J\big(\mathbf{r}(\mathbf{a})\big)^{+}\mathbf{r}(\mathbf{a}) \\ &= -\mathbf{r}(\mathbf{a})^T\mathbf{P}_{J(\mathbf{r}(\mathbf{a}))}\mathbf{r}(\mathbf{a}) \\ &= -\mathbf{r}(\mathbf{a})^T\mathbf{P}^T_{J(\mathbf{r}(\mathbf{a}))}\mathbf{P}_{J(\mathbf{r}(\mathbf{a}))}\mathbf{r}(\mathbf{a}) \\ &= -\|\mathbf{P}_{J(\mathbf{r}(\mathbf{a}))}\mathbf{r}(\mathbf{a})\|^2_2 , \end{aligned}$$

since the projector $\mathbf{P}_{J(\mathbf{r}(\mathbf{a}))}$ is a symmetric and idempotent matrix. Further, $\nabla\psi(\mathbf{a}) \neq \mathbf{0}^{p.k}$ by hypothesis, then $J(\mathbf{r}(\mathbf{a}))^T\mathbf{r}(\mathbf{a}) \neq \mathbf{0}^{p.k}$ and $\mathbf{r}(\mathbf{a})$ is not in the null space of $J(\mathbf{r}(\mathbf{a}))^T$ and is not the zero-vector. Since

$$null\Big(J\big(\mathbf{r}(\mathbf{a})\big)^T\Big) = ran\Big(J\big(\mathbf{r}(\mathbf{a})\big)\Big)^{\perp} = null(\mathbf{P}_{J(\mathbf{r}(\mathbf{a}))}),$$

it follows that $\mathbf{P}_{J(\mathbf{r}(\mathbf{a}))}\mathbf{r}(\mathbf{a}) \neq \mathbf{0}^{p.n}$ and $d\mathbf{a}_{gn}^T \nabla\psi(\mathbf{a}) = -\|\mathbf{P}_{J(\mathbf{r}(\mathbf{a}))}\mathbf{r}(\mathbf{a})\|_2^2 < 0$. This proves the first assertion. The second assertion if the Gauss-Newton direction $d\mathbf{a}_{gn}$ is defined from the pseudo-inverse of the approximate Jacobian matrix $-\mathbf{M}(\mathbf{a})$ is verified by the same way since

$$\nabla\psi(\mathbf{a}) = -\mathbf{M}(\mathbf{a})^T\mathbf{r}(\mathbf{a}),$$

as proved in Theorem 5.7.

□

These results show that it is appropriate to use a line search algorithm in the Gauss-Newton methods described in Subsection 5.2 in order to obtain global convergence even though the Jacobian matrix or its approximation are always singular. Moreover, similar results hold for the Levenberg-Marquardt directions defined in Subsection 5.2 as demonstrated in the following corollary.

Corollary 5.8. Let $\mathbf{A} \in \mathbb{R}_k^{p\times k}$, $\mathbf{a} = vec(\mathbf{A}^T) \in \mathbb{R}^{p.k}$, $\lambda \in \mathbb{R}_{+*}$ be the Marquardt damping parameter and $\mathbf{D}$ be a diagonal scaling matrix of order $p.k$ with diagonal elements $\mathbf{D}_{ii} > 0$.

If $\nabla\psi(\mathbf{a}) \neq \mathbf{0}^{k.p}$, e.g., if $\mathbf{a}$ is not a first-order stationary point of $\psi(.)$, the Levenberg-Marquardt directions

$$\begin{aligned}
d\mathbf{a}_{lm} &= -\left(J\big(\mathbf{r}(\mathbf{a})\big)^T J\big(\mathbf{r}(\mathbf{a})\big) + \lambda\mathbf{D}^T\mathbf{D}\right)^{-1} J\big(\mathbf{r}(\mathbf{a})\big)^T\mathbf{r}(\mathbf{a}) \,, \\
d\mathbf{a}_{lm} &= \left(\mathbf{M}(\mathbf{a})^T\mathbf{M}(\mathbf{a}) + \lambda\mathbf{D}^T\mathbf{D}\right)^{-1}\mathbf{M}(\mathbf{a})^T\mathbf{r}(\mathbf{a}) \,, \\
d\mathbf{a}_{lm} &= \left(\mathbf{M}(\mathbf{a})^T\mathbf{M}(\mathbf{a}) + \mathbf{N}\mathbf{N}^T + \lambda\mathbf{D}^T\mathbf{D}\right)^{-1}\mathbf{M}(\mathbf{a})^T\mathbf{r}(\mathbf{a}) \,, \\
d\mathbf{a}_{lm} &= -\left(J\big(\mathbf{r}(\mathbf{a})\big)^T J\big(\mathbf{r}(\mathbf{a})\big) + \mathbf{N}\mathbf{N}^T + \lambda\mathbf{D}^T\mathbf{D}\right)^{-1} J\big(\mathbf{r}(\mathbf{a})\big)^T\mathbf{r}(\mathbf{a}) \,,
\end{aligned}$$

where $\mathbf{N} = \mathbf{K}_{(p,k)}(\mathbf{I}_k \otimes \mathbf{A})$, are well defined and are also in a descent direction for $\psi(.)$.

In addition, if $\lambda = 0$ and $rank(J(\mathbf{r}(\mathbf{a}))) = rank(\mathbf{M}(\mathbf{a})) = (p-k).k$, these two last Levenberg-Marquardt directions are also well-defined, again in a descent direction for $\psi(.)$ and equal to the corresponding Gauss-Newton directions defined in Corollary 5.7.

Proof. To prove the first part of the Corollary, we note that $\lambda \neq 0$ and all the elements of the diagonal matrix $\mathbf{D}$ are strictly positive by hypothesis. In these conditions, the matrices

$$\begin{aligned}
&\left(J\big(\mathbf{r}(\mathbf{a})\big)^T J\big(\mathbf{r}(\mathbf{a})\big) + \lambda\mathbf{D}^T\mathbf{D}\right) && , \left(\mathbf{M}(\mathbf{a})^T\mathbf{M}(\mathbf{a}) + \lambda\mathbf{D}^T\mathbf{D}\right), \\
&\left(J\big(\mathbf{r}(\mathbf{a})\big)^T J\big(\mathbf{r}(\mathbf{a})\big) + \lambda\mathbf{D}^T\mathbf{D} + \mathbf{N}\mathbf{N}^T\right) && , \left(\mathbf{M}(\mathbf{a})^T\mathbf{M}(\mathbf{a}) + \lambda\mathbf{D}^T\mathbf{D} + \mathbf{N}\mathbf{N}^T\right) ,
\end{aligned}$$

are all positive definite and the associated the Levenberg-Marquardt directions are thus well-defined. Furthermore, in these conditions, if $\mathbf{C}$ represents any of these positive definite matrices, we have immediately

$$d\mathbf{a}_{lm}^T\nabla\psi(\mathbf{a}) = d\mathbf{a}_{lm}^T J\big(\mathbf{r}(\mathbf{a})\big)^T\mathbf{r}(\mathbf{a}) = -d\mathbf{a}_{lm}^T\mathbf{C}d\mathbf{a}_{lm} < 0 \,,$$

since the hypothesis $\nabla\psi(\mathbf{a}) \neq \mathbf{0}^{p.k}$ implies that $d\mathbf{a}_{lm} \neq \mathbf{0}^{p.k}$ and $\mathbf{C}$ is positive definite. In other words, all the associated Levenberg-Marquardt directions are in a descent direction for $\psi(.)$ if $\lambda \neq 0$, which proves the first part of the Corollary.

On the other hand, if $\lambda = 0$ and $rank(J(\mathbf{r}(\mathbf{a}))) = rank(\mathbf{M}(\mathbf{a})) = (p-k).k$, we have

$$d\mathbf{a}_{lm} = -\left(J\big(\mathbf{r}(\mathbf{a})\big)^T J\big(\mathbf{r}(\mathbf{a})\big) + \mathbf{N}\mathbf{N}^T\right)^{-1} J\big(\mathbf{r}(\mathbf{a})\big)^T\mathbf{r}(\mathbf{a})$$

or

$$d\mathbf{a}_{lm} = \left(\mathbf{M}(\mathbf{a})^T\mathbf{M}(\mathbf{a}) + \mathbf{N}\mathbf{N}^T\right)^{-1}\mathbf{M}(\mathbf{a})^T\mathbf{r}(\mathbf{a}) \,.$$

Further, noting that the matrices

$$\begin{bmatrix} J\big(\mathbf{r}(\mathbf{a})\big) \\ \mathbf{N}^T \end{bmatrix} \text{ and } \begin{bmatrix} \mathbf{M}(\mathbf{a}) \\ \mathbf{N}^T \end{bmatrix}$$

are of full column rank, as demonstrated in equation (5.24) of Subsection 5.2, we deduce that the matrices

$$\Big(J\big(\mathbf{r}(\mathbf{a})\big)^T J\big(\mathbf{r}(\mathbf{a})\big) + \mathbf{N}\mathbf{N}^T\Big) \text{ and } \big(\mathbf{M}(\mathbf{a})^T\mathbf{M}(\mathbf{a}) + \mathbf{N}\mathbf{N}^T\big)$$

are both positive definite. This implies that the corresponding Levenberg-Marquardt directions are again well defined and still in a descent direction for $\psi(.)$ by similar arguments as used in the first part of the demonstration.

Finally, the linear systems

$$\Big(J\big(\mathbf{r}(\mathbf{a})\big)^T J\big(\mathbf{r}(\mathbf{a})\big) + \mathbf{N}\mathbf{N}^T\Big)\, d\mathbf{a}_{lm} = -J\big(\mathbf{r}(\mathbf{a})\big)^T \mathbf{r}(\mathbf{a})$$

and

$$\big(\mathbf{M}(\mathbf{a})^T\mathbf{M}(\mathbf{a}) + \mathbf{N}\mathbf{N}^T\big)\, d\mathbf{a}_{lm} = \mathbf{M}(\mathbf{a})^T\mathbf{r}(\mathbf{a})$$

are, respectively, the normal equations of the linear least-squares problems

$$\min_{d\mathbf{a}\in\mathbb{R}^{p.k}} \frac{1}{2}\Big\| \begin{bmatrix} \mathbf{r}(\mathbf{a}) \\ \mathbf{0}^{k_{\mathbf{A}}.k} \end{bmatrix} + \begin{bmatrix} J\big(\mathbf{r}(\mathbf{a})\big) \\ \mathbf{N}^T \end{bmatrix} d\mathbf{a}\Big\|_2^2 \text{ and } \min_{d\mathbf{a}\in\mathbb{R}^{p.k}} \frac{1}{2}\Big\| \begin{bmatrix} \mathbf{r}(\mathbf{a}) \\ \mathbf{0}^{k_{\mathbf{A}}.k} \end{bmatrix} - \begin{bmatrix} \mathbf{M}(\mathbf{a}) \\ \mathbf{N}^T \end{bmatrix} d\mathbf{a}\Big\|_2^2 ,$$

which both have an unique solution, as the associated coefficient matrices are of full column rank, and these solutions are, respectively, the minimum 2-norm solutions of the rank deficient linear least-squares problems

$$\min_{d\mathbf{a}\in\mathbb{R}^{p.k}} \frac{1}{2}\big\|\mathbf{r}(\mathbf{a}) + J\big(\mathbf{r}(\mathbf{a})\big)d\mathbf{a}\big\|_2^2 \text{ and } \min_{d\mathbf{a}\in\mathbb{R}^{p.k}} \frac{1}{2}\big\|\mathbf{r}(\mathbf{a}) - \mathbf{M}(\mathbf{a})d\mathbf{a}\big\|_2^2 ,$$

as demonstrated in Subsection 5.2. In other words, we have

$$d\mathbf{a}_{lm} = -J\big(\mathbf{r}(\mathbf{a})\big)^+\mathbf{r}(\mathbf{a}) = d\mathbf{a}_{gn} \text{ or } d\mathbf{a}_{lm} = \mathbf{M}(\mathbf{a})^+\mathbf{r}(\mathbf{a}) = d\mathbf{a}_{gn} ,$$

if an approximate Jacobian matrix $-\mathbf{M}(\mathbf{a})$ is used. This concludes the demonstration of the Corollary.

□

We now explore the relationships between the Euclidean gradient $\nabla\psi(\mathbf{a})$ of $\psi(.)$ at $\mathbf{a}$, considered as a real function from $\mathbb{R}^{p.k}$ into $\mathbb{R}$ (e.g., of the vectorized form of the $\mathbf{A}$ matrix, see equation (3.23)), and the Riemannian gradient of the unvectorized form of $\psi(.)$ (e.g., $\psi \circ h^{-1}(.)$ where $h^{-1}(.)$ is defined in equation (3.29) of Subsection 3.4 with $h^{-1}(\mathbf{A}) = vec(\mathbf{A}^T) = \mathbf{a}, \forall \mathbf{A} \in \mathbb{R}_k^{p\times k}$), when this cost function is considered as defined on the Grassmann manifold $\mathrm{Gr}(p,k)$, as already discussed at the end of Subsection 5.2.

To this end, we require that $\mathbf{W} \in \mathbb{R}_{+*}^{p\times n}$ (as otherwise $\psi \circ h^{-1}(.)$ is not smooth on its whole domain $\mathbb{R}_k^{p\times k}$) and that each element of $\mathrm{Gr}(p,k)$ is represented by an element of the compact Stiefel manifold $\mathrm{St}(p,k) = \mathbb{O}^{p\times k}$ instead of $\mathbb{R}_k^{p\times k}$ as it is customary for simplicity and numerical reasons in previous works on Riemannian optimization on $\mathrm{Gr}(p,k)$ [47][14][11]. With these requirements, we first observe that the restriction of the mapping $\psi \circ h^{-1}(.)$ to the domain $\mathrm{St}(p,k) = \mathbb{O}^{p\times k}$ can be considered as a smooth map defined on the Stiefel manifold and, by extension, also on the Grassmann manifold as the Grassmann manifold $\mathrm{Gr}(p,k)$ is a Riemannian quotient manifold of $\mathrm{St}(p,k)$ by the action of the orthogonal group $\mathbb{O}^{k\times k}$, see Subsections 2.4 and 5.2, and [3][11] for more details. In these conditions, the Riemannian gradient of this smooth map defined on the Grassmann manifold $\mathrm{Gr}(p,k)$ at $\mathbf{O} \in \mathrm{St}(p,k)$, considered as the matrix representation of $\mathring{\mathbf{O}} =$

ran$(\mathbf{O}) \in \mathrm{Gr}(p,k)$, is an element of the tangent space of $\mathrm{Gr}(p,k)$ at $\mathring{\mathbf{O}}$, $\mathcal{T}_{\mathring{\mathbf{O}}}\mathrm{Gr}(p,k)$, which is a linear subspace of dimension *dim*$\big(\mathrm{Gr}(p,k)\big) = k.(p-k)$. Moreover, any element of $\mathcal{T}_{\mathring{\mathbf{O}}}\mathrm{Gr}(p,k)$ can be represented uniquely by a matrix $\mathbf{D} \in \mathbb{R}^{p\times k}$ verifying $\mathbf{O}^T\mathbf{D} = \mathbf{0}^{k\times k}$, as already noted in Subsection 5.2. This subset of $\mathbb{R}^{p\times k}$ is noted $\mathcal{T}_{\mathbf{O}}\mathrm{Gr}(p,k)$ and is nothing else than the horizontal space $\mathcal{H}_{\mathbf{O}}\mathbb{O}^{p\times k}$ of $\mathrm{St}(p,k) = \mathbb{O}^{p\times k}$ at $\mathbf{O} \in \mathrm{St}(p,k)$.

Thus, the Riemannian gradient of $\psi \circ h^{-1}(.)$ at $\mathring{\mathbf{O}} \in \mathrm{Gr}(p,k)$, noted $\nabla_R\psi \circ h^{-1}(\mathring{\mathbf{O}})$, can be represented uniquely by one element of $\mathcal{T}_{\mathbf{O}}\mathrm{Gr}(p,k)$ and, according to equation (2.51) in Subsection 2.4, this Riemannian gradient is given, with a slight abuse of notation, by

$$\begin{aligned}\nabla_R\psi \circ h^{-1}(\mathring{\mathbf{O}}) &= \mathbf{P}_{\mathcal{H}_{\mathbf{O}}\mathbb{O}^{p\times k}}\nabla_F\psi \circ h^{-1}(\mathbf{O}) \\ &= (\mathbf{I}_p - \mathbf{O}\mathbf{O}^T)\nabla_F\psi \circ h^{-1}(\mathbf{O}) \;, \end{aligned} \tag{5.27}$$

where $\mathbf{O} \in \mathrm{St}(p,k) = \mathbb{O}^{p\times k}$, $\mathring{\mathbf{O}} \in \mathrm{Gr}(p,k)$, $\mathbf{P}_{\mathcal{H}_{\mathbf{O}}\mathbb{O}^{p\times k}} = \mathbf{I}_p - \mathbf{O}\mathbf{O}^T$ is the orthogonal projector onto the orthogonal of the range of $\mathbf{O}$ in $\mathbb{R}^p$ (e.g., *ran*$(\mathbf{O})^\perp$) and also the orthogonal projector on the horizontal space of $\mathrm{St}(p,k) = \mathbb{O}^{p\times k}$ at $\mathbf{O}$ (in the linear spcace $\mathbb{R}^{p\times k}$) and, finally, $\nabla_F\psi \circ h^{-1}(\mathbf{O})$ is the usual Frobenius gradient of $\psi \circ h^{-1}(.)$ at $\mathbf{O}$ when $\psi \circ h^{-1}(.)$ is considered as a real function defined on the linear space $\mathbb{R}^{p\times k}$. See [3][14][11] for a derivation of this standard result on the geometry of the Stiefel and Grassmann manifolds.

Next, we first observe that

$$\nabla\psi(\mathbf{o}) = vec\Big(\big(\nabla_F\psi \circ h^{-1}(\mathbf{O})\big)^T\Big) \in \mathbb{R}^{p.k} \;,$$

since following the conventions used throughout the monograph, we have defined $\mathbf{a} = vec(\mathbf{A}^T)$, $\forall\mathbf{A} \in \mathbb{R}^{p\times k}$, instead of $\mathbf{a} = vec(\mathbf{A})$ as it is commonly used in past works on the WLRA problem. In words, $\nabla\psi(\mathbf{o})$ is the vectorized form of the transpose of the Frobenius gradient of $\psi \circ h^{-1}(.)$ at $\mathbf{O}$. Using a similar and consistent convention, we now define

$$\nabla_R\psi(\mathbf{o}) = vec\Big(\big(\nabla_R\psi \circ h^{-1}(\mathring{\mathbf{O}})\big)^T\Big) \in \mathbb{R}^{p.k} \;,$$

e.g., $\nabla_R\psi(\mathbf{o})$ is the vectorized form of the transpose of the Riemannian gradient of $\psi \circ h^{-1}(.)$ at $\mathring{\mathbf{O}} \in \mathrm{Gr}(p,k)$. In these conditions, the equality (5.27) defining the Riemannian gradient of $\psi \circ h^{-1}(.)$ at $\mathring{\mathbf{O}}$ is equivalent to the matrix equality

$$\big(\nabla_R\psi \circ h^{-1}(\mathring{\mathbf{O}})\big)^T = \big(\nabla_F\psi \circ h^{-1}(\mathbf{O})\big)^T\big(\mathbf{I}_p - \mathbf{O}\mathbf{O}^T\big) \;,$$

as an orthogonal projector is a symmetric matrix (see equation (2.13)), and also to the vector equality

$$\begin{aligned} vec\Big(\big(\nabla_R\psi \circ h^{-1}(\mathring{\mathbf{O}})\big)^T\Big) &= vec\Big(\big(\nabla_F\psi \circ h^{-1}(\mathbf{O})\big)^T(\mathbf{I}_p - \mathbf{O}\mathbf{O}^T)\Big) \\ &= \big((\mathbf{I}_p - \mathbf{O}\mathbf{O}^T) \otimes \mathbf{I}_k\big)vec\Big(\big(\nabla_F\psi \circ h^{-1}(\mathbf{O})\big)^T\Big) \;, \end{aligned}$$

where we have used again the symmetry of the orthogonal projector and equation (2.33). In other words, the vectorized form of the Riemannnian gradient of $\psi \circ h^{-1}(.)$ at $\mathring{\mathbf{O}}$ is given by

$$\nabla_R\psi(\mathbf{o}) = \big((\mathbf{I}_p - \mathbf{O}\mathbf{O}^T) \otimes \mathbf{I}_k\big)\nabla\psi(\mathbf{o}) \;.$$

Before proceeding further, we now precise the nature of the $p.k \times p.k$ symmetric matrix $(\mathbf{I}_p - \mathbf{O}\mathbf{O}^T) \otimes \mathbf{I}_k$. More precisely, we show that this symmetric matrix is the orthogonal projector onto $null(J(\mathbf{r}(\mathbf{o})))^\perp = null(\mathbf{M}(\mathbf{o}))^\perp$ when $\mathbf{W} \in \mathbb{R}^{p\times n}_{+*}$ and $\mathbf{O} \in \mathrm{St}(p,k) = \mathbb{O}^{p\times k}$. To this end, we first recall that, in these conditions and according to Corollary 5.6, the columns of the matrix $\bar{\mathbf{O}} = \mathbf{K}_{(p,k)}(\mathbf{I}_k \otimes \mathbf{O})$ form an orthonormal basis of *null*$\big(J(\mathbf{r}(\mathbf{o}))\big) = null(\mathbf{M}(\mathbf{o}))$ and that the columns of $\bar{\mathbf{O}}^\perp = \mathbf{K}_{(p,k)}(\mathbf{I}_k \otimes \mathbf{O}^\perp)$ are an orthonormal basis of *null*$\big(J(\mathbf{r}(\mathbf{o}))\big)^\perp = null(\mathbf{M}(\mathbf{o}))^\perp$

as soon as the columns of $\mathbf{O}$ and $\mathbf{O}^{\perp}$ are orthonormal bases of $ran(\mathbf{O})$ and $ran(\mathbf{O})^{\perp}$, respectively. Next, using equation (2.29) in Subsection 2.2, we have

$$(\mathbf{I}_p - \mathbf{O}\mathbf{O}^T) \otimes \mathbf{I}_k = \mathbf{I}_{p.k} - (\mathbf{O}\mathbf{O}^T \otimes \mathbf{I}_k) \; .$$

Furthermore, using the identities (2.32) and (2.36) again in Subsection 2.2, we deduce that

$$\begin{aligned} \mathbf{O}\mathbf{O}^T \otimes \mathbf{I}_k &= (\mathbf{O} \otimes \mathbf{I}_k)(\mathbf{O}^T \otimes \mathbf{I}_k) \\ &= (\mathbf{O} \otimes \mathbf{I}_k)\mathbf{K}_{(k,k)}\mathbf{K}_{(k,k)}(\mathbf{O}^T \otimes \mathbf{I}_k) \\ &= \big(\mathbf{K}_{(p,k)}(\mathbf{I}_k \otimes \mathbf{O})\big)\big(\mathbf{K}_{(p,k)}(\mathbf{I}_k \otimes \mathbf{O})\big)^T \\ &= \bar{\mathbf{O}}\bar{\mathbf{O}}^T \; . \end{aligned}$$

Thus, $\mathbf{O}\mathbf{O}^T \otimes \mathbf{I}_k = \bar{\mathbf{O}}\bar{\mathbf{O}}^T$ is the orthogonal projector onto $null\big(J(\mathbf{r}(\mathbf{o}))\big) = null(\mathbf{M}(\mathbf{o}))$ and

$$\mathbf{I}_{p.k} - (\mathbf{O}\mathbf{O}^T \otimes \mathbf{I}_k) = \mathbf{I}_{p.k} - \bar{\mathbf{O}}\bar{\mathbf{O}}^T = \bar{\mathbf{O}}^{\perp}(\bar{\mathbf{O}}^{\perp})^T$$

is the orthogonal projector onto $null\big(J(\mathbf{r}(\mathbf{o}))\big)^{\perp} = null(\mathbf{M}(\mathbf{o}))^{\perp}$ as stated above. Using these results, Theorem 5.7, and remembering that $\bar{\mathbf{O}}$ is an orthogonal basis of $null(\mathbf{M}(\mathbf{o}))$ and, thus, that $\mathbf{M}(\mathbf{o})\bar{\mathbf{O}} = \mathbf{0}^{p.n \times k.k}$, we deduce that

$$\begin{aligned} \nabla_R\psi(\mathbf{o}) &= (\mathbf{I}_{p.k} - \bar{\mathbf{O}}\bar{\mathbf{O}}^T)\nabla\psi(\mathbf{o}) \\ &= \nabla\psi(\mathbf{o}) - \bar{\mathbf{O}}\bar{\mathbf{O}}^T\nabla\psi(\mathbf{o}) \\ &= \nabla\psi(\mathbf{o}) + \bar{\mathbf{O}}\bar{\mathbf{O}}^T\mathbf{M}(\mathbf{o})^T\mathbf{r}(\mathbf{o}) \\ &= \nabla\psi(\mathbf{o}) + \bar{\mathbf{O}}\big(\mathbf{M}(\mathbf{o})\bar{\mathbf{O}}\big)^T\mathbf{r}(\mathbf{o}) \\ &= \nabla\psi(\mathbf{o}) \; . \end{aligned}$$

From the vectorized equality $\nabla_R\psi(\mathbf{o}) = \nabla\psi(\mathbf{o})$, by identification and unicity of the Frobenius (or Riemaniann) gradient, we deduce finally that

$$\nabla_R\psi \circ h^{-1}(\mathring{\mathbf{O}}) = \nabla_F\psi \circ h^{-1}(\mathbf{O}) \; .$$

In other words, the Riemannian gradient of $\psi \circ h^{-1}(.)$ at $\mathring{\mathbf{O}}$ is simply equal to its Frobenius gradient at $\mathbf{O}$, which again illustrates the tight relationships between the variable projection method used here and the Riemannian optimization framework on the Grassmann manifold developed in Boumal and Absil [13][14] despite the derivations are completely different.

To conclude this paragraph on the comparison of the Euclidean and Riemannian gradients of the cost function $\psi(.)$, we now derive an explicit matrix form of $\nabla_F\psi \circ h^{-1}(\mathbf{A})$ for $\mathbf{A} \in \mathbb{R}_k^{p \times k}$ starting from the vectorized equality

$$\nabla\psi(\mathbf{a}) = vec\Big(\big(\nabla_F\psi \circ h^{-1}(\mathbf{A})\big)^T\Big) \; .$$

Using Theorems 4.3 and 5.7, we first recall that

$$\nabla\psi(\mathbf{a}) = \mathbf{G}(\widehat{\mathbf{b}})^T\big(\mathbf{G}(\widehat{\mathbf{b}})\mathbf{a} - \mathbf{z}\big),$$

where, as usual,

$$\begin{aligned} \mathbf{a} &= vec(\mathbf{A}^T) \; , \\ \widehat{\mathbf{b}} &= \mathbf{F}(\mathbf{a})^{+}\mathbf{x} = \Big(diag\big(vec(\sqrt{\mathbf{W}})\big)\big(\mathbf{I}_n \otimes \mathbf{A}\big)\Big)^{+}\mathbf{x} \; , \\ \mathbf{z} &= vec\big((\sqrt{\mathbf{W}} \odot \mathbf{X})^T\big) = diag\big(vec(\sqrt{\mathbf{W}}^T)\big)vec(\mathbf{X}^T) \; , \\ \mathbf{G}(\widehat{\mathbf{b}}) &= diag\big(vec(\sqrt{\mathbf{W}}^T)\big)(\mathbf{I}_p \otimes \widehat{\mathbf{B}}^T) \; . \end{aligned}$$

Using these results and equation (2.33) in Subsection 2.2, we deduce that

$$\begin{aligned}
\nabla\psi(\mathbf{a}) &= (\mathbf{I}_p \otimes \widehat{\mathbf{B}})\mathit{diag}\big(\mathit{vec}(\mathbf{W}^T)\big)\big((\mathbf{I}_p \otimes \widehat{\mathbf{B}}^T)\mathit{vec}(\mathbf{A}^T) - \mathit{vec}(\mathbf{X}^T)\big) \\
&= (\mathbf{I}_p \otimes \widehat{\mathbf{B}})\mathit{diag}\big(\mathit{vec}(\mathbf{W}^T)\big)\big(\mathit{vec}(\widehat{\mathbf{B}}^T\mathbf{A}^T) - \mathit{vec}(\mathbf{X}^T)\big) \\
&= (\mathbf{I}_p \otimes \widehat{\mathbf{B}})\mathit{diag}\big(\mathit{vec}(\mathbf{W}^T)\big)\mathit{vec}(\widehat{\mathbf{B}}^T\mathbf{A}^T - \mathbf{X}^T) \\
&= (\mathbf{I}_p \otimes \widehat{\mathbf{B}})\mathit{vec}\big(\mathbf{W}^T \odot (\mathbf{A}\widehat{\mathbf{B}} - \mathbf{X})^T\big) \\
&= (\mathbf{I}_p \otimes \widehat{\mathbf{B}})\mathit{vec}\Big(\big(\mathbf{W} \odot (\mathbf{A}\widehat{\mathbf{B}} - \mathbf{X})\big)^T\Big) \\
&= \mathit{vec}\Big(\widehat{\mathbf{B}}\big(\mathbf{W} \odot (\mathbf{A}\widehat{\mathbf{B}} - \mathbf{X})\big)^T\Big) \, .
\end{aligned}$$

By identification and the unicity of the (Frobenius) gradient and equation (3.11) of Subsection 3.2, we obtain, finally, the following unvectorized form of $\nabla\psi(\mathbf{a})$ as:

$$\nabla_F\psi \circ h^{-1}(\mathbf{A}) = \big(\mathbf{W} \odot (\mathbf{A}\widehat{\mathbf{B}} - \mathbf{X})\big)\widehat{\mathbf{B}}^T = \nabla\varphi^*_{\mathbf{A}}(\mathbf{A}, \widehat{\mathbf{B}}) \, ,$$

which is entirely consistent with the results in Theorem 5.7.

Consistent with the fact that $\nabla_R\psi \circ h^{-1}(\mathbf{O}) = \nabla_F\psi \circ h^{-1}(\mathbf{O})$ derived above, when $\mathbf{O} \in \mathrm{St}(p,k) = \mathbb{O}^{p\times k}$ and $\mathbf{W} \in \mathbb{R}^{p\times n}_{+*}$, it can be further verified that this formulae for $\nabla_F\psi \circ h^{-1}(\mathbf{A})$ is equal to the Riemannian gradient $\mathrm{grad}f(\mathbf{A})$ (which corresponds to $\nabla_R\psi \circ h^{-1}(\mathring{\mathbf{A}})$ in our notations) derived in equations 23 and 25 of Boumal and Absil [14], if we assumed that $\mathbf{A}$ is an orthonormal matrix, $\mathbf{W} = \mathbf{W}_\lambda \in \mathbb{R}^{p\times n}_{+*}$ and $\mathbf{X} = \mathbf{X}_\Omega$ as defined, respectively, in equations (3.16) and (3.17) of Subsection 3.3, e.g., when a regularization parameter $\lambda > 0$ is used in the Riemannian optimization method on the Grassmann manifold developed by Boumal and Absil [14] to minimize their cost function $f(.)$, which is nothing else than a variable projection form of the cost function $g_\lambda(.)$ defined in equation (3.18) already discussed in Subsection 3.3.

We summarize these different results about the unvectorized forms of the Frobenius and Riemannian gradients of $\psi(.)$ in the following Theorem:

Theorem 5.8. Let $\psi(.)$ and $h^{-1}(.)$ be defined, respectively, as in equations (3.23) and (3.29) of Subsection 3.4, $\mathbf{A} \in \mathbb{R}^{p\times k}_k$ and $\mathbf{a} = h^{-1}(\mathbf{A}) = \mathit{vec}(\mathbf{A}^T) \in \mathbb{R}^{p.k}$, and further assume that $\mathbf{a}$ belongs to an open set $\Omega \subset \mathbb{R}^{p.k}$ in which the matrix function $\mathbf{F}(\mathbf{a})$, defined in equation (3.20), has a constant rank, so that the Jacobian matrix $J(\mathbf{r}(\mathbf{a}))$ derived in Subsection 5.2, the Euclidean gradient $\nabla\psi(\mathbf{a})$ derived in Theorem 5.7 and the Frobenius gradient $\nabla_F\psi \circ h^{-1}(\mathbf{A})$ are all well-defined. Then

$$\nabla_F\psi \circ h^{-1}(\mathbf{A}) = \big(\mathbf{W} \odot (\mathbf{A}\widehat{\mathbf{B}} - \mathbf{X})\big)\widehat{\mathbf{B}}^T = \nabla\varphi^*_{\mathbf{A}}(\mathbf{A}, \widehat{\mathbf{B}}) \, ,$$

where $\mathbf{X} \in \mathbb{R}^{p\times n}$ and $\mathbf{W} \in \mathbb{R}^{p\times n}_+$ are, respectively, the data and weight matrices of the WLRA problem (P1) and $\widehat{\mathbf{B}} \in \mathbb{R}^{k\times n}$ is such that $\mathit{vec}(\widehat{\mathbf{B}}) = \mathbf{F}(\mathbf{a})^+\mathbf{x}$, with $\mathbf{x} = \mathit{vec}(\sqrt{\mathbf{W}} \odot \mathbf{X})$.

Furthermore, if we assume that $\mathbf{W} \in \mathbb{R}^{p\times n}_{+*}$, $\mathbf{O} \in \mathrm{St}(p,k) = \mathbb{O}^{p\times k}$ and $\mathring{\mathbf{O}} = \mathit{ran}(\mathbf{O}) \in \mathrm{Gr}(p,k)$, we have

$$\nabla_R\psi \circ h^{-1}(\mathring{\mathbf{O}}) = \nabla_F\psi \circ h^{-1}(\mathbf{O}) \, ,$$

where $\nabla_R\psi \circ h^{-1}(\mathring{\mathbf{O}})$ is the Riemannian gradient of the smooth cost function $\psi \circ h^{-1}(.)$ (defined now on the Grassmann manifold $\mathrm{Gr}(p,k)$) at $\mathring{\mathbf{O}} \in \mathrm{Gr}(p,k)$.

□

To conclude, we highlight that the above results about the comparison of the gradients in the variable projection and Grassmann manifold optimization frameworks are a slight extension of results derived in a different way in [82], who do not consider the case where a regularization parameter λ is used in the Boumal and Absil algorithms [13][14].

We now derive an explicit form for the Hessian matrix $\mathbf{H} = \nabla^2\psi(\mathbf{a})$ where $\mathbf{a} = h^{-1}(\mathbf{A}) = \mathit{vec}(\mathbf{A}^T) \in \mathbb{R}^{p.k}$ with $\mathbf{A} \in \mathbb{R}^{p\times k}_k$. To this end, we follow the derivation of the Newton step presented in Borges [10] for solving a general variable projection NLLS problem. In the context of

the WLRA problem, this Hessian matrix has been already explicitly derived in different forms by several authors [28][169][14], but these forms apply mainly to the case of binary weights or are not convenient to derive important properties of $\nabla^2\psi(\mathbf{a})$. As an illustration, Chen [28] has derived $\mathbf{H}$ (see its equation 25) for the case where $\mathbf{W}$ is a presence-absence matrix by expressing $\psi(\mathbf{a})$ as the sum of the j^{th} atomic functions $\psi_j(.)$ defined in equation (3.25) of Subsection 3.4:

$$\begin{aligned}\psi(\mathbf{a}) &= \frac{1}{2}\sum_{j=1}^{n}\psi_j(\mathbf{a}) \\ &= \frac{1}{2}\sum_{j=1}^{n}\big\|\big(\mathbf{I}_p - \mathbf{F}_j(\mathbf{a})\mathbf{F}_j(\mathbf{a})^+\big)\mathbf{x}_j\big\|_2^2 \\ &= \frac{1}{2}\sum_{j=1}^{n}\Big\|\Big(\mathbf{I}_p - \Big(\mathit{diag}(\sqrt{\mathbf{W}}_{.j})\mathbf{A}\Big)\Big(\mathit{diag}(\sqrt{\mathbf{W}}_{.j})\mathbf{A}\Big)^+\Big)\,(\sqrt{\mathbf{W}}_{.j}\odot\mathbf{X}_{.j})\Big\|_2^2\end{aligned}$$

and computing the Hessian matrices of $\psi_j(.)$ at $\mathbf{a}$, for $j = 1, \cdots, n$, and summing all these Hessian matrices. However, this formulation of $\nabla^2\psi(\mathbf{a})$ does not give a compact form for $\mathbf{H}$ and is not convenient to derive important properties of the Hessian matrix later in this subsection. The two other formulations of $\mathbf{H}$ derived in [169][14] apply, respectively, only to the binary weights case or only to the variable projection formulation of the regularized cost function $g_\lambda(.)$, defined in equation (3.18) of Subsection 3.3, and this one lacks also of generality as $\mathbf{H}$ is not obtained explicitly, but in the form of a directional derivative [14], which is used in an inexact subproblem solver based on a truncated conjugate gradient method to compute an approximate (Riemannian) Newton step at each iteration [14].

To start with, we recall that the Hessian matrix is given formally by

$$\mathbf{H} = J\big(\mathbf{r}(\mathbf{a})\big)^T J\big(\mathbf{r}(\mathbf{a})\big) + \sum_{l=1}^{n.p}\mathbf{r}_l(\mathbf{a})\nabla^2\mathbf{r}_l(\mathbf{a})\;, \tag{5.28}$$

where $\nabla^2\mathbf{r}_l(\mathbf{a})$ is the Hessian matrix of the functional $\mathbf{r}_l(\mathbf{a})$ for $l = 1, \cdots, n.p$ and is a $p.k \times p.k$ matrix given by

$$[\nabla^2\mathbf{r}_l(\mathbf{a})]_{ij} = \frac{\partial^2\mathbf{r}_l(\mathbf{a})}{\partial\mathbf{a}_i\partial\mathbf{a}_j} \text{ for } i = 1, \cdots, p.k \text{ and } j = 1, \cdots, p.k\;.$$

We first show that the calculation of the first term of $\nabla^2\psi(\mathbf{a})$, involving only the Jacobian matrix $J(\mathbf{r}(\mathbf{a}))$ and which corresponds exactly to the Gauss-Newton approximation of $\nabla^2\psi(\mathbf{a})$, can be simplified as for the gradient of $\psi(.)$. Using the notations and equation (5.22) of Subsection 5.2, we have

$$J(\mathbf{r}(\mathbf{a})) = -\big(\mathbf{M}(\mathbf{a}) + \mathbf{L}(\mathbf{a})\big)\;,$$

and it follows that

$$\begin{aligned}J\big(\mathbf{r}(\mathbf{a})\big)^T J\big(\mathbf{r}(\mathbf{a})\big) &= \big(\mathbf{M}(\mathbf{a}) + \mathbf{L}(\mathbf{a})\big)^T\big(\mathbf{M}(\mathbf{a}) + \mathbf{L}(\mathbf{a})\big) \\ &= \mathbf{M}(\mathbf{a})^T\mathbf{M}(\mathbf{a}) + \mathbf{M}(\mathbf{a})^T\mathbf{L}(\mathbf{a}) + \mathbf{L}(\mathbf{a})^T\mathbf{M}(\mathbf{a}) + \mathbf{L}(\mathbf{a})^T\mathbf{L}(\mathbf{a}) \\ &= \mathbf{M}(\mathbf{a})^T\mathbf{M}(\mathbf{a}) + \mathbf{L}(\mathbf{a})^T\mathbf{L}(\mathbf{a})\;,\end{aligned} \tag{5.29}$$

since $\mathit{ran}(\mathbf{M}(\mathbf{a})) \subset \mathit{ran}(\mathbf{F}(\mathbf{a}))^\perp$ and $\mathit{ran}(\mathbf{L}(\mathbf{a})) \subset \mathit{ran}(\mathbf{F}(\mathbf{a}))$, see Subsection 5.2 for more details.

We now derive an explicit expression for the second term of the Hessian matrix, given formally by

$$\mathbf{S} = \sum_{l=1}^{n.p}\mathbf{r}_l(\mathbf{a})\nabla^2\mathbf{r}_l(\mathbf{a})\;. \tag{5.30}$$

To this end, we first recall that the Jacobian matrix may also be expressed as

$$J(\mathbf{r}(\mathbf{a})) = -D(\mathbf{P}_{\mathbf{F}(\mathbf{a})})\mathbf{x}\ ,$$

where $\mathbf{x} = vec(\sqrt{\mathbf{W}} \odot \mathbf{X})$ (see equation (5.13) in Subsection 5.2) and this implies that the i^{th} column of $J(\mathbf{r}(\mathbf{a}))$ is given by

$$[J(\mathbf{r}(\mathbf{a}))]_{.j} = -\frac{\partial \mathbf{P}_{\mathbf{F}(\mathbf{a})}}{\partial \mathbf{a}_i}\mathbf{x} = \frac{\partial \mathbf{r}(\mathbf{a})}{\partial \mathbf{a}_i}\ ,$$

and it follows that

$$\frac{\partial^2 \mathbf{r}(\mathbf{a})}{\partial \mathbf{a}_i \partial \mathbf{a}_j} = -\frac{\partial^2 \mathbf{P}_{\mathbf{F}(\mathbf{a})}}{\partial \mathbf{a}_i \partial \mathbf{a}_j}\mathbf{x} \text{ for } i = 1, \cdots, k.p \text{ and } j = 1, \cdots, k.p\ .$$

Using this last equality and the definition of $\mathbf{S}$, it is not difficult to see that the ij element of $\mathbf{S}$ is given by

$$\mathbf{S}_{ij} = -\mathbf{r}(\mathbf{a})^T \frac{\partial^2 \mathbf{P}_{\mathbf{F}(\mathbf{a})}}{\partial \mathbf{a}_i \partial \mathbf{a}_j}\mathbf{x} \text{ for } i = 1, \cdots, k.p \text{ and } j = 1, \cdots, k.p\ .$$

Thus, in order to evaluate $\mathbf{S}$, we need an explicit expression for all the second partial derivatives of $\mathbf{P}_{\mathbf{F}(\mathbf{a})}$

$$\frac{\partial^2 \mathbf{P}_{\mathbf{F}(\mathbf{a})}}{\partial \mathbf{a}_i \partial \mathbf{a}_j} \in \mathcal{L}\big(\mathbb{R}, \mathcal{L}(\mathbb{R}, \mathbb{R}^{n.p \times n.p})\big) \text{ for } i = 1, \cdots, k.p \text{ and } j = 1, \cdots, k.p\ .$$

From the expression of $D(\mathbf{P}_{\mathbf{F}(\mathbf{a})})$ given in Corollary 5.1, it follows that

$$\begin{aligned}\frac{\partial \mathbf{P}_{\mathbf{F}(\mathbf{a})}}{\partial \mathbf{a}_i} &= \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\frac{\partial \mathbf{F}(\mathbf{a})}{\partial \mathbf{a}_i}\mathbf{F}(\mathbf{a})^+ + \Big(\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\frac{\partial \mathbf{F}(\mathbf{a})}{\partial \mathbf{a}_i}\mathbf{F}(\mathbf{a})^+\Big)^T \\ &= \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\frac{\partial \mathbf{F}(\mathbf{a})}{\partial \mathbf{a}_i}\mathbf{F}(\mathbf{a})^+ + \big(\mathbf{F}(\mathbf{a})^+\big)^T\frac{\partial \mathbf{F}(\mathbf{a})^T}{\partial \mathbf{a}_i}\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\ . \end{aligned} \tag{5.31}$$

Using the product rule to take the partial derivative of $\partial \mathbf{P}_{\mathbf{F}(\mathbf{a})}/\partial \mathbf{a}_i$ with respect to $\mathbf{a}_j$ [26], we obtain

$$\begin{aligned}\frac{\partial^2 \mathbf{P}_{\mathbf{F}(\mathbf{a})}}{\partial \mathbf{a}_i \partial \mathbf{a}_j} = &\frac{\partial \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}}{\partial \mathbf{a}_j}\frac{\partial \mathbf{F}(\mathbf{a})}{\partial \mathbf{a}_i}\mathbf{F}(\mathbf{a})^+ \quad + \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\frac{\partial^2 \mathbf{F}(\mathbf{a})}{\partial \mathbf{a}_i \partial \mathbf{a}_j}\mathbf{F}(\mathbf{a})^+ + \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\frac{\partial \mathbf{F}(\mathbf{a})}{\partial \mathbf{a}_i}\frac{\partial \mathbf{F}(\mathbf{a})^+}{\partial \mathbf{a}_j} \\ &+ \Big(\frac{\partial \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}}{\partial \mathbf{a}_j}\frac{\partial \mathbf{F}(\mathbf{a})}{\partial \mathbf{a}_i}\mathbf{F}(\mathbf{a})^+ \quad + \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\frac{\partial^2 \mathbf{F}(\mathbf{a})}{\partial \mathbf{a}_i \partial \mathbf{a}_j}\mathbf{F}(\mathbf{a})^+ + \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\frac{\partial \mathbf{F}(\mathbf{a})}{\partial \mathbf{a}_i}\frac{\partial \mathbf{F}(\mathbf{a})^+}{\partial \mathbf{a}_j}\Big)^T\ .\end{aligned}$$

For the sake of brevity, we define

$$\mathbf{D}_{(i,j)} = \frac{\partial \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}}{\partial \mathbf{a}_j}\frac{\partial \mathbf{F}(\mathbf{a})}{\partial \mathbf{a}_i}\mathbf{F}(\mathbf{a})^+ + \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\frac{\partial^2 \mathbf{F}(\mathbf{a})}{\partial \mathbf{a}_i \partial \mathbf{a}_j}\mathbf{F}(\mathbf{a})^+ + \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\frac{\partial \mathbf{F}(\mathbf{a})}{\partial \mathbf{a}_i}\frac{\partial \mathbf{F}(\mathbf{a})^+}{\partial \mathbf{a}_j}\ ,$$

and, thus,

$$\frac{\partial^2 \mathbf{P}_{\mathbf{F}(\mathbf{a})}}{\partial \mathbf{a}_i \partial \mathbf{a}_j} = \mathbf{D}_{(i,j)} + \mathbf{D}^T_{(i,j)}\ .$$

We will now simplify the computation of $\mathbf{D}_{(i,j)}$. First, using the fact that

$$\frac{\partial \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}}{\partial \mathbf{a}_j} = -\frac{\partial \mathbf{P}_{\mathbf{F}(\mathbf{a})}}{\partial \mathbf{a}_j}\ ,$$

we deduce that

$$\mathbf{D}_{(i,j)} = -\frac{\partial \mathbf{P}_{\mathbf{F}(\mathbf{a})}}{\partial \mathbf{a}_j}\frac{\partial \mathbf{F}(\mathbf{a})}{\partial \mathbf{a}_i}\mathbf{F}(\mathbf{a})^+ + \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\frac{\partial^2 \mathbf{F}(\mathbf{a})}{\partial \mathbf{a}_i \partial \mathbf{a}_j}\mathbf{F}(\mathbf{a})^+ + \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\frac{\partial \mathbf{F}(\mathbf{a})}{\partial \mathbf{a}_i}\frac{\partial \mathbf{F}(\mathbf{a})^+}{\partial \mathbf{a}_j}\ .$$

Next, we recall that $\mathbf{F}(.)$ is a linear transformation from $\mathbb{R}^{k.p}$ into $\mathbb{R}^{n.p\times n.k}$ (see equation (3.20) in Subsection 3.4). This implies, in particular, that $D\big(\mathbf{F}(\mathbf{a})\big) = \mathbf{F}(.)$, for all $\mathbf{a} \in \mathbb{R}^{k.p}$, meaning that $D\big(\mathbf{F}(.)\big)$ is a constant function from $\mathbb{R}^{k.p}$ into $\pounds(\mathbb{R}^{k.p}, \mathbb{R}^{n.p\times n.k})$. From this, we deduce that $D^2\big(\mathbf{F}(\mathbf{a})\big)$ is the null application for every $\mathbf{a} \in \mathbb{R}^{k.p}$, which implies that all the second partial derivatives $\partial^2\mathbf{F}(\mathbf{a})/\partial\mathbf{a}_i\partial\mathbf{a}_j$ are identically zero.

This may be used to simplify $\mathbf{D}_{(i,j)}$ again, giving

$$\mathbf{D}_{(i,j)} = -\frac{\partial\mathbf{P}_{\mathbf{F}(\mathbf{a})}}{\partial\mathbf{a}_j}\frac{\partial\mathbf{F}(\mathbf{a})}{\partial\mathbf{a}_i}\mathbf{F}(\mathbf{a})^+ + \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\frac{\partial\mathbf{F}(\mathbf{a})}{\partial\mathbf{a}_i}\frac{\partial\mathbf{F}(\mathbf{a})^+}{\partial\mathbf{a}_j} \ .$$

Now, we need to compute the derivative of $\mathbf{F}(\mathbf{a})^+$ which is given in Theorem 5.1 under the assumption that $\mathbf{F}(\mathbf{a})$ is of local constant rank at any point $\mathbf{a}$ in which differentiation is to be performed (see Theorem 5.1 for details):

$$\begin{aligned} D(\mathbf{F}(\mathbf{a})^+) = &- \mathbf{F}(\mathbf{a})^+D(\mathbf{F}(\mathbf{a}))\mathbf{F}(\mathbf{a})^+ + \mathbf{F}(\mathbf{a})^+(\mathbf{F}(\mathbf{a})^+)^TD(\mathbf{F}(\mathbf{a})^T)\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})} \\ &+ \big(\mathbf{I}_{n.k} - \mathbf{F}(\mathbf{a})^+\mathbf{F}(\mathbf{a})\big)D(\mathbf{F}(\mathbf{a})^T)(\mathbf{F}(\mathbf{a})^+)^T\mathbf{F}(\mathbf{a})^+ \ . \end{aligned}$$

In order to simplify the derivation, we will now further assume that $\mathbf{F}(\mathbf{a})$ has full column-rank, leaving the general case for future research. For practical applications, this hypothesis implies that the rank of $\mathbf{F}(\mathbf{a})$ is $k.n$ and that each column of $\mathbf{X}$ must have at least k "nonmissing" elements. In other words, using the $p \times n$ incidence matrix, $\boldsymbol{\delta}$, associated with the matrix $\mathbf{X}$ (see equation (5.26) in Subsection 5.2), we must have

$$\sum_{l=1}^{p} \boldsymbol{\delta}_{lj} > k \quad \text{, for all } j = 1, \cdots, n \ .$$

Note also that this hypothesis is automatically verified if $\mathbf{W} \in \mathbb{R}^{p\times n}_{+*}$ and $\mathbf{A}$ is of full column rank since

$$\mathbf{F}(\mathbf{a}) = \bigoplus_{j=1}^{n} \mathbf{F}_j(\mathbf{a}) \text{ where } \mathbf{F}_j(\mathbf{a}) = \mathit{diag}(\sqrt{\mathbf{W}}_{.j})\mathbf{A} \ .$$

From this hypothesis, we deduce that

$$\mathbf{F}(\mathbf{a})^+\mathbf{F}(\mathbf{a}) = \mathbf{I}_{n.k} \ ,$$

yielding a simplified formula for the derivative of $\mathbf{F}(\mathbf{a})^+$

$$D(\mathbf{F}(\mathbf{a})^+) = -\mathbf{F}(\mathbf{a})^+D(\mathbf{F}(\mathbf{a}))\mathbf{F}(\mathbf{a})^+ + \mathbf{F}(\mathbf{a})^+(\mathbf{F}(\mathbf{a})^+)^TD(\mathbf{F}(\mathbf{a})^T)\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}$$

and its partial derivatives

$$\frac{\partial\mathbf{F}(\mathbf{a})^+}{\partial\mathbf{a}_j} = -\mathbf{F}(\mathbf{a})^+\frac{\partial\mathbf{F}(\mathbf{a})}{\partial\mathbf{a}_j}\mathbf{F}(\mathbf{a})^+ + \mathbf{F}(\mathbf{a})^+(\mathbf{F}(\mathbf{a})^+)^T\frac{\partial\mathbf{F}(\mathbf{a})^T}{\partial\mathbf{a}_j}\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})} \ .$$

Substituting this definition into $\mathbf{D}_{(i,j)}$ gives

$$\begin{aligned} \mathbf{D}_{(i,j)} = &- \frac{\partial\mathbf{P}_{\mathbf{F}(\mathbf{a})}}{\partial\mathbf{a}_j}\frac{\partial\mathbf{F}(\mathbf{a})}{\partial\mathbf{a}_i}\mathbf{F}(\mathbf{a})^+ \\ &+ \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\frac{\partial\mathbf{F}(\mathbf{a})}{\partial\mathbf{a}_i}\Big(-\mathbf{F}(\mathbf{a})^+\frac{\partial\mathbf{F}(\mathbf{a})}{\partial\mathbf{a}_j}\mathbf{F}(\mathbf{a})^+ + \mathbf{F}(\mathbf{a})^+(\mathbf{F}(\mathbf{a})^+)^T\frac{\partial\mathbf{F}(\mathbf{a})^T}{\partial\mathbf{a}_j}\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\Big) \\ = &- \frac{\partial\mathbf{P}_{\mathbf{F}(\mathbf{a})}}{\partial\mathbf{a}_j}\frac{\partial\mathbf{F}(\mathbf{a})}{\partial\mathbf{a}_i}\mathbf{F}(\mathbf{a})^+ \\ &+ \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\frac{\partial\mathbf{F}(\mathbf{a})}{\partial\mathbf{a}_i}\mathbf{F}(\mathbf{a})^+\Big((\mathbf{F}(\mathbf{a})^+)^T\frac{\partial\mathbf{F}(\mathbf{a})^T}{\partial\mathbf{a}_j}\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})} - \frac{\partial\mathbf{F}(\mathbf{a})}{\partial\mathbf{a}_j}\mathbf{F}(\mathbf{a})^+\Big) \ . \end{aligned}$$

Using this result, we are now in the position to deduce an explicit formula for $\mathbf{S}_{ij}$

$$\begin{aligned}\mathbf{S}_{ij} &= -\mathbf{r}(\mathbf{a})^T\big(\mathbf{D}_{(i,j)} + \mathbf{D}_{(i,j)}^T\big)\mathbf{x} \\ &= -\mathbf{r}(\mathbf{a})^T\mathbf{D}_{(i,j)}\mathbf{x} - \mathbf{r}(\mathbf{a})^T\mathbf{D}_{(i,j)}^T\mathbf{x} \\ &= -\mathbf{r}(\mathbf{a})^T\mathbf{D}_{(i,j)}\mathbf{x} - \mathbf{x}^T\mathbf{D}_{(i,j)}\mathbf{r}(\mathbf{a})\ .\end{aligned}$$

Using the facts that $\mathbf{r}(\mathbf{a}) = \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\mathbf{x}$, $\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\mathbf{r}(\mathbf{a}) = \mathbf{r}(\mathbf{a})$ and $\mathbf{b} = \mathbf{F}(\mathbf{a})^+\mathbf{x}$, the first term in the right hand side of this equation reduces to

$$\begin{aligned}-\mathbf{r}(\mathbf{a})^T\mathbf{D}_{(i,j)}\mathbf{x} &= \mathbf{r}(\mathbf{a})^T\frac{\partial\mathbf{P}_{\mathbf{F}(\mathbf{a})}}{\partial\mathbf{a}_j}\frac{\partial\mathbf{F}(\mathbf{a})}{\partial\mathbf{a}_i}\mathbf{b} \\ &\quad - \ \mathbf{r}(\mathbf{a})^T\frac{\partial\mathbf{F}(\mathbf{a})}{\partial\mathbf{a}_i}\mathbf{F}(\mathbf{a})^+\Big((\mathbf{F}(\mathbf{a})^+)^T\frac{\partial\mathbf{F}(\mathbf{a})^T}{\partial\mathbf{a}_j}\mathbf{r}(\mathbf{a}) - \frac{\partial\mathbf{F}(\mathbf{a})}{\partial\mathbf{a}_j}\mathbf{b}\Big) \\ &= \mathbf{r}(\mathbf{a})^T\frac{\partial\mathbf{P}_{\mathbf{F}(\mathbf{a})}}{\partial\mathbf{a}_j}\frac{\partial\mathbf{F}(\mathbf{a})}{\partial\mathbf{a}_i}\mathbf{b} - \mathbf{r}(\mathbf{a})^T\frac{\partial\mathbf{F}(\mathbf{a})}{\partial\mathbf{a}_i}\mathbf{F}(\mathbf{a})^+(\mathbf{F}(\mathbf{a})^+)^T\frac{\partial\mathbf{F}(\mathbf{a})^T}{\partial\mathbf{a}_j}\mathbf{r}(\mathbf{a}) \\ &\quad + \ \mathbf{r}(\mathbf{a})^T\frac{\partial\mathbf{F}(\mathbf{a})}{\partial\mathbf{a}_i}\mathbf{F}(\mathbf{a})^+\frac{\partial\mathbf{F}(\mathbf{a})}{\partial\mathbf{a}_j}\mathbf{b}\ .\end{aligned}$$

Substituting $\partial\mathbf{P}_{\mathbf{F}(\mathbf{a})}/\partial\mathbf{a}_j$ (see equation (5.31) above) yields

$$\begin{aligned}-\mathbf{r}(\mathbf{a})^T\mathbf{D}_{(i,j)}\mathbf{x} &= \mathbf{r}(\mathbf{a})^T\Big(\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\frac{\partial\mathbf{F}(\mathbf{a})}{\partial\mathbf{a}_j}\mathbf{F}(\mathbf{a})^+ + \big(\mathbf{F}(\mathbf{a})^+\big)^T\frac{\partial\mathbf{F}(\mathbf{a})^T}{\partial\mathbf{a}_j}\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\Big)\frac{\partial\mathbf{F}(\mathbf{a})}{\partial\mathbf{a}_i}\mathbf{b} \\ &\quad - \ \mathbf{r}(\mathbf{a})^T\frac{\partial\mathbf{F}(\mathbf{a})}{\partial\mathbf{a}_i}\mathbf{F}(\mathbf{a})^+(\mathbf{F}(\mathbf{a})^+)^T\frac{\partial\mathbf{F}(\mathbf{a})^T}{\partial\mathbf{a}_j}\mathbf{r}(\mathbf{a}) \\ &\quad + \ \mathbf{r}(\mathbf{a})^T\frac{\partial\mathbf{F}(\mathbf{a})}{\partial\mathbf{a}_i}\mathbf{F}(\mathbf{a})^+\frac{\partial\mathbf{F}(\mathbf{a})}{\partial\mathbf{a}_j}\mathbf{b}\ .\end{aligned}$$

Noting that $\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\mathbf{r}(\mathbf{a}) = \mathbf{r}(\mathbf{a})$ and $\mathbf{F}(\mathbf{a})^+\mathbf{r}(\mathbf{a}) = \mathbf{0}^{k.n}$, this reduces to

$$\begin{aligned}-\mathbf{r}(\mathbf{a})^T\mathbf{D}_{(i,j)}\mathbf{x} &= \mathbf{r}(\mathbf{a})^T\frac{\partial\mathbf{F}(\mathbf{a})}{\partial\mathbf{a}_j}\mathbf{F}(\mathbf{a})^+\frac{\partial\mathbf{F}(\mathbf{a})}{\partial\mathbf{a}_i}\mathbf{b} \\ &\quad - \ \mathbf{r}(\mathbf{a})^T\frac{\partial\mathbf{F}(\mathbf{a})}{\partial\mathbf{a}_i}\mathbf{F}(\mathbf{a})^+(\mathbf{F}(\mathbf{a})^+)^T\frac{\partial\mathbf{F}(\mathbf{a})^T}{\partial\mathbf{a}_j}\mathbf{r}(\mathbf{a}) \\ &\quad + \ \mathbf{r}(\mathbf{a})^T\frac{\partial\mathbf{F}(\mathbf{a})}{\partial\mathbf{a}_i}\mathbf{F}(\mathbf{a})^+\frac{\partial\mathbf{F}(\mathbf{a})}{\partial\mathbf{a}_j}\mathbf{b}\ .\end{aligned}$$

Next, if we consider the second term in $\mathbf{S}_{ij}$, it reduces to

$$\begin{aligned}-\mathbf{x}^T\mathbf{D}_{(i,j)}\mathbf{r}(\mathbf{a}) &= \mathbf{x}^T\frac{\partial\mathbf{P}_{\mathbf{F}(\mathbf{a})}}{\partial\mathbf{a}_j}\frac{\partial\mathbf{F}(\mathbf{a})}{\partial\mathbf{a}_i}\mathbf{F}(\mathbf{a})^+\mathbf{r}(\mathbf{a}) \\ &\quad - \ \mathbf{x}^T\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\frac{\partial\mathbf{F}(\mathbf{a})}{\partial\mathbf{a}_i}\mathbf{F}(\mathbf{a})^+(\mathbf{F}(\mathbf{a})^+)^T\frac{\partial\mathbf{F}(\mathbf{a})^T}{\partial\mathbf{a}_j}\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\mathbf{r}(\mathbf{a}) \\ &\quad + \ \mathbf{x}^T\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\frac{\partial\mathbf{F}(\mathbf{a})}{\partial\mathbf{a}_i}\mathbf{F}(\mathbf{a})^+\frac{\partial\mathbf{F}(\mathbf{a})}{\partial\mathbf{a}_j}\mathbf{F}(\mathbf{a})^+\mathbf{r}(\mathbf{a}) \\ &= \ -\mathbf{r}(\mathbf{a})^T\frac{\partial\mathbf{F}(\mathbf{a})}{\partial\mathbf{a}_i}\mathbf{F}(\mathbf{a})^+(\mathbf{F}(\mathbf{a})^+)^T\frac{\partial\mathbf{F}(\mathbf{a})^T}{\partial\mathbf{a}_j}\mathbf{r}(\mathbf{a})\ ,\end{aligned}$$

using again the facts that $\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\mathbf{r}(\mathbf{a}) = \mathbf{r}(\mathbf{a})$, $\mathbf{F}(\mathbf{a})^+\mathbf{r}(\mathbf{a}) = \mathbf{0}^{k.n}$ and $\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})} = (\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})})^T$.

Collecting all these results together, we deduce that

$$\begin{aligned}\mathbf{S}_{ij} = \mathbf{r}(\mathbf{a})^T \frac{\partial \mathbf{F}(\mathbf{a})}{\partial \mathbf{a}_j} \mathbf{F}(\mathbf{a})^+ \frac{\partial \mathbf{F}(\mathbf{a})}{\partial \mathbf{a}_i} \mathbf{b} \\ + \ \mathbf{r}(\mathbf{a})^T \frac{\partial \mathbf{F}(\mathbf{a})}{\partial \mathbf{a}_i} \mathbf{F}(\mathbf{a})^+ \frac{\partial \mathbf{F}(\mathbf{a})}{\partial \mathbf{a}_j} \mathbf{b} \\ - \ 2\mathbf{r}(\mathbf{a})^T \frac{\partial \mathbf{F}(\mathbf{a})}{\partial \mathbf{a}_i} \mathbf{F}(\mathbf{a})^+ (\mathbf{F}(\mathbf{a})^+)^T \frac{\partial \mathbf{F}(\mathbf{a})^T}{\partial \mathbf{a}_j} \mathbf{r}(\mathbf{a}) \ .\end{aligned}$$

Now from equations (5.14) and (5.18) in Subsection 5.2, $\forall\ \mathbf{a}, \triangle\mathbf{a} \in \mathbb{R}^{p.k}$, we have

$$\big(D(\mathbf{F}(\mathbf{a}))(\triangle\mathbf{a})\big)\mathbf{b} = \mathbf{U}(\mathbf{a})\triangle\mathbf{a} \quad \text{and} \quad \big(D(\mathbf{F}(\mathbf{a}))(\triangle\mathbf{a})\big)^T\mathbf{r}(\mathbf{a}) = \mathbf{V}(\mathbf{a})\triangle\mathbf{a} \ ,$$

where $\mathbf{U}(\mathbf{a})$ and $\mathbf{V}(\mathbf{a})$ are defined in equations (5.20) and (5.21) of Subsection 5.2. In these conditions, it is evident that

$$\frac{\partial \mathbf{F}(\mathbf{a})}{\partial \mathbf{a}_i}\mathbf{b} = \mathbf{U}(\mathbf{a})_{.i} \quad \text{and} \quad \frac{\partial \mathbf{F}(\mathbf{a})^T}{\partial \mathbf{a}_i}\mathbf{r}(\mathbf{a}) = \mathbf{V}(\mathbf{a})_{.i} \quad \text{for all } i = 1, \cdots, p.k \ .$$

Thus,

$$\begin{aligned}\mathbf{S}_{ij} &= \big(\mathbf{V}(\mathbf{a})_{.j}\big)^T\mathbf{F}(\mathbf{a})^+\mathbf{U}(\mathbf{a})_{.i} + \big(\mathbf{V}(\mathbf{a})_{.i}\big)^T\mathbf{F}(\mathbf{a})^+\mathbf{U}(\mathbf{a})_{.j} - 2.\big(\mathbf{V}(\mathbf{a})_{.i}\big)^T\mathbf{F}\big(\mathbf{a}\big)^+\big(\mathbf{F}(\mathbf{a})^+\big)^T\mathbf{V}(\mathbf{a})_{.j} \\ &= \big(\mathbf{L}(\mathbf{a})_{.j}\big)^T\mathbf{U}(\mathbf{a})_{.i} + \big(\mathbf{L}(\mathbf{a})_{.i}\big)^T\mathbf{U}(\mathbf{a})_{.j} - 2.\big(\mathbf{L}(\mathbf{a})_{.i}\big)^T\mathbf{L}(\mathbf{a})_{.j} \\ &= \big(\mathbf{U}(\mathbf{a})_{.i}\big)^T\mathbf{L}(\mathbf{a})_{.j} + \big(\mathbf{L}(\mathbf{a})_{.i}\big)^T\mathbf{U}(\mathbf{a})_{.j} - 2.\big(\mathbf{L}(\mathbf{a})_{.i}\big)^T\mathbf{L}(\mathbf{a})_{.j} \ ,\end{aligned}$$

where $\mathbf{L}(\mathbf{a})$ is defined in equation (5.19) of Subsection 5.2. This implies, finally, that

$$\mathbf{S} = \mathbf{U}(\mathbf{a})^T\mathbf{L}(\mathbf{a}) + \mathbf{L}(\mathbf{a})^T\mathbf{U}(\mathbf{a}) - 2.\mathbf{L}(\mathbf{a})^T\mathbf{L}(\mathbf{a}) \ . \tag{5.32}$$

Using the fact that $J(\mathbf{r}(\mathbf{a}))^T J(\mathbf{r}(\mathbf{a})) = \mathbf{M}(\mathbf{a})^T\mathbf{M}(\mathbf{a}) + \mathbf{L}(\mathbf{a})^T\mathbf{L}(\mathbf{a})$ established in equation (5.29) above, we finally obtain the following explicit and compact expression for the Hessian matrix $\nabla^2\psi(\mathbf{a})$ as the sum of three symmetric matrix terms

$$\mathbf{H} = \mathbf{M}(\mathbf{a})^T\mathbf{M}(\mathbf{a}) - \mathbf{L}(\mathbf{a})^T\mathbf{L}(\mathbf{a}) + \big(\mathbf{U}(\mathbf{a})^T\mathbf{L}(\mathbf{a}) + \mathbf{L}(\mathbf{a})^T\mathbf{U}(\mathbf{a})\big) \ , \tag{5.33}$$

under the hypothesis that $\mathbf{F}(.)$ has full column-rank in a neighborhood of $\mathbf{a}$.

We first note that $\mathbf{H}$ is relatively cheap to evaluate as all the matrix terms involved in the above expression of $\mathbf{H}$ are also needed to compute the Jacobian matrix $J(\mathbf{r}(\mathbf{a}))$ and are, thus, available in most cases. This small difference between a full Newton and a Gauss-Newton approach in terms of computational load is due to the fact that all the second partial derivatives of $\mathbf{F}(.)$ are identically zero everywhere and that, consequently, all the mixed partial derivatives matrix terms, which are normally present in the Hessian matrix, vanish here [10]. This small overhead of a full Newton method for solving the WLRA problem with a variable projection or Grassmann manifold approaches has already been highlighted by Boumal and Absil [14] when solving a regularized version of the WLRA problem when zero weights are present with a Riemannian method.

Next, we observe that the Gauss-Newton approximation of the Hessian matrix is

$$\nabla^2\psi(\mathbf{a}) \approx \mathbf{M}(\mathbf{a})^T\mathbf{M}(\mathbf{a}) + \mathbf{L}(\mathbf{a})^T\mathbf{L}(\mathbf{a}) \ ,$$

while the first two symmetric terms in the exact formulation of $\nabla^2\psi(\mathbf{a})$ derived above are

$$\mathbf{M}(\mathbf{a})^T\mathbf{M}(\mathbf{a}) - \mathbf{L}(\mathbf{a})^T\mathbf{L}(\mathbf{a}) \ .$$

This suggests that the term $\mathbf{M}(\mathbf{a})^T\mathbf{M}(\mathbf{a})$ can be eventually a much better estimate of the full Hessian matrix $\nabla^2\psi(\mathbf{a})$ than its Gauss-Newton estimate surprisingly. This feature is not specific of

the WLRA problem, but rather related to the variable projection framework, and remains true for any separable NLLS problem [10]. This is also verified in the comparison experiments of Hong et al. [81] in which a Levenberg-Marquardt algorithm using $-\mathbf{M}(\mathbf{a})$ as a simplified Jacobian matrix performs equally or even better than the one using the full Jacobian matrix $-(\mathbf{M}(\mathbf{a})+\mathbf{L}(\mathbf{a}))$. As we will explain in Subsection 6.1, approximating the Jacobian matrix by the term $-\mathbf{M}(\mathbf{a})$ corresponds also to the simplification of the standard variable projection Gauss-Newton and Levenberg-Marquardt algorithms introduced by Kaufman [96] and Ruhe and Wedin [166]. These results were also given given in Hong and Fitzgibbon [81], but their notations are different from those used here.

Based on similar arguments, we can also develop an efficient quasi-Newton method in which we drop the last symmetric term, e.g., $\mathbf{U}(\mathbf{a})^T\mathbf{L}(\mathbf{a}) + \mathbf{L}(\mathbf{a})^T\mathbf{U}(\mathbf{a})$ in the full Hessian, and use the symmetric matrix

$$\bar{\mathbf{H}} = \mathbf{M}(\mathbf{a})^T\mathbf{M}(\mathbf{a}) - \mathbf{L}(\mathbf{a})^T\mathbf{L}(\mathbf{a}) \; , \tag{5.34}$$

as an approximate Hessian matrix. This approach is new and has never been proposed in the literature to the best of our knowledge. Obviously, we are not assure that this approximation is always positive semi-definite as in any Newton-like method, but this will be generally the rule as the term $\mathbf{M}(\mathbf{a})$ usually dominates the term $\mathbf{L}(\mathbf{a})$ in the Jacobian matrix as discussed above. However, an iterative algorithm based on this approximate Hessian $\bar{\mathbf{H}}$ should perform similarly or better than any Gauss-Newton- or Levenberg-Marquardt-like methods using the full Jacobian $-(\mathbf{M}(\mathbf{a})+\mathbf{L}(\mathbf{a}))$ or its approximation $-\mathbf{M}(\mathbf{a})$ as discussed above.

The approximate Hessian matrix $\bar{\mathbf{H}}$ also inherits of the singularity of the Jacobian matrix and of its two terms (see Theorems 5.2 and 5.3) since $\mathit{null}(J(\mathbf{r}(\mathbf{a}))) = \mathit{null}(\mathbf{M}(\mathbf{a})) \cap \mathit{null}(\mathbf{L}(\mathbf{a}))$ is always included in the null space of $\bar{\mathbf{H}}$ and this can be considered at first sight as a disadvantage. However, we will demonstrate below that a Newton method using the full Hessian matrix $\mathbf{H}$ has also to overcome similar singularity and ill-conditioning problems in a small neighborhood of a first-order stationary point of $\psi(.)$. Furthermore, we will finally illustrate that, in both cases, we can handle these difficulties for many practical cases with the same techniques developed for the Gauss-Newton or Levenberg-Marquardt algorithms in Subsection 5.2), e.g., by restricting the search directions for the Newton correction to $\mathit{ran}(\mathbf{A})^{\perp}$ at each iteration.

The results in Subsection 5.2 show that the problem of minimizing $\psi(.)$ is an ill-posed NLLS problem in the sense that the Jacobian matrix, $J(\mathbf{r}(\mathbf{a}))$, is exactly rank-deficient everywhere in the solution space $\mathbb{R}_k^{p\times k}$. We first derive some KKT theorems which give a characterization of a local minimum of $\psi(.)$ for this very special class of NLLS problems.

Theorem 5.9. Let $\widehat{\mathbf{a}} = \mathit{vec}(\widehat{\mathbf{A}})^T$, with $\widehat{\mathbf{A}} \in \mathbb{R}_k^{p\times k}$, be a first-order stationary point of $\psi(.)$ and $\mathbf{F}(.)$ has full column-rank in a neighborhood of $\widehat{\mathbf{a}}$. Then, the Hessian matrix of $\psi(.)$ at $\widehat{\mathbf{a}}$, $\widehat{\mathbf{H}} = \nabla^2\psi(\widehat{\mathbf{a}})$, is a matrix of rank less than or equal to $(p-k).k$ as for the Jacobian matrix $J(\mathbf{r}(\widehat{\mathbf{a}}))$ (see Corollary 5.3).

Proof. First, the hypothesis that $\mathbf{F}(\mathbf{a})$ has full column-rank in a neighborhood of $\widehat{\mathbf{a}}$ implies the existence of $\widehat{\mathbf{H}}$ as demonstrated above. Furthermore, from equation (5.33), $\widehat{\mathbf{H}}$ is given by

$$\widehat{\mathbf{H}} = \mathbf{M}(\widehat{\mathbf{a}})^T\mathbf{M}(\widehat{\mathbf{a}}) - \mathbf{L}(\widehat{\mathbf{a}})^T\mathbf{L}(\widehat{\mathbf{a}}) + \mathbf{U}(\widehat{\mathbf{a}})^T\mathbf{L}(\widehat{\mathbf{a}}) + \mathbf{L}(\widehat{\mathbf{a}})^T\mathbf{U}(\widehat{\mathbf{a}}) \; .$$

Now, as in Theorem 5.2, we consider the matrix

$$\widehat{\mathbf{N}} = \mathbf{K}_{(p,k)}(\mathbf{I}_k \otimes \widehat{\mathbf{A}}) \; .$$

From Theorem 5.2, we have $\mathit{rank}(\widehat{\mathbf{N}}) = k.k$ and the relation

$$\mathit{ran}(\widehat{\mathbf{N}}) \subset \mathit{null}(\mathbf{M}(\widehat{\mathbf{a}})) \cap \mathit{null}(\mathbf{L}(\widehat{\mathbf{a}})) \; .$$

We also observe that the theorem will be proved if the relation $\mathit{ran}(\widehat{\mathbf{N}}) \subset \mathit{null}(\widehat{\mathbf{H}})$ is verified.

To prove this inclusion, we first note that, if $\mathbf{y} \in \mathit{ran}(\widehat{\mathbf{N}})$, we have

$$\mathbf{M}(\widehat{\mathbf{a}})\mathbf{y} = \mathbf{L}(\widehat{\mathbf{a}})\mathbf{y} = \mathbf{0}^{n.p} \; .$$

From this, we deduce that

$$\begin{aligned}\widehat{\mathbf{H}}\mathbf{y} &= \mathbf{M}(\widehat{\mathbf{a}})^T\mathbf{M}(\widehat{\mathbf{a}})\mathbf{y} - \mathbf{L}(\widehat{\mathbf{a}})^T\mathbf{L}(\widehat{\mathbf{a}})\mathbf{y} + \mathbf{U}(\widehat{\mathbf{a}})^T\mathbf{L}(\widehat{\mathbf{a}})\mathbf{y} + \mathbf{L}(\widehat{\mathbf{a}})^T\mathbf{U}(\widehat{\mathbf{a}})\mathbf{y} \\ &= \mathbf{L}(\widehat{\mathbf{a}})^T\mathbf{U}(\widehat{\mathbf{a}})\mathbf{y} \; .\end{aligned}$$

Thus, to demonstrate that $\mathbf{y} \in \mathit{null}(\widehat{\mathbf{H}})$, it suffices to show that $\mathbf{L}(\widehat{\mathbf{a}})^T\mathbf{U}(\widehat{\mathbf{a}})\mathbf{y} = \mathbf{0}^{k.p}$.

If $\mathbf{y} \in \mathit{ran}(\widehat{\mathbf{N}})$, then $\exists \mathbf{Z} \in \mathbb{R}^{k\times k}$ such that

$$\mathbf{y} = \widehat{\mathbf{N}}\mathit{vec}(\mathbf{Z}) = \mathbf{K}_{(p,k)}(\mathbf{I}_k \otimes \widehat{\mathbf{A}})\mathit{vec}(\mathbf{Z}) \; .$$

Consequently, using the definition of $\mathbf{U}(\widehat{\mathbf{a}})$ (see Subsection 5.2), we have

$$\begin{aligned}\mathbf{U}(\widehat{\mathbf{a}})\mathbf{y} &= \mathit{diag}\big(\mathit{vec}(\sqrt{\mathbf{W}})\big)(\widehat{\mathbf{B}}^T \otimes \mathbf{I}_p)\mathbf{K}_{(k,p)}\mathbf{K}_{(p,k)}(\mathbf{I}_k \otimes \widehat{\mathbf{A}})\mathit{vec}(\mathbf{Z}) \\ &= \mathit{diag}\big(\mathit{vec}(\sqrt{\mathbf{W}})\big)(\widehat{\mathbf{B}}^T \otimes \mathbf{I}_p)(\mathbf{I}_k \otimes \widehat{\mathbf{A}})\mathit{vec}(\mathbf{Z}) \\ &= \mathit{diag}\big(\mathit{vec}(\sqrt{\mathbf{W}})\big)(\widehat{\mathbf{B}}^T \otimes \widehat{\mathbf{A}})\mathit{vec}(\mathbf{Z}) \\ &= \mathit{diag}\big(\mathit{vec}(\sqrt{\mathbf{W}})\big)\mathit{vec}(\widehat{\mathbf{A}}\mathbf{Z}\widehat{\mathbf{B}}) \\ &= \mathit{diag}\big(\mathit{vec}(\sqrt{\mathbf{W}})\big)(\mathbf{I}_n \otimes \widehat{\mathbf{A}})\mathit{vec}(\mathbf{Z}\widehat{\mathbf{B}}) \\ &= \mathbf{F}(\widehat{\mathbf{a}})\mathit{vec}(\mathbf{Z}\widehat{\mathbf{B}}) \; .\end{aligned}$$

Now, as in Subsection 5.2, using the projection operator, $P_\Omega(.)$, associated with the $p \times n$ weight matrix $\mathbf{W}$, we have

$$\big[P_\Omega(\mathbf{X} - \widehat{\mathbf{A}}\widehat{\mathbf{B}})\big]_{ij} = \begin{cases} \mathbf{X}_{ij} - \displaystyle\sum_{l=1}^{k} \widehat{\mathbf{A}}_{il}\widehat{\mathbf{B}}_{lj} & \text{if } \mathbf{W}_{ij} \neq 0, \\ 0 & \text{if } \mathbf{W}_{ij} = 0. \end{cases} ,$$

where $\widehat{\mathbf{b}} = \mathbf{F}(\widehat{\mathbf{a}})^+\mathbf{x}$ and $\widehat{\mathbf{B}} = \mathit{mat}(\widehat{\mathbf{b}})$, and the variable projection residual vector of $\mathbf{X}$ at $\widehat{\mathbf{A}}$ can be written as

$$\mathbf{r}(\widehat{\mathbf{a}}) = \mathit{vec}\big(\sqrt{\mathbf{W}} \odot P_\Omega(\mathbf{X} - \widehat{\mathbf{A}}\widehat{\mathbf{B}})\big) \; .$$

Thus, using the definition of $\mathbf{L}(\widehat{\mathbf{a}})$ (see equation (5.19) in Section 5.2) and the fact that $\mathbf{F}(\widehat{\mathbf{a}})^+\mathbf{F}(\widehat{\mathbf{a}}) = \mathbf{I}_{n.k}$, which results from the hypothesis that $\mathbf{F}(\widehat{\mathbf{a}})$ has full column-rank, we have

$$\begin{aligned}\mathbf{L}(\widehat{\mathbf{a}})^T\mathbf{U}(\widehat{\mathbf{a}})\mathbf{y} &= \big((\mathbf{W} \odot P_\Omega(\mathbf{X} - \widehat{\mathbf{A}}\widehat{\mathbf{B}})) \otimes \mathbf{I}_k\big)\mathbf{F}(\widehat{\mathbf{a}})^+\mathbf{F}(\widehat{\mathbf{a}})\mathit{vec}(\mathbf{Z}\widehat{\mathbf{B}}) \\ &= \big((\mathbf{W} \odot P_\Omega(\mathbf{X} - \widehat{\mathbf{A}}\widehat{\mathbf{B}})) \otimes \mathbf{I}_k\big)\mathit{vec}(\mathbf{Z}\widehat{\mathbf{B}}) \; ,\end{aligned}$$

Finally, this leads to the equalities

$$
\begin{aligned}
\widehat{\mathbf{L}}^T\mathbf{U}(\widehat{\mathbf{a}})\mathbf{y} &= vec\big(\mathbf{Z}\widehat{\mathbf{B}}(\mathbf{W}\odot P_\Omega(\mathbf{X}-\widehat{\mathbf{A}}\widehat{\mathbf{B}}))^T\big)\\
&= (\mathbf{I}_p\otimes\mathbf{Z})vec\big(\widehat{\mathbf{B}}(\mathbf{W}\odot P_\Omega(\mathbf{X}-\widehat{\mathbf{A}}\widehat{\mathbf{B}}))^T\big)\\
&= (\mathbf{I}_p\otimes\mathbf{Z})(\mathbf{I}_p\otimes\widehat{\mathbf{B}})vec\big((\mathbf{W}\odot P_\Omega(\mathbf{X}-\widehat{\mathbf{A}}\widehat{\mathbf{B}}))^T\big)\\
&= (\mathbf{I}_p\otimes\mathbf{Z})(\mathbf{I}_p\otimes\widehat{\mathbf{B}})diag\big(vec(\sqrt{\mathbf{W}}^T)\big)vec\big((\sqrt{\mathbf{W}}\odot P_\Omega(\mathbf{X}-\widehat{\mathbf{A}}\widehat{\mathbf{B}}))^T\big)\\
&= (\mathbf{I}_p\otimes\mathbf{Z})(\mathbf{I}_p\otimes\widehat{\mathbf{B}})diag\big(vec(\sqrt{\mathbf{W}}^T)\big)vec\big((\sqrt{\mathbf{W}}\odot P_\Omega(\mathbf{X}-\widehat{\mathbf{A}}\widehat{\mathbf{B}}))^T\big)\\
&= (\mathbf{I}_p\otimes\mathbf{Z})\Big(diag\big(vec(\sqrt{\mathbf{W}}^T)\big)(\mathbf{I}_p\otimes\widehat{\mathbf{B}})^T\Big)^T vec\big((\sqrt{\mathbf{W}}\odot P_\Omega(\mathbf{X}-\widehat{\mathbf{A}}\widehat{\mathbf{B}}))^T\big)\\
&= (\mathbf{I}_p\otimes\mathbf{Z})\Big(diag\big(vec(\sqrt{\mathbf{W}}^T)\big)(\mathbf{I}_p\otimes\widehat{\mathbf{B}}^T)\Big)^T vec\big((\sqrt{\mathbf{W}}\odot P_\Omega(\mathbf{X}-\widehat{\mathbf{A}}\widehat{\mathbf{B}}))^T\big)\\
&= (\mathbf{I}_p\otimes\mathbf{Z})\mathbf{G}(\widehat{\mathbf{b}})^T vec\big((\sqrt{\mathbf{W}}\odot P_\Omega(\mathbf{X}-\widehat{\mathbf{A}}\widehat{\mathbf{B}}))^T\big)\\
&= (\mathbf{I}_p\otimes\mathbf{Z})\mathbf{G}(\widehat{\mathbf{b}})^T\mathbf{K}_{(p,n)}vec(\sqrt{\mathbf{W}}\odot P_\Omega(\mathbf{X}-\widehat{\mathbf{A}}\widehat{\mathbf{B}}))\\
&= (\mathbf{I}_p\otimes\mathbf{Z})\mathbf{G}(\widehat{\mathbf{b}})^T\mathbf{K}_{(p,n)}\mathbf{r}(\widehat{\mathbf{a}})\\
&= -(\mathbf{I}_p\otimes\mathbf{Z})\nabla\psi(\widehat{\mathbf{a}})\\
&= \mathbf{0}^{k.p}\ ,
\end{aligned}
$$

the two last equalities resulting from Theorem 5.7 and the hypothesis $\nabla\psi(\widehat{\mathbf{a}}) = \mathbf{0}^{k.p}$. Thus, $\widehat{\mathbf{H}}\mathbf{y} = \mathbf{0}^{k.p}$ and $ran(\widehat{\mathbf{N}}) \subset null(\widehat{\mathbf{H}})$. This implies, finally, that $k.k \leqslant dim(null(\widehat{\mathbf{H}}))$ and $rank(\widehat{\mathbf{H}}) \leqslant (p-k).k$ as claimed in the theorem.

□

Furthermore, if $rank(J(\mathbf{r}(\widehat{\mathbf{a}})) = (p-k).k$, for example if the hypotheses of Theorem 5.3 are satisfied, we have the following corollary:

Corollary 5.9. Let $\widehat{\mathbf{a}} = vec(\widehat{\mathbf{A}})^T$, with $\widehat{\mathbf{A}} \in \mathbb{R}_k^{p\times k}$, be a first-order stationary point of $\psi(.)$ and suppose that $\mathbf{F}(\mathbf{a})$ has full column-rank in a neighborhood of $\widehat{\mathbf{a}}$ and the rank of the Jacobian matrix $J(\mathbf{r}(\widehat{\mathbf{a}}))$ is equal to $r = (p-k).k$. Then, the Hessian matrix of $\psi(.)$ at $\widehat{\mathbf{a}}$, $\widehat{\mathbf{H}} = \nabla^2\psi(\widehat{\mathbf{a}})$, is a matrix of rank less than or equal to r with its null space containing the null space of $J(\mathbf{r}(\widehat{a}))$.

Proof. Using the same notations as in Theorem 5.9, the hypotheses imply the relation

$$dim\Big(null\big(J(\mathbf{r}(\widehat{\mathbf{a}}))\big)\Big) = k.k = rank(\widehat{\mathbf{N}})\ .$$

From the relation $ran(\widehat{\mathbf{N}}) \subset null(\mathbf{M}(\widehat{\mathbf{a}})) \cap null(\mathbf{L}(\widehat{\mathbf{a}})) = null(J\big(\mathbf{r}(\widehat{\mathbf{a}})\big))$, we then deduce that $ran(\widehat{\mathbf{N}}) = null(J\big(\mathbf{r}(\widehat{\mathbf{a}})\big))$ and all the results are a direct consequence of Theorem 5.9.

□

Remark 5.3. Theorem 5.9 and Corollary 5.9 are similar to Theorem 2.2 of Eriksson et al. [54] and its consequences, which deal with NLLS problems that are uniformly rank-deficient, i.e., with a Jacobian matrix that have the same deficient rank in the neighborhood of a solution. However, it is important to highlight that the strong assumption of uniform rank deficiency of the Jacobian matrix in the neighborhood of a solution has not been used here to derive Theorem 5.9, while this hypothesis is central in the results of Eriksson et al. [54] through the use of the Constant-Rank Theorem [26]. ■

Remark 5.4. Theorem 5.9 and Corollary 5.9 also explain why full Newton variable projection algorithms (without any specific adaptation or regularization) perform poorly in the comparison experiments of Okatani et al. [150] and Hong et al. [81]. As soon as we are in a small neighborhood

of a first-order stationary point $\widehat{\mathbf{a}}$ of $\psi(.)$, the Hessian matrix $\nabla^2\psi(\mathbf{a})$ may become nearly singular and ill-conditioned according to Theorem 5.9 and Corollary 5.9 leading to an erratic behaviour and a dramatic lost of accuracy in the final steps of the Newton method. ■

Importantly, Theorem 5.9 and Corollary 5.9 can also be used to illustrate that the sufficient condition for the existence of a strict local minima of $\psi(.)$ (see Subsection 2.4 for details) are never meet, but non strict local minima may still exist though. To see this, suppose that $\psi(.)$ is twice continuously differentiable at $\widehat{\mathbf{a}}$ and that $\widehat{\mathbf{a}}$ is a first-order stationary point of $\psi(.)$, then $\nabla\psi(\widehat{\mathbf{a}}) = \mathbf{0}^{k.p}$ and the second-order Taylor expansion of $\psi(.)$ at $\widehat{\mathbf{a}}$ becomes

$$\psi(\widehat{\mathbf{a}} + d\mathbf{a}) = \psi(\widehat{\mathbf{a}}) + d\mathbf{a}^T\nabla^2\psi(\widehat{\mathbf{a}})d\mathbf{a} + \mathcal{O}(\|d\mathbf{a}\|_2^3) \; .$$

However, for $d\mathbf{a} \in \mathit{null}(\widehat{\mathbf{N}})$, where $\widehat{\mathbf{N}} = \mathbf{K}_{(p,k)}(\mathbf{I}_k \otimes \widehat{\mathbf{A}})$, and $d\mathbf{a}$ sufficiently small, this Taylor expansion reduces to

$$\psi(\widehat{\mathbf{a}} + d\mathbf{a}) = \psi(\widehat{\mathbf{a}}) + \mathcal{O}(\|d\mathbf{a}\|_2^3) \; ,$$

according to Theorem 5.9. This last equation and the fact that the matrix $\nabla^2\psi(\widehat{\mathbf{a}})$ is never positive definite, are consistent with the property that, if the (VP1) problem admits a solution $\widehat{\mathbf{a}}$, then there exist infinitely many solutions near $\widehat{\mathbf{a}}$ (see Remark 3.5 for details). Thus, having $J(\mathbf{r}(\mathbf{a}))$ rank deficient everywhere makes the minimization of $\psi(.)$ an ill-posed problem since if this problem has a solution $\widehat{\mathbf{a}}$ than this solution is never unique and, in addition, there are an infinite set of solutions in any neighborhood of $\widehat{\mathbf{a}}$.

To conclude this section, we now give sufficient second-order KKT conditions for $\psi(.)$ to have a (non strict) minimum under the exact and constant rank-deficient condition of $J(\mathbf{r}(\mathbf{a}))$ in a neighborhood of a first-order stationary point in the following theorem, which is a reformulation and an extension in our context of results demonstrated in Eriksson et al. [54].

Theorem 5.10. Let $\widehat{\mathbf{a}} \in \mathbb{R}^{k.p}$, with $\widehat{\mathbf{A}} = (\mathit{mat}_{k\times p}(\widehat{\mathbf{a}}))^T \in \mathbb{R}_k^{p\times k}$, be a first-order stationary point of $\psi(.)$, and suppose that $\mathbf{F}(\mathbf{a})$ has full column-rank and the rank of the Jacobian matrix $J(\mathbf{r}(\mathbf{a}))$ is equal to $r = (p-k).k$ in a neighborhood of $\widehat{\mathbf{a}}$. Let further $\mathbf{Q} \in \mathbb{R}_r^{k.p\times r}$ such that the columns of $\mathbf{Q}$ form a basis for $\mathit{ran}(J(\mathbf{r}(\widehat{\mathbf{a}}))^T) = \mathit{null}(J(\mathbf{r}(\widehat{\mathbf{a}})))^\perp$. Then, define the $r \times r$ symmetric matrix

$$\begin{aligned}\widehat{\mathbf{T}} &= \mathbf{Q}^T\nabla^2\psi(\widehat{\mathbf{a}})\mathbf{Q} \\ &= \mathbf{Q}^T\widehat{\mathbf{H}}\mathbf{Q} \\ &= \mathbf{Q}^T\left(\mathbf{M}(\widehat{\mathbf{a}})^T\mathbf{M}(\widehat{\mathbf{a}}) - \mathbf{L}(\widehat{\mathbf{a}})^T\mathbf{L}(\widehat{\mathbf{a}}) + \mathbf{U}(\widehat{\mathbf{a}})^T\mathbf{L}(\widehat{\mathbf{a}}) + \mathbf{L}(\widehat{\mathbf{a}})^T\mathbf{U}(\widehat{\mathbf{a}})\right)\mathbf{Q} \; . \end{aligned} \tag{5.35}$$

In these conditions, if $\psi(.)$ has a local minimum at $\widehat{\mathbf{a}}$ then $\widehat{\mathbf{T}}$ is positive semi-definite. Reciprocally, if $\widehat{\mathbf{T}}$ is positive definite then $\psi(.)$ has a (non strict) local minimizer at $\widehat{\mathbf{a}}$ and we have the equalities

$$\mathit{null}\big(\nabla^2\psi(\widehat{\mathbf{a}})\big) = \mathit{null}\big(J(\mathbf{r}(\widehat{\mathbf{a}}))\big) \quad \text{and} \quad \mathit{rank}\big(\nabla^2\psi(\widehat{\mathbf{a}})\big) = r \; .$$

Proof. We first verify that the condition is necessary if $\widehat{\mathbf{a}} \in \mathbb{R}^{k.p}$ is a local minimizer of $\psi(.)$. To this end, let $\mathbf{P} \in \mathbb{R}_{k.k}^{k.p\times k.k}$ such that the columns of $\mathbf{P}$ form a basis for $\mathit{null}(J(\mathbf{r}(\widehat{\mathbf{a}})))$. We first note that orthonormal matrices $\mathbf{P}$ and $\mathbf{Q}$ can be easily constructed from the results of Corollary 5.6 and that, with this choice, the columns of the partitioned matrix

$$\begin{bmatrix}\mathbf{P} & \mathbf{Q}\end{bmatrix} = \begin{bmatrix}\bar{\mathbf{O}} & \bar{\mathbf{O}}^\perp\end{bmatrix}$$

is an orthonormal basis of $\mathbb{R}^{k.p}$ since $\mathit{null}(J(\mathbf{r}(\widehat{\mathbf{a}})))$ and $\mathit{null}(J(\mathbf{r}(\widehat{\mathbf{a}})))^\perp$ are orthogonal subspaces of $\mathbb{R}^{k.p}$ and

$$\mathbb{R}^{k.p} = \mathit{null}\Big(J\big(\mathbf{r}(\widehat{\mathbf{a}})\big)\Big) \oplus \mathit{null}\Big(J\big(\mathbf{r}(\widehat{\mathbf{a}})\big)\Big)^\perp \; ,$$

where $\oplus$ stands for the direct sum. Thus, $\forall \mathbf{y} \in \mathbb{R}^{k.p}$, $\mathbf{y}$ can be decomposed as

$$\mathbf{y} = \begin{bmatrix}\mathbf{P} & \mathbf{Q}\end{bmatrix}\begin{bmatrix}\mathbf{y}_P \\ \mathbf{y}_Q\end{bmatrix} = \mathbf{P}\mathbf{y}_P + \mathbf{Q}\mathbf{y}_Q \text{ with } \mathbf{y}_P \in \mathbb{R}^{k.k} \text{ and } \mathbf{y}_Q \in \mathbb{R}^{(p-k).k} \; .$$

Now, as $\widehat{\mathbf{a}}$ is a local minimizer of $\psi(.)$, $\nabla^2\psi(\widehat{\mathbf{a}})$ is a positive semi-definite matrix (see Subsection2.4), which implies that, $\forall \mathbf{y} \in \mathbb{R}^{k.p}$, we have

$$\mathbf{y}^T\nabla^2\psi(\widehat{\mathbf{a}})\mathbf{y} \geq 0 \; .$$

On the other hand, since the null space of $J(\mathbf{r}(\widehat{a}))$ is included in the null space of $\nabla^2\psi(\widehat{\mathbf{a}})$ according to Corollary 5.9 and taking into account the symmetry of $\nabla^2\psi(\widehat{\mathbf{a}})$, we have

$$\begin{aligned}\mathbf{y}^T\nabla^2\psi(\widehat{\mathbf{a}})\mathbf{y} &= \left(\mathbf{P}\mathbf{y}_P + \mathbf{Q}\mathbf{y}_Q\right)^T \nabla^2\psi(\widehat{\mathbf{a}}) \left(\mathbf{P}\mathbf{y}_P + \mathbf{Q}\mathbf{y}_Q\right) \\ &= \left(\mathbf{Q}\mathbf{y}_Q\right)^T \nabla^2\psi(\widehat{\mathbf{a}}) \left(\mathbf{Q}\mathbf{y}_Q\right) \\ &= \mathbf{y}_Q^T\widehat{\mathbf{T}}\mathbf{y}_Q \geq 0 \; ,\end{aligned}$$

which demonstrates that $\widehat{\mathbf{T}}$ is a positive semi-definite matrix as claimed in the theorem.

We now give a sketch of the proof for the sufficient condition and we refer the interested reader to Eriksson et al. [52][53][54] for more details. First, the hypothesis that $\widehat{\mathbf{A}}$ is of full column rank implies that it exists an open neighborhood of $\widehat{\mathbf{a}}$, say Υ, in which all its elements have also full-column rank, as $\mathbb{R}_k^{p\times k}$ is open in $\mathbb{R}^{p\times k}$ according to Theorem 2.3. By eventually restricting this open neighborhood Υ, the hypotheses that $\mathbf{F}(\mathbf{a})$ has full column-rank and that the rank of the Jacobian matrix $J(\mathbf{r}(\mathbf{a}))$ is equal to $r = (p-k).k$ in a neighborhood of $\widehat{\mathbf{a}}$ imply that the residual function $\mathbf{r}(.)$ from $\mathbb{R}^{k.p}$ to $\mathbb{R}^{p.n}$ is infinitely differentiable in an open neighborhood Υ of $\widehat{\mathbf{a}}$. This allows us to apply the Constant-Rank Theorem (see Theorem 2.1 in [54] or [26]) to demonstrate that there exist two functions, e.g.,

$$\mathbf{z} : \mathbb{R}^{k.p} \longrightarrow \mathbb{R}^r \quad \text{and} \quad \mathbf{h} : \mathbb{R}^r \longrightarrow \mathbb{R}^{n.p} \; ,$$

which are twice continuously differentiable functions, respectively, over an open neighborhood Υ of $\widehat{\mathbf{a}}$ and over $\mathbb{R}^r$, such that $\mathbf{r}(\mathbf{a}) = \mathbf{h}(\mathbf{z}(\mathbf{a}))$.

Using the chain rule for computing derivatives [26], we get, for all $\mathbf{a} \in \Upsilon$, the relation

$$J\big(\mathbf{r}(\mathbf{a})\big) = J\big(\mathbf{h}(\mathbf{z}(\mathbf{a}))\big)J\big(\mathbf{z}(\mathbf{a})\big) \; . \tag{5.36}$$

In these conditions, using the hypothesis that the rank of the Jacobian matrix $J(\mathbf{r}(\mathbf{a}))$ is equal to $r = (p-k).k, \forall \mathbf{a} \in \Upsilon$, and equation (2.2) in Subsection 2.1, we deduce immediately that the Jacobian matrices, $J(\mathbf{z}(\mathbf{a}))$ and $J(\mathbf{h}(\mathbf{z}(\mathbf{a})))$, have also a constant (full) rank equals to $r = (p-k).k$ for all $\mathbf{a} \in \Upsilon$. Furthermore, since both $J(\mathbf{h}(\mathbf{z}(\mathbf{a})))$ and $J(\mathbf{z}(\mathbf{a}))$ have full rank r, which is also equal to the rank of $J(\mathbf{r}(\mathbf{a}))$, we have the equalities

$$\mathit{null}\big(J(\mathbf{r}(\mathbf{a}))\big) = \mathit{null}\big(J(\mathbf{z}(\mathbf{a}))\big) \quad \text{and} \quad \mathit{ran}\big(J(\mathbf{r}(\mathbf{a}))\big) = \mathit{ran}\big(J(\mathbf{h}(\mathbf{z}(\mathbf{a})))\big) \; ,$$

and also the "transposed" versions of these equalities

$$\mathit{null}\big(J(\mathbf{r}(\mathbf{a}))^T\big) = \mathit{null}\big(J(\mathbf{h}(\mathbf{z}(\mathbf{a})))^T\big) \quad \text{and} \quad \mathit{ran}\big(J(\mathbf{r}(\mathbf{a}))^T\big) = \mathit{ran}\big(J(\mathbf{z}(\mathbf{a}))^T\big) \; ,$$

for all $\mathbf{a} \in \Upsilon$.

Now, the equality $\mathit{null}\left(J(\mathbf{r}(\widehat{\mathbf{a}}))^T\right) = \mathit{null}\left(J(\mathbf{h}(\mathbf{z}(\widehat{\mathbf{a}})))^T\right)$ and the hypothesis that $\widehat{\mathbf{a}}$ is a first-order stationary point of $\psi(.)$, i.e.,

$$\nabla\psi(\widehat{\mathbf{a}}) = J\big(\mathbf{r}(\widehat{\mathbf{a}})\big)^T\mathbf{r}(\widehat{\mathbf{a}}) = \mathbf{0}^{p.k} \; ,$$

imply that

$$J\big(\mathbf{h}(\mathbf{z}(\widehat{\mathbf{a}}))\big)^T\mathbf{h}(\mathbf{z}(\widehat{\mathbf{a}})) = J\big(\mathbf{h}(\mathbf{z}(\widehat{\mathbf{a}}))\big)^T\mathbf{r}(\widehat{\mathbf{a}}) = \mathbf{0}^{p.k} \; .$$

This demonstrates that $\mathbf{z}(\widehat{\mathbf{a}})$ is a first-order stationary point of the twice continuously differentiable real function $\phi(.)$ defined by

$$\phi : \mathbb{R}^r \longrightarrow \mathbb{R} : \mathbf{o} \mapsto \frac{1}{2}\|\mathbf{h}(\mathbf{o})\|_2^2 = \frac{1}{2}\mathbf{h}(\mathbf{o})^T\mathbf{h}(\mathbf{o}) \; .$$

In these conditions, a sufficient condition for the first-order stationary point $\mathbf{z}(\widehat{\mathbf{a}})$ to be a strict local minimizer of $\phi(.)$ is that the Hessian matrix $\nabla^2\phi(\mathbf{z}(\widehat{\mathbf{a}}))$ is positive definite (see Subsection 2.4) and this will also imply that $\widehat{\mathbf{a}}$ is a non strict local minimizer of $\psi(.)$ since

$$\psi(\mathbf{a}) = \frac{1}{2}\mathbf{r}(\mathbf{a})^T\mathbf{r}(\mathbf{a}) = \frac{1}{2}\mathbf{h}\big(\mathbf{z}(\mathbf{a})\big)^T\mathbf{h}\big(\mathbf{z}(\mathbf{a})\big) = \phi\big(\mathbf{z}(\mathbf{a})\big)\ ,$$

for all $\mathbf{a} \in \Upsilon$. The Hessian matrix $\nabla^2\phi(\mathbf{z}(\widehat{\mathbf{a}}))$ is positive definite if and only if

$$\mathbf{o}^T\left(\nabla^2\phi(\mathbf{z}(\widehat{\mathbf{a}}))\right)\mathbf{o} > 0 \text{ for all } \mathbf{o} \in \mathbb{R}^r \text{ with } \mathbf{o} \neq \mathbf{0}^r\ .$$

Now, the equality $\mathit{null}\,(J(\mathbf{r}(\widehat{\mathbf{a}}))) = \mathit{null}\,(J(\mathbf{z}(\widehat{\mathbf{a}})))$ demonstrated above implies that

$$\mathit{null}\big(J(\mathbf{r}(\widehat{\mathbf{a}}))\big)^\perp = \mathit{null}\big(J(\mathbf{z}(\widehat{\mathbf{a}}))\big)^\perp\ ,$$

and this shows that the $r \times r$ matrix $J(\mathbf{z}(\widehat{\mathbf{a}}))\mathbf{Q}$ is of full rank r as the r columns of $\mathbf{Q}$ form a basis of $\mathit{null}\,(J(\mathbf{z}(\widehat{\mathbf{a}})))^\perp$. In these conditions, the columns of $J(\mathbf{z}(\widehat{\mathbf{a}}))\mathbf{Q}$ form a basis of $\mathbb{R}^r$ and any $\mathbf{o} \in \mathbb{R}^r$ can be written as $\mathbf{o} = J(\mathbf{z}(\widehat{\mathbf{a}}))\mathbf{Q}\mathbf{w}$ for some $\mathbf{w} \in \mathbb{R}^r$ and the proposition that $\nabla^2\phi(\mathbf{z}(\widehat{\mathbf{a}}))$ is positive definite, is equivalent to

$$\mathbf{w}^T\left(\mathbf{Q}^T J\big(\mathbf{z}(\widehat{\mathbf{a}})\big)^T\Big(\nabla^2\phi\big(\mathbf{z}(\widehat{\mathbf{a}})\big)\Big)J\big(\mathbf{z}(\widehat{\mathbf{a}})\big)\mathbf{Q}\right)\mathbf{w} > 0 \text{ for all } \mathbf{w} \in \mathbb{R}^r \text{ with } \mathbf{w} \neq \mathbf{0}^r\ ,$$

e.g., that the $r \times r$ symmetric matrix $\left(\mathbf{Q}^T J\big(\mathbf{z}(\widehat{\mathbf{a}})\big)^T\Big(\nabla^2\phi\big(\mathbf{z}(\widehat{\mathbf{a}})\big)\Big)J\big(\mathbf{z}(\widehat{\mathbf{a}})\big)\mathbf{Q}\right)$ is definite positive. We will now show that this symmetric matrix is nothing else than the $r \times r$ symmetric matrix $\widehat{\mathbf{T}}$, defined in equation (5.35), and which is assumed to be positive definite by hypothesis.

Using equations 2.66 and 5.36 above, we have

$$\begin{aligned}
&J\big(\mathbf{z}(\widehat{\mathbf{a}})\big)^T\Big(\nabla^2\phi(\mathbf{z}(\widehat{\mathbf{a}}))\Big)J\big(\mathbf{z}(\widehat{\mathbf{a}})\big)\\
&= J\big(\mathbf{z}(\widehat{\mathbf{a}})\big)^T\left(J(\mathbf{h}(\mathbf{z}(\widehat{\mathbf{a}})))^T J(\mathbf{h}(\mathbf{z}(\widehat{\mathbf{a}}))) + \sum_{l=1}^{n.p}\mathbf{h}_l(\mathbf{z}(\widehat{\mathbf{a}}))\nabla^2\mathbf{h}_l(\mathbf{z}(\widehat{\mathbf{a}}))\right)J\big(\mathbf{z}(\widehat{\mathbf{a}})\big)\\
&= J\big(\mathbf{r}(\widehat{\mathbf{a}})\big)^T J\big(\mathbf{r}(\widehat{\mathbf{a}})\big) + J\big(\mathbf{z}(\widehat{\mathbf{a}})\big)^T\left(\sum_{l=1}^{n.p}\mathbf{r}_l(\widehat{\mathbf{a}})\nabla^2\mathbf{h}_l(\mathbf{z}(\widehat{\mathbf{a}}))\right)J\big(\mathbf{z}(\widehat{\mathbf{a}})\big)\ .
\end{aligned}$$

Next, using again the chain rule for computing the second derivatives of $\mathbf{r}_l(.)$ and the fact that

$$J\big(\mathbf{h}(\mathbf{z}(\widehat{\mathbf{a}}))\big)^T\mathbf{r}(\widehat{\mathbf{a}}) = \mathbf{0}^{p.k}\ ,$$

demonstrated above, we obtain the following expression for the second term of the Hessian matrix $\widehat{\mathbf{H}} = \nabla^2\psi(\widehat{\mathbf{a}})$, defined in equations 5.30 and 5.32, in terms of the derivatives of $\mathbf{z}(.)$ and $\mathbf{h}(.)$:

$$\widehat{\mathbf{S}} = \sum_{l=1}^{n.p}\mathbf{r}_l(\widehat{\mathbf{a}})\nabla^2\mathbf{r}_l(\widehat{\mathbf{a}}) = J\big(\mathbf{z}(\widehat{\mathbf{a}})\big)^T\left(\sum_{l=1}^{n.p}\mathbf{r}_l(\widehat{\mathbf{a}})\nabla^2\mathbf{h}_l(\mathbf{z}(\widehat{\mathbf{a}}))\right)J\big(\mathbf{z}(\widehat{\mathbf{a}})\big)\ .$$

This implies the equality

$$J\big(\mathbf{z}(\widehat{\mathbf{a}})\big)^T\Big(\nabla^2\phi\big(\mathbf{z}(\widehat{\mathbf{a}})\big)\Big)J\big(\mathbf{z}(\widehat{\mathbf{a}})\big) = J\big(\mathbf{r}(\widehat{\mathbf{a}})\big)^T J\big(\mathbf{r}(\widehat{\mathbf{a}})\big) + \widehat{\mathbf{S}} = \widehat{\mathbf{H}}\ ,$$

from which we deduce that

$$\mathbf{Q}^T J\big(\mathbf{z}(\widehat{\mathbf{a}})\big)^T\Big(\nabla^2\phi\big(\mathbf{z}(\widehat{\mathbf{a}})\big)\Big)J\big(\mathbf{z}(\widehat{\mathbf{a}})\big)\mathbf{Q} = \mathbf{Q}^T\widehat{\mathbf{H}}\mathbf{Q} = \widehat{\mathbf{T}}\ .$$

As the matrix $\widehat{\mathbf{T}}$ is positive definite by hypothesis, this also shows that $\nabla^2\phi(\mathbf{z}(\widehat{\mathbf{a}}))$ is definite positive since the proposition that $\nabla^2\phi(\mathbf{z}(\widehat{\mathbf{a}}))$ is definite positive is equivalent to the proposition that $\widehat{\mathbf{T}}$ is

definite positive as noted above. Furthermore, this implies immediately that the first-order stationary point $\mathbf{z}(\widehat{\mathbf{a}})$ is a strict local minimizer of $\phi(.)$ and that the first-order stationary point $\widehat{\mathbf{a}}$ is a non strict local minimizer of $\psi(.)$ as claimed in the theorem.

Finally, since $\nabla^2\phi(\mathbf{z}(\widehat{\mathbf{a}}))$ is definite positive and we have the equality

$$\widehat{\mathbf{H}} = J\big(\mathbf{z}(\widehat{\mathbf{a}})\big)^T \Big(\nabla^2\phi\big(\mathbf{z}(\widehat{\mathbf{a}})\big)\Big) J\big(\mathbf{z}(\widehat{\mathbf{a}})\big) \ ,$$

it is easy to deduce that $\mathit{null}(\widehat{\mathbf{H}}) = \mathit{null}\big(J(\mathbf{z}(\widehat{\mathbf{a}}))\big) = \mathit{null}\big(J(\mathbf{r}(\widehat{\mathbf{a}}))\big)$ and $\mathit{dim}\big(\mathit{null}(\widehat{\mathbf{H}})\big) = k.k$, using the hypothesis $\mathit{rank}\big(J\big(\mathbf{r}(\widehat{\mathbf{a}})\big)\big) = r$ and the rank-nullity relationship 2.1. This last equality implies, finally, that $\mathit{rank}(\widehat{\mathbf{H}}) = r$ again by the rank-nullity theorem 2.1, which concludes the demonstration of the theorem.

□

Interestingly, under the same hypotheses as used in Theorem 5.10 and when, in addition, the $r \times r$ symmetric matrix $\widehat{\mathbf{T}}$ is definite positive and $\widehat{\mathbf{a}}$ is thus a (local) minimizer of $\psi(.)$, the continuum of (local) minimizers of $\psi(.)$ which achieve the same minimum value of $\psi(\widehat{\mathbf{a}})$, say $\mathcal{S}(\widehat{\mathbf{a}})$ defined formally as follows,

$$\mathcal{S}(\widehat{\mathbf{a}}) = \Big\{ \mathbf{a} \in \mathbb{R}^{p.k} \,/\, \psi(\mathbf{a}) = \psi(\widehat{\mathbf{a}}) \text{ and } \mathbf{A} = \big(\mathit{mat}_{k\times p}(\mathbf{a})\big)^T \in \mathbb{R}_k^{p\times k} \Big\} \ , \tag{5.37}$$

is locally well-behaved around $\widehat{\mathbf{a}}$ in the sense that this continuum forms a smooth (e.g., C^1) submanifold of $\mathbb{R}^{k.p}$ (locally) around $\widehat{\mathbf{a}}$ of dimension $k.k$ and the tangent linear space to $\mathcal{S}(\widehat{\mathbf{a}})$ at $\widehat{\mathbf{a}}$ is exactly the kernel of $\nabla^2\psi(\widehat{\mathbf{a}})$, i.e.,

$$\mathit{null}\big(\nabla^2\psi(\widehat{\mathbf{a}})\big) = \mathcal{T}_{\widehat{\mathbf{a}}}\mathcal{S}(\widehat{\mathbf{a}}) \ .$$

Note that, this is the best, we can hope for smooth functions like $\psi(.)$ with singular (local) minima (because of over-parameterization) and this expresses the fact that the restriction of $\nabla^2\psi(\widehat{\mathbf{a}})$ to the normal linear space of $\mathcal{S}(\widehat{\mathbf{a}})$ at $\widehat{\mathbf{a}}$, say $\mathcal{N}_{\widehat{\mathbf{a}}}\mathcal{S}(\widehat{\mathbf{a}})$ (i.e., $\mathcal{N}_{\widehat{\mathbf{a}}}\mathcal{S}(\widehat{\mathbf{a}}) = \big(\mathcal{T}_{\widehat{\mathbf{a}}}\mathcal{S}(\widehat{\mathbf{a}})\big)^{\perp}$), is definite positive (e.g., $\nabla^2\psi(\widehat{\mathbf{a}})$ is definite positive along its normal directions) as demonstrated in Theorem 5.10.

This nice geometrical property is called the Morse-Bott property in Rebjock and Boumal [163], see their Definition 1.1. They further show that this local Morse-Bott property is essentially equivalent to various (local) structural assumptions like the Polyak–Lojasiewicz condition, Quadratic Growth and Error Bound properties, which have been used in the past to explain the surprising local convergence with a fast local rate of (quasi-)Newton methods in some neighborhood of non-isolated optima; e.g., for minimizing C^2 cost functions like $\psi(.)$ for which the Hessian at (local) minima is at best positive semi-definite, but never positive definite because local minima are never isolated due to over-parameterization.

In order to demonstrate that $\psi(.)$ satisfies effectively the Morse-Bott property at $\widehat{\mathbf{a}}$, let us first give a precise definition of a smooth (i.e., C^1) submanifold around one of its points in $\mathbb{R}^n$, taken from Example 6.8 in [165].

Definition 5.1. Let $\mathcal{C}$ be a nonempty subset of $\mathbb{R}^n$.

We say that $\mathcal{C}$ is a d-dimensional smooth submanifold in $\mathbb{R}^n$ around the point $\widehat{\mathbf{a}} \in \mathcal{C}$, if $\mathcal{C}$ can be represented relative to an open neighborhood U of $\widehat{\mathbf{a}}$ as the set of solutions to $g(\mathbf{a}) = \mathbf{0}^m$ where $g(.)$ is a C^1 mapping from U to $\mathbb{R}^m$ with its differential at $\widehat{\mathbf{a}}$, $g^{'}(\widehat{\mathbf{a}})$, being a surjective linear operator, which is equivalent to say that the Jacobian matrix $J(g(\widehat{\mathbf{a}})) \in \mathbb{R}^{m\times n}$ is of full rank m, where $m = n - d$.

Thus, equivalently, $\mathcal{C}$ is a d-dimensional smooth submanifold in $\mathbb{R}^n$ around the point $\widehat{\mathbf{a}} \in \mathcal{C}$, if

$$\mathcal{C} \cap U = \{\mathbf{a} \in U \,/\, g(\mathbf{a}) = \mathbf{0}^m\} \quad \text{ and } \quad \mathit{rank}\Big(J\big(g(\widehat{\mathbf{a}})\big)\Big) = m = n - d \ .$$

Such function $g(.)$, if it exists, is called a local defining function for $\mathcal{C}$ at $\widehat{\mathbf{a}}$ as in the Definition 2.4 of a classical C^p embedded submanifold of $\mathbb{R}^n$ given in Subsection 2.4.

Obviously, if $\mathcal{M}$ is a smooth embedded submanifold of $\mathbb{R}^n$ in the sense of Definition 2.4, $\mathcal{M}$ is also a C^1 submanifold around each of its points in the sense of Definition 5.1, but the converse is not true in general. Thus, a smooth submanifold around one of its points, is just the "local" version of a classical smooth submanifold embedded in $\mathbb{R}^n$.

Next, we will use the definition of tangent vectors to an arbitrary nonempty subset of a normed vector space, given in Definition 2.5, and a finite version of the Lyusternik Theorem [105] (given below), which gives a complete characterization of the set of tangent vectors to a subset of a normed vector space of the form

$$\mathcal{S} = \{\mathbf{a} \in U \,/\, g(\mathbf{a}) = \mathbf{0}^m\} \,,$$

where U is an open set of $\mathbb{R}^n$ and $g(.)$ is a function from U to $\mathbb{R}^m$ of class C^1.

Theorem 5.11. Let U be an open subset of $\mathbb{R}^n$ and $g(.)$ a C^1 function from U to $\mathbb{R}^m$. Further, define the subset of $\mathbb{R}^n$

$$\mathcal{S} = \{\mathbf{a} \in U \,/\, g(\mathbf{a}) = \mathbf{0}^m\} \,,$$

and let $\widehat{\mathbf{a}} \in \mathcal{S}$.

If the differential of $g(.)$ at $\widehat{\mathbf{a}}$, $g^{'}(\widehat{\mathbf{a}})$, is a surjective linear operator from $\mathbb{R}^n$ to $\mathbb{R}^m$, which is equivalent to say that *rank*$(J(g(\widehat{\mathbf{a}}))) = m$, then

$$\textit{null}\big(J(g(\widehat{\mathbf{a}}))\big) = \mathcal{T}_{\widehat{\mathbf{a}}}\mathcal{S} \,,$$

where $J(g(\widehat{\mathbf{a}}))$ is the Jacobian matrix of $g(.)$ at $\widehat{\mathbf{a}}$ and $\mathcal{T}_{\widehat{\mathbf{a}}}\mathcal{S}$ is the set of tangent vectors to $\mathcal{S}$ at $\widehat{\mathbf{a}}$ in the sense of Definition 2.5, which is thus a linear subspace of $\mathbb{R}^n$ of dimension $d = n - m$.

Proof. Omitted. See [105] and [168] for proof. □

The Lyusternik Theorem 5.11 is the angular stone behind the Definition 5.1 of an embedded C^1 submanifold $\mathcal{C}$ around one of its points in $\mathbb{R}^n$. Furthermore, it allows an easy identification of the tangent space to such "local" submanifolds as the kernel of the Jacobian of the local defining function for $\mathcal{C}$ at this particular point and also of the dimension of such smooth submanifold around one of its points. These "local" results are further consistent with those concerning classical smooth submanifolds embedded in normed vector spaces discussed in Subsection 2.4.

With these preliminaries, we are now in position to demonstrate that the continuum of (local) minima of $\psi(.)$, $\mathcal{S}(\widehat{\mathbf{a}})$, defined in equation (5.37), is effectively a smooth (e.g., C^1) submanifold of $\mathbb{R}^{k.p}$ (locally) around $\widehat{\mathbf{a}}$ of dimension $k.k$ in the sense of Definition 5.1 and that

$$\textit{null}\big(\nabla^2\psi(\widehat{\mathbf{a}})\big) = \mathcal{T}_{\widehat{\mathbf{a}}}\mathcal{S}(\widehat{\mathbf{a}}) \,,$$

if we assume, as in Theorem 5.10, that:

- $\widehat{\mathbf{a}} \in \mathbb{R}^{k.p}$, with $\widehat{\mathbf{A}} = (\textit{mat}_{k\times p}(\widehat{\mathbf{a}}))^T \in \mathbb{R}_k^{p\times k}$, is a first-order stationary point of $\psi(.)$;

- $\mathbf{F}(\mathbf{a})$ has full column-rank and the rank of the Jacobian matrix $J(\mathbf{r}(\mathbf{a}))$ is equal to $r = (p-k).k$ in a neighborhood of $\widehat{\mathbf{a}}$.

- and, finally, the $r \times r$ symmetric matrix $\widehat{\mathbf{T}}$ defined in equation (5.35) is positive definite.

Under these hypotheses, we know from Theorem 5.10, that $\widehat{\mathbf{a}}$ is a (local) minimizer of $\psi(.)$ and we can assert that the set $\mathcal{S}(\widehat{\mathbf{a}})$ defined in equation (5.37) is nonempty and even forms a continuum of points. Furthermore, again as a direct consequence of Theorem 5.10, we also have

$$\textit{null}\big(\nabla^2\psi(\widehat{\mathbf{a}})\big) = \textit{null}\big(J(\mathbf{r}(\widehat{\mathbf{a}}))\big) \quad \text{and} \quad \textit{dim}\big(\textit{null}\big(\nabla^2\psi(\widehat{\mathbf{a}})\big)\big) = k.k \,.$$

We now first demonstrate that $\mathcal{S}(\widehat{\mathbf{a}})$ is effectively an embedded $k.k$-dimensional C^1 submanifold in $\mathbb{R}^{k.p}$ around the point $\widehat{\mathbf{a}} \in \mathcal{S}(\widehat{\mathbf{a}})$ in the sense of Definition 5.1 and that the tangent space to $\mathcal{S}(\widehat{\mathbf{a}})$ around $\widehat{\mathbf{a}}$ is nothing else than the kernel of $J(\mathbf{r}(\widehat{\mathbf{a}}))$ using the Lyusternik Theorem 5.11.

To verify Definition 5.1 for $\mathcal{S}(\widehat{\mathbf{a}})$, we need to find an open neighborhood U of $\widehat{\mathbf{a}}$ in $\mathbb{R}^{k.p}$ and a C^1 mapping $g(.)$ from U to $\mathbb{R}^r$ with $r = k.(p-k)$ such that

$$\mathcal{S}(\widehat{\mathbf{a}}) \cap U = \{\mathbf{a} \in U \,/\, g(\mathbf{a}) = \mathbf{0}^r\} \quad \text{and} \quad \mathit{rank}\big(J(\mathbf{g}(\widehat{\mathbf{a}}))\big) = r \; .$$

However, as in the demonstration of Theorem 5.10, the above hypotheses imply that there exist an open neighborhood Υ of $\widehat{\mathbf{a}}$ in $\mathbb{R}^{k.p}$, a twice continuously differentiable function $\mathbf{z}(.)$ from Υ to $\mathbb{R}^r$ and a twice continuously differentiable function $\mathbf{h}(.)$ from $\mathbb{R}^r$ to $\mathbb{R}^{k.p}$ such that $\mathbf{r}(\mathbf{a}) = \mathbf{h}(\mathbf{z}(\mathbf{a}))$ and the Jacobian matrices, $J(\mathbf{z}(\mathbf{a}))$ and $J(\mathbf{h}(\mathbf{z}(\mathbf{a})))$, have a constant rank equals to $r = (p-k).k$ for all $\mathbf{a} \in \Upsilon$.

Furthermore, we can also define a twice continuously differentiable real function $\phi(.)$ from $\mathbb{R}^r$ to $\mathbb{R}$ such that $\phi(\mathbf{o}) = \frac{1}{2}\|\mathbf{h}(\mathbf{o})\|_2^2$, $\forall \mathbf{o} \in \mathbb{R}^r$, and, again according to the demonstration of Theorem 5.10, the point $\mathbf{z}(\widehat{\mathbf{a}})$ is a strict (local) minimizer of $\phi(.)$. This implies the existence of an open neighborhood V of $\mathbf{z}(\widehat{\mathbf{a}})$ in $\mathbb{R}^r$ such that ,$\forall \mathbf{o} \in V$, we have $\phi(\mathbf{o}) > \phi(\mathbf{z}(\widehat{\mathbf{a}}))$. As the function $\mathbf{z}(.)$ is continuous over Υ, the set $U = \mathbf{z}^{-1}(V)$ is of the form $U = \Upsilon \cap W$, where W is an open set of $\mathbb{R}^{k.p}$, and U is thus an open neighborhood of $\widehat{\mathbf{a}}$ in $\mathbb{R}^{k.p}$. Clearly, the elements $\mathbf{a} \in \mathcal{S}(\widehat{\mathbf{a}}) \cap U$ verify $\mathbf{z}(\mathbf{a}) = \mathbf{z}(\widehat{\mathbf{a}})$.

In these conditions, U is an open neighborhood of $\widehat{\mathbf{a}}$ and we can define a twice continuously differentiable function $g(.)$ from U to $\mathbb{R}^r$ by

$$g : U \longrightarrow \mathbb{R}^r : \mathbf{a} \mapsto g(\mathbf{a}) = \mathbf{z}(\mathbf{a}) - \mathbf{z}(\widehat{\mathbf{a}}) \; ,$$

and we have effectively

$$\mathcal{S}(\widehat{\mathbf{a}}) \cap U = \{\mathbf{a} \in U \,/\, g(\mathbf{a}) = \mathbf{0}^m\} \; .$$

Furthermore, the Jacobian matrix $J\big(g(\mathbf{a})\big) = J\big(\mathbf{z}(\mathbf{a})\big) \in \mathbb{R}^{r \times k.p}$ verifies $\mathit{rank}\Big(J\big(g(\mathbf{a})\big)\Big) = r, \forall \mathbf{a} \in U$ (see the demonstration of Theorem 5.10). This implies in particular that $\mathit{rank}\Big(J\big(g(\widehat{\mathbf{a}})\big)\Big) = r$ and $g^{'}(\widehat{\mathbf{a}})$ is a surjective linear mapping from $\mathbb{R}^{k.p}$ to $\mathbb{R}^r$ and we conclude that $\mathcal{S}(\widehat{\mathbf{a}})$ is effectively an embedded C^1 submanifold around the (local) minimizer $\widehat{\mathbf{a}}$ in $\mathbb{R}^{k.p}$ of dimension $k.p - r = k.k$.

Moreover, a closer look at the preceding demonstration further shows that the set $\mathcal{S}(\widehat{\mathbf{a}}) \cap U$ is in fact also an embedded $k.k$-dimensional C^2 submanifold in $\mathbb{R}^{k.p}$ in the sense of Definition 2.4 because the mapping $g(.)$ from U to $\mathbb{R}^r$, defined above, is of class C^2 and a valid local defining function for $\mathcal{S}(\widehat{\mathbf{a}}) \cap U$ at all $\mathbf{a} \in \mathcal{S}(\widehat{\mathbf{a}}) \cap U$.

Let us now determine the tangent space to $\mathcal{S}(\widehat{\mathbf{a}}) \cap U$ at $\mathbf{a} \in \mathcal{S}(\widehat{\mathbf{a}}) \cap U$. By applying the Lyusternik Theorem 5.11 to the set $\mathcal{S}(\widehat{\mathbf{a}}) \cap U$, we have immediately

$$\mathit{null}\Big(J\big(g(\mathbf{a})\big)\Big) = \mathit{null}\Big(J\big(\mathbf{z}(\mathbf{a})\big)\Big) = \mathcal{T}_{\mathbf{a}}\big(\mathcal{S}(\widehat{\mathbf{a}}) \cap U\big) \quad , \forall \mathbf{a} \in \mathcal{S}(\widehat{\mathbf{a}}) \cap U \; .$$

Finally, using the equality

$$\mathit{null}\big(J(\mathbf{r}(\mathbf{a}))\big) = \mathit{null}\big(J(\mathbf{z}(\mathbf{a}))\big) \, , \forall \mathbf{a} \in \Upsilon \; ,$$

established in the demonstration of Theorem 5.10), we get the equality

$$\mathit{null}\big(J(\mathbf{r}(\mathbf{a}))\big) = \mathcal{T}_{\mathbf{a}}\big(\mathcal{S}(\widehat{\mathbf{a}}) \cap U\big) \quad , \forall \mathbf{a} \in \mathcal{S}(\widehat{\mathbf{a}}) \cap U \; .$$

This shows that the tangent space to $\mathcal{S}(\widehat{\mathbf{a}}) \cap U$ at $\mathbf{a} \in \mathcal{S}(\widehat{\mathbf{a}}) \cap U$ is nothing else than the kernel of the Jacobian matrix $J(\mathbf{r}(\mathbf{a}))$, $\forall \mathbf{a} \in \mathcal{S}(\widehat{\mathbf{a}}) \cap U$. This implies in particular that

$$\mathit{null}\big(J(\mathbf{r}(\widehat{\mathbf{a}}))\big) = \mathcal{T}_{\widehat{\mathbf{a}}}\big(\mathcal{S}(\widehat{\mathbf{a}}) \cap U\big) \; .$$

Finally, using the hypothesis that $\widehat{\mathbf{T}}$ (defined in equation (5.35)) is positive definite and Theorem 5.10, we obtain the equality

$$\mathit{null}\big(\nabla^2\psi(\widehat{\mathbf{a}})\big) = \mathit{null}\big(J(\mathbf{r}(\widehat{\mathbf{a}}))\big) = \mathcal{T}_{\widehat{\mathbf{a}}}\big(\mathcal{S}(\widehat{\mathbf{a}}) \cap U\big) \; ,$$

which demonstrates that the cost function $\psi(.)$ effectively verifies the so-called Morse-Bott property at the (local) minimizer $\widehat{\mathbf{a}}$ under the hypotheses of Theorem 5.10 and when, in addition, the $r \times r$ symmetric matrix $\widehat{\mathbf{T}}$ defined in equation (5.35) is positive definite as claimed above. Note that if $\widehat{\mathbf{a}}$ is a (local) minimizer of $\psi(.)$, but we don't assume the hypothesis $\widehat{\mathbf{T}}$ is positive definite, we can still get the inclusion

$$\mathcal{T}_{\widehat{\mathbf{a}}}\big(\mathcal{S}(\widehat{\mathbf{a}}) \cap U\big) \subset \mathit{null}\big(\nabla^2\psi(\widehat{\mathbf{a}})\big) \; ,$$

by using Theorem 5.9 and Corollary 5.9, and the fact that $\widehat{\mathbf{a}}$ is a first-order stationary point of $\psi(.)$.

In order to show now that Theorem 5.10 covers a large body of real applications, we note that the matrix variable $\mathbf{A}$ is always requested to have full column rank to assure the differentiability of $\psi(.)$ and that $\mathbf{F}(\mathbf{a})$ is of full column rank in most cases as soon as every line or column of the data matrix $\mathbf{X}$ has at least k non-missing entries. Furthermore, the hypothesis that the rank of the Jacobian matrix $J(\mathbf{r}(\mathbf{a}))$ is constant and equal to $r = (p-k).k$ in a neighborhood of $\widehat{\mathbf{a}}$ is automatically verified if the hypotheses of Theorem 5.3 are fulfilled in a neighborhood of $\widehat{\mathbf{a}}$. Finally, we note that orthonormal bases given, respectively, by the columns of $\mathbf{P}$ and $\mathbf{Q}$ for the null space of $J(\mathbf{r}(\widehat{\mathbf{a}}))$ and its orthogonal complement, used in Theorem 5.10, can be easily computed from the results of Corollary 5.6.

Obviously, Theorem 5.10 also justifies the extension of the regularization techniques, developed in Subsection 5.2 to overcome the systematic singularity of the Jacobian matrix (or its approximation) in the Gauss-Newton or Levenberg-Marquardt methods, to the Newton method using the full Hessian matrix $\mathbf{H}$ or its two-term approximation $\bar{\mathbf{H}}$ derived above. In other words, to avoid the singularity or ill-conditioning of these two symmetric matrices near first-order stationary points of $\psi(.)$, the Newton correction vector $d\mathbf{a}_n \in \mathbb{R}^{k.p}$ can be found in a two-step procedure at each iteration, as for the Gauss-Newton algorithm described in Subsection 5.2. In a first step, we solve the $(p-k).k \times (p-k).k$ symmetric linear system

$$(\bar{\mathbf{O}}^\perp)^T\mathbf{H}\bar{\mathbf{O}}^\perp d\bar{\mathbf{a}}_n = -(\bar{\mathbf{O}}^\perp)^T J\big(\mathbf{r}(\mathbf{a})\big)^T\mathbf{r}(\mathbf{a}) = \big(\mathbf{M}(\mathbf{a})\bar{\mathbf{O}}^\perp\big)^T\mathbf{r}(\mathbf{a}) \; , \tag{5.38}$$

or, if we use the two-term approximation of the Hessian, $\bar{\mathbf{H}}$ defined in equation (5.34),

$$(\bar{\mathbf{O}}^\perp)^T\bar{\mathbf{H}}\bar{\mathbf{O}}^\perp d\bar{\mathbf{a}}_n = \big(\mathbf{M}(\mathbf{a})\bar{\mathbf{O}}^\perp\big)^T\mathbf{r}(\mathbf{a}) \; , \tag{5.39}$$

for $d\bar{\mathbf{a}}_n \in \mathbb{R}^{(p-k).k}$ and where $\bar{\mathbf{O}}^\perp = \mathbf{K}_{(p,k)}(\mathbf{I}_k \otimes \mathbf{O}^\perp) \in \mathbb{O}^{k.p\times(p-k).k}$ and $\mathbf{O}^\perp \in \mathbb{O}^{p\times(p-k)}$ are orthonormal matrices whose columns form, respectively, a basis of $\mathit{null}\big(J(\mathbf{r}(\mathbf{a}))\big)^\perp$ and $\mathit{ran}(\mathbf{A})^\perp$, see Corollary 5.6 and Theorem 5.7 for details. In a second step, to get the Newton correction vector we need to compute the following matrix-vector product

$$d\mathbf{a}_n = \bar{\mathbf{O}}^\perp d\bar{\mathbf{a}}_n \in \mathbb{R}^{k.p} \; ,$$

or, equivalently, in matrix format,

$$d\mathbf{A}_n = \mathbf{O}^\perp d\bar{\mathbf{A}}_n \in \mathbb{R}^{p\times k} \; .$$

This modification of the Newton algorithm in the context of the (VP1) problem, first suggested by Chen [28], has a strong theoretical justification as it can be interpreted as a Riemannian Newton operating on the (quotient) Grassmann manifold $\mathrm{Gr}(p,k)$ as we will show below, but it is also computationally very expensive as $\bar{\mathbf{O}}^\perp$ is huge matrix in most cases.

Alternatively, we can use a cheaper alternative based on the orthogonality constraint

$$\mathbf{N}^T d\mathbf{a}_n = \mathbf{0}^{k.k} ,$$

where $\mathbf{N} = \mathbf{K}_{(p,k)}(\mathbf{I}_k \otimes \mathbf{A})$ or $\mathbf{N} = \mathbf{K}_{(p,k)}(\mathbf{I}_k \otimes \mathbf{O})$, and the columns of the matrix $\mathbf{A}$ ($\mathbf{O}$) form a (orthonormal) basis of *ran*$(\mathbf{A})$ and the columns of $\mathbf{N}$ is a (orthonormal) basis of *null*$\big(J(\mathbf{r}(\mathbf{a}))\big) =$ *null*$\big(\mathbf{M}(\mathbf{a})\big)$, as demonstrated in Corollary 5.6 and discussed in Subsection 5.2, and proceed in one step by solving the damped symmetric linear system

$$\big(\mathbf{H} + \mathbf{N}\mathbf{N}^T\big) d\mathbf{a}_n = -\nabla\psi(\mathbf{a}) = \mathbf{M}(\mathbf{a})^T\mathbf{r}(\mathbf{a}) \tag{5.40}$$

or

$$\big(\bar{\mathbf{H}} + \mathbf{N}\mathbf{N}^T\big) d\mathbf{a}_n = -\nabla\psi(\mathbf{a}) = \mathbf{M}(\mathbf{a})^T\mathbf{r}(\mathbf{a}) , \tag{5.41}$$

if we use the two-term approximation of the Hessian matrix. Note that, in both cases, the term $\mathbf{N}\mathbf{N}^T$ can be efficiently evaluated as

$$\mathbf{N}\mathbf{N}^T = \mathbf{K}_{(p,k)}(\mathbf{I}_k \otimes \mathbf{A}\mathbf{A}^T)\mathbf{K}_{(k,p)} \quad \text{or} \quad \mathbf{N}\mathbf{N}^T = \mathbf{K}_{(p,k)}(\mathbf{I}_k \otimes \mathbf{O}\mathbf{O}^T)\mathbf{K}_{(k,p)} . \tag{5.42}$$

This last approach using a linear constraint is new in the context of Newton methods, but is simply an extension to (quasi-)Newton methods of the technique first proposed by Okatani et al. [150] for Gauss-Newton and Levenberg-Marquardt methods. The variant of the Newton method, defined by equation (5.41), can also be interpreted as a Riemannian quasi-Newton operating on the (quotient) Grassmann manifold $\text{Gr}(p, k)$, but not the first one defined by equation (5.40) as we will see below.

Assuming that the dimension of the null space of $\mathbf{H}$ or $\bar{\mathbf{H}}$ at first-order stationary points of $\psi(.)$ is equal to the dimension of the null space of the Jacobian matrix and that both are equal to $k.k$, these different approaches will efficiently overcome the systematic singularity of $\bar{\mathbf{H}}$ or those of $\mathbf{H}$ at these first-order stationary points of $\psi(.)$. In these conditions, the above symmetric linear systems will have an unique solution and the (quasi-)Newton direction is thus well defined. Of course, as always in Newton methods, one common weakness of these two second-order approaches for solving the (VP1) problem is that $\mathbf{H}$ or its two-term approximation $\bar{\mathbf{H}}$ may not be positive definite at some points in a region of mixed curvature. In such condition, the Newton direction may not be in a descent direction and the above linear systems cannot be solved by a simple Cholesky factorization. However, many standard techniques are available to deal with this classical problem [139][123] and can be applied here in our specific WLRA context without any modification as we will discussed in Subsection 6.3.

Alternatively, we can again recast the WLRA problem in its variable projection formulation as an optimization problem on the (quotient) Grassmann manifold $\text{Gr}(p, k)$ [3][11] and use a Riemannian Newton method to solve it as was done for example in [13][14]. To clarify the differences and similarities between the Riemannian Newton method operating on $\text{Gr}(p, k)$ and the above (quasi)-Newton algorithms operating on $\mathbb{R}_k^{p\times k}$ (or on $\text{St}(p, k) = \mathbb{O}^{p\times k}$), we now investigate the relationships between the Euclidean Hessian of $\psi(.)$ at $\mathbf{a}$, given in equation (5.33) (e.g., when $\psi(.)$ is considered as a real function from $\mathbb{R}^{p.k}$ to $\mathbb{R}$, see equation (3.23)), and the Riemannian Hessian of the unvectorized form of $\psi(.)$ at $\mathbf{A} \in \mathbb{R}_k^{p\times k}$ (e.g., the real function $\psi \circ h^{-1}(.)$ from $\mathbb{R}_k^{p\times k}$ to $\mathbb{R}$, where $h^{-1}(.)$ is defined in equation (3.29) of Subsection 3.4 with $h^{-1}(\mathbf{A}) = vec(\mathbf{A}^T) = \mathbf{a}, \forall \mathbf{A} \in \mathbb{R}_k^{p\times k}$), when this cost function is considered abusively as defined on the Grassmann manifold $\text{Gr}(p, k)$, as already discussed at the end of Subsection 5.2 and at the beginning of this subsection. The results we present now are a slight extension, with our notations, of those given in [82].

As in our previous discussion on the connections between the Euclidean gradient of $\psi(.)$ and the Riemannian gradient of $\psi \circ h^{-1}(.)$, to simplify the presentation we require that $\mathbf{W} \in \mathbb{R}_{+*}^{p\times n}$ and that each element $\mathring{\mathbf{O}} \in \text{Gr}(p, k)$ is represented by an element of the compact Stiefel manifold $\mathbf{O} \in \text{St}(p, k)$, in line with previous works on Riemannian optimization on $\text{Gr}(p, k)$ [47][14][11].

Recall also that any element of the tangent space of $\mathrm{Gr}(p,k)$ at $\mathring{\mathbf{O}}$, $\mathcal{T}_{\mathring{\mathbf{O}}}\mathrm{Gr}(p,k)$, can be represented uniquely by an element of the following linear subspace of $\mathbb{R}^{p\times k}$ of dimension $(p-k).k$:

$$\mathcal{T}_{\mathbf{O}}\mathrm{Gr}(p,k) = \left\{\mathbf{D} \in \mathbb{R}^{p\times k} \,/\, \mathbf{O}^T\mathbf{D} = \mathbf{0}^{k\times k}\right\} ,$$

as already noted in Subsection 5.2. $\mathcal{T}_{\mathbf{O}}\mathrm{Gr}(p,k)$ is nothing else than the horizontal space of $\mathrm{St}(p,k)$ at $\mathbf{O} \in \mathrm{St}(p,k)$, noted $\mathcal{H}_{\mathbf{O}}\mathrm{St}(p,k)$ or $\mathcal{H}_{\mathbf{O}}\mathbb{O}^{p\times k}$ in Subsection 2.4. In these conditions, using Theorem 5.8, we have

$$\nabla_R\psi \circ h^{-1}(\mathring{\mathbf{O}}) = \left(\mathbf{I}_p - \mathbf{O}\mathbf{O}^T\right) \nabla_F\psi \circ h^{-1}(\mathbf{O}) = \nabla_F\psi \circ h^{-1}(\mathbf{O}) ,$$

where $\nabla_R\psi \circ h^{-1}(\mathring{\mathbf{O}})$ is the Riemannian gradient of the smooth cost function $\psi \circ h^{-1}(.)$ (defined on the Grassmann manifold $\mathrm{Gr}(p,k)$) at $\mathring{\mathbf{O}} \in \mathrm{Gr}(p,k)$ and $\nabla_F\psi \circ h^{-1}(\mathbf{O})$ is the standard Frobenius gradient of $\psi \circ h^{-1}(.)$ at $\mathbf{O} \in \mathrm{St}(p,k)$.

In this framework, the Riemannian Hessian of the smooth map $\psi \circ h^{-1}(.)$ defined on the Grassmann manifold $\mathrm{Gr}(p,k)$ at $\mathring{\mathbf{O}} = \mathit{ran}(\mathbf{O}) \in \mathrm{Gr}(p,k)$, is thus a linear transformation from $\mathcal{T}_{\mathbf{O}}\mathrm{Gr}(p,k)$ to $\mathcal{T}_{\mathbf{O}}\mathrm{Gr}(p,k)$. Furthermore, using the above equality between the Riemannian and Frobenius gradients and equation (2.52) in Subsection 2.4, $\forall \mathbf{D} \in \mathcal{T}_{\mathbf{O}}\mathrm{Gr}(p,k)$, the image of $\mathbf{D}$ by the Hessian $\nabla^2_R\psi \circ h^{-1}(\mathring{\mathbf{O}})$ considered as a linear operator from $\mathcal{T}_{\mathbf{O}}\mathrm{Gr}(p,k)$ to $\mathcal{T}_{\mathbf{O}}\mathrm{Gr}(p,k)$ is given by

$$[\nabla^2_R\psi \circ h^{-1}(\mathring{\mathbf{O}})](\mathbf{D}) = \left(\mathbf{I}_p - \mathbf{O}\mathbf{O}^T\right) [\nabla^2_F\psi \circ h^{-1}(\mathbf{O})](\mathbf{D}) , \tag{5.43}$$

where $\nabla^2_F\psi \circ h^{-1}(\mathbf{O})$ is the standard Frobenius Hessian of $\psi \circ h^{-1}(.)$ at $\mathbf{O} \in \mathrm{St}(p,k)$ and $\left(\mathbf{I}_p - \mathbf{O}\mathbf{O}^T\right)$ is the orthogonal projector onto $\mathit{ran}(\mathbf{O})$ in $\mathbb{R}^p$, but also the orthogonal projector onto $\mathcal{T}_{\mathbf{O}}\mathrm{Gr}(p,k)$ in $\mathbb{R}^{p\times k}$, when $\mathcal{T}_{\mathbf{O}}\mathrm{Gr}(p,k)$ is considered as a subspace of $\mathbb{R}^{p\times k}$ of dimension $(p-k).k$, which is also the horizontal space of $\mathrm{St}(p,k)$ at $\mathbf{O} \in \mathrm{St}(p,k)$ as noted above.

Thus, we have $\mathbf{D} \in \mathcal{T}_{\mathbf{O}}\mathrm{Gr}(p,k)$ and $[\nabla^2_R\psi \circ h^{-1}(\mathring{\mathbf{O}})](\mathbf{D}) \in \mathcal{T}_{\mathbf{O}}\mathrm{Gr}(p,k)$, and both are considered as $p \times k$ matrices in equation (5.43). In these conditions, as noted in Subsection 5.2, equivalently, we can vectorize both $\mathbf{D}$ and $[\nabla^2_R\psi \circ h^{-1}(\mathring{\mathbf{O}})](\mathbf{D})$ as

$$\mathbf{h} = \mathit{vec}\left(\mathbf{D}^T\right) \in \mathbb{R}^{p.k} \quad \text{and} \quad \nabla^2_R\psi(\mathring{\mathbf{o}})\mathbf{d} = \mathit{vec}\left(\left([\nabla^2_R\psi \circ h^{-1}(\mathring{\mathbf{O}})](\mathbf{D})\right)^T\right) \in \mathbb{R}^{p.k} ,$$

where now $\nabla^2_R\psi(\mathring{\mathbf{o}})$ is a $p.k \times p.k$ (asymmetric) matrix.

Furthermore, using equations (5.43), (2.32), (2.36) and Corollary 5.6, we have, $\forall \mathbf{d} \in \mathbb{R}^{p.k}$,

$$\begin{aligned}
\nabla^2_R\psi(\mathring{\mathbf{o}})\mathbf{d} &= \mathit{vec}\left(\left(\left(\mathbf{I}_p - \mathbf{O}\mathbf{O}^T\right) [\nabla^2_F\psi \circ h^{-1}(\mathbf{O})](\mathbf{D})\right)^T\right) \\
&= \mathit{vec}\left(\left([\nabla^2_F\psi \circ h^{-1}(\mathbf{O})](\mathbf{D})\right)^T \left(\mathbf{I}_p - \mathbf{O}\mathbf{O}^T\right)\right) \\
&= \left(\left(\mathbf{I}_p - \mathbf{O}\mathbf{O}^T\right) \otimes \mathbf{I}_k\right) \mathit{vec}\left(\left([\nabla^2_F\psi \circ h^{-1}(\mathbf{O})](\mathbf{D})\right)^T\right) \\
&= \left(\mathbf{I}_{p.k} - \left(\mathbf{O}\mathbf{O}^T \otimes \mathbf{I}_k\right)\right) \mathit{vec}\left(\left([\nabla^2_F\psi \circ h^{-1}(\mathbf{O})](\mathbf{D})\right)^T\right) \\
&= \left(\mathbf{I}_{p.k} - \bar{\mathbf{O}}\bar{\mathbf{O}}^T\right) \mathit{vec}\left(\left([\nabla^2_F\psi \circ h^{-1}(\mathbf{O})](\mathbf{D})\right)^T\right) \\
&= \left(\mathbf{I}_{p.k} - \bar{\mathbf{O}}\bar{\mathbf{O}}^T\right) \nabla^2\psi(\mathbf{o})\mathbf{d} ,
\end{aligned}$$

where $\nabla^2\psi(\mathbf{o})$ is a $p.k \times p.k$ symmetric matrix and the orthogonal projector $\left(\mathbf{I}_p - \mathbf{O}\mathbf{O}^T\right)$ onto $\mathcal{T}_{\mathbf{O}}\mathrm{Gr}(p,k)$ is represented in vectorized form by the orthogonal projector onto $\mathit{null}\big(J(\mathbf{r}(\mathbf{o}))\big)^{\perp} = \mathit{null}\big(\mathbf{M}(\mathbf{o})\big)^{\perp}$ given by $\left(\mathbf{I}_{p.k} - \bar{\mathbf{O}}\bar{\mathbf{O}}^T\right)$, where the columns of $\bar{\mathbf{O}} = \mathbf{K}_{(p,k)}(\mathbf{I}_k \otimes \mathbf{O})$ are an orthonormal basis of $\mathit{null}\big(J(\mathbf{r}(\mathbf{o}))\big) = \mathit{null}\big(\mathbf{M}(\mathbf{o})\big)$, as demonstrated in Corollary 5.6 of Subsection 5.2.

This leads to the matrix equality

$$\nabla^2_R\psi(\mathring{\mathbf{o}}) = \left(\mathbf{I}_{p.k} - \bar{\mathbf{O}}\bar{\mathbf{O}}^T\right) \nabla^2\psi(\mathbf{o}) ,$$

and, using the explicit form of $\nabla^2\psi(\mathbf{o})$ given by equation (5.33), we obtain:

$$\begin{aligned}\nabla^2_R\psi(\mathring{\mathbf{o}}) &= \left(\mathbf{I}_{p.k} - \bar{\mathbf{O}}\bar{\mathbf{O}}^T\right)\left(\mathbf{M}(\mathbf{o})^T\mathbf{M}(\mathbf{o}) - \mathbf{L}(\mathbf{o})^T\mathbf{L}(\mathbf{o}) + \mathbf{U}(\mathbf{o})^T\mathbf{L}(\mathbf{o}) + \mathbf{L}(\mathbf{o})^T\mathbf{U}(\mathbf{o})\right)\\ &= \mathbf{M}(\mathbf{o})^T\mathbf{M}(\mathbf{o}) - \mathbf{L}(\mathbf{o})^T\mathbf{L}(\mathbf{o}) + \left(\mathbf{I}_{p.k} - \bar{\mathbf{O}}\bar{\mathbf{O}}^T\right)\mathbf{U}(\mathbf{o})^T\mathbf{L}(\mathbf{o}) + \mathbf{L}(\mathbf{o})^T\mathbf{U}(\mathbf{o})\ , \quad (5.44)\end{aligned}$$

since

$$\begin{aligned}\left(\mathbf{I}_{p.k} - \bar{\mathbf{O}}\bar{\mathbf{O}}^T\right)\mathbf{M}(\mathbf{o})^T &= \mathbf{M}(\mathbf{o})^T - \bar{\mathbf{O}}(\bar{\mathbf{O}}^T\mathbf{M}(\mathbf{o})^T)\\ &= \mathbf{M}(\mathbf{o})^T - \bar{\mathbf{O}}(\mathbf{M}(\mathbf{o})\bar{\mathbf{O}})^T\\ &= \mathbf{M}(\mathbf{o})^T\ ,\end{aligned}$$

as the columns of $\bar{\mathbf{O}}$ form an orthonormal basis of $\mathit{null}\big(\mathbf{M}(\mathbf{o})\big)$ and, similarly,

$$\begin{aligned}\left(\mathbf{I}_{p.k} - \bar{\mathbf{O}}\bar{\mathbf{O}}^T\right)\mathbf{L}(\mathbf{o})^T &= \mathbf{L}(\mathbf{o})^T - \bar{\mathbf{O}}(\bar{\mathbf{O}}^T\mathbf{L}(\mathbf{o})^T)\\ &= \mathbf{L}(\mathbf{o})^T - \bar{\mathbf{O}}(\mathbf{L}(\mathbf{o})\bar{\mathbf{O}})^T\\ &= \mathbf{L}(\mathbf{o})^T\ ,\end{aligned}$$

as $\mathit{null}\big(J(\mathbf{r}(\mathbf{o}))\big) = \mathit{null}\big(\mathbf{M}(\mathbf{o})\big) \cap \mathit{null}\big(\mathbf{L}(\mathbf{o})\big) \subset \mathit{null}\big(\mathbf{L}(\mathbf{o})\big)$ and, thus, each column vector of $\bar{\mathbf{O}}$ is also an element of $\mathit{null}\big(\mathbf{L}(\mathbf{o})\big)$ when $\mathbf{W} \in \mathbb{R}^{p\times n}_{+*}$ according to Corollary 5.5.

Since $\nabla^2_R\psi \circ h^{-1}(\mathring{\mathbf{O}})$ is a linear operator from $\mathcal{T}_{\mathbf{O}}\mathrm{Gr}(p,k)$ to $\mathcal{T}_{\mathbf{O}}\mathrm{Gr}(p,k)$, which is of dimension $(p-k).k$, the $p.k \times p.k$ matrix $\nabla^2_R\psi(\mathring{\mathbf{o}})$ represents a linear mapping from $\mathit{null}\big(J(\mathbf{r}(\mathbf{a}))\big)^\perp$ to $\mathit{null}\big(J(\mathbf{r}(\mathbf{a}))\big)^\perp$ (or equivalently from $\mathit{null}\big(\mathbf{M}(\mathbf{o})\big)^\perp$ to $\mathit{null}\big(\mathbf{M}(\mathbf{o})\big)^\perp$) and we can represent this linear mapping in terms of the orthonormal basis of $\mathit{null}\big(J(\mathbf{r}(\mathbf{a}))\big)^\perp$ given by the columns of $\bar{\mathbf{O}}^\perp = \mathbf{K}_{(p,k)}(\mathbf{I}_k \otimes \mathbf{O}^\perp) \in \mathbb{O}^{k.p\times(p-k).k}$. When we do so the Riemannian Newton direction vector, $d\bar{\mathbf{o}}_{r-n}$, in $\mathit{null}\big(\mathbf{M}(\mathbf{o})\big)^\perp$ can be computed as the solution of the following $(p-k).k \times (p-k).k$ symmetric linear system of equations

$$(\bar{\mathbf{O}}^\perp)^T\nabla^2_R\psi(\mathring{\mathbf{o}})\bar{\mathbf{O}}^\perp d\bar{\mathbf{o}}_{r-n} = -(\bar{\mathbf{O}}^\perp)^T\nabla_R\psi(\mathring{\mathbf{o}})\ , \quad (5.45)$$

which is exactly equivalent to the symmetric linear system given in equation 5.38 since

$$\begin{aligned}(\bar{\mathbf{O}}^\perp)^T\nabla^2_R\psi(\mathring{\mathbf{o}}) &= (\bar{\mathbf{O}}^\perp)^T\left(\mathbf{I}_{p.k} - \bar{\mathbf{O}}\bar{\mathbf{O}}^T\right)\nabla^2\psi(\mathbf{o})\\ &= (\bar{\mathbf{O}}^\perp)^T\bar{\mathbf{O}}^\perp(\bar{\mathbf{O}}^\perp)^T\nabla^2\psi(\mathbf{o})\\ &= (\bar{\mathbf{O}}^\perp)^T\nabla^2\psi(\mathbf{o})\ ,\end{aligned}$$

and

$$\begin{aligned}-(\bar{\mathbf{O}}^\perp)^T\nabla_R\psi(\mathring{\mathbf{o}}) &= -(\bar{\mathbf{O}}^\perp)^T\nabla\psi(\mathbf{o})\\ &= -(\bar{\mathbf{O}}^\perp)^T J\big(\mathbf{r}(\mathbf{o})\big)^T\mathbf{r}(\mathbf{o})\\ &= \big(\mathbf{M}(\mathbf{o})\bar{\mathbf{O}}^\perp\big)^T\mathbf{r}(\mathbf{o})\ .\end{aligned}$$

In a final step, we can also get the Riemannian Newton direction vector, $d\mathbf{o}_{r-n}$, in $\mathbb{R}^{k.p}$ as

$$d\mathbf{o}_{r-n} = \bar{\mathbf{O}}^\perp d\bar{\mathbf{o}}_{r-n}\ ,$$

and we have both $d\mathbf{o}_{r-n} = d\mathbf{o}_n$ and $d\bar{\mathbf{o}}_{r-n} = d\bar{\mathbf{o}}_n$. Thus, the Newton iteration defined by equation (5.38) can effectively be considered as a Riemannian Newton method if the next Newton iterate is computed with the help of a proper retraction onto the Stiefel manifold $\mathrm{St}(p,k)$, as defined in equation (5.25) of Subsection 5.2. Obviously, a similar argument shows that the quasi-Newton iteration defined by equation (5.39) can also be interpreted as a Riemannian quasi-Newton method.

We now show that the Newton iteration defined by equation (5.40) does not share this nice property in general. Since $\nabla^2_R\psi(\mathring{\mathbf{o}})$ is asymmetric and is considered to be a linear operator from $\mathcal{T}_{\mathbf{O}}\text{Gr}(p,k)$ to $\mathcal{T}_{\mathbf{O}}\text{Gr}(p,k)$ (more precisely from $null\big(J(\mathbf{r}(\mathbf{a}))\big)^{\perp}$ to $null\big(J(\mathbf{r}(\mathbf{a}))\big)^{\perp}$), it is first convenient to define the $p.k \times p.k.$ symmetric projected Riemannian Hessian as

$$\begin{aligned}
&\nabla^2_R\psi(\mathring{\mathbf{o}})\left(\mathbf{I}_{p.k} - \bar{\mathbf{O}}\bar{\mathbf{O}}^T\right)\\
&= \left(\mathbf{I}_{p.k} - \bar{\mathbf{O}}\bar{\mathbf{O}}^T\right)\left(\mathbf{M}(\mathbf{o})^T\mathbf{M}(\mathbf{o}) - \mathbf{L}(\mathbf{o})^T\mathbf{L}(\mathbf{o}) + \mathbf{U}(\mathbf{o})^T\mathbf{L}(\mathbf{o}) + \mathbf{L}(\mathbf{o})^T\mathbf{U}(\mathbf{o})\right)\left(\mathbf{I}_{p.k} - \bar{\mathbf{O}}\bar{\mathbf{O}}^T\right)\\
&= \mathbf{M}(\mathbf{o})^T\mathbf{M}(\mathbf{o}) - \mathbf{L}(\mathbf{o})^T\mathbf{L}(\mathbf{o}) + \left(\mathbf{I}_{p.k} - \bar{\mathbf{O}}\bar{\mathbf{O}}^T\right)\mathbf{U}(\mathbf{o})^T\mathbf{L}(\mathbf{o}) + \mathbf{L}(\mathbf{o})^T\mathbf{U}(\mathbf{o})\left(\mathbf{I}_{p.k} - \bar{\mathbf{O}}\bar{\mathbf{O}}^T\right)\\
&= \left(\mathbf{I}_{p.k} - \bar{\mathbf{O}}\bar{\mathbf{O}}^T\right)\nabla^2\psi(\mathbf{o})\left(\mathbf{I}_{p.k} - \bar{\mathbf{O}}\bar{\mathbf{O}}^T\right)\ , \qquad (5.46)
\end{aligned}$$

which is now a symmetric linear mapping from $\mathbb{R}^{p.k}$ to $\mathbb{R}^{p.k}$, but this one has, however, the inconvenient to be always rank-deficient with its null space equals to $null(J(\mathbf{r}(\mathbf{o}))) = null(\mathbf{M}(\mathbf{o}))$ in regular cases, e.g., when the hypotheses of Corollary 5.5 are fulfilled and the matrix function $\mathbf{F}(.)$ has full column rank in a neighborhood of $\mathbf{o}$ so that $\nabla^2\psi(\mathbf{o})$ exists.

Next, adding the term $\mathbf{N}\mathbf{N}^T$, defined in equation (5.42), to this symmetric projected Riemannian Hessian matrix will remove its systematic rank degeneracy [82] and provides an alternative method to compute the Riemannian Newton direction vector $d\mathbf{o}_{r-n}$, defined above, as the unique solution vector of the following $p.k \times p.k$ symmetric linear system in regular cases

$$\left(\mathbf{I}_{p.k} - \bar{\mathbf{O}}\bar{\mathbf{O}}^T\right)\nabla^2\psi(\mathbf{o})\left(\mathbf{I}_{p.k} - \bar{\mathbf{O}}\bar{\mathbf{O}}^T\right)d\mathbf{o}_{r-n} = \mathbf{M}(\mathbf{o})^T\mathbf{r}(\mathbf{o})\ .$$

However, we observe that Riemannian Newton iteration based on this damped version of the symmetric projected Riemannian Hessian will differ in general from the damped Newton iteration based on equation (5.40) despite the damping term is the same. This is obvious from equation (5.46) defining the symmetric projected Riemannian Hessian, which shows that the third symmetric term in the symmetric projected Riemannian Hessian differs from the third symmetric term in the formulation of the Euclidean Hessian given in equation (5.33). On the other hand, as the first two-terms of the symmetric projected Riemannian and Euclidean Hessians are identical, we can conclude that the damped quasi-Newton iteration based on a two-term approximation of the Euclidean Hessian $\bar{\mathbf{H}}$, defined in equation (5.41), can still be interpreted as a Riemannian quasi-Newton method operating on the (quotient) Grassmann manifold $\text{Gr}(p,k)$ as claimed above.

To conclude that subsection, we now give an overview of the second-order trust-region method (RTRMC2) proposed by Boumal and Absil [14] to minimize the cost function $\psi \circ h^{-1}(.) = g_\lambda(.)$ on the Grassmann manifold $\text{Gr}(p,k)$, where $g_\lambda(.)$ is defined in equation (3.18), in order to contrast its advantages and drawbacks compared to the quasi-Newton methods just described above. To this end and as above, we assume that $\mathbf{W} \in \mathbb{R}^{p\times n}_{+*}$, that each element $\mathring{\mathbf{O}} \in \text{Gr}(p,k)$ is represented by an element of the compact Stiefel manifold $\mathbf{O} \in \text{St}(p,k)$ and that $\psi \circ h^{-1}(.)$ is a smooth function on $\text{St}(p,k)$ such that both the Riemannian gradient and Hessian of $\psi \circ h^{-1}(.)$ exist $\forall \mathbf{O} \in \text{St}(p,k)$. In this framework, the RTRMC2 method can be both interpreted as a variable projection method and a Riemannian second-order optimization method operating on the quotient Grassmann manifold $\text{Gr}(p,k)$.

First, Boumal and Absil have avoided the direct computation of the vectorized forms of the Euclidean or Riemannian Hessian of $\psi(.)$ as we did in equations (5.33) and (5.44), respectively. Instead, they have derived only the directional derivative in the direction of $\mathbf{D} \in \mathcal{T}_{\mathbf{O}}\text{Gr}(p,k) = \mathcal{H}_{\mathbf{O}}\text{St}(p,k)$ of the Riemannian gradient of $\psi \circ h^{-1}(.)$ at $\mathring{\mathbf{O}} \in \text{Gr}(p,k)$ in a compact formulae, see equation 27 in [14], which is relatively inexpensive and efficient compared to the evaluation of the full Hessians in equations (5.33) and (5.44). In other words and in our notations, this expression uses unvectorized variables and is essentially equivalent to equation (5.43) given above. Furthermore, this compact expression is sufficient for implementing efficiently an (inexact) iterative inner subsolver inside the RTRMC2 method, as we will describe now.

Besides, the RTRMC2 method generates a sequence of iterates $\mathbf{O}^i \in \text{St}(p,k)$ (more precisely of iterates $\mathring{\mathbf{O}}^i \in \text{Gr}(p,k)$) together with a sequence of trust-region radii $\delta^i > 0$. At each iteration i,

a subproblem solver computes a new step $d\mathbf{O}^i \in \mathcal{T}_{\mathbf{O}}\text{Gr}(p,k) = \mathcal{H}_{\mathbf{O}}\text{St}(p,k)$ and the next iterate $\mathbf{O}^{i+1} \in \text{St}(p,k)$ is obtained by performing a retraction $\text{Retraction}_{\mathbf{O}^i}(d\mathbf{O}^i)$, as defined in equation (5.25), to go back to the correct manifold, e.g., $\text{St}(p,k)$ (or more precisely $\text{Gr}(p,k)$ [14]). The RTRMC2 method proceeds by computing $d\mathbf{O}^i$ via minimizing an approximate second-order Taylor expansion of the objective function $\psi \circ h^{-1}(.)$ within a "trust-region" $\|d\mathbf{O}^i\|_F \leq \delta^i$ with adaptively chosen radii δ^i [3][14][11]. Depending on the performance of the inner subsolver, the outer RTRMC2 algorithm decides to accept or reject the proposed step $d\mathbf{O}^i$, and, possibly, decides to increase or reduce the trust-region radii. More precisely, in the inner subsolver, $d\mathbf{O}^i$ is computed to approximately minimize the following quadratic model function $m^i(d\mathbf{O}) : \mathcal{T}_{\mathbf{O}}\text{Gr}(p,k) \mapsto \mathbb{R}$ defined by

$$m^i(d\mathbf{O}) = \psi \circ h^{-1}\left(\mathbf{O}^i\right) + \langle d\mathbf{O}, \nabla_R \psi \circ h^{-1}(\mathring{\mathbf{O}})\rangle_F + \frac{1}{2}\langle d\mathbf{O}, [\nabla_R^2 \psi \circ h^{-1}(\mathring{\mathbf{O}})](d\mathbf{O})\rangle_F \,,$$

under the constraint $\|d\mathbf{O}\|_F \leq \delta^i$ and so that $m^i(\mathbf{0}^{p\times k}) = \psi \circ h^{-1}\left(\mathbf{O}^i\right)$ and $m^i(d\mathbf{O}^i) \approx \psi \circ h^{-1}\left(\mathbf{O}^{i+1}\right)$. The selected inner subsolver is inexact and uses truncated conjugate gradient iterations to attempt to minimize $m^i(.)$ over $\mathcal{T}_{\mathbf{O}}\text{Gr}(p,k)$ [14][11] .

Despite the use of an iterative inner subsolver inside of an outer solver, the RTRMC2 method has a medium per-iteration complexity cost of the order of $\mathcal{O}((p.k)^2)$, while the (quasi-)Newton methods given in equations (5.38), (5.39), (5.40), (5.41) have a much high per-iteration complexity of order at least $\mathcal{O}((p.k)^3)$ for inverting the damped Hessian, giving a clear advantage to the RTRMC2 method in terms of speed and efficiency. On the other hand, the RTRMC2 method has more severe instability issues because its iterative inner solver based on (truncated) conjugate gradient iterations is not always robust when the Hessian is singular or ill-conditioned, which is always the case at first-order stationarity points and near the (local) non-isolated minima of $\psi \circ h^{-1}(.)$ where the Hessian has always vanishingly small, possibly negative eigenvalues [162]. This is verified experimentally in Hong et al. [81] and is explained by the theory developed in [162]. Thus, with respect to accuracy and robustness, the Gauss-Newton, Levenberg-Marquardt and (quasi-)Newton methods developed in the previous and present subsections have a clear advantage as illustrated in the comparative studies of Okatani et al. [150] or Hong et al. [81].

This suggests finally that a fruitful area of future research to get the advantages of the two worlds, for solving efficiently and accurately difficult WLRA problems, may be to use a damped version of the Riemannian Hessian or of its two-term approximation (e.g., with a damping term equivalent to the one defined in equation (5.42) for $\psi(.)$) in the quadratic model function $m^i(.)$ minimized by the inner solver of RTRMC2. This may eventually help to reduce its instability issues near the non-isolated minima of $\psi \circ h^{-1}(.)$, while keeping its lower per-iteration complexity.

6 Implementation of variable projection NLLS methods for solving the WLRA problem

This section is concerned with the formulation of practical and effective second-order algorithms for minimizing the cost function $\psi(.)$, i.e., the description of variable projection (pseudo-)second-order algorithms designed to solve the (VP1) formulation of the WLRA problem using the theoretical results established in the previous sections, especially the systematic rank-deficient nature of the Jacobian matrix $J(\mathbf{r}(\mathbf{a}))$ everywhere in the search space and the exactly rank-deficient nature of the Hessian matrix $\nabla^2\psi(\widehat{\mathbf{a}})$ if $\widehat{\mathbf{a}}$ is a stationary point of $\psi(.)$.

Standard results for the convergence of (quasi-)Newton methods suppose that the cost function is smooth, the target local minimum has a positive definite Hessian and that the algorithm is initialized in a neighborhood of this minimum [45][139]. However, this hypothesis is always violated here for the cost function $\psi(.)$ used in the (VP1) formulation of the WLRA problem as local minima of $\psi(.)$ are never isolated and can even form a continuum or a smooth manifold in some circumstances (e.g., when $\psi(.)$ locally verified the Morse-Bott property as illustrated in the previous section). In

such deteriorated conditions, standard NLLS methods such as the Gauss-Newton or (quasi-)Newton methods would have difficulties [163][199]. The Levenberg-Marquardt and trust-region Gauss-Newton methods can be used without modifications if the Jacobian $J(\mathbf{r}(\mathbf{a}))$ is not of full rank if the Marquardt damping parameter λ or the radius Δ of the trust-region are controlled appropriately to not approach zero during the iterations and, especially, in the neighborhood of a critical point $\widehat{\mathbf{a}}$. However, these methods may have very slow convergence if the Jacobian matrix $J(\mathbf{r}(\mathbf{a}))$ is singular everywhere in the search space and the damping factor is controlled so as not to tend to zero during the iterations, as in this scenario, we loose the quadratic or superlinear convergence speed of these methods near a critical point and the attainable accuracy is also limited [199]. Moreover, most convergence results for all these methods essentially depend on the assumption that, at a solution point $\widehat{\mathbf{a}}$, the Jacobian $J(\mathbf{r}(\widehat{\mathbf{a}}))$ is nonsingular [45][139].

However, some convergence results are also available for the exactly rank-deficient Jacobian case in the neighborhood of a solution point $\widehat{\mathbf{a}}$ for Gauss-Newton- or Levenberg-Marquardt-like methods [5][6][55][111][9][73][195][52][53][196][56][54][34][19] or for a singular Hessian at local minima $\widehat{\mathbf{a}}$ for (quasi-)Newton methods [5][6][159][40][67][59][161][137][36][2][42][163][199][43]. Moreover, in practice, some of these (quasi-)second-order methods preserve their favourable faster convergence compared to first-order methods for non-isolated minima, a surprising result, which has been explained under mild assumptions like the Polyak–Lojasiewicz, Quadratic Growth and Error Bound conditions [199][162][163][43]. Recently, Rebjock and Boumal [163] unified these different results by demonstrated that these three conditions are essentially equivalent to the Morse-Bott property if the objective function is a C^2 function in a neighborhood of the target local minimum like the cost function $\psi(.)$ used in the (VP1) formulation of the WLRA problem under some circumstances (see the previous section for more details).

In the case of the Newton method, one of the most successful approaches in the case of a singular Hessian is the use of bordering techniques introduced by Griewank and co-workers [59][161] and the adaptation of the Newton method described in the previous section to solve the (VP1) form of the WLRA problem falls in this category. On the other hand, if $\mathbf{a}_s$ is the current approximate solution, recall that the Gauss-Newton method for minimizing $\psi(.)$ is based on a linear approximation of the residual function $\mathbf{r}(.)$ from the Taylor's expansion around $\mathbf{a}_s$

$$\mathbf{r}(\mathbf{a}) = \mathbf{r}(\mathbf{a}_s) + J\big(\mathbf{r}(\mathbf{a}_s)\big)(\mathbf{a} - \mathbf{a}_s) + \mathcal{O}(\|\mathbf{a} - \mathbf{a}_s\|_2^2) \; .$$

The Gauss-Newton step $d\mathbf{a}_{gn} = \mathbf{a} - \mathbf{a}_s$ is then the solution of the linear least-squares problem

$$d\mathbf{a}_{gn} = \text{Arg} \min_{d\mathbf{a} \in \mathbb{R}^{p.k}} \frac{1}{2} \|\mathbf{r}(\mathbf{a}_s) + J\big(\mathbf{r}(\mathbf{a}_s)\big) d\mathbf{a}\|_2^2 \; .$$

Since $J(\mathbf{r}(\mathbf{a}_s))$ is always rank deficient, the solution of this linear least-squares problem is not unique. However, a natural choice is to take $d\mathbf{a}_{gn}$ to be the unique minimum 2-norm solution, which is given by

$$d\mathbf{a}_{gn} = -J\big(\mathbf{r}(\mathbf{a}_s)\big)^+ \mathbf{r}(\mathbf{a}_s) \; ,$$

since the linearization argument used to derive the Gauss-Newton iteration is only valid in a "small" neighborhood of $\mathbf{a}_s$. This leads to the generalized (Gauss-)Newton or Ben-Israel iterative method

$$\mathbf{a}_{s+1} = \mathbf{a}_s + d\mathbf{a}_{gn} = \mathbf{a}_s - J\big(\mathbf{r}(\mathbf{a}_s)\big)^+ \mathbf{r}(\mathbf{a}_s) \; .$$

The local and global convergence properties of this algorithm when $J(\mathbf{r}(\mathbf{a}))$ does not have full rank in the neighborhood of a solution point $\widehat{\mathbf{a}}$ have been first investigated by Ben-Israel [5][6] and then by several authors afterward [9][73][137][36][42][199]. On assumptions like that $J(\mathbf{r}(\mathbf{a}))^+$ is Lipschitz-continuous, the Jacobian $J(\mathbf{r}(\mathbf{a}))$ is of constant rank in some neighborhood of $\widehat{\mathbf{a}}$ or the cost function $\psi(.)$ verifies a Morse-Bott-like property in a neighborhood of $\widehat{\mathbf{a}}$, they were able to show the convergence of this generalized (Gauss-)Newton method to a stationary point of $\psi(.)$. In practice, it is necessary to include some strategy to estimate the numerical rank of $J(\mathbf{r}(\mathbf{a}_s))$

and the assigned rank can have a decisive influence on the success of the method. Thus, a QR- or COD-factorization, or even a SVD-decomposition, of the matrix $J(\mathbf{r}(\mathbf{a}_s))$ are natural tools for the Ben-Israel iteration. This implies that the Ben-Israel method is relatively expensive. An alternative to this problem is to use a Tikhonov regularized version of the Ben-Israel iteration as in Levenberg-Marquardt or trust-region methods (see Subsection 5.1), but the choice of the regularization parameter is also a challenge in the rank-deficient case. Fortunately, the specific properties of the Jacobian $J(\mathbf{r}(\mathbf{a}))$ derived in Subsection 5.2 can be used for this purpose and also to reduce drastically the cost of the Ben-Israel iterative method or its regularized version as we will illustrate below.

Alternatively, Menzel [121] has proposed to reformulate the exactly rank deficient problem as an auxiliary least-squares problem of higher dimension which can be shown to be a well-posed one if the rank deficiency of $J(\mathbf{r}(\mathbf{a}))$ is small. Moreover, he was able to prove that for arbitrary rank deficiency in the consistent case $\psi(\widehat{\mathbf{a}}) = 0$, the Gauss-Newton sequence for his auxiliary least-squares problem converges at least superlinearly to $\widehat{\mathbf{a}}$. However, his technique, which expands considerably the size of the problem, is useful in practice only for small dimensions and small rank deficiency values and cannot be applied to our WLRA problem where both the dimensions and the rank deficiency values may be high.

More recently, Eriksson and Wedin [52][53][54] also considered the NLLS problem with an exactly rank-deficient Jacobian matrix for all points in a neighborhood of a local solution. As this situation corresponds exactly to the minimization of the cost function $\psi(.)$ in the (VP1) formulation of the WLRA problem, as demonstrated in the previous sections, we discuss now their method in some details. They suggested the following reformulation (in our notations) of the (VP1) variant of the WLRA problem in order to obtain a uniquely defined solution

$$\widehat{\mathbf{a}} = \begin{cases} \text{Arg}\min_{\mathbf{a}\in\mathbb{R}^{p.k}} \ \frac{1}{2}\|\mathbf{a}\|_2^2 \\ \text{s.t. Arg}\min_{\mathbf{a}\in\mathbb{R}^{p.k}} \ \frac{1}{2}\psi(\mathbf{a}) = \frac{1}{2}\|\mathbf{r}(\mathbf{a})\|_2^2 \end{cases} ,$$

and they proposed two different iterative methods to solve this problem: a Gauss-Newton method and a Tikhonov regularization method. In words, what Eriksson and Wedin [52][53][54] suggested is to actually obtain the minimum Euclidean norm solution to the problem of minimizing $\psi(.)$. A similar approach was already mentioned by Boggs [9]. Using a Taylor series through two terms around $\mathbf{a}_s$, we get the linearized version of this problem and the following iterative method

$$\mathbf{a}_{s+1} = \begin{cases} \text{Arg}\min_{\mathbf{a}\in\mathbb{R}^{p.k}} \ \frac{1}{2}\|\mathbf{a}\|_2^2 \\ \text{s.t. Arg}\min_{\mathbf{a}\in\mathbb{R}^{p.k}} \ \frac{1}{2}\|\mathbf{r}(\mathbf{a}_s) + J\big(\mathbf{r}(\mathbf{a}_s)\big)(\mathbf{a}-\mathbf{a}_s)\|_2^2 \end{cases} .$$

Following Pes and Rodriguez [153][154], we will denote this as the Minimal-Norm Gauss-Newton (MNGN) method. It must be noted that the Ben-Israel iterative method has no predisposition toward the minimum 2-norm solution of minimizing $\psi(.)$ in the sense that any limit point generated by the Ben-Israel iteration is a least-squares solution, but not in general the minimum 2-norm solution and, consequently, the Ben-Israel and MNGN solutions will differ in general [9][54][34][153]. Since

$$\mathbf{r}(\mathbf{a}_s) + J\big(\mathbf{r}(\mathbf{a}_s)\big)(\mathbf{a}-\mathbf{a}_s) = \Big(\mathbf{r}(\mathbf{a}_s) - J\big(\mathbf{r}(\mathbf{a}_s)\big)\mathbf{a}_s\Big) + J\big(\mathbf{r}(\mathbf{a}_s)\big)\mathbf{a} ,$$

the minimum 2-norm solution $\mathbf{a}_{s+1}$ of the problem

$$\min_{\mathbf{a}\in\mathbb{R}^{p.k}} \ \frac{1}{2}\|\mathbf{r}(\mathbf{a}_s) + J\big(\mathbf{r}(\mathbf{a}_s)\big)(\mathbf{a}-\mathbf{a}_s)\|_2^2$$

is

$$\begin{aligned}\mathbf{a}_{s+1} &= -J\big(\mathbf{r}(\mathbf{a}_s)\big)^+\Big(\mathbf{r}(\mathbf{a}_s) - J\big(\mathbf{r}(\mathbf{a}_s)\big)\mathbf{a}_s\Big)\\ &= -J\big(\mathbf{r}(\mathbf{a}_s)\big)^+\mathbf{r}(\mathbf{a}_s) + J\big(\mathbf{r}(\mathbf{a}_s)\big)^+ J\big(\mathbf{r}(\mathbf{a}_s)\big)\mathbf{a}_s\\ &= d\mathbf{a}_{gn} + \mathbf{P}_{J(\mathbf{r}(\mathbf{a}_s))^T}\mathbf{a}_s\\ &= \mathbf{P}_{J(\mathbf{r}(\mathbf{a}_s))^T}\big(\mathbf{a}_s + d\mathbf{a}_{gn}\big)\ ,\end{aligned}$$

where $\mathbf{P}_{J(\mathbf{r}(\mathbf{a}_s))^T}$ is the orthogonal projector onto $\mathit{ran}(J(\mathbf{r}(\mathbf{a}_s))^T) = \mathit{null}(J(\mathbf{r}(\mathbf{a}_s)))^\perp$, e.g., onto the row space of $J(\mathbf{r}(\mathbf{a}_s))$ and the last equality results from equation (2.9). Thus, to ensure computation of the minimal 2-norm solution, at the s^{th} iteration of the MNGN method, the $standard$ Gauss-Newton approximation $\mathbf{a}_s + d\mathbf{a}_{gn}$, computed by the Ben-Israel method, is (orthogonally) projected onto the orthogonal of the null space of $J(\mathbf{r}(\mathbf{a}_s))$, e.g., $\mathit{null}(J(\mathbf{r}(\mathbf{a}_s)))^\perp$. Alternatively, the MNGN iteration proposed by Eriksson and Wedin [53][54] can be written as

$$\mathbf{a}_{s+1} = \mathbf{a}_s + d\mathbf{a}_{mngn}\ ,$$

with the MNGN step computed as

$$\begin{aligned}d\mathbf{a}_{mngn} &= -J\big(\mathbf{r}(\mathbf{a}_s)\big)^+\mathbf{r}(\mathbf{a}_s) + \mathbf{P}_{J(\mathbf{r}(\mathbf{a}_s))^T}\mathbf{a}_s - \mathbf{a}_s\\ &= d\mathbf{a}_{gn} - \mathbf{P}^\perp_{J(\mathbf{r}(\mathbf{a}_s))^T}\mathbf{a}_s\\ &= d\mathbf{a}_{gn} - \mathbf{P}_{\mathit{null}(J(\mathbf{r}(\mathbf{a}_s)))}\mathbf{a}_s\ ,\end{aligned}$$

and where $\mathbf{P}_{\mathit{null}(J(\mathbf{r}(\mathbf{a}_s)))}$ is the orthogonal projector onto the null space of $J(\mathbf{r}(\mathbf{a}_s))$. See Subsection 2.1 for details how the orthogonal projectors $\mathbf{P}_{\mathit{null}(J(\mathbf{r}(\mathbf{a}_s)))}$ and $\mathbf{P}_{J(\mathbf{r}(\mathbf{a}_s))^T}$ can be represented using the SVD of $J(\mathbf{r}(\mathbf{a}_s))$ or much more efficiently using a COD of this matrix. Furthermore, note that these orthogonal projectors can be easily computed using the orthonormal bases of $\mathit{null}(J(\mathbf{r}(\mathbf{a}_s)))$ and its orthogonal complement identified in Corollary 5.6 (assuming that the rank of $J(\mathbf{r}(\mathbf{a}_s))$ is equal to $r = k.(p-k)$).

In order to ensure global convergence of the MNGN method, Campbell et al. [34] and Pes and Rodriguez [154] have also considered the inclusion of relaxation (or damping) parameters in the MNGN method, giving the iterative methods

$$\mathbf{a}_{s+1} = \mathbf{a}_s + \alpha_s d\mathbf{a}_{gn} - \mathbf{P}_{\mathit{null}(J(\mathbf{r}(\mathbf{a}_s)))}\mathbf{a}_s \qquad \text{(MNGN1)}$$

or

$$\mathbf{a}_{s+1} = \mathbf{a}_s + d\mathbf{a}_{gn} - \beta_s\mathbf{P}_{\mathit{null}(J(\mathbf{r}(\mathbf{a}_s)))}\mathbf{a}_s \qquad \text{(MNGN2)}$$

and also

$$\mathbf{a}_{s+1} = \mathbf{a}_s + \alpha_s d\mathbf{a}_{gn} - \beta_s\mathbf{P}_{\mathit{null}(J(\mathbf{r}(\mathbf{a}_s)))}\mathbf{a}_s\ , \qquad \text{(MNGN3)}$$

where α_s and β_s are step length parameters, which can be determined by a line search and specific strategies described in [34][154].

As suggested by Eriksson and Wedin [52][53][54] and Pes and Rodriguez [153], a good approximate Minimal-Norm Levenberg-Marquardt step, $d\mathbf{a}_{mnlm}$, can also be found by solving the regularized linear least-squares problem

$$d\mathbf{a}_{mnlm} = \mathrm{Arg}\min_{d\mathbf{a}\in\mathbb{R}^{p.k}} \frac{1}{2}\left\| \begin{bmatrix}\mathbf{r}(\mathbf{a}_s)\\ \mu\mathbf{a}_s\end{bmatrix} + \begin{bmatrix}J\big(\mathbf{r}(\mathbf{a}_s)\big)\\ \mu\mathbf{I}_{k.p}\end{bmatrix} d\mathbf{a}\right\|_2^2\ , \qquad (6.1)$$

with a sufficiently small Tikhonov regularization parameter μ since this regularized problem has an unique solution

$$\begin{aligned}d\mathbf{a}_{mnlm} = &-\Big(J\big(\mathbf{r}(\mathbf{a}_s)\big)^T J\big(\mathbf{r}(\mathbf{a}_s)\big) + \mu^2\mathbf{I}_{k.p}\Big)^{-1} J\big(\mathbf{r}(\mathbf{a}_s)\big)^T\mathbf{r}(\mathbf{a}_s)\\ &-\Big(J\big(\mathbf{r}(\mathbf{a}_s)\big)^T J\big(\mathbf{r}(\mathbf{a}_s)\big) + \mu^2\mathbf{I}_{k.p}\Big)^{-1}\mu^2\mathbf{a}_s\end{aligned}$$

and

$$\lim_{\mu\to 0}\left(J\big(\mathbf{r}(\mathbf{a}_s)\big)^T J\big(\mathbf{r}(\mathbf{a}_s)\big)+\mu^2\mathbf{I}_{k.p}\right)^{-1} J\big(\mathbf{r}(\mathbf{a}_s)\big)^T = J\big(\mathbf{r}(\mathbf{a}_s)\big)^+ ,$$

$$\lim_{\mu\to 0}\left(J\big(\mathbf{r}(\mathbf{a}_s)\big)^T J\big(\mathbf{r}(\mathbf{a}_s)\big)+\mu^2\mathbf{I}_{k.p}\right)^{-1} \mu^2 = \mathbf{P}_{null(J(\mathbf{r}(\mathbf{a}_s)))} ,$$

as demonstrated in [70]. The Tikhonov method proposed by Eriksson and Wedin [54], and also considered by Pes and Rodriguez [153], then approximately solves a sequence of Tikhonov regularized NLLS problems for a sequence of decreasing regularization parameter μ_t where the index t does not necessarily equal the iteration index s. The approximate solution of one regularized NLLS problem with Tikhonov parameter μ_t is taken as the starting point for the next regularized problem with Tikhonov parameter $\mu_{t+1} < \mu_t$. Eriksson and Wedin [54] proved that both the MNGN and Tikhonov methods converge to a local minimum 2-norm solution if the iterations are started in a neighborhood of the solution.

These two methods are directly applicable to an exactly rank deficient problem such as the (VP1) formulation of the WLRA problem. However, we note that the different damped variants of the MNGN method may fail to converge in many cases because projecting the GN solution orthogonally to the null space of $J(\mathbf{r}(\mathbf{a}_s))$ may cause the residual to increase during the iterations for the MNGN1 variant or because the MNGN2 variant is equivalent to the application of the undamped Gauss-Newton method, whose convergence is not theoretically guaranteed [154][34]. Note further that in the case of the MNGN3 variant, convergence can be generally obtained if α_s and β_s are suitably chosen, but in that case the method does not converge to the minimum 2-norm solution unless $\beta_s = 1$ for s close to convergence [154]. Moreover, it is not of interest, nor natural to find the minimum 2-norm least-squares solution of our WLRA problem due to its very special structure, separability properties and over-parameterization (e.g., because it is an optimization problem on manifolds or subspaces). Additionally, the MNGN and Tikhonov methods involve the extra-computations of the terms

$$\mathbf{P}_{null(J(\mathbf{r}(\mathbf{a}_s)))}\mathbf{a}_s \quad \text{and} \quad \big(J\big((\mathbf{r}(\mathbf{a}_s)\big()^T J\big((\mathbf{r}(\mathbf{a}_s)\big()+\mu^2\mathbf{I}_{k.p}\big)^{-1}\mu^2\mathbf{a}_s ,$$

in each iteration, respectively. This suggests that simple Ben-Israel or regularized Tikhonov methods will be more appropriate to minimize $\psi(.)$.

With these considerations, we are now in position to give a full formal description of different variations of the Gauss-Newton, Levenberg-Marquardt and (quasi-)Newton algorithms which may be used to minimize $\psi(.)$ in practice.

6.1 Variable projection Gauss-Newton algorithms

Using similar notations and definitions as in previous sections, an outline of the variable projection Gauss-Newton algorithms is as follows:

Gauss-Newton algorithms 1.

Choose starting matrix $\mathbf{A}_1 \in \mathbb{R}^{p\times k}$, $\varepsilon_1, \varepsilon_2, \varepsilon_3 \in \mathbb{R}_{+*}$ and $i_{max}, j_{max} \in \mathbb{N}_*$, appropriately

For $i = 1, 2, \ldots$ **until convergence do**

(0) Optionally, compute a QRCP of $\mathbf{A}_i$ (see equation (2.15)) to determine $k_i = \mathit{rank}(\mathbf{A}_i)$ and an orthonormal basis of $\mathit{ran}(\mathbf{A}_i)$:

$$\mathbf{Q}_i\mathbf{A}_i\mathbf{P}_i = \begin{bmatrix}\mathbf{R}_i & \mathbf{S}_i \\ \mathbf{0}^{(p-k_i)\times k_i} & \mathbf{0}^{(p-k_i)\times(k-k_i)}\end{bmatrix},$$

where $\mathbf{Q}_i$ is an $p \times p$ orthogonal matrix, $\mathbf{P}_i$ is an $k \times k$ permutation matrix, $\mathbf{R}_i$ is an $k_i \times k_i$ nonsingular upper triangular matrix (with diagonal elements of decreasing absolute magnitude) and $\mathbf{S}_i$ an $k_i \times (k - k_i)$ full matrix, which is vacuous if $k_i = k$. If $k_i < k$, complete the

orthonormal basis of $ran(\mathbf{A}_i)$ with $k - k_i$ orthonormal vectors by using the $p \times p$ orthogonal matrix $\mathbf{Q}_i$ computed implicitly during the QRCP of $\mathbf{A}_i$.

In other words, in all cases, compute a $p \times k$ matrix $\mathbf{O}_i$ with orthonormal columns as the first k columns of $\mathbf{Q}_i$ (i.e., such that $ran(\mathbf{A}_i) \subset ran(\mathbf{O}_i)$ if $k_i < k$ and $ran(\mathbf{A}_i) = ran(\mathbf{O}_i)$ if $k_i = k$) and set

$$\mathbf{A}_i = \mathbf{O}_i \ .$$

This optional orthogonalization step is a safe-guard as the condition $k_i = k$ is a necessary condition for the differentiability of $\psi(.)$ at a point $\mathbf{A}_i$ and is also useful to limit the occurrence of overflows and underflows in the next steps by enforcing that the matrix variable $\mathbf{A}_i \in \mathbb{O}^{p \times k}$.

(1) Determine (implicitly) the block diagonal matrix

$$\mathbf{F}(\mathbf{a}_i) = diag\big(vec(\sqrt{\mathbf{W}})\big)\big(\mathbf{I}_n \otimes \mathbf{A}_i\big) \ ,$$

where $\mathbf{a}_i = vec(\mathbf{A}_i^T)$.

(2) Compute (implicitly) a QRCP of $\mathbf{F}(\mathbf{a}_i)$ to determine $\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a}_i)}$ and $\mathbf{F}(\mathbf{a}_i)^-$ (see equations (2.18) and (2.19)) or, alternatively, a COD of $\mathbf{F}(\mathbf{a}_i)$ to determine $\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a}_i)}$ and $\mathbf{F}(\mathbf{a}_i)^+$ (see equations (2.18) and (2.21)).

Note also that $\mathbf{F}(\mathbf{a}_i)^- = \mathbf{F}(\mathbf{a}_i)^+$ if $\mathbf{F}(\mathbf{a}_i)$ is of full column rank and that $\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a}_i)}$, $\mathbf{F}(\mathbf{a}_i)^-$ and $\mathbf{F}(\mathbf{a}_i)^+$ are also block diagonal matrices.

(3) Solve the block diagonal linear least-squares problem

$$\mathbf{b}_i = \text{Arg}\min_{\mathbf{b} \in \mathbb{R}^{k.n}} \ \|\mathbf{x} - \mathbf{F}(\mathbf{a}_i)\mathbf{b}\|_2^2 \ ,$$

e.g., compute

$$\mathbf{b}_i = \begin{cases} \mathbf{F}(\mathbf{a}_i)^-\mathbf{x} & \{\text{if a QRCP of } \mathbf{F}(\mathbf{a}_i) \text{ is used in step } \textbf{(2)}\} \\ \mathbf{F}(\mathbf{a}_i)^+\mathbf{x} & \{\text{if a COD of } \mathbf{F}(\mathbf{a}_i) \text{ is used in step } \textbf{(2)}\} \end{cases} \ .$$

(4) Determine:

$$\mathbf{r}(\mathbf{a}_i) = \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a}_i)}\mathbf{x} \ \{\text{current residual vector}\}$$

$$\psi(\mathbf{a}_i) = \tfrac{1}{2}\|\mathbf{r}(\mathbf{a}_i)\|_2^2 \ \{\text{current value of the cost function }\}$$

$$\nabla\psi(\mathbf{a}_i) = \mathbf{G}(\mathbf{b}_i)^T\mathbf{G}(\mathbf{b}_i)\mathbf{a}_i - \mathbf{G}(\mathbf{b}_i)^T\mathbf{z} \ \{\text{see Theorems 4.3 and 5.7}\}$$

Note that the steps **(1)** to **(4)** above can be very easily parallelized using the block diagonal structure of $\mathbf{F}(\mathbf{a}_i)$.

(5) Check for convergence. Relevant convergence criteria in the algorithms are of the form:

- $\|\nabla\psi(\mathbf{a}_i)\|_2 \leq \varepsilon_1$
- $\|\mathbf{a}_i - \mathbf{a}_{i-1}\|_2 \leq \varepsilon_2(\varepsilon_2 + \|\mathbf{a}_i\|_2)$ $\{\text{if } i \neq 1\}$

  If step **(0)** is used, this last convergence condition can be simplified as:

  $\|\mathbf{a}_i - \mathbf{a}_{i-1}\|_2 \leq \varepsilon_2\|\mathbf{a}_i\|_2 = \varepsilon_2\sqrt{k}$
- $|\psi(\mathbf{a}_{i-1}) - \psi(\mathbf{a}_i)| \leq \varepsilon_3(\varepsilon_3 + \psi(\mathbf{a}_i))$ $\{\text{if } i \neq 1\}$
- $i \geq i_{max}$ {e.g., give up if the number of iterations is too large}

where $\varepsilon_1, \varepsilon_2, \varepsilon_3$ and i_{max} are constants chosen by the user.

Exit if convergence. **Otherwise, go to** step **(6)**

(6) Compute the Gauss-Newton correction vector $d\mathbf{a}_{gn}$ as the minimum 2-norm solution of one of the following linear least-squares problems:

Golub-Pereyra step: Golub and Pereyra [63], Ruhe and Wedin [166]

$$\begin{aligned} d\mathbf{a}_{gp-gn} &= \big(\mathbf{M}(\mathbf{a}_i) + \mathbf{L}(\mathbf{a}_i)\big)^+ \mathbf{r}(\mathbf{a}_i) \\ &= \begin{cases} \text{Arg min}_{d\mathbf{a}\in\mathbb{R}^{p.k}} \ \|d\mathbf{a}\|_2^2 \\ \text{s.t. Arg min}_{d\mathbf{a}\in\mathbb{R}^{p.k}} \ \|\mathbf{r}(\mathbf{a}_i) - \big(\mathbf{M}(\mathbf{a}_i) + \mathbf{L}(\mathbf{a}_i)\big)d\mathbf{a}\|_2^2 \end{cases} \end{aligned}$$

Kaufman step: Kaufman [96], Ruhe and Wedin [166]

$$\begin{aligned} d\mathbf{a}_{k-gn} &= \mathbf{M}(\mathbf{a}_i)^+ \mathbf{r}(\mathbf{a}_i) \\ &= \begin{cases} \text{Arg min}_{d\mathbf{a}\in\mathbb{R}^{p.k}} \ \|d\mathbf{a}\|_2^2 \\ \text{s.t. Arg min}_{d\mathbf{a}\in\mathbb{R}^{p.k}} \ \|\mathbf{r}(\mathbf{a}_i) - \mathbf{M}(\mathbf{a}_i)d\mathbf{a}\|_2^2 \end{cases} \end{aligned}$$

Gauss-Seidel step: Ruhe and Wedin [166]

$$\begin{aligned} d\mathbf{a}_{gs-gn} &= \big(\mathbf{K}_{(n,p)}\mathbf{G}(\mathbf{b}_i)\big)^+ \mathbf{r}(\mathbf{a}_i) \\ &= \begin{cases} \text{Arg min}_{d\mathbf{a}\in\mathbb{R}^{p.k}} \ \|d\mathbf{a}\|_2^2 \\ \text{s.t. Arg min}_{d\mathbf{a}\in\mathbb{R}^{p.k}} \ \|\mathbf{r}(\mathbf{a}_i) - \mathbf{K}_{(n,p)}\mathbf{G}(\mathbf{b}_i)d\mathbf{a}\|_2^2 \end{cases} \end{aligned}$$

(7) Increment $\mathbf{a}_i = vec(\mathbf{A}_i^T)$, e.g., compute $\mathbf{a}_{i+1} = vec(\mathbf{A}_{i+1}^T)$ such that $\psi(\mathbf{a}_{i+1}) < \psi(\mathbf{a}_i)$ in order to obtain global convergence.

(7.1) To this end, first compute

$$\mathbf{a}_{i+1} = \mathbf{a}_i + d\mathbf{a}_{gn}$$

and if a Golub-Pereyra or Kaufman step is used in step **(6)**

$$\psi(\mathbf{a}_{i+1}) = \tfrac{1}{2}\|\mathbf{r}(\mathbf{a}_{i+1})\|_2^2 = \tfrac{1}{2}\|\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a}_{i+1})}\mathbf{x}\|_2^2 \,,$$

using (implicitly) a QRCP of the block diagonal matrix $\mathbf{F}(\mathbf{a}_{i+1})$.

(7.2) **If** a Golub-Pereyra or Kaufman step is used in step **(6)** and $\psi(\mathbf{a}_{i+1}) > \psi(\mathbf{a}_i)$ **then** recompute $\mathbf{a}_{i+1}$ by one of the following methods:

Gauss-Seidel: $\mathbf{a}_{i+1} = \mathbf{a}_i + d\mathbf{a}_{gs-gn}$ where $d\mathbf{a}_{gs-gn}$ is a Gauss-Seidel step [166] defined as

$$\begin{aligned} d\mathbf{a}_{gs-gn} &= \big(\mathbf{K}_{(n,p)}\mathbf{G}(\mathbf{b}_i)\big)^+ \mathbf{r}(\mathbf{a}_i) \\ &= \begin{cases} \text{Arg min}_{d\mathbf{a}\in\mathbb{R}^{p.k}} \ \|d\mathbf{a}\|_2^2 \\ \text{s.t. Arg min}_{d\mathbf{a}\in\mathbb{R}^{p.k}} \ \|\mathbf{r}(\mathbf{a}_i) - \mathbf{K}_{(n,p)}\mathbf{G}(\mathbf{b}_i)d\mathbf{a}\|_2^2 \end{cases} \end{aligned}$$

Block alternating least-squares:

$$\begin{aligned} \mathbf{a}_{i+1} &= \mathbf{G}(\mathbf{b}_i)^+ \mathbf{z} \\ &= \begin{cases} \text{Arg min}_{\mathbf{a}\in\mathbb{R}^{p.k}} \ \|\mathbf{a}\|_2^2 \\ \text{s.t. Arg min}_{\mathbf{a}\in\mathbb{R}^{p.k}} \ \|\mathbf{z} - \mathbf{G}(\mathbf{b}_i)\mathbf{a}\|_2^2 \end{cases} \end{aligned}$$

Line search:

$$\mathbf{a}_{i+1} = \mathbf{a}_i + \alpha_i d\mathbf{a}_{gn} \ ,$$

where $\alpha_i < 1$ is determined by a line search to make the algorithm a descent method (i.e., such that $\psi(\mathbf{a}_{i+1}) < \psi(\mathbf{a}_i)$). This is always possible as the correction vector $d\mathbf{a}_{gn}$ is in a descent direction for $\psi(.)$ if $\|\nabla\psi(\mathbf{a}_i)\|_2 \neq 0$, see Corollary 5.7.

As an illustration, a simple, but still efficient, strategy is to first shorten the correction step to half the Gauss-Newton length, compute the new trial value for $\psi(\mathbf{a}_{i+1})$ and, if it is still worse, continue to reduce the step until we get a step short enough such that $\psi(\mathbf{a}_{i+1}) < \psi(\mathbf{a}_i)$. The following loop incorporates this simple step-shortening algorithm:

For $j = 1, 2, \ldots$ **while**$\big(\psi(\mathbf{a}_{i+1}) > \psi(\mathbf{a}_i)\big)$

$d\mathbf{a}_{gn} = \frac{1}{2} d\mathbf{a}_{gn}$

$\mathbf{a}_{i+1} = \mathbf{a}_i + d\mathbf{a}_{gn}$

$\psi(\mathbf{a}_{i+1}) = \frac{1}{2}\|\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a}_{i+1})}\mathbf{x}\|_2^2$ {using a QRCP of the matrix $\mathbf{F}(\mathbf{a}_{i+1})$}

If $j > j_{max}$ **exit** {e.g., give up if the number of iterations is too large}

End do

End do

For the convenience of the reader, we first recall the shape and definition of the vector and matrix variables used in these Gauss-Newton algorithms. We have: $\mathbf{X} \in \mathbb{R}^{p\times n}$, $\mathbf{W} \in \mathbb{R}^{p\times n}_{+}$, $\mathbf{A}_i \in \mathbb{R}^{p\times k}$, $\mathbf{B}_i \in \mathbb{R}^{k\times n}$, $\mathbf{O}_i \in \mathbb{O}^{p\times k}$ and

$$\begin{aligned}
\mathbf{x} &= vec(\sqrt{\mathbf{W}} \odot \mathbf{X}) \ , \\
\mathbf{z} &= vec\big((\sqrt{\mathbf{W}} \odot \mathbf{X})^T\big) \ , \\
\mathbf{a}_i &= vec(\mathbf{A}_i^T) \ , \\
\mathbf{F}(\mathbf{a}_i) &= diag\big(vec(\sqrt{\mathbf{W}})\big)\big(\mathbf{I}_n \otimes \mathbf{A}_i\big) \ , \\
\mathbf{b}_i = vec(\mathbf{B}_i) &= \begin{cases} \mathbf{F}(\mathbf{a}_i)^{-}\mathbf{x} & \{\text{if a QRCP is used in step } \textbf{(2)}\} \\ \mathbf{F}(\mathbf{a}_i)^{+}\mathbf{x} & \{\text{if a COD is used in step } \textbf{(2)}\} \end{cases} \ , \\
\mathbf{G}(\mathbf{b}_i) &= diag\big(vec(\sqrt{\mathbf{W}}^T)\big)\big(\mathbf{I}_p \otimes \mathbf{B}_i^T\big) \ , \\
\mathbf{U}(\mathbf{a}_i) &= diag(vec(\sqrt{\mathbf{W}}))(\mathbf{B}_i^T \otimes \mathbf{I}_p)\mathbf{K}_{(k,p)} \ , \\
\mathbf{V}(\mathbf{a}_i) &= \big((\mathbf{W} \odot P_\Omega(\mathbf{X} - \mathbf{A}_i\mathbf{B}_i))^T \otimes \mathbf{I}_k\big)\{\text{where } P_\Omega(.) \text{ is defined in equation (3.17)}\} \ , \\
\mathbf{M}(\mathbf{a}_i) &= \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a}_i)}\mathbf{U}(\mathbf{a}_i) = \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a}_i)}\mathbf{K}_{(n,p)}\mathbf{G}(\mathbf{b}_i) \ , \\
\mathbf{L}(\mathbf{a}_i) &= \begin{cases} \mathbf{F}(\mathbf{a}_i)^{-T}\mathbf{V}(\mathbf{a}_i) & \{\text{if a QRCP is used in step } \textbf{(2)}\} \\ \mathbf{F}(\mathbf{a}_i)^{+T}\mathbf{V}(\mathbf{a}_i) & \{\text{if a COD is used in step } \textbf{(2)}\} \end{cases} \ .
\end{aligned}$$

In these Gauss-Newton algorithms, the iterations are terminated either when one or several of the convergence criteria listed in step **(5)** are satisfied, or when the iteration count exceeds the predetermined number i_{max}.

The Golub-Pereyra step $d\mathbf{a}_{gp-gn}$ corresponds exactly to the standard Gauss-Newton step $d\mathbf{a}_{gn}$ applied to the minimization of the variable projection functional $\psi(.)$, which is introduced in Subsection 5.1.

The following is a brief review of the basic ideas underlying the simplification introduced by the Kaufman step $d\mathbf{a}_{k-gn}$ in the Gauss-Newton algorithms (1) and also in the Marquardt-Levenberg algorithms (2) described in the next subsection. As stated in Subsection 5.1, the Gauss-Newton

method may be interpreted as a variation of Newton's method to find a zero of the gradient of $\psi(.)$. More precisely, dropping the iteration index of the algorithm in order to simplify the notation, the Gauss-Newton algorithm approximates the Hessian matrix $\mathbf{H} = \nabla^2\psi(\mathbf{a})$ with the cross-product matrix $J(\mathbf{r}(\mathbf{a}))^T J(\mathbf{r}(\mathbf{a}))$ at each iteration based on the assumption that the components of the residual vector $|\mathbf{r}_l(\mathbf{a})|$ are small near a solution and using the fact that the second term of the Hessian matrix $\nabla^2\psi(\mathbf{a})$ given by

$$\mathbf{S} = \sum_{l=1}^{n.p} \mathbf{r}_l(\mathbf{a})\nabla^2\mathbf{r}_l(\mathbf{a})$$

is of order $\mathcal{O}(\|\mathbf{r}(\mathbf{a})\|_2)$ (see equation (5.32) for the explicit form of $\mathbf{S}$ in our WLRA context). Now, the Gauss-Newton approximation of the Hessian matrix is

$$\nabla^2\psi(\mathbf{a}) \approx J(\mathbf{r}(\mathbf{a}))^T J(\mathbf{r}(\mathbf{a}) = \mathbf{M}(\mathbf{a})^T\mathbf{M}(\mathbf{a}) + \mathbf{L}(\mathbf{a})^T\mathbf{L}(\mathbf{a}) \ ,$$

as demonstrated in Subsection 5.3 (see equation (5.29) for details). Hence, the term $-\mathbf{L}(\mathbf{a})$ does not contribute to the gradient $\nabla\psi(\mathbf{a})$ (see Theorem 5.7) and changes the Gauss-Newton approximation of the Hessian only by the term $\mathbf{L}(\mathbf{a})^T\mathbf{L}(\mathbf{a})$, which is of order $\mathcal{O}(\|\mathbf{r}(\mathbf{a})\|_2^2)$, see equation (5.19). If $\|\mathbf{r}(\mathbf{a})\|_2$ is small then this term is smaller than the second term of $\nabla^2\psi(\mathbf{a})$, $\mathbf{S}$, which is of order $\mathcal{O}(\|\mathbf{r}(\mathbf{a})\|_2)$, and which is already dropped in the Gauss-Newton and Levenberg-Marquardt methods. Following the Gauss-Newton philosophy, it is then natural to drop this term here too. In addition, since the approximation of the Hessian matrix by its first two symmetric exact terms is given by

$$\nabla^2\psi(\mathbf{a}) \approx \mathbf{M}(\mathbf{a})^T\mathbf{M}(\mathbf{a}) - \mathbf{L}(\mathbf{a})^T\mathbf{L}(\mathbf{a}) \ ,$$

see equation (5.34), we may expect that approximating the Hessian matrix by the cross-product matrix $\mathbf{M}(\mathbf{a})^T\mathbf{M}(\mathbf{a})$ (or equivalently the Jacobian matrix by $-\mathbf{M}(\mathbf{a})$) can perform even better than the Gauss-Newton approximation of the Hessian matrix, see Subsection 5.3 and equation (5.33) for details. This leads to the following linear least-squares problems for computing the simplified Kaufman correction step in the Gauss-Newton and Levenberg-Marquardt methods, respectively,

$$\begin{aligned} d\mathbf{a}_{k-gn} &= \text{Arg}\min_{d\mathbf{a}\in\mathbb{R}^{p.k}} \frac{1}{2}\|\mathbf{r}(\mathbf{a}) - \mathbf{M}(\mathbf{a})d\mathbf{a}\|_2^2 \ , \\ d\mathbf{a}_{k-lm} &= \text{Arg}\min_{d\mathbf{a}\in\mathbb{R}^{p.k}} \frac{1}{2}\|\mathbf{r}(\mathbf{a}) - \mathbf{M}(\mathbf{a})d\mathbf{a}\|_2^2 + \frac{\lambda}{2}\left\|\mathbf{D}d\mathbf{a}\right\|_2^2 \ , \end{aligned}$$

where λ is the damping Marquardt parameter and $\mathbf{D}$ is a diagonal scaling matrix of dimension $k.p$, or their variants including a linear constraint to deal with the singularity of the $\mathbf{M}(\mathbf{a})$ matrix as described in Subsection 5.2.

For more details in a general variable projection context, see Kaufman [96], where this simplification has been derived for the first time using a QR factorization of the matrix $\mathbf{F}(\mathbf{a})$ and differentiation of orthogonal matrices, Ruhe and Wedin [166] where a more direct derivation and generalizations are given, and also O'Leary and Rust [149] for a recent discussion of the respective merits of this Jacobian approximation against the true Jacobian matrix, again for general variable projection NLLS algorithms. Note that, in the computer vision's community, this Kaufman variant of the Gauss-Newton or Levenberg-Marquardt algorithms has been already extensively used for solving Structure-From-Motion (SFM) tasks [147][27][28][37][150][66][81][88] and is incorrectly called the Wiberg's algorithm in reference of the conference paper [190]. However, the first application of this algorithm to solve WLRA problems (with binary weights) is in fact due to Ruhe [158]. Furthermore, again in the computer vision community, this algorithm has frequently be assumed to be similar to the block ALS algorithm [176][15] (described in Section 4), which is also incorrect as shown above. The first correct derivation of this variant of the Gauss-Newton algorithm to solve WLRA problems with binary weights in the computer vision field is due to Okatani and Deguchi [147], see also the excellent Master thesis of Daskalov [37] in which this derivation is revisited.

Now, the Gauss-Seidel step, $d\mathbf{a}_{gs-gn}$, is closely related to the Kaufman step, $d\mathbf{a}_{k-gn}$, since it corresponds to applying the Gauss-Seidel iteration to the linear system appearing in the Kaufman-Gauss-Newton iteration, see [166] for details. It is also closely related to the block ALS algorithm described in Section 4, as we will demonstrate now.

The Gauss-Seidel step, $d\mathbf{a}_{gs-gn}$, is computed as the minimum 2-norm solution of the linear least-squares problem

$$\min_{d\mathbf{a}\in\mathbb{R}^{p.k}} \|\mathbf{r}(\mathbf{a}_i) - \mathbf{K}_{(n,p)}\mathbf{G}(\mathbf{b}_i)d\mathbf{a}\|_2^2$$

and we have

$$\mathbf{r}(\mathbf{a}_i) = \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a}_i)}\mathbf{x} = \mathbf{x} - \mathbf{F}(\mathbf{a}_i)\mathbf{b}_i = \mathbf{x} - \mathbf{K}_{(n,p)}\mathbf{G}(\mathbf{b}_i)\mathbf{a}_i \ ,$$

as demonstrated in Subsection 3.4. Hence,

$$\begin{aligned}\mathbf{r}(\mathbf{a}_i) - \mathbf{K}_{(n,p)}\mathbf{G}(\mathbf{b}_i)d\mathbf{a} &= \mathbf{x} - \mathbf{K}_{(n,p)}\mathbf{G}(\mathbf{b}_i)\mathbf{a}_i - \mathbf{K}_{(n,p)}\mathbf{G}(\mathbf{b}_i)d\mathbf{a} \\ &= \mathbf{x} - \mathbf{K}_{(n,p)}\mathbf{G}(\mathbf{b}_i)\big(\mathbf{a}_i + d\mathbf{a}\big) \\ &= \mathbf{K}_{(n,p)}\Big(\mathbf{z} - \mathbf{G}(\mathbf{b}_i)\big(\mathbf{a}_i + d\mathbf{a}\big)\Big) \ .\end{aligned}$$

Furthermore,

$$\|\mathbf{r}(\mathbf{a}_i) - \mathbf{K}_{(n,p)}\mathbf{G}(\mathbf{b}_i)d\mathbf{a}\|_2^2 = \|\mathbf{z} - \mathbf{G}(\mathbf{b}_i)\big(\mathbf{a}_i + d\mathbf{a}\big)\|_2^2 \ ,$$

since $\mathbf{K}_{(n,p)}$ is an orthogonal matrix, and we see that the Gauss-Seidel step, $d\mathbf{a}_{gs-gn}$, solves the linear least-squares problem

$$\min_{d\mathbf{a}\in\mathbb{R}^{p.k}} \|\mathbf{z} - \mathbf{G}(\mathbf{b}_i)\big(\mathbf{a}_i + d\mathbf{a}\big)\|_2^2 = \|\big(\mathbf{z} - \mathbf{G}(\mathbf{b}_i)\mathbf{a}_i\big) - \mathbf{G}(\mathbf{b}_i)d\mathbf{a}\|_2^2 \ ,$$

while the ALS iteration computes $\mathbf{a}_{i+1}$ as the minimum 2-norm solution of the linear least-squares problem

$$\min_{\mathbf{a}\in\mathbb{R}^{p.k}} \|\mathbf{z} - \mathbf{G}(\mathbf{b}_i)\mathbf{a}\|_2^2 \ .$$

If $\mathbf{G}(\mathbf{b}_i)$ is a full column-rank matrix, we then have

$$\mathbf{a}_{i+1} = \mathbf{G}(\mathbf{b}_i)^{+}\mathbf{z} = \big(\mathbf{G}(\mathbf{b}_i)^T\mathbf{G}(\mathbf{b}_i)\big)^{-1}\mathbf{G}(\mathbf{b}_i)^T\mathbf{z}$$

and

$$\begin{aligned}d\mathbf{a}_{gs-gn} &= \mathbf{G}(\mathbf{b}_i)^{+}\big(\mathbf{z} - \mathbf{G}(\mathbf{b}_i)\mathbf{a}_i\big) \\ &= \big(\mathbf{G}(\mathbf{b}_i)^T\mathbf{G}(\mathbf{b}_i)\big)^{-1}\mathbf{G}(\mathbf{b}_i)^T\big(\mathbf{z} - \mathbf{G}(\mathbf{b}_i)\mathbf{a}_i\big) \\ &= \big(\mathbf{G}(\mathbf{b}_i)^T\mathbf{G}(\mathbf{b}_i)\big)^{-1}\mathbf{G}(\mathbf{b}_i)^T\mathbf{z} - \mathbf{a}_i \\ &= \mathbf{a}_{i+1} - \mathbf{a}_i \ ,\end{aligned}$$

and, in these conditions, the ALS and Gauss-Seidel-Gauss-Newton algorithms generate exactly the same iterates. On the other hand, if $\mathbf{G}(\mathbf{b}_i)$ is not a full column-rank matrix, we still have the equality

$$\|\mathbf{z} - \mathbf{G}(\mathbf{b}_i)\mathbf{a}_{i+1}\|_2^2 = \|\big(\mathbf{z} - \mathbf{G}(\mathbf{b}_i)\mathbf{a}_i\big) - \mathbf{G}(\mathbf{b}_i)d\mathbf{a}_{gs-gn}\|_2^2 \ .$$

However, in general $\mathbf{a}_{i+1} \neq \mathbf{a}_i + d\mathbf{a}_{gs-gn}$ since the Gauss-Seidel iteration produces the minimum 2-norm correction vector $d\mathbf{a}_{gs-gn}$ to the above linear least-squares problem while the ALS algorithm obtains the minimum 2-norm solution $\mathbf{a}_{i+1}$ of this linear least-squares problem. Thus, unless $\mathbf{a}_{i+1}$ belongs to the correct manifold, the Gauss-Seidel and ALS steps do not produce the same iterate when the matrix $\mathbf{G}(\mathbf{b}_i)$ is not of full column-rank.

We finally observe that a line search in step **(7.2)** of the Gauss-Newton algorithms (1) is not required for the Gauss-Seidel correction $d\mathbf{a}_{gs-gn}$ in order to obtain the inequality $\psi(\mathbf{a}_{i+1}) \leq \psi(\mathbf{a}_i)$ and global convergence of the iterations, see Section 4 for details. On the other hand, for both the Golub-Pereyra correction $d\mathbf{a}_{gp-gn}$ and the Kaufman correction $d\mathbf{a}_{k-gn}$, it may happen that $\psi(\mathbf{a}_{i+1}) >$

$\psi(\mathbf{a}_i)$ meaning that the $\psi(.)$ surface is not reliably approximated by a quadratic function. In other words, the quadratic approximation is only good in the local neighborhood of $\mathbf{a}_i$, not at the bottom of the quadratic valley that the Gauss-Newton approach uses. In such cases, a line search algorithm to determine α_i at each iteration such that $\psi(\mathbf{a}_{i+1}) < \psi(\mathbf{a}_i)$ must be incorporated in step **(7.2)** of the Gauss-Newton algorithms (1) in order to obtain global convergence. Note also that this is always possible despite the singularity of the Jacobian matrix $-\big(\mathbf{M}(\mathbf{a}_i) + \mathbf{L}(\mathbf{a}_i)\big)$ or its Kaufman approximation $-\mathbf{M}(\mathbf{a}_i)$ (see Theorem 5.2) as the correction vectors $d\mathbf{a}_{gp-gn}$ and $d\mathbf{a}_{k-gn}$ are in a descent direction for $\psi(.)$ if $\nabla\psi(\mathbf{a}_i) \neq \mathbf{0}^{k.p}$ (see Corollary 5.7). However, to develop damped versions of these Gauss-Newton algorithms by implementing a line search algorithm, we have to perform the second part of step **(4)** of the Gauss-Newton algorithms (1), every time we want to get $\psi(\mathbf{a}_{i+1})$ for a new trial value of $\mathbf{a}_{i+1}$, since

$$\psi(\mathbf{a}_{i+1}) = \frac{1}{2}\|\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a}_{i+1})}\mathbf{x}\|_2^2 \,,$$

and a line search can involve many extra evaluations of $\psi(.)$, which do not get us closer to the solution. Furthermore, if a line search is incorporated, the Gauss-Newton algorithms (1) must be slightly reorganized to avoid duplicate computations in steps **(4)** and **(7)**, but we omit these details here.

In these conditions, to obtain global convergence, it is tempting to perform one iteration with a Gauss-Seidel step $d\mathbf{a}_{gs-gn}$ or even several iterations with the fast block ALS method described in Section 4 to compute $\mathbf{a}_{i+1}$ in step **(7.2)** (e.g., if $\psi(\mathbf{a}_i + d\mathbf{a}_{gn}) > \psi(\mathbf{a}_i)$) instead of using a more costly line search. In other words, if a full Gauss-Newton step gives a sufficient decrease of $\psi(.)$, we accept this point as the new iterate. Otherwise we switch to the fast Gauss-Seidel or block ALS methods.

We now explain how the matrices $\mathbf{M}(\mathbf{a}_i)$ and $-J(\mathbf{r}(\mathbf{a}_i)) = \mathbf{M}(\mathbf{a}_i) + \mathbf{L}(\mathbf{a}_i)$ and their QR factorizations can be computed efficiently and with reduced storage in order to obtain the correction vectors $d\mathbf{a}_{k-gn}$ or $d\mathbf{a}_{gp-gn}$ at each iteration of the Gauss-Newton algorithms (1). To this end, we first recall from the results of Subsection 5.2 that we have the following explicit expressions for these matrices:

$$\begin{aligned}
\mathbf{M}(\mathbf{a}) &= \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\mathbf{U}(\mathbf{a}) \,, \\
\mathbf{L}(\mathbf{a}) &= (\mathbf{F}(\mathbf{a})^{+})^T\mathbf{V}(\mathbf{a}) \,, \\
-J(\mathbf{r}(\mathbf{a})) &= \mathbf{M}(\mathbf{a}) + \mathbf{L}(\mathbf{a}) = \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\mathbf{U}(\mathbf{a}) + (\mathbf{F}(\mathbf{a})^{+})^T\mathbf{V}(\mathbf{a}) \,,
\end{aligned}$$

with

$$\begin{aligned}
\mathbf{U}(\mathbf{a}) &= \mathit{diag}(\mathit{vec}(\sqrt{\mathbf{W}}))(\widehat{\mathbf{B}}^T \otimes \mathbf{I}_p)\mathbf{K}_{(k,p)} = \mathbf{K}_{(n,p)}\mathbf{G}(\widehat{\mathbf{b}}) \,, \\
\mathbf{V}(\mathbf{a}) &= (\mathbf{W} \odot P_{\Omega}(\mathbf{X} - \mathbf{A}\widehat{\mathbf{B}}))^T \otimes \mathbf{I}_k \,,
\end{aligned}$$

as stated in equations (5.20), (5.21) and (5.22). The notations here are exactly the same as in Subsection 5.2 and we have drop again the iteration index of the Gauss-Newton iterations in order to simplify the notations in the rest of this section. Note that, in the definition of $\mathbf{L}(\mathbf{a})$, $\mathbf{F}(\mathbf{a})^{+}$ can be replaced by a symmetric generalized inverse $\mathbf{F}(\mathbf{a})^{-}$ at our convenience. We also recall that the matrices $-J(\mathbf{r}(\mathbf{a}))$, $\mathbf{M}(\mathbf{a})$ and $\mathbf{L}(\mathbf{a})$ have $n.p$ rows if $\mathbf{W} \in \mathbb{R}^{p\times n}_{+*}$, but only $nobs$ rows if $\mathbf{W} \in \mathbb{R}^{p\times n}_{+}$, where $nobs$ is the number of non-zero rows of $\mathbf{F}(\mathbf{a})$. As explained in Subsection 5.2, $nobs$ is simply the number of "non-missing" elements in the data matrix $\mathbf{X}$ or, equivalently, the number of non-zero weights in the weight matrix $\mathbf{W}$, namely,

$$nobs = \sum_{ij} \boldsymbol{\delta}_{ij} \,,$$

where $\boldsymbol{\delta}$ is the incidence matrix associated the weight matrix (also defined in Subsection 5.2). From the above equations, we see that the matrix $-J(\mathbf{r}(\mathbf{a}))$ and its two matrix components are tall and

skinny in most cases, even if the number of missing values is high, as k is expected to be much smaller than $\min(p,n)$ and that their evaluations require the computations of the orthogonal projector $\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}$, and, of $\mathbf{F}(\mathbf{a})^{+}$ and $\widehat{\mathbf{b}} = \mathbf{F}(\mathbf{a})^{+}\mathbf{x}$ if a COD is used in step **(2)** of the Gauss-Newton algorithms (1) or, alternatively, of $\mathbf{F}(\mathbf{a})^{-}$ and $\widehat{\mathbf{b}} = \mathbf{F}(\mathbf{a})^{-}\mathbf{x}$ if a QRCP is used in this step **(2)**.

The key-observation to compute efficiently these different matrices is to remember that $\mathbf{F}(\mathbf{a})$ is a block-diagonal matrix, namely,

$$\mathbf{F}(\mathbf{a}) = \bigoplus_{j=1}^{n} \mathbf{F}_j(\mathbf{a}) \text{ , where } \mathbf{F}_j(\mathbf{a}) = \mathit{diag}(\sqrt{\mathbf{W}}_{.j})\mathbf{A} \text{ .}$$

In these conditions, as already discussed in Subsection 5.2, we have

$$\mathit{rank}(\mathbf{F}(\mathbf{a})) = \sum_{j=1}^{n} \mathit{rank}(\mathbf{F}_j(\mathbf{a})) = \sum_{j=1}^{n} r_j = r_{\mathbf{F}(\mathbf{a})}$$

and

$$\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})} = \bigoplus_{j=1}^{n} \mathbf{P}^{\perp}_{\mathbf{F}_j(\mathbf{a})} \text{ , } \mathbf{F}(\mathbf{a})^{+} = \bigoplus_{j=1}^{n} \mathbf{F}_j(\mathbf{a})^{+} \text{ and } \mathbf{F}(\mathbf{a})^{-} = \bigoplus_{j=1}^{n} \mathbf{F}_j(\mathbf{a})^{-}.$$

Taking advantage of these block-structures of $\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}$ and of the generalized inverses of $\mathbf{F}(\mathbf{a})$ can reduce drastically the required storage and allows us to use efficient parallelization techniques for reducing the computing time needed to solve large WLRA problems using variable projection second-order algorithms as we will illustrate below. Interestingly, the techniques used for this purpose are very similar to those developed for solving large and dense structured linear least-squares problems arising in the context of separable NLLS problems with multiple right hand sides (e.g., NLLS problems in which a linear combination of nonlinear functions is fit linearly to data in many datasets); see Kaufman and Silvester [102], Kaufman et al. [103] and Kaufman [97] for more details.

Taking into account the block-structures of $\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}$, $\mathbf{F}(\mathbf{a})^{+}$ and $\mathbf{F}(\mathbf{a})^{-}$, we first observe that the $n.p \times k.p$ matrices $-J(\mathbf{r}(\mathbf{a}))$, $\mathbf{M}(\mathbf{a})$ and $\mathbf{L}(\mathbf{a})$ can be divided into n blocks, each of shape $p \times k.p$ if $\mathbf{W} \in \mathbb{R}^{p\times n}_{+*}$:

$$-J(\mathbf{r}(\mathbf{a})) = \begin{bmatrix} \mathbf{J}_1 \\ \vdots \\ \mathbf{J}_j \\ \vdots \\ \mathbf{J}_n \end{bmatrix} \text{ , } \mathbf{M}(\mathbf{a}) = \begin{bmatrix} \mathbf{M}_1 \\ \vdots \\ \mathbf{M}_j \\ \vdots \\ \mathbf{M}_n \end{bmatrix} \text{ and } \mathbf{L}(\mathbf{a}) = \begin{bmatrix} \mathbf{L}_1 \\ \vdots \\ \mathbf{L}_j \\ \vdots \\ \mathbf{L}_n \end{bmatrix} \text{ .} \tag{6.2}$$

If $\mathbf{W} \in \mathbb{R}^{p\times n}_{+}$, these matrices can also be divided into n blocks, but the number of rows in each block $\mathbf{J}_j$, $\mathbf{M}_j$ and $\mathbf{L}_j$ will differ and will be equal to the number of non-missing elements in the corresponding column of the data matrix $\mathbf{X}$. In order to simplify the exposition, but without loss of generality, we will assume in the rest of this section that $\mathbf{W} \in \mathbb{R}^{p\times n}_{+*}$. Obviously, if $\mathbf{W} \in \mathbb{R}^{p\times n}_{+}$, the zero-rows of these submatrices should be eliminated in real computations.

For the same reasons, the matrices $\mathbf{U}(\mathbf{a})$ and $\mathbf{V}(\mathbf{a})$ involved, respectively, in the definitions of $\mathbf{M}(\mathbf{a})$ and $\mathbf{L}(\mathbf{a})$ can also be considered as stacks of n blocks. We also observe that these two matrices are very sparse as $\mathbf{U}(\mathbf{a})$ is a row-permuted block diagonal matrix and $\mathbf{V}(\mathbf{a})$ is the Kronecker product of a matrix with $\mathbf{I}_k$, the identity matrix of order k. However, they have also a well-defined regular structure for the positions of their non-zero elements, which can be exploited in practical computations. As an illustration, it is easily checked, using the equality, $\mathbf{U}(\mathbf{a}) = \mathbf{K}_{(n,p)}\mathbf{G}(\widehat{\mathbf{b}})$, that the matrix $\mathbf{U}(\mathbf{a})$ as the following block structure with at most k non-zero elements in each of its rows

and at most n non-zero elements in each of its columns:

$$\mathbf{U}(\mathbf{a}) = \begin{bmatrix} \mathbf{U}_1 \\ \vdots \\ \mathbf{U}_j \\ \vdots \\ \mathbf{U}_n \end{bmatrix} \text{ with } \mathbf{U}_j \in \mathbb{R}^{p \times k.p} \text{ for } j = 1, \cdots, n \text{ and} \tag{6.3}$$

$$\mathbf{U}_j = \begin{bmatrix} \sqrt{\mathbf{W}}_{1j}(\widehat{\mathbf{B}}_{.j})^T & 0 & \dots & 0 & 0 \\ 0 & \sqrt{\mathbf{W}}_{2j}(\widehat{\mathbf{B}}_{.j})^T & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & \sqrt{\mathbf{W}}_{(p-1)j}(\widehat{\mathbf{B}}_{.j})^T & 0 \\ 0 & 0 & \dots & 0 & \sqrt{\mathbf{W}}_{pj}(\widehat{\mathbf{B}}_{.j})^T \end{bmatrix} .$$

Similarly, it is easily verified that the matrix $\mathbf{V}(\mathbf{a})$ has the following block structure with at most n non-zero elements in each of its columns and at most p non-zero elements in each of its rows:

$$\mathbf{V}(\mathbf{a}) = \begin{bmatrix} \mathbf{V}_1 \\ \vdots \\ \mathbf{V}_j \\ \vdots \\ \mathbf{V}_n \end{bmatrix} \text{ with } \mathbf{V}_j \in \mathbb{R}^{k \times k.p} \text{ for } j = 1, \cdots, n \text{ and } \mathbf{V}_j = \begin{bmatrix} \mathbf{Z}_1 & \cdots & \mathbf{Z}_i & \cdots & \mathbf{Z}_p \end{bmatrix} , \tag{6.4}$$

with $\mathbf{Z}_i \in \mathbb{R}^{k \times k}$ and $\mathbf{Z}_i = \beta_i \mathbf{I}_k$ for $i = 1, \cdots, p$, where $\beta_i = \mathbf{W}_{ij}(\bar{\mathbf{X}}_{ij} - \sum_{l=1}^{k} \mathbf{A}_{il}\widehat{\mathbf{B}}_{lj})$, $\mathbf{I}_k$ is the identity matrix of order k and

$$\bar{\mathbf{X}}_{ij} = \begin{cases} \mathbf{X}_{ij} & \text{if } \mathbf{W}_{ij} \neq 0 \\ 0 & \text{if } \mathbf{W}_{ij} = 0 \end{cases} .$$

Finally, if $\mathbf{W} \in \mathbb{R}^{p \times n}_{+*}$, the residual vector $\mathbf{r}(\mathbf{a}) = \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a})}\mathbf{x}$ can also be considered as a stack of n p-vectors with

$$\mathbf{r}(\mathbf{a}) = \begin{bmatrix} \mathbf{r}_1(\mathbf{a}) \\ \vdots \\ \mathbf{r}_j(\mathbf{a}) \\ \vdots \\ \mathbf{r}_n(\mathbf{a}) \end{bmatrix} \text{ with } \mathbf{r}_j(\mathbf{a}) = \mathbf{P}^{\perp}_{\mathbf{F}_j(\mathbf{a})}\mathbf{x}_j \text{ and } \mathbf{x}_j = \sqrt{\mathbf{W}}_{.j} \odot \mathbf{X}_{.j} . \tag{6.5}$$

The subvector $\mathbf{r}_j(\mathbf{a})$ is the j^{th} residual vector associated with the j^{th} atomic function $\psi_j(.)$ defined in equation (3.25) of Subsection 3.4. Using these different block-structures, we can use the following strategy for computing $\mathbf{M}(\mathbf{a})$ and $-J(\mathbf{r}(\mathbf{a}))$ in n independent steps.

At the j^{th} step, we compute the blocks $\mathbf{M}_j$, $\mathbf{L}_j$ and $\mathbf{J}_j$ defined above, namely,

$$\begin{aligned} \mathbf{J}_j &= \mathbf{M}_j + \mathbf{L}_j , \\ \mathbf{M}_j &= \mathbf{P}^{\perp}_{\mathbf{F}_j(\mathbf{a})}\mathbf{U}_j , \\ \mathbf{L}_j &= \mathbf{F}_j(\mathbf{a})^{+T}\mathbf{V}_j \text{ or } \mathbf{L}_j = \mathbf{F}_j(\mathbf{a})^{-T}\mathbf{V}_j . \end{aligned}$$

To this end, we need to process the j^{th} columns of $\mathbf{X}$ and $\mathbf{W}$, and we first compute the matrix $\mathbf{F}_j(\mathbf{a})$ as

$$\mathbf{F}_j(\mathbf{a}) = \mathit{diag}(\sqrt{\mathbf{W}}_{.j})\mathbf{A} ,$$

where $\mathbf{A}$ is the current estimate for this matrix variable of the factor model and we eliminate eventually the zero rows in $\mathbf{F}_j(\mathbf{a})$ if some elements of the weight column-vector $\mathbf{W}_{.j}$ are equal to zero.

Next, from the above equations, we see that the computation of $\mathbf{J}_j$ and its two matrix components requires the computations of $\mathbf{P}^{\perp}_{\mathbf{F}_j(\mathbf{a})}$, $\mathbf{F}_j(\mathbf{a})^+$ and $\widehat{\mathbf{B}}_{.j} = \mathbf{F}_j(\mathbf{a})^+\mathbf{x}_j$, or, $\mathbf{F}_j(\mathbf{a})^-$ and $\widehat{\mathbf{B}}_{.j} = \mathbf{F}_j(\mathbf{a})^-\mathbf{x}_j$, where $\mathbf{x}_j = \sqrt{\mathbf{W}}_{.j} \odot \mathbf{X}_{.j}$. This implies that the two matrix components of $\mathbf{J}_j$ can be obtained from a QRCP or a COD (see equations 2.15 and 2.20 in Subsection 2.1) of the $p \times k$ matrix $\mathbf{F}_j(\mathbf{a})$. See also Golub and Pereyra [63] [64], Krogh [95], Kaufman [96] and Gay and Kaufman [61] for more details in a more general framework dealing with general separable NLLS problems. Thus, we first compute the QRCP of $\mathbf{F}_j(\mathbf{a})$ as

$$\mathbf{F}_j(\mathbf{a}) = \mathbf{Q}_j^T \begin{bmatrix} \mathbf{R}_j & \mathbf{S}_j \\ \mathbf{0}^{(p-r_j)\times r_j} & \mathbf{0}^{(p-r_j)\times(k-r_j)} \end{bmatrix} \mathbf{P}_j^T \ ,$$

where $\mathbf{Q}_j$ is an $p \times p$ orthogonal matrix, $\mathbf{R}_j$ is an $r_j \times r_j$ nonsingular upper triangular matrix with $r_j = \mathit{rank}(\mathbf{F}_j(\mathbf{a}))$, which can be estimated during the QRCP, $\mathbf{S}_j$ is vacuous unless $\mathbf{F}_j(\mathbf{a})$ is rank deficient and $\mathbf{P}_j$ is a $k \times k$ permutation matrix. Note that $\mathbf{F}_j(\mathbf{a})$ is a dense matrix, so that its QRCP can be computed efficiently and cheaply with the help of standard dense methods, see Subsection 2.1 and [71][8] for details. From the above QRCP of $\mathbf{F}_j(\mathbf{a})$, we can compute $\mathbf{P}^{\perp}_{\mathbf{F}_j(\mathbf{a})}$ as

$$\mathbf{P}^{\perp}_{\mathbf{F}_j(\mathbf{a})} = \mathbf{Q}_j^T \begin{bmatrix} \mathbf{0}^{r_j\times r_j} & \mathbf{0}^{r_j\times(p-r_j)} \\ \mathbf{0}^{(p-r_j)\times r_j} & \mathbf{I}^{p-r_j} \end{bmatrix} \mathbf{Q}_j$$

and also a symmetric generalized inverse of $\mathbf{F}_j(\mathbf{a})$ as

$$\mathbf{F}_j(\mathbf{a})^- = \mathbf{P}_j \begin{bmatrix} \mathbf{R}_j^{-1} & \mathbf{0}_{r_j\times(p-r_j)} \\ \mathbf{0}^{(k-r_j)\times r_j} & \mathbf{0}^{(k-r_j)\times(p-r_j)} \end{bmatrix} \mathbf{Q}_j \ .$$

See again Subsection 2.1 for more information. Next, from this formulation of $\mathbf{F}_j(\mathbf{a})^-$, the j^{th} column of $\widehat{\mathbf{B}}$ can be computed as

$$\widehat{\mathbf{B}}_{.j} = \mathbf{F}_j(\mathbf{a})^-\mathbf{x}_j = \mathbf{P}_j^1\mathbf{R}_j^{-1}\mathbf{Q}_j^1\mathbf{x}_j \ .$$

In this last equation, the orthogonal matrices $\mathbf{Q}_j$ and $\mathbf{P}_j$ have been partitioned as

$$\mathbf{Q}_j = \begin{bmatrix} \mathbf{Q}_j^1 \\ \mathbf{Q}_j^2 \end{bmatrix} \text{ and } \begin{bmatrix} \mathbf{P}_j^1 & \mathbf{P}_j^2 \end{bmatrix} \ ,$$

where

- $\mathbf{Q}_j^1$ and $\mathbf{Q}_j^2$ have, respectively, r_j and $p - r_j$ rows ,
- $\mathbf{P}_j^1$ and $\mathbf{P}_j^2$ have, respectively, r_j and $k - r_j$ columns .

If $r_j = k$ then $\mathbf{F}_j(\mathbf{a})^+ = \mathbf{F}_j(\mathbf{a})^-$. On the other hand, if $\mathbf{F}_j(\mathbf{a})$ is singular, we can optionally compute its COD from its QRCP in order to obtain $\mathbf{F}_j(\mathbf{a})^+$ and compute $\widehat{\mathbf{B}}_{.j}$ as $\mathbf{F}_j(\mathbf{a})^+\mathbf{x}_j$. However, taking into account the special structure and indeterminacy associated with the solutions $(\widehat{\mathbf{A}}, \widehat{\mathbf{B}})$ of the WLRA problem (see Section 3), we don't really need to compute $\mathbf{F}_j(\mathbf{a})^+$ even if $\mathbf{F}_j(\mathbf{a})$ is rank-deficient, so we omit here the details of the optional computation of the COD of $\mathbf{F}_j(\mathbf{a})$.

Once $\widehat{\mathbf{B}}_{.j}$ has been computed, the submatrices $\mathbf{U}_j$ and $\mathbf{V}_j$ defined above can then be evaluated, or more precisely are available, to compute $\mathbf{M}_j$ and $\mathbf{L}_j$. Finally, $\mathbf{r}_j(\mathbf{a})$ can be computed as follows

$$\mathbf{r}_j(\mathbf{a}) = \mathbf{P}^{\perp}_{\mathbf{F}_j(\mathbf{a}_i)}\mathbf{x}_j = \mathbf{Q}_j^T \begin{bmatrix} \mathbf{0}^{r_j} \\ \mathbf{Q}_j^2\mathbf{x}_j \end{bmatrix} = (\mathbf{Q}_j^2)^T(\mathbf{Q}_j^2\mathbf{x}_j) \ ,$$

using the above results, and similarly if a QRCP or COD of $\mathbf{F}_j(\mathbf{a})$ is available. Inserting now the

above expressions for $\mathbf{P}^{\perp}_{\mathbf{F}_j(\mathbf{a})}$ and $\mathbf{F}_j(\mathbf{a})^-$ in the definition of $\mathbf{J}_j$, we obtain

$$\begin{aligned}
\mathbf{J}_j &= \mathbf{P}^{\perp}_{\mathbf{F}_j(\mathbf{a})}\mathbf{U}_j + (\mathbf{F}_j(\mathbf{a})^-)^T\mathbf{V}_j \\
&= \mathbf{Q}_j^T \begin{bmatrix} \mathbf{0}^{r_j\times r_j} & \mathbf{0}^{r_j\times(p-r_j)} \\ \mathbf{0}^{(p-r_j)\times r_j} & \mathbf{I}_{p-r_j} \end{bmatrix} \mathbf{Q}_j\mathbf{U}_j + \mathbf{Q}_j^T \begin{bmatrix} \mathbf{R}_j^{-T} & \mathbf{0}^{r_j\times(k-r_j)} \\ \mathbf{0}^{(p-r_j)\times r_j} & \mathbf{0}^{(p-r_j)\times(k-r_j)} \end{bmatrix} \mathbf{P}_j^T\mathbf{V}_j \\
&= \mathbf{Q}_j^T \Big(\begin{bmatrix} \mathbf{0}^{r_j\times r_j} & \mathbf{0}^{r_j\times(p-r_j)} \\ \mathbf{0}^{(p-r_j)\times r_j} & \mathbf{I}_{p-r_j} \end{bmatrix} \mathbf{Q}_j\mathbf{U}_j + \begin{bmatrix} \mathbf{R}_j^{-T} & \mathbf{0}^{r_j\times(k-r_j)} \\ \mathbf{0}^{(p-r_j)\times r_j} & \mathbf{0}^{(p-r_j)\times(k-r_j)} \end{bmatrix} \mathbf{P}_j^T\mathbf{V}_j \Big) \\
&= \mathbf{Q}_j^T \Big(\begin{bmatrix} \mathbf{0}^{r_j\times k.p} \\ \mathbf{Q}_j^2\mathbf{U}_j \end{bmatrix} + \begin{bmatrix} \mathbf{R}_j^{-T}(\mathbf{P}_j^1)^T\mathbf{V}_j \\ \mathbf{0}^{(p-r_j)\times k.p} \end{bmatrix} \Big) \\
&= \mathbf{Q}_j^T \begin{bmatrix} \mathbf{R}_j^{-T}(\mathbf{P}_j^1)^T\mathbf{V}_j \\ \mathbf{Q}_j^2\mathbf{U}_j \end{bmatrix}.
\end{aligned}$$

In these conditions, for $d\mathbf{a} \in \mathbb{R}^{k.p}$, we have

$$\mathbf{r}_j(\mathbf{a}) - \mathbf{J}_j d\mathbf{a} = \mathbf{Q}_j^T \Big(\begin{bmatrix} \mathbf{0}^{r_j} \\ \mathbf{Q}_j^2\mathbf{x}_j \end{bmatrix} - \begin{bmatrix} \mathbf{R}_j^{-T}(\mathbf{P}_j^1)^T\mathbf{V}_j \\ \mathbf{Q}_j^2\mathbf{U}_j \end{bmatrix} d\mathbf{a} \Big) .$$

At this point, we introduce several new block matrix and vector definitions again to simplify the notation going forward:

$$\widetilde{J}(\mathbf{r}(\mathbf{a})) = \begin{bmatrix} \widetilde{\mathbf{J}}_1 \\ \vdots \\ \widetilde{\mathbf{J}}_j \\ \vdots \\ \widetilde{\mathbf{J}}_n \end{bmatrix} \text{ with } \widetilde{\mathbf{J}}_j = \begin{bmatrix} \mathbf{R}_j^{-T}(\mathbf{P}_j^1)^T\mathbf{V}_j \\ \mathbf{Q}_j^2\mathbf{U}_j \end{bmatrix} \tag{6.6}$$

and

$$\widetilde{\mathbf{r}}(\mathbf{a}) = \begin{bmatrix} \widetilde{\mathbf{r}}_1 \\ \vdots \\ \widetilde{\mathbf{r}}_j \\ \vdots \\ \widetilde{\mathbf{r}}_n \end{bmatrix} \text{ with } \widetilde{\mathbf{r}}_j = \begin{bmatrix} \mathbf{0}^{r_j} \\ \mathbf{Q}_j^2\mathbf{x}_j \end{bmatrix} . \tag{6.7}$$

Thus, $\widetilde{\mathbf{r}}(\mathbf{a})$ is an $n.p$-vector, which is a stack of n $(p-r_j)$-subvectors, separated by r_j zero elements in sequential order. Finally, we conceptually define the orthogonal block diagonal matrix

$$\mathbf{Q}_\mathbf{F} = \bigoplus_{j=1}^{n} \mathbf{Q}_j . \tag{6.8}$$

With these new notations and the preceding results, we have

$$-J(\mathbf{r}(\mathbf{a})) = \mathbf{Q}_\mathbf{F}\widetilde{J}(\mathbf{r}(\mathbf{a})) \text{ and } \mathbf{r}(\mathbf{a}) = \mathbf{Q}_\mathbf{F}\widetilde{\mathbf{r}}(\mathbf{a}) . \tag{6.9}$$

Now, as the 2-norm is unitarily invariant, $\forall d\mathbf{a} \in \mathbb{R}^{k.p}$, we have

$$\|\mathbf{r}(\mathbf{a}) + J\big(\mathbf{r}(\mathbf{a})\big)d\mathbf{a}\|_2 = \|\mathbf{r}(\mathbf{a}) - \big(\mathbf{M}(\mathbf{a}) + \mathbf{L}(\mathbf{a})\big)d\mathbf{a}\|_2 = \|\widetilde{\mathbf{r}}(\mathbf{a}) - \widetilde{J}\big(\mathbf{r}(\mathbf{a})\big)d\mathbf{a}\|_2 ,$$

as $\mathbf{Q}_\mathbf{F}$ is an $p.n \times p.n$ orthogonal matrix. In other words, the linear least-squares problem

$$\min_{d\mathbf{a}\in\mathbb{R}^{p.k}} \|\mathbf{r}(\mathbf{a}) - \big(\mathbf{M}(\mathbf{a}) + \mathbf{L}(\mathbf{a})\big)d\mathbf{a}\|_2^2 , \tag{6.10}$$

which must be solved at each iteration of the Gauss-Newton algorithm if a Golub-Pereyra step $d\mathbf{a}_{gp-gn}$ is used, is equivalent to the linear least-squares problem

$$\min_{d\mathbf{a}\in\mathbb{R}^{p.k}} \|\widetilde{\mathbf{r}}(\mathbf{a}) - \widetilde{J}(\mathbf{r}(\mathbf{a}))d\mathbf{a}\|_2^2 . \tag{6.11}$$

Similarly, if we use the Kaufman variant at each iteration of the Gauss-Newton algorithm (1), it is not difficult to verify using similar arguments that computing a QRCP of $\mathbf{F}_j(\mathbf{a})$ at step **(2)** of this algorithm is again sufficient and that the associated linear least-squares problem to solve for computing the correction vector $d\mathbf{a}_{k-gn}$, namely,

$$\min_{d\mathbf{a}\in\mathbb{R}^{p.k}} \|\mathbf{r}(\mathbf{a}) - \mathbf{M}(\mathbf{a})d\mathbf{a}\|_2^2 \,, \tag{6.12}$$

is equivalent to the linear least-squares problem

$$\min_{d\mathbf{a}\in\mathbb{R}^{p.k}} \|\widetilde{\mathbf{r}}(\mathbf{a}) - \widetilde{\mathbf{M}}(\mathbf{a})d\mathbf{a}\|_2^2 \,,$$

where $\widetilde{\mathbf{M}}(\mathbf{a})$ has the following block structure

$$\widetilde{\mathbf{M}}(\mathbf{a}) = \begin{bmatrix} \widetilde{\mathbf{M}}_1 \\ \vdots \\ \widetilde{\mathbf{M}}_j \\ \vdots \\ \widetilde{\mathbf{M}}_n \end{bmatrix} \text{ with } \widetilde{\mathbf{M}}_j = \begin{bmatrix} \mathbf{0}^{r_j \times k.p} \\ \mathbf{Q}_j^2 \mathbf{U}_j \end{bmatrix} \,, \tag{6.13}$$

and we also have the matrix equality

$$\mathbf{M}(\mathbf{a}) = \mathbf{Q}_{\mathbf{F}} \widetilde{\mathbf{M}}(\mathbf{a}) \,. \tag{6.14}$$

Furthermore, as zero rows appearing in the coefficient matrix of a linear least-squares problem do not affect the solution of this linear least-squares problem, these zero rows can be deleted and the linear least-squares problem to be solved at each iteration of the Kaufman variant of the Gauss-Newton algorithm (1) reduces, finally, to

$$\min_{d\mathbf{a}\in\mathbb{R}^{p.k}} \|\bar{\mathbf{r}}(\mathbf{a}) - \bar{\mathbf{M}}(\mathbf{a})d\mathbf{a}\|_2^2 \,, \tag{6.15}$$

where

$$\bar{\mathbf{M}}(\mathbf{a}) = \begin{bmatrix} \bar{\mathbf{M}}_1 \\ \vdots \\ \bar{\mathbf{M}}_j \\ \vdots \\ \bar{\mathbf{M}}_n \end{bmatrix} \text{ with } \bar{\mathbf{M}}_j = \mathbf{Q}_j^2 \mathbf{U}_j \tag{6.16}$$

and

$$\bar{\mathbf{r}}(\mathbf{a}) = \begin{bmatrix} \bar{\mathbf{r}}_1 \\ \vdots \\ \bar{\mathbf{r}}_j \\ \vdots \\ \bar{\mathbf{r}}_n \end{bmatrix} \text{ with } \bar{\mathbf{r}}_j = \mathbf{Q}_j^2 \mathbf{x}_j \,. \tag{6.17}$$

$\bar{\mathbf{M}}(\mathbf{a})$ and $\bar{\mathbf{r}}(\mathbf{a})$ have, respectively, only $n.p - r_{\mathbf{F}(\mathbf{a})}$ rows and elements if $\mathbf{W} \in \mathbb{R}_{+*}^{p\times n}$, and, $nobs - r_{\mathbf{F}(\mathbf{a})}$ rows and elements if $\mathbf{W} \in \mathbb{R}_{+}^{p\times n}$. Thus, the work involved in solving this linear least-squares problem is further reduced in addition to the simplifications introduced by the use of the approximated Jacobian matrix $-\mathbf{M}(\mathbf{a})$ as the coefficient matrix of the linear least-squares problem.

In all the above alternative formulations of the linear least-squares problems involving the Jacobian matrix or its approximation, which must be solved at each iteration of the Gauss-Newton

algorithms (1), we observe that both the coefficient matrix and the right hand-side vector of the associated linear least-squares problems can be computed independently in n steps, which may offer some important speed-up in a parallel environment.

The next critical step is to solve the linear least-squares problems involving the huge, but tall and skinny, matrices $\widetilde{J}(\mathbf{r}(\mathbf{a}))$ or $\bar{\mathbf{M}}(\mathbf{a})$ in a computationally responsible manner. If $\mathbf{W} \in \mathbb{R}^{p\times n}_{+*}$, $\widetilde{J}(\mathbf{r}(\mathbf{a}))$ and $\bar{\mathbf{M}}(\mathbf{a})$ will have, respectively, $n.p$ and $n.p - r_{\mathbf{F}(\mathbf{a})}$ rows and $k.p$ columns (with $k \ll \min(n, n)$) and it is not conceivable to store such huge matrices in main memory to compute their SVD, QRCP or COD by standard methods as soon as both p and n are relatively large numbers. We suggest two different strategies to alleviate this problem. The first one uses a QR decomposition of the transformed matrices $\widetilde{J}(\mathbf{r}(\mathbf{a}))$ or $\bar{\mathbf{M}}(\mathbf{a})$ as a preliminary step to reduce the size of the problems and the second one consists in solving the normal equations associated with the linear least-squares problems (6.10) and (6.12) or their transformed versions (6.11) and (6.15).

For most NLLS problems, separable or not, using a QR decomposition of the Jacobian matrix takes at least twice as long as using the normal equations, but gives improved accuracy when the Jacobian matrix is ill-conditioned. As we already know that $-J(\mathbf{r}(\mathbf{a}))$ and $\mathbf{M}(\mathbf{a})$, or their transformed versions, are always rank-deficient matrices (see Theorem 5.2), this may favor a QR approach for more reliable results. However, computing a Cholesky factor of the Gauss-Newton approximations of the Hessian matrix is substantially faster than computing a QR factorization of the (approximated) Jacobian matrix, especially when this matrix is tall and skinny. This explains why the normal equations approach has been favored in past studies [150][81][88] despite the inherent difficulties in this approach to deal efficiently and accurately with the singularity of the Gauss-Newton approximations of the Hessian matrix.

We first describe the iterative methods, which aim at computing the thin QR decomposition of the $n.p \times k.p$ transformed Jacobian matrix

$$\widetilde{J}(\mathbf{r}(\mathbf{a})) = \widetilde{\mathbf{Q}}_J \mathbf{R}_J \ , \tag{6.18}$$

where $\widetilde{\mathbf{Q}}_J$ is an $n.p \times k.p$ matrix (if $\mathbf{W} \in \mathbb{R}^{p\times n}_{+*}$) with orthonormal columns and $\mathbf{R}_J$ is an $k.p \times k.p$ (singular) upper-triangular matrix, or of its Kaufman approximation

$$\bar{\mathbf{M}}(\mathbf{a}) = \bar{\mathbf{Q}}_{\mathbf{M}} \mathbf{R}_{\mathbf{M}} \ , \tag{6.19}$$

where $\bar{\mathbf{Q}}_{\mathbf{M}}$ and $\mathbf{R}_{\mathbf{M}}$ have similar shapes than $\bar{\mathbf{M}}(\mathbf{a})$ and $\mathbf{R}_J$, respectively. In most practical applications k is much smaller than $\min(p, n)$ and the matrices $\widetilde{J}(\mathbf{r}(\mathbf{a}))$ and $\bar{\mathbf{M}}(\mathbf{a})$ are tall and skinny, as discussed above, for which highly efficient parallel QR algorithms have been proposed in the literature [48]. These dedicated Tall and Skinny QR (TSQR) algorithms are often referred as communication avoiding algorithms and outperform significantly the conventional Householder QR algorithm for an $m \times n$ matrix with $m \gg n$. Furthermore, these TSQR algorithms can be combined or more precisely "fused" with the n independent steps described above for the computation of $\widetilde{J}(\mathbf{r}(\mathbf{a}))$ or $\bar{\mathbf{M}}(\mathbf{a})$. This allows us to get the triangular factors $\mathbf{R}_J$ or $\mathbf{R}_{\mathbf{M}}$ in the standard QR factorizations of $\widetilde{J}(\mathbf{r}(\mathbf{a}))$ or $\bar{\mathbf{M}}(\mathbf{a})$ without storing in main memory these huge matrices or explicitly computing the orthonormal matrices $\widetilde{\mathbf{Q}}_J$ or $\bar{\mathbf{Q}}_{\mathbf{M}}$.

We now focus on two TSQR algorithms for computing these triangular factors and also the matrix-vector products $\widetilde{\mathbf{Q}}_J^T \widetilde{\mathbf{r}}(\mathbf{a})$ or $\bar{\mathbf{Q}}_{\mathbf{M}}^T \bar{\mathbf{r}}(\mathbf{a})$, which are needed to solve the associated linear least-squares problems involving $\widetilde{J}(\mathbf{r}(\mathbf{a}))$ and $\bar{\mathbf{M}}(\mathbf{a})$ in a final step. The first TSQR method is a serial algorithm and the second one is a parallel algorithm. Both of them processes the rows of $\widetilde{J}(\mathbf{r}(\mathbf{a}))$ or $\bar{\mathbf{M}}(\mathbf{a})$ in n steps as described above, but in different order and with different computational kernels as we will see now.

For the sake of convenience, we first describe the serial TSQR, which proceeds the n steps in sequential order from $j = 1$ to n. Again, to simplify the presentation, but without loss of generality, we will also assume that $\mathbf{W} \in \mathbb{R}^{p\times n}_{+*}$, so that $\widetilde{J}(\mathbf{r}(\mathbf{a}))$ have $n.p$ rows and each of the n steps processes exactly p rows of $\widetilde{J}(\mathbf{r}(\mathbf{a}))$, namely, the j^{th} step processes the submatrix $\widetilde{J}_j$ defined above.

In the first step, the submatrix $\widetilde{J}_1$ and the subvector $\widetilde{\mathbf{r}}_1$ are evaluated exactly as described above. Then, a standard Householder QR algorithm is applied to $\widetilde{J}_1$ to transform this target matrix into an upper triangular (if $k = 1$) or trapezoidal (if $k > 1$) matrix $\widetilde{\mathbf{R}}_1$:

$$\widetilde{J}_1 = \widetilde{\mathbf{Q}}_1 \widetilde{\mathbf{R}}_1 \ ,$$

where $\widetilde{\mathbf{Q}}_1$ is an $p \times p$ orthogonal matrix and $\widetilde{\mathbf{R}}_1$ is an $p \times k.p$ upper triangular or trapezoidal matrix. This is performed by the application of a sequence of p Householder transformations, whose product implicitly represents the orthogonal matrix $\widetilde{\mathbf{Q}}_1$, see Subsection 2.1 for more details. The algorithm consists of the iterations of two steps: generation of the Householder transformation from the target column vector of $\widetilde{J}_1$ and application of this Householder transformation to the trailing part of $\widetilde{J}_1$. Next, the right hand-side subvector $\widetilde{\mathbf{r}}_1$ is pre-multiplied by the transpose of $\widetilde{\mathbf{Q}}_1$. Note that, in a practical implementation of this TSQR algorithm, it is convenient to concatenate and store $\widetilde{J}_1$ and $\widetilde{\mathbf{r}}_1$ in the same matrix array (with p rows and $k.p + 1$ columns if $\mathbf{W} \in \mathbb{R}^{p \times n}_{+*}$) so that the matrix-vector product $\widetilde{\mathbf{Q}}_1^T \widetilde{\mathbf{r}}_1$ is directly computed when the Householder transformations are applied to the trailing part of $\widetilde{J}_1$ during its QR factorization.

In the second step of the TSQR algorithm, the submatrix $\widetilde{J}_2$ and the subvector $\widetilde{\mathbf{r}}_2$ are first evaluated and then combined with the results of the first step as follows:

$$\widetilde{\widetilde{J}}_2 = \begin{bmatrix} \widetilde{\mathbf{R}}_1 \\ \widetilde{J}_2 \end{bmatrix} \text{ and } \widetilde{\widetilde{\mathbf{r}}}_2 = \begin{bmatrix} \widetilde{\mathbf{Q}}_1^T \widetilde{\mathbf{r}}_1 \\ \widetilde{\mathbf{r}}_2 \end{bmatrix} \ .$$

Then, a structured and thin QR factorization of the matrix $\widetilde{\widetilde{J}}_2$ is performed

$$\widetilde{\widetilde{J}}_2 = \widetilde{\mathbf{Q}}_2 \widetilde{\mathbf{R}}_2 \ ,$$

where $\widetilde{\mathbf{Q}}_2$ is a matrix with $\min(2.p, k.p)$ orthonormal columns and $\widetilde{\mathbf{R}}_2$ is an upper triangular or trapezoidal matrix. This reduction to triangular or trapezoidal form of $\widetilde{\widetilde{J}}_2$ can be accomplished by a special sequence of $\min(2.p, k.p)$ Householder transformations in which the i^{th} transformation is designed to annihilate the nonzero subdiagonal elements in the i^{th} column of $\widetilde{\widetilde{J}}_2$. Note that no fill-in occurs during this process because the columns of $\widetilde{\widetilde{J}}_2$ are reduced from left to right. Finally, the vector $\widetilde{\widetilde{\mathbf{r}}}_2$ is pre-multiplied by the transpose of $\widetilde{\mathbf{Q}}_2$ and this ends the second step.

At the j^{th} step, the submatrix $\widetilde{J}_j$ and subvector $\widetilde{\mathbf{r}}_j$ are evaluated and concatenated with the outputs of the $j-1$ step as

$$\widetilde{\widetilde{J}}_j = \begin{bmatrix} \widetilde{\mathbf{R}}_{j-1} \\ \widetilde{J}_j \end{bmatrix} \text{ and } \widetilde{\widetilde{\mathbf{r}}}_j = \begin{bmatrix} \widetilde{\mathbf{Q}}_{j-1}^T \widetilde{\mathbf{r}}_{j-1} \\ \widetilde{\mathbf{r}}_j \end{bmatrix} \ ,$$

and a structured and thin QR factorization of $\widetilde{\widetilde{J}}_j$ is performed as

$$\widetilde{\widetilde{J}}_j = \widetilde{\mathbf{Q}}_j \widetilde{\mathbf{R}}_j \ ,$$

where $\widetilde{\mathbf{Q}}_j$ is a matrix with $\min(j.p, k.p)$ orthonormal columns and $\widetilde{\mathbf{R}}_j$ is an upper triangular or trapezoidal matrix. The vector $\widetilde{\widetilde{\mathbf{r}}}_j$ is also pre-multiplied by the transpose of $\widetilde{\mathbf{Q}}_j$ after this new structured QR factorization.

Then, the following steps are exactly similar to the j^{th} step and this process continues in blocks of p rows of $\widetilde{J}(\mathbf{r}(\mathbf{a}))$ until there are no more rows of $\widetilde{J}(\mathbf{r}(\mathbf{a}))$ left. In exact arithmetic without roundoff errors, it can be shown that the final triangular factor $\widetilde{\mathbf{R}}_n$ obtained by this recursive algorithm is the upper triangular factor $\mathbf{R}_J$ of the standard thin QR factorization of $\widetilde{J}(\mathbf{r}(\mathbf{a}))$ defined in equation (6.18). Furthermore, the associated right hand-side vector $\widetilde{\mathbf{Q}}_n^T \widetilde{\widetilde{\mathbf{r}}}_n$ in output of the recursive process is also equal the matrix-vector product $\widetilde{\mathbf{Q}}_J^T \widetilde{\mathbf{r}}(\mathbf{a})$, where $\widetilde{\mathbf{Q}}_J$ is also defined in equation (6.18) and $\widetilde{\mathbf{r}}(\mathbf{a})$ in equation (6.7).

For the description of the parallel TSQR algorithm, we assume that t processors are available with their own memory. Then, the n columns of the matrices $\mathbf{X}$ and $\mathbf{W}$ and the n steps are distributed equally among these t processors (or eventually such that the rows of $\widetilde{J}(\mathbf{r}(\mathbf{a}))$ are partitioned equally among the t processors if $\mathbf{W} \in \mathbb{R}_+^{p \times n}$).

Next, each processor i processes its own n_i steps independently without any communication between the processors as in the serial TSQR algorithm. The obtained triangular factors and transformed right hand-side vectors, say,

$$\widetilde{\mathbf{R}(i)} = \widetilde{\mathbf{R}(i)}_{n_i} \text{ and } \widetilde{\mathbf{r}(i)}(\mathbf{a}) = \widetilde{\mathbf{Q}(i)}_{n_i}^T \widetilde{\mathbf{r}(i)}_{n_i} \text{ for } i = 1 \text{ to } t \,,$$

are then reduced into an unique triangular factor $\mathbf{R}_J$ and an unique transformed right hand-side vector $\widetilde{\mathbf{Q}}_J^T \widetilde{\mathbf{r}}(\mathbf{a})$ by the QR factorizations (in parallel) of a sequence of matrices built by coupling two upper-triangular factors $\widetilde{\mathbf{R}(i)}$ and $\widetilde{\mathbf{R}(j)}$ on top of each other. We also call such special QR factorization, a structured QR factorization, and this QR factorization can also be performed by a special sequence of Householder transformations [111][48]. In this second part of the parallel TSQR algorithm, we note that several reduction trees are available to obtain the final triangular factor $\mathbf{R}_J$ and transformed right hand-side vector $\widetilde{\mathbf{Q}}_J^T \widetilde{\mathbf{r}}(\mathbf{a})$ [48]. But, for simplicity, we further assume that t is a power of two and that a binary reduction tree is used. In other words, the pairs of processors used in the second part of the parallel TSQR algorithm are given by simply grouping together, initially, a processor and its neighbour, e.g., $(0,1), (2,3), \cdots, (t-2, t-1)$. Then, in the next recursion level, we group the left most processor of a pair, e.g., $(0,2), (4,6), \cdots$ and proceed in this fashion until the last pair is $(0, t/2)$ with the result that the triangular factor and right hand-side vector stored in processor 0 contains $\mathbf{R}_J$ and the matrix-vector product $\widetilde{\mathbf{Q}}_J^T \widetilde{\mathbf{r}}(\mathbf{a})$, e.g., the actual triangular factor in the QR factorization of $\widetilde{J}(\mathbf{r}(\mathbf{a}))$ and the associated transformed right hand-side vector.

Thus, in the parallel version of the TSQR algorithm, after each processor i has computed its own triangular factor $\widetilde{\mathbf{R}(i)}$, one repeats sending, receiving one's $\widetilde{\mathbf{R}(j)}$ to/from one's processor neighbour and calculating structured QR factorizations in parallel. Then, another round of a reduced number of structured QR factorizations is performed in parallel and this process repeats until the final $\mathbf{R}_J$ factor is obtained. Obviously, other groupings of processors can be used to take advantage of a given topology of a network of processors. As noted in Demmel et al. [48], any sequence of tree of (structured) QR factorizations between the serial and parallel versions of the TSQR algorithm will work. Our binary tree version shows how to compute the TSQR factorization with maximum parallelism while the serial version does not exhibit any parallelism, but requires less memory. Note, finally, that the two structured QR factorizations used in the serial and parallel TSQR algorithms, namely,

$$\begin{bmatrix} \mathbf{R}_1 \\ \mathbf{R}_2 \end{bmatrix} = \mathbf{QR} \quad \text{and} \quad \begin{bmatrix} \mathbf{R}_1 \\ \mathbf{X} \end{bmatrix} = \mathbf{QR} \,,$$

where $\mathbf{R}_1, \mathbf{R}_2$ and $\mathbf{R}$ are upper triangular or trapezoidal matrices, $\mathbf{Q}$ is an orthogonal matrix and $\mathbf{X}$ is a full matrix, can be computed very efficiently using BLAS3 algorithms exploiting the triangular structure of the $\mathbf{R}_1$ and $\mathbf{R}_2$ matrices appearing in these structured QR factorizations [48]. As an illustration, such computational kernels are already available in the recent versions of the LAPACK library.

Furthermore, the orthonormal matrix $\widetilde{\mathbf{Q}}_J$ in the thin QR factorization of $\widetilde{J}(\mathbf{r}(\mathbf{a}))$ (see equation (6.18)) is never explicitly formed in both the serial and parallel TSQR algorithms, but pre-multiplying the vector $\widetilde{\mathbf{r}}(\mathbf{a})$ by $\widetilde{\mathbf{Q}}_J^T$ can be done also recursively and efficiently during the TSQR algorithms (as illustrated above) and this is sufficient to solve the linear least-squares problems involving $\widetilde{J}(\mathbf{r}(\mathbf{a}))$ as a coefficient matrix, or as a block of the coefficient matrix to deal with its singularity as we will illustrate below. Obviously, the same recursive TSQR methods can be used to compute the thin QR factorization of $\bar{\mathbf{M}}(\mathbf{a})$ if a Kaufman step is used in the Gauss-Newton algorithms (1), but we omit the details here as the steps are essentially the same.

In the previous paragraphs, we show how variable projection WLRA solvers, which request the Jacobian matrix (or its approximation in the case of the Kaufman variant) and perform a QR decomposition of it for solving the linear least-squares problems (6.10) or (6.12) at each iteration, can be implemented efficiently and with reduced memory requirements by exploiting the block diagonal structure of $\mathbf{F}(\mathbf{a})$ and using a parallel two-step TSQR algorithm. We now consider variable projection WLRA solvers, which, at each iteration, solve the linear least-squares problems (6.10) or (6.12) by computing the $k.p \times k.p$ cross-product positive semi-definite matrices

$$\Delta = J(\mathbf{r}(\mathbf{a}))^T J(\mathbf{r}(\mathbf{a})) \text{ or } \Lambda = \mathbf{M}(\mathbf{a})^T \mathbf{M}(\mathbf{a})$$

and solving the associated normal equations, e.g.,

$$\Delta d\mathbf{a} = -\nabla\psi(\mathbf{a}) \text{ or } \Lambda d\mathbf{a} = -\nabla\psi(\mathbf{a}) \ ,$$

where $\nabla\psi(\mathbf{a}) = J(\mathbf{r}(\mathbf{a}))^T \mathbf{r}(\mathbf{a})$.

In order to present in more details this Cholesky approach, we first recall from the results in Section 5.2 that

$$J(\mathbf{r}(\mathbf{a})) = -\big(\mathbf{M}(\mathbf{a}) + \mathbf{L}(\mathbf{a})\big)$$

and that the columns of $\mathbf{L}(\mathbf{a})$ lie in *ran*$(\mathbf{F}(\mathbf{a}))$, the range of $\mathbf{F}(\mathbf{a})$, and those of $\mathbf{M}(\mathbf{a})$ lie in *ran*$(\mathbf{F}(\mathbf{a}))^\perp$. This implies the equalities

$$\mathbf{M}(\mathbf{a})^T \mathbf{L}(\mathbf{a}) = \mathbf{0}^{k.p \times k.p} \quad \text{and} \quad \mathbf{L}(\mathbf{a})^T \mathbf{r}(\mathbf{a}) = \mathbf{0}^{k.p} \ ,$$

from which we deduce that

$$\Delta = \mathbf{M}(\mathbf{a})^T \mathbf{M}(\mathbf{a}) + \mathbf{L}(\mathbf{a})^T \mathbf{L}(\mathbf{a}) \quad \text{and} \quad \nabla\psi(\mathbf{a}) = \mathbf{M}(\mathbf{a})^T \mathbf{r}(\mathbf{a}) \ .$$

Thus, Δ is the sum of two positive semi-definite matrices and $\mathbf{L}(\mathbf{a})$ does not contribute in $\nabla\psi(\mathbf{a})$, see Section 5.3 for details. As in the QR approach, to speed up and parallelize the computations, it is convenient to consider the matrices $-J(\mathbf{r}(\mathbf{a}))$, $\mathbf{M}(\mathbf{a})$ and $\mathbf{L}(\mathbf{a})$ as stacks of n submatrices (each of p rows if $\mathbf{W} \in \mathbb{R}^{p \times n}_{+*}$) as in equation (6.2). As before, the j^{th} blocks $\mathbf{J}_j$, $\mathbf{M}_j$ and $\mathbf{L}_j$ are defined by

$$\begin{aligned} \mathbf{J}_j &= \mathbf{M}_j + \mathbf{L}_j \ , \\ \mathbf{M}_j &= \mathbf{P}^\perp_{\mathbf{F}_j(\mathbf{a})} \mathbf{U}_j \ , \\ \mathbf{L}_j &= \mathbf{L}_j = \mathbf{F}_j(\mathbf{a})^{-T} \mathbf{V}_j \ . \end{aligned}$$

and can be computed from the j^{th} columns of the matrices $\mathbf{X}$ and $\mathbf{W}$. Similarly, it is useful to consider the $n.p$-vector $\mathbf{r}(\mathbf{a})$ as a stack of n subvectors of dimension p (if $\mathbf{W} \in \mathbb{R}^{p \times n}_{+*}$) as in equation (6.5). With these block-structures of $-J(\mathbf{r}(\mathbf{a}))$, $\mathbf{M}(\mathbf{a})$, $\mathbf{L}(\mathbf{a})$ and $\mathbf{r}(\mathbf{a})$, we have the equalities

$$\Lambda = \mathbf{M}(\mathbf{a})^T \mathbf{M}(\mathbf{a}) = \sum_{j=1}^{n} \mathbf{M}_j^T \mathbf{M}_j \quad \text{and} \quad \mathbf{L}(\mathbf{a})^T \mathbf{L}(\mathbf{a}) = \sum_{j=1}^{n} \mathbf{L}_j^T \mathbf{L}_j$$

and also

$$\Delta = \big(\sum_{j=1}^{n} \mathbf{M}_j^T \mathbf{M}_j + \sum_{j=1}^{n} \mathbf{L}_j^T \mathbf{L}_j\big) \quad \text{and} \quad \nabla\psi(\mathbf{a}) = \sum_{j=1}^{n} \mathbf{M}_j^T \mathbf{r}_j(\mathbf{a}) \ ,$$

which show that the computations of Δ, Λ and $\nabla\psi(\mathbf{a})$ can be easily parallelized if several processors are available.

As in the QR approach, we can also use the transformed versions of $-J(\mathbf{r}(\mathbf{a}))$ and $\mathbf{M}(\mathbf{a})$ (see equations (6.8), (6.9) and (6.14)) defined as

$$\widetilde{J}\big(\mathbf{r}(\mathbf{a})\big) = \mathbf{Q}_\mathbf{F}^T J\big(\mathbf{r}(\mathbf{a})\big) \quad \text{and} \quad \widetilde{\mathbf{M}}(\mathbf{a}) = \mathbf{Q}_\mathbf{F}^T \mathbf{M}(\mathbf{a})$$

since

$$\Delta = \widetilde{J}\big(\mathbf{r}(\mathbf{a})\big)^T \widetilde{J}\big(\mathbf{r}(\mathbf{a})\big) \quad \text{and} \quad \Lambda = \widetilde{\mathbf{M}}(\mathbf{a})^T \widetilde{\mathbf{M}}(\mathbf{a}) = \bar{\mathbf{M}}(\mathbf{a})^T \bar{\mathbf{M}}(\mathbf{a}) \; .$$

Furthermore, taking again into account the respective block-structures of $\widetilde{J}(\mathbf{r}(\mathbf{a}))$, $\widetilde{\mathbf{M}}(\mathbf{a})$ and $\bar{\mathbf{M}}$, defined, respectively, in equations (6.6), (6.13) and (6.16), we have the equalities

$$\Delta = \sum_{j=1}^{n} \widetilde{\mathbf{J}}_j^T \widetilde{\mathbf{J}}_j \quad \text{and} \quad \Lambda = \sum_{j=1}^{n} \widetilde{\mathbf{M}}_j^T \widetilde{\mathbf{M}}_j = \sum_{j=1}^{n} \bar{\mathbf{M}}_j^T \bar{\mathbf{M}}_j \; ,$$

which demonstrate that the evaluation of these transformed forms of Δ and Λ can also be easily parallelized. Similarly, using the block-structures of $\widetilde{\mathbf{r}}(\mathbf{a})$ and $\bar{\mathbf{r}}(\mathbf{a})$ (defined in equations (6.7) and (6.17)) and equations (6.13), (6.14) and (6.16), we can also express $\nabla\psi(\mathbf{a})$ as

$$\nabla\psi(\mathbf{a}) = \sum_{j=1}^{n} \widetilde{\mathbf{M}}_j^T \widetilde{\mathbf{r}}_j = \sum_{j=1}^{n} \bar{\mathbf{M}}_j^T \bar{\mathbf{r}}_j \; .$$

This demonstrates that the evaluation of $\nabla\psi(\mathbf{a})$ can also be easily evaluated and parallelized in the normal-equation framework. An alternative approach to evaluate $\nabla\psi(\mathbf{a})$ in parallel is to use Theorems 4.3 and 5.7 and the equality

$$\nabla\psi(\mathbf{a}) = \frac{\partial\varphi^*(\mathbf{A}, \widehat{\mathbf{B}})}{\partial \mathbf{a}} = \mathbf{G}(\widehat{\mathbf{b}})^T \mathbf{G}(\widehat{\mathbf{b}})\mathbf{a} - \mathbf{G}(\widehat{\mathbf{b}})^T \mathbf{z} \; ,$$

where $\mathbf{z} = vec\big((\sqrt{\mathbf{W}} \odot \mathbf{X})^T\big)$, as already indicated in the formal description of the variable projection Gauss-Newton algorithms (1) at the beginning of this section. Taking into account the diagonal block-structure of $\mathbf{G}(\widehat{\mathbf{b}})$ (see equation (3.22)), the evaluation of $\nabla\psi(\mathbf{a})$ using this last formulation can also be parallelized very efficiently.

We now explain how to compute the correction vectors $d\mathbf{a}_{gp-gn}$ and $d\mathbf{a}_{k-gn}$ in the Gauss-Newton algorithms (1) using both the TSQR and normal-equation approaches described above.

In the TSQR approach, we compute implicitly a thin QR decomposition of $-J(\mathbf{r}(\mathbf{a}))$ or $\mathbf{M}(\mathbf{a})$ in two stages. For the Jacobian matrix $J(\mathbf{r}(\mathbf{a}))$, these two steps are as follows

(a) $-J\big(\mathbf{r}(\mathbf{a})\big) = \mathbf{Q}_\mathbf{F} \widetilde{J}\big(\mathbf{r}(\mathbf{a})\big)$,

(b) $\widetilde{J}\big(\mathbf{r}(\mathbf{a})\big) = \widetilde{\mathbf{Q}}_J \mathbf{R}_J$,

giving the thin QR factorization of $-J(\mathbf{r}(\mathbf{a})) = \mathbf{M}(\mathbf{a}) + \mathbf{L}(\mathbf{a})$ as

$$\mathbf{M}(\mathbf{a}) + \mathbf{L}(\mathbf{a}) = (\mathbf{Q}_\mathbf{F} \widetilde{\mathbf{Q}}_J)\mathbf{R}_J = \mathbf{Q}_J \mathbf{R}_J \; , \tag{6.20}$$

where $\mathbf{Q}_J$ is an $p.n \times k.p$ matrix with orthonormal columns and $\mathbf{R}_J$ is an $k.p \times k.p$ upper triangular matrix. Similarly, if we use the approximate Jacobian matrix $-\mathbf{M}(\mathbf{a})$, we have the two steps:

(a) $\mathbf{M}(\mathbf{a}) = \mathbf{Q}_\mathbf{F} \widetilde{\mathbf{M}}(\mathbf{a})$,

(b) $\widetilde{\mathbf{M}}(\mathbf{a}) = \widetilde{\mathbf{Q}}_\mathbf{M} \mathbf{R}_\mathbf{M}$,

giving the thin QR factorization of $\mathbf{M}(\mathbf{a})$ as

$$\mathbf{M}(\mathbf{a}) = (\mathbf{Q}_\mathbf{F} \widetilde{\mathbf{Q}}_\mathbf{M})\mathbf{R}_\mathbf{M} = \mathbf{Q}_\mathbf{M} \mathbf{R}_\mathbf{M} \; , \tag{6.21}$$

where, again, $\mathbf{Q}_\mathbf{M}$ is an $p.n \times k.p$ matrix with orthonormal columns and $\mathbf{R}_\mathbf{M}$ is an $k.p \times k.p$ upper triangular matrix.

As noted above, $\widetilde{\mathbf{M}}(\mathbf{a})$ has $r_{\mathbf{F}(\mathbf{a})}$ zero rows and these zero rows can be eliminated in the second step, giving the following simplified step

(b') $\bar{\mathbf{M}}(\mathbf{a}) = \bar{\mathbf{Q}}_\mathbf{M} \mathbf{R}_\mathbf{M}$,

for the computation of the upper triangular factor $\mathbf{R_M}$ in the Kaufman variant of the Gauss-Newton algorithm. Note that the matrices $\mathbf{Q}_J$ and $\mathbf{Q_M}$ are never explicitly computed and all that is needed to solve the associated linear least-squares problems (6.10) and (6.12), and compute the correction vectors $d\mathbf{a}_{gp-gn}$ and $d\mathbf{a}_{k-gn}$ are the triangular factor $\mathbf{R}_J$ and the matrix-vector product $\mathbf{Q}_J^T\mathbf{r}(\mathbf{a})$ in the case of the Golub-Pereyra variant or the triangular factor $\mathbf{R_M}$ and the matrix-vector product $\mathbf{Q}_\mathbf{M}^T\mathbf{r}(\mathbf{a})$ in the case of the Kaufman variant. These two matrix-vector products are given, respectively, by

$$\mathbf{Q}_J^T\mathbf{r}(\mathbf{a}) = \widetilde{\mathbf{Q}}_J^T\mathbf{Q}_\mathbf{F}^T\mathbf{r}(\mathbf{a}) = \widetilde{\mathbf{Q}}_J^T\tilde{\mathbf{r}}(\mathbf{a}) \tag{6.22}$$

and

$$\mathbf{Q}_\mathbf{M}^T\mathbf{r}(\mathbf{a}) = \widetilde{\mathbf{Q}}_\mathbf{M}^T\mathbf{Q}_\mathbf{F}^T\mathbf{r}(\mathbf{a}) = \widetilde{\mathbf{Q}}_\mathbf{M}^T\tilde{\mathbf{r}}(\mathbf{a}) = \bar{\mathbf{Q}}_\mathbf{M}^T\bar{\mathbf{r}}(\mathbf{a}) \; . \tag{6.23}$$

As explained in the previous paragraphs, both $\mathbf{Q}_J^T\mathbf{r}(\mathbf{a})$ and $\mathbf{Q}_\mathbf{M}^T\mathbf{r}(\mathbf{a})$ can be computed recursively, as the triangular factors $\mathbf{R}_J$ and and $\mathbf{R_M}$, and without explicitly computing the matrices $\mathbf{Q}_J$ and $\mathbf{Q_M}$ with the help of the (parallel) TSQR algorithm.

Now, if we assume that the triangular matrices $\mathbf{R}_J$ and $\mathbf{R_M}$ are of full rank (which is equivalent to assume that $-J(\mathbf{r}(\mathbf{a}))$ and $\mathbf{M}(\mathbf{a})$ are of full column-rank), the unique solutions of the associated linear least-squares problems (6.10) and (6.12) are simply found by solving the following upper triangular systems, for $d\mathbf{a} \in \mathbb{R}^{k.p}$,

$$\mathbf{R}_J d\mathbf{a} = \mathbf{Q}_J^T\mathbf{r}(\mathbf{a}) \quad \text{and} \quad \mathbf{R_M}d\mathbf{a} = \mathbf{Q}_\mathbf{M}^T\mathbf{r}(\mathbf{a}) \; ,$$

since the 2-norm is unitarily invariant. Thus, in this case, we obtain

$$d\mathbf{a}_{gp-gn} = \mathbf{R}_J^{-1}\mathbf{Q}_J^T\mathbf{r}(\mathbf{a}) \quad \text{and} \quad d\mathbf{a}_{k-gn} = \mathbf{R}_\mathbf{M}^{-1}\mathbf{Q}_\mathbf{M}^T\mathbf{r}(\mathbf{a}) \; .$$

Of course, in our case, this simple approach cannot be used as we already know that the triangular factors $\mathbf{R}_J$ and $\mathbf{R_M}$ are rank deficient since $-J(\mathbf{r}(\mathbf{a}))$ and $\mathbf{M}(\mathbf{a})$ are always column rank deficient (see Theorem 5.2). Thus, we will discuss how to proceed in order to compute the minimum 2-norm solutions of these triangular systems (and, thus, of the associated linear least-squares problems (6.10) and (6.12)) using the theoretical results of Subsection 5.2 once we have presented how the correction vectors $d\mathbf{a}_{gp-gn}$ and $d\mathbf{a}_{k-gn}$ can be computed in the normal-equation approach.

In the normal-equation approach, the correction vectors $d\mathbf{a}_{gp-gn}$ and $d\mathbf{a}_{k-gn}$ can be found by computing, respectively, the matrices Δ and Λ, as described above, and solving the associated normal equations, namely,

$$\Delta d\mathbf{a} = -\nabla\psi(\mathbf{a}) \quad \text{or} \quad \Lambda d\mathbf{a} = -\nabla\psi(\mathbf{a}) \; ,$$

where $\nabla\psi(\mathbf{a}) = J(\mathbf{r}(\mathbf{a}))^T\mathbf{r}(\mathbf{a})$ is also evaluated by one of the methods described in the preceding paragraphs. Since Δ and Λ are symmetric and positive semi-definite matrices, these normal equations are usually solved by computing the Cholesky factorization of Δ and Λ, namely,

$$\Delta = \mathbf{R}_\Delta^T\mathbf{R}_\Delta \quad \text{or} \quad \Lambda = \mathbf{R}_\Lambda^T\mathbf{R}_\Lambda \; ,$$

where $\mathbf{R}_\Delta$ and $\mathbf{R}_\Lambda$ are $k.p \times k.p$ upper-triangular matrices. Then, if we assume that Δ and Λ are of full rank and, thus, positive definite (which is again equivalent to assume that $J(\mathbf{r}(\mathbf{a}))$ and $\mathbf{M}(\mathbf{a})$ are of full column-rank), the normal equations can be solved by backward and forward substitutions using these Cholesky triangular factors. For example, assuming that Δ is nonsingular, we first solve, for $d\mathbf{a}_\Delta \in \mathbb{R}^{k.p}$, the triangular system

$$\mathbf{R}_\Delta^T d\mathbf{a}_\Delta = -J\big(\mathbf{r}(\mathbf{a})\big)^T\mathbf{r}(\mathbf{a}) \Longrightarrow d\mathbf{a}_\Delta = -\mathbf{R}_\Delta^{-T}J\big(\mathbf{r}(\mathbf{a})\big)^T\mathbf{r}(\mathbf{a}) \; ,$$

and, then, solve for $d\mathbf{a}_{gp-gn} \in \mathbb{R}^{k.p}$,

$$\mathbf{R}_\Delta d\mathbf{a}_{gp-gn} = d\mathbf{a}_\Delta \Longrightarrow d\mathbf{a}_{gp-gn} = \mathbf{R}_\Delta^{-1}d\mathbf{a}_\Delta \; .$$

Note that, up to the sign of the rows of the matrices, we have the equality

$$\mathbf{R}_\Delta = \mathbf{R}_J \ ,$$

and, also up to the sign of the elements of the vectors, the equality

$$d\mathbf{a}_\Delta = \mathbf{Q}_J^T \mathbf{r}(\mathbf{a}) \ ,$$

where $\mathbf{Q}_J$ and $\mathbf{R}_J$ are, respectively, a $n.p \times k.p$ matrix with orthonormal columns and n $k.p \times k.p$ upper triangular matrix, which define the two matrix factors of the QR decomposition of $-J(\mathbf{r}(\mathbf{a}))$ in equation (6.20). Obviously, similar results are valid for Λ. In such conditions, we have, thus, the equivalences

$$\mathbf{R}_\Delta d\mathbf{a}_{gp-gn} = d\mathbf{a}_\Delta \Longleftrightarrow \mathbf{R}_J d\mathbf{a}_{gp-gn} = \mathbf{Q}_J^T \mathbf{r}(\mathbf{a})$$

and

$$\mathbf{R}_\Lambda d\mathbf{a}_{k-gn} = d\mathbf{a}_\Lambda \Longleftrightarrow \mathbf{R}_\mathbf{M} d\mathbf{a}_{k-gn} = \mathbf{Q}_\mathbf{M}^T \mathbf{r}(\mathbf{a}) \ ,$$

which establish the equivalence between the QR and Cholesky approaches when we assume that $-J(\mathbf{r}(\mathbf{a}))$ and $\mathbf{M}(\mathbf{a})$ are of full column rank and that the computations are performed with exact arithmetic without any roundoff errors. However, again, $d\mathbf{a}_{gp-gn}$ and $d\mathbf{a}_{k-gn}$ cannot be found in this simple way with the normal-equation approach as we already know that the symmetric matrices Δ and Λ are only positive semi-definite since $-J(\mathbf{r}(\mathbf{a}))$ and $\mathbf{M}(\mathbf{a})$ are always rank-deficient and we will also come back to this problem after discussing the respective merits of the QR and normal-equation approaches in more details.

Because $n.p$ is in most cases much larger than $k.p$ (assuming that $p < n$ and $k \ll p$) and the computation of the QR decomposition of the (approximated) Jacobian matrix is the major portion of the time for the variable projection WLRA solvers, which use this QR approach (despite of the use of a parallel TSQR algorithm and BLAS3 kernels for the structured QR decompositions in it), a standard normal-equation approach can be much faster than the QR approach. However, while the normal-equation approach is extremely efficient in terms of work and speed, it may be also numerically unreliable for ill-conditioned linear least-squares problems as it is well know [111][71][87][8]. Furthermore, as $-J(\mathbf{r}(\mathbf{a}))$ and $\mathbf{M}(\mathbf{a})$ are always rank-deficient, but their precise rank is not known if $\mathbf{W} \in \mathbb{R}_+^{p \times n}$ and the number of zero weights in $\mathbf{W}$ is large (see Section 5), it is important to have reliable information about the rank of these matrices, which can be obtained theoretically from the triangular factors $\mathbf{R}_J$ and $\mathbf{R}_\mathbf{M}$ in the QR approach or $\mathbf{R}_\Delta$ and $\mathbf{R}_\Lambda$ in the normal equations approach. However, the errors in the computation of $\mathbf{R}_J$ and $\mathbf{R}_\mathbf{M}$ depend on the condition number of $-J(\mathbf{r}(\mathbf{a}))$ and $\mathbf{M}(\mathbf{a})$, while that of $\mathbf{R}_\Delta$ and $\mathbf{R}_\Lambda$ depend on the square of the condition numbers of $-J(\mathbf{r}(\mathbf{a}))$ and $\mathbf{M}(\mathbf{a})$ [111][71][87][8]. Thus, the QR approach can still be preferred for stability and accuracy reasons, especially when the number of zero weights in $\mathbf{W}$ is large.

More precisely, from the results of Subsection 5.2, we know that the matrices $-J(\mathbf{r}(\mathbf{a}))$ and $\mathbf{M}(\mathbf{a})$ have a rank r at most equal to $(p-k).k$ if *rank*$(\mathbf{A}) = k$ (see Theorem 5.2 and Corollary 5.3; recall also that the condition *rank*$(\mathbf{A}) = k$ is required for the continuity and differentiability of $\mathbf{P}_{\mathbf{F}(.)}^\perp$, $\mathbf{F}(.)^+$ and $\mathbf{F}(.)^-$ at $\mathbf{a}$), but that r can be smaller than $(p-k).k$ depending on the number of zero weights or missing values in $\mathbf{X}$ (see Theorems 5.5 and 5.6 for details). Standard tools for solving linear least-squares problems with such deficient matrices are the SVD or the COD, as outlined in Subsection 2.1, which both allow to estimate the solution vectors $d\mathbf{a}_{gp-gn}$ and $d\mathbf{a}_{k-gn}$ of minimum 2-norm of problems (6.11) and (6.15) with the help of the generalized inverse of $\mathbf{R}_J$ or $\mathbf{R}_\mathbf{M}$ in the QR approach.

As an illustration, if the SVD of the upper triangular matrix $\mathbf{R}_J$ is given by

$$\mathbf{R}_J = \boldsymbol{U_J} \boldsymbol{D_J} \boldsymbol{V_J}^T \ ,$$

where $\boldsymbol{U_J}$ and $\boldsymbol{V_J}$ are $k.p \times k.p$ orthogonal matrices and $\boldsymbol{D_J}$ is a $k.p \times k.p$ diagonal matrix with its diagonal elements equal to the singular values of $\mathbf{R}_J$ in decreasing order with at least its $k.k$

last diagonal elements equal to zero according to Theorem 5.2. As stated in equation (2.12), we have

$$\mathbf{R}_J^+ = \boldsymbol{V_J} \boldsymbol{D_J}^+ \boldsymbol{U_J}^T ,$$

where $[\boldsymbol{D_J}^+]_{ii} = [\boldsymbol{D_J}]_{ii}^{-1}$ if $[\boldsymbol{D_J}]_{ii} \neq 0$ and $[\boldsymbol{D_J}^+]_{ii} = 0$ if $[\boldsymbol{D_J}]_{ii} = 0$ and the correction vector $d\mathbf{a}_{gp-gn}$ of minimum 2-norm can be computed as

$$d\mathbf{a}_{gp-gn} = \mathbf{R}_J^+ \mathbf{Q}_J^T \mathbf{r}(\mathbf{a})$$

and, similarly, if we use the Kaufman variant of the Gauss-Newton algorithm, the correction vector $d\mathbf{a}_{k-gn}$ of minimum 2-norm can be estimated as

$$d\mathbf{a}_{k-gn} = \mathbf{R}_\mathbf{M}^+ \mathbf{Q}_\mathbf{M}^T \mathbf{r}(\mathbf{a}) = \mathbf{R}_\mathbf{M}^+ \bar{\mathbf{Q}}_\mathbf{M}^T \bar{\mathbf{r}}(\mathbf{a}) .$$

Note that if any $[\boldsymbol{D_J}]_{ii}$ or $[\boldsymbol{D}_\mathbf{M}]_{ii}$ is small, but non-zero, these computations can be numerically unstable, which makes important to consider approximate methods, which can provide control over the size of $d\mathbf{a}_{gp-gn}$ or $d\mathbf{a}_{k-gn}$. This leads to consider low rank estimates of $\mathbf{R}_J$ and $\mathbf{R}_\mathbf{M}$ by considering only their singular values, which are above a suitable threshold $\nu \in \mathbb{R}_{+*}$ in the SVDs of $\mathbf{R}_J$ and $\mathbf{R}_\mathbf{M}$, and, finally, in the computations of $\mathbf{R}_J^+$ and $d\mathbf{a}_{gp-gn}$ or, alternatively, of $\mathbf{R}_M^+$ and $d\mathbf{a}_{k-gn}$.

As an illustration, for the Golub-Pereyra variant of the Gauss-Newton algorithm, without a threshold, we will have

$$d\mathbf{a}_{gp-gn} = \sum_{[\boldsymbol{D_J}]_{ii}>0} \frac{\left(\mathbf{Q}_J^T \mathbf{r}(\mathbf{a})\right)^T [\boldsymbol{U_J}]_{.i}}{[\boldsymbol{D_J}]_{ii}} [\boldsymbol{V_J}]_{.i} ,$$

but, using the threshold ν, we will get

$$d\mathbf{a}_{gp-gn} = \sum_{[\boldsymbol{D_J}]_{ii}>\nu} \frac{\left(\mathbf{Q}_J^T \mathbf{r}(\mathbf{a})\right)^T [\boldsymbol{U_J}]_{.i}}{[\boldsymbol{D_J}]_{ii}} [\boldsymbol{V_J}]_{.i} ,$$

which obviously limits the potential occurrence of large elements in $d\mathbf{a}_{gp-gn}$. Alternatively, we can estimate $\mathbf{R}_J^+$ and $\mathbf{R}_\mathbf{M}^+$ by a COD of $\mathbf{R}_J$ and $\mathbf{R}_\mathbf{M}$ as described in Subsection 2.1. This will be less time consuming, but also less reliable in estimating precisely the ranks of $-J(\mathbf{r}(\mathbf{a}))$ and $\mathbf{M}(\mathbf{a})$. However, recent investigations in the context of general NLLS problems suggest that using a COD can even be more reliable than the truncated SVD approach described above, see [90] for details.

Similarly, in the normal-equation approach, we can compute the Eigenvalue-Vector Decomposition (EVD) of the positive semi-definite matrices Δ (or Λ) and use a truncated EVD to estimate its pseudo-inverses Δ^+ (or Λ^+) and, finally, $d\mathbf{a}_{gp-gn}$ (or $d\mathbf{a}_{k-gn}$) as

$$d\mathbf{a}_{gp-gn} = -\sum_{[\boldsymbol{D_J}]_{ii}^2>\nu^2} \frac{\left(J(\mathbf{r}(\mathbf{a}))^T \mathbf{r}(\mathbf{a})\right)^T [\boldsymbol{V_J}]_{.i}}{[\boldsymbol{D_J}]_{ii}^2} [\boldsymbol{V_J}]_{.i} ,$$

where the EVD of Δ is given by

$$\Delta = \boldsymbol{V_J} \boldsymbol{D_J}^2 \boldsymbol{V_J}^T ,$$

where the matrices $\boldsymbol{V_J}$ and $\boldsymbol{D_J}$ have the same meaning as in the SVD of $\mathbf{R}_J$.

However, if p is large and the rank k of the WLRA matrix approximation we are seeking is also not small, the dimensions of $\mathbf{R}_J$ and $\mathbf{R}_\mathbf{M}$ in the TSQR approach, or, Δ and Λ in the normal-equation approach, can be very large. In these conditions, computing a SVD, or even a COD, of $\mathbf{R}_J$ (or $\mathbf{R}_\mathbf{M}$) in the TSQR approach or an EVD of Δ (or Λ) in the normal-equation approach, at each iteration of the variable projection Gauss-Newton algorithms (1), can be very costly and, consequently, must be avoided as much as possible.

In many practical applications, for example if $\mathbf{W} \in \mathbb{R}^{p \times n}_{+*}$ or if the number of observed values in each column and row of the data matrix $\mathbf{X}$ is larger than the rank k of the matrix approximation we are seeking, we also know that $r = \textit{rank}(J(\mathbf{r}(\mathbf{a})))$ will be exactly equal to $(p-k).k$ with high probability (see Theorem 5.3 and Corollaries 5.4 and 5.5), the last $k.k$ columns of $J(\mathbf{r}(\mathbf{a}))$ and $\mathbf{M}(\mathbf{a})$ are linearly dependent upon the first $(p-k).k$ columns of these matrices (see Theorem 5.4) and, finally, that it is easy to compute an orthonormal basis of the null-space of $J(\mathbf{r}(\mathbf{a}))$ and $\mathbf{M}(\mathbf{a})$ from an orthonormal basis of $\mathbf{A}$ if $\textit{rank}(\mathbf{A}) = k$ (see Corollary 5.6). Collectively, these different results suggest that the minimum 2-norm solutions of the linear least-squares problems (6.10) and (6.12) involving the rank deficient matrices $-J(\mathbf{r}(\mathbf{a}))$ and $\mathbf{M}(\mathbf{a})$ can still be found accurately and much more efficiently without resorting to costly techniques like the SVD or a full COD in the TSQR approach or the EVD in the normal-equation framework as already illustrated in Subsection 5.2.

We first reconsider the TSQR approach in which we want to find the minimum 2-norm solutions $d\mathbf{a}_{gp-gn}$ or $d\mathbf{a}_{k-gn}$ of the rank-deficient, but consistent upper triangular systems,

$$\mathbf{R}_J d\mathbf{a}_{gp-gn} = \mathbf{Q}_J^T \mathbf{r}(\mathbf{a}) \text{ or } \mathbf{R}_\mathbf{M} d\mathbf{a}_{k-gn} = \mathbf{Q}_\mathbf{M}^T \mathbf{r}(\mathbf{a}) \ ,$$

assuming that

$$\textit{rank}(\mathbf{R}_J) = \textit{rank}(J(\mathbf{r}(\mathbf{a}))) = (p-k).k$$

and, similarly, that

$$\textit{rank}(\mathbf{R}_\mathbf{M}) = \textit{rank}(\mathbf{M}(\mathbf{a})) = (p-k).k \ .$$

Under the hypotheses of Theorem 5.4, we know that the first $(p-k).k$ columns of $-J(\mathbf{r}(\mathbf{a}))$ and $\mathbf{M}(\mathbf{a})$ are linearly independent and that the last $k.k$ columns of these matrices are linearly dependent onto the first $(p-k).k$ columns of these matrices. Obviously, the same relationships hold for $\mathbf{R}_J$ and $\mathbf{R}_\mathbf{M}$ are these matrices are, respectively, the triangular factors in the QR factorizations of $-J(\mathbf{r}(\mathbf{a}))$ and $\mathbf{M}(\mathbf{a})$. Then, it is possible to compute the vectors $d\mathbf{a}_{gp-gn}$ or $d\mathbf{a}_{k-gn}$ efficiently as follows.

First, we define the following partitions of the two matrix factors in the thin QR decompositions of $-J(\mathbf{r}(\mathbf{a}))$ and $\mathbf{M}(\mathbf{a})$, defined in equations (6.20) and (6.21), which correspond to the partitions of $J(\mathbf{r}(\mathbf{a}))$ and $\mathbf{M}(\mathbf{a})$ used in Theorem 5.4:

$$\mathbf{Q}_J = \begin{bmatrix} \mathbf{Q}_J^1 & \mathbf{Q}_J^2 \end{bmatrix} \ , \mathbf{R}_J = \begin{bmatrix} \mathbf{R}_J^{11} & \mathbf{R}_J^{12} \\ \mathbf{0}^{k.k \times (p-k).k} & \mathbf{R}_J^{22} \end{bmatrix} \ ,$$

and

$$\mathbf{Q}_\mathbf{M} = \begin{bmatrix} \mathbf{Q}_\mathbf{M}^1 & \mathbf{Q}_\mathbf{M}^2 \end{bmatrix} \ , \mathbf{R}_\mathbf{M} = \begin{bmatrix} \mathbf{R}_\mathbf{M}^{11} & \mathbf{R}_\mathbf{M}^{12} \\ \mathbf{0}^{k.k \times (p-k).k} & \mathbf{R}_\mathbf{M}^{22} \end{bmatrix} \ ,$$

where

- $\mathbf{Q}_J^1$ and $\mathbf{Q}_\mathbf{M}^1 \in \mathbb{O}^{p.n \times (p-k).k}$,
- $\mathbf{Q}_J^2$ and $\mathbf{Q}_\mathbf{M}^2 \in \mathbb{O}^{p.n \times k.k}$,
- $\mathbf{R}_J^{11}$ and $\mathbf{R}_\mathbf{M}^{11}$ are $(p-k).k \times (p-k).k$ upper triangular matrices ,
- $\mathbf{R}_J^{22}$ and $\mathbf{R}_\mathbf{M}^{22}$ are $k.k \times k.k$ upper triangular matrices ,
- $\mathbf{R}_J^{12}$ and $\mathbf{R}_\mathbf{M}^{12}$ are $(p-k).k \times k.k$ full matrices .

From Theorem 5.4, we deduce immediately that

$$\mathbf{R}_J^{22} = \mathbf{R}_\mathbf{M}^{22} = \mathbf{0}^{k.k \times k.k} \ ,$$

as $\mathbf{R}_J$ and $\mathbf{R}_\mathbf{M}$ exhibit the same linear dependencies as $J(\mathbf{r}(\mathbf{a}))$ and $\mathbf{M}(\mathbf{a})$.

Next, applying $(p-k).k$ Householder transformations to the right of $\begin{bmatrix}\mathbf{R}_J^{11} & \mathbf{R}_J^{12}\end{bmatrix}$ and $\begin{bmatrix}\mathbf{R}_{\mathbf{M}}^{11} & \mathbf{R}_{\mathbf{M}}^{12}\end{bmatrix}$ to annihilate $\mathbf{R}_J^{12}$ and $\mathbf{R}_{\mathbf{M}}^{12}$, we obtain the following simplified CODs of $\mathbf{R}_J$ and $\mathbf{R}_{\mathbf{M}}$:

$$\mathbf{R}_J = \begin{bmatrix}\mathbf{R}_J^{11} & \mathbf{R}_J^{12} \\ \mathbf{0}^{k.k\times(p-k).k} & \mathbf{0}^{k.k\times k.k}\end{bmatrix} = \begin{bmatrix}\mathbf{T}_J & \mathbf{0}^{(p-k).k\times k.k} \\ \mathbf{0}^{k.k\times(p-k).k} & \mathbf{0}^{k.k\times k.k}\end{bmatrix}\mathbf{Z}_J^T$$

and

$$\mathbf{R}_{\mathbf{M}} = \begin{bmatrix}\mathbf{R}_{\mathbf{M}}^{11} & \mathbf{R}_{\mathbf{M}}^{12} \\ \mathbf{0}^{k.k\times(p-k).k} & \mathbf{0}^{k.k\times k.k}\end{bmatrix} = \begin{bmatrix}\mathbf{T}_{\mathbf{M}} & \mathbf{0}^{(p-k).k\times k.k} \\ \mathbf{0}^{k.k\times(p-k).k} & \mathbf{0}^{k.k\times k.k}\end{bmatrix}\mathbf{Z}_{\mathbf{M}}^T\,,$$

where $\mathbf{T}_J$ and $\mathbf{T}_{\mathbf{M}}$ are $(p-k).k\times(p-k).k$ nonsingular upper triangular matrices, and, $\mathbf{Z}_J$ and $\mathbf{Z}_{\mathbf{M}}$ are $k.p\times k.p$ orthogonal matrices, which are the product of $(p-k).k$ Householder transformations designed to annihilate $\mathbf{R}_J^{12}$ and $\mathbf{R}_{\mathbf{M}}^{12}$, respectively. In doing so, we implicitly obtain the following CODs of $-J(\mathbf{r}(\mathbf{a}))$ or $\mathbf{M}(\mathbf{a})$ from their thin QR decompositions (defined in equations (6.20) and (6.21)) and computed by the TSQR algorithm:

$$-J(\mathbf{r}(\mathbf{a})) = \mathbf{Q}_J\mathbf{R}_J = \mathbf{Q}_J\begin{bmatrix}\mathbf{T}_J & \mathbf{0}^{(p-k).k\times k.k} \\ \mathbf{0}^{k.k\times(p-k).k} & \mathbf{0}^{k.k\times k.k}\end{bmatrix}\mathbf{Z}_J^T$$

and

$$\mathbf{M}(\mathbf{a}) = \mathbf{Q}_{\mathbf{M}}\mathbf{R}_{\mathbf{M}} = \mathbf{Q}_{\mathbf{M}}\begin{bmatrix}\mathbf{T}_{\mathbf{M}} & \mathbf{0}^{(p-k).k\times k.k} \\ \mathbf{0}^{k.k\times(p-k).k} & \mathbf{0}^{k.k\times k.k}\end{bmatrix}\mathbf{Z}_{\mathbf{M}}^T\,,$$

and we can express $-J(\mathbf{r}(\mathbf{a}))^+$ or $\mathbf{M}(\mathbf{a})^+$ as

$$-J(\mathbf{r}(\mathbf{a}))^+ = \mathbf{Z}_J\begin{bmatrix}\mathbf{T}_J^{-1} & \mathbf{0}^{(p-k).k\times k.k} \\ \mathbf{0}^{k.k\times(p-k).k} & \mathbf{0}^{k.k\times k.k}\end{bmatrix}\mathbf{Q}_J^T = \mathbf{Z}_J\begin{bmatrix}\mathbf{T}_J^{-1}(\mathbf{Q}_J^1)^T \\ \mathbf{0}^{k.k\times(p-k).k}\end{bmatrix}$$

and

$$\mathbf{M}(\mathbf{a})^+ = \mathbf{Z}_{\mathbf{M}}\begin{bmatrix}\mathbf{T}_{\mathbf{M}}^{-1} & \mathbf{0}^{(p-k).k\times k.k} \\ \mathbf{0}^{k.k\times(p-k).k} & \mathbf{0}^{k.k\times k.k}\end{bmatrix}\mathbf{Q}_{\mathbf{M}}^T = \mathbf{Z}_{\mathbf{M}}\begin{bmatrix}\mathbf{T}_{\mathbf{M}}^{-1}(\mathbf{Q}_{\mathbf{M}}^1)^T \\ \mathbf{0}^{k.k\times(p-k).k}\end{bmatrix}\,.$$

Finally, $d\mathbf{a}_{gp-gn}$ and $d\mathbf{a}_{k-gn}$ can be computed as

$$d\mathbf{a}_{gp-gn} = -J(\mathbf{r}(\mathbf{a}))^+\mathbf{r}(\mathbf{a}) = \mathbf{Z}_J\begin{bmatrix}\mathbf{T}_J^{-1}(\mathbf{Q}_J^1)^T\mathbf{r}(\mathbf{a}) \\ \mathbf{0}^{k.k\times(p-k).k}\end{bmatrix} = \mathbf{Z}_J\begin{bmatrix}\mathbf{T}_J^{-1}(\widetilde{\mathbf{Q}}_J^1)^T\widetilde{\mathbf{r}}(\mathbf{a}) \\ \mathbf{0}^{k.k\times(p-k).k}\end{bmatrix}$$

and

$$d\mathbf{a}_{k-gn} = \mathbf{M}(\mathbf{a})^+\mathbf{r}(\mathbf{a}) = \mathbf{Z}_{\mathbf{M}}\begin{bmatrix}\mathbf{T}_{\mathbf{M}}^{-1}(\mathbf{Q}_{\mathbf{M}}^1)^T\mathbf{r}(\mathbf{a}) \\ \mathbf{0}^{k.k\times(p-k).k}\end{bmatrix} = \mathbf{Z}_{\mathbf{M}}\begin{bmatrix}\mathbf{T}_{\mathbf{M}}^{-1}(\bar{\mathbf{Q}}_{\mathbf{M}}^1)^T\bar{\mathbf{r}}(\mathbf{a}) \\ \mathbf{0}^{k.k\times(p-k).k}\end{bmatrix},$$

where we have used equations (6.22) and (6.23), and, in both cases, the matrix expressions on the right hand-side of these equalities are available on output of the TSQR algorithm.

Alternatively, the correction vectors $d\mathbf{a}_{gp-gn}$ and $d\mathbf{a}_{k-gn}$ can be computed with the help of Theorem 5.3 and Corollary 5.6, which state that the matrix

$$\mathbf{N} = \mathbf{K}_{(p,k)}(\mathbf{I}_k \otimes \mathbf{A})$$

is a matrix of full column rank and that the columns of $\mathbf{N}$ form a basis of $\mathit{null}\big(J(\mathbf{r}(\mathbf{a}))\big) = \mathit{null}(\mathbf{M}(\mathbf{a}))$, if $\mathit{rank}(\mathbf{A}) = k$ and the hypotheses of Theorem 5.3 are verified. Note that the condition $\mathit{rank}(\mathbf{A}) = k$ is always verified if step **(0)** of the Gauss-Newton algorithms (1) is performed at each iteration. On the other hand, the hypotheses of Theorem 5.3 can be violated or, more generally, the condition $\mathit{rank}(J(\mathbf{r}(\mathbf{a}))) = \mathit{rank}(\mathbf{M}(\mathbf{a})) = (p-k).k$ can be not verified, especially, in the case of a very large number of missing values in $\mathbf{X}$ or zero weights in $\mathbf{W}$ as demonstrated in Theorems 5.5 and 5.6. However, as noted at the end of Subsection 5.2 , if we restrict the set of WLRA problems by imposing the condition $\sum_{l=1}^{p}\boldsymbol{\delta}_{lj} > k$ for all $j = 1,\cdots,n$, where $\boldsymbol{\delta}$ is the incidence matrix associated with the matrix $\mathbf{X}$, the condition $\mathit{rank}(J(\mathbf{r}(\mathbf{a}))) = \mathit{rank}(\mathbf{M}(\mathbf{a})) = (p-k).k$ will be also verified in

the majority of practical applications. In this scenario, the columns of $\mathbf{N}$ form also an orthonormal basis of $null\big(J(\mathbf{r}(\mathbf{a}))\big) = null(\mathbf{M}(\mathbf{a}))$ if step **(0)** of the Gauss-Newton algorithms (1) is used at each iteration according to Corollary 5.6, since $\mathbf{A}$ has orthonormal columns after step **(0)** is performed. Then, using the results of Section 5.2, it is not difficult to see that the vector $d\mathbf{a}_{gp-gn}$ is the unique solution of the structured linear system

$$\begin{bmatrix} \mathbf{R}_J \\ \mathbf{N}^T \end{bmatrix} d\mathbf{a}_{gp-gn} = \begin{bmatrix} \widetilde{\mathbf{Q}}_J^T \widetilde{\mathbf{r}}(\mathbf{a}) \\ \mathbf{0}^{k.k \times k.p} \end{bmatrix} ,$$

which can be solved by computing the structured thin QR decomposition of $\begin{bmatrix} \mathbf{R}_J \\ \mathbf{N}^T \end{bmatrix}$ in a first step as

$$\begin{bmatrix} \mathbf{R}_J \\ \mathbf{N}^T \end{bmatrix} = \mathbf{Q}_J(\mathbf{N})\mathbf{R}_J(\mathbf{N}) ,$$

where $\mathbf{Q}_J(\mathbf{N})$ is an $k.(p+k) \times k.p$ matrix with orthonormal columns and $\mathbf{R}_J(\mathbf{N})$ is an $k.p \times k.p$ nonsingular upper triangular matrix. In these conditions, $d\mathbf{a}_{gp-gn}$ is the unique solution of the upper triangular system

$$\mathbf{R}_J(\mathbf{N}) d\mathbf{a}_{gp-gn} = \mathbf{Q}_J(\mathbf{N})^T \begin{bmatrix} \widetilde{\mathbf{Q}}_J^T \widetilde{\mathbf{r}}(\mathbf{a}) \\ \mathbf{0}^{k.k \times k.p} \end{bmatrix} ,$$

which can be easily solved by backward substitution. In a similar fashion and using the same notations, $d\mathbf{a}_{k-gn}$ can be evaluated by computing the structured thin QR decomposition of $\begin{bmatrix} \mathbf{R}_\mathbf{M} \\ \mathbf{N}^T \end{bmatrix}$ in a first step

$$\begin{bmatrix} \mathbf{R}_\mathbf{M} \\ \mathbf{N}^T \end{bmatrix} = \mathbf{Q}_\mathbf{M}(\mathbf{N})\mathbf{R}_\mathbf{M}(\mathbf{N}) ,$$

and by solving the following nonsingular upper triangular system in the second step

$$\mathbf{R}_\mathbf{M}(\mathbf{N}) d\mathbf{a}_{k-gn} = \mathbf{Q}_\mathbf{M}(\mathbf{N})^T \begin{bmatrix} \bar{\mathbf{Q}}_\mathbf{M}^T \bar{\mathbf{r}}(\mathbf{a}) \\ \mathbf{0}^{k.k \times k.p} \end{bmatrix} .$$

Recall, finally, that solving these upper triangular systems for $d\mathbf{a}_{gp-gn}$ and $d\mathbf{a}_{k-gn}$ in output of the TSQR algorithm is equivalent to find, respectively, the unique solutions of the following "constrained" linear least-squares problems

$$d\mathbf{a}_{gp-gn} = \text{Arg} \min_{d\mathbf{a} \in \mathbb{R}^{p.k}} \frac{1}{2} \Big\| \begin{bmatrix} \mathbf{r}(\mathbf{a}) \\ \mathbf{0}^{k.k} \end{bmatrix} - \begin{bmatrix} \mathbf{M}(\mathbf{a}) + \mathbf{L}(\mathbf{a}) \\ \mathbf{N}^T \end{bmatrix} d\mathbf{a} \Big\|_2^2 \tag{6.24}$$

and

$$d\mathbf{a}_{k-gn} = \text{Arg} \min_{d\mathbf{a} \in \mathbb{R}^{p.k}} \frac{1}{2} \Big\| \begin{bmatrix} \mathbf{r}(\mathbf{a}) \\ \mathbf{0}^{k.k} \end{bmatrix} - \begin{bmatrix} \mathbf{M}(\mathbf{a}) \\ \mathbf{N}^T \end{bmatrix} d\mathbf{a} \Big\|_2^2, \tag{6.25}$$

in three steps at each iteration of the Golub-Pereyra or Kaufman variants of the Gauss-Newton algorithms (1) (as discussed in Section 5.2).

Remark 6.1. A third solution for computing the correction vectors $d\mathbf{a}_{gp-gn}$ and $d\mathbf{a}_{k-gn}$ in the Gauss-Newton algorithms (1), if we assumed again that $rank(\mathbf{A}) = k$ and $rank(J(\mathbf{r}(\mathbf{a}))) = rank(\mathbf{M}(\mathbf{a})) = (p-k).k$, is to apply the TSQR algorithm to the matrices $-J(\mathbf{r}(\mathbf{a}))\bar{\mathbf{O}}^\perp$ and $\mathbf{M}(\mathbf{a})\bar{\mathbf{O}}^\perp$ defined in Corollary 5.6 instead to $-J(\mathbf{r}(\mathbf{a})$ and $\mathbf{M}(\mathbf{a})$ as described above. In these conditions, the upper triangular matrices obtained in the output of the TSQR algorithm are also nonsingular and can, thus, be directly solved by backward substitution. Finally, the correction vectors $d\mathbf{a}_{gp-gn}$ and $d\mathbf{a}_{k-gn}$ can be computed by a simple matrix-vector product as described in Subsection 5.2. ■

Similarly, in the normal-equation approach, several faster alternative methods can be used to find the correction vectors $d\mathbf{a}_{gp-gn}$ and $d\mathbf{a}_{k-gn}$ from the symmetric positive semi-definite matrices

$$\Delta = J\big(\mathbf{r}(\mathbf{a})\big)^T J\big(\mathbf{r}(\mathbf{a})\big) = \widetilde{J}\big(\mathbf{r}(\mathbf{a})\big)^T \widetilde{J}\big(\mathbf{r}(\mathbf{a})\big)$$

and

$$\Lambda = \mathbf{M}(\mathbf{a})^T \mathbf{M}(\mathbf{a}) = \bar{\mathbf{M}}(\mathbf{a})^T \bar{\mathbf{M}}(\mathbf{a}) \ ,$$

if we assume that $\mathit{rank}(\mathbf{A}) = k$ and $\mathit{rank}(J(\mathbf{r}(\mathbf{a}))) = \mathit{rank}(\mathbf{M}(\mathbf{a})) = (p-k).k$. Remember again that the first condition is always true if step **(0)** of the Gauss-Newton algorithms (1) is performed at each iteration. Under these hypotheses, we deduce immediately that $\mathit{rank}(\Delta) = \mathit{rank}(\Lambda) = (p-k).k$ and to cope with this uniform rank deficiency of Δ and Λ, we can again use the results of Corollary 5.6, which state that the matrices $\mathbf{N} = \mathbf{K}_{(p,k)}(\mathbf{I}_k \otimes \mathbf{A})$ and $\bar{\mathbf{O}} = \mathbf{K}_{(p,k)}(\mathbf{I}_k \otimes \mathbf{O})$ (where $\mathbf{O}$ is an orthonormal basis of $\mathit{ran}(\mathbf{A})$) are, respectively, a basis and an orthonormal basis of $\mathit{null}\big(J(\mathbf{r}(\mathbf{a}))\big) = \mathit{null}(\mathbf{M}(\mathbf{a}))$ and, thus, also of the null spaces of the matrices Δ and Λ.

In these conditions, as first suggested by Okatani et al. [150], we can compute

$$\mathbf{N}\mathbf{N}^T = \mathbf{K}_{(p,k)}(\mathbf{I}_k \otimes \mathbf{A}\mathbf{A}^T)\mathbf{K}_{(k,p)}$$

or

$$\bar{\mathbf{O}}\bar{\mathbf{O}}^T = \mathbf{K}_{(p,k)}(\mathbf{I}_k \otimes \mathbf{O}\mathbf{O}^T)\mathbf{K}_{(k,p)} \ ,$$

and add these positive semi-definite matrices of rank $k.k$ to Δ and Λ. Again, note that, if step **(0)** of the Gauss-Newton algorithms (1) is performed, we have $\mathbf{N} = \bar{\mathbf{O}}$ and, thus, $\mathbf{N}\mathbf{N}^T = \bar{\mathbf{O}}\bar{\mathbf{O}}^T$. Next, we can compute

$$\Delta(\mathbf{N}) = \Delta + \mathbf{N}\mathbf{N}^T = \begin{bmatrix} J(\mathbf{r}(\mathbf{a})) \\ \mathbf{N}^T \end{bmatrix}^T \begin{bmatrix} J(\mathbf{r}(\mathbf{a})) \\ \mathbf{N}^T \end{bmatrix} \text{ or } \Lambda(\mathbf{N}) = \Lambda + \mathbf{N}\mathbf{N}^T = \begin{bmatrix} \mathbf{M}(\mathbf{a}) \\ \mathbf{N}^T \end{bmatrix}^T \begin{bmatrix} \mathbf{M}(\mathbf{a}) \\ \mathbf{N}^T \end{bmatrix} .$$

Since $\mathit{ran}(\mathbf{N}\mathbf{N}^T) = \mathit{ran}(\Delta)^\perp = \mathit{ran}(\Lambda)^\perp$ if $\mathit{rank}(J(\mathbf{r}(\mathbf{a}))) = \mathit{rank}(\mathbf{M}(\mathbf{a})) = (p-k).k$, we have the relationships

$$\mathit{dim}(\Delta(\mathbf{N})) = \mathit{dim}(\Delta) + \mathit{dim}(\mathbf{N}\mathbf{N}^T) = (p-k).k + k.k = k.p \ ,$$

and, similarly, $\mathit{dim}(\Lambda(\mathbf{N})) = k.p$. In other words, the matrices $\Delta(\mathbf{N})$ and $\Lambda(\mathbf{N})$ are positive definite and, thus, of full rank. In these conditions, the normal equations

$$\Delta(\mathbf{N})d\mathbf{a}_{gp-gn} = -\nabla\psi(\mathbf{a}) \text{ or } \Lambda(\mathbf{N})d\mathbf{a}_{k-gn} = -\nabla\psi(\mathbf{a})$$

have an unique solution, which are also the solutions of the associated linear least-square problems (6.24) and (6.25) solved in the TSQR approach in order to find $d\mathbf{a}_{gp-gn}$ and $d\mathbf{a}_{k-gn}$. Thus, when using the normal-equation approach, we finally need to compute the Cholesky factorizations of $\Delta(\mathbf{N})$ or $\Lambda(\mathbf{N})$ and solve the above positive definite systems by forward and backward substitutions using these triangular Cholesky factors as described in Okatani et al. [150].

Remark 6.2. As for the TSQR method, an alternative solution for computing the correction vectors $d\mathbf{a}_{gp-gn}$ and $d\mathbf{a}_{k-gn}$ in the Gauss-Newton algorithms (1), if we assumed that $\mathit{rank}(J(\mathbf{r}(\mathbf{a}))) = \mathit{rank}(\mathbf{M}(\mathbf{a})) = (p-k).k$, is to apply the normal-equation algorithm to the matrices $-J(\mathbf{r}(\mathbf{a}))\bar{\mathbf{O}}^\perp$ and $\mathbf{M}(\mathbf{a})\bar{\mathbf{O}}^\perp$ defined in Corollary 5.6 instead to $-J(\mathbf{r}(\mathbf{a})$ and $\mathbf{M}(\mathbf{a})$ as described above. In these conditions, the upper triangular matrices obtained in the output of the Cholesky factorization are nonsingular and can, thus, be directly solved by forward and backward substitutions. Finally, the correction vectors $d\mathbf{a}_{gp-gn}$ and $d\mathbf{a}_{k-gn}$ can also be computed by a simple matrix-vector product as described in Subsection 5.2. ■

6.2 Variable projection Levenberg-Marquardt algorithms

This subsection describes and investigates variable projection Levenberg-Marquardt methods for the solution of the WLRA problem. Using similar notations as in the Gauss-Newton algorithms (1) described in the previous subsection, an outline of a first version of the variable projection Levenberg-Marquardt algorithms is as follows:

Levenberg-Marquardt algorithms 2.

Choose starting matrix $\mathbf{A}_1 \in \mathbb{R}^{p\times k}$, $\varepsilon_1, \varepsilon_2, \varepsilon_3, \beta, \|\nabla\psi\|_{min} \in \mathbb{R}_{+*}$ and $i_{max}, j_{max} \in \mathbb{N}_*$, appropriately

For $i = 1, 2, \ldots$ **until convergence do**

(0) Optionally, compute a QRCP of $\mathbf{A}_i$ (see equation (2.15)) to determine $k_i = \mathit{rank}(\mathbf{A}_i)$ and an orthonormal basis of $\mathit{ran}(\mathbf{A}_i)$:

$$\mathbf{Q}_i\mathbf{A}_i\mathbf{P}_i = \begin{bmatrix} \mathbf{R}_i & \mathbf{S}_i \\ \mathbf{0}^{(p-k_i)\times k_i} & \mathbf{0}^{(p-k_i)\times(k-k_i)} \end{bmatrix},$$

where $\mathbf{Q}_i$ is an $p \times p$ orthogonal matrix, $\mathbf{P}_i$ is an $k \times k$ permutation matrix, $\mathbf{R}_i$ is an $k_i \times k_i$ nonsingular upper triangular matrix (with diagonal elements of decreasing absolute magnitude) and $\mathbf{S}_i$ an $k_i \times (k - k_i)$ full matrix, which is vacuous if $k_i = k$.

In all cases, compute an $p \times k$ matrix $\mathbf{O}_i$ with orthonormal columns as the first k columns of $\mathbf{Q}_i$ (i.e., such that $\mathit{ran}(\mathbf{A}_i) \subset \mathit{ran}(\mathbf{O}_i)$ if $k_i < k$ and $\mathit{ran}(\mathbf{A}_i) = \mathit{ran}(\mathbf{O}_i)$ if $k_i = k$) and set

$$\mathbf{A}_i = \mathbf{O}_i \, .$$

This optional orthogonalization step is a safe-guard as the condition $k_i = k$ is a necessary condition for the differentiability of $\psi(.)$ at a point $\mathbf{A}_i$ and is also to limit the occurrence of overflows and underflows in the next steps by enforcing that the matrix variable $\mathbf{A}_i \in \mathbb{O}^{p\times k}$.

(1) Determine (implicitly) the block diagonal matrix

$$\mathbf{F}(\mathbf{a}_i) = \mathit{diag}\big(\mathit{vec}(\sqrt{\mathbf{W}})\big)\big(\mathbf{I}_n \otimes \mathbf{A}_i\big) \, ,$$

where $\mathbf{a}_i = \mathit{vec}(\mathbf{A}_i^T)$.

(2) Compute (implicitly) a QRCP of $\mathbf{F}(\mathbf{a}_i)$ to determine $\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a}_i)}$ and $\mathbf{F}(\mathbf{a}_i)^-$ (see equations (2.18) and (2.19)) or, alternatively, a COD of $\mathbf{F}(\mathbf{a}_i)$ to determine $\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a}_i)}$ and $\mathbf{F}(\mathbf{a}_i)^+$ (see equations (2.18) and (2.21)).

Note also that $\mathbf{F}(\mathbf{a}_i)^- = \mathbf{F}(\mathbf{a}_i)^+$ if $\mathbf{F}(\mathbf{a}_i)$ is of full column rank and that $\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a}_i)}$, $\mathbf{F}(\mathbf{a}_i)^-$ and $\mathbf{F}(\mathbf{a}_i)^+$ are also block diagonal matrices.

(3) Solve the block diagonal linear least-squares problem

$$\mathbf{b}_i = \mathrm{Arg}\min_{\mathbf{b}\in\mathbb{R}^{k.n}} \|\mathbf{x} - \mathbf{F}(\mathbf{a}_i)\mathbf{b}\|_2^2 \, ,$$

e.g., compute

$$\mathbf{b}_i = \begin{cases} \mathbf{F}(\mathbf{a}_i)^-\mathbf{x} & \{\text{if a QRCP of } \mathbf{F}(\mathbf{a}_i) \text{ is used in step } \mathbf{(2)}\} \\ \mathbf{F}(\mathbf{a}_i)^+\mathbf{x} & \{\text{if a COD of } \mathbf{F}(\mathbf{a}_i) \text{ is used in step } \mathbf{(2)}\} \end{cases} .$$

(4) Determine:

$$\mathbf{r}(\mathbf{a}_i) = \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a}_i)}\mathbf{x} \ \{\text{current residual vector}\}$$

$$\psi(\mathbf{a}_i) = \tfrac{1}{2}\|\mathbf{r}(\mathbf{a}_i)\|_2^2 \ \{\text{current value of the cost function }\}$$

$$\nabla\psi(\mathbf{a}_i) = \mathbf{G}(\mathbf{b}_i)^T\mathbf{G}(\mathbf{b}_i)\mathbf{a}_i - \mathbf{G}(\mathbf{b}_i)^T\mathbf{z} \ \{\text{see Theorems 4.3 and 5.7}\}$$

$\lambda_i = \beta \|\nabla\psi(\mathbf{a}_i)\|_2^2$ {set ridge parameter proportional to the squared 2-norm of the gradient}

Note that the steps **(1)** to **(4)** above can be very easily parallelized using the block diagonal structure of $\mathbf{F}(\mathbf{a}_i)$.

(5) Check for convergence. Relevant convergence criteria in the algorithms are of the form:

- $\|\nabla\psi(\mathbf{a}_i)\|_2 \leq \varepsilon_1$
- $\|\mathbf{a}_i - \mathbf{a}_{i-1}\|_2 \leq \varepsilon_2(\varepsilon_2 + \|\mathbf{a}_i\|_2)$ {if $i \neq 1$}

  If step **(0)** is used, this last convergence condition can be simplified as:

  $\|\mathbf{a}_i - \mathbf{a}_{i-1}\|_2 \leq \varepsilon_2\|\mathbf{a}_i\|_2 = \varepsilon_2\sqrt{k}$
- $|\psi(\mathbf{a}_{i-1}) - \psi(\mathbf{a}_i)| \leq \varepsilon_3(\varepsilon_3 + \psi(\mathbf{a}_i))$ {if $i \neq 1$}
- $i \geq i_{max}$ {e.g., give up if the number of iterations is too large}

where $\varepsilon_1, \varepsilon_2, \varepsilon_3$ and i_{max} are constants chosen by the user.

Exit if convergence. **Otherwise, go to** step **(6)**

(6) Compute the Levenberg-Marquardt correction vector $d\mathbf{a}_{lm}$ as the (minimum 2-norm) solution of one of the following (regularized) linear least-squares problems:

(6.1) **If** $\|\nabla\psi(\mathbf{a}_i)\|_2 \geq \|\nabla\psi\|_{min}$ **then**

Golub-Pereyra Levenberg-Marquardt step: Golub and Pereyra [63]

$$\begin{aligned} d\mathbf{a}_{gp-lm} &= \begin{bmatrix} \mathbf{M}(\mathbf{a}_i) + \mathbf{L}(\mathbf{a}_i) \\ \sqrt{\lambda_i}\mathbf{D}_i \end{bmatrix}^+ \begin{bmatrix} \mathbf{r}(\mathbf{a}_i) \\ \mathbf{0}^{k.p} \end{bmatrix} \\ &= \text{Arg}\min_{d\mathbf{a}\in\mathbb{R}^{p.k}} \|\mathbf{r}(\mathbf{a}_i) - \big(\mathbf{M}(\mathbf{a}_i) + \mathbf{L}(\mathbf{a}_i)\big)d\mathbf{a}\|_2^2 + \lambda_i\|\mathbf{D}_i d\mathbf{a}\|_2^2 \end{aligned}$$

Kaufman Levenberg-Marquardt step: Kaufman [96]

$$\begin{aligned} d\mathbf{a}_{k-lm} &= \begin{bmatrix} \mathbf{M}(\mathbf{a}_i) \\ \sqrt{\lambda_i}\mathbf{D}_i \end{bmatrix}^+ \begin{bmatrix} \mathbf{r}(\mathbf{a}_i) \\ \mathbf{0}^{k.p} \end{bmatrix} \\ &= \text{Arg}\min_{d\mathbf{a}\in\mathbb{R}^{p.k}} \|\mathbf{r}(\mathbf{a}_i) - \mathbf{M}(\mathbf{a}_i)d\mathbf{a}\|_2^2 + \lambda_i\|\mathbf{D}_i d\mathbf{a}\|_2^2 \end{aligned}$$

(6.2) **Else**

Golub-Pereyra Gauss-Newton step: Golub and Pereyra [63], Ruhe and Wedin [166]

$$\begin{aligned} d\mathbf{a}_{gp-gn} &= \big(\mathbf{M}(\mathbf{a}_i) + \mathbf{L}(\mathbf{a}_i)\big)^+\mathbf{r}(\mathbf{a}_i) \\ &= \begin{cases} \text{Arg}\min_{d\mathbf{a}\in\mathbb{R}^{p.k}} \|d\mathbf{a}\|_2^2 \\ \text{s.t. Arg}\min_{d\mathbf{a}\in\mathbb{R}^{p.k}} \|\mathbf{r}(\mathbf{a}_i) - \big(\mathbf{M}(\mathbf{a}_i) + \mathbf{L}(\mathbf{a}_i)\big)d\mathbf{a}\|_2^2 \end{cases} \end{aligned}$$

Kaufman Gauss-Newton step: Kaufman [96], Ruhe and Wedin [166]

$$\begin{aligned} d\mathbf{a}_{k-gn} &= \mathbf{M}(\mathbf{a}_i)^+\mathbf{r}(\mathbf{a}_i) \\ &= \begin{cases} \text{Arg}\min_{d\mathbf{a}\in\mathbb{R}^{p.k}} \|d\mathbf{a}\|_2^2 \\ \text{s.t. Arg}\min_{d\mathbf{a}\in\mathbb{R}^{p.k}} \|\mathbf{r}(\mathbf{a}_i) - \mathbf{M}(\mathbf{a}_i)d\mathbf{a}\|_2^2 \end{cases} \end{aligned}$$

(7) Increment $\mathbf{a}_i = vec(\mathbf{A}_i^T)$, e.g., compute $\mathbf{a}_{i+1} = vec(\mathbf{A}_{i+1}^T)$ such that $\psi(\mathbf{a}_{i+1}) < \psi(\mathbf{a}_i)$ in order to obtain global convergence.

(7.1) To this end, first compute

$$\mathbf{a}_{i+1} = \mathbf{a}_i + d\mathbf{a}_{lm}$$
$$\psi(\mathbf{a}_{i+1}) = \tfrac{1}{2}\|\mathbf{r}(\mathbf{a}_{i+1})\|_2^2 = \tfrac{1}{2}\|\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a}_{i+1})}\mathbf{x}\|_2^2 \,,$$

using (implicitly) a QRCP of the block diagonal matrix $\mathbf{F}(\mathbf{a}_{i+1})$.

(7.2) **If** $\psi(\mathbf{a}_{i+1}) > \psi(\mathbf{a}_i)$ **then** recompute $\mathbf{a}_{i+1}$ by one of the following methods:

Gauss-Seidel: $\mathbf{a}_{i+1} = \mathbf{a}_i + d\mathbf{a}_{gs-gn}$ where $d\mathbf{a}_{gs-gn}$ is a Gauss-Seidel step [166]

$$\begin{aligned} d\mathbf{a}_{gs-gn} &= \big(\mathbf{K}_{(n,p)}\mathbf{G}(\mathbf{b}_i)\big)^{+}\mathbf{r}(\mathbf{a}_i) \\ &= \begin{cases} \text{Arg}\min_{d\mathbf{a}\in\mathbb{R}^{p.k}} \|d\mathbf{a}\|_2^2 \\ \text{s.t. Arg}\min_{d\mathbf{a}\in\mathbb{R}^{p.k}} \|\mathbf{r}(\mathbf{a}_i) - \mathbf{K}_{(n,p)}\mathbf{G}(\mathbf{b}_i)d\mathbf{a}\|_2^2 \end{cases} \end{aligned}$$

Block alternating least-squares:

$$\begin{aligned} \mathbf{a}_{i+1} &= \mathbf{G}(\mathbf{b}_i)^{+}\mathbf{z} \\ &= \begin{cases} \text{Arg}\min_{\mathbf{a}\in\mathbb{R}^{p.k}} \|\mathbf{a}\|_2^2 \\ \text{s.t. Arg}\min_{\mathbf{a}\in\mathbb{R}^{p.k}} \|\mathbf{z} - \mathbf{G}(\mathbf{b}_i)\mathbf{a}\|_2^2 \end{cases} \end{aligned}$$

Line search:

$$\mathbf{a}_{i+1} = \mathbf{a}_i + \alpha_i d\mathbf{a}_{lm}$$

where $\alpha_i < 1$ is determined by a line search to make the algorithm a descent method (i.e, such that $\psi(\mathbf{a}_{i+1}) < \psi(\mathbf{a}_i)$). This is always possible as the correction vector $d\mathbf{a}_{lm}$ is in a descent direction for $\psi(.)$ if $\|\nabla\psi(\mathbf{a}_i)\|_2 \neq 0$, see Corollaries 5.7 and 5.8.

As an illustration, a simple, but still efficient, strategy is to first shorten the correction step to half the Levenberg-Marquardt length (or Gauss-Newton length if $\|\nabla\psi(\mathbf{a}_i)\|_2 < \|\nabla\psi\|_{min}$), compute the new trial value for $\psi(\mathbf{a}_{i+1})$ and, if it is still worse, continue to reduce the step until we get a step short enough such that $\psi(\mathbf{a}_{i+1}) < \psi(\mathbf{a}_i)$. The following loop incorporates this simple step-shortening algorithm:

For $j = 1, 2, \ldots$ **while**$\big(\psi(\mathbf{a}_{i+1}) > \psi(\mathbf{a}_i)\big)$

$d\mathbf{a}_{lm} = \frac{1}{2}d\mathbf{a}_{lm}$

$\mathbf{a}_{i+1} = \mathbf{a}_i + d\mathbf{a}_{lm}$

$\psi(\mathbf{a}_{i+1}) = \frac{1}{2}\|\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a}_{i+1})}\mathbf{x}\|_2^2$ {using a QRCP of the matrix $\mathbf{F}(\mathbf{a}_{i+1})$}

If $j > j_{max}$ **exit** {give up if the number of iterations is too large}

End do

End do

In this version of the Levenberg-Marquardt algorithms (2), the shape and definition of the different vector and matrix variables are exactly the same as in the Gauss-Newton algorithms (1) described in the previous subsection. Furthermore, if, during the iterations, $\|\nabla\psi(\mathbf{a}_i)\|_2 < \|\nabla\psi\|_{min}$ where

$\|\nabla\psi\|_{min}$ is a positive real constant greater than ε_1 chosen by the user, we use a Gauss-Newton correction step as also described in the previous subsection. In addition, as in the Gauss-Newton algorithm, the computations in the above Levenberg-Marquardt algorithms (2) are terminated either when one or several of the convergence criteria listed in step **(5)** are satisfied, or when the iteration count exceeds the predetermined number i_{max}. Obviously, this version (2) of the Levenberg-Marquardt algorithms is, thus, similar to the Gauss-Newton algorithms (1), except in step **(6.1)**, when $\|\nabla\psi(\mathbf{a}_i)\|_2 \geq \|\nabla\psi\|_{min}$.

This approach is first justified by the fact that, when $\varepsilon_1 < \|\nabla\psi(\mathbf{a}_i)\|_2 < \|\nabla\psi\|_{min}$, we are near a stationary or local solution point of our minimization problem, in which case, we want to benefit from the faster convergence of the Gauss-Newton method, see Subsection 5.1 for more details. On the other hand, if $\|\nabla\psi(\mathbf{a}_i)\|_2 \geq \|\nabla\psi\|_{min}$, we consider that we are far away from a stationary or solution point in which case the Gauss-Newton method may be much less satisfactorily and we prefer to use a more robust correction step, which will be more in the steepest descent direction, in order to widen the basin of convergence of the method. With these considerations in mind, when $\|\nabla\psi(\mathbf{a}_i)\|_2 \geq \|\nabla\psi\|_{min}$, we introduce both a strictly positive damping parameter λ_i (e.g. the Marquardt parameter), which takes into account how far we are from a solution and, optionally, a strictly positive scaling diagonal matrix $\mathbf{D}_i \in \mathbb{R}_+^{k.p\times k.p}$, which may be useful to render the algorithm invariant under diagonal scaling of the solution vector $\hat{\mathbf{a}}$ and even more robust when λ_i becomes very large as discussed also in Subsection 5.1.

The choice of λ_i influences both the direction and the size of the correction vector $d\mathbf{a}_{gp-lm}$ and $d\mathbf{a}_{k-lm}$. If λ_i tends to zero, $d\mathbf{a}_{gp-lm}$ and $d\mathbf{a}_{k-lm}$ will tend, respectively, to the corresponding Gauss-Newton steps $d\mathbf{a}_{gp-gn}$ and $d\mathbf{a}_{k-gn}$. On the other hand, if λ_i tends to infinity, then $d\mathbf{a}_{gp-lm}$ and $d\mathbf{a}_{k-lm}$ will tend to a short step in the steepest descent direction, e.g., $-\frac{1}{\lambda_i}\nabla\psi(\mathbf{a}_i)$, see Subsection 5.1 for more information. Thus, the choice of the Marquardt parameter λ_i is based on the following considerations: if we are close to a local solution then we want the faster convergence of the Gauss-Newton method while it is safe to choose the steepest descent method when we are far from the solution. In other words, the selection procedure

$$\lambda_i = \beta\|\nabla\psi(\mathbf{a}_i)\|_2^2 \,,$$

used in step **(6.1)** of our Levenberg-Marquardt algorithms (2), is first motivated by the fact that the method of steepest descent has global convergence not held by the Gauss-Newton method. When one is far away from the solution (i.e. $\|\nabla\psi(\mathbf{a}_i)\|_2$ is large), λ_i is chosen to be large in order to weight the descent part of the correction. As the iterates proceed toward the solution (i.e. $\|\nabla\psi(\mathbf{a}_i)\|_2$ is small), λ_i is decreased to weight the Gauss-Newton part of the correction. When we are far from the solution we are interested in the stability of the steepest descent method; when we are close, we strive for the rapidity of convergence of the Gauss-Newton method.

However, taking into account the systematic rank deficiency of the Jacobian matrix $J(\mathbf{r}(\mathbf{a}_i))$ or its Kaufman approximation $-\mathbf{M}(\mathbf{a}_i)$ demonstrated in the previous sections, we cannot let λ_i tends to zero freely and we need to control it appropriately in order to avoid numerical instability when computing the correction steps $d\mathbf{a}_{gp-lm}$ or $d\mathbf{a}_{k-lm}$ if λ_i approaches zero. Thus, in an actual computer implementation, the condition test $\lambda_i = 0$ must be replaced by the condition $\lambda_i \leq \lambda_{min}$ (with $\lambda_{min} \in \mathbb{R}_{+*}$) to switch to the Gauss-Newton method, where λ_{min} is a suitably chosen real constant such that the matrices

$$\begin{bmatrix} \mathbf{M}(\mathbf{a}_i) + \mathbf{L}(\mathbf{a}_i) \\ \sqrt{\lambda_i}\mathbf{D}_i \end{bmatrix} \text{ and } \begin{bmatrix} \mathbf{M}(\mathbf{a}_i) \\ \sqrt{\lambda_i}\mathbf{D}_i \end{bmatrix}$$

do not become nearly singular or ill-conditioned when λ_i approaches zero as it is expected after some iterations of the Levenberg-Marquardt algorithms (2). Equivalently, in our version (2) of the Levenberg-Marquardt algorithms, such numerical test is also performed in step **(6.1)**, on $\|\nabla\psi(\mathbf{a})\|_2$ using the user defined threshold $\|\nabla\psi\|_{min} \in \mathbb{R}_{+*}$ rather on λ_i. This is justified by the fact that $\lambda_i = \beta\|\nabla\psi(\mathbf{a}_i)\|_2$, where β is a strictly positive real constant also chosen by the user. Obviously, the choice of $\|\nabla\psi\|_{min}$ (or alternatively λ_{min}) can be tricky as it depends on the scaling of the problem.

Moreover, it must be done with care to avoid numerical instabilities when computing $d\mathbf{a}_{gp-lm}$ or $d\mathbf{a}_{k-lm}$, and, at the same time, maintain the global convergence properties of the Levenberg-Marquardt algorithms (2) .

In addition, we have introduced a strictly positive diagonal matrix $\mathbf{D}_i \in \mathbb{R}_+^{k.p \times k.p}$ in step **(6.1)** of the Levenberg-Marquardt algorithms. A common simple choice for this scaling diagonal matrix is to set $\mathbf{D}_i = \mathbf{I}_{k.p}$, the identity matrix of order $k.p$. This choice together with a suitable strategy to update λ_i across the iterations gives the original Levenberg algorithm [107]. Note that $\mathbf{D}_i$ can also vary during the iterations and permits for example to introduce some scaling in order to take into account the relative sizes of the columns of the Jacobian matrix or its Kaufman approximation. Thus, as first suggested by Marquardt [120], we can also set

$$\big[\mathbf{D}_i\big]_{jj} = \begin{cases} \|\big[\mathbf{M}(\mathbf{a}_i) + \mathbf{L}(\mathbf{a}_i)\big]_{.j}\|_2 & \{\text{if } d\mathbf{a}_{lm} = d\mathbf{a}_{gp-lm} \text{ in step } \mathbf{(6.1)}\} \\ \|\big[\mathbf{M}(\mathbf{a}_i)\big]_{.j}\|_2 & \{\text{if } d\mathbf{a}_{lm} = d\mathbf{a}_{k-lm} \text{ in step } \mathbf{(6.1)}\} \end{cases},$$

for $j = 1, \cdots, k.p$, see Subsection 5.1 for further details. Note, however, that this last choice for the scaling matrix $\mathbf{D}_i$ implies that the conditions stated in Theorem 5.6 are not verified as otherwise some of the elements of the diagonal of $\mathbf{D}_i$ will be equal to zero during the whole iterative process.

The Golub-Pereyra step $d\mathbf{a}_{gp-lm}$ corresponds exactly to the standard Levenberg-Marquardt step $d\mathbf{a}_{lm}$ applied to the minimization of the variable projection functional $\psi(.)$, which is introduced in Subsection 5.1. The philosophy behind the Kaufman step $d\mathbf{a}_{k-lm}$ is exactly similar to the one detailed for the Gauss-Newton algorithms (2): in most cases, approximating the Jacobian matrix by $-\mathbf{M}(\mathbf{a})$ can perform even better than to use the exact Jacobian matrix, taking into account the particular form of the Hessian matrix $\nabla^2\psi(\mathbf{a})$ derived in Subsection 5.3.

The Golub-Pereyra and Kaufman variants in the Gauss-Newton algorithms (1) generate a sequence $\{\mathbf{a}_i\}$ by setting $\mathbf{a}_{i+1} = \mathbf{a}_i + \alpha_i d\mathbf{a}_i$, where $d\mathbf{a}_i$ is the minimum 2-norm solution of one of the linearized subproblems

$$d\mathbf{a}_i = \text{Arg} \min_{d\mathbf{a} \in \mathbb{R}^{p.k}} \frac{1}{2}\|l(d\mathbf{a})\|_2^2 ,$$

with

$$l(d\mathbf{a}) = \begin{cases} \mathbf{r}(\mathbf{a}_i) - \big(\mathbf{M}(\mathbf{a}_i) + \mathbf{L}(\mathbf{a}_i)\big)d\mathbf{a} & \{\text{if } d\mathbf{a}_{gp-gn} \text{ is used in step } \mathbf{(6)}\} \\ \mathbf{r}(\mathbf{a}_i) - \mathbf{M}(\mathbf{a}_i)d\mathbf{a} & \{\text{if } d\mathbf{a}_{k-gn} \text{ is used in step } \mathbf{(6)}\} \end{cases}, \tag{6.26}$$

as explained in the previous subsection. However, we know from the results of the previous sections that a solution of the WLRA problem, if it exists, is never unique and isolated. Furthermore, since we also know that the above linear least-squares subproblems are always rank-deficient, Levenberg-Marquardt methods, which replace them by regularized linearized subproblems of the form

$$d\mathbf{a}_i = \text{Arg} \min_{d\mathbf{a} \in \mathbb{R}^{p.k}} \frac{1}{2}\|l(d\mathbf{a})\|_2^2 + \frac{\lambda_i}{2}\|\mathbf{D}_i d\mathbf{a}\|_2^2 ,$$

where λ_i is a strictly positive parameter and $\mathbf{D}_i$ is a (positive) diagonal matrix, are an interesting alternative. Equivalently, this means that the correction step $d\mathbf{a}_i$ in the Levenberg-Marquardt approach minimizes one of the following quadratic models

$$L_{\lambda_i}(d\mathbf{a}) = \begin{cases} \psi(\mathbf{a}_i) + d\mathbf{a}^T\nabla\psi(\mathbf{a}_i) + \frac{1}{2}d\mathbf{a}^T\big(\mathbf{M}(\mathbf{a}_i)^T\mathbf{M}(\mathbf{a}_i) + \mathbf{L}(\mathbf{a}_i)^T\mathbf{L}(\mathbf{a}_i) + \lambda_i\mathbf{D}_i^2\big)d\mathbf{a} \\ \psi(\mathbf{a}_i) + d\mathbf{a}^T\nabla\psi(\mathbf{a}_i) + \frac{1}{2}d\mathbf{a}^T\big(\mathbf{M}(\mathbf{a}_i)^T\mathbf{M}(\mathbf{a}_i) + \lambda_i\mathbf{D}_i^2\big)d\mathbf{a} \end{cases},$$

since $\nabla\psi(\mathbf{a}_i) = -\big(\mathbf{M}(\mathbf{a}_i) + \mathbf{L}(\mathbf{a}_i)\big)^T\mathbf{r}(\mathbf{a}_i) = -\mathbf{M}(\mathbf{a}_i)^T\mathbf{r}(\mathbf{a}_i)$. On the other hand, the correction step $d\mathbf{a}_i$ in the Gauss-Newton methods is based on the simpler quadratic models

$$G(d\mathbf{a}) = L_0(d\mathbf{a}) = \begin{cases} \psi(\mathbf{a}_i) + d\mathbf{a}^T\nabla\psi(\mathbf{a}_i) + \frac{1}{2}d\mathbf{a}^T\big(\mathbf{M}(\mathbf{a}_i)^T\mathbf{M}(\mathbf{a}_i) + \mathbf{L}(\mathbf{a}_i)^T\mathbf{L}(\mathbf{a}_i)\big)d\mathbf{a} \\ \psi(\mathbf{a}_i) + d\mathbf{a}^T\nabla\psi(\mathbf{a}_i) + \frac{1}{2}d\mathbf{a}^T\mathbf{M}(\mathbf{a}_i)^T\mathbf{M}(\mathbf{a}_i)d\mathbf{a} \end{cases}.$$

The gradients of the quadratic functions $L_{\lambda_i}(.)$ are, respectively,

$$\nabla L_{\lambda_i}(d\mathbf{a}) = \begin{cases} \nabla\psi(\mathbf{a}_i) + \big(\mathbf{M}(\mathbf{a}_i)^T\mathbf{M}(\mathbf{a}_i) + \mathbf{L}(\mathbf{a}_i)^T\mathbf{L}(\mathbf{a}_i) + \lambda_i\mathbf{D}_i^2\big)d\mathbf{a} \\ \nabla\psi(\mathbf{a}_i) + \big(\mathbf{M}(\mathbf{a}_i)^T\mathbf{M}(\mathbf{a}_i) + \lambda_i\mathbf{D}_i^2\big)d\mathbf{a} \end{cases} ,$$

and, by setting these gradients equal to zero, we get $d\mathbf{a}_{gp-lm}$ and $d\mathbf{a}_{k-lm}$ as the solutions to the linear systems

$$\big(\mathbf{M}(\mathbf{a}_i)^T\mathbf{M}(\mathbf{a}_i) + \mathbf{L}(\mathbf{a}_i)^T\mathbf{L}(\mathbf{a}_i) + \lambda_i\mathbf{D}_i^2\big)d\mathbf{a}_{gp-lm} = -\nabla\psi(\mathbf{a}_i) = \mathbf{M}(\mathbf{a}_i)^T\mathbf{r}(\mathbf{a}_i)$$

and

$$\big(\mathbf{M}(\mathbf{a}_i)^T\mathbf{M}(\mathbf{a}_i) + \lambda_i\mathbf{D}_i^2\big)d\mathbf{a}_{k-lm} = -\nabla\psi(\mathbf{a}_i) = \mathbf{M}(\mathbf{a}_i)^T\mathbf{r}(\mathbf{a}_i) ,$$

which are, respectively, the normal equations for the damped linear least-squares problems

$$\min_{d\mathbf{a}\in\mathbb{R}^{p.k}} \left\| \begin{bmatrix} \mathbf{r}(\mathbf{a}_i) \\ \mathbf{0}^{p.k} \end{bmatrix} - \begin{bmatrix} \mathbf{M}(\mathbf{a}_i) + \mathbf{L}(\mathbf{a}_i) \\ \sqrt{\lambda_i}\mathbf{D}_i \end{bmatrix} d\mathbf{a} \right\|_2^2$$

and

$$\min_{d\mathbf{a}\in\mathbb{R}^{p.k}} \left\| \begin{bmatrix} \mathbf{r}(\mathbf{a}_i) \\ \mathbf{0}^{p.k} \end{bmatrix} - \begin{bmatrix} \mathbf{M}(\mathbf{a}_i) \\ \sqrt{\lambda_i}\mathbf{D}_i \end{bmatrix} d\mathbf{a} \right\|_2^2 .$$

Furthermore, as the coefficient matrices of the above normal equations are always positive definite if $\lambda_i > 0$, these linear systems have always an unique solution, which are the global minimizers of the associated linear least-squares problems or quadratic model functions, and these quadratic functions are also strictly convex. These nice properties are important numerically and are also an another advantage of the Levenberg-Marquardt methods over a simple Gauss-Newton approach. These results also show that the correction vectors $d\mathbf{a}_{gp-lm}$ and $d\mathbf{a}_{k-lm}$ can be computed, alternatively, by a normal-equation or a more stable QR method as for the Gauss-Newton correction vectors $d\mathbf{a}_{gp-gn}$ and $d\mathbf{a}_{k-gn}$ and we will discuss this matter in more details below after we derive a second and third versions of the Levenberg-Marquardt algorithms.

One disadvantage with the simple strategy used in the Levenberg-Marquardt algorithms (2) for updating the Marquardt parameter λ_i is, however, that strict descent (i.e., $\psi(\mathbf{a}_{i+1}) < \psi(\mathbf{a}_i)$) is not guaranteed if a line search or alternative strategies for computing $\mathbf{a}_{i+1}$ are not incorporated in step **(7.2)** of the algorithms. However, as for the Gauss-Newton algorithms (1), in order to implement a line search algorithm, we have to perform the second part of step **(4)** of the algorithm, every time we want to get $\psi(\mathbf{a}_{i+1})$ for a new trial value of α_i since

$$\psi(\mathbf{a}_{i+1}) = \frac{1}{2}\|\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a}_{i+1})}\mathbf{x}\|_2^2 ,$$

and this line search can involve many extra evaluations of $\psi(.)$, which do not get us closer to an acceptable solution. In these conditions, it is again tempting to perform one or several iterations with the fast Gauss-Seidel or block ALS methods to compute $\mathbf{a}_{i+1}$ in step **(7.2)** in case we have $\psi(\mathbf{a}_i + d\mathbf{a}_{lm}) > \psi(\mathbf{a}_i)$ instead of using a more costly line search algorithm. In other words, if a full Levenberg-Marquardt step gives a sufficient decrease of $\psi(.)$, we accept the point $\mathbf{a}_i + d\mathbf{a}_{lm}$ as the new iterate. Otherwise we switch to the fast Gauss-Seidel or block ALS methods.

Alternatively, it is well-known that a line search can be completely avoided in Levenberg-Marquardt methods by using a more sophisticated strategy for updating λ during the iterations since the choice of the Marquardt parameter influences both the direction and the size of the correction vector $d\mathbf{a}_{lm}$ [139][123]. Furthermore, it is always possible to find a λ such that $\psi(\mathbf{a}_i + d\mathbf{a}_{lm}) < \psi(\mathbf{a}_i)$ if $\|\nabla\psi(\mathbf{a}_i)\|_2 \neq 0$ [139][123]. Thus, by a proper adjustment of the damping parameter λ we have also a direct method for ensuring the descent condition $\psi(\mathbf{a}_i + d\mathbf{a}_{lm}) < \psi(\mathbf{a}_i)$.

As first suggested by Marquardt [120], one such strategy is to start with λ sets at a small value, 10^{-8} for example. Whenever a step is unsuccessful, λ gets multiplied by 10 to force smaller steps until

$\psi(\mathbf{a}_i + d\mathbf{a}_{lm}) < \psi(\mathbf{a}_i)$. On the other hand, when the steps become successful λ is divided by 10. This simple strategy results in a fully adaptive technique that behaves just like Gauss-Newton when Gauss-Newton is successful, but shifts in the steepest descent direction and shortens steps when the steps are not successful.

More sophisticated strategies for updating the Marquardt parameter λ during the iterations are based on the so-called gain factor

$$\rho = \frac{\psi(\mathbf{a}_i) - \psi(\mathbf{a}_i + d\mathbf{a})}{G(\mathbf{0}^{k.p}) - G(d\mathbf{a})} = \frac{\psi(\mathbf{a}_i) - \psi(\mathbf{a}_i + d\mathbf{a})}{L_0(\mathbf{0}^{k.p}) - L_0(d\mathbf{a})} ,$$

where

$$G(d\mathbf{a}) = L_0(d\mathbf{a}) = \frac{1}{2} l(d\mathbf{a})^T l(d\mathbf{a}) ,$$

and $l(.)$ is defined in equation (6.26), see [141][122][139][123] for a discussion of this strategy in a general NLLS context. $G(d\mathbf{a})$ is assumed to be a good approximation to $\psi(\mathbf{a}_i + d\mathbf{a})$ when $d\mathbf{a}$ is sufficiently small since $l(d\mathbf{a})$ is based on first order Taylor's expansions for the residual function $\mathbf{r}(\mathbf{a})$ around the current iterate $\mathbf{a}_i$. Note further that $G(d\mathbf{a}) = L_0(d\mathbf{a})$ is the quadratic model, which is used to approximate $\psi(\mathbf{a})$ in the neighborhood of the current iterate $\mathbf{a}_i$ and is minimized at each iteration of the variable projection Gauss-Newton algorithms (1) as discussed above. Now, we have for the Kaufman variant of the variable projection Levenberg-Marquardt algorithm the equalities

$$\begin{aligned}
G(\mathbf{0}^{k.p}) - G(d\mathbf{a}_{k-lm}) &= \frac{1}{2} l(\mathbf{0}^{k.p})^T l(\mathbf{0}^{k.p}) - \frac{1}{2} l(d\mathbf{a}_{k-lm})^T l(d\mathbf{a}_{k-lm}) \\
&= \psi(\mathbf{a}_i) - \frac{1}{2} \big(\mathbf{r}(\mathbf{a}_i) - \mathbf{M}(\mathbf{a}_i) d\mathbf{a}_{k-lm}\big)^T \big(\mathbf{r}(\mathbf{a}_i) - \mathbf{M}(\mathbf{a}_i) d\mathbf{a}_{k-lm}\big) \\
&= -d\mathbf{a}_{k-lm}^T \nabla\psi(\mathbf{a}_i) - \frac{1}{2} d\mathbf{a}_{k-lm}^T \mathbf{M}(\mathbf{a}_i)^T \mathbf{M}(\mathbf{a}_i) d\mathbf{a}_{k-lm} \\
&= -\frac{1}{2} d\mathbf{a}_{k-lm}^T \big(2\nabla\psi(\mathbf{a}_i) + \mathbf{M}(\mathbf{a}_i)^T \mathbf{M}(\mathbf{a}_i) d\mathbf{a}_{k-lm}\big) \\
&= -\frac{1}{2} d\mathbf{a}_{k-lm}^T \Big(2\nabla\psi(\mathbf{a}_i) + \big(\mathbf{M}(\mathbf{a}_i)^T \mathbf{M}(\mathbf{a}_i) + \lambda_i \mathbf{D}_i^2 - \lambda_i \mathbf{D}_i^2\big) d\mathbf{a}_{k-lm}\Big) \\
&= -\frac{1}{2} d\mathbf{a}_{k-lm}^T \Big(2\nabla\psi(\mathbf{a}_i) - \nabla\psi(\mathbf{a}_i) - \lambda_i \mathbf{D}_i^2 d\mathbf{a}_{k-lm}\Big) \\
&= \frac{1}{2} d\mathbf{a}_{k-lm}^T \Big(\lambda_i \mathbf{D}_i^2 d\mathbf{a}_{k-lm} - \nabla\psi(\mathbf{a}_i)\Big) \\
&= \frac{1}{2} \Big(\lambda_i \|\mathbf{D}_i d\mathbf{a}_{k-lm}\|_2^2 - d\mathbf{a}_{k-lm}^T \nabla\psi(\mathbf{a}_i)\Big) ,
\end{aligned}$$

where we have used the equalities

$$\nabla\psi(\mathbf{a}_i) = -\mathbf{M}(\mathbf{a}_i)^T \mathbf{r}(\mathbf{a}_i) \text{ and } \big(\mathbf{M}(\mathbf{a}_i)^T \mathbf{M}(\mathbf{a}_i) + \lambda_i \mathbf{D}_i^2\big) d\mathbf{a}_{k-lm} = -\nabla\psi(\mathbf{a}_i) ,$$

derived in Theorem 5.7 and Corollary 5.8. Moreover, the same equality holds for the Golub-Pereyra variant since

$$\mathit{ran}(\mathbf{M}(\mathbf{a}_i)) \subset \mathit{ran}(\mathbf{F}(\mathbf{a}_i))^{\perp} \text{ and } \mathit{ran}(\mathbf{L}(\mathbf{a}_i)) \subset \mathit{ran}(\mathbf{F}(\mathbf{a}_i)) .$$

Thus, in both cases, $G(\mathbf{0}^{k.p}) - G(d\mathbf{a}_{lm})$ and the gain factor ρ can be easily computed at each iteration of the Levenberg-Marquardt algorithms as the terms $\lambda_i \|\mathbf{D}_i d\mathbf{a}_{lm}\|_2^2$ and $\nabla\psi(\mathbf{a}_i)$ are already available before the gain factor must be evaluated.

Furthermore, $G(\mathbf{0}^{k.p}) - G(d\mathbf{a}_{lm})$ is always guaranteed to be positive as both $\|\mathbf{D}_i d\mathbf{a}_{lm}\|_2^2$ and $-d\mathbf{a}_{lm}^T \nabla\psi(\mathbf{a}_i)$ are positive if $\|\nabla\psi(\mathbf{a}_i)\|_2 \neq 0$ since $d\mathbf{a}_{lm}$ is in a descent direction for $\psi(.)$, as demonstrated in Corollary 5.7. In these conditions, it follows that the condition $\rho > 0$ is equivalent to the descending condition $\psi(\mathbf{a}_i + d\mathbf{a}_{lm}) < \psi(\mathbf{a}_i)$ if $\|\nabla\psi(\mathbf{a}_i)\|_2 \neq 0$ for both the Golub-Pereyra and Kaufman variants of the Levenberg-Marquardt algorithm. Using these results, the following

clever strategy proposed by Madsen and Nielsen [123] may be used to update λ at each iteration of Levenberg-Marquardt methods:

For $i = 1, 2, \ldots$ **until convergence do**

(0) $\cdots$

$\vdots$

(6) Compute the Levenberg-Marquardt correction vector $d\mathbf{a}_{lm}$ as the solution of one of the following constrained and damped linear least-squares problems:

Golub-Pereyra Levenberg-Marquardt step:

$$d\mathbf{a}_{gp-lm} = \text{Arg} \min_{d\mathbf{a} \in \mathbb{R}^{p.k}} \|\mathbf{r}(\mathbf{a}_i) - \big(\mathbf{M}(\mathbf{a}_i) + \mathbf{L}(\mathbf{a}_i)\big) d\mathbf{a}\|_2^2 + \lambda \|\mathbf{D}_i d\mathbf{a}\|_2^2 \,,$$

Kaufman Levenberg-Marquardt step:

$$d\mathbf{a}_{k-lm} = \text{Arg} \min_{d\mathbf{a} \in \mathbb{R}^{p.k}} \|\mathbf{r}(\mathbf{a}_i) - \mathbf{M}(\mathbf{a}_i) d\mathbf{a}\|_2^2 + \lambda \|\mathbf{D}_i d\mathbf{a}\|_2^2 \,,$$

(7) Increment $\mathbf{a}_i = vec(\mathbf{A}_i^T)$, e.g., compute $\mathbf{a}_{i+1} = vec(\mathbf{A}_{i+1}^T)$ such that $\psi(\mathbf{a}_{i+1}) < \psi(\mathbf{a}_i)$ in order to obtain global convergence. To this end, first compute the gain factor

$$\rho = \frac{\psi(\mathbf{a}_i) - \psi(\mathbf{a}_i + d\mathbf{a}_{lm})}{G(\mathbf{0}^{k.p}) - G(d\mathbf{a}_{lm})} = \frac{\psi(\mathbf{a}_i) - \psi(\mathbf{a}_i + d\mathbf{a}_{lm})}{\frac{1}{2}\big(\lambda \|\mathbf{D}_i d\mathbf{a}_{lm}\|_2^2 - d\mathbf{a}_{lm}^T \nabla \psi(\mathbf{a}_i)\big)}$$

If $\rho > 0$ **then**

$\mathbf{a}_{i+1} = \mathbf{a}_i + d\mathbf{a}_{lm}$

$\lambda = \lambda.max\big(\frac{1}{3}, 1 - (2.\rho - 1)^3\big)$

$\nu = 2$

Else

$\lambda = \nu.\lambda$

$\nu = 2.\nu$

Go to step **(6)**

End do

In this algorithm, the factor ν is initialized to 2 and the Marquardt parameter λ is again initialized to a small value like 10^{-8} (see below for more details). With this updating strategy, if $\rho \leq 0$ then $\mathbf{A}_i$ is kept fixed, but we increase λ quickly with the twofold purpose of getting closer to the steepest descent direction and reduce the step length. On the other hand, if $\rho > 0$ we accept the new point $\mathbf{a}_{i+1} = \mathbf{a}_i + d\mathbf{a}_{lm}$. However, if ρ is small, the quadratic model $G(d\mathbf{a})$ is considered not to be a good approximation to the function $\psi(\mathbf{a}_i + d\mathbf{a})$ in the neighborhood of $\mathbf{a}_i$ and λ is increased. On the other hand, if ρ is large the quadratic model $G(d\mathbf{a})$ is considered to be a good approximation to the function $\psi(\mathbf{a}_i + d\mathbf{a})$ and λ is decreased in order to get closer to the Gauss-Newton direction in the next iteration step. Thus, this more sophisticated strategy results also in a fully adaptive technique that behaves just like Gauss-Newton when it is successful, but shifts smoothly in the steepest descent direction and shortens steps when the steps are not successful [123].

However, in our specific WLRA context, both updating schemes of the Marquardt parameter must be adapted to take care of the systematic rank deficiency of the coefficient matrices $\mathbf{M}(\mathbf{a}_i) + \mathbf{L}(\mathbf{a}_i)$ or $\mathbf{M}(\mathbf{a}_i)$ across the iterations and this without using a specific threshold like λ_{min} or $\|\nabla\psi\|_{min}$ to shift to a Gauss-Newton step when λ approaches zero as this will break the incremental nature of the updating schemes if we set $\lambda = 0$ when a full Gauss-Newton step is used at one particular

iteration. Furthermore, as before, we must also avoid the near singularity and ill-conditioning of the regularized coefficient matrices

$$\begin{bmatrix} \mathbf{M}(\mathbf{a}_i) + \mathbf{L}(\mathbf{a}_i) \\ \sqrt{\lambda}\mathbf{D}_i \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \mathbf{M}(\mathbf{a}_i) \\ \sqrt{\lambda}\mathbf{D}_i \end{bmatrix} ,$$

when λ approaches zero.

A clever way to avoid the use of a threshold and to deal efficiently with the uniform rank deficiency of the Jacobian matrix $J(\mathbf{r}(\mathbf{a}_i))$ or its Kaufman approximation $-\mathbf{M}(\mathbf{a}_i)$, if we want to incorporate these updating schemes in the variable projection Levenberg-Marquardt algorithms, is to solve at step **(6)** the constrained and regularized problems

$$d\mathbf{a}_{gp-lm} = \text{Arg} \min_{d\mathbf{a}\in\mathbb{R}^{p.k}} \|\mathbf{r}(\mathbf{a}_i) - \big(\mathbf{M}(\mathbf{a}_i) + \mathbf{L}(\mathbf{a}_i)\big)d\mathbf{a}\|_2^2 + \|\mathbf{N}_i^T d\mathbf{a}\|_2^2 + \lambda_i\|\mathbf{D}_i d\mathbf{a}\|_2^2$$

or

$$d\mathbf{a}_{k-lm} = \text{Arg} \min_{d\mathbf{a}\in\mathbb{R}^{p.k}} \|\mathbf{r}(\mathbf{a}_i) - \mathbf{M}(\mathbf{a}_i)d\mathbf{a}\|_2^2 + \|\mathbf{N}_i^T d\mathbf{a}\|_2^2 + \lambda_i\|\mathbf{D}_i d\mathbf{a}\|_2^2 ,$$

where the columns of $\mathbf{N}_i = \mathbf{K}_{(p,k)}(\mathbf{I}_k \otimes \mathbf{A}_i)$ are a (orthonormal) basis of $\mathit{null}\big(\mathbf{M}(\mathbf{a}_i) + \mathbf{L}(\mathbf{a}_i)\big) = \mathit{null}(\mathbf{M}(\mathbf{a}_i))$ if $\mathit{rank}(\mathbf{A}_i) = k$ and $\mathit{rank}(J(\mathbf{r}(\mathbf{a}_i))) = \mathit{rank}(\mathbf{M}(\mathbf{a}_i)) = k.(p-k)$ (see Corollary 5.6 for details). The associated quadratic models are

$$L_{\lambda_i}^{\mathbf{N}_i}(d\mathbf{a}) = \psi(\mathbf{a}_i) + d\mathbf{a}^T\nabla\psi(\mathbf{a}_i) + \frac{1}{2}d\mathbf{a}^T\big(\mathbf{M}(\mathbf{a}_i)^T\mathbf{M}(\mathbf{a}_i) + \mathbf{L}(\mathbf{a}_i)^T\mathbf{L}(\mathbf{a}_i) + \mathbf{N}_i\mathbf{N}_i^T + \lambda_i\mathbf{D}_i^2\big)d\mathbf{a}$$

or

$$L_{\lambda_i}^{\mathbf{N}_i}(d\mathbf{a}) = \psi(\mathbf{a}_i) + d\mathbf{a}^T\nabla\psi(\mathbf{a}_i) + \frac{1}{2}d\mathbf{a}^T\big(\mathbf{M}(\mathbf{a}_i)^T\mathbf{M}(\mathbf{a}_i) + \mathbf{N}_i\mathbf{N}_i^T + \lambda_i\mathbf{D}_i^2\big)d\mathbf{a} ,$$

which can be minimized by solving the symmetric linear systems

$$\big(\mathbf{M}(\mathbf{a}_i)^T\mathbf{M}(\mathbf{a}_i) + \mathbf{L}(\mathbf{a}_i)^T\mathbf{L}(\mathbf{a}_i) + \mathbf{N}_i\mathbf{N}_i^T + \lambda_i\mathbf{D}_i^2\big)d\mathbf{a}_{gp-lm} = \mathbf{M}(\mathbf{a}_i)^T\mathbf{r}(\mathbf{a}_i)$$

or

$$\big(\mathbf{M}(\mathbf{a}_i)^T\mathbf{M}(\mathbf{a}_i) + \mathbf{N}_i\mathbf{N}_i^T + \lambda_i\mathbf{D}_i^2\big)d\mathbf{a}_{k-lm} = \mathbf{M}(\mathbf{a}_i)^T\mathbf{r}(\mathbf{a}_i) ,$$

which, in turn, are the normal equations for the constrained and damped linear systems

$$\min_{d\mathbf{a}\in\mathbb{R}^{p.k}} \left\| \begin{bmatrix} \mathbf{r}(\mathbf{a}_i) \\ \mathbf{0}^{k.k} \\ \mathbf{0}^{p.k} \end{bmatrix} - \begin{bmatrix} \mathbf{M}(\mathbf{a}_i) + \mathbf{L}(\mathbf{a}_i) \\ \mathbf{N}_i^T \\ \sqrt{\lambda_i}\mathbf{D}_i \end{bmatrix} d\mathbf{a} \right\|_2^2$$

and

$$\min_{d\mathbf{a}\in\mathbb{R}^{p.k}} \left\| \begin{bmatrix} \mathbf{r}(\mathbf{a}_i) \\ \mathbf{0}^{k.k} \\ \mathbf{0}^{p.k} \end{bmatrix} - \begin{bmatrix} \mathbf{M}(\mathbf{a}_i) \\ \mathbf{N}_i^T \\ \sqrt{\lambda_i}\mathbf{D}_i \end{bmatrix} d\mathbf{a} \right\|_2^2 .$$

This approach was first suggested by Okatani et al. [150]. These two linear least-squares problems always have an unique solution independently of the value of λ if we assume that

$$\mathit{rank}(\mathbf{A}_i) = k \quad \text{and} \quad \mathit{rank}(J(\mathbf{r}(\mathbf{a}_i))) = \mathit{rank}(\mathbf{M}(\mathbf{a}_i)) = k.(p-k) ,$$

during the iterations or, at least, that these conditions are verified as soon as $\lambda = 0$. If $\lambda = 0$, we again obtained $d\mathbf{a}_{gp-lm} = d\mathbf{a}_{gp-gn}$ and $d\mathbf{a}_{k-lm} = d\mathbf{a}_{k-gn}$ as in version (2) of the Levenberg-Marquardt algorithms if the above assumptions are verified. However, the key-difference with this version (2) is that these two linear least-squares problems remain nonsingular and well-conditioned when λ approaches zero if the above assumptions are verified thanks to the inclusion of the block $\mathbf{N}_i^T$ in the coefficient matrix of these linear least-squares problems or to the addition of the term $\mathbf{N}_i\mathbf{N}_i^T$ in the associated normal equations if we use a normal-equation approach to solve them.

Finally, when $\lambda \gg 0$, and we want to shift $d\mathbf{a}_{gp-lm}$ or $d\mathbf{a}_{k-lm}$ in the steepest descent direction in order to benefit of the good global convergence ability of the gradient descent method, the inclusion of the block $\mathbf{N}_i^T$ or the term $\mathbf{N}_i\mathbf{N}_i^T$ is of secondary importance and will not impair the performance of the algorithm as they just try to constrain the columns of the perturbation matrices $d\mathbf{A}_{gp-lm}$ or $d\mathbf{A}_{k-lm}$ to belong to $ran(\mathbf{A}_i)^\perp$, see the discussion after Corollary 5.6 for details.

Finally, note that the computations of the gain factor ρ at step **(7)** must also be slightly modified as follows

$$\rho = \frac{\psi(\mathbf{a}_i) - \psi(\mathbf{a}_i + d\mathbf{a}_{lm})}{G(\mathbf{0}^{k.p}) - G(d\mathbf{a}_{lm})} = \frac{\psi(\mathbf{a}_i) - \psi(\mathbf{a}_i + d\mathbf{a}_{lm})}{\frac{1}{2}\big(\|\mathbf{N}_i^T d\mathbf{a}_{lm}\|_2^2 + \lambda\|\mathbf{D}_i d\mathbf{a}_{lm}\|_2^2 - d\mathbf{a}_{lm}^T \nabla\psi(\mathbf{a}_i)\big)} ,$$

when the above constrained and regularized linear least-squares problems are solved in step **(6)**. This is justified by the fact that we cannot assume that $\|\mathbf{N}_i^T d\mathbf{a}_{lm}\|_2 = 0$ if a damping term $\lambda\|\mathbf{D}_i d\mathbf{a}_{lm}\|_2^2$ with $\lambda > 0$ is also used in the algorithms.

These different considerations lead to our second and third versions of the Levenberg-Marquardt algorithms, which use, respectively, the simple and more sophisticated updating strategies of λ discussed above:

Levenberg-Marquardt algorithms 3.

Choose starting matrix $\mathbf{A}_1 \in \mathbb{R}^{p\times k}$, $\varepsilon_1, \varepsilon_2, \varepsilon_3, \lambda \in \mathbb{R}_{+*}$ and $i_{max}, j_{max} \in \mathbb{N}_*$, appropriately

For $i = 1, 2, \ldots$ **until convergence do**

(0) Optionally, compute a QRCP of $\mathbf{A}_i$ (see equation (2.15)) to determine $k_i = rank(\mathbf{A}_i)$ and an orthonormal basis of $ran(\mathbf{A}_i)$:

$$\mathbf{Q}_i\mathbf{A}_i\mathbf{P}_i = \begin{bmatrix} \mathbf{R}_i & \mathbf{S}_i \\ \mathbf{0}^{(p-k_i)\times k_i} & \mathbf{0}^{(p-k_i)\times(k-k_i)} \end{bmatrix} ,$$

where $\mathbf{Q}_i$ is an $p \times p$ orthogonal matrix, $\mathbf{P}_i$ is an $k \times k$ permutation matrix, $\mathbf{R}_i$ is an $k_i \times k_i$ nonsingular upper triangular matrix (with diagonal elements of decreasing absolute magnitude) and $\mathbf{S}_i$ an $k_i \times (k - k_i)$ full matrix, which is vacuous if $k_i = k$.

In all cases, compute an $p \times k$ matrix $\mathbf{O}_i$ with orthonormal columns as the first k columns of $\mathbf{Q}_i$ (i.e., such that $ran(\mathbf{A}_i) \subset ran(\mathbf{O}_i)$ if $k_i < k$ and $ran(\mathbf{A}_i) = ran(\mathbf{O}_i)$ if $k_i = k$) and set

$$\mathbf{A}_i = \mathbf{O}_i .$$

This optional orthogonalization step is a safe-guard as the condition $k_i = k$ is a necessary condition for the differentiability of $\psi(.)$ at a point $\mathbf{A}_i$ and also to limit the occurence of overflows and underflows in the next steps by enforcing that the matrix variable $\mathbf{A}_i \in \mathbb{O}^{p\times k}$.

(1) Determine (implicitly) the block diagonal matrix

$$\mathbf{F}(\mathbf{a}_i) = diag\big(vec(\sqrt{\mathbf{W}})\big)\big(\mathbf{I}_n \otimes \mathbf{A}_i\big) ,$$

where $\mathbf{a}_i = vec(\mathbf{A}_i^T)$.

(2) Compute (implicitly) a QRCP of $\mathbf{F}(\mathbf{a}_i)$ to determine $\mathbf{P}^\perp_{\mathbf{F}(\mathbf{a}_i)}$ and $\mathbf{F}(\mathbf{a}_i)^-$ (see equations (2.18) and (2.19)) or, alternatively, a COD of $\mathbf{F}(\mathbf{a}_i)$ to determine $\mathbf{P}^\perp_{\mathbf{F}(\mathbf{a}_i)}$ and $\mathbf{F}(\mathbf{a}_i)^+$ (see equations (2.18) and (2.21)).

Note also that $\mathbf{F}(\mathbf{a}_i)^- = \mathbf{F}(\mathbf{a}_i)^+$ if $\mathbf{F}(\mathbf{a}_i)$ is of full column rank and that $\mathbf{P}^\perp_{\mathbf{F}(\mathbf{a}_i)}$, $\mathbf{F}(\mathbf{a}_i)^-$ and $\mathbf{F}(\mathbf{a}_i)^+$ are also block diagonal matrices.

(3) Solve the block diagonal linear least-squares problem

$$\mathbf{b}_i = \mathrm{Arg\,min}_{\mathbf{b}\in\mathbb{R}^{k.n}} \ \|\mathbf{x} - \mathbf{F}(\mathbf{a}_i)\mathbf{b}\|_2^2 ,$$

e.g., compute

$$\mathbf{b}_i = \begin{cases} \mathbf{F}(\mathbf{a}_i)^{-}\mathbf{x} & \{\text{if a QRCP of } \mathbf{F}(\mathbf{a}_i) \text{ is used in step } \mathbf{(2)}\} \\ \mathbf{F}(\mathbf{a}_i)^{+}\mathbf{x} & \{\text{if a COD of } \mathbf{F}(\mathbf{a}_i) \text{ is used in step } \mathbf{(2)}\} \end{cases} .$$

(4) Determine and set:

$\mathbf{r}(\mathbf{a}_i) = \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a}_i)}\mathbf{x}$ {current residual vector}

$\psi(\mathbf{a}_i) = \frac{1}{2}\|\mathbf{r}(\mathbf{a}_i)\|_2^2$ {current value of the cost function }

$\nabla\psi(\mathbf{a}_i) = \mathbf{G}(\mathbf{b}_i)^T\mathbf{G}(\mathbf{b}_i)\mathbf{a}_i - \mathbf{G}(\mathbf{b}_i)^T\mathbf{z}$ {see Theorems 4.3 and 5.7}

$j = 0$ {initialize counter for the ridge scaling subiterations}

Note that the steps **(1)** to **(4)** above can be very easily parallelized using the block diagonal structure of $\mathbf{F}(\mathbf{a}_i)$.

(5) Check for convergence. Relevant convergence criteria in the algorithms are of the form:

- $\|\nabla\psi(\mathbf{a}_i)\|_2 \leq \varepsilon_1$
- $\|\mathbf{a}_i - \mathbf{a}_{i-1}\|_2 \leq \varepsilon_2(\varepsilon_2 + \|\mathbf{a}_i\|_2)$ {if $i \neq 1$}

  If step **(0)** is used, this convergence condition can be simplified as:

  $\|\mathbf{a}_i - \mathbf{a}_{i-1}\|_2 \leq \varepsilon_2\|\mathbf{a}_i\|_2 = \varepsilon_2\sqrt{k}$
- $|\psi(\mathbf{a}_{i-1}) - \psi(\mathbf{a}_i)| \leq \varepsilon_3(\varepsilon_3 + \psi(\mathbf{a}_i))$ {if $i \neq 1$}
- $i \geq i_{max}$ {e.g., give up if the number of iterations is too large}

where $\varepsilon_1, \varepsilon_2, \varepsilon_3$ and i_{max} are constants chosen by the user.

Exit if convergence. **Otherwise, go to** step **(6)**

(6) Compute the Levenberg-Marquardt correction vector $d\mathbf{a}_{lm}$ as the solution of one of the following constrained and damped linear least-squares problems:

Golub-Pereyra Levenberg-Marquardt step:

$$\begin{aligned} d\mathbf{a}_{gp-lm} &= \begin{bmatrix} \mathbf{M}(\mathbf{a}_i) + \mathbf{L}(\mathbf{a}_i) \\ \mathbf{N}_i^T \\ \sqrt{\lambda}\mathbf{D}_i \end{bmatrix}^{+} \begin{bmatrix} \mathbf{r}(\mathbf{a}_i) \\ \mathbf{0}^{k.k} \\ \mathbf{0}^{k.p} \end{bmatrix} \\ &= \text{Arg}\min_{d\mathbf{a}\in\mathbb{R}^{p.k}} \|\mathbf{r}(\mathbf{a}_i) - \big(\mathbf{M}(\mathbf{a}_i) + \mathbf{L}(\mathbf{a}_i)\big)d\mathbf{a}\|_2^2 + \|\mathbf{N}_i^T d\mathbf{a}\|_2^2 + \lambda\|\mathbf{D}_i d\mathbf{a}\|_2^2 \end{aligned}$$

Kaufman Levenberg-Marquardt step: Okatani et al. [150]

$$\begin{aligned} d\mathbf{a}_{k-lm} &= \begin{bmatrix} \mathbf{M}(\mathbf{a}_i) \\ \mathbf{N}_i^T \\ \sqrt{\lambda}\mathbf{D}_i \end{bmatrix}^{+} \begin{bmatrix} \mathbf{r}(\mathbf{a}_i) \\ \mathbf{0}^{k.k} \\ \mathbf{0}^{k.p} \end{bmatrix} \\ &= \text{Arg}\min_{d\mathbf{a}\in\mathbb{R}^{p.k}} \|\mathbf{r}(\mathbf{a}_i) - \mathbf{M}(\mathbf{a}_i)d\mathbf{a}\|_2^2 + \|\mathbf{N}_i^T d\mathbf{a}\|_2^2 + \lambda\|\mathbf{D}_i d\mathbf{a}\|_2^2 \end{aligned}$$

where the columns of $\mathbf{N}_i = \mathbf{K}_{(p,k)}(\mathbf{I}_k \otimes \mathbf{A}_i)$ are a (orthonormal) basis of $null\big(\mathbf{M}(\mathbf{a}_i) + \mathbf{L}(\mathbf{a}_i)\big) = null(\mathbf{M}(\mathbf{a}_i))$, see Corollary 5.6.

(7) Compute $\mathbf{a}_{i+1} = vec(\mathbf{A}_{i+1}^T)$ such that $\psi(\mathbf{a}_{i+1}) < \psi(\mathbf{a}_i)$ in order to obtain global convergence.

(7.1) To this end, first compute

$$\psi(\mathbf{a}_i + d\mathbf{a}_{lm}) = \tfrac{1}{2}\|\mathbf{r}(\mathbf{a}_i + d\mathbf{a}_{lm})\|_2^2 = \tfrac{1}{2}\|\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a}_i + d\mathbf{a}_{lm})}\mathbf{x}\|_2^2 \,,$$

using (implicitly) a QRCP of the block diagonal matrix $\mathbf{F}(\mathbf{a}_i + d\mathbf{a}_{lm})$.

(7.2) **If** $\psi(\mathbf{a}_i + d\mathbf{a}_{lm}) > \psi(\mathbf{a}_i)$ **then** {step rejected}

$j = j + 1$

$\lambda = 10.\lambda$ {scale up the ridge parameter}

If $j \leq j_{max}$ **go to** step **(6)** {recompute $d\mathbf{a}_{lm}$ with inflated diagonal}

(7.3) **Else** {step acceptable}

If $j = 0$ **then** $\lambda = \lambda/10$ {scale down the ridge parameter if step is successful}

(7.4) Increment $\mathbf{a}_i$:

$\mathbf{a}_{i+1} = \mathbf{a}_i + d\mathbf{a}_{lm}$ {compute new iterate}

End do

Levenberg-Marquardt algorithms 4.

Choose starting matrix $\mathbf{A}_1 \in \mathbb{R}^{p\times k}$, $\varepsilon_1, \varepsilon_2, \varepsilon_3, \lambda \in \mathbb{R}_{+*}$ and $i_{max}, j_{max} \in \mathbb{N}_*$, appropriately, and initialize $\nu = 2$

For $i = 1, 2, \ldots$ **until convergence do**

(0) Optionally, compute a QRCP of $\mathbf{A}_i$ (see equation (2.15)) to determine $k_i = \mathit{rank}(\mathbf{A}_i)$ and an orthonormal basis of $\mathit{ran}(\mathbf{A}_i)$:

$$\mathbf{Q}_i\mathbf{A}_i\mathbf{P}_i = \begin{bmatrix} \mathbf{R}_i & \mathbf{S}_i \\ \mathbf{0}^{(p-k_i)\times k_i} & \mathbf{0}^{(p-k_i)\times(k-k_i)} \end{bmatrix} ,$$

where $\mathbf{Q}_i$ is an $p \times p$ orthogonal matrix, $\mathbf{P}_i$ is an $k \times k$ permutation matrix, $\mathbf{R}_i$ is an $k_i \times k_i$ nonsingular upper triangular matrix (with diagonal elements of decreasing absolute magnitude) and $\mathbf{S}_i$ an $k_i \times (k - k_i)$ full matrix, which is vacuous if $k_i = k$.

In all cases, compute an $p \times k$ matrix $\mathbf{O}_i$ with orthonormal columns as the first k columns of $\mathbf{Q}_i$ (i.e., such that $\mathit{ran}(\mathbf{A}_i) \subset \mathit{ran}(\mathbf{O}_i)$ if $k_i < k$ and $\mathit{ran}(\mathbf{A}_i) = \mathit{ran}(\mathbf{O}_i)$ if $k_i = k$) and set

$\mathbf{A}_i = \mathbf{O}_i$.

This optional orthogonalization step is a safe-guard as the condition $k_i = k$ is a necessary condition for the differentiability of $\psi(.)$ at a point $\mathbf{A}_i$ and also to limit the occurence of overflows and underflows in the next steps by enforcing that the matrix variable $\mathbf{A}_i \in \mathbb{O}^{p\times k}$.

(1) Determine (implicitly) the block diagonal matrix

$$\mathbf{F}(\mathbf{a}_i) = \mathit{diag}\big(\mathit{vec}(\sqrt{\mathbf{W}})\big)\big(\mathbf{I}_n \otimes \mathbf{A}_i\big) \,,$$

where $\mathbf{a}_i = \mathit{vec}(\mathbf{A}_i^T)$.

(2) Compute (implicitly) a QRCP of $\mathbf{F}(\mathbf{a}_i)$ to determine $\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a}_i)}$ and $\mathbf{F}(\mathbf{a}_i)^-$ (see equations (2.18) and (2.19)) or, alternatively, a COD of $\mathbf{F}(\mathbf{a}_i)$ to determine $\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a}_i)}$ and $\mathbf{F}(\mathbf{a}_i)^+$ (see equations (2.18) and (2.21)).

Note also that $\mathbf{F}(\mathbf{a}_i)^- = \mathbf{F}(\mathbf{a}_i)^+$ if $\mathbf{F}(\mathbf{a}_i)$ is of full column rank and that $\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a}_i)}$, $\mathbf{F}(\mathbf{a}_i)^-$ and $\mathbf{F}(\mathbf{a}_i)^+$ are also block diagonal matrices.

(3) Solve the block diagonal linear least-squares problem

$$\mathbf{b}_i = \text{Arg}\min_{\mathbf{b}\in\mathbb{R}^{k.n}} \|\mathbf{x} - \mathbf{F}(\mathbf{a}_i)\mathbf{b}\|_2^2,$$

e.g., compute

$$\mathbf{b}_i = \begin{cases} \mathbf{F}(\mathbf{a}_i)^{-}\mathbf{x} & \{\text{if a QRCP of } \mathbf{F}(\mathbf{a}_i) \text{ is used in step } \textbf{(2)}\} \\ \mathbf{F}(\mathbf{a}_i)^{+}\mathbf{x} & \{\text{if a COD of } \mathbf{F}(\mathbf{a}_i) \text{ is used in step } \textbf{(2)}\} \end{cases}.$$

(4) Determine and set:

$\mathbf{r}(\mathbf{a}_i) = \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a}_i)}\mathbf{x}$ {current residual vector}

$\psi(\mathbf{a}_i) = \frac{1}{2}\|\mathbf{r}(\mathbf{a}_i)\|_2^2$ {current value of the cost function }

$\nabla\psi(\mathbf{a}_i) = \mathbf{G}(\mathbf{b}_i)^T\mathbf{G}(\mathbf{b}_i)\mathbf{a}_i - \mathbf{G}(\mathbf{b}_i)^T\mathbf{z}$ {see Theorems 4.3 and 5.7}

$j = 0$ {initialize counter for the ridge scaling subiterations}

Note that the steps **(1)** to **(4)** above can be very easily parallelized using the block diagonal structure of $\mathbf{F}(\mathbf{a}_i)$.

(5) Check for convergence. Relevant convergence criteria in the algorithms are of the form:

- $\|\nabla\psi(\mathbf{a}_i)\|_2 \le \varepsilon_1$
- $\|\mathbf{a}_i - \mathbf{a}_{i-1}\|_2 \le \varepsilon_2(\varepsilon_2 + \|\mathbf{a}_i\|_2)$ {if $i \ne 1$}

  If step **(0)** is used, this convergence condition can be simplified as:

  $\|\mathbf{a}_i - \mathbf{a}_{i-1}\|_2 \le \varepsilon_2\|\mathbf{a}_i\|_2 = \varepsilon_2\sqrt{k}$
- $|\psi(\mathbf{a}_{i-1}) - \psi(\mathbf{a}_i)| \le \varepsilon_3(\varepsilon_3 + \psi(\mathbf{a}_i))$ {if $i \ne 1$}
- $i \ge i_{max}$ {e.g., give up if the number of iterations is too large}

where $\varepsilon_1, \varepsilon_2, \varepsilon_3$ and i_{max} are constants chosen by the user.

Exit if convergence. **Otherwise, go to** step **(6)**

(6) Compute the Levenberg-Marquardt correction vector $d\mathbf{a}_{lm}$ as the solution of one of the following constrained and damped linear least-squares problems:

Golub-Pereyra Levenberg-Marquardt step:

$$\begin{aligned} d\mathbf{a}_{gp-lm} &= \begin{bmatrix} \mathbf{M}(\mathbf{a}_i) + \mathbf{L}(\mathbf{a}_i) \\ \mathbf{N}_i^T \\ \sqrt{\lambda}\mathbf{D}_i \end{bmatrix}^{+} \begin{bmatrix} \mathbf{r}(\mathbf{a}_i) \\ \mathbf{0}^{k.k} \\ \mathbf{0}^{k.p} \end{bmatrix} \\ &= \text{Arg}\min_{d\mathbf{a}\in\mathbb{R}^{p.k}} \|\mathbf{r}(\mathbf{a}_i) - \big(\mathbf{M}(\mathbf{a}_i) + \mathbf{L}(\mathbf{a}_i)\big)d\mathbf{a}\|_2^2 + \|\mathbf{N}_i^T d\mathbf{a}\|_2^2 + \lambda\|\mathbf{D}_i d\mathbf{a}\|_2^2 \end{aligned}$$

Kaufman Levenberg-Marquardt step: Okatani et al. [150]

$$\begin{aligned} d\mathbf{a}_{k-lm} &= \begin{bmatrix} \mathbf{M}(\mathbf{a}_i) \\ \mathbf{N}_i^T \\ \sqrt{\lambda}\mathbf{D}_i \end{bmatrix}^{+} \begin{bmatrix} \mathbf{r}(\mathbf{a}_i) \\ \mathbf{0}^{k.k} \\ \mathbf{0}^{k.p} \end{bmatrix} \\ &= \text{Arg}\min_{d\mathbf{a}\in\mathbb{R}^{p.k}} \|\mathbf{r}(\mathbf{a}_i) - \mathbf{M}(\mathbf{a}_i)d\mathbf{a}\|_2^2 + \|\mathbf{N}_i^T d\mathbf{a}\|_2^2 + \lambda\|\mathbf{D}_i d\mathbf{a}\|_2^2 \end{aligned}$$

where the columns of $\mathbf{N}_i = \mathbf{K}_{(p,k)}(\mathbf{I}_k \otimes \mathbf{A}_i)$ are a (orthonormal) basis of $\mathit{null}\big(\mathbf{M}(\mathbf{a}_i) + \mathbf{L}(\mathbf{a}_i)\big) = \mathit{null}(\mathbf{M}(\mathbf{a}_i))$, see Corollary 5.6.

(7) Compute $\mathbf{a}_{i+1} = \mathit{vec}(\mathbf{A}_{i+1}^T)$ such that $\psi(\mathbf{a}_{i+1}) < \psi(\mathbf{a}_i)$ in order to obtain global convergence.

(7.1) To this end, first compute

$$\psi(\mathbf{a}_i + d\mathbf{a}_{lm}) = \tfrac{1}{2}\|\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a}_i+d\mathbf{a}_{lm})}\mathbf{x}\|_2^2 \,,$$

using (implicitly) a QRCP of the block diagonal matrix $\mathbf{F}(\mathbf{a}_i + d\mathbf{a}_{lm})$, and the gain factor

$$\rho = \frac{\psi(\mathbf{a}_i)-\psi(\mathbf{a}_i+d\mathbf{a}_{lm})}{G(\mathbf{0}^{k.p})-G(d\mathbf{a}_{lm})} = \frac{\psi(\mathbf{a}_i)-\psi(\mathbf{a}_i+d\mathbf{a}_{lm})}{\frac{1}{2}\left(\|\mathbf{N}_i^T d\mathbf{a}_{lm}\|_2^2+\lambda\|\mathbf{D}_i d\mathbf{a}_{lm}\|_2^2-d\mathbf{a}_{lm}^T\nabla\psi(\mathbf{a}_i)\right)}$$

(7.2) **If** $\rho > 0$ **then** $\{$step acceptable$\}$

$\lambda = \lambda.max\big(\frac{1}{3}, 1-(2.\rho-1)^3\big)$ $\{$scale down the ridge parameter$\}$

$\nu = 2$ $\{$reinitialize the growth factor of the ridge parameter$\}$

(7.3) **Else** $\{$step rejected$\}$

$j = j + 1$

$\lambda = \nu.\lambda$ $\{$scale up the ridge parameter$\}$

$\nu = 2.\nu$ $\{$increase the growth factor of the ridge parameter$\}$

If $j \leq j_{max}$ **go to** step **(6)** $\{$recompute $d\mathbf{a}_{lm}$ with inflated diagonal$\}$

(7.4) Increment $\mathbf{a}_i$:

$\mathbf{a}_{i+1} = \mathbf{a}_i + d\mathbf{a}_{lm}\{$compute new iterate$\}$

End do

In the Levenberg-Marquardt algorithms (3) and (4), the Marquardt parameter λ is initialized to $\lambda = \tau$ if

$$\big[\mathbf{D}_i\big]_{jj} = \begin{cases} \|\big[\mathbf{M}(\mathbf{a}_i)+\mathbf{L}(\mathbf{a}_i)\big]_{.j}\|_2 & \{\text{when } d\mathbf{a}_{lm} = d\mathbf{a}_{gp-lm} \text{ in step } \mathbf{(6)}\} \\ \|\big[\mathbf{M}(\mathbf{a}_i)\big]_{.j}\|_2 & \{\text{when } d\mathbf{a}_{lm} = d\mathbf{a}_{k-lm} \text{ in step } \mathbf{(6)}\}, \end{cases}$$

for $j = 1, \cdots, k.p$ during the iterations, or to

$$\lambda = \tau.\begin{cases} max_{j=1,\cdots,k.p}\|\big[(\mathbf{M}(\mathbf{a}_1)+\mathbf{L}(\mathbf{a}_1)\big]_{.j}\|_2^2) & \{\text{when } d\mathbf{a}_{lm} = d\mathbf{a}_{gp-lm} \text{ in step } \mathbf{(6)}\} \\ max_{j=1,\cdots,k.p}\|\big[(\mathbf{M}(\mathbf{a}_1)\big]_{.j}\|_2^2) & \{\text{when } d\mathbf{a}_{lm} = d\mathbf{a}_{k-lm} \text{ in step } \mathbf{(6)}\}, \end{cases}$$

if $\mathbf{D}_i$ is set to the identity matrix during the iterations. In both cases τ is taken in the interval $[10^{-8}\, 1]$ and a small value of τ is selected if we believe that $\mathbf{A}_1$ is close to a solution (say $\tau = 10^{-6}$). Otherwise, we can use $\tau = 10^{-3}$ or even 1. The algorithms are not very sensitive to this initial choice of τ as λ is quickly updated during the iterations in both Levenberg-Marquardt algorithms (3) and (4). Version (4) of the Levenberg-Marquardt algorithms also uses a growth factor ν for the ridge parameter, which is initialized to 2 at the start of the algorithm and reinitialized to this initial value in step **(7.2)** when a step is successful.

Remark 6.3. An alternative for computing the correction vectors $d\mathbf{a}_{gp-lm}$ and $d\mathbf{a}_{k-lm}$ in step **(6)** of the Levenberg-Marquardt algorithms (3) and (4), if we assume again that $rank(\mathbf{A}_i) = k$ and $rank(J(\mathbf{r}(\mathbf{a}_i))) = rank(\mathbf{M}(\mathbf{a}_i)) = (p-k).k$, is to first find the unique solutions of the following "reduced" and damped linear least-squares problems

$$d\bar{\mathbf{a}}_{gp-lm} = \text{Arg}\min_{d\bar{\mathbf{a}}\in\mathbb{R}^{(p-k).k}} \|\mathbf{r}(\mathbf{a}_i) - \big(\mathbf{M}(\mathbf{a}_i)+\mathbf{L}(\mathbf{a}_i)\big)\bar{\mathbf{O}}_i^{\perp} d\bar{\mathbf{a}}\|_2^2 + \lambda\|\bar{\mathbf{D}}_i d\bar{\mathbf{a}}\|_2^2 \,,$$

or

$$d\bar{\mathbf{a}}_{k-lm} = \text{Arg}\min_{d\bar{\mathbf{a}}\in\mathbb{R}^{(p-k).k}} \|\mathbf{r}(\mathbf{a}_i) - \mathbf{M}(\mathbf{a}_i)\bar{\mathbf{O}}_i^{\perp} d\bar{\mathbf{a}}\|_2^2 + \lambda\|\bar{\mathbf{D}}_i d\bar{\mathbf{a}}\|_2^2 \,,$$

where $\bar{\mathbf{O}}_i^{\perp}$ is an orthonormal basis of $null\big(J(\mathbf{r}(\mathbf{a}_i))\big)^{\perp} = null(\mathbf{M}(\mathbf{a}_i))^{\perp}$ and $\bar{\mathbf{D}}_i$ is now a positive diagonal matrix of order $(p-k).k$, see Corollary 5.6 for more information. Finally, in an additional step just before incrementing $\mathbf{a}_i$ in step **(7.4)**, the correction vectors $d\mathbf{a}_{gp-lm}$ and $d\mathbf{a}_{k-lm}$ can be computed by the matrix-vector products

$$d\mathbf{a}_{gp-lm} = \bar{\mathbf{O}}_i^{\perp} d\bar{\mathbf{a}}_{gp-lm} \quad \text{and} \quad d\mathbf{a}_{k-lm} = \bar{\mathbf{O}}_i^{\perp} d\bar{\mathbf{a}}_{k-lm} \; ,$$

or, equivalently, the matrix-matrix products

$$d\mathbf{A}_{gp-lm} = \mathbf{O}_i^{\perp} d\bar{\mathbf{A}}_{gp-lm} \quad \text{and} \quad d\mathbf{A}_{k-lm} = \mathbf{O}_i^{\perp} d\bar{\mathbf{A}}_{k-lm} \; ,$$

where $\mathbf{O}_i^{\perp}$ is an orthonormal basis of $ran(\mathbf{A}_i)^{\perp}$, as also described in Subsection 5.2. ■

We now consider in more details how to compute the correction vectors $d\mathbf{a}_{gp-lm}$ and $d\mathbf{a}_{k-lm}$ in the variable projection Levenberg-Marquardt algorithms (2), (3) and (4) using the normal-equation or TSQR approaches.

Using the normal-equation framework, the first step to obtain the correction vectors $d\mathbf{a}_{gp-lm}$ or $d\mathbf{a}_{k-lm}$ in all the Levenberg-Marquardt algorithms is to form the cross-product positive semi-definite matrices (e.g., the Gauss-Newton approximations of the Hessian matrix)

$$\Delta = J\big(\mathbf{r}(\mathbf{a})\big)^T J\big(\mathbf{r}(\mathbf{a})\big) = \mathbf{M}(\mathbf{a})^T \mathbf{M}(\mathbf{a}) + \mathbf{L}(\mathbf{a})^T \mathbf{L}(\mathbf{a}) \quad \text{or} \quad \Lambda = \mathbf{M}(\mathbf{a})^T \mathbf{M}(\mathbf{a}) \; ,$$

exactly as in the Gauss-Newton algorithms (1) described in the previous subsection. Note that, in these last equations and the rest of this subsection, we have drop the iteration index i of the Levenberg-Marquardt algorithms for notational convenience and the notations are exactly similar as in the Gauss-Newton methods described in the previous subsection. Furthermore, this first and costly step can be easily parallelized as for the Gauss-Newton algorithms (1).

For the Levenberg-Marquardt algorithms (2), in a second stage, we just need to regularize (or damp) these positive semi-definite matrices by adding the diagonal matrix $\lambda \mathbf{D}^2$ to these Gauss-Newton approximations of the Hessian matrix:

$$\Delta(\lambda) = \Delta + \lambda \mathbf{D}^2 \quad \text{or} \quad \Lambda(\lambda) = \Lambda + \lambda \mathbf{D}^2 \; ,$$

where $\lambda > 0$ is a ridge parameter, which will control both the magnitude and the direction of the correction vectors $d\mathbf{a}_{gp-lm}$ and $d\mathbf{a}_{k-lm}$, and $\mathbf{D}$ is a strictly positive diagonal matrix. Finally, in the third and last step of the Levenberg-Marquardt algorithms (2), we have to solve the consistent linear systems

$$\Delta(\lambda) d\mathbf{a}_{gp-lm} = -J\big(\mathbf{r}(\mathbf{a})\big)^T \mathbf{r}(\mathbf{a}) = \mathbf{M}(\mathbf{a})^T \mathbf{r}(\mathbf{a}) \tag{6.27}$$

or

$$\Lambda(\lambda) d\mathbf{a}_{k-lm} = -J\big(\mathbf{r}(\mathbf{a})\big)^T \mathbf{r}(\mathbf{a}) = \mathbf{M}(\mathbf{a})^T \mathbf{r}(\mathbf{a}) \; , \tag{6.28}$$

in order to obtain the correction vectors $d\mathbf{a}_{gp-lm}$ and $d\mathbf{a}_{k-lm}$. As $\lambda > 0$, $\Delta(\lambda)$ and $\Lambda(\lambda)$ are positive definite matrices and we can simply compute their Cholesky factorizations as

$$\Delta(\lambda) = \mathbf{R}_{\Delta}(\lambda)^T \mathbf{R}_{\Delta}(\lambda) \quad \text{or} \quad \Lambda(\lambda) = \mathbf{R}_{\Lambda}(\lambda)^T \mathbf{R}_{\Lambda}(\lambda) \; ,$$

where $\mathbf{R}_{\Delta}(\lambda)$ and $\mathbf{R}_{\Lambda}(\lambda)$ are $k.p \times k.p$ nonsingular upper triangular matrices. Then, in a final step, $d\mathbf{a}_{gp-lm}$ and $d\mathbf{a}_{k-lm}$ can be obtained by forward and backward substitutions in the usual manner, using these Cholesky factors, as

$$d\mathbf{a}_{gp-lm} = \mathbf{R}_{\Delta}(\lambda)^{-1} \mathbf{R}_{\Delta}(\lambda)^{-T} \mathbf{M}(\mathbf{a})^T \mathbf{r}(\mathbf{a}) \quad \text{and} \quad d\mathbf{a}_{k-lm} = \mathbf{R}_{\Lambda}(\lambda)^{-1} \mathbf{R}_{\Lambda}(\lambda)^{-T} \mathbf{M}(\mathbf{a})^T \mathbf{r}(\mathbf{a}) \; .$$

On the other hand, if we use a normal-equation approach in both the Levenberg-Marquardt algorithms (3) and (4), we have first to compute the constrained and damped cross-product (approximated) Jacobian matrices

$$\Delta(\mathbf{N}, \lambda) = \Delta + \mathbf{N}\mathbf{N}^T + \lambda \mathbf{D}^2 \quad \text{or} \quad \Lambda(\mathbf{N}, \lambda) = \Lambda + \mathbf{N}\mathbf{N}^T + \lambda \mathbf{D}^2 \; ,$$

perform their Cholesky decompositions as

$$\Delta(\mathbf{N},\lambda) = \mathbf{R}_\Delta(\mathbf{N},\lambda)^T\mathbf{R}_\Delta(\mathbf{N},\lambda) \quad \text{or} \quad \Lambda(\mathbf{N},\lambda) = \mathbf{R}_\Lambda(\mathbf{N},\lambda)^T\mathbf{R}_\Lambda(\mathbf{N},\lambda)\ ,$$

where $\mathbf{R}_\Delta(\mathbf{N},\lambda)$ and $\mathbf{R}_\Lambda(\mathbf{N},\lambda)$ are $k.p \times k.p$ upper triangular matrices. and, finally, solve the consistent linear systems

$$\Delta(\mathbf{N},\lambda)d\mathbf{a}_{gp-lm} = \mathbf{M}(\mathbf{a})^T\mathbf{r}(\mathbf{a}) \tag{6.29}$$

and

$$\Lambda(\mathbf{N},\lambda)d\mathbf{a}_{k-lm} = \mathbf{M}(\mathbf{a})^T\mathbf{r}(\mathbf{a})\ , \tag{6.30}$$

using these Cholesky factorizations. These normal equations always have an unique solution independently of the value of λ if we assume that $\mathit{rank}(\mathbf{A}) = k$ and $\mathit{rank}(J(\mathbf{r}(\mathbf{a}))) = \mathit{rank}(\mathbf{M}(\mathbf{a})) = k.(p-k)$ during the iterations (or at least that these conditions are verified as soon as $\lambda = 0$). If $\lambda = 0$, we simply obtained $d\mathbf{a}_{gp-lm} = d\mathbf{a}_{gp-gn}$ and $d\mathbf{a}_{k-lm} = d\mathbf{a}_{k-gn}$ as in version (2) of the Levenberg-Marquardt algorithm when $\|\nabla\psi(\mathbf{a})\|_2 < \|\nabla\psi\|_{min}$. However, the key-difference with version (2) of the Levenberg-Marquardt algorithms is that these two linear systems remain nonsingular and well-conditioned when λ approaches zero thanks to the addition of the term $\mathbf{N}\mathbf{N}^T$ in the coefficient matrix of these normal equations. In other words, in these conditions, $\mathbf{R}_\Delta(\mathbf{N},\lambda)$ and $\mathbf{R}_\Lambda(\mathbf{N},\lambda)$ are nonsingular upper triangular matrices and the normal equations can be solved by forward and backward substitutions to get $d\mathbf{a}_{gp-lm}$ and $d\mathbf{a}_{k-lm}$ in step **(6)** of the Levenberg-Marquardt algorithms (3) and (4), respectively,:

$$d\mathbf{a}_{gp-lm} = \mathbf{R}_\Delta(\mathbf{N},\lambda)^{-1}\mathbf{R}_\Delta(\mathbf{N},\lambda)^{-T}\mathbf{M}(\mathbf{a})^T\mathbf{r}(\mathbf{a})$$

or

$$d\mathbf{a}_{k-lm} = \mathbf{R}_\Lambda(\mathbf{N},\lambda)^{-1}\mathbf{R}_\Lambda(\mathbf{N},\lambda)^{-T}\mathbf{M}(\mathbf{a})^T\mathbf{r}(\mathbf{a})\ .$$

Note that, as this step **(6)** of the Levenberg-Marquardt algorithms (3) and (4) has to be performed several times with different values of λ, but the same matrix $\mathbf{N}$, for some particular iterations i of these algorithms, it is convenient to first add the cross-product matrix $\mathbf{N}\mathbf{N}^T$ to Δ or Λ at each iteration and, then update only the diagonal of these intermediate matrices with $\lambda\mathbf{D}^2$ before computing the Cholesky decomposition for a new λ value at each subiteration j of the algorithms.

If we want to use a more accurate QR approach in the Levenberg-Marquardt algorithms (2), (3) and (4), the key-observation is to recognize that the linear systems (6.27), (6.28), (6.29) and (6.30) solved in the above Cholesky approach are, respectively, the normal equations of the damped linear least-squares problems

$$d\mathbf{a}_{gp-lm} = \text{Arg}\min_{d\mathbf{a}\in\mathbb{R}^{p.k}} \left\| \begin{bmatrix}\mathbf{r}(\mathbf{a})\\ \mathbf{0}^{p.k}\end{bmatrix} - \begin{bmatrix}\mathbf{M}(\mathbf{a})+\mathbf{L}(\mathbf{a})\\ \sqrt{\lambda}\mathbf{D}\end{bmatrix} d\mathbf{a}\right\|_2^2 \tag{6.31}$$

$$d\mathbf{a}_{k-lm} = \text{Arg}\min_{d\mathbf{a}\in\mathbb{R}^{p.k}} \left\| \begin{bmatrix}\mathbf{r}(\mathbf{a})\\ \mathbf{0}^{p.k}\end{bmatrix} - \begin{bmatrix}\mathbf{M}(\mathbf{a})\\ \sqrt{\lambda}\mathbf{D}\end{bmatrix} d\mathbf{a}\right\|_2^2, \tag{6.32}$$

for the Levenberg-Marquardt algorithms (2), and of the constrained and damped linear least-squares problems

$$d\mathbf{a}_{gp-lm} = \text{Arg}\min_{d\mathbf{a}\in\mathbb{R}^{p.k}} \left\| \begin{bmatrix}\mathbf{r}(\mathbf{a})\\ \mathbf{0}^{k.k}\\ \mathbf{0}^{p.k}\end{bmatrix} - \begin{bmatrix}\mathbf{M}(\mathbf{a})+\mathbf{L}(\mathbf{a})\\ \mathbf{N}^T\\ \sqrt{\lambda}\mathbf{D}\end{bmatrix} d\mathbf{a}\right\|_2^2 \tag{6.33}$$

$$d\mathbf{a}_{k-lm} = \text{Arg}\min_{d\mathbf{a}\in\mathbb{R}^{p.k}} \left\| \begin{bmatrix}\mathbf{r}(\mathbf{a})\\ \mathbf{0}^{k.k}\\ \mathbf{0}^{p.k}\end{bmatrix} - \begin{bmatrix}\mathbf{M}(\mathbf{a})\\ \mathbf{N}^T\\ \sqrt{\lambda}\mathbf{D}\end{bmatrix} d\mathbf{a}\right\|_2^2, \tag{6.34}$$

for the Levenberg-Marquardt algorithms (3) and (4).

Thus, the linear least-squares problems (6.31), (6.32), (6.33) and (6.34) can be solved by computing, respectively, a thin QR decomposition of the "damped" Jacobian matrices

$$J(\mathbf{r}(\mathbf{a}))(\lambda) = \begin{bmatrix} \mathbf{M}(\mathbf{a}) + \mathbf{L}(\mathbf{a}) \\ \sqrt{\lambda}\mathbf{D} \end{bmatrix} \quad \text{or} \quad \mathbf{M}(\mathbf{a})(\lambda) = \begin{bmatrix} \mathbf{M}(\mathbf{a}) \\ \sqrt{\lambda}\mathbf{D} \end{bmatrix} , \tag{6.35}$$

for the Levenberg-Marquardt algorithms (2) and a thin QR decomposition of the "constrained" and damped" Jacobian matrices

$$J(\mathbf{r}(\mathbf{a}))(\mathbf{N}, \lambda) = \begin{bmatrix} \mathbf{M}(\mathbf{a}) + \mathbf{L}(\mathbf{a}) \\ \mathbf{N}^T \\ \sqrt{\lambda}\mathbf{D} \end{bmatrix} \quad \text{or} \quad \mathbf{M}(\mathbf{a})(\mathbf{N}, \lambda) = \begin{bmatrix} \mathbf{M}(\mathbf{a}) \\ \mathbf{N}^T \\ \sqrt{\lambda}\mathbf{D} \end{bmatrix} , \tag{6.36}$$

for the Levenberg-Marquardt algorithms (3) and (4). Furthermore, these different thin QR decompositions can again be done in several steps if we take into account the block-column structure of these constrained and damped Jacobian matrices in order to reduce the memory footprint of the algorithms.

In the first stage, for all the algorithms, we compute the QR decomposition of $\mathbf{M}(\mathbf{a}) + \mathbf{L}(\mathbf{a})$ or $\mathbf{M}(\mathbf{a})$ (without column pivoting) with the same TSQR algorithms as used in the Gauss-Newton methods. This produces implicitly the thin QR factorizations

$$\mathbf{M}(\mathbf{a}) + \mathbf{L}(\mathbf{a}) = \mathbf{Q}_J\mathbf{R}_J \text{ or } \mathbf{M}(\mathbf{a}) = \mathbf{Q}_\mathbf{M}\mathbf{R}_\mathbf{M} ,$$

where $\mathbf{Q}_J$ and $\mathbf{Q}_\mathbf{M}$ are $n.p \times k.p$ matrices with orthonormal columns, and, $\mathbf{R}_J$ and $\mathbf{R}_\mathbf{M}$ are $k.p \times k.p$ singular upper triangular matrices as discussed in the previous subsection.

Next, in step **(6.1)** of the Levenberg-Marquardt algorithms (2) in which $\lambda > 0$, we first note, using the above thin QR decomposition of $\mathbf{M}(\mathbf{a}) + \mathbf{L}(\mathbf{a})$ or $\mathbf{M}(\mathbf{a})$, that we have

$$J\big(\mathbf{r}(\mathbf{a})\big)(\lambda) = \begin{bmatrix} \mathbf{Q}_J\mathbf{R}_J \\ \sqrt{\lambda}\mathbf{D} \end{bmatrix} = \begin{bmatrix} \mathbf{Q}_J & \mathbf{0}^{n.p \times k.p} \\ \mathbf{0}^{k.p \times k.p} & \mathbf{I}_{k.p} \end{bmatrix} \begin{bmatrix} \mathbf{R}_J \\ \sqrt{\lambda}\mathbf{D} \end{bmatrix}$$

and

$$\mathbf{M}(\mathbf{a})(\lambda) = \begin{bmatrix} \mathbf{Q}_\mathbf{M}\mathbf{R}_\mathbf{M} \\ \sqrt{\lambda}\mathbf{D} \end{bmatrix} = \begin{bmatrix} \mathbf{Q}_\mathbf{M} & \mathbf{0}^{n.p \times k.p} \\ \mathbf{0}^{k.p \times k.p} & \mathbf{I}_{k.p} \end{bmatrix} \begin{bmatrix} \mathbf{R}_\mathbf{M} \\ \sqrt{\lambda}\mathbf{D} \end{bmatrix} .$$

Thus, in a second stage, we can factorize the block-column matrices $\begin{bmatrix} \mathbf{R}_J \\ \sqrt{\lambda}\mathbf{D} \end{bmatrix}$ or $\begin{bmatrix} \mathbf{R}_\mathbf{M} \\ \sqrt{\lambda}\mathbf{D} \end{bmatrix}$ on the right hand side of these equations into the product of an orthogonal matrix times a rectangular matrix in upper triangular form. This can be done efficiently with a sequence of $k.p.(k.p+1)$ Givens rotations applied to the left of these block-column matrices to annihilate their bottom diagonal block, $\sqrt{\lambda}\mathbf{D}$. These Givens rotations use the diagonal elements of $\mathbf{R}_J$ and $\mathbf{R}_\mathbf{M}$ to eliminate the diagonal elements of $\sqrt{\lambda}\mathbf{D}$ and reduce the fill-in in this process. Note further that, in this recursive process, the bands of zeros introduced in the previous stages are unaffected by the subsequent stages thanks to the use of Givens rotations in the calculations; see Section 10.3 of Nocedal and Wright [139] for a good account of this computing scheme. At the end, we get the matrix equations

$$\mathbf{W}_J(\lambda) \begin{bmatrix} \mathbf{R}_J \\ \sqrt{\lambda}\mathbf{D} \end{bmatrix} = \begin{bmatrix} \mathbf{R}_J(\lambda) \\ \mathbf{0}^{k.p \times k.p} \end{bmatrix} \text{ or } \mathbf{W}_\mathbf{M}(\lambda) \begin{bmatrix} \mathbf{R}_\mathbf{M} \\ \sqrt{\lambda}\mathbf{D} \end{bmatrix} = \begin{bmatrix} \mathbf{R}_\mathbf{M}(\lambda) \\ \mathbf{0}^{k.p \times k.p} \end{bmatrix} .$$

Here, $\mathbf{W}_J(\lambda)$ and $\mathbf{W}_\mathbf{M}(\lambda)$ are $2.k.p \times 2.k.p$ orthogonal matrices, which are the products of $k.p.(k.p + 1)$ Givens rotations, and, $\mathbf{R}_J(\lambda)$ and $\mathbf{R}_\mathbf{M}(\lambda)$ are nonsingular $k.p \times k.p$ upper triangular matrices as $\lambda > 0$ and $\mathbf{D}$ has no zero elements on its diagonal.

Proceeding in this way, we implicitly build up thin QR decompositions of $J(\mathbf{r}(\mathbf{a}))(\lambda)$ or $\mathbf{M}(\mathbf{a})(\lambda)$ in several stages since

$$\begin{aligned} J\big(\mathbf{r}(\mathbf{a})\big)(\lambda) &= \begin{bmatrix} \mathbf{Q}_J & \mathbf{0}^{n.p\times k.p} \\ \mathbf{0}^{k.p\times k.p} & \mathbf{I}_{k.p} \end{bmatrix} \mathbf{W}_J(\lambda)^T \begin{bmatrix} \mathbf{R}_J(\lambda) \\ \mathbf{0}^{k.p\times k.p} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{Q}_J\mathbf{W}_J^{11}(\lambda)^T \\ \mathbf{W}_J^{12}(\lambda)^T \end{bmatrix} \mathbf{R}_J(\lambda) \\ &= \mathbf{Q}_J(\lambda)\mathbf{R}_J(\lambda) \end{aligned}$$

and

$$\begin{aligned} \mathbf{M}(\mathbf{a})(\lambda) &= \begin{bmatrix} \mathbf{Q}_\mathbf{M} & \mathbf{0}^{n.p\times k.p} \\ \mathbf{0}^{k.p\times k.p} & \mathbf{I}_{k.p} \end{bmatrix} \mathbf{W}_\mathbf{M}(\lambda)^T \begin{bmatrix} \mathbf{R}_\mathbf{M}(\lambda) \\ \mathbf{0}^{k.p\times k.p} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{Q}_\mathbf{M}\mathbf{W}_\mathbf{M}^{11}(\lambda)^T \\ \mathbf{W}_\mathbf{M}^{12}(\lambda)^T \end{bmatrix} \mathbf{R}_\mathbf{M}(\lambda) \\ &= \mathbf{Q}_\mathbf{M}(\lambda)\mathbf{R}_\mathbf{M}(\lambda)\ , \end{aligned}$$

where the $2.k.p \times 2.k.p$ orthogonal matrices $\mathbf{W}_J(\lambda)$ and $\mathbf{W}_\mathbf{M}(\lambda)$ have been partitioned in four blocks of $k.p$ rows and columns each:

$$\mathbf{W}_J(\lambda) = \begin{bmatrix} \mathbf{W}_J^{11}(\lambda) & \mathbf{W}_J^{12}(\lambda) \\ \mathbf{W}_J^{21}(\lambda) & \mathbf{W}_J^{22}(\lambda) \end{bmatrix},\ \mathbf{W}_\mathbf{M}(\lambda) = \begin{bmatrix} \mathbf{W}_\mathbf{M}^{11}(\lambda) & \mathbf{W}_\mathbf{M}^{12}(\lambda) \\ \mathbf{W}_\mathbf{M}^{21}(\lambda) & \mathbf{W}_\mathbf{M}^{22}(\lambda) \end{bmatrix},$$

and the $(p.n+k.p)\times k.p$ matrices $\mathbf{Q}_J(\lambda)$ and $\mathbf{Q}_\mathbf{M}(\lambda)$ have orthonormal columns since $\mathbf{W}_J(\lambda)$ and $\mathbf{W}_\mathbf{M}(\lambda)$ are orthogonal matrices, and, the matrices $\mathbf{Q}_J$ and $\mathbf{Q}_\mathbf{M}$ have orthonormal columns.

Finally, using these thin QR decompositions of $J(\mathbf{r}(\mathbf{a}))(\lambda)$ and $\mathbf{M}(\mathbf{a})(\lambda)$, the solutions $d\mathbf{a}_{gp-lm}$ and $d\mathbf{a}_{k-lm}$ of the damped linear least-square problems (6.31) and (6.32), which must be solved in step **(6.1)** of the Levenberg-Marquardt algorithms (2), can be easily computed in a last step as $\mathbf{R}_J(\lambda)$ and $\mathbf{R}_\mathbf{M}(\lambda)$ are nonsingular upper triangular matrices, e.g.,

$$d\mathbf{a}_{gp-lm} = \mathbf{R}_J(\lambda)^{-1}\mathbf{Q}_J(\lambda)^T \begin{bmatrix} \mathbf{r}(\mathbf{a}) \\ \mathbf{0}^{k.p} \end{bmatrix} \text{ or } d\mathbf{a}_{k-lm} = \mathbf{R}_\mathbf{M}(\lambda)^{-1}\mathbf{Q}_\mathbf{M}(\lambda)^T \begin{bmatrix} \mathbf{r}(\mathbf{a}) \\ \mathbf{0}^{k.p} \end{bmatrix}\ .$$

Next, if we want to use a QR approach in the Levenberg-Marquardt algorithms (3) and (4), we have to solve the constrained and damped linear least-squares problems (6.33) or (6.34) at each iteration. This can also be done by computing the thin QR decomposition of the block-column matrices $J(\mathbf{r}(\mathbf{a}))(\mathbf{N},\lambda)$ and $\mathbf{M}(\mathbf{a})(\mathbf{N},\lambda)$ defined in equation (6.36) in several steps to reduce the memory footprint of the algorithms. More precisely with one more steps compared to the structured QR algorithm used in step **(6.1)** of the Levenberg-Marquardt algorithms (2) to reduce the matrices $J(\mathbf{r}(\mathbf{a}))(\lambda)$ and $\mathbf{M}(\mathbf{a})(\lambda)$ (defined in equation (6.35)) to triangular form.

Thus, after using the same first stage as before to get the QR decomposition of $\mathbf{M}(\mathbf{a}) + \mathbf{L}(\mathbf{a})$ or $\mathbf{M}(\mathbf{a})$ with the TSQR algorithms, we next compute implicitly the thin QR factorizations of

$$J\big(\mathbf{r}(\mathbf{a})\big)(\mathbf{N}) = \begin{bmatrix} \mathbf{M}(\mathbf{a}) + \mathbf{L}(\mathbf{a}) \\ \mathbf{N}^T \end{bmatrix} = \mathbf{Q}_J(\mathbf{N})\mathbf{R}_J(\mathbf{N})$$

or

$$\mathbf{M}(\mathbf{a})(\mathbf{N}) = \begin{bmatrix} \mathbf{M}(\mathbf{a}) \\ \mathbf{N}^T \end{bmatrix} = \mathbf{Q}_\mathbf{M}(\mathbf{N})\mathbf{R}_\mathbf{M}(\mathbf{N})\ ,$$

where $\mathbf{Q}_J(\mathbf{N})$ and $\mathbf{Q}_\mathbf{M}(\mathbf{N})$ are $(p.n + k.k) \times k.p$ matrices with orthonormal columns and $\mathbf{R}_J(\mathbf{N})$ and $\mathbf{R}_\mathbf{M}(\mathbf{N})$ are $k.p \times k.p$ upper triangular matrices. Since

$$J\big(\mathbf{r}(\mathbf{a})\big)(\mathbf{N}) = \begin{bmatrix} \mathbf{Q}_J\mathbf{R}_J \\ \mathbf{N}^T \end{bmatrix} = \begin{bmatrix} \mathbf{Q}_J & \mathbf{0}^{n.p\times k.k} \\ \mathbf{0}^{k.k\times k.p} & \mathbf{I}_{k.k} \end{bmatrix} \begin{bmatrix} \mathbf{R}_J \\ \mathbf{N}^T \end{bmatrix}$$

and

$$\mathbf{M}(\mathbf{a})(\mathbf{N}) = \begin{bmatrix} \mathbf{Q_M R_M} \\ \mathbf{N}^T \end{bmatrix} = \begin{bmatrix} \mathbf{Q_M} & \mathbf{0}^{n.p \times k.k} \\ \mathbf{0}^{k.k \times k.p} & \mathbf{I}_{k.k} \end{bmatrix} \begin{bmatrix} \mathbf{R_M} \\ \mathbf{N}^T \end{bmatrix} ,$$

this can be done by performing structured and thin QR factorizations of the matrices $\begin{bmatrix} \mathbf{R}_J \\ \mathbf{N}^T \end{bmatrix}$ and $\begin{bmatrix} \mathbf{R_M} \\ \mathbf{N}^T \end{bmatrix}$; more precisely, by applying a sequence of $k.p$ dedicated Householder transformations on the left of the matrices. These Householder transformations are designed to annihilate the lower block $\mathbf{N}^T$ of these matrices, giving the matrix equalities,

$$\mathbf{W}_J(\mathbf{N}) \begin{bmatrix} \mathbf{R}_J \\ \mathbf{N}^T \end{bmatrix} = \begin{bmatrix} \mathbf{R}_J(\mathbf{N}) \\ \mathbf{0}^{k.k \times k.p} \end{bmatrix} \quad \text{or} \quad \mathbf{W_M}(\mathbf{N}) \begin{bmatrix} \mathbf{R_M} \\ \mathbf{N}^T \end{bmatrix} = \begin{bmatrix} \mathbf{R_M}(\mathbf{N}) \\ \mathbf{0}^{k.k \times k.p} \end{bmatrix} ,$$

where $\mathbf{W}_J(\mathbf{N})$ and $\mathbf{W_M}(\mathbf{N})$ are $k.(p+k) \times k.(p+k)$ orthogonal matrices composed of the product of these $k.p$ elementary Householder transformations, and $\mathbf{R}_J(\mathbf{N})$ and $\mathbf{R_M}(\mathbf{N})$ are nonsingular upper triangular matrices if *rank*$(\mathbf{A}) = k$ and *rank*$(\mathbf{M}(\mathbf{a}) + \mathbf{L}(\mathbf{a}))$ = *rank*$(\mathbf{M}(\mathbf{a})) = k.(p-k)$. This reduction to triangular from is exactly similar to one of the steps of the serial TSQR algorithm described in Subsection 6.1, after the first one. Note also that, in this recursive process, the band of zeros introduced in the previous stages or preceding Householder transformations are unaffected by the subsequent stages if dedicated Householder transformations are used in the calculations.

Using this computational sequence, we finally obtain a thin QR factorization of $J(\mathbf{r}(\mathbf{a}))(\mathbf{N})$ or $\mathbf{M}(\mathbf{a})(\mathbf{N})$ since

$$\begin{aligned} J\big(\mathbf{r}(\mathbf{a})\big)(\mathbf{N}) &= \begin{bmatrix} \mathbf{Q}_J & \mathbf{0}^{n.p \times k.k} \\ \mathbf{0}^{k.k \times k.p} & \mathbf{I}_{k.k} \end{bmatrix} \mathbf{W}_J(\mathbf{N})^T \begin{bmatrix} \mathbf{R}_J(\mathbf{N}) \\ \mathbf{0}^{k.k \times k.p} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{Q}_J \mathbf{W}_J^{11}(\mathbf{N})^T \\ \mathbf{W}_J^{12}(\mathbf{N})^T \end{bmatrix} \mathbf{R}_J(\mathbf{N}) \\ &= \mathbf{Q}_J(\mathbf{N}) \mathbf{R}_J(\mathbf{N}) \end{aligned}$$

and

$$\begin{aligned} \mathbf{M}(\mathbf{a})(\mathbf{N}) &= \begin{bmatrix} \mathbf{Q_M} & \mathbf{0}^{n.p \times k.k} \\ \mathbf{0}^{k.k \times k.p} & \mathbf{I}_{k.k} \end{bmatrix} \mathbf{W_M}(\mathbf{N})^T \begin{bmatrix} \mathbf{R_M}(\mathbf{N}) \\ \mathbf{0}^{k.k \times k.p} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{Q_M} \mathbf{W_M^{11}}(\mathbf{N})^T \\ \mathbf{W_M^{12}}(\mathbf{N})^T \end{bmatrix} \mathbf{R_M}(\mathbf{N}) \\ &= \mathbf{Q_M}(\mathbf{N}) \mathbf{R_M}(\mathbf{N}) , \end{aligned}$$

where the $k.(p+k) \times k.(p+k)$ orthogonal matrices $\mathbf{W}_J(\mathbf{N})$ and $\mathbf{W_M}(\mathbf{N})$ have been partitioned in four blocks as

$$\mathbf{W}_J(\mathbf{N}) = \begin{bmatrix} \mathbf{W}_J^{11}(\mathbf{N}) & \mathbf{W}_J^{12}(\mathbf{N}) \\ \mathbf{W}_J^{21}(\mathbf{N}) & \mathbf{W}_J^{22}(\mathbf{N}) \end{bmatrix} \text{ and } \mathbf{W_M}(\mathbf{N}) = \begin{bmatrix} \mathbf{W_M^{11}}(\mathbf{N}) & \mathbf{W_M^{12}}(\mathbf{N}) \\ \mathbf{W_M^{21}}(\mathbf{N}) & \mathbf{W_M^{22}}(\mathbf{N}) \end{bmatrix} ,$$

where

$$\begin{aligned} &\mathbf{W}_J^{11}(\mathbf{N}), \mathbf{W_M^{11}}(\mathbf{N}) \in \mathbb{R}^{k.p \times k.p} , \\ &\mathbf{W}_J^{12}(\mathbf{N}), \mathbf{W_M^{12}}(\mathbf{N}) \in \mathbb{R}^{k.p \times k.k} , \\ &\mathbf{W}_J^{21}(\mathbf{N}), \mathbf{W_M^{21}}(\mathbf{N}) \in \mathbb{R}^{k.k \times k.p} , \\ &\mathbf{W}_J^{22}(\mathbf{N}), \mathbf{W_M^{22}}(\mathbf{N}) \in \mathbb{R}^{k.k \times k.k} , \end{aligned}$$

and the $(n.p + k.k) \times k.p$ matrices $\mathbf{Q}_J(\mathbf{N})$ and $\mathbf{Q_M}(\mathbf{N})$ have orthonormal columns since $\mathbf{W}_J(\mathbf{N})$ and $\mathbf{W_M}(\mathbf{N})$ are orthogonal matrices and the $n.p \times k.p$ matrices $\mathbf{Q}_J$ and $\mathbf{Q_M}$ have orthonormal columns.

Using these results, we can write, finally,

$$J\big(\mathbf{r}(\mathbf{a})\big)(\mathbf{N},\lambda) = \begin{bmatrix} J(\mathbf{r}(\mathbf{a}))(\mathbf{N}) \\ \sqrt{\lambda}\mathbf{D} \end{bmatrix} = \begin{bmatrix} \mathbf{Q}_J(\mathbf{N})\mathbf{R}_J(\mathbf{N}) \\ \sqrt{\lambda}\mathbf{D} \end{bmatrix} = \begin{bmatrix} \mathbf{Q}_J(\mathbf{N}) & \mathbf{0}^{(n.p+k.k)\times k.p} \\ \mathbf{0}^{k.p\times k.p} & \mathbf{I}_{k.p} \end{bmatrix} \begin{bmatrix} \mathbf{R}_J(\mathbf{N}) \\ \sqrt{\lambda}\mathbf{D} \end{bmatrix}$$

and

$$\mathbf{M}(\mathbf{a})(\mathbf{N},\lambda) = \begin{bmatrix} \mathbf{M}(\mathbf{a})(\mathbf{N}) \\ \sqrt{\lambda}\mathbf{D} \end{bmatrix} = \begin{bmatrix} \mathbf{Q}_\mathbf{M}(\mathbf{N})\mathbf{R}_\mathbf{M}(\mathbf{N}) \\ \sqrt{\lambda}\mathbf{D} \end{bmatrix} = \begin{bmatrix} \mathbf{Q}_\mathbf{M}(\mathbf{N}) & \mathbf{0}^{(n.p+k.k)\times k.p} \\ \mathbf{0}^{k.p\times k.p} & \mathbf{I}_{k.p} \end{bmatrix} \begin{bmatrix} \mathbf{R}_\mathbf{M}(\mathbf{N}) \\ \sqrt{\lambda}\mathbf{D} \end{bmatrix},$$

and the thin QR decompositions of $J(\mathbf{r}(\mathbf{a}))(\mathbf{N},\lambda)$ and $\mathbf{M}(\mathbf{a})(\mathbf{N},\lambda)$ can be obtained by computing the structured and thin QR decompositions of the column-block matrices $\begin{bmatrix} \mathbf{R}_J(\mathbf{N}) \\ \sqrt{\lambda}\mathbf{D} \end{bmatrix}$ and $\begin{bmatrix} \mathbf{R}_\mathbf{M}(\mathbf{N}) \\ \sqrt{\lambda}\mathbf{D} \end{bmatrix}$ using the same computing sequence as performed in the last stage of step **(6.1)** of the Levenberg-Marquardt algorithm (2) described above.

In other words, the elements of the diagonal matrix $\sqrt{\lambda}\mathbf{D}$ can be eliminated by a sequence of $k.p.(k.p+1)$ Givens rotations and, at the end of this process, we get

$$\mathbf{W}_J(\mathbf{N},\lambda)\begin{bmatrix} \mathbf{R}_J(\mathbf{N}) \\ \sqrt{\lambda}\mathbf{D} \end{bmatrix} = \begin{bmatrix} \mathbf{R}_J(\mathbf{N},\lambda) \\ \mathbf{0}^{k.p\times k.p} \end{bmatrix} \text{ or } \mathbf{W}_\mathbf{M}(\mathbf{N},\lambda)\begin{bmatrix} \mathbf{R}_\mathbf{M}(\mathbf{N}) \\ \sqrt{\lambda}\mathbf{D} \end{bmatrix} = \begin{bmatrix} \mathbf{R}_\mathbf{M}(\mathbf{N},\lambda) \\ \mathbf{0}^{k.p\times k.p} \end{bmatrix}.$$

where $\mathbf{W}_J(\mathbf{N},\lambda)$ and $\mathbf{W}_\mathbf{M}(\mathbf{N},\lambda)$ are $2.k.p \times 2.k.p$ orthogonal matrices, which are the products of these $k.p.(k.p+1)$ Givens rotations, and, $\mathbf{R}_J(\mathbf{N},\lambda)$ and $\mathbf{R}_\mathbf{M}(\mathbf{N},\lambda)$ are nonsingular $k.p\times k.p$ upper triangular matrices. Finally, using these matrices equalities, we obtain the thin QR factorizations of $J(\mathbf{r}(\mathbf{a}))(\mathbf{N},\lambda)$ and $\mathbf{M}(\mathbf{a})(\mathbf{N},\lambda)$ since

$$\begin{aligned} J\big(\mathbf{r}(\mathbf{a})\big)(\mathbf{N},\lambda) &= \begin{bmatrix} \mathbf{Q}_J(\mathbf{N}) & \mathbf{0}^{(n.p+k.k)\times k.p} \\ \mathbf{0}^{k.p\times k.p} & \mathbf{I}_{k.p} \end{bmatrix} \mathbf{W}_J(\mathbf{N},\lambda)^T \begin{bmatrix} \mathbf{R}_J(\mathbf{N},\lambda) \\ \mathbf{0}^{k.p\times k.p} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{Q}_J(\mathbf{N})\mathbf{W}_J^{11}(\mathbf{N},\lambda)^T \\ \mathbf{W}_J^{12}(\mathbf{N},\lambda)^T \end{bmatrix} \mathbf{R}_J(\mathbf{N},\lambda) \\ &= \mathbf{Q}_J(\mathbf{N},\lambda)\mathbf{R}_J(\mathbf{N},\lambda) \end{aligned}$$

and

$$\begin{aligned} \mathbf{M}(\mathbf{a})(\mathbf{N},\lambda) &= \begin{bmatrix} \mathbf{Q}_\mathbf{M}(\mathbf{N}) & \mathbf{0}^{(n.p+k.k)\times k.p} \\ \mathbf{0}^{k.p\times k.p} & \mathbf{I}_{k.p} \end{bmatrix} \mathbf{W}_\mathbf{M}(\mathbf{N},\lambda)^T \begin{bmatrix} \mathbf{R}_\mathbf{M}(\mathbf{N},\lambda) \\ \mathbf{0}^{k.p\times k.p} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{Q}_\mathbf{M}(\mathbf{N})\mathbf{W}_\mathbf{M}^{11}(\mathbf{N},\lambda)^T \\ \mathbf{W}_\mathbf{M}^{12}(\mathbf{N},\lambda)^T \end{bmatrix} \mathbf{R}_\mathbf{M}(\mathbf{N},\lambda) \\ &= \mathbf{Q}_\mathbf{M}(\mathbf{N},\lambda)\mathbf{R}_\mathbf{M}(\mathbf{N},\lambda)\,, \end{aligned}$$

where the $2.k.p \times 2.k.p$ orthogonal matrices $\mathbf{W}_J(\mathbf{N},\lambda)$ and $\mathbf{W}_\mathbf{M}(\mathbf{N},\lambda)$ have been partitioned in four blocks of $k.p$ rows and columns each, and the $(p.n + k.k + k.p) \times k.p$ matrices $\mathbf{Q}_J(\mathbf{N},\lambda)$ and $\mathbf{Q}_\mathbf{M}(\mathbf{N},\lambda)$ have orthonormal columns since $\mathbf{W}_J(\mathbf{N},\lambda)$ and $\mathbf{W}_\mathbf{M}(\mathbf{N},\lambda)$ are orthogonal matrices, and, the matrices $\mathbf{Q}_J(\mathbf{N})$ and $\mathbf{Q}_\mathbf{M}(\mathbf{N})$ have orthonormal columns.

Using these thin QR factorizations of $J(\mathbf{r}(\mathbf{a}))(\mathbf{N},\lambda)$ and $\mathbf{M}(\mathbf{a})(\mathbf{N},\lambda)$, the correction vectors $d\mathbf{a}_{gp-lm}$ and $d\mathbf{a}_{k-lm}$ in step **(6)** of the Levenberg-Marquardt algorithms (3) and (4) can then be computed as

$$d\mathbf{a}_{gp-lm} = \mathbf{R}_J(\mathbf{N},\lambda)^{-1}\mathbf{Q}_J(\mathbf{N},\lambda)^T \begin{bmatrix} \mathbf{r}(\mathbf{a}) \\ \mathbf{0}^{k.k+k.p} \end{bmatrix}$$

and

$$d\mathbf{a}_{k-lm} = \mathbf{R}_\mathbf{M}(\mathbf{N},\lambda)^{-1}\mathbf{Q}_\mathbf{M}(\mathbf{N},\lambda)^T \begin{bmatrix} \mathbf{r}(\mathbf{a}) \\ \mathbf{0}^{k.k+k.p} \end{bmatrix},$$

if we use a QR approach in these algorithms.

Note, finally, that if, at a particular iteration i of the Levenberg-Marquardt algorithms (3) and (4), we have to solve several times the constrained and damped linear least squares problems (6.33) and (6.34) for the same value of $\mathbf{a}_i$, but different values of λ in order to ensure the descending condition $\psi(\mathbf{a}_{i+1}) < \psi(\mathbf{a}_i)$ and the convergence of the algorithms, only the last stage of the structured and thin QR factorizations of $J(\mathbf{r}(\mathbf{a}_i))(\mathbf{N}_i, \lambda)$ and $\mathbf{M}(\mathbf{a}_i)(\mathbf{N}_i, \lambda)$ involving the damping parameter λ has to be performed again as the first two stages remain identical if $\mathbf{a}_i$ is not changed. This obviously can save a lot of computing time as $p.n \gg k.p$ for most WLRA problems encountered in practice. This is an interesting feature of the QR approach, see Section 10.3 of Nocedal and Wright [139] for further discussion on these algorithmic, but important, details in a more general NLLS context.

6.3 Variable projection Newton and quasi-Newton algorithms

For large residuals WLRA problems, the variable projection Gauss-Newton and Levenberg-Marquardt algorithms described in Subsections 6.1 and 6.2 can be much less efficient as they do not include second-order derivative information from the Hessian matrix $\nabla^2\psi(\mathbf{a})$ in their associated quadratic models for the variations of $\psi(.)$ in a neighborhood of $\mathbf{a}$. Thus, their asymptotic convergence rates are expected to be only linear in these conditions [45][139]. Fortunately, from the results of Subsection 5.3, we have a compact expression for the Hessian matrix $\nabla^2\psi(\mathbf{a})$ (see equation (5.33))

$$\mathbf{H} = \mathbf{M}(\mathbf{a})^T\mathbf{M}(\mathbf{a}) - \mathbf{L}(\mathbf{a})^T\mathbf{L}(\mathbf{a}) + \left(\mathbf{U}(\mathbf{a})^T\mathbf{L}(\mathbf{a}) + \mathbf{L}(\mathbf{a})^T\mathbf{U}(\mathbf{a})\right) ,$$

under the hypothesis that $\mathbf{F}(.)$ has full column-rank in a neighborhood of $\mathbf{a}$, and, also, all the machinery to implement full Newton algorithms based on this exact three-term expression of $\nabla^2\psi(\mathbf{a})$ or quasi-Newton methods based on its two-term approximation (see equation (5.34))

$$\bar{\mathbf{H}} = \mathbf{M}(\mathbf{a})^T\mathbf{M}(\mathbf{a}) - \mathbf{L}(\mathbf{a})^T\mathbf{L}(\mathbf{a}) .$$

Here $\mathbf{M}(\mathbf{a}), \mathbf{L}(\mathbf{a})$ and $\mathbf{U}(\mathbf{a})$ are defined, respectively, in equations (5.15), (5.19) and (5.20). Note further that, as we assume that $\mathbf{F}(\mathbf{a})$ is of full column rank, e.g., $r_{\mathbf{F}(\mathbf{a})} = \mathit{rank}(\mathbf{F}(\mathbf{a})) = k.p$, we have $\mathbf{F}(\mathbf{a})^+ = \mathbf{F}(\mathbf{a})^-$ and the computation of the three matrices $\mathbf{M}(\mathbf{a}), \mathbf{L}(\mathbf{a})$ and $\mathbf{U}(\mathbf{a})$ can be simplified accordingly. Finally, the cost of the above quasi-Newton methods is similar to those of the Golub-Pereyra Gauss-Newton algorithms using a Cholesky approach as both methods compute exactly the same symmetric matrix terms, e.g., $\mathbf{M}(\mathbf{a})^T\mathbf{M}(\mathbf{a})$ and $\mathbf{L}(\mathbf{a})^T\mathbf{L}(\mathbf{a})$.

Furthermore, as demonstrated in the Cholesky approach of the Gauss-Newton algorithms (1), the computations of these two symmetric terms can be easily parallelized since

$$\mathbf{M}(\mathbf{a})^T\mathbf{M}(\mathbf{a}) = \sum_{j=1}^{n} \mathbf{M}_j^T\mathbf{M}_j \text{ and } \mathbf{L}(\mathbf{a})^T\mathbf{L}(\mathbf{a}) = \sum_{j=1}^{n} \mathbf{L}_j^T\mathbf{L}_j ,$$

where $\mathbf{M}(\mathbf{a})$ and $\mathbf{L}(\mathbf{a})$ have been divided into n blocks $\mathbf{M}_j$ and $\mathbf{L}_j$ as defined in equation (6.2). The last symmetric term $\mathbf{U}(\mathbf{a})^T\mathbf{L}(\mathbf{a}) + \mathbf{L}(\mathbf{a})^T\mathbf{U}(\mathbf{a})$ in the above expression of $\nabla^2\psi(\mathbf{a})$ can also be efficiently computed in two stages. First, by computing in n parallel steps, the term

$$\mathbf{L}(\mathbf{a})^T\mathbf{U}(\mathbf{a}) = \mathbf{V}(\mathbf{a})^T\mathbf{F}(\mathbf{a})^-\mathbf{U}(\mathbf{a}) = \sum_{j=1}^{n} \mathbf{V}_j^T\mathbf{F}_j(\mathbf{a})^-\mathbf{U}_j ,$$

where $\mathbf{V}(\mathbf{a})$ is defined in equation (5.21) and the matrices $\mathbf{U}(\mathbf{a})$ and $\mathbf{V}(\mathbf{a})$ have also been partitioned into n blocks $\mathbf{V}_j$ and $\mathbf{U}_j$ as defined in equations (6.3) and (6.4). Note that, in this last equation, we have used again a symmetric generalized inverse of $\mathbf{F}(\mathbf{a})$ instead of the Moore-Penrose inverse $\mathbf{F}(\mathbf{a})^+$ as we assume here that $\mathbf{F}(\mathbf{a})$ is of full column rank, e.g., that $r_{\mathbf{F}(\mathbf{a})} = \mathit{rank}(\mathbf{F}(\mathbf{a})) = k.p$. In a second stage, we just transpose this squared matrix and sum this squared matrix and its transpose to get the third symmetric term in the above expression of $\nabla^2\psi(\mathbf{a})$.

Using the matrix $\mathbf{H}$ or its two-term approximation $\bar{\mathbf{H}}$, a basic variable projection Newton or quasi-Newton algorithm for the WLRA problem has to solve at each iteration, respectively, the constrained linear systems (see equations (5.40) and (5.41))

$$\big(\mathbf{H} + \mathbf{N}\mathbf{N}^T\big)d\mathbf{a}_n = -\nabla\psi(\mathbf{a}) = \mathbf{M}(\mathbf{a})^T\mathbf{r}(\mathbf{a})$$

and

$$\big(\bar{\mathbf{H}} + \mathbf{N}\mathbf{N}^T\big)d\mathbf{a}_n = -\nabla\psi(\mathbf{a}) = \mathbf{M}(\mathbf{a})^T\mathbf{r}(\mathbf{a})\ ,$$

where the columns of the matrix $\mathbf{N} = \mathbf{K}_{(p,k)}(\mathbf{I}_k \otimes \mathbf{A})$ form a (orthonormal) basis of the null space of $J(\mathbf{r}(\mathbf{a}))$. As already discussed in the previous subsections, such a basis can be easily computed with the help of Corollary 5.6 under the hypothesis that *rank*$(\mathbf{A}) = k$ and *rank*$(J(\mathbf{r}(\mathbf{a}))) =$ *rank*$(\mathbf{M}(\mathbf{a})) = k.(p-k)$. Furthermore, the cross-product matrix $\mathbf{N}\mathbf{N}^T$ can be evaluated efficiently with equation (5.42) and the matrix-product $\mathbf{M}(\mathbf{a})^T\mathbf{r}(\mathbf{a}) = -\nabla\psi(\mathbf{a})$ can also be computed in n parallel steps as the matrices $\mathbf{H}$ and $\bar{\mathbf{H}}$ (see Subsection 6.1 for details).

As discussed at the end of Subsection 5.3, adding the symmetric matrix $\mathbf{N}\mathbf{N}^T$ to $\mathbf{H}$ and $\bar{\mathbf{H}}$ will guarantee that the minimum 2-norm solutions of the consistent linear systems

$$\mathbf{H}d\mathbf{a}_n = \mathbf{M}(\mathbf{a})^T\mathbf{r}(\mathbf{a}) \quad \text{and} \quad \bar{\mathbf{H}}d\mathbf{a}_n = \mathbf{M}(\mathbf{a})^T\mathbf{r}(\mathbf{a})$$

are computed if *rank*$(\mathbf{A}) = k$ and *rank*$(J(\mathbf{r}(\mathbf{a}))) =$ *rank*$(\mathbf{M}(\mathbf{a})) = k.(p-k)$ and, thus, will overcome the systematic singularity and ill-conditioning of $\bar{\mathbf{H}}$ or those of $\mathbf{H}$ at the stationary points of $\psi(.)$ in most cases without using any pre-conditioner for the Hessian matrix or its approximation as suggested in [14]. Note that these two constrained linear systems are based, respectively, on the quadratic approximation models

$$\psi(\mathbf{a} + d\mathbf{a}) \approx N^{\mathbf{N}}(d\mathbf{a}) = \psi(\mathbf{a}) + d\mathbf{a}^T\nabla\psi(\mathbf{a}) + \frac{1}{2}d\mathbf{a}^T\big(\mathbf{H} + \mathbf{N}\mathbf{N}^T\big)d\mathbf{a}$$

and

$$\psi(\mathbf{a} + d\mathbf{a}) \approx N^{\mathbf{N}}(d\mathbf{a}) = \psi(\mathbf{a}) + d\mathbf{a}^T\nabla\psi(\mathbf{a}) + \frac{1}{2}d\mathbf{a}^T\big(\bar{\mathbf{H}} + \mathbf{N}\mathbf{N}^T\big)d\mathbf{a}\ .$$

These quadratic functions are more accurate then the corresponding Golub-Pereyra and Kaufman quadratic models

$$G^{\mathbf{N}}(d\mathbf{a}) = \psi(\mathbf{a}) + d\mathbf{a}^T\nabla\psi(\mathbf{a}) + \frac{1}{2}d\mathbf{a}^T\big(\mathbf{M}(\mathbf{a})^T\mathbf{M}(\mathbf{a}) + \mathbf{L}(\mathbf{a})^T\mathbf{L}(\mathbf{a}) + \mathbf{N}\mathbf{N}^T\big)d\mathbf{a}$$

and

$$G^{\mathbf{N}}(d\mathbf{a}) = \psi(\mathbf{a}) + d\mathbf{a}^T\nabla\psi(\mathbf{a}) + \frac{1}{2}d\mathbf{a}^T\big(\mathbf{M}(\mathbf{a})^T\mathbf{M}(\mathbf{a}) + \mathbf{N}\mathbf{N}^T\big)d\mathbf{a}\ ,$$

which are used in the variable projection Gauss-Newton methods, as information on second-order derivatives of $\psi(.)$ is included in $N^{\mathbf{N}}(.)$, but not in $G^{\mathbf{N}}(.)$.

However, if the addition of the term $\mathbf{N}\mathbf{N}^T$ to the Hessian matrix $\mathbf{H}$, or to its two-term approximation $\bar{\mathbf{H}}$, is useful to ensure that these matrices stay nonsingular and well-conditioned at each iteration, it is not sufficient to guarantee that these two matrices stay positive definite in all the steps and , thus, that $d\mathbf{a}_n$ is always in a descent direction for $\psi(.)$. As explained in Subsection 5.1, this property can be achieved by adding an (another) damping term $\lambda\mathbf{I}_{k.p}$ (where $\lambda > 0$) in the above Newton quadratic models

$$N_\lambda^{\mathbf{N}}(d\mathbf{a}) = \psi(\mathbf{a}) + d\mathbf{a}^T\nabla\psi(\mathbf{a}) + \frac{1}{2}d\mathbf{a}^T\big(\mathbf{H} + \mathbf{N}\mathbf{N}^T + \lambda\mathbf{I}_{k.p}\big)d\mathbf{a}$$

or

$$N_\lambda^{\mathbf{N}}(d\mathbf{a}) = \psi(\mathbf{a}) + d\mathbf{a}^T\nabla\psi(\mathbf{a}) + \frac{1}{2}d\mathbf{a}^T\big(\bar{\mathbf{H}} + \mathbf{N}\mathbf{N}^T + \lambda\mathbf{I}_{k.p}\big)d\mathbf{a}\ ,$$

which lead, respectively, to the damped and constrained linear systems

$$\big(\mathbf{H} + \mathbf{N}\mathbf{N}^T + \lambda\mathbf{I}_{k.p}\big)d\mathbf{a}_n = \mathbf{M}(\mathbf{a})^T\mathbf{r}(\mathbf{a})$$

and

$$\big(\bar{\mathbf{H}} + \mathbf{N}\mathbf{N}^T + \lambda \mathbf{I}_{k.p}\big) d\mathbf{a}_n = \mathbf{M}(\mathbf{a})^T \mathbf{r}(\mathbf{a}) \ ,$$

for the computation of the Newton or quasi-Newton correction step at each iteration.

Remark 6.4. Alternatively, as already discussed at the end of Subsection 5.3, the correction vectors in the variable projection Newton or quasi-Newton algorithms can be computed in a two-step procedure as first suggested by Chen [28]. First, by solving the reduced damped linear system

$$(\bar{\mathbf{O}}^\perp)^T \big(\mathbf{H} + \lambda \mathbf{I}_{k.p}\big) \bar{\mathbf{O}}^\perp d\bar{\mathbf{a}}_n = \big(\mathbf{M}(\mathbf{a})\bar{\mathbf{O}}^\perp\big)^T \mathbf{r}(\mathbf{a})$$

or

$$(\bar{\mathbf{O}}^\perp)^T \big(\bar{\mathbf{H}} + \lambda \mathbf{I}_{k.p}\big) \bar{\mathbf{O}}^\perp d\bar{\mathbf{a}}_n = \big(\mathbf{M}(\mathbf{a})\bar{\mathbf{O}}^\perp\big)^T \mathbf{r}(\mathbf{a}) \ ,$$

for $d\bar{\mathbf{a}}_n \in \mathbb{R}^{(p-k).k}$ and where $\bar{\mathbf{O}}^\perp = \mathbf{K}_{(p,k)}(\mathbf{I}_k \otimes \mathbf{O}^\perp) \in \mathbb{O}^{k.p \times (p-k).k}$ and $\mathbf{O}^\perp \in \mathbb{O}^{p \times (p-k)}$ is an orthonormal matrix whose columns form a basis of $ran(\mathbf{A})^\perp$. Next, $d\mathbf{a}_n$ is computed by

$$d\mathbf{A}_n = \mathbf{O}^\perp d\bar{\mathbf{A}}_n \ ,$$

in a second step. ■

In other words, we can develop variable projection $Levenberg$-$Marquardt$-type Newton and quasi-Newton algorithms with a wider basin of convergence by using the same strategies as used in the Levenberg-Marquardt algorithms described in Subsection 6.2.

Note that we can also define and use a gain factor ρ in the context of these variable projection $Levenberg$-$Marquardt$-type Newton and quasi-Newton methods, e.g.,

$$\rho = \frac{\psi(\mathbf{a}_i) - \psi(\mathbf{a}_i + d\mathbf{a})}{N(\mathbf{0}^{k.p}) - N(d\mathbf{a})} \ ,$$

where $N(.)$ is the more accurate quadratic model used by the (quasi-)Newton iterations and defined by

$$N(d\mathbf{a}) = \psi(\mathbf{a}) + d\mathbf{a}^T \nabla\psi(\mathbf{a}) + \frac{1}{2} d\mathbf{a}^T \mathbf{H} d\mathbf{a}$$

or

$$N(d\mathbf{a}) = \psi(\mathbf{a}) + d\mathbf{a}^T \nabla\psi(\mathbf{a}) + \frac{1}{2} d\mathbf{a}^T \bar{\mathbf{H}} d\mathbf{a} \ ,$$

if we use a quasi-Newton algorithm. Furthermore, in both cases, a direct computation shows that we can compute cheaply the difference $N(\mathbf{0}^{k.p}) - N(d\mathbf{a}_n)$ at each iteration of the (quasi-)Newton algorithms as

$$N(\mathbf{0}^{k.p}) - N(d\mathbf{a}_n) = \frac{1}{2}\big(\|\mathbf{N}_i^T d\mathbf{a}_n\|_2^2 + \lambda \|d\mathbf{a}_n\|_2^2 - d\mathbf{a}_n^T \nabla\psi(\mathbf{a}_i)\big) \ ,$$

similarly as for the Levenberg-Marquardt algorithms (4) developed in the last subsection. Furthermore, $N(\mathbf{0}^{k.p}) - N(d\mathbf{a}_n)$ will be guaranteed to be positive, if $d\mathbf{a}_n$ is in a descent direction for $\psi(.)$ and $\|\nabla\psi(\mathbf{a}_i)\|_2 \neq 0$.

Using these considerations and inspired by the Levenberg-Marquardt algorithms (2), (3) and (4), an outline of three different versions of the variable projection (quasi-)Newton algorithms is detailed below. The definitions and variables used in these Newton and quasi-Newton algorithms have the same meaning as in the previous Gauss-Newton and Levenberg-Marquardt algorithms.

Note that, in all these algorithms, we compute the Newton correction step $d\mathbf{a}_n$ with the exact Hessian matrix $\mathbf{H}$. However, at the user convenience, for example to reduce the computing time per iteration and the memory footprint in all the algorithms, we can eliminate the computation of the

third symmetric term of $\nabla^2\psi(\mathbf{a})$ and use the approximate Hessian matrix $\bar{\mathbf{H}}$ instead to compute the Newton correction step $d\mathbf{a}_n$ at each iteration. Thus, this simple modification defines the quasi-Newton variant for all the algorithms.

Newton algorithms 5.

Choose starting matrix $\mathbf{A}_1 \in \mathbb{R}^{p\times k}$, $\varepsilon_1, \varepsilon_2, \varepsilon_3, \beta \in \mathbb{R}_{+*}$ and $i_{max}, j_{max} \in \mathbb{N}_*$, appropriately

For $i = 1, 2, \ldots$ **until convergence do**

(0) Optionally, compute a QRCP of $\mathbf{A}_i$ (see equation (2.15)) to determine $k_i = \mathit{rank}(\mathbf{A}_i)$ and an orthonormal basis of $\mathit{ran}(\mathbf{A}_i)$:

$$\mathbf{Q}_i\mathbf{A}_i\mathbf{P}_i = \begin{bmatrix} \mathbf{R}_i & \mathbf{S}_i \\ \mathbf{0}^{(p-k_i)\times k_i} & \mathbf{0}^{(p-k_i)\times(k-k_i)} \end{bmatrix},$$

where $\mathbf{Q}_i$ is an $p \times p$ orthogonal matrix, $\mathbf{P}_i$ is an $k \times k$ permutation matrix, $\mathbf{R}_i$ is an $k_i \times k_i$ nonsingular upper triangular matrix (with diagonal elements of decreasing absolute magnitude) and $\mathbf{S}_i$ an $k_i \times (k - k_i)$ full matrix, which is vacuous if $k_i = k$.

In all cases, compute an $p \times k$ matrix $\mathbf{O}_i$ with orthonormal columns as the first k columns of $\mathbf{Q}_i$ (i.e., such that $\mathit{ran}(\mathbf{A}_i) \subset \mathit{ran}(\mathbf{O}_i)$ if $k_i < k$ and $\mathit{ran}(\mathbf{A}_i) = \mathit{ran}(\mathbf{O}_i)$ if $k_i = k$) and set

$$\mathbf{A}_i = \mathbf{O}_i \, .$$

This optional orthogonalization step is a safe-guard as the condition $k_i = k$ is a necessary condition for the differentiability of $\psi(.)$ at a point $\mathbf{A}_i$ and also to limit the occurrence of overflows and underflows in the next steps by enforcing that the matrix variable $\mathbf{A}_i \in \mathbb{O}^{p\times k}$.

(1) Determine (implicitly) the block diagonal matrix

$$\mathbf{F}(\mathbf{a}_i) = \mathit{diag}\big(\mathit{vec}(\sqrt{\mathbf{W}})\big)\big(\mathbf{I}_n \otimes \mathbf{A}_i\big) \, ,$$

where $\mathbf{a}_i = \mathit{vec}(\mathbf{A}_i^T)$.

(2) Compute (implicitly) a QRCP of $\mathbf{F}(\mathbf{a}_i)$ to determine $r_{\mathbf{F}(\mathbf{a}_i)} = \mathit{rank}(\mathbf{F}(\mathbf{a}_i))$, $\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a}_i)}$ and $\mathbf{F}(\mathbf{a}_i)^-$ (see equations (2.18) and (2.19)). Note that $\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a}_i)}$ and $\mathbf{F}(\mathbf{a}_i)^-$ are also block diagonal matrices and that the Newton and quasi-Newton algorithms assume that $r_{\mathbf{F}(\mathbf{a}_i)} = k.p$; see the derivation of the Hessian matrix $\nabla^2\psi(\mathbf{a}_i)$ in Subsection 5.3 for more details. Note that the validity of this hypothesis can be checked here in output of the QRCP of $\mathbf{F}(\mathbf{a}_i)$.

(3) Solve the block diagonal linear least-squares problem

$$\mathbf{b}_i = \mathrm{Arg}\min_{\mathbf{b}\in\mathbb{R}^{k.n}} \|\mathbf{x} - \mathbf{F}(\mathbf{a}_i)\mathbf{b}\|_2^2 \, ,$$

e.g., compute $\mathbf{b}_i = \mathbf{F}(\mathbf{a}_i)^-\mathbf{x}$.

(4) Determine:

$$\mathbf{r}(\mathbf{a}_i) = \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a}_i)}\mathbf{x} \text{ \{current residual vector\}}$$

$$\psi(\mathbf{a}_i) = \tfrac{1}{2}\|\mathbf{r}(\mathbf{a}_i)\|_2^2 \text{ \{current value of the cost function \}}$$

$$\nabla\psi(\mathbf{a}_i) = \mathbf{G}(\mathbf{b}_i)^T\mathbf{G}(\mathbf{b}_i)\mathbf{a}_i - \mathbf{G}(\mathbf{b}_i)^T\mathbf{z} \text{ \{see Theorems 4.3 and 5.7\}}$$

$$\lambda_i = \beta\|\nabla\psi(\mathbf{a}_i)\|_2^2 \text{ \{set ridge parameter proportional to the squared 2-norm of the gradient\}}$$

Note that the steps **(1)** to **(4)** above can be very easily parallelized using the block diagonal structure of $\mathbf{F}(\mathbf{a}_i)$.

(5) Check for convergence. Relevant convergence criteria in the algorithms are of the form:

- $\|\nabla\psi(\mathbf{a}_i)\|_2 \leq \varepsilon_1$

- $\|\mathbf{a}_i - \mathbf{a}_{i-1}\|_2 \leq \varepsilon_2(\varepsilon_2 + \|\mathbf{a}_i\|_2)$ {if $i \neq 1$}.

  If step **(0)** is used, this convergence condition can be simplified as:

  $\|\mathbf{a}_i - \mathbf{a}_{i-1}\|_2 \leq \varepsilon_2\|\mathbf{a}_i\|_2 = \varepsilon_2\sqrt{k}$

- $|\psi(\mathbf{a}_{i-1}) - \psi(\mathbf{a}_i)| \leq \varepsilon_3(\varepsilon_3 + \psi(\mathbf{a}_i))$ {if $i \neq 1$}
- $i \geq i_{max}$ {e.g., give up if the number of iterations is too large}

where $\varepsilon_1, \varepsilon_2, \varepsilon_3$ and i_{max} are constants chosen by the user.

Exit if convergence. **Otherwise, go to** step **(6)**

(6) Compute the Newton correction vector $d\mathbf{a}_n$.

(6.1) To this end, first compute Hessian matrix or its two-term approximation:

$$\mathbf{H}_i^1 = \mathbf{M}(\mathbf{a}_i)^T\mathbf{M}(\mathbf{a}_i) + \mathbf{N}_i\mathbf{N}_i^T + \lambda_i\mathbf{I}_{k.p}$$

$$\mathbf{H}_i^2 = \mathbf{L}(\mathbf{a}_i)^T\mathbf{L}(\mathbf{a}_i)$$

$$\mathbf{H}_i^3 = \mathbf{U}(\mathbf{a}_i)^T\mathbf{L}(\mathbf{a}_i) + \mathbf{L}(\mathbf{a}_i)^T\mathbf{U}(\mathbf{a}_i) \text{ \{only if a full Newton step is wanted\}}$$

$$\mathbf{H}_i = \begin{cases} \mathbf{H}_i^1 - \mathbf{H}_i^2 + \mathbf{H}_i^3 & \{\text{for a full Newton step}\} \\ \mathbf{H}_i^1 - \mathbf{H}_i^2 & \{\text{for a quasi-Newton step}\} \end{cases}$$

where the columns of $\mathbf{N}_i = \mathbf{K}_{(p,k)}(\mathbf{I}_k \otimes \mathbf{A}_i)$ are a (orthonormal) basis of $\mathit{null}\big(\mathbf{M}(\mathbf{a}_i) + \mathbf{L}(\mathbf{a}_i)\big) = \mathit{null}(\mathbf{M}(\mathbf{a}_i))$, see Corollary 5.6.

(6.2) **If** $\mathbf{H}_i$ is positive definite **then** {use Cholesky factorization}

Newton step: get Newton or quasi-Newton step as the solution of the linear system

$$\mathbf{H}_i d\mathbf{a}_n = -\nabla\psi(\mathbf{a}_i) = \mathbf{M}(\mathbf{a}_i)^T\mathbf{r}(\mathbf{a}_i)$$

(6.3) **Else**

Gauss-Newton step: get (Gauss-)Newton step as the solution of the linear system

$$\mathbf{H}_i^1 d\mathbf{a}_n = -\nabla\psi(\mathbf{a}_i) = \mathbf{M}(\mathbf{a}_i)^T\mathbf{r}(\mathbf{a}_i)$$

(7) Increment $\mathbf{a}_i = \mathit{vec}(\mathbf{A}_i^T)$, e.g., compute $\mathbf{a}_{i+1} = \mathit{vec}(\mathbf{A}_{i+1}^T)$ such that $\psi(\mathbf{a}_{i+1}) < \psi(\mathbf{a}_i)$ in order to obtain global convergence.

(7.1) To this end, first compute:

$$\mathbf{a}_{i+1} = \mathbf{a}_i + d\mathbf{a}_n$$

$$\psi(\mathbf{a}_{i+1}) = \tfrac{1}{2}\|\mathbf{r}(\mathbf{a}_{i+1})\|_2^2 = \tfrac{1}{2}\|\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a}_{i+1})}\mathbf{x}\|_2^2\,,$$

using (implicitly) a QRCP of the block diagonal matrix $\mathbf{F}(\mathbf{a}_{i+1})$.

(7.2) **If** $\psi(\mathbf{a}_{i+1}) > \psi(\mathbf{a}_i)$ **then** recompute $\mathbf{a}_{i+1}$ by one of the following methods:

Gauss-Seidel: $\mathbf{a}_{i+1} = \mathbf{a}_i + d\mathbf{a}_{gs-gn}$ where $d\mathbf{a}_{gs-gn}$ is a Gauss-Seidel step [166]

$$d\mathbf{a}_{gs-gn} = \big(\mathbf{K}_{(n,p)}\mathbf{G}(\mathbf{b}_i)\big)^{+}\mathbf{r}(\mathbf{a}_i)$$
$$= \begin{cases} \text{Arg min}_{d\mathbf{a}\in\mathbb{R}^{p.k}} \ \|d\mathbf{a}\|_2^2 \\ \text{s.t. Arg min}_{d\mathbf{a}\in\mathbb{R}^{p.k}} \ \|\mathbf{r}(\mathbf{a}_i) - \mathbf{K}_{(n,p)}\mathbf{G}(\mathbf{b}_i)d\mathbf{a}\|_2^2 \end{cases}$$

Block alternating least-squares:

$$\begin{aligned}\mathbf{a}_{i+1} &= \mathbf{G}(\mathbf{b}_i)^{+}\mathbf{z} \\ &= \begin{cases} \text{Arg}\min_{\mathbf{a}\in\mathbb{R}^{p.k}} \|\mathbf{a}\|_2^2 \\ \text{s.t. Arg}\min_{\mathbf{a}\in\mathbb{R}^{p.k}} \|\mathbf{z}-\mathbf{G}(\mathbf{b}_i)\mathbf{a}\|_2^2 \end{cases}\end{aligned}$$

Line search:

$$\mathbf{a}_{i+1} = \mathbf{a}_i + \alpha_i d\mathbf{a}_n$$

where $\alpha_i < 1$ is determined by a line search to make the algorithm a descent method (i.e., such that $\psi(\mathbf{a}_{i+1}) < \psi(\mathbf{a}_i)$). This is always possible as the correction vector $d\mathbf{a}_n$ is in a descent direction for $\psi(.)$ if $\|\nabla\psi(\mathbf{a}_i)\|_2 \neq 0$.

A simple strategy is to first shorten the correction step to half the Newton length (or Gauss-Newton length if $\mathbf{H}_i$ is not positive definite), compute the new trial value for $\psi(\mathbf{a}_{i+1})$ and, if it is still worse, continue to reduce the step until we get a step short enough such that $\psi(\mathbf{a}_{i+1}) < \psi(\mathbf{a}_i)$. The following loop incorporates this simple step-shortening algorithm:

For $j = 1, 2, \ldots$ **while**$\big(\psi(\mathbf{a}_{i+1}) > \psi(\mathbf{a}_i)\big)$

$d\mathbf{a}_n = \frac{1}{2} d\mathbf{a}_n$

$\mathbf{a}_{i+1} = \mathbf{a}_i + d\mathbf{a}_n$

$\psi(\mathbf{a}_{i+1}) = \frac{1}{2}\|\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a}_{i+1})}\mathbf{x}\|_2^2$ {using a QRCP of the matrix $\mathbf{F}(\mathbf{a}_{i+1})$}

If $j > j_{max}$ **exit** {give up if the number of iterations is too large}

End do

End do

Newton algorithms 6.

Choose starting matrix $\mathbf{A}_1 \in \mathbb{R}^{p\times k}$, $\varepsilon_1, \varepsilon_2, \varepsilon_3, \lambda \in \mathbb{R}_{+*}$ and $i_{max}, j_{max} \in \mathbb{N}_*$, appropriately

For $i = 1, 2, \ldots$ **until convergence do**

(0) Optionally, compute a QRCP of $\mathbf{A}_i$ (see equation (2.15)) to determine $k_i = rank(\mathbf{A}_i)$ and an orthonormal basis of $ran(\mathbf{A}_i)$:

$$\mathbf{Q}_i\mathbf{A}_i\mathbf{P}_i = \begin{bmatrix} \mathbf{R}_i & \mathbf{S}_i \\ \mathbf{0}^{(p-k_i)\times k_i} & \mathbf{0}^{(p-k_i)\times(k-k_i)} \end{bmatrix},$$

where $\mathbf{Q}_i$ is an $p \times p$ orthogonal matrix, $\mathbf{P}_i$ is an $k \times k$ permutation matrix, $\mathbf{R}_i$ is an $k_i \times k_i$ nonsingular upper triangular matrix (with diagonal elements of decreasing absolute magnitude) and $\mathbf{S}_i$ an $k_i \times (k - k_i)$ full matrix, which is vacuous if $k_i = k$.

In all cases, compute an $p \times k$ matrix $\mathbf{O}_i$ with orthonormal columns as the first k columns of $\mathbf{Q}_i$ (i.e., such that $ran(\mathbf{A}_i) \subset ran(\mathbf{O}_i)$ if $k_i < k$ and $ran(\mathbf{A}_i) = ran(\mathbf{O}_i)$ if $k_i = k$) and set

$$\mathbf{A}_i = \mathbf{O}_i \, .$$

This optional orthogonalization step is a safe-guard as the condition $k_i = k$ is a necessary condition for the differentiability of $\psi(.)$ at a point $\mathbf{A}_i$ and also to limit the occurrence of overflows and underflows in the next steps by enforcing that the matrix variable $\mathbf{A}_i \in \mathbb{O}^{p\times k}$.

(1) Determine (implicitly) the block diagonal matrix

$$\mathbf{F}(\mathbf{a}_i) = \mathit{diag}\big(\mathit{vec}(\sqrt{\mathbf{W}})\big)\big(\mathbf{I}_n \otimes \mathbf{A}_i\big)\ ,$$

where $\mathbf{a}_i = \mathit{vec}(\mathbf{A}_i^T)$

(2) Compute (implicitly) a QRCP of $\mathbf{F}(\mathbf{a}_i)$ to determine $r_{\mathbf{F}(\mathbf{a}_i)} = \mathit{rank}(\mathbf{F}(\mathbf{a}_i))$, $\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a}_i)}$ and $\mathbf{F}(\mathbf{a}_i)^{-}$ (see equations (2.18) and (2.19)). Note that $\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a}_i)}$ and $\mathbf{F}(\mathbf{a}_i)^{-}$ are also block diagonal matrices and that the Newton and quasi-Newton algorithms assume that $r_{\mathbf{F}(\mathbf{a}_i)} = k.p$; see the derivation of the Hessian matrix $\nabla^2\psi(\mathbf{a}_i)$ in Subsection 5.3 for more details. Note that, optionally, the validity of this hypothesis can be checked here in output of the QRCP of $\mathbf{F}(\mathbf{a}_i)$.

(3) Solve the block diagonal linear least-squares problem

$$\mathbf{b}_i = \mathrm{Arg}\min_{\mathbf{b}\in\mathbb{R}^{k.n}} \|\mathbf{x} - \mathbf{F}(\mathbf{a}_i)\mathbf{b}\|_2^2\ ,$$

e.g., compute $\mathbf{b}_i = \mathbf{F}(\mathbf{a}_i)^{-}\mathbf{x}$.

(4) Determine and set:

$\mathbf{r}(\mathbf{a}_i) = \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a}_i)}\mathbf{x}$ {current residual vector}

$\psi(\mathbf{a}_i) = \frac{1}{2}\|\mathbf{r}(\mathbf{a}_i)\|_2^2$ {current value of the cost function }

$\nabla\psi(\mathbf{a}_i) = \mathbf{G}(\mathbf{b}_i)^T\mathbf{G}(\mathbf{b}_i)\mathbf{a}_i - \mathbf{G}(\mathbf{b}_i)^T\mathbf{z}$ {see Theorems 4.3 and 5.7}

$j = 0$ {initialize counter for the ridge scaling subiterations}

Note that the steps **(1)** to **(4)** above can be very easily parallelized using the block diagonal structure of $\mathbf{F}(\mathbf{a}_i)$.

(5) Check for convergence. Relevant convergence criteria in the algorithms are of the form:

- $\|\nabla\psi(\mathbf{a}_i)\|_2 \leq \varepsilon_1$
- $\|\mathbf{a}_i - \mathbf{a}_{i-1}\|_2 \leq \varepsilon_2(\varepsilon_2 + \|\mathbf{a}_i\|_2)$ {if $i \neq 1$}

  If step **(0)** is used, this convergence condition can be simplified as:

  $\|\mathbf{a}_i - \mathbf{a}_{i-1}\|_2 \leq \varepsilon_2\|\mathbf{a}_i\|_2 = \varepsilon_2\sqrt{k}$
- $|\psi(\mathbf{a}_{i-1}) - \psi(\mathbf{a}_i)| \leq \varepsilon_3(\varepsilon_3 + \psi(\mathbf{a}_i))$ {if $i \neq 1$}
- $i \geq i_{max}$ {e.g., give up if the number of iterations is too large}

where $\varepsilon_1, \varepsilon_2, \varepsilon_3$ and i_{max} are constants chosen by the user.

Exit if convergence. **Otherwise, go to** step **(6)**

(6) Compute the Newton correction vector $d\mathbf{a}_n$.

(6.1) To this end, first compute Hessian matrix or its two-term approximation:

$$\mathbf{H}_i^1 = \mathbf{M}(\mathbf{a}_i)^T\mathbf{M}(\mathbf{a}_i)$$

$$\mathbf{H}_i^2 = \mathbf{L}(\mathbf{a}_i)^T\mathbf{L}(\mathbf{a}_i)$$

$$\mathbf{H}_i^3 = \mathbf{U}(\mathbf{a}_i)^T\mathbf{L}(\mathbf{a}_i) + \mathbf{L}(\mathbf{a}_i)^T\mathbf{U}(\mathbf{a}_i)\ \{\text{only if a full Newton step is wanted}\}$$

$$\mathbf{H}_i = \begin{cases} \mathbf{H}_i^1 - \mathbf{H}_i^2 + \mathbf{H}_i^3 & \{\text{for a full Newton step}\} \\ \mathbf{H}_i^1 - \mathbf{H}_i^2 & \{\text{for a quasi-Newton step}\} \end{cases}$$

$$\mathbf{H}_i = \mathbf{H}_i + \mathbf{N}_i\mathbf{N}_i^T$$

where the columns of $\mathbf{N}_i = \mathbf{K}_{(p,k)}(\mathbf{I}_k \otimes \mathbf{A}_i)$ are a (orthonormal) basis of $\mathit{null}\big(\mathbf{M}(\mathbf{a}_i) + \mathbf{L}(\mathbf{a}_i)\big) = \mathit{null}(\mathbf{M}(\mathbf{a}_i))$, see Corollary 5.6.

(6.2) Check first diagonal elements of $\mathbf{H}_i$:

$h_{min} = min_{l=1,\cdots,k.p}[\mathbf{H}_i]_{ll}$

If $h_{min} < 0$ **then** $\lambda = \lambda - h_{min}$ {scale up the ridge parameter}

(6.3) **Do while** $\mathbf{H}_i + \lambda \mathbf{I}_{k.p}$ is not positive definite {use Cholesky factorization}

$j = j + 1$

$\lambda = 10.\lambda$ {scale up the ridge parameter}

(6.4) **Newton step:** get Newton step as the solution of the positive definite linear system

$$\big(\mathbf{H}_i + \lambda \mathbf{I}_{k.p}\big) d\mathbf{a}_n = -\nabla\psi(\mathbf{a}_i) = \mathbf{M}(\mathbf{a}_i)^T \mathbf{r}(\mathbf{a}_i)$$

(7) Compute next iterate $\mathbf{a}_{i+1} = vec(\mathbf{A}_{i+1}^T)$ such that $\psi(\mathbf{a}_{i+1}) < \psi(\mathbf{a}_i)$ in order to obtain global convergence.

(7.1) To this end, first compute

$$\psi(\mathbf{a}_i + d\mathbf{a}_n) = \tfrac{1}{2}\|\mathbf{r}(\mathbf{a}_i + d\mathbf{a}_n)\|_2^2 = \tfrac{1}{2}\|\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a}_i + d\mathbf{a}_n)}\mathbf{x}\|_2^2$$

using (implicitly) a QRCP of the block diagonal matrix $\mathbf{F}(\mathbf{a}_i + d\mathbf{a}_n)$.

(7.2) **If** $\psi(\mathbf{a}_i + d\mathbf{a}_n) > \psi(\mathbf{a}_i)$ **then** {step rejected}

$j = j + 1$

$\lambda = 10.\lambda$ {scale up the ridge parameter}

If $j \leq j_{max}$ **go to** step **(6.3)** {recompute $d\mathbf{a}_n$ with inflated diagonal of $\mathbf{H}_i$}

(7.3) **Else** {step acceptable}

If $j = 0$ **then** $\lambda = \lambda/10$ {scale down the ridge parameter if step is successful}

(7.4) Increment $\mathbf{a}_i$:

$\mathbf{a}_{i+1} = \mathbf{a}_i + d\mathbf{a}_n$ {compute new iterate}

End do

Newton algorithms 7.

Choose starting matrix $\mathbf{A}_1 \in \mathbb{R}^{p\times k}$, $\varepsilon_1, \varepsilon_2, \varepsilon_3, \lambda \in \mathbb{R}_{+*}$ and $i_{max}, j_{max} \in \mathbb{N}_*$, appropriately, and initialize $\nu = 2$

For $i = 1, 2, \ldots$ **until convergence do**

(0) Optionally, compute a QRCP of $\mathbf{A}_i$ (see equation (2.15)) to determine $k_i = rank(\mathbf{A}_i)$ and an orthonormal basis of $ran(\mathbf{A}_i)$:

$$\mathbf{Q}_i \mathbf{A}_i \mathbf{P}_i = \begin{bmatrix} \mathbf{R}_i & \mathbf{S}_i \\ \mathbf{0}^{(p-k_i)\times k_i} & \mathbf{0}^{(p-k_i)\times(k-k_i)} \end{bmatrix},$$

where $\mathbf{Q}_i$ is an $p \times p$ orthogonal matrix, $\mathbf{P}_i$ is an $k \times k$ permutation matrix, $\mathbf{R}_i$ is an $k_i \times k_i$ nonsingular upper triangular matrix (with diagonal elements of decreasing absolute magnitude) and $\mathbf{S}_i$ an $k_i \times (k - k_i)$ full matrix, which is vacuous if $k_i = k$.

In all cases, compute an $p \times k$ matrix $\mathbf{O}_i$ with orthonormal columns as the first k columns of $\mathbf{Q}_i$ (i.e., such that $ran(\mathbf{A}_i) \subset ran(\mathbf{O}_i)$ if $k_i < k$ and $ran(\mathbf{A}_i) = ran(\mathbf{O}_i)$ if $k_i = k$) and set

$\mathbf{A}_i = \mathbf{O}_i$.

This optional orthogonalization step is a safe-guard as the condition $k_i = k$ is a necessary condition for the differentiability of $\psi(.)$ at a point $\mathbf{A}_i$ and also to limit the occurrence of overflows and underflows in the next steps by enforcing that the matrix variable $\mathbf{A}_i \in \mathbb{O}^{p\times k}$.

(1) Determine (implicitly) the block diagonal matrix

$$\mathbf{F}(\mathbf{a}_i) = \mathit{diag}\big(\mathit{vec}(\sqrt{\mathbf{W}})\big)\big(\mathbf{I}_n \otimes \mathbf{A}_i\big)\ ,$$

where $\mathbf{a}_i = \mathit{vec}(\mathbf{A}_i^T)$.

(2) Compute (implicitly) a QRCP of $\mathbf{F}(\mathbf{a}_i)$ to determine $r_{\mathbf{F}(\mathbf{a}_i)} = \mathit{rank}(\mathbf{F}(\mathbf{a}_i))$, $\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a}_i)}$ and $\mathbf{F}(\mathbf{a}_i)^-$ (see equations (2.18) and (2.19)). Note that $\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a}_i)}$ and $\mathbf{F}(\mathbf{a}_i)^-$ are also block diagonal matrices and that the Newton and quasi-Newton algorithms assume that $r_{\mathbf{F}(\mathbf{a}_i)} = k.p$; see the derivation of the Hessian matrix $\nabla^2\psi(\mathbf{a}_i)$ in Subsection 5.3 for more details. Note that, optionally, the validity of this hypothesis can be checked here in output of the QRCP of $\mathbf{F}(\mathbf{a}_i)$.

(3) Solve the block diagonal linear least-squares problem

$$\mathbf{b}_i = \mathrm{Arg}\min_{\mathbf{b}\in\mathbb{R}^{k.n}} \|\mathbf{x} - \mathbf{F}(\mathbf{a}_i)\mathbf{b}\|_2^2\ ,$$

e.g., compute $\mathbf{b}_i = \mathbf{F}(\mathbf{a}_i)^-\mathbf{x}$.

(4) Determine and set:

$\mathbf{r}(\mathbf{a}_i) = \mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a}_i)}\mathbf{x}$ {current residual vector}

$\psi(\mathbf{a}_i) = \frac{1}{2}\|\mathbf{r}(\mathbf{a}_i)\|_2^2$ {current value of the cost function }

$\nabla\psi(\mathbf{a}_i) = \mathbf{G}(\mathbf{b}_i)^T\mathbf{G}(\mathbf{b}_i)\mathbf{a}_i - \mathbf{G}(\mathbf{b}_i)^T\mathbf{z}$ {see Theorems 4.3 and 5.7}

$j = 0$ {initialize counter for the ridge scaling subiterations}

Note that the steps **(1)** to **(4)** above can be very easily parallelized using the block diagonal structure of $\mathbf{F}(\mathbf{a}_i)$.

(5) Check for convergence. Relevant convergence criteria in the algorithms are of the form:

- $\|\nabla\psi(\mathbf{a}_i)\|_2 \le \varepsilon_1$
- $\|\mathbf{a}_i - \mathbf{a}_{i-1}\|_2 \le \varepsilon_2(\varepsilon_2 + \|\mathbf{a}_i\|_2)$ {if $i \ne 1$}

  If step **(0)** is used, this convergence condition can be simplified as:

  $\|\mathbf{a}_i - \mathbf{a}_{i-1}\|_2 \le \varepsilon_2\|\mathbf{a}_i\|_2 = \varepsilon_2\sqrt{k}$
- $|\psi(\mathbf{a}_{i-1}) - \psi(\mathbf{a}_i)| \le \varepsilon_3(\varepsilon_3 + \psi(\mathbf{a}_i))$ {if $i \ne 1$}
- $i \ge i_{max}$ {e.g., give up if the number of iterations is too large}

where $\varepsilon_1, \varepsilon_2, \varepsilon_3$ and i_{max} are constants chosen by the user.

Exit if convergence. **Otherwise, go to** step **(6)**

(6) Compute the Newton correction vector $d\mathbf{a}_n$.

(6.1) To this end, first compute Hessian matrix or its two-term approximation:

$$\mathbf{H}_i^1 = \mathbf{M}(\mathbf{a}_i)^T\mathbf{M}(\mathbf{a}_i)$$

$$\mathbf{H}_i^2 = \mathbf{L}(\mathbf{a}_i)^T\mathbf{L}(\mathbf{a}_i)$$

$$\mathbf{H}_i^3 = \mathbf{U}(\mathbf{a}_i)^T\mathbf{L}(\mathbf{a}_i) + \mathbf{L}(\mathbf{a}_i)^T\mathbf{U}(\mathbf{a}_i)\ \{\text{only if a full Newton step is wanted}\}$$

$$\mathbf{H}_i = \begin{cases} \mathbf{H}_i^1 - \mathbf{H}_i^2 + \mathbf{H}_i^3 & \{\text{for a full Newton step}\} \\ \mathbf{H}_i^1 - \mathbf{H}_i^2 & \{\text{for a quasi-Newton step}\} \end{cases}$$

$\mathbf{H}_i = \mathbf{H}_i + \mathbf{N}_i\mathbf{N}_i^T$

where the columns of $\mathbf{N}_i = \mathbf{K}_{(p,k)}(\mathbf{I}_k \otimes \mathbf{A}_i)$ are a (orthonormal) basis of $null\big(\mathbf{M}(\mathbf{a}_i) + \mathbf{L}(\mathbf{a}_i)\big) = null(\mathbf{M}(\mathbf{a}_i))$, see Corollary 5.6.

(6.2) Check first diagonal elements of $\mathbf{H}_i$:

$h_{min} = min_{l=1,\cdots,k.p}[\mathbf{H}_i]_{ll}$

If $h_{min} < 0$ **then** $\lambda = \lambda - h_{min}$ {scale up the ridge parameter}

(6.3) **Do while** $\mathbf{H}_i + \lambda\mathbf{I}_{k.p}$ is not positive definite {use Cholesky factorization}

$j = j + 1$

$\lambda = \nu.\lambda$ {scale up the ridge parameter}

$\nu = 2.\nu$ {increase the growth factor of the ridge parameter}

(6.4) **Newton step:** get Newton step as the solution of the positive definite linear system

$$\big(\mathbf{H}_i + \lambda\mathbf{I}_{k.p}\big)d\mathbf{a}_n = -\nabla\psi(\mathbf{a}_i) = \mathbf{M}(\mathbf{a}_i)^T\mathbf{r}(\mathbf{a}_i)$$

(7) Compute next iterate $\mathbf{a}_{i+1} = vec(\mathbf{A}_{i+1}^T)$ such that $\psi(\mathbf{a}_{i+1}) < \psi(\mathbf{a}_i)$ in order to obtain global convergence.

(7.1) To this end, first compute

$\psi(\mathbf{a}_i + d\mathbf{a}_n) = \frac{1}{2}\|\mathbf{P}^{\perp}_{\mathbf{F}(\mathbf{a}_i + d\mathbf{a}_n)}\mathbf{x}\|_2^2$,

using (implicitly) a QRCP of the block diagonal matrix $\mathbf{F}(\mathbf{a}_i + d\mathbf{a}_n)$, and the gain factor

$$\rho = \frac{\psi(\mathbf{a}_i) - \psi(\mathbf{a}_i + d\mathbf{a}_n)}{N(\mathbf{0}^{k.p}) - N(d\mathbf{a}_n)} = \frac{\psi(\mathbf{a}_i) - \psi(\mathbf{a}_i + d\mathbf{a}_n)}{\frac{1}{2}\left(\|\mathbf{N}_i^T d\mathbf{a}_n\|_2^2 + \lambda\|d\mathbf{a}_n\|_2^2 - d\mathbf{a}_n^T\nabla\psi(\mathbf{a}_i)\right)}$$

(7.2) **If** $\rho > 0$ **then** {step acceptable}

$\lambda = \lambda.max\big(\frac{1}{3}, 1 - (2.\rho - 1)^3\big)$ {scale down the ridge parameter}

$\nu = 2$ {reinitialize the growth factor of the ridge parameter}

(7.3) **Else** {step rejected}

$j = j + 1$

$\lambda = \nu.\lambda$ {scale up the ridge parameter}

$\nu = 2.\nu$ {increase the growth factor of the ridge parameter}

If $j \leq j_{max}$ **go to** step **(6.3)** {recompute $d\mathbf{a}_n$ with inflated diagonal of $\mathbf{H}_i$}

(7.4) Increment $\mathbf{a}_i$:

$\mathbf{a}_{i+1} = \mathbf{a}_i + d\mathbf{a}_n$ {compute new iterate}

End do

As in the Gauss-Newton or Levenberg-Marquardt algorithms, the computations in the above Newton algorithms are terminated either when one or several of the convergence criteria listed in step **(5)** are satisfied, or when the iteration count exceeds the predetermined number i_{max}.

In both the Newton algorithms (6) and (7), the Marquardt parameter λ is taken in the interval $[10^{-8}\ 1]$ and a small value of λ is selected if we believe that $\mathbf{A}_1$ is close to a solution (say $\lambda = 10^{-6}$). Otherwise, we can use $\lambda = 10^{-3}$ or 10^{-4}, or even 1. The algorithms are not very sensitive to this initial choice of λ as this parameter is quickly updated during the iterations in both Newton algorithms (6) and (7). Version (7) of the Newton algorithms also uses a growth factor ν for the

ridge parameter, which is initialized to 2 at the start of the algorithm and reinitialized to this initial value in step **(7.2)** when a Newton step is successful.

Note that it may happens during some iterations that the exact Hessian matrix $\mathbf{H}_i$ or its two-term approximation $\bar{\mathbf{H}}_i$ may become not positive definite in which case the computed Newton correction vector $d\mathbf{a}_n$ is not in a descent direction for $\psi(.)$. The Newton algorithms (5) overcome this difficulty by using immediately a Gauss-Newton step, more precisely a Kaufman Gauss-Newton step, which is always in a descent direction if $\mathbf{a}_i$ is not a stationary point as demonstrated in Corollary 5.7. On the other hand, the Newton algorithms (6) and (7) use the simple strategy of adding a multiple of the identity to the Hessian matrix until this modified Hessian matrix becomes positive definite; see step **(6.3)** in these algorithms. The drawback in this simple approach is that each time we add a multiple of the identity to the Hessian matrix, a new Cholesky factorization of $\mathbf{H}_i + \lambda \mathbf{I}_{k.p}$ must be attempted. This can become very expensive if the process is repeated many times across the iterations. This explains why λ is multiplied by a value as large as 10 in step **(6.3)** of the Newton algorithm (6) and that the updating strategy of λ in step **(6.3)** of the Newton algorithm (7) is modified so that consecutive failures of getting a positive definite modified Hessian matrix give a very fast growth of λ. The same modified strategy with a faster growth of λ is also used in step **(7.3)** of the Newton algorithms (7) when the gain factor is negative and a trial Newton step is rejected. However, in both Newton algorithms (6) and (7), if the number of ridge scaling subiterations in steps **(6.3)** and **(7.2)** (or **(7.3)** for Newton algorithms (7)) become too important, we stop this ridge scaling process at the end of step **(7.2)** (or **(7.3)** for Newton algorithms (7)) and we try to decrease the cost function $\psi(.)$ through the main loop of the algorithms instead.

6.4 Variable projection hybrid algorithms

In the previous subsections, we have presented a large variety of variable projection WLRA solvers based on Gauss-Newton, Levenberg-Marquardt and Newton algorithms, which can all be considered as second-order or pseudo second-order methods. The complexity of these algorithms is relatively high, especially when $p \approx n$, even if $min(p, n) \gg k$, as these algorithms have to solve a large linear system or a tall and skinny linear least-squares problem at each iteration as discussed in the previous subsections. Furthermore, the pre-processing of the Hessian matrix and its different approximations in a normal-equation approach or of the coefficient matrix of the linear least-squares problem in a QR approach for the Gauss-Newton and Levenberg-Marquardt methods is also expensive and has a high memory footprint [28][150][81][88]. In the previous section, we have provided different parallel techniques to reduce both the computing time and the memory storage requirements for this expensive pre-processing step included in all these variable projection (pseudo) second-order algorithms.

However, all these (pseudo) second-order algorithms have a much higher complexity than the block ALS method (e.g., NIPALS, presented in Section 4) or other first-order algorithms based on majorization or Expectation-Maximization (EM) methods, steepest, conjugate or stochastic gradient descent (or combinations of some of them), which have been proposed in the literature to solve the WLRA problem, since most of them are based on relatively inexpensive iterative algorithms that monotonically improve the function value by sequential repetition of local optimizations as the ALS method [98][99][93][91][86][129][167][23][17][181][146]. Moreover, the block ALS method greatly reduces the memory footprint requirements and can also be parallelized easily and very efficiently taking into account the block diagonal structures of the matrices $\mathbf{F}(\mathbf{a})$ and $\mathbf{G}(\mathbf{b})$ as explained in Section 4. On the other hand, it is known that the block ALS method and its variants fail frequently to converge to an acceptable optimal solution for difficult WLRA problems without the use of a proper regularization, and are prone to flattening, especially when the percentage of missing data or/and the level of noise are high [15][28][37][150][81][88].

These different features suggest that hybrid methods combining the fast block ALS algorithm with any of the (pseudo) second-order variable projection algorithms detailed in the last subsections can performed much better than any of the previous individual algorithms as first suggested by [15]

and confirmed by Chen [28] in the specific case of variable projection Newton algorithms. In this way, we can benefit of the remarkable efficiency of the block ALS method in terms of speed and, at the same time, of the good convergence ability of the (pseudo) second-order methods to reduce drastically the computing time without sacrificing the overall performance of the (pseudo) second-order WLRA solvers.

As an illustration, one simple, but still very efficient, choice of such hybrid methods could be to add an additional step consisting of a fixed number of iterations (say between 2 and 20) of the block ALS algorithm before step **(0)** in all our (pseudo) second-order variable projection algorithms. Preliminary tests (not shown) suggest that this simple modification is always very beneficial in terms of speed without impairing the global convergence performance of the (pseudo) second-order variable projection WLRA solvers discussed here. Of course, many variations other than this simple hybrid scheme are possible [15][28].

7 Conclusions and discussion

In this monograph, we consider the difficult WLRA problem, an extension of the well-known matrix completion problem [44]. The WLRA problem is NP-hard and has no closed solution in general [62], but has an increasing number of important applications in practice.

We survey many different approaches which have been used to solve it, with a particular focus on variable projection second-order methods [158][63][166][10][150][149][81]. A large variety of low-complexity first-order methods have been already proposed in the past to solve the WLRA problem [188][186][14][86][17] [181][146], but only a few second-order methods, as these second-order methods have a very high per-iteration complexity, which preclude their use for very large datasets commonly found in recent applications. However, variable projection second-order methods perform better than first-order methods for badly conditioned WLRA problems, which are very common, and these more costly methods are thus still of interest in this context [150][81].

First, we review in detail the connections between manifold, variety, factorization and variable projection formulations of the WLRA problem, which are most often treated as disconnected approaches in the literature and establish relationships between (local) minima, first- and second-order critical points of the objective functions used in the these different formulations of the WLRA problem. These results are an illustration and slight extension in the context of the WLRA problem of recent results presented in [173][84][83][115][113] about the near equivalences of first- and second-order critical points of the objective functions in nonconvex factorization, manifold and variety formulations in more general low-rank matrix optimization.

Second, we provide an extended and original overview of the variable projection formulation of the WLRA problem both from theoretical and algorithmic perspectives. In particular, we study in detail the non-smoothness of the variable projection cost function of the WLRA problem when some weights are equal to zero (e.g., in presence of missing values) and when this cost function is not regularized. We characterize precisely its discontinuities generalizing the preliminary investigations of Dai et al. [46][47] on this topic. These points of discontinuity form barrier sets in the feasible space of solutions, which prevent low-complexity algorithms like gradient descent or alternating least squares to converge to first-order critical points for some WLRA problems when missing values are present. Up to now, most of variable projection algorithms proposed in the literature for solving the WLRA problem simply "ignore" these discontinuities when missing values are present or use a regularized objective function to eliminate them. However, such regularization may degrade significantly the compression quality of the computed low-rank matrix approximation. It is thus interesting to explore whether it is possible to detect if we move in front of these discontinuity points in the unregularized variable projection methods when missing values are present and to find ways of escaping from them during the course of the computations as was discussed in Dai et al. [46] for a gradient descent algorithm. Finally, it is also worth to explore in more depth the

landscape connections between these discontinuity points in the variable projection formulation and the corresponding points in the manifold and factorization formulations of the WLRA problem. As an illustration, the matrix incoherency hypotheses, which are often used to prove the convergence of gradient descent or alternating least-squares for matrix completion problems [187] mainly lead to the exclusion of the above barrier sets.

Next, we derive new formulae for the variable projection gradient vector, and Jacobian and Hessian matrices, which are pivotal in all the second-order variable projection methods. These new formulae also allow a better understanding of the geometric landscape of the objective function associated with the variable projection formulation. In particular, they allow to characterize precisely the systematic rank degeneracy of the variable projection Jacobian matrix and also the singularities of the variable projection Hessian at first-order stationarity points of the variable projection objective function. Furthermore, these new formulae allow us to demonstrate that most of the first- and second-order variable projection methods, which can be used to solve the WLRA problem, can be viewed as as Riemannian optimization methods operating on the Grassmann manifold when the variable projection objective function is smooth over all the points of the Grassmannian.

Our new formulae for the variable projection gradient vector, Jacobian and Hessian matrices also allow us to formulate more accurate and robust variable projection second-order algorithms based on stable orthogonal kernels to tackle the systematic rank degeneracy of the Jacobian, the singularity of the Hessian at first-order stationarity points of the variable projection functional and, finally, the ill-conditioning and indefinite nature of the Hessian at points arbitrarily closed to local minima of this functional. These singularities and instabilities are inescapable here as the (local) minima of the variable projection objective function are always non-isolated and form a differentiable submanifold of the ambient linear space around around each of the minima under some regularity hypotheses of the variable projection functional as demonstrated in Subsection 5.3.

From an algorithmic point of view, these formulae also allow us to formulate more efficient variable projection second-order algorithms to tackle WLRA problems of larger dimensions, which are now the rule in many applications, despite these variable projection second-order algorithms have still a high per-iteration complexity. In particular, we improve significantly the scalability and efficiency of the proposed variable projection second-order algorithms compared to previous studies by taking better into account the sparseness of the different matrix variables involved in each iteration of the algorithms and by using large-scale parallelization techniques and highly optimized BLAS3 kernels in the sensible parts of the algorithms. However, some parts of the variable projection algorithms still do not explicitly take into account the sparsity of the matrix variables. This concerns especially the recursive and parallel implementations of the QR decomposition of the the tall and skinny matrices generated by blocks and used in the iterations of both the Gauss-Newton and Levenberg-Marquardt algorithms. Devising more efficient recursive and parallel QR decompositions of these sparse, tall and skinny matrices is thus an important topic for future research, including possibly the use of very fast randomized QR methods [128].

Finally, here, we have assumed that the rank of the low-rank matrix approximation we are seeking is given or bounded before hand by the user. Of course, in practice, this rank is often unknown. An interesting continuation of this work would thus be to devise efficient variable projection algorithms specifically designed to adaptively select or change the rank of the low-rank matrix solution of the WLRA problem during the computations as was done for Riemannian descent methods on low-rank matrix varieties in [183][184]. Such rank-adaptive optimization strategies in which local minima of smaller rank are used as starting points for improved approximation with a larger rank will be very useful to select the best rank in practical applications and can also lead to improve efficiency and accuracy in the computations of larger, but fixed low-rank matrix approximation, especially for ill-conditioned WLRA problems. More generally, selecting accurate, but cheap, initial low-rank matrix approximations as a first guess of the costly variable projection second-order methods described here is another useful and important continuation of this work.

Obviously, the variable projection second-order algorithms described here must be compared to the many other state-of-the-art first-order methods already proposed in the literature for solving the WLRA problem [188][186][14][86][17] [181][146] in comprehensive benchmark experiments. This benchmark must be based on both synthetic and real datasets and not only restricted to the matrix completion problem as in many past experiments. In particular, we expect that the variable projection second-order algorithms grow more efficient, robust and accurate than gradient-based or alternating least-squares methods as the amount of missing values (e.g., zero weights) increase in the WLRA problem, but this must be objectively validated in comprehensive experiments. Of course, if speed is the priority, first-order (e.g. gradient descent or alternating least squares) or hybrid algorithms as described in Subsection 6.4 will be the methods of choice.

Fortran90 codes with OpenMP and BLAS supports, for the variable projection second-order algorithms described here, will be later available in the open source STATPACK library available at:

https://pagesperso.locean-ipsl.upmc.fr/terray/statpack2.3/index.html.

References

[1] P.-A. Absil, J. Malick (2012) *Projection-like retractions on matrix manifolds*, SIAM J. Optim., 22:135-158.

[2] E.L. Allgower, K. Bohmer, A. Hoy, V. Janovsky (1999) *Direct methods for solving singular nonlinear equations*, Z. Angew. Math. Mech. 79(4):219-231.

[3] P.-A. Absil, R. Mahony, R. Sepulchre (2008) *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, Princeton, NJ.

[4] H. Attouch, J. Bolte, P. Redont, A. Soubeyran (2010) *Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Lojasiewicz inequality*, Mathematics of Operations Research, 35(2):438-457.

[5] A. Ben-Israel (1965) *A modified Newton-Raphson method for the solution of systems of equations*, Israel Journal of Mathematics, 3:94-98

[6] A. Ben-Israel (1966) *A Newton-Raphson method for the solution of systems of equations*, Journal of Mathematical Analysis and Applications, 15:243-252

[7] D.P. Bertsekas (1999) *Nonlinear Programming*, 2nd Edition, Athena Scientific, Belmont, Massachusetts.

[8] A. Bjorck (2015) *Numerical Methods in Matrix Computations*, Series: Texts in Applied Mathematics, Vol. 59, Springer, 800 p.

[9] P.T. Boggs (1976) *The convergence of the Ben-Israel iteration for nonlinear least-squares problems*, Mathematics of Computation, 30(135):512-522

[10] C.F. Borges (2009) *A full-newton approach to separable nonlinear least-squares problems and its application to discrete least-squares rational approximation*, Electronic Transactions on Numerical Analysis, 35:57-68.

[11] N. Boumal (2023) *An introduction to optimization on smooth manifolds*, Cambridge University Press. Also available at https://www.nicolasboumal.net/book.

[12] G.E. Bredon (1993) *Topology and Geometry*, 2nd Edition, Graduate texts in mathematics, Springer-Verlag, New York.

[13] N. Boumal, P.-A. Absil (2011) *RTRMC: a Riemannian trust-region method for low-rank matrix completion*, in: J. Shawe-Taylor, R.S. Zemel, P. Bartlett, F.C.N. Pereira, K.Q. Weinberger (Eds.), Advances in Neural Information Processing Systems24, NIPS, 2011, pp.406-414.

[14] N. Boumal, P.-A. Absil (2015) *Low-rank matrix completion via preconditioned optimization on the Grassmann manifold*, Linear Algebra and its Applications, 475:200-239.

[15] A.M. Buchanan, A.W. Fitzgibbon (2005) *Damped Newton algorithms for matrix factorization with missing data*, in Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2:316-322.

[16] S. Bhojanapalli, P. Jain (2014) *Universal matrix completion*, in Proceedings of the 31st International Conference on Machine Learning (ICML-14), pages 1881-1889, 2014.

[17] D. Bertsimas, M.L. Li (2020) *Fast Exact Matrix Completion: A Unifed Optimization Framework for Matrix Completion*, Journal of Machine Learning Research, 21:1-43.

[18] S. Burer, R.D.C. Monteiro (2003) *A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization*, Mathematical Programming, 95(2):329-357.

[19] S. Bellavia, B. Morini (2015) *Strong local convergence properties of adaptive regularized methods for nonlinear least squares*. IMA J. Numer. Anal. 35(2):947-968.

[20] J.M. Beckers, M. Rixen (2003) *EOF calculations and data filling from incomplete oceanographic datasets*, J. Atoms. Ocean. Tech., 20(12):1839-1856.

[21] A. Beck, L. Tetruashvili (2013) *On the convergence of block coordinate descent type methods*, SIAM J. Optim., 23(4):2037-2060.

[22] . W. Bruns and U. Vetter (1988) *Determinantal rings*, Lecture Notes in Math. 1327, Springer-Verlag, Berlin.

[23] F. Ban, D. Woodruff, Q. Zhang (2019) *Regularized Weighted Low Rank Approximation*, 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

[24] T. Bendokat, R. Zimmermann, P.A. Absil (2023) *A Grassmann manifold handbook: Basic Geometry and Computational Aspects*, arXiv:2011.13699v3. See https://arxiv.org/abs/2011.13699v3.

[25] A. Bhaskara, A.K. Ruwanpathirana, M. Wijewardena (2021) *Additive Error Guarantees for Weighted Low Rank Approximation*, in Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021.

[26] H. Cartan (2017) *Differential calculus on normed spaces: A course in Analysis*, 2nd Edition, 217 pp.

[27] P. Chen (2008a) *Heteroscedastic Low-Rank Matrix Approximation by the Wiberg Algorithm*, IEEE Transactions on signal processing, 56(4):1429-1439.

[28] P. Chen (2008b) *Optimization algorithms on subspaces: Revisiting missing data problem in low-rank matrix*, Int. J. Comput. Vis., 80:125-142.

[29] S.L. Campbell, C.D. Meyer (2009) *Generalized Inverses of Linear Transformations*, Classics in Applied Mathematics, SIAM, Philadelphia.

[30] E.J. Candes, B. Recht (2009) *Exact matrix completion via convex optimization*, Found. Comput. Math. 9(6):717-772.

[31] T.P. Cason,P.A. Absil, P.V. Dooren (2013) *Iterative methods for low rank approximation of graph similarity matrices*, Linear Algebra Appl. 438(4):1863-1882.

[32] J.F. Cai, E.J. Candes, Z. Shen (2010) *A singular value thresholding algorithm for matrix completion*, SIAM J. Optim., 20(4):1956-1982.

[33] M.T. Chu, R.E. Funderlic, R.J. Plemmons (2003) *Structured low rank approximation*, Linear Algebra and its Applications, 366:157-172.

[34] S.L. Campbell, P. Kunkel, K. Bobinyec (2012) *A minimal norm corrected underdetermined Gauss-Newton procedure*, Appl. Numer. Math. 62:592-605.

[35] E.J. Candès, X. Li, Y. Ma, J. Wright (2011) *Robust principal component analysis?*, Journal of the ACM, 58(3):11.

[36] X. Chen, Z. Nashed, L. Qi (1997) *Convergence of Newton's method for singular smooth and nonsmooth equations using adaptive outer inverses*, SIAM J. Optim. 7:445-462.

[37] B.N. Daskalov (2011) *Iterative methods for matrix factorization with missing data*, Master Thesis, ETH, Zurich. See https://www.research-collection.ethz.ch/handle/20.500.11850/152839.

[38] H.P. Decell (1974) *On the derivative of the generalized inverse of a matrix*, Linear and Multilinear Algebra 1(4):357.

[39] F.R. Deutsch (2012) *Best Approximation in Inner Product Spaces*. Springer.

[40] D.W. Decker, C.T. Kelley (1980) *Newton's method for singular points, I*, Siam J. Numer. Anal., 17:66-70.

[41] D.W. Decker, C.T. Kelley (1980) *Newton's method for singular points, II*, Siam J. Numer. Anal., 17(3):465-471.

[42] J.P. Dedieu, M.H. Kim (2002) *Newton's method for analytic systems of equations with constant rank derivatives*, J. Complexity 18:187-209.

[43] D. Duan, H. Liu (2024) *A fast and efficient randomized quasi-Newton method*, OPT 2024: Optimization for Machine Learning. See https://openreview.net/forum?id=laJUMr2p3l.

[44] M.A. Davenport, J. Romberg (2016) *An Overview of Low-Rank Matrix Recovery From Incomplete Observations*, IEEE Journal of Selected Topics in Signal Processing, 10(4):608-622.

[45] J.E. Dennis, R.B. Schnabel (1983) *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice Hall.

[46] W. Dai, O. Milenkovic, E. Kerman (2011) *Subspace evolution and transfer (SET) for low-rank matrix completion*, IEEE Trans. Signal Process.59(7):3120-3132.

[47] W. Dai, E. Kerman, O. Milenkovic (2012) *A geometric approach to low-rank matrix completion*, IEEE Trans. Inform. Theory 58(1):237-247.

[48] J. Demmel, L. Grigori, M. Hoemmen, J. Langou (2012) *Communication-optimal parallel and sequential QR and LU factorizations*, SIAM J. Sci. Comp., 34:A206-A239.

[49] J.E. Dennis, D.M. Gay, R.E. Welsch (1981) *Algorithm 573 - NL2SOL, An adaptive nonlinear least-squares algorithm*, ACM Transactions on Mathematical Software, 7, pp. 348-368.

[50] A. Dutta, J. Liang, X. Li (2022) *A fast and adaptive svd-free algorithm for general weighted low-rank recovery*, arXiv:2101.00749v2. See https://arxiv.org/abs/2101.00749v2.

[51] A. Edelman, T.A. Arias, S.T. Smith (1998) *The geometry of algorithms with orthogonality constraints*, SIAM J. Matrix Anal. Applicat. 20(2):303-353.

[52] J. Eriksson (1996) *Optimization and Regularization of Nonlinear Least Squares Problems*, Ph.D. Thesis, Umea University, Sweden.

[53] J. Eriksson, P.-A. Wedin (1996) *Regularization Methods for Nonlinear Least Squares. Part 1: Exactly Rank-deficient Problems*, Technical Report UMINF-96.03, Department of Computing Science, Umea University, Umea, Sweden.

[54] J. Eriksson, P.-A. Wedin, M.E. Gulliksson, I. Soderkvist (2005) *Regularization methods for uniformly rank-deficient nonlinear least-squares problems*, J. Optimi. Theory and Appl. 127:1-26.

[55] A. Fletcher (1968) *Generalized inverse methods for the best least-squares solution of systems of non-linear equations*, Comput. J., 10:392-399.

[56] J. Fan, Y. Yuan *On the quadratic convergence of the Levenberg–Marquardt method without non-singularity assumption*. Computing 74(1):23-39.

[57] K. Gabriel (1978) *Least Squares Approximation of Matrices by Additive and Multiplicative Models*, Journal of the Royal Statistical Society, Series B (Methodological), 40(2):186-196.

[58] M. Goldberg (2017) *Continuity of seminorms on finite-dimensional vector spaces*, Linear Algebra and its Applications, 515:175-179.

[59] A. Griewank (1985) *On solving nonlinear equations with simple singularities or nearly singular solutions*, SIAM Review, 27(4):537-563.

[60] M. Guignard (1969) *Generalized Kuhn–Tucker conditions for mathematical programming problems in a Banach space*, SIAM Journal on Control, 7:232-241.

[61] D.M. Gay, L. Kaufman (1991) *Tradeoffs in Algorithms for Separable Nonlinear Least Squares*, IMACS '91, Proceedings of the 13th World Congress on Computational and Applied Mathematics, edited by R. Vichnevetsky and J. J. H. Miller, Criterion Press, Dublin, 1991.

[62] N. Gillis, F. Glineur (2011) *Low-rank matrix approximation with weights or missing data is NP-hard*, SIAM J. Matrix Anal. Appl., 32(4):1149-1165.

[63] G.H. Golub, V. Pereyra (1973) *The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate*, SIAM Journal on Numerical Analysis, 10(2):413-432.

[64] G.H. Golub, V. Pereyra (1976) *Differentiation of pseudo-inverses, separable nonlinear least squares problems and other tales*, Proceedings of an Advanced Seminar Sponsored by the Mathematics Research Center, the University of Wisconsin-Madison, October 8-10, 1973, pages 303-324.

[65] G.H. Golub, V. Pereyra (2003) *Separable nonlinear least squares: the variable projection method and its applications*, Inverse Problems, 19:R1-R26.

[66] P.F. Gotardo, A.M. Martinez (2011) *Computing smooth time trajectories for camera and deformable shape in structure from motion with occlusion*, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 33(10):2051-2065.

[67] A. Griewank, M.R. Osborne (1981) *Newton's method for singular problems when the dimension of the null space is > 1*, SIAM J. Numer. Anal., 18(1):145-149.

[68] L. Grippo, M. Sciandrone (1999) *Globally convergent block-coordinate techniques for unconstrained optimization*, Optim. Methods Soft., 10: 587-637.

[69] L. Grippo, M. Sciandrone (2000) *On the convergence of the block nonlinear Gauss–Seidel method under convex constraints*, Operations research letters, 26(3):127-136.

[70] M.E. Gulliksson, P.A. Wedin (2000) *The Use and Properties of Tikhonov Filter Matrices*, SIAM Journal on Matrix Analysis and Applications 22:276-281.

[71] G.H. Golub, C. Van Loan (1996) *Matrix Computation*, 3rd Edition, The Johns Hopkins University Press, Baltimore, MD.

[72] K. Gabriel, S. Zamir (1979) *Lower rank approximation of matrices by least squares with any choice of weights*, Technometrics, 21:489-498.

[73] W.M. Haussler (1986) *A Kantorovich-type convergence analysis for the Gauss-Newton method*, Numer. Math., 48:119-125.

[74] J. Harris(1992) *Algebraic Geometry*, volume 133 of Graduate Texts in Mathematics. Springer- Verlag New York.

[75] N.D. Ho (2010) *Nonnegative Matrix Factorization - Algorithms and Applications*, Ph.D. thesis, Universite catholique de Louvain, Louvain-la-Neuve, Belgium.

[76] M. Hardt (2014) *Understanding alternating minimization for matrix completion*, in: IEEE 55th Annual Symposium on Foundations of Computer Science, FOCS, 2014, IEEE, Oct 2014, 651-660.

[77] U. Helmke, J.B. Moore (1996) *Optimization and Dynamical Systems*, Communications and Control Engineering Series, Springer-Verlag London Ltd., London.

[78] U. Helmke, M.A. Shayman (1995) *Critical points of matrix least squares distance functions*, Linear Algebra Appl., 215:1-19.

[79] J.B. Hiriart-Urruty, C. Le Marechal (2004) *Fundamentals of convex analysis*, Springer.

[80] J.B. Hiriart-Urruty, H.Y. Le (2013) *A variational approach of the rank function*, TOP 21:207-240.

[81] J.H. Hong, A.W. Fitzgibbon (2015a) *Secrets of Matrix Factorization: Approximations, Numerics, Manifold Optimization and Random Restarts*, in Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), pages 4130-4138.

[82] J.H. Hong, A.W. Fitzgibbon (2015b) *Secrets of Matrix Factorization: Further Derivations and Comparisons*, Supplementary material to [81] in Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV).

[83] W. Ha, H. Liu, R.F. Barber (2020) *An equivalence between critical points for rank constraints versus low-rank factorizations*, SIAM Journal on Optimization, 30(4):2927-2955.

[84] S. Hosseini, D.R. Luke, A. Uschmajew (2019) *Tangent and Normal Cones for Low-Rank Matrices*, in: Hosseini, S., Mordukhovich, B., Uschmajew, A. (eds) Nonsmooth Optimization and Its Applications. International Series of Numerical Mathematics, vol 170. Birkhäuser, Cham. See https://doi.org/10.1007/978-3-030-11370-4_3.

[85] N. Halko, P.-G. Martinsson, J. Tropp (2011) *Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions*, SIAM Review, 53:217-288.

[86] T. Hastie, R. Mazumder, J.D. Lee, R. Zadeh (2015) *Matrix Completion and Low-Rank SVD via Fast Alternating Least Squares*, Journal of Machine Learning Research (JMLR), 16:3367-3402.

[87] P.C. Hansen, V. Pereyra, G. Scherer (2012) *Least Squares Data Fitting with Applications*, The Johns Hopkins University Press, Baltimore.

[88] J.H. Hong, C. Zach, A.W. Fitzgibbon (2017) *Revisiting the Variable Projection Method for Separable Nonlinear Least Squares Problems*, in Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 5939-5947.

[89] J. Hu, X. Liu, Z.W. Wen, Y.X. Yuan(2020) *A Brief Introduction to Manifold Optimization*, in Journal of the Operations Research Society of China, 8:199-248.

[90] I.C.F. Ipsen, C.T. Kelley, S.R. Pope (2011) *Rank-deficient nonlinear least squares problems and subset selection*, SIAM J. Numer. Anal., 49(3):1244-1266.

[91] A. Ilin, T. Raiko (2010) *Practical Approaches to Principal Component Analysis in the Presence of Missing Values*, Journal of Machine Learning Research (JMLR), 11:1957-2000.

[92] I.T. Joliffe (2002) *Principal Component Analysis*, Springer Series in Statistics.

[93] J. Josse, F. Husson, J. Pages (2009) *Gestion des donnees manquantes en Analyse en Composantes Principales*, Journal de la Societe Francaise de Statistique, 150(2):28-51.

[94] P. Jain, P. Netrapalli, S. Sanghavi (2013) *Low-rank matrix completion using alternating minimization*, in Proc. 45th Symposium on Theory of Computing (STOC). ACM, 2013, pp. 665-674.

[95] F.T. Krogh (1974) *Efficient implementation of a variable projection algorithm for nonlinear least squares problems*, Commun. ACM 17(3):167-169.

[96] L. Kaufman (1975) *A variable projection method for solving separable nonlinear least squares problems*, BIT Numer. Math. 15(1):49-57.

[97] L. Kaufman (2010) *Solving separable nonlinear least squares problems with multiple data sets*, Exponential Data Fitting and its Applications, Bentham e-books.

[98] H.A.L. Kiers (1997) *Weighted least squares fitting using iterative ordinary least squares algorithms*, Psychometrika 62:251:266.

[99] H.A.L. Kiers (2002) *Setting up alternating least squares and iterative majorization algorithms for solving various matrix optimization problems*, Computational Statistics and Data Analysis 41:157-170.

[100] R.H. Keshavan, A. Montanari (2010) *Regularization for matrix completion*, in: IEEE International Symposium on Information Theory Proceedings, ISIT, 2010, IEEE, 2010, pp.1503-1507.

[101] R.H. Keshavan, A. Montanari, S. Oh (2010) *Matrix completion from noisy entries*, J.Mach. Learn. Res., 99:2057-2078.

[102] L. Kaufman, G.S. Silvester (1992) *Separable nonlinear least squares with multiple right-hand sides*, SIAM journal on matrix analysis and applications, 13(1):68-89.

[103] L. Kaufman, G.S. Silvester, M.H. Wright (1994) *Structured linear least-squares problems in system identification and separable nonlinear data fitting*, SIAM Journal on Optimization, 4(4):847-871.

[104] Y. Koren, R. Bell, C. Volinsky (2009) *Matrix factorization techniques for recommender systems*, IEEE Comput. 42(8):30-37.

[105] L.A. Lyusternick(1934) *Conditional extrema of functionals*, Matem. Sb., 41(3):390-401.

[106] J.M. Lee (2003) *Introduction to Smooth Manifolds*, Grad. Texts in Math. 218, Springer-Verlag, New York.

[107] K. Levenberg (1944) *A method for the solution of certain non-linear problems in least squares*, Quart. Appl. Math., 2:164-168.

[108] D.G. Luenberger (1973) *Introduction to Linear and Nonlinear Programming*, Addison-Wesley, Reading, MA.

[109] G.G. Lukeman (2009) *Separable overdetermined nonlinear systems: an application of the Shen-Ypma Algorithm*. VDM Verlag Dr. Muller, Saarbrucken.

[110] E. Levin (2020) *Towards optimization on varieties*, Undergraduate senior thesis, Princeton University.

[111] C.L. Lawson, R.J. Hanson (1974) *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs, NJ.

[112] E. Levin, J. Kileel, N. Boumal (2023) *Finding stationary points on bounded-rank matrices: a geometric hurdle and a smooth remedy*. Mathematical Programming, 199:831-864.

[113] E. Levin, J. Kileel, N. Boumal (2025) *The effect of smooth parametrizations on nonconvex optimization landscapes*. Mathematical Programming, 209:63-111.

[114] Y. Li, Y. Liang, A. Risteski (2016) *Recovery guarantee of weighted low rank approximation via alternating minimization*, In International Conference on Machine Learning, pages 2358-2367, 2016.

[115] Y. Luo, X. Li, A.R. Zhang (2024) *Nonconvex Factorization and Manifold Formulations Are Almost Equivalent in Low-Rank Matrix Optimization*, INFORMS Journal on Optimization 0(0). See https://doi.org/10.1287/ijoo.2022.0030.

[116] X. Li, W. Song, N. Xiu (2019) *Optimality conditions for rank-constrained matrix optimization*, Journal of the Operations Research Society of China, 7:285-301. See https://doi.org/10.1007/s40305-019-00245-0.

[117] Q. Li, Z. Zhu, G. Tang (2019) *The non-convex geometry of low-rank matrix optimization*, Information and Inference: A Journal of the IMA, 8:51-96.

[118] Q. Li, Z. Zhu, G. Tang (2019) *Alternating minimizations converge to second-order optimal solutions*, in Proc. Int. Conf. Mach. Learn., 2019, pp. 3935-3943.

[119] H. Li, G.C. Linderman, A. Szlam, K.P. Stanton, Y. Kluger, M. Tygert (2017) *Algorithm 971: An implementation of a randomized algorithm for principal component analysis*, ACM Transactions on Mathematical Software (TOMS), 43:3, Article 28.

[120] D.W. Marquardt(1963) *An algorithm for least-squares estimation of nonlinear parameters*, Journal of the Society for Industrial and Applied Mathematics, 11:431-441.

[121] R. Menzel (1985) *On Solving Nonlinear Least-Squares Problems in case of Rank deficient Jacobians*, Computing, 34:63-72.

[122] J.J. More (1985) *The Levenberg-Marquardt algorithm: Implementation and theory*, in Lecture Notes in Mathematics, No. 630 - Numerical Analysis, G.Watson, ed., Springer-Verlag, pp. 105-116.

[123] K. Madsen, H. B. Nielsen (2010) *Introduction to Optimization and Data Fitting*, Lecture notes, Informatics and Mathematical Modelling, Technical University of Denmark, Lyngby.

[124] J.R. Magnus, H. Neudecker (2019) *Matrix Differential Calculus with Applications in Statistics and Econometrics*, 3rd edition.

[125] J.H. Manton, R. Mahony, Y. Hua (2003) *The geometry of weighted low-rank approximations*, IEEE Transactions on Signal Processing, 51(2):500-514.

[126] G. Marsaglia, G.P.H. Styan (1974) *Equalities and Inequalities for Ranks of Matrices*, Linear and Multilinear Algebra, 2(3):269-292.

[127] I. Markovsky, K. Usevich (2013) *Structured low-rank approximation with missing data*, SIAM J. Matrix Anal. Appl. 34(2):814-830.

[128] R. Murray, J. Demmel, M.W.. Mahoney, N.B. Erichson, M. Melnichenko, O.A. Malik, L. Grigori, P. Luszczek, M. Derezinski, M.E.. Lopes, T. Liang, H. Luo, J. Dongarra (2023) *Randomized Numerical Linear Algebra : A Perspective on the Field With an Eye to Software*, arXiv:2302.11474v2. See https://arxiv.org/abs/2302.11474.

[129] R. Mazumder, T. Hastie, R. Tibshirani (2010) *Spectral regularization algorithms for learning large incomplete matrices*, Journal of Machine Learning Research (JMLR), 11:2287-2322.

[130] B. Mishra, K.A. Apuroop, R. Sepulchre (2012) *A Riemannian geometry for low-rank matrix completion*, arXiv preprint arXiv:1211.1550.

[131] B. Mishra, G. Meyer, F. Bach, R. Sepulchre (2013) *Low-rank optimization with trace norm penalty*, SIAM J. Optim. 23(4):2124-2149.

[132] B. Mishra, G. Meyer, S. Bonnabel, R. Sepulchre (2014) *Fixed-rank matrix factorizations and Riemannian low-rank optimization*, Comput Stat 29:591-621.

[133] B. Mishra, R. Sepulchre (2016) *Riemannian preconditioning*, SIAM J. Optim. 26(1):635-660.

[134] K. Madsen, H.B. Nielsen, O. Tingleff (2004) *Methods for non-linear least-squares problems*, 2nd Edition, IMM, DTU, Lecture note IMM3215, 60 pp.

[135] C. Musco, C. Musco, D.P. Woodruff (2021) *Simple heuristics yield provable algorithms for masked low-rank approximation*, in 12th Innovations in Theoretical Computer Science Conference (ITCS 2021).

[136] H.B. Nielsen (2000) *Separable Nonlinear Least Squares*, IMM, DTU, Report MM-REP-2000-01, 2000.

[137] M.Z. Nashed, X. Chen (1993) *Convergence of Newton-like methods for singular operator equations using outer inverses*, Numerische Mathematik, 66:235-257.

[138] P. Netrapalli, U.N.. Niranjan, S. Sanghavi, A. Anandkumar, P. Jain (2014) *Non-convex robust PCA*, in Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada, pages 1107–1115. See http://papers.nips.cc/paper/5430-non-convex-robust-pca.

[139] J. Nocedal, S.J. Wright (2006) *Numerical Optimization*, second edition, Springer-Verlag, Berlin.

[140] L.T. Nguyen, J. Kim, B. Shim (2019) *Low-rank matrix completion: A contemporary survey*, IEEE Access, 7:94215-94237.

[141] M.R. Osborne (1976) *Nonlinear least squares - the Levenberg algorithm revisited*, J. Austral. Math. Soc. (Series B), 19:343-357.

[142] G. Olikier, P.A. Absil (2022) *On the Continuity of the Tangent Cone to the Determinantal Variety*, Set-Valued Var. Anal 30:769-788.

[143] G. Olikier, P.A. Absil (2024) *Computing Bouligand stationary points efficiently in low-rank optimization*, arXiv:2409.12298v1. See https://doi.org/10.48550/arXiv.2409.12298.

[144] G. Olikier, I. Waldspurger (2024) *Projected gradient descent accumulates at Bouligand stationary points*, arXiv:2403.02530v2. See https://doi.org/10.48550/arXiv.2403.02530.

[145] G. Olikier, K.A. Gallivan, P.A. Absil (2024) *Low-rank optimization methods based on projected-projected gradient descent that accumulate at Bouligand stationary points*, arXiv:2201.03962v2. See https://arxiv.org/abs/2201.03962v2.

[146] G. Olikier, A. Uschmajew, B. Vandereycken (2023) *Gauss–Southwell type descent methods for low-rank matrix optimization*, arXiv:2306.00897v2. See https://doi.org/10.48550/arXiv.2306.00897.

[147] T. Okatani, K. Deguchi (2007) *On the Wiberg algorithm for matrix factorization in the presence of missing components*, International Journal of Computer Vision (IJCV), 72(3):329-337.

[148] J.M. Ortega, W.C. Rheinboldt (1970) *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, San Diego.

[149] D.P. O Leary, B.W. Rust (2013) *Variable projection for nonlinear least squares problems*, Computational Optimization and Applications, 54(3):579-593.

[150] T. Okatani, T. Yoshida, K. Deguchi (2011) *Efficient algorithm for low-rank matrix factorization with missing components and performance comparison of latest algorithms*, In Proceedings of the 2011 IEEE International Conference on Computer Vision (ICCV), 842-849.

[151] E. Pauwels (2013) *Generic Frechet stationarity in constrained optimization*, arXiv:2402.09831v2. See https://arxiv.org/abs/2402.09831v2.

[152] M.J.D. Powell (1973) *On search directions for minimization algorithms*, Math. Programming, 4:193-201.

[153] F. Pes, G. Rodriguez (2020) *The minimal-norm Gauss-Newton method and some of its regularized variants*, Electron. Trans. Numer. Anal. 53:459-480.

[154] F. Pes, G. Rodriguez (2022) *A doubly relaxed minimal-norm Gauss-Newton method for underdetermined nonlinear least-squares problems*, Applied Numerical Mathematics 171:233-248.

[155] P. Paatero, U. Tapper (1994) *Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values*, Environmetrics, 5:111-126.

[156] D. Park, A. Kyrillidis, C. Carmanis, S. Sanghavi (2017) *Non-square matrix sensing without spurious local minima via the Burer-Monteiro approach*, in Proc. Artif. Intell. Statist., 2017, pp. 65-74.

[157] J. Rennie, N. Srebro (2005) *Fast maximum margin matrix factorization for collaborative prediction*, in Proceedings of the 22nd International Conference on Machine Learning, pages 713-719. ACM, 2005.

[158] A. Ruhe (1974) *Numerical computation of principal components when several observations are missing*, Technical report, UMINF-48, Umea, Sweden.

[159] G.W. Reddien (1980) *Newton's method and high order singularities*, Comput. Math. Appl., 5:79-86.

[160] A. Ruszczynski (2006) *Nonlinear Optimization*. Princeton University Press.

[161] P.J. Rabier, G.W. Reddien (1986) *Characterization and computation of singular points with maximum rank deficiency*, SIAM J. Numer. Anal. 23:1040-1051.

[162] Q. Rebjock, N. Boumal (2024) *Fast convergence of trust-regions for non-isolated minima via analysis of CG on indefinite matrices*. Math. Program., online.

[163] Q. Rebjock, N. Boumal (2024) *Fast convergence to non-isolated minima: four equivalent conditions for C^2 functions*. Math. Program., online.

[164] J.W. Robbin, D.A. Salomon (2022) *Introduction to differential geometry*, Springer Spektrum Berlin, Heidelberg, 418 pp.

[165] R.T. Rockafellar, R.J.B. Wets (1998) *Variational Analysis*, vol. 317 of Grundlehren der mathematischen Wissenschaften, Springer-Verlag Berlin Heidelberg. Corrected 3rd printing 2009.

[166] A. Ruhe, P.-A. Wedin (1980) *Algorithms for Separable Nonlinear Least Squares Problems*, SIAM Review, 22(3):318-337.

[167] I. Razenshteyn, Z. Song, D.P. Woodruff (2016) *Weighted Low Rank Approximations with Provable Guarantees*, STOC'16, June 19-21, 2016, Cambridge, MA, USA.

[168] M.F. Sukhinin (1973) *Conditional extrema of functionals in topological linear spaces*, Mathematical Notes of the Academy of Sciences of the USSR 14:775-779.

[169] L. Simonsson, L. Elden (2010) *Grassmann algorithms for low rank approximation of matrices with missing values*, BIT Numer. Math., 50:173-191.

[170] R. Sun, Z.Q. Luo (2016) *Guaranteed Matrix Completion via Non-Convex Factorization*, in IEEE Transactions on Information Theory, vol. 62, no. 11, pp. 6535-6579, Nov. 2016.

[171] N. Srebro, T. Jaakkola (2004) *Weighted low-rank approximations*, in Proceedings of the 20th International Conference on Machine Learning, pp. 720-727.

[172] G.W. Stewart, J. Sun (1990) *Matrix perturbation Theory*, Academic Press, INC, New York.

[173] S. Schneider, A. Uschmajew (2015) *Convergence results for projected line-search methods on varieties of low-rank matrices via Lojasiewicz inequality*, SIAM Journal on Optimization, 25:622-646.

[174] T. Schramm, B. Weitz (2015) *Low-rank matrix completion with adversarial missing entries*, arXiv:1506.03137v1. See https://arxiv.org/abs/1506.03137v1.

[175] Y. Shen, T.J. Ypma (2019) *Solving separable nonlinear least squares problems using the QR factorization*. J. Comput. Appl. Math. 345:48-58.

[176] H. Shum, K. Ikeuchi, R. Reddy (1995) *Principal component analysis with missing data and its application to polyhedral object modeling*, IEEE Transaction on Pattern Analysis and Machine Intelligence, 17(9):855-867.

[177] N. Srebro, J. Rennie, T. Jaakkola (2005) *Maximum-margin matrix factorization*, in Advances in Neural Information Processing Systems 17 (NIPS 2004) pp. 1329-1336. MIT Press.

[178] A. Szlam, A. Tulloch, M. Tygert (2017) *Accurate low-rank approximations via a few iterations of alternating least squares*, SIAM Journal on Matrix Analysis and Applications, 38(2):425-433.

[179] P. Terray (1995) *Space/Time structure of monsoons interannual variability*, Journal of Climate, 8:2595-2619.

[180] P. Terray (2002) *Application of Weighted Empirical Orthogonal Function Analysis to ship's datasets*, Compte-Rendu de la IVème journée Statistique IPSL (Classification et Analyse spatiale), NAI no 23. pp. 11-28. ISSN 1626-8334.

[181] E. Tuzhilina, T. Hastie (2021) *Weighted Low Rank Matrix Approximation and Acceleration*, arXiv:2109.11057v1. See https://arxiv.org/abs/2109.11057v1.

[182] K. Usevich, I. Markovsky (2014) *Variable projection methods for affinely structured low-rank approximation in weighted 2-norms*, Journal of Computational and Applied Mathematics, 272:430-448.

[183] A. Uschmajew, R. Vandereycken (2014) *Line-search methods and rank increase on low-rank matrix varieties*, in: 2014 International Symposium on Nonlinear Theory and its Applications (NOLTA2014), Luzern, Switzerland, September 14-18, 2014, 52-55 (IEICE: Luzern, 2014)

[184] A. Uschmajew, R. Vandereycken (2015) *Greedy rank updates combined with Riemannian descent methods for low-rank optimization*, in: 2015 International Conference on Sampling Theory and Applications (SampTA), 25-29 May 2015, Washington, DC, USA, ed. by Stephen Casey, Kevin Duke, Michael Robinson, 420-424 (IEEE: Piscataway, NJ, 2015)

[185] R. Vandereycken (2013) *Low-rank matrix completion by Riemannian optimization-extended version*, arXiv:1209.3834v1. See https://arxiv.org/abs/1209.3834v1.

[186] R. Vandereycken (2013) *Low-rank matrix completion by Riemannian optimization*, SIAM J. Optim. 23(2):1214-1236.

[187] R. Vidal, Y. Ma, S.S. Sastry (2016) *Generalized Principal Component Analysis*, New York, NY: Springer New York, DOI: 10.1007/978-0-387-87811-9. ISSN: 21969973: 09396047.

[188] Z. Wen, W. Yin, Y. Zhang (2012) *Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm*, Math. Program. Comput. 4(4):333-361.

[189] P.A. Wedin (1973) *Perturbation Theory for Pseudo-Inverses*, BIT, 13:217-232.

[190] T. Wiberg (1976) *Computation of principal components when data are missing*, in Proceedings of the 2nd Symposium of Computational Statistics, pages 229-326, 1976.

[191] H. Wold (1966) *Nonlinear estimation by iterative least squares procedures*, Research Papers in Statistics (F. N. David Ed.), pp. 411-444, New York: Wiley.

[192] H. Wold, E. Lyttkens (1969) *Nonlinear Iterative Partial Least Squares (NIPALS) Estimation Procedures*, Bulletin of the International Statistical Institute, 43:29-51.

[193] L. Wang, X. Zhang, Q. Gu (2017) *A unified computational and statistical framework for nonconvex low-rank matrix estimation*, in Proc. Artif. Intell. Statist., 2017, pp. 981-990.

[194] Y. Xu, W. Yin (2013) *A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion*, SIAM Journal on imaging sciences, 6(3):1758-1789.

[195] Y. Yamamoto (1989) *Uniqueness of the Solution in a Kantorovich-Type Theorem of Hausler for the Gauss-Newton Method*, Japan J. Appl. Math., 6:77-81.

[196] N. Yamashita, M. Fukushima (2001) *On the rate of convergence of the Levenberg–Marquardt method*. In: Topics in Numerical Analysis, pp 239-249. Springer.

[197] W.H. Yang, L.H. Zhang, R. Song (2014) *Optimality conditions for the nonlinear programming problems on riemannian manifolds*, Pacific Journal of Optimization, 10(2):415-434.

[198] G. Zhou (2015) *Rank-contrained optimization: A Riemannian manifold approach*, Ph.D. thesis, Florida State University.

[199] Z. Zeng (2024) *A Newton's Iteration Converges Quadratically to Non-isolated Solutions Too*, arXiv:2101.09180v4. See https://arxiv.org/abs/2101.09180v4.

[200] Z. Zhu, Q. Li, G. Tang, M.B. Wakin (2018) *Global optimality in low-rank matrix optimization*, IEEE Transactions on Signal Processing, 66(13):3614-3628.

[201] Z. Zhu, Q. Li, G. Tang, M.B. Wakin (2021) *The Global Optimization Geometry of low-rank matrix optimization*, IEEE Transactions on Information Theory, 67(2):1308-1331.