HAL
open science

A proximal approach for a class of matrix optimization problems

Alessandro Benfenati, Emilie Chouzenoux, Jean-Christophe Pesquet

To cite this version:

Alessandro Benfenati, Emilie Chouzenoux, Jean-Christophe Pesquet. A proximal approach for a class of matrix optimization problems. 2017. hal-01673027

HAL Id: hal-01673027

https://hal.science/hal-01673027

Preprint submitted on 28 Dec 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A PROXIMAL APPROACH FOR A CLASS OF MATRIX OPTIMIZATION PROBLEMS*

ALESSANDRO BENFENATI†, EMILIE CHOUZENOUX‡, AND JEAN–CHRISTOPHE PESQUET‡

Abstract. In recent years, there has been a growing interest in mathematical models leading to the minimization, in a symmetric matrix space, of a Bregman divergence coupled with a regularization term. We address problems of this type within a general framework where the regularization term is split in two parts, one being a spectral function while the other is arbitrary. A Douglas–Rachford approach is proposed to address such problems and a list of proximity operators is provided allowing us to consider various choices for the fit–to–data functional and for the regularization term. Numerical experiments show the validity of this approach for solving convex optimization problems encountered in the context of sparse covariance matrix estimation. Based on our theoretical results, an algorithm is also proposed for noisy graphical lasso where a precision matrix has to be estimated in the presence of noise. The nonconvexity of the resulting objective function is dealt with a majorization–minimization approach, i.e. by building a sequence of convex surrogates and solving the inner optimization subproblems via the aforementioned Douglas–Rachford procedure. We establish conditions for the convergence of this iterative scheme and we illustrate its good numerical performance with respect to state–of–the–art approaches.

Key words. Covariance estimation, Graphical Lasso, matrix optimization, Douglas-Rachford method, majorization-minimization, Bregman divergence

AMS subject classifications. 15A18, 15B48, 62J10, 65K10, 90C06, 90C25, 90C26, 90C35

1. Introduction.

In recent years, various applications such as shape classification models [30], gene expression [44], model selection [3, 18], computer vision [33], inverse covariance estimation [31, 29, 68, 28, 62], graph estimation [48, 53, 67], social network and corporate inter-relationships analysis [2], or brain network analysis [65] have led to matrix variational formulations of the form:

$$\underset{\mathbf{C}\in\mathcal{S}_n}{\text{minimize}}\ f(\mathbf{C}) - \operatorname{trace}(\mathbf{TC}) + g(\mathbf{C}), \tag{1}$$

where $\mathcal{S}_n$ is the set of real symmetric matrices of dimension $n \times n$, $\mathbf{T}$ is a given $n \times n$ real matrix (without loss of generality, it will be assumed to be symmetric), and $f\colon \mathcal{S}_n \to]-\infty, +\infty]$ and $g\colon : \mathcal{S}_n \to]-\infty, +\infty]$ are lower-semicontinuous functions which are proper, in the sense that they are finite at least in one point.
It is worth noticing that the notion of Bregman divergence [13] gives a particular insight into Problem (1). Indeed, suppose that f is a convex function differentiable on the interior of its domain $\operatorname{int}(\operatorname{dom} f) \neq \varnothing$. Let us recall that, in $\mathcal{S}_n$ endowed with the Frobenius norm, the f-Bregman divergence between $\mathbf{C} \in \mathcal{S}_n$ and $\mathbf{Y} \in \operatorname{int}(\operatorname{dom} f)$ is

$$D^f(\mathbf{C}, \mathbf{Y}) = f(\mathbf{C}) - f(\mathbf{Y}) - \operatorname{trace}(\mathbf{T}(\mathbf{C} - \mathbf{Y})), \tag{2}$$

*Submitted to the editors DATE.

Funding: This work was funded by the Agence Nationale de la Recherche under grant ANR-14-CE27-0001 GRAPHSIP.

†Laboratoire d'Informatique Gaspard Monge, ESIEE Paris, University Paris-Est, FR (alessandro.benfenati@esiee.fr).

‡Center for Visual Computing, INRIA Saclay and CentraleSupélec, University Paris-Saclay, FR (emilie.chouzenoux@centralesupelec.fr, jean-christophe@pesquet.eu).

This manuscript is for review purposes only.

where $\mathbf{T} = \nabla f(\mathbf{Y})$ is the gradient of f at $\mathbf{Y}$. Hence, the original problem (1) is equivalently expressed as

$$\underset{\mathbf{C}\in\mathcal{S}_n}{\text{minimize}}\ g(\mathbf{C}) + D^f(\mathbf{C},\mathbf{Y}). \tag{3}$$

Solving Problem (3) amounts to computing the proximity operator of g at $\mathbf{Y}$ with respect to the divergence D^f [5, 7] in the space $\mathcal{S}_n$. In the vector case, such kind of proximity operator has been found to be useful in a number of recent works regarding, for example, image restoration [14, 8, 9, 70], image reconstruction [71], and compressive sensing problems [66, 32].

In this paper, it will be assumed that f belongs to the class of *spectral functions* [11, Chapter 5, Section 2], i.e., for every permutation matrix $\mathbf{\Sigma} \in \mathbb{R}^{n\times n}$,

$$(\forall \mathbf{C} \in \mathcal{S}_n) \quad f(\mathbf{C}) = \varphi(\mathbf{\Sigma}\mathbf{d}), \tag{4}$$

where $\varphi\colon \mathbb{R}^n \to]-\infty,+\infty]$ is a proper lower semi-continuous convex function and $\mathbf{d}$ is a vector of eigenvalues of $\mathbf{C}$.

Due to the nature of the problems, in many of the aforementioned applications, g is a regularization function promoting the sparsity of $\mathbf{C}$. We consider here a more generic class of regularization functions obtained by decomposing g as $g_0 + g_1$, where g_0 is a spectral function, i.e., for every permutation matrix $\mathbf{\Sigma} \in \mathbb{R}^{n\times n}$,

$$(\forall \mathbf{C} \in \mathcal{S}_n) \quad g_0(\mathbf{C}) = \psi(\mathbf{\Sigma}\mathbf{d}), \tag{5}$$

with $\psi\colon \mathbb{R}^n \to]-\infty,+\infty]$ a proper lower semi–continuous function, $\mathbf{d}$ still denoting a vector of the eigenvalues of $\mathbf{C}$, while $g_1\colon \mathcal{S}_n \to]-\infty,+\infty]$ is a proper lower semi–continuous function which cannot be expressed under a spectral form.

A very popular and useful example encompassed by our framework is the graphical lasso (GLASSO) problem, where f is the minus log-determinant function, g_1 is a component–wise ℓ_1 norm (of the matrix elements), and $g_0 \equiv 0$. Various algorithms have been proposed to solve Problem (1) in this context, including the popular GLASSO algorithm [31] and some of its recent variants [47]. We can also mention the dual block coordinate ascent method from [3], the SPICE algorithm [57], the gradient projection method in [30], the Refitted CLIME algorithm [17], various algorithms [28, 42, 43] based on Nesterov's smooth gradient approach [50], ADMM approaches [68, 58], an inexact Newton method [62], and interior point methods [67, 40]. A related model is addressed in [44, 18], with the additional assumption that the sought solution can be split as $\mathbf{C}_1 + \mathbf{C}_2$, where $\mathbf{C}_1$ is sparse and $\mathbf{C}_2$ is low–rank. Finally, let us mention the ADMM algorithm from [72], and the incremental proximal gradient approach from [54], both addressing Problem (1) when f is the squared Frobenius norm, g_0 is a nuclear norm, and g_1 is an element–wise ℓ_1 norm.

The main goal of this paper is to propose numerical approaches for solving Problem (1). Two settings will be investigated, namely (i) $g_1 \equiv 0$, i.e. the whole cost function is a spectral one, (ii) $g_1 \not\equiv 0$. In the former case, some general results concerning the D^f-proximity operator of g_0 are established. In the latter case, a Douglas–Rachford optimization method is proposed, which leads us to calculate the proximity operators of several spectral functions of interest. We then consider applications of our results to the estimation of (possibly low-rank) covariance matrices from noisy observations of multivalued random variables. Two variational approaches are proposed for estimating the unknown covariance matrix, depending on the prior assumptions made on it. We show that the cost function arising from the first formulation can be minimized through our proposed Douglas-Rachford procedure under

mild assumptions on the involved regularization functions. The second formulation of the problem aims at preserving desirable sparsity properties of the inverse covariance (i.e., precision) matrix. We establish that the proposed objective function is a difference of convex terms, and we introduce a novel majorization-minimization (MM) algorithm to optimize it.

The paper is organized as follows. Section 2 is devoted to the solution of the particular instance of Problem (1) corresponding to $g_1 \equiv 0$. Section 3 describes a a proximal to address the problem when $g_1 \not\equiv 0$. Its implementation is discussed for a bunch of useful choices for the involved functionals. Section 4 presents two new approaches for estimating covariance matrices from noisy data. Finally, in Section 5, numerical experiments illustrate the applicability of the proposed methods, and its good performance with respect to the state-of-the-art, in two distinct scenarios.

Notation: Greek letters usually designate real numbers, bold letters designate vectors in a Euclidean space, capital bold letters indicate matrices. The i–th element of the vector $\mathbf{d}$ is denoted by d_i. $\mathrm{Diag}(\mathbf{d})$ denotes the diagonal matrix whose diagonal elements are the components of $\mathbf{d}$. $\mathcal{D}_n$ is the cone of vectors $\mathbf{d} \in \mathbb{R}^n$ whose components are ordered by decreasing values. The symbol $\mathrm{vect}(\mathbf{C})$ denotes the vector resulting from a column–wise ordering of the elements of matrix $\mathbf{C}$. The product $\mathbf{A} \otimes \mathbf{B}$ denotes the classical Kronecker product of matrices $\mathbf{A}$ and $\mathbf{B}$. Let $\mathcal{H}$ be a real Hilbert space endowed with an inner product $\langle \cdot, \cdot \rangle$ and a norm $\|\cdot\|$, the domain of a function $f \colon \mathcal{H} \to]-\infty, +\infty]$ is $\mathrm{dom}\, f = \{x \in \mathcal{H} \mid f(x) < +\infty\}$. f is coercive if $\lim_{\|x\| \to +\infty} f(x) = +\infty$ and supercoercive if $\lim_{\|x\| \to +\infty} f(x)/\|x\| = +\infty$. The Moreau subdifferential of f at $x \in \mathcal{H}$ is $\partial f(x) = \{t \in \mathcal{H} \mid (\forall y \in \mathcal{H}) f(y) \geqslant f(x) + \langle t, y - x \rangle\}$. $\Gamma_0(\mathcal{H})$ denotes the class of lower-semicontinuous convex functions from $\mathcal{H}$ to $]-\infty, +\infty]$ with a nonempty domain (proper). If $f \in \Gamma_0(\mathcal{H})$ is (Gâteaux) differentiable at $x \in \mathcal{H}$, then $\partial f(x) = \{\nabla f(x)\}$ where $\nabla f(x)$ is the gradient of f at x. If a function $f \colon \mathcal{H} \to]-\infty, +\infty]$ possesses a unique minimizer on a set $E \subset \mathcal{H}$, it will be denoted by $\underset{x \in E}{\mathrm{argmin}}\, f(x)$. If there are possibly several minimizers, their set will be denoted by $\underset{x \in E}{\mathrm{Argmin}}\, f(x)$. Given a set E, $\mathrm{int}(E)$ designates the interior of E and ι_E denotes the indicator function of the set, which is equal to 0 over this set and $+\infty$ otherwise. In the remainder of the paper, the underlying Hilbert space will be $\mathcal{S}_n$, the set of real symmetric matrices equipped with the Frobenius norm, denoted by $\|\cdot\|_{\mathrm{F}}$. The matrix spectral norm is denoted by $\|\cdot\|_{\mathrm{S}}$, the ℓ_1 norm of a matrix $\mathbf{A} = (A_{i,j})_{i,j}$ is $\|\mathbf{A}\|_1 = \sum_{i,j} |A_{i,j}|$. For every $p \in [1, +\infty[$, $\mathcal{R}_p(\cdot)$ denotes the Schatten p–norm, the nuclear norm being obtained when $p = 1$. $\mathcal{O}_n$ denotes the set of orthogonal matrices of dimension n with real elements; $\mathcal{S}_n^+$ and $\mathcal{S}_n^{++}$ denote the set of real symmetric positive semidefinite, and symmetric positive definite matrices, respectively, of dimension n. $\mathrm{I_d}$ denotes the identity matrix whose dimension will be clear from the context. The soft thresholding operator soft_μ and the hard thresholding operator hard_μ of parameter $\mu \in [0, +\infty[$ are given by

$$(\forall \xi \in \mathbb{R}) \qquad \mathrm{soft}_\mu(\xi) = \begin{cases} \xi - \mu & \text{if } \xi > \mu \\ \xi + \mu & \text{if } \xi < -\mu \\ 0 & \text{otherwise} \end{cases}, \qquad \mathrm{hard}_\mu(\xi) = \begin{cases} \xi & \text{if } |\xi| > \mu \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

2. Spectral Approach. In this section, we show that, in the particular case when $g_1 \equiv 0$, Problem (1) reduces to the optimization of a function defined on $\mathbb{R}^n$. Indeed, the problem then reads:

$$\underset{\mathbf{C} \in \mathcal{S}_n}{\mathrm{minimize}}\; f(\mathbf{C}) - \mathrm{trace}\,(\mathbf{T}\mathbf{C}) + g_0(\mathbf{C}), \tag{7}$$

where the spectral forms of f and g_0 allow us to take advantage of the eigendecompositions of $\mathbf{C}$ and $\mathbf{T}$ in order to simplify the optimization problem, as stated below.

THEOREM 2.1. *Let $\mathbf{t} \in \mathbb{R}^n$ be a vector of eigenvalues of $\mathbf{T}$ and let $\mathbf{U_T} \in \mathcal{O}_n$ be such that $\mathbf{T} = \mathbf{U}_T \operatorname{Diag}(\mathbf{t})\mathbf{U}_T^\top$. Let f and g_0 be functions satisfying (4) and (5), respectively, where φ and ψ are lower-semicontinuous functions. Assume that $\operatorname{dom}\varphi \cap \operatorname{dom}\psi \neq \varnothing$ and that the function $\mathbf{d} \mapsto \varphi(\mathbf{d}) - \mathbf{d}^\top\mathbf{t} + \psi(\mathbf{d})$ is coercive. Then a solution to Problem (7) exists, which is given by*

$$\widehat{\mathbf{C}} = \mathbf{U_T} \operatorname{Diag}(\widehat{\mathbf{d}})\mathbf{U_T}^\top \tag{8}$$

where $\widehat{\mathbf{d}}$ is any solution to the following problem:

$$\underset{\mathbf{d}\in\mathbb{R}^n}{\text{minimize}}\ \varphi(\mathbf{d}) - \mathbf{d}^\top\mathbf{t} + \psi(\mathbf{d}). \tag{9}$$

For the sake of clarity, before establishing this result, we recall two useful lemmas from linear algebra.

LEMMA 2.2. [46, Chapter 9, Sec. H, p. 340] *Let $\mathbf{C} \in \mathcal{S}_n$ and let $\mathbf{d} \in \mathcal{D}_n$ be a vector of ordered eigenvalues of this matrix. Let $\mathbf{T} \in \mathcal{S}_n$ and let $\mathbf{t} \in \mathcal{D}_n$ be a vector of ordered eigenvalues of this matrix. The following inequality holds:*

$$\operatorname{trace}(\mathbf{CT}) \leqslant \mathbf{d}^\top\mathbf{t}. \tag{10}$$

In addition, the upper bound is reached if and only if $\mathbf{T}$ and $\mathbf{C}$ share the same eigenbasis, i.e. there exists $\mathbf{U} \in \mathcal{O}_n$ such that $\mathbf{C} = \mathbf{U}\operatorname{Diag}(\mathbf{d})\mathbf{U}^\top$ and $\mathbf{T} = \mathbf{U}\operatorname{Diag}(\mathbf{t})\mathbf{U}^\top$.

The subsequent lemma is also known as the *rearrangement inequality*:

LEMMA 2.3. [34, Section 10.2, Theorem 368] *Let $\mathbf{a} \in \mathcal{D}_n$ and $\mathbf{b} \in \mathcal{D}_n$. Then, for every permutation matrix $\mathbf{P}$ of dimension $n \times n$,*

$$\mathbf{a}^\top\mathbf{P}\mathbf{b} \leqslant \mathbf{a}^\top\mathbf{b}. \tag{11}$$

We are now ready to prove Theorem 2.1.

Proof of Theorem 2.1. Due to the assumptions made on f and g_0, Problem (7) can be reformulated as

$$\underset{\mathbf{d}\in\mathcal{D}_n, \mathbf{U_C}\in\mathcal{O}_n}{\text{minimize}}\ \varphi(\mathbf{d}) - \operatorname{trace}\left(\mathbf{U_C}\operatorname{Diag}(\mathbf{d})\mathbf{U_C}^\top\mathbf{T}\right) + \psi(\mathbf{d}).$$

According to the first claim in Lemma 2.2,

$$\inf_{\mathbf{d}\in\mathcal{D}_n, \mathbf{U_C}\in\mathcal{O}_n} \varphi(\mathbf{d}) - \operatorname{trace}\left(\mathbf{U_C}\operatorname{Diag}(\mathbf{d})\mathbf{U_C}^\top\mathbf{T}\right) + \psi(\mathbf{d}) \geqslant \inf_{\mathbf{d}\in\mathcal{D}_n} \varphi(\mathbf{d}) - \mathbf{d}^\top\widetilde{\mathbf{t}} + \psi(\mathbf{d}),$$

where $\widetilde{\mathbf{t}} \in \mathcal{D}_n$ is the vector of ordered eigenvalues of $\mathbf{T} = \widetilde{\mathbf{U}}_\mathbf{T}\operatorname{Diag}(\widetilde{\mathbf{t}})\widetilde{\mathbf{U}}_\mathbf{T}^\top$ with $\widetilde{\mathbf{U}}_\mathbf{T} \in \mathcal{O}_n$. In addition, the last claim in Lemma 2.2 allows us to conclude that the lower bound is attained when $\mathbf{U_C} = \widetilde{\mathbf{U}}_\mathbf{T}$. This proves that

$$\inf_{\mathbf{C}\in\mathcal{S}_n} f(\mathbf{C}) - \operatorname{trace}(\mathbf{TC}) + g_0(\mathbf{C}) = \inf_{\mathbf{d}\in\mathcal{D}_n} \varphi(\mathbf{d}) - \mathbf{d}^\top\widetilde{\mathbf{t}} + \psi(\mathbf{d}). \tag{12}$$

Let us now show that ordering the eigenvalues is unnecessary for our purposes. Let $\mathbf{t} \in \mathbb{R}^n$ be a vector of non necessarily ordered eigenvalues of $\mathbf{T}$. Then, $\mathbf{T} = \mathbf{U_T}\operatorname{Diag}(\mathbf{t})\mathbf{U_T}^\top$

with $\mathbf{U_T} \in \mathcal{O}_n$ and there exists a permutation matrix $\mathbf{Q}$ such that $\mathbf{t} = \mathbf{Q}\widetilde{\mathbf{t}}$. For every vector $\mathbf{d} \in \mathcal{D}_n$ and for every permutation matrix $\mathbf{P}$ of dimension $n \times n$, we have then

$$\begin{aligned}\varphi(\mathbf{Pd}) - (\mathbf{Pd})^\top \mathbf{t} + \psi(\mathbf{Pd}) &= \varphi(\mathbf{Pd}) - (\mathbf{Pd})^\top \mathbf{Q}\widetilde{\mathbf{t}} + \psi(\mathbf{Pd}) \\ &= \varphi(\mathbf{d}) - (\mathbf{Q}^\top \mathbf{Pd})^\top \widetilde{\mathbf{t}} + \psi(\mathbf{d}) \\ &\geqslant \varphi(\mathbf{d}) - \mathbf{d}^\top \widetilde{\mathbf{t}} + \psi(\mathbf{d}),\end{aligned} \tag{13}$$

where the last inequality is a direct consequence of Lemma 2.3. In addition, the equality is obviously reached if $\mathbf{P} = \mathbf{Q}$. Since every vector in $\mathbb{R}^n$ can be expressed as permutation of a vector in $\mathcal{D}_n$, we deduce that

$$\inf_{\mathbf{d} \in \mathbb{R}^n} \varphi(\mathbf{d}) - \mathbf{d}^\top \mathbf{t} + \psi(\mathbf{d}) = \inf_{\mathbf{d} \in \mathcal{D}_n} \varphi(\mathbf{d}) - \mathbf{d}^\top \widetilde{\mathbf{t}} + \psi(\mathbf{d}). \tag{14}$$

Altogether, (12) and (14) lead to

$$\inf_{\mathbf{C} \in \mathcal{S}_n} f(\mathbf{C}) - \operatorname{trace}(\mathbf{TC}) + g_0(\mathbf{C}) = \inf_{\mathbf{d} \in \mathbb{R}^n} \varphi(\mathbf{d}) - \mathbf{d}^\top \mathbf{t} + \psi(\mathbf{d}). \tag{15}$$

Since the function $\mathbf{d} \mapsto \varphi(\mathbf{d}) - \mathbf{d}^\top \mathbf{t} + \psi(\mathbf{d})$ is proper, lower-semicontinuous, and coercive, it follows from [56, Theorem 1.9] that there exists $\widehat{\mathbf{d}} \in \mathbb{R}^n$ such that

$$\varphi(\widehat{\mathbf{d}}) - \widehat{\mathbf{d}}^\top \mathbf{t} + \psi(\widehat{\mathbf{d}}) = \inf_{\mathbf{d} \in \mathbb{R}^n} \varphi(\mathbf{d}) - \mathbf{d}^\top \mathbf{t} + \psi(\mathbf{d}). \tag{16}$$

In addition, it is easy to check that if $\widehat{\mathbf{C}}$ is given by (8) then

$$f(\widehat{\mathbf{C}}) - \operatorname{trace}\left(\mathbf{T}\widehat{\mathbf{C}}\right) + g_0(\widehat{\mathbf{C}}) = \varphi(\widehat{\mathbf{d}}) - \widehat{\mathbf{d}}^\top \mathbf{t} + \psi(\widehat{\mathbf{d}}), \tag{17}$$

which yields the desired result. ∎

Before deriving a main consequence of this result, we need to recall some definitions from convex analysis [55, Chapter 26] [5, Section 3.4]:

Definition 2.4. *Let $\mathcal{H}$ be a finite dimensional real Hilbert space with norm $\|\cdot\|$ and scalar product $\langle \cdot, \cdot \rangle$. Let $h \colon \mathcal{H} \to]-\infty, +\infty]$ be a proper convex function.*

- *h is essentially smooth if h is differentiable on $\operatorname{int}(\operatorname{dom} h) \neq \varnothing$ and $\lim_{n \to +\infty} \|\nabla h(x_n)\| = +\infty$ for every sequence $(x_n)_{n \in \mathbb{N}}$ of $\operatorname{int}(\operatorname{dom} h)$ converging to a point on the boundary of $\operatorname{dom} h$.*
- *h is essentially strictly convex if h is strictly convex on every convex subset of the domain of its subdifferential.*
- *h is a Legendre function if it is both essentially smooth and essentially strictly convex.*
- *If h is differentiable on $\operatorname{int}(\operatorname{dom} h) \neq \varnothing$, the h-Bregman divergence is the function D^h defined on $\mathcal{H}^2$ as*

$$(\forall (x, y) \in \mathcal{H}^2) \quad D^h(x, y) = \begin{cases} h(x) - h(y) - \langle \nabla h(y), x - y \rangle & \text{if } y \in \operatorname{int}(\operatorname{dom} f) \\ +\infty & \text{otherwise.} \end{cases} \tag{18}$$

- *Assume that h is a lower-semicontinuous Legendre function and that ℓ is a lower-semicontinuous convex function such that $\operatorname{int}(\operatorname{dom} h) \cap \operatorname{dom} \ell \neq \varnothing$*

and either ℓ is bounded from below or $h+\ell$ is supercoercive. Then, the D^h-proximity operator of ℓ is

$$\operatorname{prox}_\ell^h\colon \operatorname{int}(\operatorname{dom} h) \to \operatorname{int}(\operatorname{dom} h) \cap \operatorname{dom} \ell \qquad y \mapsto \underset{x\in\mathcal{H}}{\operatorname{argmin}}\ \ell(x) + D^h(x,y). \tag{19}$$

In this definition, when $h = \|\cdot\|^2/2$, we recover the classical definition of the proximity operator in [49], which is defined over $\mathcal{H}$, for every function $\ell \in \Gamma_0(\mathcal{H})$, and that will be simply denoted by prox_ℓ.

We will also need the following result:

Lemma 2.5. *Let f be a function satisfying (4) where $\varphi\colon \mathbb{R}^n \to]-\infty,+\infty]$. Let $\mathbf{C} \in \mathcal{S}_n$ and let $\mathbf{d} \in \mathbb{R}^n$ be a vector of eigenvalues of this matrix. The following hold:*

(i) *$\mathbf{C} \in \operatorname{dom} f$ if and only if $\mathbf{d} \in \operatorname{dom}\varphi$;*

(ii) *$\mathbf{C} \in \operatorname{int}(\operatorname{dom} f)$ if and only if $\mathbf{d} \in \operatorname{int}(\operatorname{dom}\varphi)$.*

Proof. (i) obviously holds since f is a spectral function.
Let us now prove (ii). If $\mathbf{C} \in \operatorname{int}(\operatorname{dom} f)$, then $\mathbf{d} \in \operatorname{dom}\varphi$. In addition, there exists $\rho \in]0,+\infty[$ such that, for every $\mathbf{C}' \in \mathcal{S}_n$, if $\|\mathbf{C}'-\mathbf{C}\|_{\mathrm{F}} \leqslant \rho$, then $\mathbf{C}' \in \operatorname{dom} f$. Let $\mathbf{U}_{\mathbf{C}} \in \mathcal{O}_n$ be such that $\mathbf{C} = \mathbf{U}_{\mathbf{C}} \operatorname{Diag}(\mathbf{d})\mathbf{U}_{\mathbf{C}}^\top$ and let us choose $\mathbf{C}' = \mathbf{U}_{\mathbf{C}} \operatorname{Diag}(\mathbf{d}')\mathbf{U}_{\mathbf{C}}^\top$ with $\mathbf{d}' \in \mathbb{R}^n$. Since $\mathbf{C}$ and $\mathbf{C}'$ share the same eigenbasis,

$$\|\mathbf{C}'-\mathbf{C}\|_{\mathrm{F}} = \|\mathbf{d}'-\mathbf{d}\|. \tag{20}$$

Hence, for any $\mathbf{d}' \in \mathbb{R}^n$ such that $\|\mathbf{d}'-\mathbf{d}\| \leqslant \rho$, $\mathbf{C}' \in \operatorname{dom} f$, hence $\mathbf{d}' \in \operatorname{dom}\varphi$. This shows that $\mathbf{d} \in \operatorname{int}(\operatorname{dom}\varphi)$.
Conversely, let us assume that $\mathbf{d} = (d_i)_{1\leqslant i\leqslant n} \in \operatorname{int}(\operatorname{dom}\varphi)$. Without loss of generality, it can be assumed that $\mathbf{d} \in \mathcal{D}_n$. There thus exists $\rho \in]0,+\infty[$ such that for every $\mathbf{d}' = (d_i')_{1\leqslant i\leqslant n} \in \mathcal{D}_n$, if

$$(\forall i \in \{1,\ldots,n\}) \qquad |d_i'-d_i| \leqslant \rho, \tag{21}$$

then $\mathbf{d}' \in \operatorname{dom}\varphi$. Furthermore, let $\mathbf{C}'$ be any matrix in $\mathcal{S}_n$ such that

$$\|\mathbf{C}'-\mathbf{C}\|_{\mathrm{F}} \leqslant \rho \tag{22}$$

and let $\mathbf{d}' = (d_i')_{1\leqslant i\leqslant n} \in \mathcal{D}_n$ be a vector of eigenvalues of $\mathbf{C}$. It follows from Weyl's inequality [46] that

$$(\forall i \in \{1,\ldots,n\}) \qquad |d_i'-d_i| \leqslant \|\mathbf{C}'-\mathbf{C}\|_{\mathrm{S}} \leqslant \|\mathbf{C}'-\mathbf{C}\|_{\mathrm{F}} \leqslant \rho. \tag{23}$$

We deduce that $\mathbf{d}' \in \operatorname{dom}\varphi$ and, consequently $\mathbf{C}' \in \operatorname{dom} f$. This shows that $\mathbf{C} \in \operatorname{int}(\operatorname{dom} f)$. ∎

As an offspring of Theorem 2.1, we then get:

Corollary 2.6. *Let f and g_0 be functions satisfying (4) and (5), respectively, where $\varphi \in \Gamma_0(\mathbb{R}^n)$ is a Legendre function, $\psi \in \Gamma_0(\mathbb{R}^n)$, $\operatorname{int}(\operatorname{dom}\varphi) \cap \operatorname{dom}\psi \neq \varnothing$, and either ψ is bounded from below or $\varphi+\psi$ is supercoercive. Then, the D^f-proximity operator of g_0 is defined at every $\mathbf{Y} \in \mathcal{S}_n$ such that $\mathbf{Y} = \mathbf{U}_{\mathbf{Y}} \operatorname{Diag}(\mathbf{y})\mathbf{U}_{\mathbf{Y}}^\top$ with $\mathbf{U}_{\mathbf{Y}} \in \mathcal{O}_n$ and $\mathbf{y} \in \operatorname{int}(\operatorname{dom}\varphi)$, and it is expressed as*

$$\operatorname{prox}_{g_0}^f(\mathbf{Y}) = \mathbf{U}_{\mathbf{Y}} \operatorname{Diag}(\operatorname{prox}_\psi^\varphi(\mathbf{y}))\mathbf{U}_{\mathbf{Y}}^\top. \tag{24}$$

Proof. According to the properties of spectral functions [38, Corollary 2.7],

$$\varphi \in \Gamma_0(\mathbb{R}^n) \text{ (resp. } \psi \in \Gamma_0(\mathbb{R}^n)) \quad \Rightarrow \quad f \in \Gamma_0(\mathcal{S}_n) \text{ (resp. } g_0 \in \Gamma_0(\mathcal{S}_n)). \tag{25}$$

In addition, according to [38, Corollaries 3.3&3.5], since φ is a Legendre function, f is a Legendre function. It is also straightforward to check that, when ψ is lower bounded, then g_0 is lower bounded and, when $\varphi + \psi$ is supercoercive, then $f + g_0$ is supercoercive. It also follows from Lemma 2.5 that $\operatorname{int}(\operatorname{dom}\varphi) \cap \operatorname{dom}\psi \neq \varnothing \Leftrightarrow \operatorname{int}(\operatorname{dom} f) \cap \operatorname{dom} g_0 \neq \varnothing$.

The above results show that the D^f-proximity operator of g_0 is properly defined as follows:

$$\begin{aligned} \operatorname{prox}_{g_0}^f \colon \operatorname{int}(\operatorname{dom} f) &\to \operatorname{int}(\operatorname{dom} f) \cap \operatorname{dom} g_0 \\ \mathbf{Y} &\mapsto \operatorname*{argmin}_{\mathbf{C} \in \mathcal{S}_n} g_0(\mathbf{C}) + D^f(\mathbf{C}, \mathbf{Y}). \end{aligned} \tag{26}$$

This implies that computing the D^f-proximity operator of g_0 at $\mathbf{Y} \in \operatorname{int}(\operatorname{dom} f)$ amounts to finding the unique solution to Problem (7) where $\mathbf{T} = \nabla f(\mathbf{Y})$. Let $\mathbf{Y} = \mathbf{U_Y} \operatorname{Diag}(\mathbf{y}) \mathbf{U_Y}^\top$ with $\mathbf{U_Y} \in \mathcal{O}_n$ and $\mathbf{y} \in \mathbb{R}^n$. By Lemma 2.5(ii), $\mathbf{Y} \in \operatorname{int}(\operatorname{dom} f) \Leftrightarrow \mathbf{y} \in \operatorname{int}(\operatorname{dom}(\varphi))$ and, according to [38, Corollary 3.3], $\mathbf{T} = \mathbf{U_Y} \operatorname{Diag}(\mathbf{t}) \mathbf{U_Y}^\top$ with $\mathbf{t} = \nabla\varphi(\mathbf{y})$.

Furthermore, as φ is essentially strictly convex, it follows from [4, Theorem 5.9(ii)] that $\mathbf{t} = \nabla\varphi(\mathbf{y}) \in \operatorname{int}(\operatorname{dom} f^*)$, which according to [6, Theorem 14.17] is equivalent to the fact that $\mathbf{d} \mapsto \varphi(\mathbf{d}) - \mathbf{d}^\top \mathbf{t}$ is coercive. So, if ψ is lower-bounded, $\mathbf{d} \mapsto \varphi(\mathbf{d}) - \mathbf{d}^\top \mathbf{t} + \psi(\mathbf{d})$ is coercive. The same conclusion obviously holds if $\varphi + \psi$ is supercoercive. This shows that the assumptions of Theorem 2.1 are met. Consequently, applying this theorem yields

$$\operatorname{prox}_{g_0}^f(\mathbf{Y}) = \mathbf{U_Y} \operatorname{Diag}(\widehat{\mathbf{d}}) \mathbf{U_Y}^\top, \tag{27}$$

where $\widehat{\mathbf{d}}$ minimizes

$$\mathbf{d} \mapsto \varphi(\mathbf{d}) - \mathbf{d}^\top \mathbf{t} + \psi(\mathbf{d}) \tag{28}$$

or, equivalently,

$$\mathbf{d} \mapsto \psi(\mathbf{d}) + D^\varphi(\mathbf{d}, \mathbf{y}). \tag{29}$$

This shows that $\widehat{\mathbf{d}} = \operatorname{prox}_\psi^\varphi(\mathbf{y})$. ∎

Remark 2.7. Corollary 2.6 extends known results concerning the case when $f = \|\cdot\|_{\mathrm{F}}/2$ [16]. A rigorous derivation of the proximity operator of spectral functions in $\Gamma_0(\mathcal{S}_n)$ for the standard Frobenius metric can be found in [6, Corollary 24.65]. Our proof allows us to recover a similar result by adopting a more general approach. In particular, it is worth noticing that Theorem 2.1 does not require any convexity assumption.

3. Proximal Iterative Approach. Let us now turn to the more general case of the resolution of Problem (1) when $f \in \Gamma_0(\mathcal{S}_n)$ and $g_1 \not\equiv 0$. Proximal splitting approaches for finding a minimizer of a sum of non-necessarily smooth functions have attracted a large interest in the last years [24, 51, 37, 15]. In these methods, the functions can be dealt with either via their gradient or their proximity operator depending on their differentiability properties. In this section, we first list a number of

proximity operators of scaled versions of $f - \operatorname{trace}(\mathbf{T}\,\cdot) + g_0$, where f and g_0, satisfying (4) and (5), are chosen among several options that can be useful in a wide range of practical scenarios. Based on these results, we then propose a proximal splitting Douglas-Rachford algorithm to solve Problem (1).

3.1. Proximity Operators. By definition, computing the proximity operator of $\gamma\,(f - \operatorname{trace}(\mathbf{T}\,\cdot) + g_0)$ with $\gamma \in]0, +\infty[$ at $\overline{\mathbf{C}} \in \mathcal{S}_n$ amounts to find a minimizer of the function

$$\mathbf{C} \mapsto f(\mathbf{C}) - \operatorname{trace}(\mathbf{T}\mathbf{C}) + g_0(\mathbf{C}) + \frac{1}{2\gamma}\|\mathbf{C} - \overline{\mathbf{C}}\|_{\mathrm{F}}^2 \tag{30}$$

over $\mathcal{S}_n$. The (possibly empty) set of such minimizers is denoted by $\operatorname{Prox}_{\gamma(f-\operatorname{trace}(\mathbf{T}\,\cdot)+g_0)}(\overline{\mathbf{C}})$. As pointed out in Section 2, if $f + g_0 \in \Gamma_0(\mathcal{S}_n)$ then this set is a singleton $\{\operatorname{prox}_{\gamma(f-\operatorname{trace}(\mathbf{T}\,\cdot)+g_0)}(\overline{\mathbf{C}})\}$. We have the following characterization of this proximity operator:

PROPOSITION 3.1. *Let $\gamma \in]0, +\infty[$ and $\overline{\mathbf{C}} \in \mathcal{S}_n$. Let f and g_0 be functions satisfying (4) and (5), respectively, where $\varphi \in \Gamma_0(\mathbb{R}^n)$ and ψ is a lower-semicontinuous function such that $\operatorname{dom}\varphi \cap \operatorname{dom}\psi \neq \varnothing$. Let $\boldsymbol{\lambda} \in \mathbb{R}^n$ and $\mathbf{U} \in \mathcal{O}_n$ be such that $\overline{\mathbf{C}} + \gamma\mathbf{T} = \mathbf{U}\operatorname{Diag}(\boldsymbol{\lambda})\mathbf{U}^\top$.*

(i) *If ψ is lower bounded by an affine function then $\operatorname{Prox}_{\gamma(\varphi+\psi)}(\boldsymbol{\lambda}) \neq \varnothing$ and, for every $\widehat{\boldsymbol{\lambda}} \in \operatorname{Prox}_{\gamma(\varphi+\psi)}(\boldsymbol{\lambda})$,*

$$\mathbf{U}\operatorname{Diag}(\widehat{\boldsymbol{\lambda}})\mathbf{U}^\top \in \operatorname{Prox}_{\gamma(f-\operatorname{trace}(\mathbf{T}\,\cdot)+g_0)}(\overline{\mathbf{C}}). \tag{31}$$

(ii) *If ψ is convex, then*

$$\operatorname{prox}_{\gamma(f-\operatorname{trace}(\mathbf{T}\,\cdot)+g_0)}(\overline{\mathbf{C}}) = \mathbf{U}\operatorname{Diag}\Big(\operatorname{prox}_{\gamma(\varphi+\psi)}(\boldsymbol{\lambda})\Big)\mathbf{U}^\top. \tag{32}$$

Proof. (i): Since it has been assumed that f and g_0 are spectral functions, we have

$$(\forall \mathbf{C} \in \mathcal{S}_n) \quad f(\mathbf{C}) + g_0(\mathbf{C}) = \varphi(\mathbf{d}) + \psi(\mathbf{d}), \tag{33}$$

where $\mathbf{d} \in \mathbb{R}^n$ is a vector of the eigenvalues of $\mathbf{C}$. It can be noticed that minimizing (30) is obviously equivalent to minimize $\widetilde{f} - \gamma^{-1}\operatorname{trace}\big((\overline{\mathbf{C}} + \gamma\mathbf{T})\;\cdot\big) + g_0$ where $\widetilde{f} = f + \|\cdot\|_{\mathrm{F}}^2/(2\gamma)$. Then

$$\widetilde{f}(\mathbf{C}) = \widetilde{\varphi}(\mathbf{d}), \tag{34}$$

where $\widetilde{\varphi} = \varphi + \|\cdot\|^2/(2\gamma)$. Since we have assumed that $\varphi \in \Gamma_0(\mathbb{R}^n)$, $\widetilde{\varphi}$ is proper, lower-semicontinuous, and strongly convex. As ψ is lower bounded by an affine function, it follows that

$$\mathbf{d} \mapsto \widetilde{\varphi}(\mathbf{d}) - \gamma^{-1}\boldsymbol{\lambda}^\top\mathbf{d} + \psi(\mathbf{d}) \tag{35}$$

is lower bounded by a strongly convex function and it is thus coercive. In addition, $\operatorname{dom}\widetilde{\varphi} = \operatorname{dom}\varphi$, hence $\operatorname{dom}\widetilde{\varphi} \cap \operatorname{dom}\psi \neq \varnothing$. Let us now apply Theorem 2.1. Let $\widehat{\boldsymbol{\lambda}}$ be a minimizer of (35). It can be claimed that $\widehat{\mathbf{C}} = \mathbf{U}\operatorname{Diag}(\widehat{\boldsymbol{\lambda}})\mathbf{U}^\top$ is a minimizer of (30). On the other hand, minimizing (35) is equivalent to minimize $\gamma(\varphi + \psi) + \frac{1}{2}\|\cdot - \boldsymbol{\lambda}\|^2$, which shows that $\widehat{\boldsymbol{\lambda}} \in \operatorname{Prox}_{\gamma(\varphi+\psi)}(\boldsymbol{\lambda})$.

(ii): If $\psi \in \Gamma_0(\mathbb{R}^n)$, then it is lower bounded by an affine function [6, Theorem 9.20]. Furthermore, $\varphi + \psi \in \Gamma_0(\mathbb{R}^n)$ and the proximity operator of $\gamma(\varphi + \psi)$ is thus single valued. On the other hand, we also have $\gamma(f - \operatorname{trace}(\mathbf{T}\,\cdot) + g_0) \in \Gamma_0(\mathcal{S}_n)$ [38, Corollary 2.7], and the proximity operator of this function is single valued too. The result directly follows from (i). ∎

We will next focus on the use of Proposition 3.1 for three choices for f, namely the classical squared Frobenius norm, the minus log det functional, and the Von Neumann entropy, each choice being coupled with various possible choices for g_0.

3.1.1. Squared Frobenius Norm. A suitable choice in Problem (1) is $f = \|\cdot\|_{\mathrm{F}}^2/2$ [72, 54, 19]. The squared Froebenius norm is the spectral function associated with the function $\varphi = \|\cdot\|^2/2$. It is worth mentioning that this choice for f allows us to rewrite the original Problem (1) under the form (3), where

$$\left(\forall(\mathbf{C},\mathbf{Y}) \in \mathcal{S}_n^2\right) \quad D^f(\mathbf{C},\mathbf{Y}) = \frac{1}{2}\|\mathbf{C}-\mathbf{Y}\|_{\mathrm{F}}^2. \tag{36}$$

We have thus re-expressed Problem (1) as the determination of a proximal point of function g at $\mathbf{T}$ in the Frobenius metric.

Table 1 presents several examples of spectral functions g_0 and the expression of the proximity operator of $\gamma(\varphi + \psi)$ with $\gamma \in]0, +\infty[$. These expressions were established by using the properties of proximity operators of functions defined on $\mathbb{R}^n$ (see [20, Example 4.4] and [24, Tables 10.1 and 10.2]).

Remark 3.2. Another option for g_0 is to choose it equal to $\mu\|\cdot\|_{\mathrm{S}}$ where $\mu \in]0, +\infty[$. For every $\gamma \in]0, +\infty[$, we have then

$$(\forall \boldsymbol{\lambda} \in \mathbb{R}^n) \qquad \operatorname{prox}_{\gamma(\varphi+\psi)}(\boldsymbol{\lambda}) = \operatorname{prox}_{\frac{\mu\gamma}{1+\gamma}\|\cdot\|_{+\infty}}\left(\frac{\boldsymbol{\lambda}}{1+\gamma}\right), \tag{37}$$

where $\|\cdot\|_{+\infty}$ is the infinity norm of $\mathbb{R}^n$. By noticing that $\|\cdot\|_{+\infty}$ is the conjugate function of the indicator function of B_{ℓ^1}, the unit ℓ^1 ball centered at 0 of $\mathbb{R}^n$, and using Moreau's decomposition formula, [6, Proposition 24.8(ix)] yields

$$(\forall \boldsymbol{\lambda} \in \mathbb{R}^n) \qquad \operatorname{prox}_{\gamma(\varphi+\psi)}(\boldsymbol{\lambda}) = \frac{1}{1+\gamma}\left(\boldsymbol{\lambda} - \mu\gamma \operatorname{proj}_{B_{\ell^1}}\left(\frac{\boldsymbol{\lambda}}{\mu\gamma}\right)\right). \tag{38}$$

The required projection onto B_{ℓ^1} can be computed through efficient algorithms [61, 25].

3.1.2. Logdet Function. Another popular choice for f is the negative logarithmic determinant function [30, 58, 44, 48, 3, 31, 67, 18], which is defined as follows

$$(\forall \mathbf{C} \in \mathcal{S}_n) \quad f(\mathbf{C}) = \begin{cases} -\log\det(\mathbf{C}) & \text{if } \mathbf{C} \in \mathcal{S}_n^{++} \\ +\infty & \text{otherwise.} \end{cases} \tag{39}$$

The above function satisfies property (5) with

$$\left(\forall \boldsymbol{\lambda} = (\lambda_i)_{1\leqslant i\leqslant n} \in \mathbb{R}^n\right) \qquad \varphi(\boldsymbol{\lambda}) = \begin{cases} -\sum_{i=1}^{n}\log(\lambda_i) & \text{if } \boldsymbol{\lambda} \in]0, +\infty[^n \\ +\infty & \text{otherwise.} \end{cases} \tag{40}$$

TABLE 1

Proximity operators of $\gamma(\frac{1}{2}\|\cdot\|_{\mathrm{F}}^2 + g_0)$ with $\gamma > 0$ evaluated at symmetric matrix with vector of eigenvalues $\boldsymbol{\lambda} = (\lambda_i)_{1\leqslant i\leqslant n}$. For the inverse Schatten penalty, the function is set to $+\infty$ when the argument $\mathbf{C}$ is not positive definite. E_1 denotes the set of matrices in $\mathcal{S}_n$ with Frobenius norm less than or equal to α and E_2 the set of matrices in $\mathcal{S}_n$ with eigenvalues between α and β. In the last line, the i-th component of the proximity operator is obtained by searching among the nonnegative roots of a third order polynomial those minimizing $\lambda_i' \mapsto \frac{1}{2}(\lambda_i' - |\lambda_i|)^2 + \gamma\big(\frac{1}{2}(\lambda_i')^2 + \mu\log((\lambda_i')^2 + \varepsilon)\big)$.

$g_0(\mathbf{C})$, $\mu > 0$	$\mathrm{prox}_{\gamma(\varphi+\psi)}(\boldsymbol{\lambda})$
Nuclear norm $\mu\mathcal{R}_1(\mathbf{C})$	$\Big(\mathrm{soft}_{\frac{\mu\gamma}{\gamma+1}}\Big(\frac{\lambda_i}{\gamma+1}\Big)\Big)_{1\leqslant i\leqslant n}$
Frobenius norm $\mu\|\mathbf{C}\|_{\mathrm{F}}$	$\Big(1 - \frac{\gamma\mu}{\|\boldsymbol{\lambda}\|}\Big)\frac{\boldsymbol{\lambda}}{1+\gamma}$ if $\|\boldsymbol{\lambda}\| > \gamma\mu$ and $\mathbf{0}$ otherwise
Squared Frobenius norm $\mu\|\mathbf{C}\|_{\mathrm{F}}^2$	$\dfrac{\boldsymbol{\lambda}}{1+\gamma\,(1+2\mu)}$
Schatten 3–penalty $\mu\mathcal{R}_3^3(\mathbf{C})$	$(6\gamma\mu)^{-1}\Big(\mathrm{sign}\,(\lambda_i)\sqrt{(\gamma+1)^2 + 12\lvert\lambda_i\rvert\gamma\mu} - \gamma - 1\Big)_{1\leqslant i\leqslant n}$
Schatten 4–penalty $\mu\mathcal{R}_4^4(\mathbf{C})$	$(8\gamma\mu)^{-1/3}\left(\sqrt[3]{\lambda_i + \sqrt{\lambda_i^2+\zeta}} + \sqrt[3]{\lambda_i - \sqrt{\lambda_i^2+\zeta}}\right)_{1\leqslant i\leqslant n}$ with $\zeta = \frac{(\gamma+1)^3}{27\gamma\mu}$
Schatten 4/3–penalty $\mu\mathcal{R}_{4/3}^{4/3}(\mathbf{C})$	$\frac{1}{1+\gamma}\left(\lambda_i + \frac{4\gamma\mu}{3\sqrt[3]{2(1+\gamma)}}\Big(\sqrt[3]{\sqrt{\lambda_i^2+\zeta} - \lambda_i} - \sqrt[3]{\sqrt{\lambda_i^2+\zeta} + \lambda_i}\Big)\right)_{1\leqslant i\leqslant n}$ with $\zeta = \frac{256(\gamma\mu)^3}{729(1+\gamma)}$
Schatten 3/2–penalty $\mu\mathcal{R}_{3/2}^{3/2}(\mathbf{C})$	$\frac{1}{1+\gamma}\left(\lambda_i + \frac{9\gamma^2\mu^2}{8(1+\gamma)}\,\mathrm{sign}(\lambda_i)\Big(1 - \sqrt{1 + \frac{16(1+\gamma)}{9\gamma^2\mu^2}\lvert\lambda_i\rvert}\Big)\right)_{1\leqslant i\leqslant n}$
Schatten p–penalty $\mu\mathcal{R}_p^p(\mathbf{C})$, $p \geqslant 1$	$\big(\mathrm{sign}(\lambda_i)d_i\big)_{1\leqslant i\leqslant n}$ with $(\forall i \in \{1,\dots,n\})$ $d_i \geqslant 0$ and $\mu\gamma p d_i^{p-1} + (\gamma+1)d_i = \lambda_i$
Inverse Schatten p–penalty $\mu\mathcal{R}_p^p(\mathbf{C}^{-1})$, $p > 0$	$\big(d_i\big)_{1\leqslant i\leqslant n}$ with $(\forall i \in \{1,\dots,n\})$ $d_i > 0$ and $(\gamma+1)d_i^{p+2} - \lambda_i d_i^{p+1} = \mu\gamma p$
Bound on the Frobenius norm $\iota_{E_1}(\mathbf{C})$	$\alpha\dfrac{\boldsymbol{\lambda}}{\|\boldsymbol{\lambda}\|}$ if $\|\boldsymbol{\lambda}\| > \alpha(1+\gamma)$ and $\dfrac{\boldsymbol{\lambda}}{1+\gamma}$ otherwise, $\alpha \in [0,+\infty[$
Bounds on eigenvalues $\iota_{E_2}(\mathbf{C})$	$(\min(\max(\lambda_i/(\gamma+1),\alpha),\beta))_{1\leqslant i\leqslant n}$, $[\alpha,\beta] \subset [-\infty,+\infty]$
Rank $\mu\,\mathrm{rank}(\mathbf{C})$	$\left(\mathrm{hard}_{\sqrt{\frac{2\mu\gamma}{1+\gamma}}}\left(\dfrac{\lambda_i}{1+\gamma}\right)\right)_{1\leqslant i\leqslant n}$
Cauchy $\mu\log\det(\mathbf{C}^2 + \varepsilon\mathrm{I_d})$, $\varepsilon > 0$	$\in \big\{(\mathrm{sign}(\lambda_i)d_i)_{1\leqslant i\leqslant n} \mid (\forall i \in \{1,\dots,n\})\ d_i \geqslant 0$ and $(\gamma+1)d_i^3 - \lvert\lambda_i\rvert d_i^2 + \big(2\gamma\mu + \varepsilon(\gamma+1)\big)d_i = \lvert\lambda_i\rvert\varepsilon\big\}$

Actually, for a given positive definite matrix, the value of function (39) simply reduces to the Burg entropy of its eigenvalues. Hereagain, if $\mathbf{Y} \in \mathcal{S}_n^{++}$ and $\mathbf{T} = -\mathbf{Y}^{-1}$, we can rewrite Problem (1) under the form (3), so that it becomes equivalent to the computation of the proximity operator of g with respect to the Bregman divergence given by

$$(\forall \mathbf{C} \in \mathcal{S}_n) \quad D^f(\mathbf{C},\mathbf{Y}) = \begin{cases} \log\left(\dfrac{\det(\mathbf{Y})}{\det(\mathbf{C})}\right) + \mathrm{trace}\left(\mathbf{Y}^{-1}\mathbf{C}\right) - n & \text{if } \mathbf{C} \in \mathcal{S}_n^{++} \\ +\infty & \text{otherwise.} \end{cases} \tag{41}$$

In Table 2, we list some particular choices for g_0, and provide the associated closed form expression of the proximity operator $\mathrm{prox}_{\gamma(\varphi+\psi)}$ for $\gamma \in]0,+\infty[$, where φ is defined in (40). These expressions were derived from [24, Table 10.2].

Remark 3.3. Let g_0 be any of the convex spectral functions listed in Table 2. Let $\mathbf{W}$ be an invertible matrix in $\mathbb{R}^{n\times n}$, and let $\overline{\mathbf{C}} \in \mathcal{S}_n$ From the above results, one can deduce the minimizer of $\mathbf{C} \mapsto \gamma(f(\mathbf{C}) + g_0(\mathbf{W}\mathbf{C}\mathbf{W}^\top)) + \frac{1}{2}\|\mathbf{W}\mathbf{C}\mathbf{W}^\top - \overline{\mathbf{C}}\|_{\mathrm{F}}^2$ where $\gamma \in]0, +\infty[$. Indeed, by making a change of variable and by using basic properties of the log det function, this minimizer is equal to $\mathbf{W}^{-1}\operatorname{prox}_{\gamma(f+g_0)}(\overline{\mathbf{C}})(\mathbf{W}^{-1})^\top$.

Table 2

Proximity operators of $\gamma(f + g_0)$ with $\gamma > 0$ and f given by (39), evaluated at a symmetric matrix with vector of eigenvalues $\boldsymbol{\lambda} = (\lambda_i)_{1\leqslant i\leqslant n}$. For the inverse Schatten penalty, the function is set to $+\infty$ when the argument $\mathbf{C}$ is not positive definite. E_2 denotes the set of matrices in $\mathcal{S}_n$ with eigenvalues between α and β. In the last line, the i-th component of the proximity operator is obtained by searching among the positive roots of a fourth order polynomial those minimizing $\lambda_i' \mapsto \frac{1}{2}(\lambda_i' - \lambda_i)^2 + \gamma\big(\mu\log((\lambda_i')^2 + \varepsilon) - \log\lambda_i'\big)$.

$g_0(\mathbf{C})$, $\mu > 0$	$\operatorname{prox}_{\gamma(\varphi+\psi)}(\boldsymbol{\lambda})$
Nuclear norm $\mu\mathcal{R}_1(\mathbf{C})$	$\frac{1}{2}\left(\lambda_i - \gamma\mu + \sqrt{(\lambda_i - \gamma\mu)^2 + 4\gamma}\right)_{1\leqslant i\leqslant n}$
Squared Frobenius norm $\mu\|\mathbf{C}\|_{\mathrm{F}}^2$	$\dfrac{1}{2(2\gamma\mu+1)}\left(\lambda_i + \sqrt{\lambda_i^2 + 4\gamma(2\gamma\mu+1)}\right)_{1\leqslant i\leqslant n}$
Schatten p–penalty $\mu\mathcal{R}_p^p(\mathbf{C})$, $p \geqslant 1$	$(d_i)_{1\leqslant i\leqslant n}$ with $(\forall i \in \{1,\ldots,n\})\ d_i > 0$ and $\mu\gamma p d_i^p + d_i^2 - \lambda_i d_i = \gamma$
Inverse Schatten p–penalty $\mu\mathcal{R}_p^p(\mathbf{C}^{-1})$, $p > 0$	$(d_i)_{1\leqslant i\leqslant n}$ with $(\forall i \in \{1,\ldots,n\})\ d_i > 0$ and $d_i^{p+2} - \lambda_i d_i^{p+1} - \gamma d_i^p = \mu\gamma p$
Bounds on eigenvalues $\iota_{E_2}(\mathbf{C})$	$\left(\min\left(\max\left(\frac{1}{2}(\lambda_i + \sqrt{\lambda_i^2 + 4\gamma}), \alpha\right), \beta\right)\right)_{1\leqslant i\leqslant n}$, $[\alpha, \beta] \subset [0, +\infty]$
Cauchy $\mu\log\det(\mathbf{C}^2 + \varepsilon\mathrm{I_d})$, $\varepsilon > 0$	$\in \big\{(d_i)_{1\leqslant i\leqslant n} \mid (\forall i \in \{1,\ldots,n\})\ d_i > 0$ and $d_i^4 - \lambda d_i^3 + \big(\varepsilon + \gamma(2\mu - 1)\big)d_i^2 - \varepsilon\lambda_i d_i = \gamma\varepsilon\big\}$

3.1.3. Von Neumann Entropy. Our third example is the negative Von Neumann entropy, which appears to be useful in some quantum mechanics problems [10]. It is defined as

$$(\forall \mathbf{C} \in \mathcal{S}_n) \quad f(\mathbf{C}) = \begin{cases} \operatorname{trace}(\mathbf{C}\log(\mathbf{C})) & \text{if } \mathbf{C} \in \mathcal{S}_n^+ \\ +\infty & \text{otherwise.} \end{cases} \tag{42}$$

In the above expression, if $\mathbf{C} = \mathbf{U}\operatorname{Diag}(\boldsymbol{\lambda})\mathbf{U}^\top$ with $\boldsymbol{\lambda} = (\lambda_i)_{1\leqslant i\leqslant n} \in]0, +\infty[^n$ and $\mathbf{U} \in \mathcal{O}_n$, then $\log(\mathbf{C}) = \mathbf{U}\operatorname{Diag}\big((\log\lambda_i)_{1\leqslant i\leqslant n}\big)\mathbf{U}^\top$. The logarithm of a symmetric definite positive matrix is uniquely defined and the function $\mathbf{C} \mapsto \mathbf{C}\log(\mathbf{C})$ can be extended by continuity on $\mathcal{S}_n^+$ similarly to the case when $n = 1$. Thus, f is the spectral function associated with

$$\big(\forall \boldsymbol{\lambda} = (\lambda_i)_{1\leqslant i\leqslant n} \in \mathbb{R}^n\big) \qquad \varphi(\boldsymbol{\lambda}) = \begin{cases} \displaystyle\sum_{i=1}^n \lambda_i\log(\lambda_i) & \text{if } \boldsymbol{\lambda} \in [0, +\infty[^n \\ +\infty & \text{otherwise.} \end{cases} \tag{43}$$

Note that the Von Neumann entropy defined for symmetric matrices is simply equal to the well–known Shannon entropy [27] of the input eigenvalues. With this choice for function f, by setting $\mathbf{T} = \log(\mathbf{Y}) + \mathrm{I_d}$ where $\mathbf{Y} \in \mathcal{S}_n^{++}$, Problem (1) can be recast under the form (3), so that it becomes equivalent to the computation of the

proximity operator of g with respect to the Bregman divergence associated with the Von Neumann entropy:

$$(\forall \mathbf{C} \in \mathcal{S}_n) \quad D^f(\mathbf{C}, \mathbf{Y}) = \begin{cases} \operatorname{trace}\left(\mathbf{C}\log(\mathbf{C}) - \mathbf{Y}\log(\mathbf{Y}) - (\log(\mathbf{Y}) + \mathrm{I_d})\,(\mathbf{C} - \mathbf{Y})\right) & \text{if } \mathbf{C} \in \mathcal{S}_n^+ \\ +\infty & \text{otherwise.} \end{cases}$$

We provide in Table 3 a list of closed form expressions of the proximity operator of $\gamma(f + g_0)$ for several choices of the spectral function g_0.

TABLE 3

Proximity operators of $\gamma(f+g_0)$ with $\gamma > 0$ and f given by (42), evaluated at a symmetric matrix with vector of eigenvalues $\boldsymbol{\lambda} = (\lambda_i)_{1\leqslant i\leqslant n}$. E_2 denotes the set of matrices in $\mathcal{S}_n$ with eigenvalues between α and β. $\mathrm{W}(\cdot)$ denotes the W-Lambert function [26].

$g_0(\mathbf{C})$, $\mu > 0$	$\operatorname{prox}_{\gamma(\varphi+\psi)}(\boldsymbol{\lambda})$
Nuclear norm $\mu\mathcal{R}_1(\mathbf{C})$	$\gamma\left(\mathrm{W}\left(\frac{1}{\gamma}\exp\left(\frac{\lambda_i}{\gamma} - \mu - 1\right)\right)\right)_{1\leqslant i\leqslant n}$
Squared Frobenius norm $\mu\lVert\mathbf{C}\rVert_{\mathrm{F}}^2$	$\frac{\gamma}{2\mu\gamma+1}\left(\mathrm{W}\left(\frac{2\mu\gamma+1}{\gamma}\exp\left(\frac{\lambda_i}{\gamma} - 1\right)\right)\right)_{1\leqslant i\leqslant n}$
Schatten p–penalty $\mu\mathcal{R}_p^p(\mathbf{C})$, $p \geqslant 1$	$(d_i)_{1\leqslant i\leqslant n}$ with $(\forall i \in \{1,\ldots,n\})\ d_i > 0$ and $p\mu\gamma d_i^{p-1} + d_i + \gamma\log d_i + \gamma = \lambda_i$
Bounds on eigenvalues $\iota_{E_2}(\mathbf{C})$	$\left(\min\left(\max\left(\gamma\mathrm{W}\left(\frac{1}{\gamma}\exp\left(\frac{\lambda_i}{\gamma} - 1\right)\right), \alpha\right), \beta\right)\right)_{1\leqslant i\leqslant n}$, $[\alpha, \beta] \subset [0, +\infty]$
Rank $\mu\operatorname{rank}(\mathbf{C})$	$(d_i)_{1\leqslant i\leqslant n}$ with $(\forall i \in \{1,\ldots,n\})\ d_i = \begin{cases} \rho_i & \text{if } \rho_i > \chi \\ 0 \text{ or } \rho_i & \text{if } \rho_i = \chi \\ 0 & \text{otherwise} \end{cases}$ and $\begin{cases} \chi = \sqrt{\gamma(\gamma + 2\mu)} - \gamma, \\ \rho_i = \gamma\mathrm{W}\left(\frac{1}{\gamma}\exp\left(\frac{\lambda_i}{\gamma} - 1\right)\right) \end{cases}$

3.2. Douglas-Rachford Algorithm. We now propose a Douglas-Rachford (DR) approach ([41, 24, 23]) for numerically solving Problem (1). The DR method minimizes the sum of $f - \operatorname{trace}(\mathbf{T}\cdot) + g_0$ and g_1 by alternately computing proximity operators of each of these functions. Proposition 3.1 allows us to calculate the proximity operator of $\gamma(f - \operatorname{trace}(\mathbf{T}\cdot) + g_0)$ with $\gamma \in]0, +\infty[$, by possibly using the expressions listed in Tables 1, 2, and 3. Since g_1 is not a spectral function, $\operatorname{prox}_{\gamma g_1}$ has to be derived from other expressions of proximity operators. For instance, if g_1 is a separable sum of functions of its elements, e.g. $g = \lVert\cdot\rVert_1$, standard expressions for the proximity operator of vector functions can be employed [20, 24].[1]

The computations to be performed are summarized in Algorithm 1. We state a convergence theorem in the matrix framework, which is an offspring of existing results in arbitrary Hilbert spaces (see, for example, [24] and [52, Proposition 3.5]).

THEOREM 3.4. *Let f and g_0 be functions satisfying (4) and (5), respectively, where $\varphi \in \Gamma_0(\mathbb{R}^n)$ and $\psi \in \Gamma_0(\mathbb{R}^n)$. Let $g_1 \in \Gamma_0(\mathcal{S}_n)$ be such that $f - \operatorname{trace}(\mathbf{T}\cdot) + g_0 + g_1$ is coercive. Assume that the intersection of the relative interiors of the domains of $f + g_0$ and g_1 is non empty. Let $(\alpha^{(k)})_{k\geqslant 0}$ be a sequence in $[0, 2]$ such that $\sum_{k=0}^{+\infty}\alpha^{(k)}(2 - \alpha^{(k)}) = +\infty$. Then, the sequences $(\mathbf{C}^{(k+\frac{1}{2})})_{k\geqslant 0}$ and $\left(\operatorname{prox}_{\gamma g_1}(2\mathbf{C}^{(k+\frac{1}{2})} - \mathbf{C}^{(k)})\right)_{k\geqslant 0}$ generated by Algorithm 1 converge to a solution to Problem (1) where $g = g_0 + g_1$.*

[1] See also http://proximity-operator.net.

Algorithm 1 Douglas–Rachford Algorithm for solving Problem (1)

Let $\mathbf{T}$ be a given matrix in $\mathcal{S}_n$, set $\gamma > 0$ and $\mathbf{C}^{(0)} \in \mathcal{S}_n$.
for $k = 0, 1, \ldots$ **do**
  Diagonalize $\mathbf{C}^{(k)} + \gamma\mathbf{T}$, i.e. find $\mathbf{U}^{(k)} \in \mathcal{O}_n$ and $\boldsymbol{\lambda}^{(k)} \in \mathbb{R}^n$ such that

$$\mathbf{C}^{(k)} + \gamma\mathbf{T} = \mathbf{U}^{(k)} \operatorname{Diag}(\boldsymbol{\lambda}^{(k)})(\mathbf{U}^{(k)})^\top$$

  $\mathbf{d}^{(k+\frac{1}{2})} \in \operatorname{Prox}_{\gamma(\varphi+\psi)}\left(\boldsymbol{\lambda}^{(k)}\right)$
  $\mathbf{C}^{(k+\frac{1}{2})} = \mathbf{U}^{(k)} \operatorname{Diag}(\mathbf{d}^{(k+\frac{1}{2})})(\mathbf{U}^{(k)})^\top$
  Choose $\alpha^{(k)} \in [0, 2]$
  $\mathbf{C}^{(k+1)} \in \mathbf{C}^{(k)} + \alpha^{(k)}\Big(\operatorname{Prox}_{\gamma g_1}(2\mathbf{C}^{(k+\frac{1}{2})} - \mathbf{C}^{(k)}) - \mathbf{C}^{(k+\frac{1}{2})}\Big)$.
end for

We have restricted the above convergence analysis to the convex case. Note however that recent convergence results for the DR algorithm in a non-convex setting are available in [1, 39] for specific choices of the involved functionals.

3.3. Positive Semi-Definite Constraint. Instead of solving Problem (1), one may be interested in:

$$\underset{\mathbf{C}\in\mathcal{S}_n^+}{\text{minimize}}\ f(\mathbf{C}) - \operatorname{trace}(\mathbf{C}\mathbf{T}) + g(\mathbf{C}), \tag{44}$$

when $\operatorname{dom} f \cap \operatorname{dom} g \not\subset \mathcal{S}_n^+$. This problem can be recast as minimizing over $\mathcal{S}_n$ $f - \operatorname{trace}(\cdot\mathbf{T}) + \widetilde{g}_0 + g_1$ where $\widetilde{g}_0 = g_0 + \iota_{\mathcal{S}_n^+}$. We are thus coming back to the original formulation where $\widetilde{g}_0$ has been substituted for g_0. In order to solve this problem with the proposed proximal approach, a useful result is stated below.

PROPOSITION 3.5. *Let $\gamma \in]0, +\infty[$ and $\overline{\mathbf{C}} \in \mathcal{S}_n$. Let f and g_0 be functions satisfying (4) and (5), respectively, where $\varphi \in \Gamma_0(\mathbb{R}^n)$ and $\psi \in \Gamma_0(\mathbb{R}^n)$. Assume that*

$$\left(\forall \boldsymbol{\lambda}' = (\lambda_i')_{1\leqslant i\leqslant n} \in \mathbb{R}^n\right) \quad \varphi(\boldsymbol{\lambda}') + \psi(\boldsymbol{\lambda}') = \sum_{i=1}^n \rho_i(\lambda_i') \tag{45}$$

where, for every $i \in \{1, \ldots, n\}$, $\rho_i \colon \mathbb{R} \to]-\infty, +\infty]$ is such that $\operatorname{dom}\rho_i \cap [0, +\infty[\neq \varnothing$. Let $\boldsymbol{\lambda} = (\lambda_i)_{1\leqslant i\leqslant n} \in \mathbb{R}^n$ and $\mathbf{U} \in \mathcal{O}_n$ be such that $\overline{\mathbf{C}} + \gamma\mathbf{T} = \mathbf{U}\operatorname{Diag}(\boldsymbol{\lambda})\mathbf{U}^\top$. Then

$$\operatorname{prox}_{\gamma(f-\operatorname{trace}(\mathbf{T}\,\cdot)+\widetilde{g}_0)}(\overline{\mathbf{C}}) = \mathbf{U}\operatorname{Diag}\left(\left(\max(0, \operatorname{prox}_{\gamma\rho_i}(\lambda_i))\right)_{1\leqslant i\leqslant n}\right)\mathbf{U}^\top. \tag{46}$$

Proof. Expression (46) readily follows from Proposition 3.1(ii) and [21, Proposition 2.2]. ∎

4. Application to Covariance Matrix Estimation. Estimating the covariance matrix of a random vector is a key problem in statistics, signal processing over graphs, and machine learning. Nonetheless, in existing optimization techniques, little attention is usually paid to the presence of noise corrupting the available observations. We show in this section how the results obtained in the previous sections can be used to tackle this problem in various contexts.

4.1. Model and Proposed Approaches. Let $\mathbf{S} \in \mathcal{S}_n^+$ be a sample estimate of a covariance matrix $\mathbf{\Sigma}$ which is assumed to be decomposed as

$$\mathbf{\Sigma} = \mathbf{Y}^* + \sigma^2 \mathrm{I_d} \tag{47}$$

where $\sigma \in [0, +\infty[$ and $\mathbf{Y}^* \in \mathcal{S}_n^+$ may have a low-rank structure. Our objective in this section will be to propose variational methods to provide an estimate of $\mathbf{Y}^*$ from $\mathbf{S}$ by assuming that σ is known. Such a problem arises when considering the following observation model [59]:

$$(\forall i \in \{1, \ldots, N\}) \quad \mathbf{x}^{(i)} = \mathbf{A}\mathbf{s}^{(i)} + \mathbf{e}^{(i)} \tag{48}$$

where $\mathbf{A} \in \mathbb{R}^{n \times m}$ with $m \leqslant n$ and, for every $i \in \{1, \ldots, N\}$, $\mathbf{s}^{(i)} \in \mathbb{R}^m$ and $\mathbf{e}^{(i)} \in \mathbb{R}^n$ are realizations of mutually independent identically distributed Gaussian multivalued random variables with zero mean and covariance matrices $\mathbf{P} \in \mathcal{S}_m^{++}$ and $\sigma^2 \mathrm{I_d}$, respectively. This model has been employed for instance in [60, 63] in the context of the "Relevant Vector Machine problem". The covariance matrix $\mathbf{\Sigma}$ of the noisy input data $\left(\mathbf{x}^{(i)}\right)_{1 \leqslant i \leqslant N}$ takes the form (47) with $\mathbf{Y}^* = \mathbf{A}\mathbf{P}\mathbf{A}^\top$. On the other hand, a simple estimate of $\mathbf{\Sigma}$ from the observed data $\left(\mathbf{x}^{(i)}\right)_{1 \leqslant i \leqslant N}$ is

$$\mathbf{S} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}^{(i)} \left(\mathbf{x}^{(i)}\right)^\top. \tag{49}$$

Covariance-based model. A first estimate $\widehat{\mathbf{Y}}$ of $\mathbf{Y}^*$ is given by

$$\widehat{\mathbf{Y}} = \underset{\mathbf{Y} \in \mathcal{S}_n^+}{\operatorname{argmin}} \ \frac{1}{2} \|\mathbf{Y} - \mathbf{S} + \sigma^2 \mathrm{I_d}\|_{\mathrm{F}}^2 + g_0(\mathbf{Y}) + g_1(\mathbf{Y}), \tag{50}$$

where $\mathbf{S}$ is the empirical covariance matrix, g_0 satisfies (5) with $\psi \in \Gamma_0(\mathbb{R}^n)$, $g_1 \in \Gamma_0(\mathcal{S}_n)$, and the intersection of the relative interiors of the domains of g_0 and g_1 is assumed to be non empty. A particular instance of this model with $\sigma = 0$, $g_0 = \mu_0 \mathcal{R}_1$, $g_1 = \mu_1 \|\cdot\|_1$, and $(\mu_0, \mu_1) \in [0, +\infty[^2$ was investigated in [72] and [54] for estimating sparse low-rank covariance matrices. In the latter reference, an application to real data processing arising from protein interaction and social network analysis is presented. One can observe that Problem (50) takes the form (44) by setting $f = \frac{1}{2}\|\cdot\|_{\mathrm{F}}^2$ and $\mathbf{T} = \mathbf{S} - \sigma^2 \mathrm{I_d}$. This allows us to solve (50) with Algorithm 1. Since it is assumed that g_0 satisfies (5), the proximity step on $f + g_0 + \iota_{\mathcal{S}_n^+}$ can be performed by employing Proposition 3.5 and formulas from Table 1. The resulting Douglas–Rachford procedure can thus be viewed as an alternative to the methods developed in [54] and [72]. Let us emphasize that these two algorithms were devised to solve an instance of (50) corresponding to the aforementioned specific choices for g_0 and g_1, while our approach leaves more freedom in the choice of the regularization functions.

Precision-based model. An alternative strategy consists of focusing on the estimation of the inverse of the covariance matrix, i.e. the *precision* matrix $\mathbf{C}^* = (\mathbf{Y}^*)^{-1}$ by assuming that $\mathbf{Y}^* \in \mathcal{S}_n^{++}$ but may have very small eigenvalues in order to model a possible low-rank structure. Tackling the problem from this viewpoint leads us to propose the following penalized negative log-likelihood cost function:

$$(\forall \mathbf{C} \in \mathcal{S}_n) \qquad \mathcal{F}(\mathbf{C}) = f(\mathbf{C}) + \mathcal{T}_{\mathbf{S}}(\mathbf{C}) + g_0(\mathbf{C}) + g_1(\mathbf{C}) \tag{51}$$

where

$$(\forall \mathbf{C} \in \mathcal{S}_n) \qquad f(\mathbf{C}) = \begin{cases} \log\det\left(\mathbf{C}^{-1} + \sigma^2 \mathrm{I_d}\right) & \text{if } \mathbf{C} \in \mathcal{S}_n^{++} \\ +\infty & \text{otherwise,} \end{cases} \tag{52}$$

$$(\forall \mathbf{C} \in \mathcal{S}_n) \qquad \mathcal{T}_{\mathbf{S}}(\mathbf{C}) = \begin{cases} \operatorname{trace}\left(\left(\mathrm{I_d} + \sigma^2 \mathbf{C}\right)^{-1} \mathbf{C}\mathbf{S}\right) & \text{if } \mathbf{C} \in \mathcal{S}_n^{+} \\ +\infty & \text{otherwise,} \end{cases} \tag{53}$$

$g_0 \in \Gamma_0(\mathcal{S}_n)$ satisfies (5) with $\psi \in \Gamma_0(\mathbb{R}^n)$, and $g_1 \in \Gamma_0(\mathcal{S}_n)$. Typical choices of interest for the latter two functions are

$$(\forall \mathbf{C} \in \mathcal{S}_n) \qquad g_0(\mathbf{C}) = \begin{cases} \mu_0 \mathcal{R}_1(\mathbf{C}^{-1}) & \text{if } \mathbf{C} \in \mathcal{S}_n^{++} \\ +\infty & \text{otherwise,} \end{cases} \tag{54}$$

and $g_1 = \mu_1 \|\cdot\|_1$ with $(\mu_0, \mu_1) \in [0, +\infty[^2$. The first function serves to promote a desired low-rank property by penalizing small eigenvalues of the precision matrix, whereas the second one enforces the sparsity of this matrix as it is usual in graph inference problems. This constitutes a main difference with respect to the covariance-based model which is more suitable to estimate sparse covariance matrices. Note that the standard Graphical Lasso framework [31] is then recovered by setting $\sigma = 0$ and $\mu_0 = 0$. The advantage of our formulation is that it allows us to consider more flexible variational models while accounting for the presence of noise corrupting the observed data. The main difficulty however is that Algorithm 1 cannot be directly applied to minimize $\mathcal{F}$. In Subsection 4.2, we will study in more details the properties of the cost function. This will allow us to derive a novel optimization algorithm making use of our previously developed Douglas-Rachford scheme for its inner steps

4.2. Study of Objective Function $\mathcal{F}$. The following lemma will reveal useful in our subsequent analysis.

LEMMA 4.1. *Let $\sigma \in]0, +\infty[$. Let $h\colon]0, \sigma^{-2}[\to \mathbb{R}$ be a twice differentiable function and let*

$$u\colon [0, +\infty[\to \mathbb{R}\colon \lambda \mapsto \frac{\lambda}{1 + \sigma^2 \lambda}. \tag{55}$$

The composition $h \circ u$ is convex on $]0, +\infty[$ if and only if

$$(\forall \upsilon \in]0, \sigma^{-2}[) \quad \ddot{h}(\upsilon)(1 - \sigma^2 \upsilon) - 2\sigma^2 \dot{h}(\upsilon) \geqslant 0, \tag{56}$$

where $\dot{h}$ (resp. $\ddot{h}$) denotes the first (resp. second) derivative of h.

Proof. The result directly follows from the calculation of the second-order derivative of $h \circ u$. ∎

Let us now note that f is a spectral function fulfilling (4) with

$$\left(\forall \boldsymbol{\lambda} = (\lambda_i)_{1 \leqslant i \leqslant n} \in \mathbb{R}^n\right) \qquad \varphi(\boldsymbol{\lambda}) = \begin{cases} -\sum_{i=1}^{n} \log\left(u(\lambda_i)\right) & \text{if } \boldsymbol{\lambda} \in]0, +\infty[^n \\ +\infty & \text{otherwise,} \end{cases} \tag{57}$$

where u is defined by (55). According to Lemma 4.1 (with $h = -\log$), $f \in \Gamma_0(\mathcal{S}_n)$. Thus, the assumptions made on g_0 and g_1, allow us to deduce that $f + g_0 + g_1$ is convex and lower-semicontinuous on $\mathcal{S}_n$.

Let us now focus on the properties of the second term in (51).

LEMMA 4.2. *Let* $\mathbf{S} \in \mathcal{S}_n^+$. *The function* $\mathcal{T}_{\mathbf{S}}$ *in* (53) *is concave on* $\mathcal{S}_n^+$.

Proof. By using differential calculus rules in [45], we will show that the Hessian of $-\mathcal{T}_{\mathbf{S}}$ evaluated at any matrix in $\mathcal{S}_n^{++}$ is a positive semidefinite operator. In order to lighten our notation, for every invertible matrix $\mathbf{C}$, let us define $\mathbf{M} = \mathbf{C}^{-1} + \sigma^2 \mathrm{I_d}$. Then, the first-order differential of $\mathcal{T}_{\mathbf{S}}$ at every $\mathbf{C} \in \mathcal{S}_n^{++}$ is

$$\begin{aligned}
\mathrm{d}\,\mathrm{trace}\,(\mathcal{T}_{\mathbf{S}}(\mathbf{C})) &= \mathrm{trace}\left(\left(\mathrm{d}\,\mathbf{M}^{-1}\right)\mathbf{S}\right) \\
&= \mathrm{trace}\left(-\mathbf{M}^{-1}(\mathrm{d}\,\mathbf{M})\mathbf{M}^{-1}\mathbf{S}\right) \\
&= \mathrm{trace}\left(\left(\mathbf{C}^{-1} + \sigma^2\mathrm{I_d}\right)^{-1}\mathbf{S}\left(\mathbf{C}^{-1} + \sigma^2\mathrm{I_d}\right)^{-1}\mathbf{C}^{-1}(\mathrm{d}\,\mathbf{C})\mathbf{C}^{-1}\right) \\
&= \mathrm{trace}\left(\left(\mathrm{I_d} + \sigma^2\mathbf{C}\right)^{-1}\mathbf{S}\left(\mathrm{I_d} + \sigma^2\mathbf{C}\right)^{-1}(\mathrm{d}\,\mathbf{C})\right). \qquad (58)
\end{aligned}$$

We have used the expression of the differential of the inverse [45, Chapter 8, Theorem 3] and the invariance of the trace with respect to cyclic permutations. It follows from (58) that the gradient of $\mathcal{T}_{\mathbf{S}}$ reads

$$(\forall \mathbf{C} \in \mathcal{S}_n^{++}) \quad \nabla\mathcal{T}_{\mathbf{S}}(\mathbf{C}) = \left(\mathrm{I_d} + \sigma^2\mathbf{C}\right)^{-1}\mathbf{S}\left(\mathrm{I_d} + \sigma^2\mathbf{C}\right)^{-1}. \qquad (59)$$

In order to calculate the Hessian $\mathfrak{H}$ of $\mathcal{T}_{\mathbf{S}}$, we calculate the differential of $\nabla\mathcal{T}_{\mathbf{S}}$. Again, in order to simplify our notation, for every matrix $\mathbf{C}$, we define

$$\mathbf{N} = \mathrm{I_d} + \sigma^2\mathbf{C} \quad \Rightarrow \quad \mathrm{d}\,\mathbf{N} = \sigma^2\,\mathrm{d}\,\mathbf{C}. \qquad (60)$$

The differential of $\nabla\mathcal{T}_{\mathbf{S}}$ at every $\mathbf{C} \in \mathcal{S}_n^{++}$ then reads

$$\begin{aligned}
\mathrm{d}\,\mathrm{vect}\,(\nabla\mathcal{T}_{\mathbf{S}}(\mathbf{C})) &= \mathrm{vect}\left(\mathrm{d}(\mathbf{N}^{-1}\mathbf{S}\mathbf{N}^{-1})\right) \\
&= \mathrm{vect}\left((\mathrm{d}\,\mathbf{N}^{-1})\mathbf{S}\mathbf{N}^{-1} + \mathbf{N}^{-1}(\mathrm{d}\,\mathbf{S}\mathbf{N}^{-1})\right) \\
&= -\,\mathrm{vect}(\mathbf{N}^{-1}(\mathrm{d}\,\mathbf{N})\mathbf{N}^{-1}\mathbf{S}\mathbf{N}^{-1}) - \mathrm{vect}\left(\mathbf{N}^{-1}\mathbf{S}\mathbf{N}^{-1}(\mathrm{d}\,\mathbf{N})\mathbf{N}^{-1}\right) \\
&= -\left(\left(\mathbf{N}^{-1}\mathbf{S}\mathbf{N}^{-1}\right)^\top \otimes \mathbf{N}^{-1}\right)\mathrm{vect}(\mathrm{d}\,\mathbf{N}) - \left(\left(\mathbf{N}^{-1}\right)^\top \otimes \mathbf{N}^{-1}\mathbf{S}\mathbf{N}^{-1}\right)\mathrm{vect}(\mathrm{d}\,\mathbf{N}) \\
&= -\left(\left(\mathbf{N}^{-1}\mathbf{S}\mathbf{N}^{-1}\right) \otimes \mathbf{N}^{-1} + \mathbf{N}^{-1} \otimes \left(\mathbf{N}^{-1}\mathbf{S}\mathbf{N}^{-1}\right)\right)\mathrm{d}\,\mathrm{vect}(\mathbf{N}) \\
&= \mathfrak{H}(\mathbf{C})\,\mathrm{d}\,\mathrm{vect}(\mathbf{C})
\end{aligned}$$

with

$$\mathfrak{H}(\mathbf{C}) = -\sigma^2\left(\nabla\mathcal{T}_{\mathbf{S}}(\mathbf{C}) \otimes \left(\mathrm{I_d} + \sigma^2\mathbf{C}\right)^{-1} + \left(\mathrm{I_d} + \sigma^2\mathbf{C}\right)^{-1} \otimes \nabla\mathcal{T}_{\mathbf{S}}(\mathbf{C})\right). \qquad (61)$$

To derive the above expression, we have used the facts that, for every $\mathbf{A} \in \mathbb{R}^{n\times m}$, $\mathbf{X} \in \mathbb{R}^{m\times p}$, and $\mathbf{B} \in \mathbb{R}^{p\times q}$, $\mathrm{vect}\,(\mathbf{AXB}) = \left(\mathbf{B}^\top \otimes \mathbf{A}\right)\mathrm{vect}\,\mathbf{X}$ [45, Chapter 2,Theorem 2] and that matrices $\mathbf{N}$ and $\mathbf{S}$ are symmetric.

Let us now check that, for every $\mathbf{C} \in \mathcal{S}_n^{++}$, $\mathfrak{H}(\mathbf{C})$ is negative semidefinite. It follows from expression (59), the symmetry of $\mathbf{C}$, and the positive semidefiniteness of $\mathbf{S}$ that $\nabla\mathcal{T}_{\mathbf{S}}(\mathbf{C})$ belongs to $\mathcal{S}_n^+$. Since

$$\begin{aligned}
\left(\nabla\mathcal{T}_{\mathbf{S}}(\mathbf{C}) \otimes \left(\mathrm{I_d} + \sigma^2\mathbf{C}\right)^{-1}\right)^\top &= \left(\nabla\mathcal{T}_{\mathbf{S}}(\mathbf{C})\right)^\top \otimes \left(\left(\mathrm{I_d} + \sigma^2\mathbf{C}\right)^{-1}\right)^\top \\
&= \nabla\mathcal{T}_{\mathbf{S}}(\mathbf{C}) \otimes \left(\mathrm{I_d} + \sigma^2\mathbf{C}\right)^{-1},
\end{aligned}$$

$\nabla\mathcal{T}_{\mathbf{S}}(\mathbf{C}) \otimes \left(\mathrm{I_d} + \sigma^2\mathbf{C}\right)^{-1}$ is symmetric. Let us denote by $(\gamma_i)_{1\leqslant i\leqslant n} \in [0,+\infty[^n$ the eigenvalues of $\nabla\mathcal{T}_{\mathbf{S}}(\mathbf{C})$ and by $(\zeta_i)_{1\leqslant i\leqslant n} \in [0,+\infty[^n$ those of of $\mathbf{C}$. According to [45, Chapter 2, Theorem 1], the eigenvalues of $\nabla\mathcal{T}_{\mathbf{S}}(\mathbf{C}) \otimes \left(\mathrm{I_d} + \sigma^2\mathbf{C}\right)^{-1}$ are

$\big(\gamma_i/(1+\sigma^2\zeta_j)\big)_{1\leqslant i,j\leqslant n}$ and they are therefore nonnegative. This allows us to claim that $\nabla\mathcal{T}_{\mathbf{S}}(\mathbf{C})\otimes\big(\mathrm{I_d}+\sigma^2\mathbf{C}\big)^{-1}$ belongs to $\mathcal{S}_{n^2}^{+}$. For similar reasons, $\big(\mathrm{I_d}+\sigma^2\mathbf{C}\big)^{-1}\otimes\nabla\mathcal{T}_{\mathbf{S}}(\mathbf{C})\in\mathcal{S}_{n^2}^{+}$, which allows us to conclude that $-\mathfrak{H}(\mathbf{C})\in\mathcal{S}_{n^2}^{+}$. Hence, we have proved that $\mathcal{T}_{\mathbf{S}}$ is concave on $\mathcal{S}_n^{++}$. By continuity of $\mathcal{T}_{\mathbf{S}}$ relative to $\mathcal{S}_n^{+}$, the concavity property extends on $\mathcal{S}_n^{+}$. ∎

As a last worth mentioning property, $\mathcal{T}_{\mathbf{S}}$ is bounded on $\mathcal{S}_n^{++}$. So, if $\operatorname{dom} f\cap\operatorname{dom} g_0\cap\operatorname{dom} g_1\neq\varnothing$ and $f+g_0+g_1$ is coercive, then there exists a minimizer of $\mathcal{F}$. Because of the form of f, the coercivity condition is satisfied if g_0+g_1 is lower bounded and $\lim_{\mathbf{C}\in\mathcal{S}_n^{+},\|\mathbf{C}\|\to+\infty} g_0(\mathbf{C})+g_1(\mathbf{C})=+\infty$.

4.3. Minimization Algorithm for $\mathcal{F}$. In order to find a minimizer of $\mathcal{F}$, we propose a *Majorize–Minimize* (MM) approach, following the ideas in [22, 59, 35, 36]. At each iteration of an MM algorithm, one constructs a tangent function that majorizes the given cost function and is equal to it at the current iterate. The next iterate is obtained by minimizing this tangent majorant function, resulting in a sequence of iterates that reduces the cost function value monotonically. According to the results stated in the previous section, our objective function reads as a difference of convex terms. We propose to build a majorizing approximation of function $\mathcal{T}_{\mathbf{S}}$ at $\mathbf{C}'\in\mathcal{S}_n^{++}$ by exploiting Lemma 4.2 and the classical concavity inequality on $\mathcal{T}_{\mathbf{S}}$:

$$(\forall\mathbf{C}\in\mathcal{S}_n^{++})\quad \mathcal{T}_{\mathbf{S}}(\mathbf{C})\leqslant\mathcal{T}_{\mathbf{S}}(\mathbf{C}')+\operatorname{trace}\left(\nabla\mathcal{T}_{\mathbf{S}}(\mathbf{C}')\,(\mathbf{C}-\mathbf{C}')\right). \tag{62}$$

As f is finite only on $\mathcal{S}_n^{++}$, a tangent majorant of the cost function (51) at $\mathbf{C}'$ reads:

$$(\forall\mathbf{C}\in\mathcal{S}_n)\quad \mathcal{G}(\mathbf{C}\mid\mathbf{C}')=f(\mathbf{C})+\mathcal{T}_{\mathbf{S}}(\mathbf{C}')+\operatorname{trace}\left(\nabla\mathcal{T}_{\mathbf{S}}(\mathbf{C}')\,(\mathbf{C}-\mathbf{C}')\right)+g_0(\mathbf{C})+g_1(\mathbf{C}).$$

This leads to the general MM scheme:

$$(\forall\ell\in\mathbb{N})\quad \mathbf{C}^{(\ell+1)}\in\underset{\mathbf{C}\in\mathcal{S}_n}{\operatorname{Argmin}}\ f(\mathbf{C})+\operatorname{trace}\big(\nabla\mathcal{T}_{\mathbf{S}}(\mathbf{C}^{(\ell)})\mathbf{C}\big)+g_0(\mathbf{C})+g_1(\mathbf{C}) \tag{63}$$

with $\mathbf{C}^{(0)}\in\mathcal{S}_n^{++}$. At each iteration of the MM algorithm, we have then to solve a convex optimization problem of the form (1). In the case when $g_1\equiv 0$, we can employ the procedure described in Section 2 to perform this task in a direct manner. The presence of a regularization term $g_1\not\equiv 0$ usually prevents us to have an explicit solution to the inner minimization problem involved in the MM procedure. We then propose in Algorithm 2 to resort to the Douglas–Rachford approach in Section 3 to solve it iteratively.

Algorithm 2 MM algorithm with DR inner steps

Let $\mathbf{S} \in \mathcal{S}_n^+$ be the data matrix. Let φ be as in (57), let $\psi \in \Gamma_0(\mathbb{R}^n)$ be associated with g_0. Let $(\gamma_\ell)_{\ell\in\mathbb{N}}$ be a sequence in $]0,+\infty[$. Set $\mathbf{C}^{(0,0)} = \mathbf{C}^{(0)} \in \mathcal{S}_n^{++}$.
for $\ell = 0, 1, \ldots$ **do**
  for $k = 0, 1, \ldots$ **do**
    Compute $\mathbf{U}^{(\ell,k)} \in \mathcal{O}_n$ and $\boldsymbol{\lambda}^{(\ell,k)} \in \mathbb{R}^n$ such that

$$\mathbf{C}^{(\ell,k)} - \gamma_\ell \nabla \mathcal{T}_{\mathbf{S}}(\mathbf{C}^{(\ell)}) = \mathbf{U}^{(\ell,k)} \operatorname{Diag}(\boldsymbol{\lambda}^{(\ell,k)}) \left(\mathbf{U}^{(\ell,k)}\right)^\top$$

    $\mathbf{d}^{\left(\ell,k+\frac{1}{2}\right)} = \operatorname{prox}_{\gamma_\ell(\varphi+\psi)}\left(\boldsymbol{\lambda}^{(\ell,k)}\right)$
    $\mathbf{C}^{\left(\ell,k+\frac{1}{2}\right)} = \mathbf{U}^{(\ell,k)} \operatorname{Diag}\left(\mathbf{d}^{\left(\ell,k+\frac{1}{2}\right)}\right) \left(\mathbf{U}^{(\ell,k)}\right)^\top$
    if Convergence of MM sub-iteration is reached **then**
      $\mathbf{C}^{(\ell+1)} = \mathbf{C}^{\left(\ell,k+\frac{1}{2}\right)}$
      $\mathbf{C}^{(\ell+1,0)} = \mathbf{C}^{(\ell,k)}$
      exit inner loop
    end if
    Choose $\alpha_{\ell,k} \in]0,2[$
    $\mathbf{C}^{(\ell,k+1)} = \mathbf{C}^{(\ell,k)} + \alpha_{\ell,k}\left(\operatorname{prox}_{\gamma_\ell g_1}\left(2\mathbf{C}^{\left(\ell,k+\frac{1}{2}\right)} - \mathbf{C}^{(\ell,k)}\right) - \mathbf{C}^{\left(\ell,k+\frac{1}{2}\right)}\right)$
  end for
end for

A convergence result is next stated, which is inspired from [64] (itself relying on [69, p. 6]), but does not require the differentiability of $g_0 + g_1$.

THEOREM 4.3. *Let $(\mathbf{C}^{(\ell)})_{\ell\geqslant 0}$ be a sequence generated by (63). Assume that $\operatorname{dom} f \cap \operatorname{dom} g_0 \cap \operatorname{dom} g_1 \neq \varnothing$, $f + g_0 + g_1$ is coercive, and $E = \{\mathbf{C} \in \mathcal{S}_n \mid \mathcal{F}(\mathbf{C}) \leqslant \mathcal{F}(\mathbf{C}^{(0)})\}$ is a subset of the relative interior of $\operatorname{dom} g_0 \cap \operatorname{dom} g_1$. Then, the following properties hold:*

(i) *$\left(\mathcal{F}(\mathbf{C}^{(\ell)})\right)_{\ell\geqslant 0}$ is a decaying sequence converging to $\widehat{\mathcal{F}} \in \mathbb{R}$.*
(ii) *$(\mathbf{C}^{(\ell)})_{\ell\geqslant 0}$ has a cluster point.*
(iii) *Every cluster point $\widehat{\mathbf{C}}$ of $(\mathbf{C}^{(\ell)})_{\ell\geqslant 0}$ is such that $\mathcal{F}(\widehat{\mathbf{C}}) = \widehat{\mathcal{F}}$ and it is a critical point of $\mathcal{F}$, i.e. $-\nabla f(\widehat{\mathbf{C}}) - \nabla \mathcal{T}_{\mathbf{S}}(\widehat{\mathbf{C}}) \in \partial(g_0 + g_1)(\widehat{\mathbf{C}})$.*

Proof. First note that $(\mathbf{C}^{(\ell)})_{\ell\geqslant 0}$ is properly defined by (63) since, for every $\mathbf{C} \in \mathcal{S}_n^{++}$, $\mathcal{G}(\cdot \mid \mathbf{C})$ is a coercive lower-semicontinuous function. It indeed majorizes $\mathcal{F}$ which is coercive, since $f + g_0 + g_1$ has been assumed coercive.
(i): As a known property of MM strategies, $\left(\mathcal{F}(\mathbf{C}^{(\ell)})\right)_{\ell\geqslant 0}$ is a decaying sequence [36]. Under our assumptions, we have already seen that $\mathcal{F}$ has a minimizer. We deduce that $\left(\mathcal{F}(\mathbf{C}^{(\ell)})\right)_{\ell\geqslant 0}$ is lower bounded, hence convergent.
(ii): Since $\left(\mathcal{F}(\mathbf{C}^{(\ell)})\right)_{\ell\geqslant 0}$ is a decaying sequence, $(\forall \ell \geqslant 0)\ \mathbf{C}^{(\ell)} \in E$. Since $\mathcal{F}$ is proper, lower-semicontinuous, and coercive, E is a nonempty compact set and $(\mathbf{C}^{(\ell)})_{\ell\geqslant 0}$ admits a cluster point in E.
(iii): If $\widehat{\mathbf{C}}$ is a cluster point of $(\mathbf{C}^{(\ell)})_{\ell\geqslant 0}$, then there exists a subsequence $(\mathbf{C}^{(\ell_k)})_{k\geqslant 0}$ converging to $\widehat{\mathbf{C}}$. Since E is a nonempty subset of the relative interior of $\operatorname{dom} g_0 \cap \operatorname{dom} g_1$ and $g_0 + g_1 \in \Gamma_0(\mathcal{S}_n)$, $g_0 + g_1$ is continuous relative to E [6, Corollary 8.41]. As $f + \mathcal{T}_{\mathbf{S}}$ is continuous on $\operatorname{dom} f \cap \operatorname{dom} \mathcal{T}_{\mathbf{S}} = \mathcal{S}_n^{++}$, $\mathcal{F}$ is continuous relative to E. Hence, $\widehat{\mathcal{F}} = \lim_{k\to+\infty} \mathcal{F}(\mathbf{C}^{(\ell_k)}) = \mathcal{F}(\widehat{\mathbf{C}})$. On the other hand, by similar arguments applied to

sequence $(\mathbf{C}^{(\ell_k+1)})_{k\geqslant 0}$, there exists a subsequence $(\mathbf{C}^{(\ell_{k_q}+1)})_{q\geqslant 0}$ converging to some $\widehat{\mathbf{C}}' \in E$ such that $\widehat{\mathcal{F}} = \mathcal{F}(\widehat{\mathbf{C}}')$. In addition, thanks to (63), we have

$$(\forall \mathbf{C} \in \mathcal{S}_n)(\forall q \in \mathbb{N}) \quad \mathcal{G}(\mathbf{C}^{(\ell_{k_q}+1)} \mid \mathbf{C}^{(\ell_{k_q})}) \leqslant \mathcal{G}(\mathbf{C} \mid \mathbf{C}^{(\ell_{k_q})}). \tag{64}$$

By continuity of f and $\nabla \mathcal{T}_{\mathbf{S}}$ on $\mathcal{S}_n^{++}$ and by continuity of $g_0 + g_1$ relative to E,

$$(\forall \mathbf{C} \in \mathcal{S}_n) \quad \mathcal{G}(\widehat{\mathbf{C}}' \mid \widehat{\mathbf{C}}) \leqslant \mathcal{G}(\mathbf{C} \mid \widehat{\mathbf{C}}). \tag{65}$$

Let us now suppose that $\widehat{\mathbf{C}}$ is not a critical point of $\mathcal{F}$. Since the subdifferential of $\mathcal{G}(\cdot \mid \widehat{\mathbf{C}})$ at $\widehat{\mathbf{C}}$ is $\nabla f(\widehat{\mathbf{C}}) + \nabla \mathcal{T}_{\mathbf{S}}(\widehat{\mathbf{C}}) + \partial(g_0 + g_1)(\widehat{\mathbf{C}})$ [6, Corollary 16.48(ii)], the null matrix does not belong to this subdifferential, which means that $\widehat{\mathbf{C}}$ is not a minimizer of $\mathcal{G}(\cdot \mid \widehat{\mathbf{C}})$ [6, Theorem 16.3]. It follows from (65) and standard MM properties that $\mathcal{F}(\widehat{\mathbf{C}}') \leqslant \mathcal{G}(\widehat{\mathbf{C}}' \mid \widehat{\mathbf{C}}) < \mathcal{G}(\widehat{\mathbf{C}} \mid \widehat{\mathbf{C}}) = \mathcal{F}(\widehat{\mathbf{C}})$. The resulting strict inequality contradicts the already established fact that $\mathcal{F}(\widehat{\mathbf{C}}') = \mathcal{F}(\widehat{\mathbf{C}})$. ∎

5. Numerical Experiments. This section presents some numerical tests illustrating the validity of the proposed algorithms. More specifically, in Subsection 5.1 the Douglas–Rachford (DR) approach of Section 3 is compared with other state–of–the–art algorithms previously mentioned, namely Incremental Proximal Descent (IPD) [54] and ADMM [72], on a problem of covariance matrix estimation. In Subsection 5.2, we present an application of the MM approach from Section 4 to a graphical lasso problem in the presence of noisy data. All the experiments were conducted on a MacBook Pro equipped with an Intel Core i7 at 2.2 GHz, 16 Gb of RAM (DDR3 1600 MHz), and Matlab R2015b.

5.1. Application to Sparse Covariance Matrix Estimation. We first consider the application of the DR algorithm from Section 3 to the sparse covariance matrix estimation problem introduced in [54]. The objective is to retrieve an estimate of a low rank covariance matrix $\mathbf{Y}^* \in \mathcal{S}_n^+$ from N noisy realizations $(\mathbf{x}^{(i)})_{1\leqslant i\leqslant N}$ of a Gaussian multivalued random vector with zero mean and covariance matrix $\mathbf{Y}^* + \sigma^2 \mathrm{I_d}$, with $\sigma > 0$. As we have shown in Subsection 4.1, a solution to this problem can be obtained by solving the penalized least-squares problem (50), where $\mathbf{S}$ is the empirical covariance matrix defined in (49), and the regularization terms are $g_0 = \mu_0 \mathcal{R}_1$ and $g_1 = \mu_1 \|\cdot\|_1$. We propose to compare the performance of the DR approach from Subsection 3.2, with the IPD algorithm [54] and the ADMM procedure [72], for solving this convex optimization problem.

The synthetic data are generated using a procedure similar to the one in [54]. A block-diagonal covariance matrix $\mathbf{Y}^*$ is considered, composed with r blocks with dimensions $(r_j)_{1\leqslant j\leqslant r}$, so that $n = \sum_{j=1}^r r_j$. The j-th diagonal block of $\mathbf{Y}^*$ reads as a product $\mathbf{a}_j \mathbf{a}_j^\top$, where the components of $\mathbf{a}_j \in \mathbb{R}^{r_j}$ are randomly drawn on $[-1, 1]$. The number of observations N is equal to n and $\sigma = 0.1$. The three algorithms are initialized with $\mathbf{S} + \mathrm{I_d}$, and stopped as soon as a relative decrease criterion on the objective function is met, i.e. when $|\mathcal{F}_{k+1} - \mathcal{F}_k|/|\mathcal{F}_k| \leqslant \varepsilon$, $\varepsilon > 0$ being a given tolerance and $\mathcal{F}_k$ denoting the objective function value at iteration k. The maximum number of iterations is set to 2000. The penalty parameters μ_1 and μ_0 are chosen in order to get a reliable estimation of the original covariance matrix. The gradient stepsize for IPD is set to k^{-1}. In Algorithm 1, α_k is set to 1.5. In ADMM, the initial Lagrange multiplier is set to a matrix with all entries equal to one, and the parameter of the proximal step is set to 1.

Figure 1 illustrates the quality of the recovered covariance matrices when setting $\varepsilon = 10^{-10}$. Three different indicators for estimation quality are provided, namely

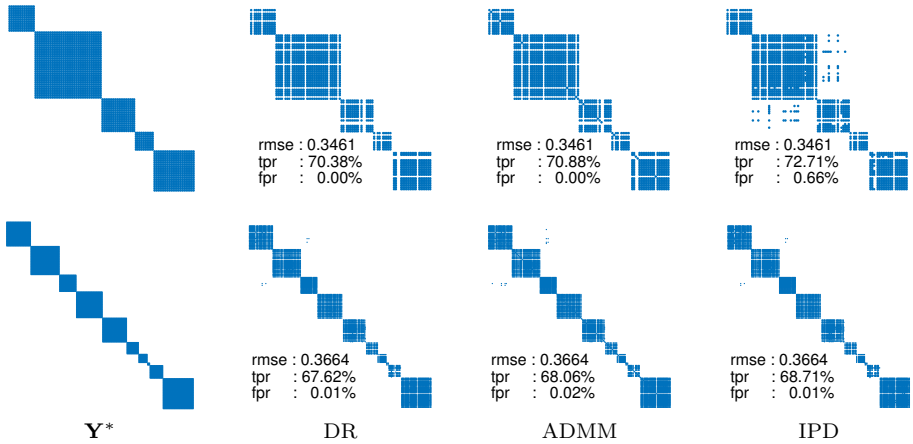


FIG. 1. *Original matrix and reconstruction results for DR, ADMM and IPD algorithms, for* $n = 100$ *(top) and* $n = 300$ *(bottom).*

the *true positive rate* (`tpr`), i.e. the correctly recognized non–zero entries, the *false positive rate* (`fpr`), i.e. the entries erroneously added to the support of the matrix, and the *relative mean square error* (`rmse`), computed as $\|\mathbf{Y}_{\text{rec}} - \mathbf{Y}^*\|_{\text{F}}^2 / \|\mathbf{Y}^*\|_{\text{F}}^2$, with $\mathbf{Y}_{\text{rec}}$ the recovered matrix. Note that the two first measurements are employed when the main interest lies in the recovery of the matrix support. A visual inspection shows that the three methods provide similar results in terms of matrix support estimation. Moreover, the reconstruction error as well as the values of `fpr` and `tpr` slightly differ.

TABLE 4

Comparison in terms of convergence speed between DR, ADMM and IPD procedures. The enlighten times refers to the shortest ones.

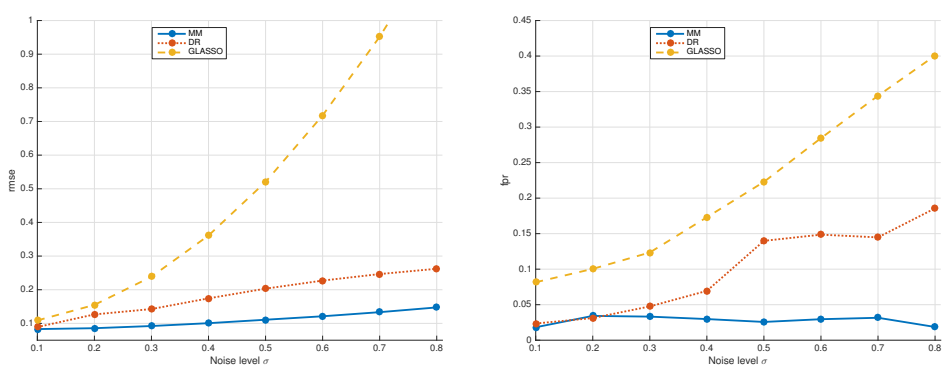
	$n = 100$, $\mu_0 = 0.2$, $\mu_1 = 0.1$, $r = 5$ $\{r_j\} = \{14, 36, 18, 10, 22\}$			$n = 300$, $\mu_0 = 0.01$, $\mu_1 = 0.12$ $r = 10$, $\{r_j\} = \{39, 46, 27, 42, 39, 19, 14, 4, 21, 49\}$		
	DR	ADMM	IPD	DR	ADMM	IPD
ε	Time(iter)	Time(iter)	Time(iter)	Time(iter)	Time(iter)	Time(iter)
10^{-6}	0.03 (23)	**0.02** (17)	0.18 (167)	0.14 (17)	**0.11** (14)	1.34 (170)
10^{-7}	0.03 (27)	**0.02** (21)	0.58 (533)	**0.32** (38)	0.34 (42)	4.35 (548)
10^{-8}	**0.03** (30)	0.04 (34)	1.83 (685)	**0.81** (95)	0.91 (115)	13.72 (1748)
10^{-9}	**0.06** (56)	**0.06** (54)	2.16 (2000)	**1.79** (211)	2.06 (258)	15.70 (2000)
10^{-10}	**0.07** (59)	**0.07** (58)	2.16 (2000)	**5.23** (620)	5.45 (686)	15.68 (2000)

Table 4 presents the comparative performance of the algorithms in terms of computation time (in second) and iteration number (averaged on 20 noise realizations), for two scenarios corresponding to distinct problem sizes and block distributions. It can be observed that the behaviors of ADMM and DR are similar, while IPD requires more iterations and time to reach the same precision. Furthermore, the latter fails to reach a high precision in the allowed maximum number of iterations, for both examples.

5.2. Application to Robust Graphical Lasso. Let us now illustrate the applicability of the MM approach presented in Subsection 4.3 to the problem of

precision matrix estimation introduced in (51). The test datasets have been generated by using the code available at http://stanford.edu/boyd/papers/admm/covsel/covsel_example.html. A sparse precision matrix $\mathbf{C}^*$ of dimension $n \times n$ is randomly created, where the number of non–zero entries is chosen as a proportion $p \in]0,1[$ of the total number n^2. Then, N realizations $(\mathbf{x}^{(i)})_{1 \leqslant i \leqslant N}$ of a Gaussian multivalued random variable with zero mean and covariance $\mathbf{Y}^* = (\mathbf{C}^*)^{-1}$ are generated. Gaussian noise with zero mean and covariance $\sigma^2 \mathrm{I}_{\mathrm{d}}$, $\sigma > 0$, is finally added to the $\mathbf{x}^{(i)}$'s, so that the covariance matrix $\mathbf{\Sigma}$ associated with the input data reads as in (47) with $\mathbf{A} = \mathrm{I}_{\mathrm{d}}$. As explained in Subsection 4.1, the estimation of $\mathbf{C}^*$ can be performed by using the MM algorithm from Subsection 4.3 based on the minimization of the nonconvex cost (51) with regularization functions $g_1 = \mu_1 \| \cdot \|_1$, $\mu_1 > 0$, and $(\forall \mathbf{C} \in \mathcal{S}_n^{++})$ $g_0(\mathbf{C}) = \mu_0 \mathcal{R}_1\left(\mathbf{C}^{-1}\right)$, $\mu_0 > 0$. The computation of $\mathrm{prox}_{\gamma(\varphi+\psi)}$ with $\gamma \in]0,+\infty[$ related to this particular choice for g_0 and function φ given by (57) and (55) leads to the search of the only positive root of a polynomial of degree 4.

A synthetic dataset of size $n = 100$ is created, where matrix $\mathbf{C}^*$ has 20 off-diagonal non-zero entries (i.e., $p = 10^{-3}$) and the corresponding covariance matrix has condition number 0.125. $N = 1000$ realizations are used to compute the empirical covariance matrix $\mathbf{S}$. In our MM algorithm, the inner stopping criterion (line 7 in Algorithm 2) is based on the relative difference of majorant function values with a tolerance of 10^{-10}, while the outer cycle is stopped when the relative difference of the objective function values falls below 10^{-8}. The DR algorithm is used to solve the inner subproblems, by using parameters $(\forall \ell)$ $\gamma_\ell = 1$, $(\forall k)$ $\alpha_{\ell,k} = 1$ (see Algorithm 2, lines 4–13). The allowed maximum inner (resp. outer) iteration number is 2000 (resp. 20). The quality of the results is quantified in terms of `fpr` on the precision matrix and `rmse` with respect to the true covariance matrix. The parameters μ_1 and μ_0 are set in order to obtain the best reconstruction in terms of `rmse`. For eight values of the noise standard deviation σ, Figure 2 illustrates the reconstruction quality (averaged on 20 noise realizations) obtained with our method, as well as two other approaches that do not take into account the noise in their formulation, namely the classical GLASSO approach from [12], which amounts to solve (1) with $f = -\log\det$, $g = \mu_1 \| \cdot \|_1$, and the DR approach described in Section 3, in the formulation given by (1) with $f = -\log\det$, $(\forall \mathbf{C} \in \mathcal{S}_n^{++})$ $g(\mathbf{C}) = \mu_0 \mathcal{R}_1\left(\mathbf{C}^{-1}\right) + \mu_1 \|\mathbf{C}\|_1$. For the DR approach, $\mathrm{prox}_{\gamma(\varphi+\psi)}$ with $\gamma \in]0,+\infty[$ is given by the fourth line of Table 2 (when $p = 1$).



(a) Behaviour of rmse wrt σ. (b) Behaviour of fpr wrt σ.

FIG. 2. *Estimation results for different noise levels in terms of* `rmse` *(left) and* `fpr` *(right) for MM, GLASSO and DR approaches.*

As expected, as the noise variance increases the reconstruction quality deteriorates. The GLASSO procedure is strongly impacted by the presence of noise, whereas the MM approach achieves better results, also when compared with DR algorithm. Moreover, the MM algorithm significantly outperforms both other methods in terms of support reconstruction, revealing itself very robust with respect to an increasing level of noise.

6. Conclusions. In this work, various proximal tools have been introduced to deal with optimization problems involving real symmetric matrices. We have focused on the variational framework (1) which is closely related to the computation of a proximity operator with respect to a Bregman divergence. It has been assumed that f in (3) is a convex spectral function, and g reads as $g_0 + g_1$, where g_0 is a spectral function. We have given a fully spectral solution in Section 2 when $g_1 \equiv 0$, and, in particular, Corollary 2.6 could be useful for developing algorithms involving proximity operators in other metrics than the Frobenius one. When $g_1 \not\equiv 0$, a proximal iterative approach has been presented, which is grounded on the use of the Douglas–Rachford procedure. As illustrated by the tables of proximity operators provided for a wide range of choices for f and g_0, the main advantage of the proposed algorithm is its great flexibility. The proposed framework also has allowed us to propose a nonconvex formulation of the precision matrix estimation problem arising in the context of noisy graphical lasso. The nonconvexity of the obtained objective function has been cirmcumvented through a Majorization–Minimization approach, each step of which consists of solving a convex problem by a Douglas-Rachford sub-iteration. Comparisons with state–of–the–art solutions have demonstrated the robustness of the proposed method. It is worth mentioning that all the results presented in this paper can be easily extended to complex Hermitian matrices.

REFERENCES

[1] F. J. Aragón Artacho and J. M. Borwein, *Global convergence of a non-convex Douglas–Rachford iteration*, J. Global Optim., 57 (2013), pp. 753–769, https://doi.org/10.1007/s10898-012-9958-4.

[2] M. S. Aslan, X.-W. Chen, and H. Cheng, *Analyzing and learning sparse and scale-free networks using Gaussian graphical models*, J. Mach. Learn. Res., 1 (2016), pp. 99–109, https://doi.org/10.1007/s41060-016-0009-y.

[3] O. Banerjee, L. El Ghaoui, and A. d'Aspremont, *Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data*, J. Mach. Learn. Res., 9 (2008), pp. 485–516.

[4] H. H. Bauschke, J. M. Borwein, and P. L. Combettes, *Essential smoothness, essential strict convexity, and Legendre functions in Banach spaces*, Comm. Contemp. Math, 3 (2001), pp. 615–647.

[5] H. H. Bauschke, J. M. Borwein, and P. L. Combettes, *Bregman monotone optimization algorithms*, SIAM J. Control Optim., 42 (2003), pp. 596–636, https://doi.org/10.1137/S0363012902407120.

[6] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer International Publishing, 2nd ed., 2017, https://doi.org/10.1007/978-3-319-48311-5.

[7] H. H. Bauschke, P. L. Combettes, and D. Noll, *Joint minimization with alternating Bregman proximity operators*, Pac. J. Optim., 2 (2006), pp. 401–424.

[8] A. Benfenati and V. Ruggiero, *Inexact Bregman iteration with an application to Poisson data reconstruction*, Inverse Problems, 29 (2013), pp. 1–32.

[9] A. Benfenati and V. Ruggiero, *Inexact Bregman iteration for deconvolution of superimposed extended and point sources*, Commun. Nonlinear Sci. Numer. Simul., 20 (2015), pp. 882 – 896, https://doi.org/http://dx.doi.org/10.1016/j.cnsns.2014.06.045.

[10] I. Bengtsson and K. Zyczkowski, *Geometry of Quantum States: An Introduction to Quantum Entanglement*, Cambridge University Press, Cambridge, 002 2006, https://doi.org/10.

1017/CBO9780511535048.

[11] J. Borwein and A. Lewis, *Convex Analysis and Nonlinear Optimization*, Springer, 2014.

[12] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Found. Trends Mach. Learn., 3 (2011), pp. 1–122, https://doi.org/10.1561/2200000016.

[13] L. M. Bregman, *The Relaxation Method of Finding the Common Point of Convex Sets and Its Application to the Solution of Problems in Convex Programming*, USSR Computational Mathematics and Mathematical Physics, 7 (1967), pp. 200–217.

[14] C. Brune, A. Sawatzky, and M. Burger, *Primal and dual Bregman methods with application to optical nanoscopy*, Int. J. Comput. Vis., 92 (2011), pp. 211–229, https://doi.org/10.1007/s11263-010-0339-5.

[15] M. Burger, A. Sawatzky, and G. Steidl, *First Order Algorithms in Variational Image Processing*, Springer International Publishing, Cham, 2016, pp. 345–407, https://doi.org/10.1007/978-3-319-41589-5_10.

[16] J.-F. Cai, E. J. Cands, and Z. Shen, *A singular value thresholding algorithm for matrix completion*, SIAM J. Optim., 20 (2010), pp. 1956–1982, https://doi.org/10.1137/080738970.

[17] T. Cai, W. Liu, and X. Luo, *A constrained ℓ_1 minimization approach to sparse precision matrix estimation*, J. Am. Stat. Assoc., 106 (2011), pp. 594–607, https://doi.org/10.1198/jasa.2011.tm10155.

[18] V. Chandrasekaran, P. A. Parrilo, and A. S. Willsky, *Latent variable graphical model selection via convex optimization*, Ann. Statist., 40 (2012), pp. 1935–1967, https://doi.org/10.1214/11-AOS949.

[19] R. Chartrand, *Nonconvex splitting for regularized low-rank + sparse decomposition*, IEEE Trans. Signal Process., 60 (2012), pp. 5810–5819.

[20] C. Chaux, P. L. Combettes, J.-C. Pesquet, and V. R. Wajs, *A variational formulation for frame-based inverse problems*, Inverse Problems, 23 (2007), p. 1495.

[21] C. Chaux, J.-C. Pesquet, and N. Pustelnik, *Nested iterative algorithms for convex constrained image recovery problem*, SIAM J. Imaging Sci., 2 (2009), pp. 730–762.

[22] E. Chouzenoux and J.-C. Pesquet, *Convergence Rate Analysis of the Majorize-Minimize Subspace Algorithm*, IEEE Signal Process. Lett., 23 (2016), pp. 1284 – 1288, https://doi.org/10.1109/LSP.2016.2593589.

[23] P. L. Combettes and J.-C. Pesquet, *A Douglas-Rachford splitting approach to nonsmooth convex variational signal recovery*, IEEE J. Sel. Topics Signal Process., 1 (2007), pp. 564–574.

[24] P. L. Combettes and J.-C. Pesquet, *Proximal Splitting Methods in Signal Processing*, in Fixed-Point Algorithms for Inverse Problems in Science and Engineering, Springer, 2011, pp. 185–212, https://doi.org/10.1007/978-1-4419-9569-8.

[25] L. Condat, *Fast projection onto the simplex and the ℓ_1 ball*, Math. Programm., 158 (2016), pp. 575–585, https://doi.org/10.1007/s10107-015-0946-6.

[26] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth, *On the Lambert W function*, Adv. Comput. Math., 5 (1996), pp. 329–359, https://doi.org/10.1007/BF02124750.

[27] T. Cover and J. Thomas, *Elements of Information Theory*, A Wiley-Interscience publication, Wiley, 2006.

[28] A. d'Aspremont, O. Banerjee, and L. E. Ghaoui, *First-order methods for sparse covariance selection*, SIAM J. Matrix Anal. Appl., 30 (2008), pp. 56–66, https://doi.org/10.1137/060670985.

[29] A. Dempster, *Covariance selection*, Biometrics, 28 (1972), pp. 157–175.

[30] J. C. Duchi, S. Gould, and D. Koller, *Projected Subgradient Methods for Learning Sparse Gaussians*, in UAI 2008, Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence, Helsinki, Finland, July 9-12, 2008, 2008, pp. 145–152.

[31] J. Friedman, T. Hastie, and R. Tibshirani, *Sparse inverse covariance estimation with the graphical lasso*, Biostatistics, 9 (2008), pp. 432–441, https://doi.org/10.1093/biostatistics/kxm045.

[32] T. Goldstein and S. Osher, *The split Bregman method for l1-regularized problems*, SIAM J. Imaging Sci., 2 (2009), pp. 323–343, https://doi.org/10.1137/080725891.

[33] J. Guo, E. Levina, G. Michailidis, and J. Zhu, *Joint estimation of multiple graphical models*, Biometrika, 98 (2011), p. 1, https://doi.org/10.1093/biomet/asq060.

[34] G. Hardy, J. Littlewood, and G. Pólya, *Inequalities*, Cambridge Mathematical Library, Cambridge University Press, 1952.

[35] D. R. Hunter and K. Lange, *A tutorial on MM algorithms*, Amer. Statist., 58 (2004), pp. 30–37, https://doi.org/10.1198/0003130042836.

[36] M. W. Jacobson and J. A. Fessler, *An expanded theoretical treatment of iteration-dependent majorize-minimize algorithms*, IEEE Trans. Image Process., 16 (2007), pp. 2411–2422, https://doi.org/10.1109/TIP.2007.904387.

[37] N. Komodakis and J. C. Pesquet, *Playing with duality: An overview of recent primal–dual approaches for solving large-scale optimization problems*, IEEE Signal Process. Mag., 32 (2015), pp. 31–54, https://doi.org/10.1109/MSP.2014.2377273.

[38] A. S. Lewis, *Convex analysis on the Hermitian matrices*, SIAM J. Optim., 6 (1996), pp. 164–177, https://doi.org/10.1137/0806009.

[39] G. Li and T. K. Pong, *Douglas–Rachford splitting for nonconvex optimization with application to nonconvex feasibility problems*, Math. Programm., 159 (2016), pp. 371–401, https://doi.org/10.1007/s10107-015-0963-5.

[40] L. Li and K.-C. Toh, *An inexact interior point method for ℓ_1–regularized sparse covariance selection*, Math. Program. Comput., 2 (2010), pp. 291–315, https://doi.org/10.1007/s12532-010-0020-6.

[41] P. L. Lions and B. Mercier, *Splitting algorithms for the sum of two nonlinear operators*, SIAM J. Numer. Anal., 16 (1979), pp. 964–979, https://doi.org/10.1137/0716071.

[42] Z. Lu, *Smooth optimization approach for sparse covariance selection*, SIAM J. Optim., 19 (2009), pp. 1807–1827, https://doi.org/10.1137/070695915.

[43] Z. Lu, *Adaptive first-order methods for general sparse inverse covariance selection*, SIAM J. Matrix Anal. Appl., 31 (2010), pp. 2000–2016, https://doi.org/10.1137/080742531.

[44] S. Ma, L. Xue, and H. Zou, *Alternating direction methods for latent variable Gaussian graphical model selection*, Neural Comput., 25 (2013), pp. 2172–2198, https://doi.org/10.1162/NECO_a_00379.

[45] J. R. Magnus and H. Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, John Wiley, second ed., 1999.

[46] A. W. Marshall, I. Olkin, and B. C. Arnold, *Inequalities: Theory of Majorization and its Applications*, vol. 143, Springer, second ed., 2011, https://doi.org/10.1007/978-0-387-68276-1.

[47] R. Mazumder and T. Hastie, *The graphical lasso: New insights and alternatives*, Electron. J. Stat., 6 (2012), pp. 2125–2149, https://doi.org/10.1214/12-EJS740.

[48] N. Meinshausen and P. Bhlmann, *High-dimensional graphs and variable selection with the lasso*, Ann. Statist., 34 (2006), pp. 1436–1462, https://doi.org/10.1214/009053606000000281.

[49] J. Moreau, *Proximit et dualit dans un espace hilbertien*, Bull. Soc. Math. France, 93 (1965), pp. 273–299.

[50] Y. Nesterov, *Smooth minimization of non-smooth functions*, Math. Programm., 103 (2005), pp. 127–152, https://doi.org/10.1007/s10107-004-0552-5.

[51] N. Parikh and S. Boyd, *Proximal algorithms*, Found. Trends Optim., 1 (2014), pp. 127–239, https://doi.org/10.1561/2400000003.

[52] J.-C. Pesquet and N. Pustelnik, *A parallel inertial proximal optimization method*, Pac. J. Optim., 8 (2012), pp. 273–305.

[53] P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu, *High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence*, Electron. J. Statist., 5 (2011), pp. 935–980, https://doi.org/10.1214/11-EJS631.

[54] E. Richard, P. andre Savalle, and N. Vayatis, *Estimation of simultaneously sparse and low rank matrices*, in Proceedings of the 29th International Conference on Machine Learning (ICML-12), ACM, 2012, pp. 1351–1358.

[55] R. Rockafellar, *Convex Analysis*, Princeton landmarks in mathematics and physics, Princeton University Press, 1970.

[56] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer-Verlag, 1st ed., 1997.

[57] A. J. Rothman, P. J. Bickel, E. Levina, and J. Zhu, *Sparse permutation invariant covariance estimation*, Electron. J. Statist., 2 (2008), pp. 494–515, https://doi.org/10.1214/08-EJS176.

[58] K. Scheinberg, S. Ma, and D. Goldfarb, *Sparse inverse covariance selection via alternating linearization methods*, in Advances in Neural Information Processing Systems 23, 2010, pp. 2101–2109.

[59] Y. Sun, P. Babu, and D. P. Palomar, *Majorization-Minimization algorithms in signal processing, communications, and machine learning*, IEEE Trans. Signal Process., 65 (2017), pp. 794–816, https://doi.org/10.1109/TSP.2016.2601299.

[60] M. E. Tipping, *Sparse Bayesian learning and the relevance vector machine*, J. Mach. Learn. Res., 1 (2001), pp. 211–244, https://doi.org/10.1162/15324430152748236.

[61] E. van den Berg and M. P. Friedlander, *Probing the Pareto frontier for basis pursuit solu-*

tions, SIAM J. Sci. Comput., 31 (2009), pp. 890–912, https://doi.org/10.1137/080714488.

[62] C. Wang, D. Sun, and K.-C. Toh, *Solving log-determinant optimization problems by a Newton-CG primal proximal point algorithm*, SIAM J. Optim., 20 (2010), pp. 2994–3013, https://doi.org/10.1137/090772514.

[63] D. P. Wipf and B. D. Rao, *Sparse Bayesian learning for basis selection*, IEEE Trans. Signal Process., 52 (2004), pp. 2153–2164, https://doi.org/10.1109/TSP.2004.831016.

[64] C. F. J. Wu, *On the convergence properties of the EM algorithm*, Ann. Statist., 11 (1983), pp. 95–103, https://doi.org/10.1214/aos/1176346060.

[65] S. Yang, Z. Lu, X. Shen, P. Wonka, and J. Ye, *Fused multiple graphical lasso*, SIAM J. Optim., 25 (2015), pp. 916–943, https://doi.org/10.1137/130936397.

[66] W. Yin, S. Osher, D. Goldfarb, and J. Darbon, *Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing*, SIAM J. Imaging Sci., 1 (2008), pp. 143–168, https://doi.org/10.1137/070703983.

[67] M. Yuan and Y. Lin, *Model selection and estimation in the Gaussian graphical model*, Biometrika, 94 (2007), p. 19, https://doi.org/10.1093/biomet/asm018.

[68] X. Yuan, *Alternating direction methods for sparse covariance selection*, (2009), http://www.optimization-online.org/DBFILE/2009/09/2390.pdf.

[69] W. I. Zangwill, *Nonlinear programming : a unified approach*, Englewood Cliffs, N.J. : Prentice-Hall, 1969.

[70] X. Zhang, M. Burger, X. Bresson, and S. Osher, *Bregmanized nonlocal regularization for deconvolution and sparse reconstruction*, SIAM J. Imaging Sci., 3 (2010), pp. 253–276, https://doi.org/10.1137/090746379.

[71] X. Zhang, M. Burger, and S. Osher, *A unified primal-dual algorithm framework based on Bregman iteration*, J. Sci. Comput., 46 (2011), pp. 20–46, https://doi.org/10.1007/s10915-010-9408-8.

[72] S. Zhou, N. Xiu, Z. Luo, and L. Kong, *Sparse and low-rank covariance matrices estimation*, (2014), https://arxiv.org/abs/1407.4596.