Majorization-Minimization on the Stiefel Manifold with Application to Robust Sparse PCA

Arnaud Breloy, *Member, IEEE*, Sandeep Kumar, Ying Sun, and Daniel P. Palomar, *Fellow, IEEE*

***Abstract*—This paper proposes a framework for optimizing cost functions of orthonormal basis learning problems, such as principal component analysis (PCA), subspace recovery, orthogonal dictionary learning, etc. The optimization algorithm is derived using the majorization-minimization framework in conjunction with orthogonal projection reformulations to deal with the orthonormality constraint in a systematic manner. In this scope, we derive surrogate functions for various standard objectives that can then be used as building blocks, with examples for robust learning costs and sparsity enforcing penalties. To illustrate this point, we propose a new set of algorithms for sparse PCA driven by this methodology, whose objective function is composed of an M-estimation type subspace fitting term plus a regularizer that promotes sparsity. Simulations and experiments on real data illustrate the interest of the proposed approach, both in terms of performance and computational complexity.**

***Index Terms*—Majorization-Minimization, Stiefel manifold, Grassmann manifold, Subspace learning, PCA, sparse PCA.**

I. Introduction

OPTIMIZATION problems on the Stiefel manifold, i.e., involving the orthonormality constraint on the matrix variable, are ubiquitous in signal processing and machine learning. To name a few, such problems arise in principal component analysis (PCA) [1, 2], sparse PCA [3–8], structured covariance matrix estimation [9, 10], robust subspace recovery [11–16], and subspace clustering [17]. To tackle the challenges from big data and high-dimensional settings the solutions are often desired with certain additional properties, e.g., sparsity. These additional properties are often enforced by including regularization penalties to the objective function. Except for some special cases where the solution appears as eigenvectors of a given matrix, these problems are usually nontrivial to deal with, owing to the nonconvex orthonormality constraints. The resolution of these problems thus generally calls for the use of iterative constrained optimization methods to search for their local minima. A very general framework to account for the orthonormality constraint is brought by Riemannian optimization algorithms on the Stiefel manifold [18–23]. However, the algorithms derived within this perspective can turn to be computationally costly. Additionally, ensuring sparsity and orthonormality simultaneously is a challenging goal, which generally calls for a trade-off between these two properties in the solutions obtained by existing state-of-the-art approaches (cf. [6] and reference therein). A recent line of work based on orthogonal Procrustes reformulations, however, has enabled researchers to ensure both of these properties. These works addressed specific applications, e.g., [7] for the sparse PCA and covariance estimation and [8] for low-rank matrix estimation problems.

Motivated by the applicability of these reformulations for tackling the aforementioned issues, this work aims to develop a general low-complexity algorithmic framework for large-scale sparsity-regularized optimization problems on the Stiefel Manifold. This framework will be based on the majorization-minimization (MM) algorithm [24], which proceeds with two steps: i) (majorization) finding a function that locally upperbounds the objective function up to a constant, referred to as surrogate function; ii) (minimization) minimizing this surrogate function. This procedure generates a sequence that monotonically decreases the objective value, and its main interest is that, with properly chosen surrogate functions, it can yield a sequence of subproblems that are easy to deal with. The main idea will thus be to deal with the orthonormality constraint through a systematic formulation of the surrogates minimization subproblems as a Euclidean projection onto the Stiefel manifold [23, 25, 26], as done for specific cases in, e.g., [7, 27–29]. Thereby, we propose a unified approach with convergence guarantee and a set of practical guidelines for applying this methodology that generalizes to a large class of cost functions. Notably, an emphasis is put on robust data fitting cost functions and sparsity enforcing penalties. The advantages of the proposed methodology are multiple: i) it can be applied to a large class of standard cost functions; ii) it guarantees that the iterates are orthonormal and has standard MM convergence guarantees, i.e., monotonic decrement of the objective value, and convergence to the set of critical points of the problem [30]; iii) it yields scalable algorithms, as the computational bottleneck of each iteration lies only in the computation of a thin-SVD.

As a main example to illustrate this framework, we propose a new class of algorithms for sparse PCA, referred to as RSPCA (for *robust sparse PCA*). The familly of objective functions proposed for this task combines a M-estimation-type subspace fitting function [12–15], plus a sparsity promoting penalty. The considered penalties leverage the proxies of the ℓ_0-norm proposed in [31], and allow for the optimization to be conducted with the proposed methodology. Interestingly, the resulting sparse PCA algorithms do not involve a relaxation of the orthonormality constraint, which is usually required in the existing state-of-the-art methods [3–6]. Finally, simulation

Arnaud Breloy is with LEME (EA4416), University Paris Nanterrre, France (e-mail: abreloy@parisnanterre.fr). Sandeep Kumar is with Department of Electrical Enginerring, Indian Institute of Technology Delhi, India (email: sandeep0kr@ee.iitd.ac.in). Ying Sun is with the Pennsylvania State University, USA (email: ysun@psu.edu). Daniel P. Palomar is with the Hong Kong University of Science and Technology (HKUST), Hong Kong (e-mail: palomar@ust.hk). This work was supported by the Hong Kong GRF 16207019 research grant.

results illustrate the usefulness of the proposed approaches, both in terms of performance and computational complexity.

The presentation is organized as follows. Section II presents relevant background and the motivations for this study. Section III presents a generic MM framework for optimization over the Stiefel manifold, where a systematic linear majorization is applied in order to derive simple iterations. Such majorization can be obtained for various standard cost functions, for which a catalog is presented in Section IV. Section V presents several examples of algorithms derived following this framework, and exhibits some links between them and existing state-of-the-art algorithms. Section VI illustrates the application of this framework to sparse-PCA for a flexible class of sparsity promoting penalties. This adds to the list of functions covered by Section IV while discussing methodological insights for applying the framework to non-standard costs functions. Finally, Section VII presents some simulations and experiments on real data to illustrate the interest of the proposed approach and algorithms, both in terms of performance and computational complexity.

The following notation is adopted: italic, lower case boldface, and upper case boldface indicate respectively, scalar, vector, and matrix quantities. The upperscript ^{H} denotes the transpose conjugate operator. $\mathbf{I}$ is the identity matrix of the appropriate dimension. Tr$\{\}$ is the Trace operator. $\{w_i\}_{i=1}^n$ denotes the set of elements w_i, $\forall i \in [\![1, N]\!]$. diag$(\{a_i\}_{i=1}^n)$ is the $n \times n$ diagonal matrix with diagonal entries a_n. For tall matrices $\mathbf{M} \in \mathbb{C}^{p\times k}$ $(p > k)$ the thin-SVD (TSVD) is denoted $\mathbf{M} \overset{\text{TSVD}}{=} \mathbf{U}_k\mathbf{D}_k\mathbf{V}^H$, where only the k first column vectors of $\mathbf{U}$ from the SVD are used. The set of $p \times p$ Hermitian positive semi-definite matrices is denoted $\mathcal{H}_p^+$. The set of $p\times p$ diagonal matrices is denoted $\mathcal{D}_p$.

II. Background and positioning

A. Majorization-Minimization (MM) algorithms

The MM framework is briefly reviewed below. For more complete information, we refer the reader to [24]. Consider the following optimization problem:

$$\underset{\mathbf{x}\in\mathcal{X}}{\text{minimize}}\ f(\mathbf{x}), \tag{1}$$

where $f : \mathcal{X} \to \mathbb{R}$ is a continuous function and $\mathcal{X}$ is a closed set. Given an initial point $\mathbf{x}^0 \in \mathcal{X}$, the MM procedure minimizes f over $\mathcal{X}$ by updating $\mathbf{x}$ iteratively as

$$\mathbf{x}^{t+1} \in \underset{\mathbf{x}\in\mathcal{X}}{\text{argmin}}\ g(\mathbf{x}|\mathbf{x}^t), \tag{2}$$

where $g(\cdot|\mathbf{x}^t) : \mathcal{X} \to \mathbb{R}$ is a *surrogate function* of f satisfying the following property:

$$\mathbf{x}^t \in \underset{\mathbf{x}\in\mathcal{X}}{\text{argmin}}\ g(\mathbf{x}|\mathbf{x}^t) - f(\mathbf{x}). \tag{3}$$

In other words, $g(\cdot|\mathbf{x}^t)$ upperbounds f globally over set $\mathcal{X}$ up to a constant:

$$g(\mathbf{x}|\mathbf{x}^t) - f(\mathbf{x}) \geq c^t \triangleq \{g(\mathbf{x}^t|\mathbf{x}^t) - f(\mathbf{x}^t)\}, \quad \forall \mathbf{x} \in \mathcal{X}. \tag{4}$$

The sequence $\{f(\mathbf{x}^t)\}_{t\in\mathbb{N}}$ generated by (2) is non-increasing since

$$f(\mathbf{x}^{t+1}) \overset{(4)}{\leq} g(\mathbf{x}^{t+1}|\mathbf{x}^t) - c^t \overset{(2)}{\leq} g(\mathbf{x}^t|\mathbf{x}^t) - c^t = f(\mathbf{x}^t). \tag{5}$$

The MM procedure suggests thus the possibility of minimizing f by iteratively seeking for a sequence of surrogate functions $\{g(\cdot|\mathbf{U}^t)\}_{t\in\mathbb{N}}$ that are easy to minimize over the feasible set.

B. Stiefel manifold and orthogonal projection problem

First, define the complex Stiefel manifold as:

$$\text{St}(p, k) = \left\{\mathbf{U} \in \mathbb{C}^{p\times k} \mid \mathbf{U}^H\mathbf{U} = \mathbf{I}\right\} \text{ for } k \leq p. \tag{6}$$

A point $\mathbf{U} \in \text{St}(p, k)$ is a semi-unitary[1] matrix, referred to as k-orthogonal frame. This point $\mathbf{U}$ is also an orthonormal basis that spans a k-dimensional subspace. However, notice that this representation of a subspace is not unique as $\mathbf{UQ}$ spans the same subspace for any $\mathbf{Q} \in \text{St}(k, k)$.

Let $\mathbf{X} \in \mathbb{C}^{p\times k}$ be a rank p matrix, the Euclidean projection of $\mathbf{X}$ onto the Stiefel manifold is the solution of the following problem

$$\underset{\mathbf{U}\in\text{St}(p,k)}{\text{minimize}}\ ||\mathbf{X} - \mathbf{U}||_F^2. \tag{7}$$

When $\mathbf{X}$ is full rank the problem (7) has a unique global minimizer, given by [23, Prop. 7], as

$$\mathbf{U}^\star = \mathcal{P}_{\text{St}}\left\{\mathbf{X}\right\}, \tag{8}$$

where the operator $\mathcal{P}_{\text{St}}$ is defined in Algorithm 1.

C. Optimization on $\text{St}(p, k)$: some existing solutions

Consider a generic optimization problem where the variable is constrained to the Stiefel manifold:

$$\underset{\mathbf{U}\in\text{St}(p,k)}{\text{minimize}}\ f\left(\mathbf{U}\right), \tag{9}$$

where $f : \mathbb{C}^{p\times k} \to \mathbb{R}$ is a differentiable objective function suited to an application of interest. Notice that optimization problems over the Stiefel manifold $\text{St}(p, k)$ are nonconvex due to the orthonormality constraint. Hence, they are usually hard to deal with, even for apparently simple objective functions f.

A natural way to handle the orthonormality constraint is to turn to the framework of Riemannian optimization. A review of this general topic can be found in [18], where $\text{St}(p, k)$ is used as example throughout the book. More focused reviews of manifold-oriented counterparts of classical optimization algorithms on $\text{St}(p, k)$ (steepest descent, conjugate gradient, Newton's method, etc.) can be found in [20, 23] ([21, 22] for $\text{St}(p, p)$). Notice that when the function is rotation invariant, i.e., $f(\mathbf{U}) = f(\mathbf{UQ}),\ \forall \mathbf{Q} \in \text{St}(k, k)$, it can be minimized on the Grassmann manifold $\text{Gr}(p, k)$ (the set of p-dimensional subspaces of $\mathbb{C}^p$), with corresponding Riemannian optimization algorithm [18, 20, 23], which has the advantage of reducing the dimension of the problem by exploiting its invariance. In this case, one can still apply optimization algorithms developed for $\text{St}(p, k)$, but the variable $\mathbf{U} \in \text{St}(p, k)$ simply acts as a representation of its equivalence class (a point in $\text{Gr}(p, k)$).

A second—more specific—option, can be to split f as $f = f_\text{u} + f_\text{v}$ and rewrite (9) as

$$\begin{array}{ll}\underset{\mathbf{V},\mathbf{U}\in\text{St}(p,k)}{\text{minimize}} & f_\text{u}\left(\mathbf{U}\right) + f_\text{v}\left(\mathbf{V}\right) \\ \text{subject to} & \mathbf{U} = \mathbf{V},\end{array} \tag{10}$$

[1] Or semi-orthogonal in the real valued case.

then relax the equality constraint of this problem using the augmented Lagrangian method. The aim of this reformulation, e.g., used in [8], is mainly to derive a block-coordinate algorithm where the updates with respect to $\mathbf{U}$ and $\mathbf{V}$ are separately easy to obtain.

D. Aim of this paper

In this paper, we explore the use of the MM framework to solve problems on $\mathrm{St}(p,k)$ formulated as in (9). Generally speaking, a majorization can be applied to cast subproblems that are more practical for the aforementioned algorithms to be applied. For example, the subproblems could have easy Newton steps, or a simpler update of the block $\mathbf{U}$ in problem (10). Here, our main focus will be to derive surrogates functions that formulate minimization subproblems with closed-form solutions. Interestingly this systematic approach yields low-complexity algorithms for solving problem as in (9), which can also be used as subroutine for more complex problems where the formulation (10) is involved.

III. MM algorithms on the Stiefel manifold with systematic projection reformulations

A. Generic algorithm formulation

To tackle the optimization of the cost function in (9) under the orthonormality constraint, we consider applying the MM [24] framework and minimizing f by solving a sequence of Euclidean projections on $\mathrm{St}(p,k)$. In short, we will construct surrogate functions that are linear when restricted to the feasible set $\mathrm{St}(p,k)$. The corresponding subproblem can then be recast as a Euclidean projection onto $\mathrm{St}(p,k)$ (cf. section II-B), which has a unique solution that can be obtained via thin-SVD. In the following, we assume that the objective in (9) is majorized at point $\mathbf{U}^t$ by a surrogate $g(\mathbf{U}|\mathbf{U}^t)$ that satisfies the following properties (examples are given later).

Assumption 1. The surrogate function $g : \mathbb{C}^{p\times k} \times \mathbb{C}^{p\times k} \to \mathbb{R}$ satisfies the following conditions:

i) Tightness: $g(\mathbf{U}^t|\mathbf{U}^t) = f(\mathbf{U}^t)$,

ii) Continuity: $g(\cdot\,|\,\cdot)$ is continuous on $\mathbb{C}^{p\times k} \times \mathbb{C}^{p\times k}$,

iii) Upperbound: $g(\mathbf{U}|\mathbf{U}^t) \geq f(\mathbf{U}),\ \forall \mathbf{U} \in \mathrm{St}(p,k)$,

iv) Linearity: restricting to $\mathrm{St}(p,k)$, g can be expressed as

$$\begin{aligned} g\left(\mathbf{U}|\mathbf{U}^t\right) &= -\operatorname{Tr}\left\{\mathbf{R}^H\left(\mathbf{U}^t\right)\mathbf{U}\right\} \\ &\quad -\operatorname{Tr}\left\{\mathbf{U}^H\mathbf{R}\left(\mathbf{U}^t\right)\right\} + \text{const.}, \\ &= -2\mathfrak{Re}\left\{\operatorname{Tr}\{\mathbf{U}^H\mathbf{R}\left(\mathbf{U}^t\right)\}\right\} + \text{const.}, \end{aligned} \tag{11}$$

where $\mathbf{R} : \mathbb{C}^{p\times k} \to \mathbb{C}^{p\times k}$ is a matrix function of $\mathbf{U}^t$.

Following the MM procedure described in Section II-A, an update of the parameter $\mathbf{U}$ is given by

$$\mathbf{U}^{t+1} \in \underset{\mathbf{U}\in\mathrm{St}(p,k)}{\operatorname{argmin}}\ g\left(\mathbf{U}|\mathbf{U}^t\right). \tag{12}$$

Since g is linear (cf. (18)) and $\mathbf{U} \in \mathrm{St}(p,k)$, it is not hard to see that obtaining this update is equivalent to solving

$$\underset{\mathbf{U}\in\mathrm{St}(p,k)}{\text{minimize}}\quad ||\mathbf{R}\left(\mathbf{U}^t\right) - \mathbf{U}||_F^2, \tag{13}$$

Algorithm 1 Computation of $\mathcal{P}_{\mathrm{St}}$ (projection on $\mathrm{St}(p,k)$)

1: **Entry:** $\mathbf{R} \in \mathbb{C}^{p\times k}$
2: Compute the thin-SVD: $\mathbf{R} \overset{\text{TSVD}}{=} \mathbf{V}_{\text{left}}\mathbf{D}\mathbf{V}_{\text{right}}^H$
3: Set $\mathbf{U}{=}\mathbf{V}_{\text{left}}\mathbf{V}_{\text{right}}^H$
4: **Output:** $\mathbf{U} = \mathcal{P}_{\mathrm{St}}(\mathbf{R}) \in \mathrm{St}(p,k)$

Algorithm 2 Generic MM Algorithm for $\mathrm{St}(p,k)$

1: **initialize** $t = 0$, $\mathbf{U}^{(0)} \in \mathbb{C}^{N\times R}$
2: **repeat**
3: Compute $\mathbf{R}\left(\mathbf{U}^t\right)$ from surrogate (18)
4: Update $\mathbf{U}^{t+1} = \mathcal{P}_{\mathrm{St}}\left\{\mathbf{R}\left(\mathbf{U}^t\right)\right\}$ with Algorithm 1
5: $t = t + 1$
6: **until** convergence criterion is met

which is the projection problem discussed in Section II-B. When $\mathbf{R}\left(\mathbf{U}^t\right)$ is full rank[2], the problem (13) admits a unique solution, leading to the MM update:

$$\mathbf{U}^{t+1} = \mathcal{P}_{\mathrm{St}}\left\{\mathbf{R}\left(\mathbf{U}^t\right)\right\}, \tag{14}$$

where the operator $\mathcal{P}_{\mathrm{St}}$ is defined in Algorithm 1. Eventually, solving the sequence of projection problems results in an MM procedure to optimize f under the orthonormality constraint, which is summarized in Algorithm 2.

Remark 1. Importantly, the MM approach is also applicable to objective function consisting sum of functions of the form

$$f(\mathbf{U}) = \sum_{i=1}^{I} f_i(\mathbf{U}). \tag{15}$$

Then, if each f_i can be majorized by a linear surrogate g_i of the form

$$\begin{aligned} g_i\left(\mathbf{U}|\mathbf{U}^t\right) &= -\operatorname{Tr}\left\{\mathbf{R}_i^H\left(\mathbf{U}^t\right)\mathbf{U}\right\} \\ &\quad -\operatorname{Tr}\left\{\mathbf{U}^H\mathbf{R}_i\left(\mathbf{U}^t\right)\right\} + \text{const.}, \end{aligned} \tag{16}$$

following the same steps as (12)-(14), the MM updates can simply be obtained as

$$\mathbf{U}^{t+1} = \mathcal{P}_{\mathrm{St}}\left\{\sum_{i=1}^{I}\mathbf{R}_i\left(\mathbf{U}^t\right)\right\}. \tag{17}$$

Obviously, this methodology cannot be applied to any arbitrary cost function. Still, Section IV presents a catalog of surrogate functions satisfying (18) for a large set of standard cost functions that can be used as building blocks (with several examples detailed in Section V). The method also suggests that simple algorithms can be obtained by designing meaningful proxies of the desired function that can be majorized by a linear surrogate on $\mathrm{St}(p,k)$. As a concrete example, we propose sparsity promoting penalties that satisfy this property in Section VI, leading to an original algorithm for sparse PCA.

[2]Rank deficiency of this matrix is a case we do not focus on: some pathological counter-examples can be build but they rely on either *i*) a cost function f that does not satisfy the initial regularity assumptions; *ii*) a subspace within $\mathbf{U}^t$ that has reached a local stationary point. For the second point, the proposed method can still be applied by setting the stable subspace fixed and updating only the remaining portion of $\mathbf{U}^t$ (i.e., recasting the problem with $k' < k$). In practice, the issue has not been experienced with the cost functions (and starting points) considered in this paper.

Remark 2. Quadratic surrogate functions, if existing, can also be used within the considered framework. These surrogates should satisfy Assumption 1 with the following alternate condition:

iv') Restricting to $\mathrm{St}(p,k)$, g can be expressed as

$$g\left(\mathbf{U}|\mathbf{U}^t\right) = -\operatorname{Tr}\left\{\mathbf{U}^H\mathbf{M}\left(\mathbf{U}^t\right)\mathbf{U}\mathbf{D}\left(\mathbf{U}^t\right)\right\} + \text{const.}, \tag{18}$$

where $\mathbf{M}:\mathbb{C}^{p\times k}\to\mathcal{H}_p^+$ and $\mathbf{D}:\mathbb{C}^{p\times k}\to\mathcal{D}_k$ are matrix functions of $\mathbf{U}^t$.

Indeed, such surrogates also yield tractable minimization subproblems under the orthonormality constraint: the corresponding updates are then obtained from the k eigenvectors associated to the largest eigenvalues of $\mathbf{M}\left(\mathbf{U}^t\right)$ if $\mathbf{D}\left(\mathbf{U}^t\right)\succcurlyeq\mathbf{0}$ (smallest conversely). In this paper, we focused on linear surrogates since quadratic functions always admit linear surrogates on $\mathrm{St}(p,k)$ (cf. propositions 2 and 3). Thus, functions with quadratic majorizers can also be used in conjunction with function that only have linear ones within the proposed framework, while the converse is not true. Nevertheless, quadratic surrogates can in some cases provide tighter approximations that yield a faster convergence of the MM algorithm (at the cost of handling larger matrices at each iterations). A main example discussing the two options is detailed in section V-C and VII-A.

B. Convergence analysis

The convergence analysis of Algorithm 2 will be obtained by following the one of the successive upper-bound minimization (SUM) algorithm in [30]. Note that the result of [30] does not hold directly for Algorithm 2, as the SUM framework does not cover non-convex constraints. Nevertheless, this result can be adapted to $\mathrm{St}(p,k)$ as in [7, 32, 33], leading to the following proposition.

Proposition 1. *Let $\{\mathbf{U}^t\}_{t=0}^{\infty}$ be a sequence generated by Algorithm 2. Then the following hold:*

1) *The sequence $\{f(\mathbf{U}^t)\}_{t\in\mathbb{N}}$ converges.*
2) *Every limit point $\mathbf{U}^*$ of the sequence is a critical (also referred to as Karush-Kuhn-Tucker, or KKT) point of the problem* (9).
3) *The whole sequence converges to $\mathcal{K}$, the set of KKT points of the problem* (9).

Proof. First, remark that Algorithm 2 follows the MM procedure, by (5) we have that the sequence $\{f(\mathbf{U}^t)\}_{t\in\mathbb{N}}$ is non-increasing, and thus converges. The compactness of the set $\mathrm{St}(p,k)$ then implies that the sequence $\{\mathbf{U}^t\}_{t\in\mathbb{N}}$ is bounded. Thus it admits at least one limit point $\mathbf{U}^*$. Following the same argument as [30, Theorem 1] we have

$$g\left(\mathbf{U}^*|\mathbf{U}^*\right)\leq g\left(\mathbf{U}|\mathbf{U}^*\right),\quad\forall\mathbf{U}\in\mathrm{St}(p,k), \tag{19}$$

meaning that $\mathbf{U}^*$ is a global minimizer of the problem

$$\underset{\mathbf{U}\in\mathrm{St}(p,k)}{\text{minimize}}\quad g\left(\mathbf{U}|\mathbf{U}^*\right). \tag{20}$$

Since the linear independence constraint qualification (LICQ) holds at $\mathbf{U}^*$ [7, Lemma 3], it satisfies the following conditions:

$$\begin{cases}\nabla g\left(\mathbf{U}^*|\mathbf{U}^*\right)=\mathbf{U}^*\mathbf{\Lambda}\\ (\mathbf{U}^*)^H\mathbf{U}^*=\mathbf{I},\end{cases} \tag{21}$$

where $\mathbf{\Lambda}$ is the associated Lagrange multiplier (see Appendix of [34] for more details). The continuity of f and g together with the upperbound condition of g implies that $\nabla g\left(\mathbf{U}^*|\mathbf{U}^*\right)=\nabla f(\mathbf{U}^*)$, which, combined with (21), implies that $\mathbf{U}^*$ satisfies the KKT conditions of problem (9). Now, assume that the whole sequence $\{\mathbf{U}_t\}_{t=0}^{\infty}$ does not converge to $\mathcal{K}$. Then there exists a convergent subsequence, indexed by $\{t_j\}$, such that $\lim_{j\to\infty}d^{(t_j)}(\mathcal{K})\geq c$ for $c\in\mathbb{R}_+^*$, and where $d^{(t_j)}(\mathcal{K})=\min_{\mathbf{Y}\in\mathrm{St}(p,k)}||\mathbf{U}^{t_j}-\mathbf{Y}||$. The subsequence $\{\mathbf{U}^{t_j}\}_{j=0}^{\infty}$ is bounded, and the compactness of the set $\mathrm{St}(p,k)$ again implies that it admits at least one limit point. As shown previously, this limit point is a KKT point of problem (9), which is a contradiction. Thus, the whole sequence $\{\mathbf{U}_t\}_{t=0}^{\infty}$ converges to $\mathcal{K}$. □

Remark 3. Note that the convergence to $\mathcal{K}$ does not imply the convergence of Algorithm 2 in terms of the variable $\mathbf{U}$. Establishing this property requires a case-by-case analysis which goes beyond the scope of this paper. In some cases the monotonic decrement of the objective can directly imply the convergence in terms of variable [35]. For the case of rotation invariant costs, this convergence in variable requires to be expressed in terms of subspace, e.g., as in [15].

C. Computational cost

First, recall that p denotes the dimension of the data, k refers to the dimension of the subspace of interest, and t indexes the algorithm steps. The sample size (number of columns of the data matrix) will be denoted n. The computational cost will be studied per iteration, therefore, it is to be multiplied by the total number of iterations T to obtain the overall complexity. A single iteration in Algortithm 2 essentially involves two operations:

- The computation of the matrix $\mathbf{R}(\mathbf{U})$: this step usually involves functions of the $p\times n$ data matrix and/or multiplying this matrix with the current point $\mathbf{U}^t$. Thus, this step is generally $\mathcal{O}(npk)$ (cf. examples in Section V). Also notice that this computation can, most of the time, be parallelized. Hence it does not represent the major bottleneck of Algorithm 2, contrarily to the second step.
- The computation of $\mathcal{P}_{\mathrm{St}}$: this step requires to compute the thin-SVD of a tall matrix $\mathbf{R}\in\mathbb{C}^{p\times k}$ which is $\mathcal{O}(pk^2+k^3)$. Also notice that this projection can be obtained through

$$\mathcal{P}_{\mathrm{St}}(\mathbf{R})=\mathbf{R}(\mathbf{R}^H\mathbf{R})^{-1/2}, \tag{22}$$

which does not improve the theoretical complexity but can be advantageous to implement in some practical cases.

Comparing to the existing approaches, e.g., the steepest descent on the Stiefel manifold [18], an iteration requires computing the gradient (also generally $\mathcal{O}(npk)$) and a retraction

(local mapping between a point in $\mathrm{St}(p,k)$ and its tangent space). The choice of the retraction is not unique, which leads to several options, e.g., based on geodesic paths [20], Euclidean projection [23], or QR decomposition [18]. Nevertheless, for all of the corresponding algorithms, the retraction step is $\mathcal{O}(pk^2+k^3)$. Hence, the computational complexity of an iteration of Algortithm 2 is on par with standard first-order based methods. However, this MM procedure is step-size free, thus it does not require the knowledge of any global parameter (such as the Lipschitz constant), or its adaptive estimation using a line search-type method. Compared to the latter option, this property effectively reduces the computational burden of each iteration, as it does not involve multiple computations of the retraction step (the computational bottleneck).

Establishing the total complexity of Algorithm 2 implies to study its convergence rate, which would require additional assumptions on the objective function and goes beyond the scope of this paper (pointers for some special cases are discussed in Section V-D). The convergence speed of an MM algorithm is generally expected to be sub-linear. Nevertheless, this possibly slow convergence can be compensated by the low-complexity of each iteration, resulting in a faster algorithm when considering the total computation time. More details on these remarks are illustrated with experiments in Section VII.

D. Extensions to block-MM

In addition to $\mathbf{U} \in \mathrm{St}(p,k)$ some problems may involve a set of side parameters $\boldsymbol{\theta} \in \mathbb{C}^{p'}$ (cf., e.g., [7–9]), and be formulated as

$$\begin{array}{ll} \underset{\mathbf{U},\mathbf{A}}{\text{minimize}} & f\left(\mathbf{U},\mathbf{A}\right) \\ \text{subject to} & (\mathbf{U},\mathbf{A}) \in \mathrm{St}(p,k) \times \Theta, \end{array} \tag{23}$$

where Θ is a compact of $\mathbb{C}^{p'}$. Such problems can be tackled by partitioning the variables and applying the block-MM algorithm [24], i.e., performing cyclic updates of the blocks, while keeping the others fixed. In our context, $\mathbf{U}$ typically represents a block updated by using one or several iterations of Algorithm 2. In this case, the convergence of the block-MM algorithm can be established from the analysis of the BSUM algorithm [30, 36] in conjunction with Proposition 1.

IV. Standard cost functions and their surrogates

The key to apply Algorithm 2 is to obtain a linear surrogate of the objective on the set $\mathrm{St}(p,k)$. In this section, we derive such surrogates functions for several standard minimization problems. Furthermore, maximization problems can also be tackled by considering the negative of the objective function (or equivalently, Minorization-Maximization).

A. Quadratic forms

First, define the Brockett function [18, Sec. 4.8] for $\mathbf{U} = [\mathbf{u}_1|\cdots|\mathbf{u}_k] \in \mathrm{St}(p,k)$ as

$$f_{\mathrm{B}}\left(\mathbf{U}\right) = \sum_{r=1}^{k} d_r \mathbf{u}_r^H \mathbf{M}\mathbf{u}_r = \mathrm{Tr}\left\{\mathbf{U}^H\mathbf{M}\mathbf{U}\mathbf{D}\right\}, \tag{24}$$

with $\mathbf{M} \in \mathcal{H}_p^+$, and $\mathbf{D} \in \mathcal{D}_k$ a diagonal matrix with $[\mathbf{D}]_{r,r} = d_r$ satisfying $0 \le d_1 \le \cdots \le d_k$. In the following, functions of the form f_{B} (resp. $-f_{\mathrm{B}}$) are referred to as convex (resp. concave) QFs. Note that some other expressions of QFs exist, but they can usually be rewritten as special cases or combinations (e.g., sums) of Brockett functions.

Proposition 2. ***(Majorization of concave QF)*** *The function $-f_{\mathrm{B}}$ as in* (24) *admits at point $\mathbf{U}_R^t$ a linear majorizing surrogate in the form of* (18)*, with*

$$\mathbf{R}(\mathbf{U}^t) = \mathbf{M}\mathbf{U}^t\mathbf{D}. \tag{25}$$

Equality holds at $\mathbf{U}^t$.

Proof. The function $-f_{\mathrm{B}}$ as in (24) is concave, so it can be majorized at point $\mathbf{U}^t$ by its first order Taylor expansion (cf. [24] section III.A), i.e.,

$$\begin{array}{rcl} -f_{\mathrm{B}}\left(\mathbf{U}\right) & \le & -\mathrm{Tr}\left\{\left(\mathbf{M}\mathbf{U}^t\mathbf{D}\right)^H\mathbf{U}\right\} \\ & & -\mathrm{Tr}\left\{\mathbf{U}^H\left(\mathbf{M}\mathbf{U}^t\mathbf{D}\right)\right\} + \mathrm{const.}. \end{array} \tag{26}$$

□

Remark 4. Majorizing a convex QF of $\mathbf{U}$ by a linear one seems counter-intuitive since it is not possible on the entire Euclidean space $\mathbb{C}^{N\times R}$. Nevertheless, the restriction to the set $\mathrm{St}(p,k)$ will make such an upperbound possible in Proposition 3. In order to give some insight, a visual example on $\mathbb{R}^2$ is also presented in Figure 1.

Proposition 3. ***(Majorization of convex QF)*** *The function f_{B} in* (24) *admits on $\mathrm{St}(p,k)$ and at point $\mathbf{U}^t$ a linear majorizing surrogate in the form of* (18)*, with*

$$\mathbf{R}(\mathbf{U}^t) = -\mathbf{K}\mathbf{U}^t\mathbf{D}, \tag{27}$$

where $\mathbf{K} = \mathbf{M} - \lambda_{\mathbf{M}}^{\max}\mathbf{I}$ and $\lambda_{\mathbf{M}}^{\max}$ is the largest eigenvalue of $\mathbf{M}$. Equality holds at $\mathbf{U}^t$.

Proof. The function f_{B} in (24) can be expressed as

$$\begin{array}{rcl} \mathrm{Tr}\left\{\mathbf{U}^H\mathbf{M}\mathbf{U}\mathbf{D}\right\} & = & \mathrm{Tr}\left\{\mathbf{U}^H\left(\mathbf{M}-\lambda_{\mathbf{M}}^{\max}\mathbf{I}\right)\mathbf{U}\mathbf{D}\right\} \\ & & +\mathrm{Tr}\left\{\mathbf{U}^H\left(\lambda_{\mathbf{M}}^{\max}\mathbf{I}\right)\mathbf{U}\mathbf{D}\right\}, \end{array} \tag{28}$$

where the second term is constant and equal to $\lambda_{\mathbf{M}}^{\max}\mathrm{Tr}\left\{\mathbf{D}\right\}$ for the restriction $\mathbf{U} \in \mathrm{St}(p,k)$. The first term of this expression is concave in $\mathbf{U}$ ($\mathbf{U} \in \mathbb{C}^{M\times R}$) so it can be upper-bounded by its first order Taylor expansion, thus

$$\begin{array}{rcl} f_{\mathrm{B}}\left(\mathbf{U}\right) & \le & +\mathrm{Tr}\left\{\left(\mathbf{K}\mathbf{U}^t\mathbf{D}\right)^H\mathbf{U}\right\} \\ & & +\mathrm{Tr}\left\{\mathbf{U}^H\left(\mathbf{K}\mathbf{U}^t\mathbf{D}\right)\right\} + \mathrm{const.}, \end{array} \tag{29}$$

with $\mathbf{K}$ defined as in Proposition 3. □

B. Concave compositions of quadratic forms

Compositions involving inner QFs that yield concave functions are often used in order to build robust loss functions (examples are given in section V-C). The following proposition gives a linear majorizer of concave functions composed from the Brockett function.

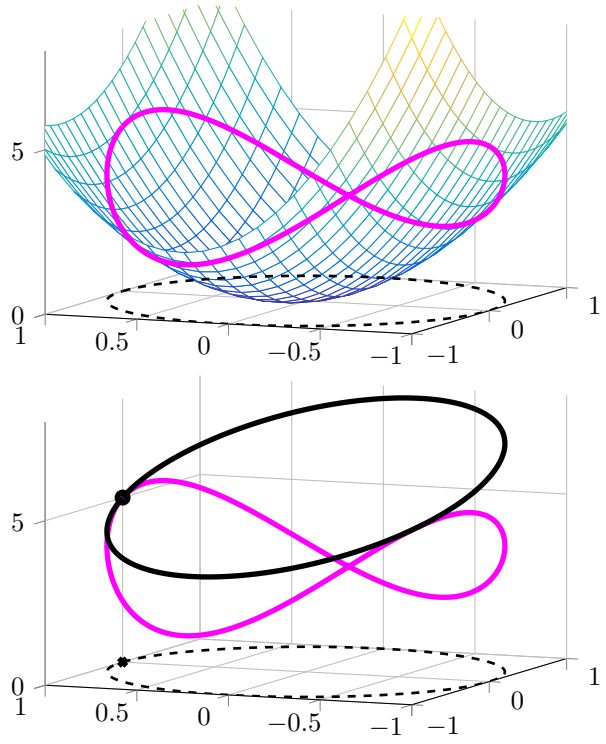

Fig. 1. Quadratic form on $\mathbb{R}^2$ and its restriction to $\mathrm{St}(2,1)$ (up). Linear majorization of this quadratic form at a given point over the set $\mathrm{St}(2,1)$ (down).

Proposition 4. ***(Majorization of concave function composed from concave QF)*** *Let $\rho : \mathbb{R} \to \mathbb{R}$ be a concave non-decreasing function. For f_{B} as in* (24)*, the function $\rho(-f_{\mathrm{B}})$ admits at point $\mathbf{U}^t$ a linear majorizing surrogate in the form of* (18)*, with*

$$\mathbf{R}(\mathbf{U}^t) = \rho'\left(-f_{\mathrm{B}}\left(\mathbf{U}^t\right)\right)\mathbf{M}\mathbf{U}^t\mathbf{D}. \tag{30}$$

Equality holds at $\mathbf{U}^t$.

Proof. The function $-f_{\mathrm{B}}$ is concave in $\mathbf{U}$ and ρ is concave non-decreasing. It follows that the function $\rho(-f_{\mathrm{B}})$ is concave [37], so it can be upper-bounded at point $\mathbf{U}^t$ by its first order Taylor expansion, i.e.,

$$\begin{aligned}\rho\left(-f_{\mathrm{B}}\left(\mathbf{U}\right)\right) \;\leq\; & -\rho'\left(-f_{\mathrm{B}}\left(\mathbf{U}^t\right)\right)\mathrm{Tr}\left\{\left(\mathbf{M}\mathbf{U}^t\mathbf{D}\right)^H\mathbf{U}\right\}\\ & -\rho'\left(-f_{\mathrm{B}}\left(\mathbf{U}^t\right)\right)\mathrm{Tr}\left\{\mathbf{U}^H\left(\mathbf{M}\mathbf{U}^t\mathbf{D}\right)\right\}\\ & +\mathrm{const}.\end{aligned} \tag{31}$$

□

Following this proof, other linear surrogates can be derived using compositions of concave (non-decreasing/non-increasing) functions and the chain rule. It is also worth noting that one can apply the reformulation of Proposition 3 to express a quadratic QF as a concave term plus a constant in order to do so (the obtained majoration is then only valid on $\mathrm{St}(p,k)$).

C. Quotients of quadratic forms

Various formulations of quotients of quadratic forms arise in generalized versions of PCA [35]. Most of them can be obtained as linear combinations of functions of the form

$$f_{\mathrm{q}}(\mathbf{U}) = -\mathrm{Tr}\left\{\left(\mathbf{U}^H\mathbf{C}\mathbf{U}\right)^{-1}\mathbf{U}^H\mathbf{A}\mathbf{U}\right\}, \tag{32}$$

where $\mathbf{C}$ is a positive definite and $\mathbf{A}$ is positive semi-definite.

Proposition 5. ***(Majorization of quotient of QFs)*** *The function f_{q} as in* (32) *admits on $\mathrm{St}(p,k)$ and at point $\mathbf{U}^t$ a linear majorizing surrogate in the form of* (18)*, with*

$$\mathbf{R}(\mathbf{U}^t) = \mathbf{T}(\mathbf{U}^t) - \left(\mathbf{K}\mathbf{U}^t\widetilde{\mathbf{T}}(\mathbf{U}^t)\right), \tag{33}$$

and

$$\begin{aligned}&\mathbf{T}(\mathbf{U}^t) = \mathbf{A}\mathbf{U}^t\left((\mathbf{U}^t)^H\mathbf{C}\mathbf{U}^t\right)^{-1},\\ &\widetilde{\mathbf{T}}(\mathbf{U}^t) = \left(\mathbf{A}^{-1/2}\mathbf{T}(\mathbf{U}^t)\right)^H\left(\mathbf{A}^{-1/2}\mathbf{T}(\mathbf{U}^t)\right),\\ &\mathbf{K} = \mathbf{C} - \lambda_{\mathbf{C}}^{\max}\mathbf{I},\end{aligned} \tag{34}$$

where $\lambda_{\mathbf{C}}^{\max}$ is the largest eigenvalue of $\mathbf{C}$. Equality holds at $\mathbf{U}^t$.

Proof. Starting from the inequality

$$\begin{aligned}&\left\|\left(\mathbf{U}^H\mathbf{C}\mathbf{U}\right)^{-1/2}\mathbf{U}^H\mathbf{A}^{1/2}\right.\\ &\left.-\left(\mathbf{U}^H\mathbf{C}\mathbf{U}\right)^{1/2}\left((\mathbf{U}^t)^H\mathbf{C}\mathbf{U}^t\right)^{-1}(\mathbf{U}^t)^H\mathbf{A}^{1/2}\right\|^2 \geq 0,\end{aligned} \tag{35}$$

we obtain

$$\begin{aligned}f_{\mathrm{q}}(\mathbf{U}) \;\leq\; & -2\mathfrak{Re}\left\{\left(\mathbf{T}(\mathbf{U}^t)\right)^H\mathbf{U}\right\}\\ & +\mathrm{Tr}\left\{\mathbf{U}^H\mathbf{C}\mathbf{U}\widetilde{\mathbf{T}}(\mathbf{U}^t)\right\},\end{aligned} \tag{36}$$

with $\mathbf{T}(\mathbf{U}^t)$ and $\widetilde{\mathbf{T}}(\mathbf{U}^t)$ as in (34), and where equality holds at $\mathbf{U}^t$. Following the proof of Proposition 3, we can majorize on $\mathrm{St}(p,k)$ the quadratic term in (36) as

$$\begin{aligned}\mathrm{Tr}\left\{\mathbf{U}^H\mathbf{C}\mathbf{U}\widetilde{\mathbf{T}}(\mathbf{U}^t)\right\} \;\leq\; & +\mathrm{Tr}\left\{\left(\mathbf{K}\mathbf{U}^t\widetilde{\mathbf{T}}(\mathbf{U}^t)\right)^H\mathbf{U}\right\}\\ & +\mathrm{Tr}\left\{\mathbf{U}^H\left(\mathbf{K}\mathbf{U}^t\widetilde{\mathbf{T}}(\mathbf{U}^t)\right)\right\}\\ & +\mathrm{const}.,\end{aligned} \tag{37}$$

with $\mathbf{K}$ as in (34), and where equality holds again at $\mathbf{U}^t$. Combining the inequalities (36) and (37) concludes the proof. □

V. Examples and links with existing algorithms

This section provides some examples of MM algorithms that leverage the proposed method with the surrogate functions from Section IV. When existing, some links with the existing state-of-the-art algorithms are highlighted.

A. Power iteration methods

Let $\mathbf{u} \in \mathrm{St}(p,1)$ and $\mathbf{M} \in \mathcal{H}_p^+$. Consider the classical problem

$$\begin{aligned} \underset{\mathbf{u}}{\text{maximize}} \quad & \mathbf{u}^H \mathbf{M} \mathbf{u} \\ \text{subject to} \quad & \mathbf{u}^H \mathbf{u} = 1, \end{aligned} \tag{38}$$

whose local solutions are the eigenvectors of $\mathbf{M}$. In order to compute those solutions, we can apply the framework proposed in Section III. First, we simply rewrite the problem as a minimization one by changing the sign of the objective:

$$\underset{\mathbf{u} \in \mathrm{St}(p,1)}{\text{minimize}} \quad \mathbf{u}^H (-\mathbf{M}) \mathbf{u}. \tag{39}$$

Second, since the objective in (39) is a concave QF, we can apply Proposition 2 in Section IV-A to obtain a majorizing linear surrogate function at the point $\mathbf{u}^t$ as:

$$g(\mathbf{u}|\mathbf{u}^t) = -\mathbf{u}^H \mathbf{M} \mathbf{u}^t - (\mathbf{u}^t)^H \mathbf{M} \mathbf{u} + \text{const}. \tag{40}$$

Finally, we recognize a surrogate function as in (18), with $\mathbf{R}(\mathbf{u}^t) = \mathbf{M}\mathbf{u}^t$. Therefore, we can apply Algorithm 2, which leads to the following MM iterations:

$$\mathbf{u}^{t+1} = \mathcal{P}_{\mathrm{Proc}} \left\{ \mathbf{M} \mathbf{u}^t \right\}. \tag{41}$$

This algorithm simply corresponds to the normalization of the iterate $\mathbf{M}\mathbf{u}^t$. Thus, Algorithm 2 for this problem yields the well-known power iteration method [38]. This algorithm also appears as special case of a steepest-descent method on the Stiefel manifold (cf. Section 4.6.6 [18]).

Let us now consider the general case $\mathbf{U} \in \mathrm{St}(p,k)$, with corresponding problem:

$$\begin{aligned} \underset{\mathbf{U}}{\text{maximize}} \quad & \mathbf{U}^H \mathbf{M} \mathbf{U} \\ \text{subject to} \quad & \mathbf{U}^H \mathbf{U} = \mathbf{I}. \end{aligned} \tag{42}$$

The stationary points of this problem are k-dimensional invariant subspaces of $\mathbf{M}$. Applying the same steps as in (39)-(41), we obtain the MM algorithm:

$$\mathbf{U}^{t+1} = \mathcal{P}_{\mathrm{Proc}} \left\{ \mathbf{M} \mathbf{U}^t \right\}, \tag{43}$$

which is a generalization of the power method similar to the orthogonal iterations [38] (that equivalently relies on the QR decomposition rather than $\mathcal{P}_{\mathrm{Proc}}$ to orthonormalize the updates).

B. MM for generic non-homogeneous QF (with a linear term)

Let $\mathbf{U} = [\mathbf{u}_1 | \cdots | \mathbf{u}_k] \in \mathrm{St}(p,k)$ and a set of k positive semi-definite Hermitian matrices $\{\mathbf{M}_r\}$. Consider the following problem (appearing for example as a subproblems in [9, 39]):

$$\begin{aligned} \underset{\mathbf{U}}{\text{maximize}} \quad & \sum_{r=1}^{k} \left[\mathbf{u}_r^H \mathbf{M}_r \mathbf{u}_r + 2\mathfrak{Re}\{\mathbf{u}_r^H \mathbf{c}_r\} \right] \\ \text{subject to} \quad & \mathbf{u}_i^H \mathbf{u}_j = \delta_{i,j} \ \forall \ i,j \in [\![1,k]\!], \end{aligned} \tag{44}$$

which has no closed form solution related to the SVD of the matrices $\mathbf{M}_r$ and the vectors $\mathbf{c}_r$. First, we rewrite the problem as

$$\underset{\mathbf{U} \in \mathrm{St}(p,k)}{\text{minimize}} \quad \sum_{r=1}^{k} \left[\mathbf{u}_r^H (-\mathbf{M}_r) \mathbf{u}_r - 2\mathfrak{Re}\{\mathbf{u}_r^H \mathbf{c}_r\} \right]. \tag{45}$$

The objective of (45) can be split into two functions and allows us to follow (15)-(17). First, notice that the linear elements of the sum are already expressed as in (18) since

$$-2\mathfrak{Re}\{\mathbf{u}_r^H \mathbf{c}_r\} = -\mathbf{u}_r^H \mathbf{c}_r - \mathbf{c}_r^H \mathbf{u}_r, \ \forall r \in [\![1,k]\!]. \tag{46}$$

Second, for each concave quadratic element of the sum, we apply the same steps as in (39)-(41). This leads to a surrogate function for the total objective in (45) in the form of (18) with

$$\mathbf{R}\left(\mathbf{U}^t\right) = \left[\mathbf{M}_1 \mathbf{u}_1^t + \mathbf{c}_1 | \cdots | \mathbf{M}_k \mathbf{u}_k^t + \mathbf{c}_k \right]. \tag{47}$$

Therefore, we can apply Algorithm 2, which leads to the following MM iterations:

$$\mathbf{U}^{t+1} = \mathcal{P}_{\mathrm{Proc}} \left\{ \left[\mathbf{M}_1 \mathbf{u}_1^t + \mathbf{c}_1 | \cdots | \mathbf{M}_k \mathbf{u}_k^t + \mathbf{c}_k \right] \right\}. \tag{48}$$

We also note that this algorithm corresponds to the one of [27], which is often called for minimizing linear-plus-quadratic terms under orthonormality constraints.

C. MM for nonconvex robust subspace recovery (RSR)

The aim of RSR is to estimate a low-dimensional subspace from a (demeaned) dataset $\{\mathbf{z}_i\}_{i=1}^n$, while being robust to potential outliers within this set. There are many approaches to tackle this problem [11]. Among them, several RSR estimators have been proposed as minimizers of a nonconvex robust regression function [12–16], with orthonormality constraint.

Following [14], a problem for such RSR can be formulated as follows:

$$\underset{\mathbf{U} \in \mathrm{St}(p,k)}{\text{minimize}} \quad \sum_{i=1}^{n} \rho\left(\mathrm{d}^2\left(\mathbf{U}, \mathbf{z}_i\right) \right), \tag{49}$$

where

$$\mathrm{d}^2\left(\mathbf{U}, \mathbf{z}\right) = \|(\mathbf{U}\mathbf{U}^H)^\perp \mathbf{z}\|_F^2 = \mathbf{z}^H \mathbf{z} - \mathrm{Tr}\left(\mathbf{U}^H \mathbf{z}\mathbf{z}^H \mathbf{U}\right) \tag{50}$$

is the Euclidean distance between a vector $\mathbf{z} \in \mathbb{C}^p$ and the subspace spanned by $\mathbf{U} \in \mathrm{St}(p,k)$, and $\rho : \mathbb{R} \to \mathbb{R}$ is a function that ensures the robustness to outliers. Here, ρ is assumed to be a concave nondecreasing function. However, this constraint is not restrictive, as it holds for a wide variety of usual robust formulations, as illustrated by the following examples:

Example 1. (ℓ_p-norm) For $p > 0$, ℓ_p-norm nonconvex RSR estimators are defined as in (49) by using the function

$$\rho_p(t) = t^{p/2}. \tag{51}$$

The least-square estimator is defined for $p = 2$. In this case, it is not hard to show that (49) is equivalent to (42) with $\mathbf{M} = \sum_{k=1}^n \mathbf{z}_k \mathbf{z}_k^H$. The solution is the k the leading eigenvectors of the sample covariance matrix, which corresponds to the standard PCA estimator.

Example 2. (Huber-type) For a parameter $T \geq 0$, Huber-type nonconvex RSR estimators are defined as in (49) by using the function

$$\rho_{\mathrm{H}}(t) = \begin{cases} t/\sqrt{T} & \text{if } t \leq T, \\ 2\sqrt{t} - \sqrt{T} & \text{if } t > T. \end{cases} \tag{52}$$

The median estimator, e.g., considered in [14, 15], corresponds to the case limit case $T \to 0$.

Example 3. (Cauchy–Lorentz-type) For a parameter $T \geq 1$, Cauchy–Lorentz-type nonconvex RSR estimators are defined as in (49) by using the function

$$\rho_{\mathrm{CL}}(t) = T \ln(T + t). \tag{53}$$

Example 4. (Geman-McClure-type) For a parameter $T \geq 0$, Geman-McClure-type nonconvex RSR estimators are defined as in (49) by using the function

$$\rho_{\mathrm{GMC}}(t) = t/(T + t), \tag{54}$$

which has been used in, e.g., [12].

From the problem in (49), we exhibit some links between algorithms from the state of the art and the proposed framework. First, notice that ρ is a concave nondecreasing and d in (50) is a concave QF of $\mathbf{U}$. Therefore each element of the sum of the objective in (49) can be majorized by using Proposition 4 in Section IV. This leads to a surrogate function in the form of (18) with $\mathbf{R}\left(\mathbf{U}^t\right) = \mathbf{M}\left(\mathbf{U}^t\right)\mathbf{U}^t$, where

$$\mathbf{M}\left(\mathbf{U}\right) = \sum_{i=1}^{n} \rho'\left(\mathrm{d}^2\left(\mathbf{U}, \mathbf{z}_i\right)\right) \mathbf{z}_i \mathbf{z}_i^H. \tag{55}$$

Applying Algorithm 2 thus yields the MM iterations

$$\mathbf{U}^{t+1} = \mathcal{P}_{\mathrm{Proc}}\left\{\mathbf{M}\left(\mathbf{U}^t\right)\mathbf{U}^t\right\}. \tag{56}$$

The resulting algorithm corresponds to R1-PCA-type fixed-point iterations, initially proposed as heuristic in [14].

Additionally, the MM perspective allows us to draw some links with other algorithms that aim at solving (49). It is shown in [15] that the objective of (49) is also majorized with equality at the point $\mathbf{U}^t$ by a concave quadratic surrogate function as

$$\sum_{i=1}^{n} \rho\left(\mathrm{d}^2\left(\mathbf{U}, \mathbf{z}_i\right)\right) \leq -\mathrm{Tr}\left\{\mathbf{U}^H \mathbf{M}\left(\mathbf{U}^t\right)\mathbf{U}\right\} + \mathrm{const}. \tag{57}$$

Minimizing the surrogate function in (57) is a problem equivalent to the one in (42) (with $\mathbf{M} = \mathbf{M}\left(\mathbf{U}^t\right)$). An update can thus be obtained by computing the k leading eigenvectors of $\mathbf{M}\left(\mathbf{U}^t\right)$. This algorithm corresponds to the fixed-point iteration proposed in [13]. An equivalent update can be obtained through the SVD of the data matrix $\mathbf{Z}^t = [\mathbf{z}_1|\cdots|\mathbf{z}_n]$ using normalized samples $\mathbf{z}_i^t = \sqrt{\rho'\left(\mathrm{d}^2\left(\mathbf{U}^t, \mathbf{z}_i\right)\right)}\mathbf{z}_i$, which corresponds to the FMS algorithm proposed in [15]. Following the example of Section V-A, a third option to compute this update is to perform an inner loop using an MM algorithm (i.e., the power method in (43)). This procedure corresponds to the algorithm referred to as *FMS with power method* in [15]. Alternatively, the surrogate function in (57) is a concave QF so it can be majorized with Proposition 2: the resulting MM algorithm is again (56). Within this perspective, we can also note that Algorithm (2) in (56) (R1-PCA-type) also appears as the FMS with power method when only one power iteration is performed at each step. Some simulations comparing these approaches are presented in Section VII.

Algorithm 3 Frank-Wolfe method

1: **initialize:** $\mathbf{x}_0 \in \mathcal{D}$
2: **repeat**
3: Compute $\mathbf{s} = \mathrm{argmin}_{\mathbf{s}\in\mathcal{D}} \ \langle \mathbf{s}, \nabla f(\mathbf{x}_t)\rangle$
4: Update $\mathbf{x}_{t+1} = (1-\gamma)\mathbf{x}_t + \gamma\mathbf{s}$ for $\gamma = 2/(t+2)$
5: $t = t + 1$
6: **until** convergence criterion is met

D. Frank-Wolfe and first-order approximation methods

The Frank-Wolfe method [40] (cf. Algorithm 3) regained popularity in machine learning related fields [41] for minimizing a convex function f (with gradient ∇f) over a convex set $\mathcal{D}$. This approach was also recently extended to the Riemannian optimiation framework in [42] as an alternative to first order projection based method (e.g., Steepest descent [18]). In this method, a descent direction is obtained by minimizing a first-order linear approximation of the objective, which is reminescent of the proposed approach (though the linear approximation is not a majorizer of the concave function). However, the problems considered in this paper are not convex (nor g-convex) so Frank-Wolfe method is not trivially transposable on the Stiefel manifold. Yet, similar first-order approximation methods have been proposed for this set. Notably, [43, Sec. 3.4] considered minimizing a concave function with the following algorithm

$$\mathbf{U}_{t+1} = \underset{\mathbf{U}\in\mathrm{St}(p,k)}{\mathrm{argmin}} \ \langle \mathbf{U}, \nabla f(\mathbf{U}_t)\rangle \tag{58}$$

which can be interpreted as a Frank-Wolfe method with a constant (maximal) step $\gamma = 1$. This algorithm also corresponds to an MM approach since the first order Taylor approximation is a linear majorizing surrogate of a concave function. Interestingly, the proposed framework also generalizes to objectives that are not necessarily concave, as it only requires a linear surrogate of the objective on $\mathrm{St}(p,k)$. Conversely, some stronger convergence results (such as the convergence rate) for Algorithm 2 can be obtained from [43] when the objective is concave.

VI. Majorization-Minimization for sparse PCA

In standard PCA, the estimated principal components are usually dense (i.e., a linear combinations of all entries of the variables). Since the principal components have an actual physical meaning in many applications, estimating sparse principal components can significantly help the interpretation, as well as the variable selection process. Many algorithms have been proposed to perform this task [3–8]. Most of the proposed methods involve a variable $\mathbf{U} \in \mathrm{St}(p,k)$, whose columns represent the principal components, and can be generically formulated through the problem

$$\underset{\mathbf{U}\in\mathrm{St}(p,k)}{\mathrm{minimize}} \quad \mathcal{L}(\mathbf{U}, \{\mathbf{z}_i\}_{i=1}^n) + \lambda\xi\left(\mathbf{U}\right), \tag{59}$$

where $\mathcal{L}$ is a data fitting term (orthogonal regression on the dataset $\{\mathbf{z}_i\}_{i=1}^n$), ξ is a sparsity promoting penalty, and λ is a regularization parameter. The introduction of the penalty usually makes the minimization in (59) hard to deal with under

the orthonormality constraint. Thus, most algorithms in the literature relax this constraint and resort to a trade-off [3–6]. In the following, this issue is addressed by using the proxies of ℓ_0-norm proposed in [31] to promote sparsity, while leveraging the MM framework of Section III to perform the optimization under the orthonormality constraint.

A. Sparse regularizers with linear surrogates on $\mathrm{St}(p,k)$

One approach to force sparsity in the principal components $\mathbf{U}=[\mathbf{u}_1|\cdots|\mathbf{u}_k]$ is to use a penalty involving the ℓ_0-norm[3]:

$$\|\mathbf{U}\|_0 = \sum_{r,i} \mathrm{sgn}\left(\ |[\mathbf{u}_r]_i|\ \right), \tag{60}$$

where $|\cdot|$ stands for the modulus of a complex number. Another approach, that promotes a row-sparse structure, is the use of the $\ell_{2,0}$-norm, defined by:

$$\|(\mathbf{U})\|_{2,0}^2 = \sum_{i=1}^{p}\left(\sum_{r=1}^{k}\mathrm{sgn}\left(\ |[\mathbf{u}_r]_i|\ \right)\right)^2. \tag{61}$$

However, both of these functions are too complex to deal with due to their discontinuity.

To alleviate this issue, we follow the approach proposed in [31], i.e., approximating the sign function by a smooth function denoted l_γ^ϵ, and defined as

$$l_\gamma^\epsilon(x) = \begin{cases} a|x|^2, & \text{if } |x|\leq\epsilon \\ l_\gamma(x)-b, & \text{if } |x|>\epsilon, \end{cases} \tag{62}$$

with appropriate constants a and b so that the approximations l_γ^ϵ are continuous and differentiable (cf. [31]), and where l_γ belongs to the family of functions:

a) ℓ_γ-norm [44–46]:
$l_\gamma(x) = |x|^\gamma,\ \gamma\in(0,1]$

b) ℓ_1-norm approximation from [47, 48]:
$l_\gamma(x) = \ln(1+|x|/\gamma)\ln(1+1/\gamma),\ \gamma>0$

c) lower bound of sign function from [49]:
$l_\gamma(x) = 1-e^{-|x|/\gamma},\ \gamma>0,$

involving a tuning parameter γ for each case. Thus, this class covers most of standard 1-dimensional sparsity forcing penalties (i.e., a proxy of the sign function). Notice that this class is still valid for data with complex entries by reading $|\cdot|$ as the modulus function. We have the following proposition (cf. [31, Section III]):

Proposition 6. *The function l_γ^ϵ in* (62) *is majorized at point x_t by the following quadratic surrogate:*

$$l_\gamma^\epsilon(x|x_t) \leq \phi(x_t)|x|^2 + \mathrm{const.} \tag{63}$$

where the function ϕ depends on l_γ^ϵ as:

a) *ℓ_γ-norm [44–46]:*
$$\phi(x_t) = \begin{cases} (\gamma/2)\epsilon^{\gamma-2}, & |x_t|\leq\epsilon \\ (\gamma/2)\,|x_t|^{\gamma/2}, & |x_t|>\epsilon \end{cases}$$

b) *ℓ_1-norm approximation [47, 48]:*
$$\phi(x_t) = \begin{cases} (\ 2\epsilon(\gamma+\epsilon)\ln(1+1/\gamma)\ )^{-1}, & |x_t|\leq\epsilon \\ (\ 2\ln(1+1/\gamma)\,|x_t|\,(|x_t|+\gamma)\ )^{-1}, & |x_t|>\epsilon \end{cases}$$

c) *lower bound of sign function [49]:*
$$\phi(x_t) = \begin{cases} e^{-\epsilon/\gamma}/2\gamma\epsilon, & |x_t|\leq\epsilon \\ e^{-|x_t|/\gamma}/2\gamma\,|x_t|, & |x_t|>\epsilon \end{cases}$$

Equality is achieved at x_t.

The functions l_γ^ϵ will serve as basic building blocks to build sparsity promoting penalties on $\mathbf{U}$. In the following we propose two functions that extend the approach of [7], with corresponding surrogates functions in the form of (18). Our aim is to give insight on how to build meaningful cost functions that can be easily handled within the framework of Section III: the key trick is to obtain a series of inequalities, notably using Proposition 6, that leads to a quadratic surrogate of $\mathbf{U}$ over $\mathrm{St}(p,k)$. At this step, Proposition 3 can be applied to obtain a linear majorizer on $\mathrm{St}(p,k)$.

1) Proxies of the ℓ_0-norm on $\mathrm{St}(p,k)$: In order to mimic a weighted ℓ_0-norm in (60), we consider the cost function:

$$r_0(\mathbf{U}) = \sum_{r=1}^{k}\omega_r\sum_{i=1}^{p} l_\gamma^\epsilon\left([\mathbf{u}_r]_i\right), \tag{64}$$

where ω_r are positive weights and with l_γ^ϵ in (62). Such type of penalty was initially proposed for Sparse PCA in [7] using the ℓ_1-norm approximation [47, 48] (i.e., l_γ^ϵ in the class b) of the considered family)).

Proposition 7. *The function r_0 in* (64) *is majorized on the set of unitary matrices $\mathrm{St}(p,k)$ at point $\mathbf{U}^t=[\mathbf{u}_1^t|\cdots|\mathbf{u}_k^t]$ as*

$$\begin{aligned} r_0(\mathbf{U}) \ \leq\ & \mathrm{Tr}\left\{\mathbf{H}^H\left(\mathbf{U}^t\right)\mathbf{U}\right\} + \mathrm{Tr}\left\{\mathbf{U}^H\mathbf{H}\left(\mathbf{U}^t\right)\right\} \\ & + \mathrm{const.}, \end{aligned} \tag{65}$$

with matrices $\mathbf{H}(\mathbf{U}^t)\in\mathbb{C}^{p\times k}$, $\mathbf{W}_r\in\mathbb{C}^{p\times p}$, $\mathbf{w}_r\in\mathbb{C}^{p\times 1}$ defined as

$$\begin{cases} \mathbf{H}\left(\mathbf{U}^t\right) = \left[\ \mathbf{W}_1\mathbf{u}_1^t\ |\ \cdots\ |\ \mathbf{W}_k\mathbf{u}_k^t\ \right] \\ \mathbf{W}_r = \mathrm{diag}(\ \mathbf{w}_r\ -\ \mathrm{w}_r^{\max}\ \mathbf{1}_p\ ) \\ [\mathbf{w}_r]_i = \omega_r\phi\left(\left[\mathbf{u}_r^t\right]_i\right) \\ \mathrm{w}_r^{\max} = \max(\mathbf{w}_r), \end{cases} \tag{66}$$

with ϕ defined (depending on the chosen l_γ^ϵ) in Proposition 6.

Proof. Proposition 6 allows to upperbound r_0 in (64) at point $\mathbf{U}^t$ as

$$r_0(\mathbf{U}) \leq \sum_{r=1}^{k}\mathbf{u}_r^H\left(\mathrm{diag}(\mathbf{w}_r)\right)\mathbf{u}_r + \mathrm{const.}, \tag{67}$$

with $\mathbf{w}_r$ defined in (66) and where equality is achieved at $\mathbf{U}^t$. Proposition 3 applied to each quadratic term of this majorizer leads to the surrogate in (65) and (66). □

2) Proxies of the mixed-norm on $\mathrm{St}(p,k)$: To mimic the $\ell_{2,0}$-norm in (61), the main idea is to involve a nonlinear coupling among the elements in a row. We consider therefore the cost function:

$$r_{2,0}(\mathbf{U}) = \sum_{i=1}^{p}\ln\left(1+\sum_{r=1}^{k} l_\gamma^\epsilon\left([\mathbf{u}_r]_i\right)\right). \tag{68}$$

Notice that the function $\ln$ is essentially used in order to obtain a surrogate satisfying (18). However, the main goal

[3]Note that this function is not a proper norm, but we still adopt the common abuse of terminology.

Algorithm 4 MM for RSPCA

initialize $t=0$, $\mathbf{U}^0$
repeat
  Compute $\mathbf{M}\left(\mathbf{U}^t\right)$ with equation (55).
  Compute the regularizer $\mathbf{K}\left(\mathbf{U}^t\right)$ with (74)
  Update $\mathbf{U}^{t+1}=\mathcal{P}_{\mathrm{St}}\left\{\frac{1}{n}\mathbf{M}\left(\mathbf{U}^t\right)\mathbf{U}^{tH}-\lambda\mathbf{K}\left(\mathbf{U}^t\right)\right\}$
  $t=t+1$
until convergence criterion is met
Output: $\hat{\mathbf{U}}$

of introducing coupling among the elements of each row is achieved with this function.

Proposition 8. *The function $r_{2,0}$ in* (68) *is majorized on the set of unitary matrices $\mathrm{St}(p,k)$ at point $\mathbf{U}^t=[\mathbf{u}_1^t|\cdots|\mathbf{u}_k^t]$ as*

$$r_{2,0}\left(\mathbf{U}\right) \;\leq\; \mathrm{Tr}\left\{\mathbf{P}^H\left(\mathbf{U}^t\right)\mathbf{U}\right\}+\mathrm{Tr}\left\{\mathbf{U}^H\mathbf{P}\left(\mathbf{U}^t\right)\right\} + \mathrm{const.}, \tag{69}$$

with

$$\begin{cases}\mathbf{P}\left(\mathbf{U}^t\right)=\left[\mathbf{X}_1\mathbf{u}_1^t|\cdots|\mathbf{X}_k\mathbf{u}_k^t,\right]\\ \mathbf{X}_r=\mathrm{diag}(\mathbf{x}_r-\mathrm{x}_r^{\max}\mathbf{1}_p)\\ [\mathbf{x}_r]_i=\omega_n^t\phi\left([\mathbf{u}_r^t]_i\right)\\ \mathrm{x}_r^{\max}=\max(\mathbf{x}_r)\\ \omega_i^t=\left(1+\sum_{r=1}^{k}l_\gamma^\epsilon([\mathbf{u}_r^t]_i)\right)^{-1}\end{cases} \tag{70}$$

with ϕ defined (depending on the chosen l_γ^ϵ) in Proposition 6.

Proof. The function $\ln$ is concave, so it can be upperbounded by its first order Taylor expansion in (68), giving:

$$r_{2,0}(\mathbf{U}|\mathbf{U}^t)\leq\sum_{i=1}^{p}\omega_n^t\sum_{r=1}^{k}l_\gamma^\epsilon\left([\mathbf{u}_r]_i\right)+\mathrm{const.} \tag{71}$$

with ω_n^t in (70). The rest of the proof follows similarly to the one of Proposition 7. □

B. Robust Sparse PCA (RSPCA): Proposed Algorithm

1) Formulation: From the generic formulation of (59) for sparse PCA, we consider the following problem:

$$\underset{\mathbf{U}\in\mathrm{St}(p,k)}{\text{minimize}}\quad\frac{1}{n}\sum_{i=1}^{n}\rho\left(\mathrm{d}^2\left(\mathbf{U},\mathbf{z}_i\right)\right)+\lambda\xi\left(\mathbf{U}\right). \tag{72}$$

where the data fitting term is a nonconvex RSR problem (cf. Section V-C), and where ξ is either r_0 in (64) or $r_{2,0}$ in (68). Hence, the proposed objective offers a quite modular formulation that encompasses various RSR methods, and all the family of regularization penalty from Section VI-A.

2) Algorithm derivation: In this section, an MM algorithm is derived to aim at solving problem (72) by following the framework proposed in Section III. To majorize the nonconvex RSR cost, we apply the same steps as in Section V-C. The penalty ξ can be majorized according to either Proposition 7 or 8. Thus, the objective in (72) can be majorized at point $\mathbf{U}^t$ by a surrogate satisfying (18) with

$$\mathbf{R}(\mathbf{U^t})=\frac{1}{n}\mathbf{M}\left(\mathbf{U}^t\right)\mathbf{U}^t-\lambda\mathbf{K}\left(\mathbf{U}^t\right) \tag{73}$$

where $\mathbf{M}$ is defined in (55) and

$$\mathbf{K}\left(\mathbf{U}^t\right)=\begin{cases}\mathbf{H}\left(\mathbf{U}^t\right)\ \text{ in (66), if }\ \xi=r_0\\ \mathbf{P}\left(\mathbf{U}^t\right)\ \text{ in (70), if }\ \xi=r_{2,0}\end{cases} \tag{74}$$

Therefore, we can apply Algorithm 2, which leads to the following MM iterations:

$$\mathbf{U}^{t+1}=\mathcal{P}_{\mathrm{St}}\left\{\frac{1}{n}\mathbf{M}\left(\mathbf{U}^t\right)\mathbf{U}^t-\lambda\mathbf{K}\left(\mathbf{U}^t\right)\right\}. \tag{75}$$

This algorithm is summed up in Algorithm 4 and will be referred to as *Robust-Sparse PCA* (RSPCA).

Remark 5. Notice that we focused on the use of the RSR functions from Section V-C in order leverage their robustness properties. However, the proposed method can be generalized to any data fitting term (e.g., more complex likelihood functions) for which Algorithm 2 applies.

VII. Simulations

A. On MM algorithms with linear majorizers for $\mathrm{St}(p,k)$

This section illustrates the computational complexity advantage of the proposed Algorithm 2 in comparison to existing state-of-the-art techniques. As an example, we consider minimizing the RSR cost in (49) with the Huber-type function (52) with parameter $T=0.1$. The data set is generated using a spiked Gaussian model, i.e., $\mathbf{z}\sim\mathcal{CN}(\mathbf{0},\mathrm{SNR}\times\mathbf{U}_0\mathbf{U}_0^H+\mathbf{I})$, with $\mathrm{SNR}=10$ and where $\mathbf{U}_0\in\mathrm{St}(p,k)$ contains the first k vectors of the canonical basis (this setting is chosen as any rotation of the samples did not affect results in terms of computational speed). The following methods are compared: i) Steepest descent on the Stiefel manifold: for this algorithm, we compare both the retraction using the Euclidean projection [23]:

$$\mathbf{U}^{t+1}\leftarrow\mathcal{P}_{\mathrm{St}}\left\{\mathbf{U}^t-\gamma_t\mathrm{grad}f|_{\mathbf{U}}\right\}, \tag{76}$$

and the retraction through the Q factor of the the QR decomposition [18] (denoted $\mathrm{qf}\{\cdot\}$):

$$\mathbf{U}^{t+1}\leftarrow\mathrm{qf}\left\{\mathbf{U}^t-\gamma_t\mathrm{grad}f|_{\mathbf{U}}\right\}, \tag{77}$$

where γ_t is a step size, and $\mathrm{grad}f|_{\mathbf{U}}$ is the Riemmanian gradient of f at $\mathbf{U}$. The step size is computed at each iteration with the standard Armijo rule. We also compare an implementation of [23] with a constant step size (selected as large as possible, while ensuring convergence); ii) MM with quadratic majorization: as discussed in section V-C, an iteration of this algorithm is solved either with the SVD of $\mathbf{M}(\mathbf{U}^t)$ in (55) (fixed point [13]), or the SVD of appropriately normalized samples (FMS [15]). In the considered scenarios $n<p$ so the second option is chosen, as it is more computationally efficient; iii) MM with linear majorization (Algorithm 2): this algorithm is detailed in (56) and also corresponds to the fixed point iterations referred to as R1-PCA in [14].

Figure 2 displays the convergence of the algorithms in terms of objective value for different scenarios with varying k ($p=1000$ and $n=p/2$). The overall performance is compared both in terms of the number of iterations and the computation time. In terms of required number of iterations, the MM with quadratic majorization appears faster. This is due to the fact

that it exploits a tighter upperbound than the MM with linear majorization (cf. Section V-C). The speed of the steepest descents could be improved by the use of an exhaustive line search. Still, the Armijo rule offers a significant improvement compared to the constant step size. When it comes to the computation time, the MM with linear majorization appears faster. Indeed, each iteration of this algorithm requires only the computation of the thin-SVD of an $p \times k$ matrix, rather than the k strongest eigenvectors of a $p \times n$ matrix for the MM with quadratic majorization. The steepest descent the Armijo rule requires to run a line search for the step size. This involves evaluating the objective, therefore calls for multiple retractions, which is computationally more intense. The constant step size iterations are fast to compute, which can compensate its slow convergence in terms of number of iterations. Nevertheless, the step size had to be manually tuned for each scenario. Thus, this algorithm corresponds to a benchmark with little practical application

Figure 3 displays total computation time, i.e., time to reach the stopping criterion $||\mathbf{U}^t - \mathbf{U}^{t+1}||_\infty < 10^{-10}$, with respect to the problem dimensions (scaled proportionally to to p). As previously, we can notice that the MM with linear majorization offers an interesting solution in terms of scalability.

B. On RSPCA algorithm

1) Validation on synthetic data: In this section, we validate several properties of the RSPCA method on synthetic examples. In order to clarify the exposition, the term RSPCA will here denote the solution of problem (72) computed with Algorithm 4. We will focus on the use of the following parameters[4]: the RSR cost is ρ_{H} (cf. Example 2) with $T = 0.1$, the sparse penalty is either r_0 or $r_{2,0}$, where l_ϵ^γ is from the proxy of the ℓ_1-norm (i.e., a)) in the familly of proposed proxies), with $\epsilon = 10^{-2}$ and $\omega_r = 1 \forall [\![\in 1, k]\!]$, and varying regularization parameter λ. the dimensions are $p = 100$, $k = 15$.

First, we study the performance of RSPCA in a standard PCA setting. Samples are drawn from $\mathbf{z} \sim \mathcal{CN}(\mathbf{0}, \mathrm{SNR} \times \mathbf{U}\mathbf{U}^H + \mathbf{I})$, where $\mathbf{U}$ are the first k eigenvectors of a the Topelitz matrix $[\mathbf{\Sigma}_T]_{i,j} = (\theta(1+i)/\sqrt{2})^{|i-j|}$ and $\mathrm{SNR} = 10$. For a given estimator $\hat{\mathbf{U}}$, the performance criterion is the average fraction of recovered energy:

$$\mathrm{AFE} = \mathbb{E}\left[\mathrm{Tr}\{\hat{\mathbf{U}}^H \mathbf{U}\mathbf{U}^H \hat{\mathbf{U}}\}/k\right], \tag{78}$$

which is evaluated through 10^2 Monte-Carlo simulations. Figure 4 (resp. 5) displays the AFE of RSPCA with r_0 (resp. $r_{2,0}$) penalty in function of n and for various λ. Note that without regularization ($\lambda = 0$) the RSPCA corresponds to RSR with the Huber cost [14, 15]. In a Gaussian setting, it is interesting to note that there is little loss due to the usage of the Huber cost function when comparing with the standard PCA, which corresponds to the maximum likelihood estimator in this case [1]. If a sparse regularization penalty is involved, two regimes can be observed. When the underlying subspace basis is actually sparse ($\theta = 0$ leads to canonical eigenvectors), this penalty significantly improves the recovery. When the underlying subspace basis is not sparse ($\theta = 0.5$), the recovery can be severely impacted by a strong penalty promoting this property. However, we can notice that RSPCA appears robust to the mismatch when λ is set low enough.

Second, we study the robustness of RSPCA to corruption by outliers in the sample set. The simulation setup is built around the so-called *haystack* model [50], which corresponds here to a mixture of orthogonal Gaussian distributions plus additive noise:

$$\begin{aligned} \{\mathbf{z}_i\}_{i=1}^n &= \{\{\mathbf{z}_i^{\mathrm{in}}\}_{i=1}^{n_{\mathrm{in}}}, \{\mathbf{z}_i^{\mathrm{out}}\}_{i=n_{\mathrm{in}}+1}^n\} \\ \mathbf{z}^{\mathrm{in}} &\sim \mathcal{CN}(\mathbf{0}, \mathrm{SNR} \times \mathbf{U}\mathbf{U}^H + \mathbf{I}) \\ \mathbf{z}^{\mathrm{out}} &\sim \mathcal{CN}(\mathbf{0}, \mathrm{ONR} \times \mathbf{U}_\perp \mathbf{U}_\perp^H + \mathbf{I}), \end{aligned} \tag{79}$$

where $\mathbf{U}$ and $\mathbf{U}_\perp$ are built from the canonical basis such that $\mathbf{U}^H \mathbf{U}_\perp = \mathbf{0}$, SNR is signal to noise ratio, and ONR is outlier to noise ratio. We will compare four algorithms: i) RSPCA with least square fitting (cf. Example 1) and $\lambda = 0$ (i.e., the standard PCA); ii) RSPCA with least square fitting and $\lambda = 100$; iv) RSPCA with Huber cost (cf. Example 2) and $\lambda = 0$ (i.e., RSR); iii) RSPCA with Huber cost and $\lambda = 1000$. The aim is to illustrate the contributions of both the robust fitting objective and the sparsity promoting penalty. Figure 6 displays the AFE of each algorithm with respect to ONR and the fraction of outliers in the sample set. We can notice that the use of a robust cost improves the performance compared to the standard PCA. Moreover, the introduction of the sparse penalty improves the results in terms of AFE (as seen in Figures 4-5), but also interestingly improves the robustness of the estimation process. Here, it is worth mentioning two critical points: a) RSPCA appears very robust when a valid signal subspace basis is actually sparse, which is probably because the sparse penalty contributes to naturally discard dense outliers. If the true subspace basis is dense, results can be degraded in practice (as also observed in Figures 4-5). b) The starting point plays an important role in the achieved robustness of all iterative algorithms (even without regularization). In these simulations, we used the spherical PCA (PCA applied on the normalized samples) as starting point. It has been observed that the achieved robustness can be lowered by using the standard PCA instead.

2) Experiment on real data: In this experiment, we compare the performance of RSPCA with state of the art sparse PCA algorithms on the Leukemia data set [51][5]. RSPCA build with a GMC cost (cf. Example 4), r_0 penalty and l_ϵ^γ is from the proxy of the lower bound of the sign function (i.e., c) in the familly of proposed proxies). Again, similar conclusions can be drawn with other objectives, up to minor changes of the parameters. RSPCA is computed with Algorithm 4 using an outer loop, decreasing ϵ from 10^{-1} to 10^{-7} in order to avoid potential local minimums. This algorithm is compared to ALSPCA [52], SPCArt and rSVD-GP from [6]. The studied

[4]We simply note that the general conclusions drawn from these examples can also be obtained with the other robust-fitting/sparse-promoting function presented in this paper, up to a good selection of the tuning parameters

[5]From https://github.com/ramhiser/datamicroarray/wiki/Golub-(1999)

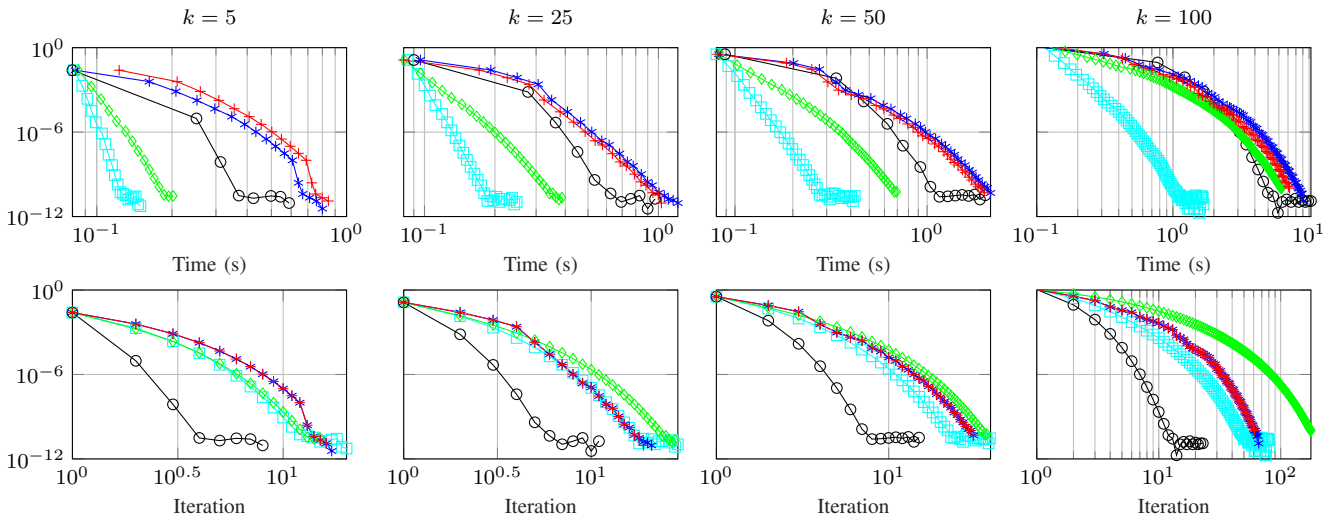

Fig. 2. Gap to minimal value (attained by any converging algorithm) versus time and number of iterations for various algorithms called on problem (49): Steepest descent using Euclidean projection retraction with constant step size (◇) or Armijo rule (∗). Steepest descent using QR retraction with Armijo rule (+). MM with quadratic majorizer (○). MM with linear majorizer (□). Simulation carried out on MATLAB R2019a with Intel(R) Core(TM) i7-8850H CPU 2.60GHz processor.

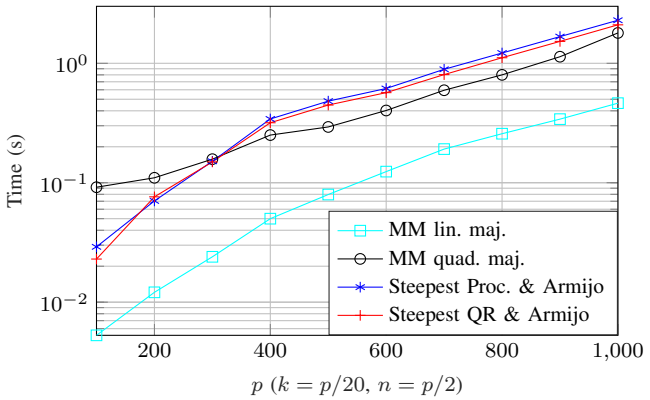

Fig. 3. Time to reach the convergence criterion $||\mathbf{U}^t - \mathbf{U}^{t+1}||_\infty < 10^{-10}$ versus dimension p (and fixed ratios $k = p/20$, $n = p/2$) for various algorithms called on problem (49).

performance criteria are the following

$$\begin{aligned} \mathrm{SP}(\hat{\mathbf{U}}) &= 1 - ||\hat{\mathbf{U}}||_0/(pk) \\ \widetilde{\mathrm{AFE}}(\hat{\mathbf{U}}) &= \mathrm{AFE}(\mathrm{span}(\hat{\mathbf{U}})) \\ \mathrm{NOR}(\hat{\mathbf{U}}) &= ||\hat{\mathbf{U}}^H\hat{\mathbf{U}} - \mathbf{I}||_F^2 \end{aligned} \tag{80}$$

which measure respectively the sparsity (expected to be close to 1, i.e., 100%), the explained variance (expected to be close to 1, i.e., 100%), and the non-orthonormality (expected to be low). Notice that the $\widetilde{\mathrm{AFE}}$ criterion slightly favors the algorithms that relax the orthonormality constraint in this case. Figure 7 displays the $\widetilde{\mathrm{AFE}}$ and NOR versus SP for the studied sparse PCA algorithms on the Leukemia data set. Interestingly, we can notice that RSPCA achieves state of the art performance when it comes to the explained variance-sparsity trade-off, but without relaxing the orthonormality constraint, as done by the other algorithms.

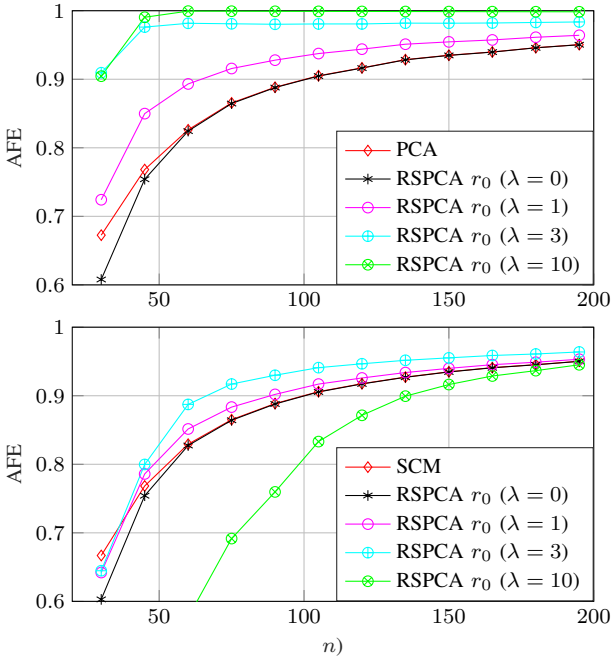

Fig. 4. AFE versus n for various algorithms in a standard PCA context. $p = 100$, $k = 10$, SNR $= 10$. RPSCA built with r_0 penalty.

VIII. Conclusion

This paper presented an optimization framework to deal with orthonormality constraints based on Majorization-Minimization. The core idea is to systematically obtain linear surrogates in order to formulate iterations as Euclidean projection problems. Such majorizers can be obtained for a large family of functions for which we presented a catalog. We also exhibited some links between this approach and well-known algorithms. Finally, the presented framework drove

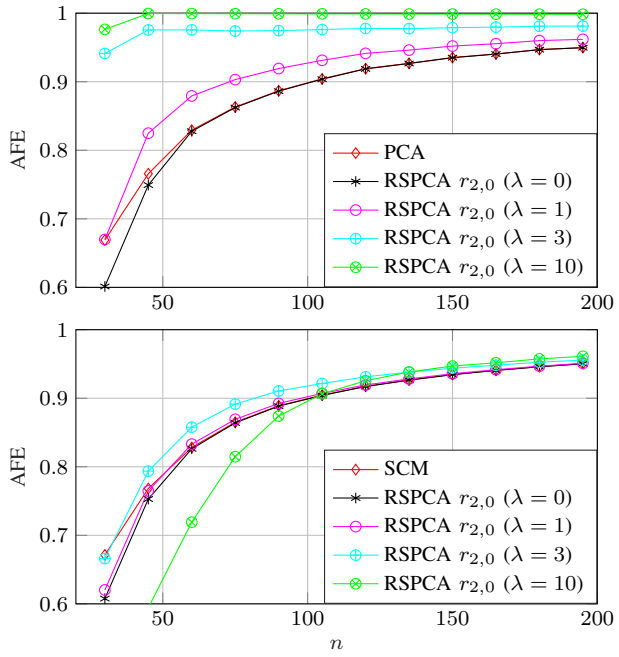

Fig. 5. AFE versus n for various algorithms in a standard PCA context. $p = 100$, $k = 10$, SNR $= 10$. RPSCA built with $r_{2,0}$ penalty.

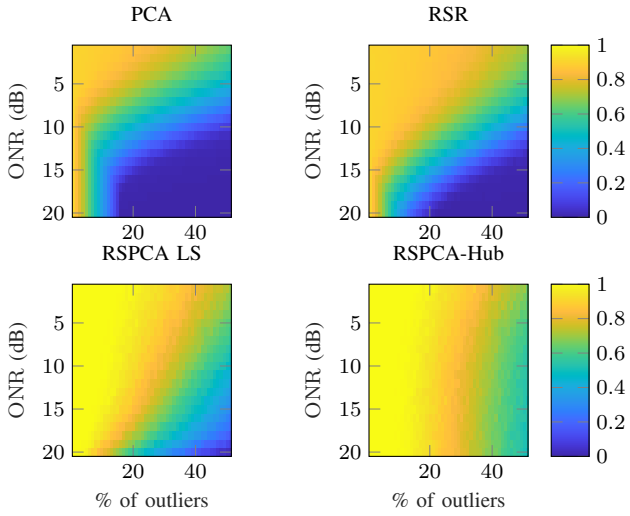

Fig. 6. AFE versus ONR and number of outliers for various algorithms. $p = 100$, $k = 15$, $n = p$, SNR $= 10$. RSPCA built with r_0 penalty.

the formulation of a novel sparse PCA algorithm (namely, RSPCA), combining a robust subspace recovery cost and sparsity promoting penalties. Several experiments illustrated the interest of the approach, both in terms of computational and estimation performance.

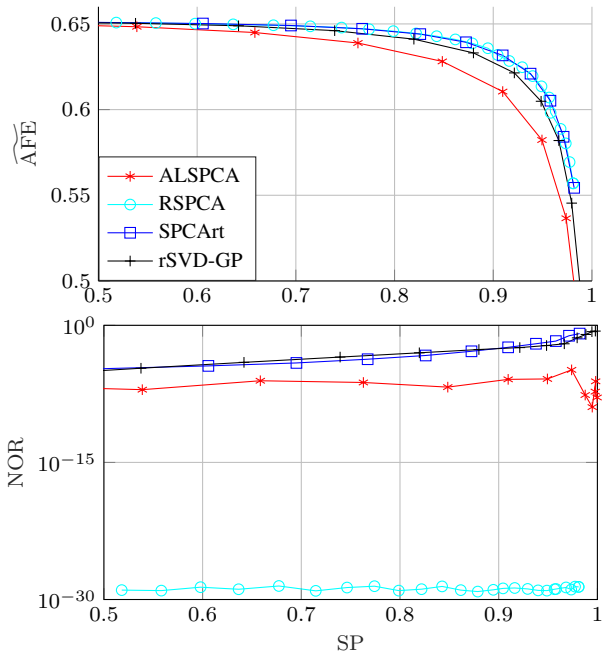

Fig. 7. $\widetilde{\text{AFE}}$ and NOR versus SP for various sparse PCA algorithms on the Leukemia data set. $p = 7129$, $k = 10$, $n = 72$.

References

[1] I. T. Jolliffe, *Principal Component Analysis and Factor Analysis*. Springer, 1986.

[2] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999.

[3] K. Chen, K.-S. Chan, and N. C. Stenseth, "Reduced rank stochastic regression with a sparse singular value decomposition," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 74, no. 2, pp. 203–221, 2012.

[4] L. Chen and J. Z. Huang, "Sparse reduced-rank regression for simultaneous dimension reduction and variable selection," *Journal of the American Statistical Association*, vol. 107, no. 500, pp. 1533–1545, 2012.

[5] F. Bunea, Y. She, and M. H. Wegkamp, "Joint variable and rank selection for parsimonious estimation of high-dimensional matrices," *The Annals of Statistics*, vol. 40, no. 5, pp. 2359–2388, 2012.

[6] Z. Hu, G. Pan, Y. Wang, and Z. Wu, "Sparse principal component analysis via rotation and truncation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 4, pp. 875–890, 2016.

[7] K. Benidis, Y. Sun, P. Babu, and D. P. Palomar, "Orthogonal sparse PCA and covariance estimation via Procrustes reformulation," *IEEE Transactions on Signal Processing*, vol. 64, no. 23, pp. 6211–6226, 2016.

[8] Y. Uematsu, Y. Fan, K. Chen, J. Lv, and W. Lin, "SOFAR: Large-scale association network learning," *IEEE Transactions on Information Theory*, vol. 65, no. 8, pp. 4924–4939, 2019.

[9] Y. Sun, A. Breloy, P. Babu, D. P. Palomar, F. Pascal, and G. Ginolhac, "Low-complexity algorithms for low rank clutter parameters estimation in radar systems," *IEEE Transactions on Signal Processing*, vol. 64, no. 8, pp. 1986–1998, April 2016.

[10] Y. Sun, P. Babu, and D. P. Palomar, "Robust estimation of structured covariance matrix for heavy-tailed elliptical distributions," *IEEE Transactions on Signal Processing*, vol. 64, no. 14, pp. 3576–3590, 2016.

[11] G. Lerman and T. Maunu, "An overview of robust subspace recovery," *Proceedings of the IEEE*, vol. 106, no. 8, pp. 1380–1410, 2018.

[12] F. De La Torre and M. J. Black, "A framework for robust subspace learning," *Intl. J. of Computer Vision*, vol. 54, no. 1-3, pp. 117–142, 2003.

[13] R. Maronna, "Principal components and orthogonal regression based on robust scales," *Technometrics*, vol. 47, no. 3, pp. 264–273, 2005. [Online]. Available: http://www.jstor.org/stable/25471020

[14] C. Ding, D. Zhou, X. He, and H. Zha, "R1-PCA: rotational invariant L1-norm principal component analysis for robust subspace factorization," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 281–288.

[15] G. Lerman and T. Maunu, "Fast, robust and non-convex subspace recovery," *Information and Inference: A Journal of the IMA*, vol. 7, no. 2, pp. 277–336, 2017.

[16] T. Maunu, T. Zhang, and G. Lerman, "A well-tempered landscape for non-convex robust subspace recovery." *Journal of Machine Learning Research*, vol. 20, no. 37, pp. 1–59, 2019.

[17] R. Vidal, "Subspace clustering," *IEEE Signal Processing Magazine*,

vol. 28, no. 2, pp. 52–68, 2011.
[18] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.
[19] N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre, "Manopt, a matlab toolbox for optimization on manifolds," *Journal of Machine Learning Research*, vol. 15, pp. 1455–1459, 2014. [Online]. Available: http://jmlr.org/papers/v15/boumal14a.html
[20] A. Edelman, T. Arias, and S. Smith, "The geometry of algorithms with orthogonality constraints," *SIAM Journal on Matrix Analysis and Applications*, vol. 20, no. 2, pp. 303–353, 1998.
[21] T. Abrudan, J. Eriksson, and V. Koivunen, "Steepest descent algorithms for optimization under unitary matrix constraint," *IEEE Transactions on Signal Processing*, vol. 56, no. 3, pp. 1134–1147, March 2008.
[22] ——, "Conjugate gradient algorithm for optimization under unitary matrix constraint," *Signal Processing*, vol. 89, no. 9, pp. 1704–1714, 2009.
[23] J. Manton, "Optimization algorithms exploiting unitary constraints," *IEEE Transactions on Signal Processing*, vol. 50, no. 3, pp. 635–650, Mar 2002.
[24] Y. Sun, P. Babu, and D. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Transactions on Signal Processing*, vol. PP, no. 99, pp. 1–1, 2016.
[25] P. H. Schönemann, "A generalized solution of the orthogonal Procrustes problem," *Psychometrika*, vol. 31, no. 1, pp. 1–10, 1966.
[26] L. Eldén and H. Park, "A Procrustes problem on the stiefel manifold," *Numerische Mathematik*, vol. 82, no. 4, pp. 599–619, 1999.
[27] M. A. Koschat and D. F. Swayne, "A weighted Procrustes criterion," *Psychometrika*, vol. 56, no. 2, pp. 229–239, 1991.
[28] H. A. Kiers, "Setting up alternating least squares and iterative majorization algorithms for solving various matrix optimization problems," *Computational statistics & data analysis*, vol. 41, no. 1, pp. 157–170, 2002.
[29] A. Breloy, Y. Sun, P. Babu, G. Ginolhac, and D. P. Palomar, "Robust rank constrained kronecker covariance matrix estimation," in *50th Asilomar Conference on Signals, Systems and Computers*, 2016, pp. 810–814.
[30] M. Razaviyayn, M. Hong, and Z. Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM on Optimization*, vol. 23, no. 2, pp. 1126–1153, 2013.
[31] J. Song, P. Babu, and D. P. Palomar, "Sparse generalized eigenvalue problem via smooth optimization," *IEEE Transactions on Signal Processing*, vol. 63, no. 7, pp. 1627–1642, April 2015.
[32] X. Fu, K. Huang, M. Hong, N. D. Sidiropoulos, and A. M. So, "Scalable and flexible multiview MAX-VAR canonical correlation analysis," *IEEE Transactions on Signal Processing*, vol. 65, no. 16, pp. 4150–4165, 2017.
[33] S. Kumar, J. Ying, J. V. de M. Cardoso, and D. P. Palomar, "A unified framework for structured graph learning via spectral constraints," *Journal of Machine Learning Research*, vol. 21, no. 22, pp. 1–60, 2020. [Online]. Available: http://jmlr.org/papers/v21/19-276.html
[34] A. Breloy, G. Ginolhac, F. Pascal, and P. Forster, "Clutter subspace estimation in low rank heterogeneous noise context," *IEEE Transactions on Signal Processing*, vol. 63, no. 9, pp. 2173–2182, May 2015.
[35] H. A. L. Kiers, "Maximization of sums of quotients of quadratic forms and some generalizations," *Psychometrika*, vol. 60, no. 2, pp. 221–245, Jun 1995.
[36] M. Hong, M. Razaviyayn, Z.-Q. Luo, and J.-S. Pang, "A unified algorithmic framework for block-structured optimization involving big data: With applications in machine learning and signal processing," *IEEE Signal Processing Magazine*, vol. 33, no. 1, pp. 57–77, 2015.
[37] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2003.
[38] G. H. Golub and C. F. Van Loan, *Matrix Computations (3rd Ed.)*. Baltimore, MD, USA: Johns Hopkins University Press, 1996.
[39] R. Ben Abdallah, A. Breloy, M. El Korso, and D. Lautru, "Bayesian signal subspace estimation with compound gaussian sources," *Signal Processing*, vol. 167, p. 107310, 2020.
[40] M. Frank, P. Wolfe *et al.*, "An algorithm for quadratic programming," *Naval research logistics quarterly*, vol. 3, no. 1-2, pp. 95–110, 1956.
[41] M. Jaggi, "Revisiting frank-wolfe: Projection-free sparse convex optimization," in *Proceedings of the 30th international conference on machine learning*, 2013, pp. 427–435.
[42] M. Weber and S. Sra, "Nonconvex stochastic optimization on manifolds via riemannian frank-wolfe methods," *arXiv preprint arXiv:1910.04194*, 2019.
[43] M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre, "Generalized power method for sparse principal component analysis." *Journal of Machine Learning Research*, vol. 11, no. 2, 2010.
[44] I. F. Gorodnitsky and B. D. Rao, "Sparse signal reconstruction from limited data using focuss: A re-weighted minimum norm algorithm," *IEEE Transactions on signal processing*, vol. 45, no. 3, pp. 600–616, 1997.
[45] R. Chartrand and Wotao Yin, "Iteratively reweighted algorithms for compressive sensing," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 3869–3872.
[46] M. Lai, Y. Xu, and W. Yin, "Improved iteratively reweighted least squares for unconstrained smoothed ℓ_q minimization," *SIAM Journal on Numerical Analysis*, vol. 51, no. 2, pp. 927–957, 2013.
[47] B. K. Sriperumbudur, D. A. Torres, and G. R. Lanckriet, "A majorization-minimization approach to the sparse generalized eigenvalue problem," *Machine learning*, vol. 85, no. 1, pp. 3–39, 2011.
[48] E. J. Candès, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 877–905, Dec 2008.
[49] O. Mangasarian, "Machine learning via polyhedral concave minimization," in *Applied Mathematics and parallel computing*. Springer, 1996, pp. 175–188.
[50] G. Lerman, M. B. McCoy, J. A. Tropp, and T. Zhang, "Robust computation of linear models by convex relaxation," *Foundations of Computational Mathematics*, vol. 15, no. 2, pp. 363–410, 2015.
[51] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring." *Science*, vol. 286, no. 5439, pp. 531–537, Oct. 1999.
[52] Z. Lu and Y. Zhang, "An augmented lagrangian approach for sparse principal component analysis," *Mathematical Programming*, vol. 135, no. 1-2, pp. 149–193, 2012.