Smooth Optimization Approach for Covariance Selection *

Zhaosong Lu†

June 30, 2007

Abstract

In this paper we study a smooth optimization approach for solving a class of non-smooth *strongly* concave maximization problems. In particular, we apply Nesterov's smooth optimization technique [17, 18] to their dual counterparts that are smooth convex problems. It is shown that the resulting approach has $\mathcal{O}(1/\sqrt{\epsilon})$ iteration complexity for finding an ϵ-optimal solution to both primal and dual problems. We then discuss the application of this approach to covariance selection that is approximately solved as a *L1-norm* penalized maximum likelihood estimation problem, and also propose a variant of this approach which has substantially outperformed the latter one in our computational experiments. We finally compare the performance of these approaches with two other first-order methods studied in [9], that is, Nesterov's $\mathcal{O}(1/\epsilon)$ smooth approximation scheme, and block-coordinate descent method for covariance selection on a set of randomly generated instances. It shows that our smooth optimization approach substantially outperforms their first method, and moreover, its variant tremendously outperforms their both methods.

Key words: Covariance selection, non-smooth strongly concave maximization, smooth minimization

1 Introduction

In [17, 18], Nesterov proposed an efficient smooth optimization method for solving convex programming problems of the form

$$\min\{f(u) : u \in U\}, \tag{1}$$

where f is a convex function with Lipschitz continuous gradient, and U is a closed convex set. It is shown that his method has $\mathcal{O}(1/\sqrt{\epsilon})$ iteration complexity bound, where $\epsilon > 0$ is

*This author gratefully acknowledges the support from SFU President's Research Grants, and NSERC Discovery Grants.

†Department of Mathematics, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada. (email: zhaosong@sfu.ca).

the absolute precision of the final objective function value. A proximal-point-type algorithm for (1) having the same complexity above has also been proposed more recently by Auslender and Teboulle [2].

In this paper, we are, motivated by [9], particularly interested in studying the use of smooth optimization approach for solving a class of non-smooth *strongly* concave maximization problems of the form (2). The key idea is to apply Nesterov's smooth optimization technique [17, 18] to their dual counterparts that are smooth convex problems. It is shown that the resulting approach has $\mathcal{O}(1/\sqrt{\epsilon})$ iteration complexity for finding an ϵ-optimal solution to both primal and dual problems.

One interesting application of the above approach is for covariance selection. Given a set of random variables with Gaussian distribution for which the true covariance matrix is unknown, covariance selection is a procedure used to estimate true covariance from a sample covariance matrix by maximizing its likelihood while imposing a certain sparsity on the inverse of the covariance estimation (e.g., see [11]). Therefore, it can be applied to determine a robust estimate of the true variance matrix, and simultaneously discover the sparse structure in the underlying model. Despite its popularity in numerous real-world applications (e.g., see [3, 9, 22] and the references therein), covariance selection itself is a challenging NP-hard combinatorial optimization problem. By an argument that is often used in regression techniques such as LASSO [20], Yuan and Lin [22], and d'Aspremont et al. [9] (see also [3]) showed that it can be approximately solved as a *L1-norm* penalized maximum likelihood estimation problem. Moreover, the authors of [9] studied two efficient first-order methods for solving this problem, that is, Nesterov's smooth approximation scheme, and block-coordinate descent method. It was shown in [9] that their first method has $\mathcal{O}(1/\epsilon)$ iteration complexity for finding an ϵ-optimal solution, but for their second method, the theoretical iteration complexity is unknown. In contrast with their methods, the smooth optimization approach proposed in this paper has a more attractive iteration complexity that is $\mathcal{O}(1/\sqrt{\epsilon})$ for finding an ϵ-optimal solution. In addition, we propose a variant of the smooth optimization approach which has substantially outperformed the latter one in our computational experiments. We also compare the performance of these approaches with their methods for covariance selection on a set of randomly generated instances. It shows that our smooth optimization approach substantially outperforms their first method, and moreover, its variant tremendously outperforms their both methods.

The paper is organized as follows. In Section 2, we introduce a class of non-smooth concave maximization problems in which we are interested, and propose a smooth optimization approach to solving them. In Section 3, we briefly introduce covariance selection, and show that it can be approximately solved as a *L1-norm* penalized maximum likelihood estimation problem. We also discuss the application of the smooth optimization approach for solving this problem, and propose a variant of this approach. In Section 4, we compare the performance of the smooth optimization approach and its variant with two other first-order methods studied in [9] for covariance selection on a set of randomly generated instances. Finally, we present some concluding remarks in Section 5.

1.1 Notation

In this paper, all vector spaces are assumed to be finite dimensional. The space of symmetric $n \times n$ matrices will be denoted by $\mathcal{S}^n$. If $X \in \mathcal{S}^n$ is positive semidefinite, we write $X \succeq 0$. Also, we write $X \preceq Y$ to mean $Y - X \succeq 0$. The cone of positive semidefinite (resp., definite) matrices is denoted by $\mathcal{S}^n_+$ (resp., $\mathcal{S}^n_{++}$). Given matrices X and Y in $\Re^{p\times q}$, the standard inner product is defined by $\langle X, Y\rangle := \mathrm{Tr}(XY^T)$, where $\mathrm{Tr}(\cdot)$ denotes the trace of a matrix. $\|\cdot\|$ denotes the Euclidean norm and its associated operator norm unless it is explicitly stated otherwise. The Frobenius norm of a real matrix X is defined as $\|X\|_F := \sqrt{\mathrm{Tr}(XX^T)}$. We denote by e the vector of all ones, and by I the identity matrix. Their dimensions should be clear from the context. For a real matrix X, we denote by $\mathrm{Card}(X)$ the cardinality of X, that is, the number of nonzero entries of X, and denote by $|X|$ the absolute value of X, that is, $|X|_{ij} = |X_{ij}|$ for all i, j. The determinant and the minimal (resp., maximal) eigenvalue of a real symmetric matrix X are denoted by $\det X$ and $\lambda_{\min}(X)$ (resp., $\lambda_{\max}(X)$), respectively. For a n-dimensional vector w, $\mathrm{diag}(w)$ denote the diagonal matrix whose i-th diagonal element is w_i for $i = 1, \dots, n$. We denote by $\mathcal{Z}_+$ the set of all nonnegative integers.

Let the space $\mathcal{F}$ be endowed with an arbitrary norm $\|\cdot\|$. The dual space of $\mathcal{F}$, denoted by $\mathcal{F}^*$, is the normed real vector space consisting of all linear functionals of $s : \mathcal{F} \to \Re$, endowed with the dual norm $\|\cdot\|^*$ defined as

$$\|s\|^* := \max_u \{\langle s, u\rangle : \ \|u\| \le 1\}, \quad \forall s \in \mathcal{F}^*,$$

where $\langle s, u\rangle := s(u)$ is the value of the linear functional s at u. Finally, given an operator $\mathcal{A} : \mathcal{F} \to \mathcal{F}^*$, we define

$$\mathcal{A}[H, H] := \langle \mathcal{A}H, H\rangle$$

for any $H \in \mathcal{F}$.

2 Smooth optimization approach

In this section, we consider a class of concave non-smooth maximization problems:

$$\max_{x\in X} g(x) := \min_{u\in U} \phi(x, u), \tag{2}$$

where X and U are nonempty convex compact sets in finite-dimensional real vector spaces $\mathcal{E}$ and $\mathcal{F}$, respectively, and $\phi(x, u) : X \times U \to \Re$ is a continuous function which is *strongly* concave in $x \in X$ for every fixed $u \in U$, and convex differentiable in $u \in U$ for every fixed $x \in X$. Therefore, for any $u \in U$, the function

$$f(u) := \max_{x\in X} \phi(x, u) \tag{3}$$

is well-defined. We also easily conclude that $f(u)$ is convex differentiable on U, and its gradient is given by

$$\nabla f(u) = \nabla_u \phi(x(u), u), \quad \forall u \in U, \tag{4}$$

where $x(u)$ denotes the unique solution of (3).

Let the space $\mathcal{F}$ be endowed with an arbitrary norm $\|\cdot\|$. We further assume that $\nabla f(u)$ is Lipschitz continuous on U with respect to $\|\cdot\|$, i.e., there exists some $L > 0$ such that

$$\|\nabla f(u) - \nabla f(\tilde{u})\|^* \leq L\|u - \tilde{u}\|, \quad \forall u, \tilde{u} \in U.$$

Under the above assumptions, we easily observe that: i) problem (2) and its dual, that is,

$$\min_u \{f(u) : \ u \in U\}, \tag{5}$$

are both solvable and have the same optimal value; and ii) the dual problem (5) can be suitably solved by Nesterov's smooth minimization approach [17, 18].

Denote by $d(u)$ a prox-function of the set U. We assume that $d(u)$ is continuous and strongly convex on U with convexity parameter $\sigma > 0$. Let u_0 be the center of the set U defined as

$$u_0 = \arg\min\{d(u) : \ u \in U\}. \tag{6}$$

Without loss of generality assume that $d(u_0) = 0$. We now describe Nesterov's smooth minimization approach [17, 18] for solving the dual problem (5), and we will show that it simultaneously solves the non-smooth concave maximization problem (2).

Smooth Minimization Algorithm:

Let $u_0 \in U$ be given in (6). Set $x_{-1} = 0$ and $k = 0$.

1) Compute $\nabla f(u_k)$ and $x(u_k)$. Set $x_k = \frac{k}{k+2} x_{k-1} + \frac{2}{k+2} x(u_k)$.

2) Find $u_k^{sd} \in \text{Argmin}\left\{\langle \nabla f(u_k), u - u_k \rangle + \frac{L}{2}\|u - u_k\|^2 : \ u \in U\right\}$.

3) Find $u_k^{ag} = \text{argmin}\left\{\frac{L}{\sigma} d(u) + \sum\limits_{i=0}^{k} \frac{i+1}{2}[f(u_i) + \langle \nabla f(u_i), u - u_i \rangle] : \ u \in U\right\}$.

4) Set $u_{k+1} = \frac{2}{k+3} u_k^{ag} + \frac{k+1}{k+3} u_k^{sd}$.

5) Set $k \leftarrow k + 1$ and go to step 1).

end

The following property of the above algorithm is established in Theorem 2 of Nesterov [18].

Theorem 2.1 *Let the sequence $\{(u_k, u_k^{sd})\}_{k=0}^{\infty} \subseteq U \times U$ be generated by the Smooth Minimization Algorithm. Then for any $k \geq 0$ we have*

$$\frac{(k+1)(k+2)}{4} f(u_k^{sd}) \leq \min_u \left\{\frac{L}{\sigma} d(u) + \sum_{i=0}^{k} \frac{i+1}{2}[f(u_i) + \langle \nabla f(u_i), u - u_i \rangle] : \ u \in U\right\}. \tag{7}$$

We are ready to establish the main convergence result of the Smooth Minimization Algorithm for solving the non-smooth concave maximization problem (2) and its dual (5).

Theorem 2.2 *After k iterations, the Smooth Minimization Algorithm generates a pair of approximate solutions (u_k^{sd}, x_k) to the problem (2) and its dual (5), respectively, which satisfy the following inequality:*

$$0 \leq f(u_k^{sd}) - g(x_k) \leq \frac{4LD}{\sigma(k+1)(k+2)}. \tag{8}$$

Thus if the termination criterion $f(u_k^{sd}) - g(x_k) \leq \epsilon$ is applied, the iteration complexity of finding an ϵ-optimal solution to the problem (2) and its dual (5) by the Smooth Minimization Algorithm does not exceed $2\sqrt{\frac{LD}{\sigma\epsilon}}$, where

$$D = \max\{d(u) : \ u \in U\}. \tag{9}$$

Proof. In view of (3), (4) and the notation $x(u)$, we have

$$f(u_i) + \langle \nabla f(u_i), u - u_i \rangle = \phi(x(u_i), u_i) + \langle \nabla_u \phi(x(u_i), u_i), u - u_i \rangle. \tag{10}$$

Invoking the fact that the function $\phi(x, \cdot)$ is convex on U for every fixed $x \in X$, we obtain

$$\phi(x(u_i), u_i) + \langle \nabla_u \phi(x(u_i), u_i), u - u_i \rangle \leq \phi(x(u_i), u). \tag{11}$$

Notice that $x_{-1} = 0$, and $x_k = \frac{k}{k+2} x_{k-1} + \frac{2}{k+2} x(u_k)$ for any $k \geq 0$, which imply

$$x_k = \sum_{i=0}^{k} \frac{2(i+1)}{(k+1)(k+2)} x(u_i). \tag{12}$$

Using (10), (11), (12) and the fact that the function $\phi(\cdot, u)$ is concave on X for every fixed $u \in U$, we have

$$\begin{aligned} \sum_{i=0}^{k} (i+1)[f(u_i) + \langle \nabla f(u_i), u - u_i \rangle] &\leq \sum_{i=0}^{k} (i+1) \phi(x(u_i), u) \\ &\leq \frac{1}{2}(k+1)(k+2)\phi(x_k, u) \end{aligned}$$

for all $u \in U$. It follows from this relation, (7), (9) and (2) that

$$\begin{aligned} f(u_k^{sd}) &\leq \frac{4LD}{\sigma(k+1)(k+2)} + \min_u \left\{ \sum_{i=0}^{k} \frac{2(i+1)}{(k+1)(k+2)} [f(u_i) + \langle \nabla f(u_i), u - u_i \rangle] : \ u \in U \right\} \\ &\leq \frac{4LD}{\sigma(k+1)(k+2)} + \min_{u \in U} \phi(x_k, u) = \frac{4LD}{\sigma(k+1)(k+2)} + g(x_k), \end{aligned}$$

and hence the inequality (8) holds. The remaining conclusion directly follows from (8). ∎

The following results will be used to develop a variant of the Smooth Minimization Algorithm for covariance selection in Subsection 3.4.

Lemma 2.3 *The problem (2) has a unique optimal solution, denoted by x^*. Moreover, for any $u^* \in \mathrm{Argmin}\{f(u) : \; u \in U\}$, we have*

$$x^* = \arg\max_{x \in X} \phi(x, u^*). \tag{13}$$

Proof. We clearly know that the problem (2) has an optimal solution. To prove its uniqueness, it suffices to show that $g(x)$ is strictly concave on X. Indeed, since $X \times U$ is a convex compact set and $\phi(x,u)$ is continuous on $X \times U$, it follows that for any $t \in (0,1)$, $x^1 \neq x^2 \in X$, there exists some $\tilde{u} \in U$ such that

$$\phi(tx^1 + (1-t)x^2, \tilde{u}) = \min_{u \in U} \phi(tx^1 + (1-t)x^2, u).$$

Recall that $\phi(\cdot, u)$ is strongly concave on X for every fixed $u \in U$. Therefore, we have

$$\begin{array}{rcl} \phi(tx^1 + (1-t)x^2, \tilde{u}) & > & t\phi(x^1, \tilde{u}) + (1-t)\phi(x^2, \tilde{u}), \\ & \geq & t \min\limits_{u \in U} \phi(x^1, u) + (1-t) \min\limits_{u \in U} \phi(x^2, u), \end{array}$$

which together with (2) implies that

$$g(tx^1 + (1-t)x^2) > tg(x^1) + (1-t)g(x^2)$$

for any $t \in (0,1)$, $x^1 \neq x^2 \in X$, and hence, $g(x)$ is strictly concave on X as desired.

Note that x^* is the optimal solution of problem (2). We clearly know that for any $u^* \in \mathrm{Argmin}\{f(u) : \; u \in U\}$, (u^*, x^*) is a saddle point for problem (2), that is,

$$\phi(x^*, u) \geq \phi(x^*, u^*) \geq \phi(x, u^*), \quad \forall (x,u) \in X \times U,$$

and hence, we have

$$x^* \in \mathrm{Arg}\max_{x \in X} \phi(x, u^*).$$

It together with the fact that $\phi(\cdot, u^*)$ is strongly concave on X, immediately yields (13).

■

Theorem 2.4 *Let x^* be the unique optimal solution of (2), and f^* be the optimal value of problems (2) and (5). Assume that the sequences $\{u_k\}_{k=0}^{\infty}$ and $\{x(u_k)\}_{k=0}^{\infty}$ are generated by the Smooth Minimization Algorithm. Then the following statements hold:*

1) $f(u_k) \to f^*$, $x(u_k) \to x^*$ as $k \to \infty$;

2) $f(u_k) - g(x(u_k)) \to 0$ as $k \to \infty$.

Proof. Recall from the Smooth Minimization Algorithm that

$$u_{k+1} = \left(2u_k^{ag} + (k+1)u_k^{sd}\right)/(k+3), \quad \forall k \geq 0.$$

Since $u_k^{sd}, u_k^{ag} \in U$ for $\forall k \geq 0$, and U is a compact set, we have $u_{k+1} - u_k^{sd} \to 0$ as $k \to \infty$. Notice that $f(u)$ is actually uniformly continuous on U. Thus, we further have $f(u_{k+1}) - f(u_k^{sd}) \to 0$ as $k \to \infty$. Also, it follows from Theorem 2.2 that $f(u_k^{sd}) \to f^*$ as $k \to \infty$. Therefore, we conclude that $f(u_k) \to f^*$ as $k \to \infty$.

Note that X is a compact set, and $x(u_k) \subseteq X$ for $\forall k \geq 0$. To prove that $x(u_k) \to x^*$ as $k \to \infty$, it suffices to show that every convergent subsequence of $\{x(u_k)\}_{k=0}^{\infty}$ converges to x^* as $k \to \infty$. Indeed, assume that $\{x(u_{n_k})\}_{k=0}^{\infty}$ is an arbitrary convergent subsequence, and $x(u_{n_k}) \to \tilde{x}^*$ as $k \to \infty$ for some $\tilde{x}^* \in X$. Without loss generality, assume that the sequence $\{u_{n_k}\}_{k=0}^{\infty} \to \tilde{u}^*$ as $k \to \infty$ for some $\tilde{u}^* \in U$ (otherwise, one can consider any convergent subsequence of $\{u_{n_k}\}_{k=0}^{\infty}$). Using the result that $f(u_k) \to f^*$, we obtain that

$$\phi\left(x(u_{n_k}), u_{n_k}\right) = f(u_{n_k}) \to f^*, \quad \text{as } k \to \infty.$$

Upon letting $k \to \infty$ and using the continuity of $\phi(\cdot,\cdot)$, we have $\phi(\tilde{x}^*, \tilde{u}^*) = f(\tilde{u}^*) = f^*$. Hence, it follows that

$$\tilde{u}^* \in \operatorname{Arg}\min_{u \in U} f(u), \quad \tilde{x}^* = \arg\max_{x \in X} \phi(x, \tilde{u}^*),$$

which together with Lemma 2.3 implies that $\tilde{x}^* = x^*$. Hence as desired, $x(u_{n_k}) \to x^*$ as $k \to \infty$.

As shown in Lemma 2.3, the function $g(x)$ is continuous on X. This result together with statement 1) immediately implies that statement 2) holds. ∎

3 Covariance selection

In this section, we discuss the application of the smooth optimization approach proposed in Section 2 to covariance selection. More specifically, we briefly introduce covariance selection in Subsection 3.1, and show that it can be approximately solved as a *L1-norm* penalized maximum likelihood estimation problem in Subsection 3.2. In Subsection 3.3, We address some implementation details of the smooth optimization approach for solving this problem, and propose a variant of this approach in Subsection 3.4.

3.1 Introduction of covariance selection

In this subsection, we briefly introduce covariance selection. For more details, see d'Aspremont et al. [9] and the references therein.

Given n variables with a Gaussian distribution $\mathcal{N}(0, C)$ for which the true covariance matrix C is unknown, we are interested in estimating C from a sample covariance matrix Σ by maximizing its likelihood while imposing a certain number of components in the inverse of the estimation of C to zero. This problem is commonly known as *covariance selection* (see [11]). Since zeros in the inverse of covariance matrix correspond to conditional independence in the model, covariance selection can be used to determine a robust estimate of the covariance matrix, and simultaneously discover the sparse structure in the underlying graphical model.

Several approaches have been proposed for covariance selection in literature. For example, Bilmes [4] proposed a method based on choosing statistical dependencies according to conditional mutual information computed from training data. The recent works [16, 13] involve identifying the Gaussian graphical models that are best supported by the data and any available prior information on the covariance matrix. Given a sample covariance matrix $\Sigma \in \mathcal{S}^n_+$, d'Aspremont et al. [9] recently formulated covariance selection as the following estimation problem:

$$\begin{array}{ll} \max\limits_{X} & \log\det X - \langle \Sigma, X\rangle - \rho \mathrm{Card}(X) \\ \text{s.t.} & \tilde{\alpha} I \preceq X \preceq \tilde{\beta} I, \end{array} \tag{14}$$

where $\rho > 0$ is a parameter controlling the trade-off between likelihood and cardinality, and $0 \le \tilde{\alpha} < \tilde{\beta} \le \infty$ are the fixed bounds on the eigenvalues of the solution. For some specific choices of ρ, the formulation (14) has been used for model selection in [1, 5], and applied to speech recognition and gene network analysis (see [4, 12]).

Note that the estimation problem (14) itself is a NP-hard combinatorial problem because of the penalty term $\mathrm{Card}(X)$. To overcome the computational difficulty, d'Aspremont et al. [9] used an argument that is often used in regression techniques (e.g., see [20, 6, 14]), where sparsity of the solution is concerned, to relax $\mathrm{Card}(X)$ to $e^T|X|e$, and obtained the following *L1-norm* penalized maximum likelihood estimation problem:

$$\begin{array}{ll} \max\limits_{X} & \log\det X - \langle \Sigma, X\rangle - \rho e^T|X|e \\ \text{s.t.} & \tilde{\alpha} I \preceq X \preceq \tilde{\beta} I, \end{array} \tag{15}$$

Recently, Yuan and Lin [22] proposed a similar estimation problem for covariance selection given as follows:

$$\begin{array}{ll} \max\limits_{X} & \log\det X - \langle \Sigma, X\rangle - \rho \sum\limits_{i\neq j} |X_{ij}| \\ \text{s.t.} & \tilde{\alpha} I \preceq X \preceq \tilde{\beta} I, \end{array} \tag{16}$$

with $\tilde{\alpha} = 0$ and $\tilde{\beta} = \infty$. They showed that problem (16) can be suitably solved by the interior point algorithm developed in Vandenberghe et al. [21]. A few other approaches have also been studied for covariance selection by solving some related maximum likelihood estimation problems in literature. For example, Huang et al. [15] proposed an iterative (heuristic) algorithm to minimize a nonconvex penalized likelihood. Dahl et al. [8, 7] applied Newton method, coordinate steepest descent method, and conjugate gradient method for the problems for which the conditional independence structure is partially known.

As shown in d'Aspremont et al. [9] (see also [3]), and Yuan and Lin [22], the *L1-norm* penalized maximum likelihood estimation problems (15) and (16) are capable of discovering effectively the sparse structure, or equivalently, the conditional independence in the underlying graphical model. Also, it is not hard to see that the estimation problem (16) becomes a special case of problem (15) if replacing Σ by $\Sigma + \rho I$ in (16). For these reasons, we focus on problem (15) only for the remaining paper.

3.2 Non-smooth strongly concave maximization reformulation

In this subsection, we show that problem (15) can be reformulated as a non-smooth strongly concave maximization problem of the form (2).

We first provide some tighter bounds on the optimal solution of problem (15) for the case where $\tilde{\alpha} = 0$ and $\tilde{\beta} = \infty$.

Proposition 3.1 *Assume that $\tilde{\alpha} = 0$ and $\tilde{\beta} = \infty$. Let $X^* \in \mathcal{S}^n_{++}$ be the unique optimal solution of problem (15). Then we have $\alpha I \preceq X^* \preceq \beta I$, where*

$$\alpha = \frac{1}{\|\Sigma\| + n\rho}, \quad \beta = \min\left\{\frac{n - \alpha \mathrm{Tr}(\Sigma)}{\rho}, \eta\right\} \tag{17}$$

with

$$\eta = \begin{cases} \min\left\{e^T|\Sigma^{-1}|e, (n - \rho\sqrt{n}\alpha)\|\Sigma^{-1}\| - (n-1)\alpha\right\}, & \text{if } \Sigma \text{ is invertible;} \\ 2e^T|(\Sigma + \frac{\rho}{2}I)^{-1}|e - \mathrm{Tr}((\Sigma + \frac{\rho}{2}I)^{-1}), & \text{otherwise.} \end{cases}$$

Proof. Let

$$\mathcal{U} := \{U \in \mathcal{S}^n : \ |U_{ij}| \le 1, \forall ij\}, \tag{18}$$

and

$$L(X, U) = \log\det X - \langle \Sigma + \rho U, X\rangle, \quad \forall (X, U) \in \mathcal{S}^n_{++} \times \mathcal{U}. \tag{19}$$

Note that $X^* \in \mathcal{S}^n_{++}$ is the optimal solution of problem (15). It can be easily shown that there exists some $U^* \in \mathcal{U}$ such that (X^*, U^*) is a saddle point of $L(\cdot,\cdot)$ on $\mathcal{S}^n_{++} \times \mathcal{U}$, that is,

$$X^* \in \arg\min_{X \in \mathcal{S}^n_{++}} L(X, U^*), \qquad U^* \in \operatorname{Arg}\min_{U \in \mathcal{U}} L(X^*, U).$$

The above relations along with (18) and (19) immediately yield

$$X^*(\Sigma + \rho U^*) = I, \qquad \langle X^*, U^*\rangle = e^T|X^*|e. \tag{20}$$

Hence, we have

$$X^* = (\Sigma + \rho U^*)^{-1} \succeq \frac{1}{\|\Sigma\| + \rho\|U^*\|} I,$$

which together with (18) and the fact $U^* \in \mathcal{U}$, implies that $X^* \succeq \frac{1}{\|\Sigma\| + n\rho} I$. Thus as desired, $X^* \succeq \alpha I$ as desired, where α is given in (17).

We next bound X^* from above. In view of (20), we have

$$\langle X^*, \Sigma\rangle + \rho e^T|X^*|e = n, \tag{21}$$

which together with the relation $X^* \succeq \alpha I$ implies that

$$e^T|X^*|e \leq \frac{n - \alpha \mathrm{Tr}(\Sigma)}{\rho}. \tag{22}$$

Now let $X(t) := (\Sigma + t\rho I)^{-1}$ for $t \in (0,1)$. By definition of X^* and $X(t)$, we clearly have

$$\begin{aligned} \log\det X^* - \langle \Sigma + t\rho I, X^* \rangle &\leq \log\det X(t) - \langle \Sigma + t\rho I, X(t) \rangle, \\ \log\det X(t) - \langle \Sigma, X(t) \rangle - \rho e^T|X(t)|e &\leq \log\det X^* - \langle \Sigma, X^* \rangle - \rho e^T|X^*|e. \end{aligned}$$

Adding the above two inequalities upon some algebraic simplification, we obtain that

$$e^T|X^*|e - t\mathrm{Tr}(X^*) \leq e^T|X(t)|e - t\mathrm{Tr}(X(t)),$$

and hence,

$$e^T|X^*|e \leq \frac{e^T|X(t)|e - t\mathrm{Tr}(X(t))}{1-t}, \quad \forall t \in (0,1). \tag{23}$$

If Σ is invertible, upon letting $t \downarrow 0$ on both sides of (23), we have

$$e^T|X^*|e \leq e^T|\Sigma^{-1}|e.$$

Otherwise, letting $t = 1/2$ in (23), we obtain

$$e^T|X^*|e \leq 2e^T|(\Sigma + \frac{\rho}{2}I)^{-1}|e - \mathrm{Tr}((\Sigma + \frac{\rho}{2}I)^{-1}).$$

Combining the above two inequalities and (22), we have

$$\|X^*\| \leq \|X^*\|_F \leq e^T|X^*|e \leq \min\left\{\frac{n - \alpha\mathrm{Tr}(\Sigma)}{\rho}, \gamma\right\}, \tag{24}$$

where

$$\gamma = \begin{cases} e^T|\Sigma^{-1}|e, & \text{if } \Sigma \text{ is invertible;} \\ 2e^T|(\Sigma + \frac{\rho}{2}I)^{-1}|e - \mathrm{Tr}((\Sigma + \frac{\rho}{2}I)^{-1}), & \text{otherwise.} \end{cases}$$

Further, using the relation $X^* \succeq \alpha I$, we obtain that

$$e^T|X^*|e \geq \|X^*\|_F \geq \sqrt{n}\alpha,$$

which together with (21) implies that

$$\mathrm{Tr}(X^*\Sigma) \leq n - \rho\sqrt{n}\alpha.$$

This inequality along with the relation $X^* \succeq \alpha I$ yields

$$\lambda_{\min}(\Sigma)((n-1)\alpha + \|X^*\|) \leq \mathrm{Tr}(X^*\Sigma) \leq n - \rho\sqrt{n}\alpha.$$

Hence if Σ is invertible, we further have

$$\|X^*\| \leq (n - \rho\sqrt{n}\alpha)\|\Sigma^{-1}\| - (n-1)\alpha.$$

This together with (24) implies that $X^* \preceq \beta I$, where β is given in (17). ∎

Remark. Some bounds on X^* were also derived in d'Aspremont et al. [9]. In contrast with their bounds, our bounds given in (17) are tighter. Moreover, our approach for deriving these bounds can be generalized to handle the case where $\tilde{\alpha} > 0$ and $\tilde{\beta} = \infty$ (this is left to the reader), but their approach cannot. ∎

From the above discussion, we conclude that problem (15) is equivalent to the following problem:

$$\begin{array}{ll} \max\limits_{X} & \log\det X - \langle \Sigma, X\rangle - \rho e^T|X|e \\ \text{s.t.} & \alpha I \preceq X \preceq \beta I, \end{array} \tag{25}$$

for some $0 < \alpha < \beta < \infty$.

We further observe that problem (25) can be rewritten as

$$\max_{X\in\mathcal{X}} \min_{U\in\mathcal{U}} \ \log\det X - \langle \Sigma + \rho U, X\rangle, \tag{26}$$

where $\mathcal{U}$ is defined in (18), and $\mathcal{X}$ is defined as follows:

$$\mathcal{X} := \{X \in \mathcal{S}^n : \ \alpha I \preceq X \preceq \beta I\}. \tag{27}$$

Therefore, we conclude that problem (15) is equivalent to (26). For the remaining paper, we will focus on problem (26) only.

3.3 Smooth optimization method for covariance selection

In this subsection, we describe the implementation details of the Smooth Minimization Algorithm proposed in Section 2 for solving problem (26). We also compare the complexity of this algorithm with interior point methods, and two other first-order methods studied in d'Aspremont et al. [9], that is, Nesterov's smooth approximation scheme, and block coordinate descent method.

We first observe that the sets $\mathcal{X}$ and $\mathcal{U}$ both lie in the space $\mathcal{S}^n$, where $\mathcal{X}$ and $\mathcal{U}$ are defined in (27) and (18), respectively. Let $\mathcal{S}^n$ be endowed with the Frobenius norm, and let $\tilde{d}(X) = \log\det X$ for $X \in \mathcal{X}$. Then for any $X \in \mathcal{X}$, we have

$$\nabla^2 \tilde{d}(X)[H,H] = -\mathrm{Tr}(X^{-1}HX^{-1}H) \leq -\beta^{-2}\|H\|_F^2$$

for all $H \in \mathcal{S}^n$, and hence, $\tilde{d}(X)$ is strongly concave on $\mathcal{X}$ with convexity parameter β^{-2}. Using this result and Theorem 1 of [18], we immediately conclude that $\nabla f(U)$ is Lipschitz continuous with constant $L = \rho^2\beta^2$ on $\mathcal{U}$, where

$$f(U) := \max_{X\in\mathcal{X}} \ \log\det X - \langle \Sigma + \rho U, X\rangle, \quad \forall U \in \mathcal{U}. \tag{28}$$

Denote the unique optimal solution of problem (28) by $X(U)$. For any $U \in \mathcal{U}$, we can compute $X(U)$, $f(U)$ and $\nabla f(U)$ as follows.

Let $\Sigma + \rho U = Q\text{diag}(\gamma)Q^T$ be an eigenvalue decomposition of $\Sigma + \rho U$ such that $QQ^T = I$. For $i = 1, \ldots, n$, let

$$\lambda_i = \begin{cases} \min\{\max\{1/\gamma_i, \alpha\}, \beta\}, & \text{if } \gamma_i > 0; \\ \beta, & \text{otherwise.} \end{cases}$$

It is not hard to show that

$$X(U) = Q\text{diag}(\lambda)Q^T, \quad f(U) = -\gamma^T\lambda + \sum_{i=1}^n \log \lambda_i, \quad \nabla f(U) = -\rho X(U). \tag{29}$$

From the above discussion, we see that problem (26) has exactly the same form as (2), and also satisfies all assumptions imposed on problem (2). Therefore, it can be suitably solved by the Smooth Minimization Algorithm proposed in Section 2. The implementation details of this algorithm for problem (26) are described as follows.

Given $U_0 \in \mathcal{U}$, let $d(U) = \|U - U_0\|_F^2/2$ be the proximal function on $\mathcal{U}$, which has convexity parameter $\sigma = 1$. For our specific choice of the norm and $d(U)$, we clearly see that steps 2) and 3) of the Smooth Minimization Algorithm can be solved as a problem of the form

$$V = \arg\min_{U \in \mathcal{U}} \langle G, U \rangle + \|U\|_F^2/2$$

for some $G \in \mathcal{S}^n$. In view of (18), we see that

$$V_{ij} = \max\{\min\{-G_{ij}, 1\}, -1\}, \ i, j = 1, \ldots, n.$$

In addition, for any $X \in \mathcal{X}$, we define

$$g(X) := \log\det X - \langle \Sigma, X \rangle - \rho e^T |X| e. \tag{30}$$

We are now ready to present a complete version of the Smooth Minimization Algorithm for solving problem (26).

Smooth Minimization Algorithm for Covariance Selection (SMACS):

Let $\epsilon > 0$ and $U_0 \in \mathcal{U}$ be given. Set $X_{-1} = 0$, $L = \rho^2\beta^2$, $\sigma = 1$, and $k = 0$.

1) Compute $\nabla f(U_k)$ and $X(U_k)$. Set $X_k = \frac{k}{k+2}X_{k-1} + \frac{2}{k+2}X(U_k)$.

2) Find $U_k^{sd} = \text{argmin}\left\{\langle \nabla f(U_k), U - U_k \rangle + \frac{L}{2}\|U - U_k\|_F^2 : \ U \in \mathcal{U}\right\}$.

3) Find $U_k^{ag} = \text{argmin}\left\{\frac{L}{2\sigma}\|U - U_0\|_F^2 + \sum_{i=0}^k \frac{i+1}{2}[f(U_i) + \langle \nabla f(U_i), U - U_i \rangle] : \ U \in \mathcal{U}\right\}$.

4) Set $U_{k+1} = \frac{2}{k+3}U_k^{ag} + \frac{k+1}{k+3}U_k^{sd}$.

5) Set $k \leftarrow k + 1$. Go to step 1) until $f(U_k^{sd}) - g(X_k) \le \epsilon$.

end

The iteration complexity of the above algorithm for solving problem (26) is established in the following theorem.

Theorem 3.2 *The iteration complexity performed by the algorithm SMACS for finding an* ϵ*-optimal solution to problem (26) and its dual does not exceed* $\sqrt{2}\rho\beta \max_{U\in\mathcal{U}} \|U-U_0\|_F/\sqrt{\epsilon}$*, and moreover, if* $U_0 = 0$*, it does not exceed* $\sqrt{2}\rho\beta n/\sqrt{\epsilon}$*.*

Proof. From the above discussion, we know that $L = \rho^2\beta^2$, $D = \max_{U\in\mathcal{U}} \|U - U_0\|_F^2/2$ and $\sigma = 1$, which together with Theorem 2.2 immediately implies that the first part of the statement holds. Further, if $U_0 = 0$, we easily obtain from (18) that $D = \max_{U\in\mathcal{U}} \|U\|_F^2/2 = n^2/2$. The second part of the statement directly follows from this result and Theorem 2.2. ∎

Alternatively, d'Aspremont et al. [9] applied Nesterov's smooth approximation scheme [18] to solve problem (26). More specifically, let $\epsilon > 0$ be the desired accuracy, and let

$$\hat{d}(U) = \|U\|_F^2/2, \quad \hat{D} = \max_{U\in\mathcal{U}} \hat{d}(U) = n^2/2.$$

As shown in [18], the non-smooth function $g(X)$ defined in (30) is uniformly approximated by the smooth function

$$g_\epsilon(X) = \min_{U\in\mathcal{U}} \ \log\det X - \langle \Sigma + \rho U, X\rangle - \frac{\epsilon}{2\hat{D}}\hat{d}(U)$$

on $\mathcal{X}$ with the error at most by $\epsilon/2$, and moreover, the function $g_\epsilon(X)$ has a Lipschitz continuous gradient on $\mathcal{X}$ with some constant $L(\epsilon) > 0$. Nesterov's smooth optimization technique [17, 18] is then applied to solve the perturbed problem $\max_{X\in\mathcal{X}} g_\epsilon(X)$, and the problem (26) is accordingly solved. It was shown in [9] that the iteration complexity of this approach for finding an ϵ-optimal solution to problem (26) does not exceed

$$\frac{2\sqrt{2}\rho\beta n^{1.5}\log\kappa}{\epsilon} + \kappa\sqrt{\frac{n\log\kappa}{\epsilon}} \tag{31}$$

where $\kappa := \beta/\alpha$.

In view of (31) and Theorem 3.2, we conclude that the smooth optimization approach improves upon Nesterov's smooth approximation scheme at least by a factor $\mathcal{O}(\sqrt{n}\log\kappa/\sqrt{\epsilon})$ in terms of the iteration complexity for solving problem (26). Moreover, the computational cost per iteration of the former approach is at least as cheap as that of the latter one.

d'Aspremont et al. [9] also studied block-coordinate descent method for solving the problem (15) with $\tilde{\alpha} = 0$ and $\tilde{\beta} = \infty$. Each iterate of this method requires computing the inverse of an $(n-1)\times(n-1)$ matrix, and solving a box constrained quadratic programming with $n-1$ variables. It shall be mentioned that the iteration complexity of this method for finding an ϵ-optimal solution is unknown. Moreover, this method is not suitable for solving problem (15) with $\tilde{\alpha} > 0$ or $\tilde{\beta} < \infty$.

In addition, we observe that problem (25) (also (15)) can be reformulated as a constrained smooth convex problem that has an explicit $\mathcal{O}(n^2)$-logarithmically homogeneous self-concordant barrier function. Thus, it can be suitably solved by interior point (IP) methods (see Nesterov and Nemirovski [19] and Vandenberghe et al. [21]). The worst-case iteration complexity of IP methods for finding an ϵ-optimal solution to (25) is $\mathcal{O}(n\log(\epsilon_0/\epsilon))$, where ϵ_0 is an initial gap. Each iterate of IP methods requires $\mathcal{O}(n^6)$ arithmetic cost for assembling and solving a typically dense Newton system. Thus, the total worst-case arithmetic cost of IP methods for finding an ϵ-optimal solution to (25) is $\mathcal{O}(n^7\log(\epsilon_0/\epsilon))$. In contrast with IP methods, the algorithm SMACS requires $\mathcal{O}(n^3)$ arithmetic cost per iteration dominated by eigenvalue decomposition and matrix multiplication of $n \times n$ matrices. Based on this observation and Theorem 3.2, we conclude that the overall worst-case arithmetic cost of the algorithm SMACS for finding an ϵ-optimal solution to (25) is $\mathcal{O}(\rho\beta n^4/\sqrt{\epsilon})$, which is substantially superior to that of IP methods, provided that $\rho\beta$ is not too large and ϵ is not too small.

3.4 Variant of Smooth Minimization Algorithm

As discussed in Subsection 3.3, the algorithm SMACS has a nice theoretical complexity in contrast with IP methods, Nesterov's smooth approximation scheme, and block-coordinate descent method. However, its practical performance is still not much attractive (see Section 4). To enhance the computational performance, we propose a variant of the algorithm SMACS for solving problem (26) in this subsection.

Our first concern of the algorithm SMACS is that the eigenvalue decomposition of two $n \times n$ matrices is required per iteration. Indeed, the eigenvalue decomposition of $\Sigma + \rho U_k$ and $\Sigma + \rho U_k^{sd}$ is needed at steps 1) and 5) to compute $\nabla f(U_k)$ and $f(U_k^{sd})$, respectively. We also know that the eigenvalue decomposition is one of major computations for the algorithm SMACS. To reduce the computational cost, we now propose a new termination criterion other than $f(U_k^{sd}) - g(X_k) \leq \epsilon$ that is used in the algorithm SMACS. In view of Theorem 2.4, we know that

$$f(U_k) - g(X(U_k)) \rightarrow 0, \text{ as } k \rightarrow \infty.$$

Thus, $f(U_k) - g(X(U_k)) \leq \epsilon$ can be used as an alternative termination criterion. Moreover, it follows from (29) that the quantity $f(U_k) - g(X(U_k))$ is readily available in step 1) of the algorithm SMACS with almost no additional cost. We easily see that the algorithm SMACS would require only one eigenvalue decomposition per iteration with this new termination criterion. In addition, as observed from computational experiments, the number of iterations performed by the algorithm SMACS based on these two different termination criteria are almost same.

For covariance selection, the penalty parameter ρ is usually small, but the parameter β can be fairly large. In view of Theorem 3.2, we know that the iteration complexity of the algorithm SMACS for solving problem (26) is proportional to β. Therefore, when β is too large, the complexity and practical performance of this algorithm become unattractive. To overcome this drawback, we will propose one strategy to dynamically update β.

Let X^* be the unique optimal solution of problem (26). For any $\hat{\beta} \in [\lambda_{\max}(X^*), \beta]$, we easily observe that X^* is also the unique optimal solution to the following problem:

$$(P_{\hat{\beta}}) \qquad \max_{X \in \mathcal{X}_{\hat{\beta}}} \min_{U \in \mathcal{U}} \log\det X - \langle \Sigma + \rho U, X \rangle, \tag{32}$$

where $\mathcal{U}$ is defined in (18), and $\mathcal{X}_{\hat{\beta}}$ is given by

$$\mathcal{X}_{\hat{\beta}} := \{X : \alpha I \preceq X \preceq \hat{\beta} I\}.$$

In view of Theorem 3.2, the iteration complexity of the algorithm SMACS for problem (32) is lower than that for problem (26) provided $\hat{\beta} \in [\lambda_{\max}(X^*), \beta)$. Hence ideally, we set $\hat{\beta} = \lambda_{\max}(X^*)$, which would give the lowest iteration complexity, but unfortunately, $\lambda_{\max}(X^*)$ is unknown. However, we can generate a sequence $\{\hat{\beta}_k\}_{k=0}^{\infty}$ that asymptotically approaches $\lambda_{\max}(X^*)$ as the algorithm progresses. Indeed, in view of Theorem 2.4, we know that $X(U_k) \to X^*$ as $k \to \infty$, and we obtain that

$$\lambda_{\max}(X(U_k)) \to \lambda_{\max}(X^*), \text{ as } k \to \infty.$$

Therefore, we see that $\{\lambda_{\max}(X(U_k))\}_{k=0}^{\infty}$ can be used to generate a sequence $\{\hat{\beta}_k\}_{k=0}^{\infty}$ that asymptotically approaches $\lambda_{\max}(X^*)$. We next propose a strategy to generate such a sequence $\{\hat{\beta}_k\}_{k=0}^{\infty}$.

For convenience of presentation, we introduce some new notations. Given any $U \in \mathcal{U}$ and $\hat{\beta} \in [\alpha, \beta]$, we define

$$X_{\hat{\beta}}(U) := \arg\max_{X \in \mathcal{X}_{\hat{\beta}}} \log\det X - \langle \Sigma + \rho U, X \rangle, \tag{33}$$

$$f_{\hat{\beta}}(U) := \max_{X \in \mathcal{X}_{\hat{\beta}}} \log\det X - \langle \Sigma + \rho U, X \rangle. \tag{34}$$

Definition 1 *Given any $U \in \mathcal{U}$ and $\hat{\beta} \in [\alpha, \beta]$, $X_{\hat{\beta}}(U)$ is called "active" if $\lambda_{\max}(X_{\hat{\beta}}(U)) = \hat{\beta}$ and $\hat{\beta} < \beta$; otherwise it is called "inactive".*

Let ς_1, $\varsigma_2 > 1$, and $\varsigma_3 \in (0, 1)$ be given and fixed. Assume that $U_k \in \mathcal{U}$ and $\hat{\beta}_k \in [\alpha, \beta]$ are given at the beginning of the kth iteration for some $k \geq 0$. We now describe the strategy for generating the next iterate U_{k+1} and $\hat{\beta}_{k+1}$ by considering the following three different cases:

1) If $X_{\hat{\beta}_k}(U_k)$ is active, find the smallest $s \in \mathcal{Z}_+$ such that $X_{\bar{\beta}}(U_k)$ is inactive, where $\bar{\beta} = \min\{\varsigma_1^s \hat{\beta}_k, \beta\}$. Set $\hat{\beta}_{k+1} = \bar{\beta}$, and apply the algorithm SMACS for problem $(P_{\hat{\beta}_{k+1}})$ starting with the point U_k and set its first iterate to be U_{k+1}.

2) If $X_{\hat{\beta}_k}(U_k)$ is inactive and $\lambda_{\max}(X_{\hat{\beta}_k}(U_k)) \leq \varsigma_3 \hat{\beta}_k$, set $\hat{\beta}_{k+1} = \max\{\min\{\varsigma_2 \lambda_{\max}(X_{\hat{\beta}_k}(U_k)), \beta\}, \alpha\}$. Apply the algorithm SMACS for problem $(P_{\hat{\beta}_{k+1}})$ starting with the point U_k, and set set its next iterate to be U_{k+1}.

3) If $X_{\hat{\beta}_k}(U_k)$ is inactive and $\lambda_{\max}(X_{\hat{\beta}_k}(U_k)) > \varsigma_3\hat{\beta}_k$, set $\hat{\beta}_{k+1} = \hat{\beta}_k$. Continue the algorithm SMACS for problem $(P_{\hat{\beta}_k})$, and set its next iterate to be U_{k+1}.

For the sequences $\{U_k\}_{k=0}^{\infty}$ and $\{\hat{\beta}_k\}_{k=0}^{\infty}$ recursively generated above, we observe that the sequence $\{X_{\hat{\beta}_{k+1}}(U_k)\}_{k=0}^{\infty}$ is always inactive. This together with (33), (34), (28) and the fact that $\hat{\beta}_k \le \beta$ for $k \ge 0$, implies that

$$f(U_k) = f_{\hat{\beta}_{k+1}}(U_k), \quad \nabla f(U_k) = \nabla f_{\hat{\beta}_{k+1}}(U_k), \quad \forall k \ge 0. \tag{35}$$

Therefore, the new termination criterion $f(U_k) - g(X(U_k)) \le \epsilon$ can be replaced by

$$f_{\hat{\beta}_{k+1}}(U_k) - g(X(U_k)) \le \epsilon \tag{36}$$

accordingly.

We now incorporate into the algorithm SMACS the new termination criterion (36) and the aforementioned strategy for generating a sequence $\{\hat{\beta}_k\}_{k=0}^{\infty}$ that asymptotically approaches $\lambda_{\max}(X^*)$, and obtain a variant of the algorithm SMACS for solving problem (26). For convenience of presentation, we omit the subscript from $\hat{\beta}_k$.

Variant of Smooth Minimization Algorithm for Covariance Selection (VSMACS):

Let $\epsilon > 0$, ς_1, $\varsigma_2 > 1$, and $\varsigma_3 \in (0,1)$ be given. Choose a $U_0 \in \mathcal{U}$. Set $\hat{\beta} = \beta$, $L = \rho^2\beta^2$, $\sigma = 1$, and $k = 0$.

1) Compute $X_{\hat{\beta}}(U_k)$ according to (29).

    1a) If $X_{\hat{\beta}}(U_k)$ is active, find the smallest $s \in \mathcal{Z}_+$ such that $X_{\bar{\beta}}(U_k)$ is inactive, where $\bar{\beta} = \min\{\varsigma_1^s\hat{\beta}, \beta\}$. Set $k = 0$, $U_0 = U_k$, $\hat{\beta} = \bar{\beta}$, $L = \rho^2\hat{\beta}^2$, and go to step 2).

    1b) If $X_{\hat{\beta}}(U_k)$ is inactive and $\lambda_{\max}(X_{\hat{\beta}}(U_k)) \le \varsigma_3\hat{\beta}$, set $k = 0$, $U_0 = U_k$, $\hat{\beta} = \max\{\min\{\varsigma_2\lambda_{\max}(X_{\hat{\beta}}(U_k)), \beta\}, \alpha\}$, and $L = \rho^2\hat{\beta}^2$.

2) If $f_{\hat{\beta}}(U_k) - g(X_{\hat{\beta}}(U_k)) \le \epsilon$, terminate. Otherwise, compute $\nabla f_{\hat{\beta}}(U_k)$ according to (29).

3) Find $U_k^{sd} = \mathrm{argmin}\left\{\langle \nabla f_{\hat{\beta}}(U_k), U - U_k\rangle + \frac{L}{2}\|U - U_k\|_F^2 : \ U \in \mathcal{U}\right\}$.

4) Find $U_k^{ag} = \mathrm{argmin}\left\{\frac{L}{2\sigma}\|U - U_0\|_F^2 + \sum\limits_{i=0}^{k}\frac{i+1}{2}[f_{\hat{\beta}}(U_i) + \langle \nabla f_{\hat{\beta}}(U_i), U - U_i\rangle] : \ U \in \mathcal{U}\right\}$.

5) Set $U_{k+1} = \frac{2}{k+3}U_k^{ag} + \frac{k+1}{k+3}U_k^{sd}$.

6) Set $k \leftarrow k + 1$, and go to step 1).

end

We next establish some preliminary convergence properties of the above algorithm.

Proposition 3.3 *For the algorithm VSMACS, the following properties hold:*

1) *Assume that the algorithm VSMACS terminates at some iterate* $(X_{\hat{\beta}}(U_k), U_k)$*. Then* $(X_{\hat{\beta}}(U_k),\, U_k)$ *is an* ϵ*-optimal solution to problem (26) and its dual.*

2) *Assume that* $\hat{\beta}$ *is updated only for a finite number of times. Then the algorithm VSMACS terminates in a finite number of iterations, and produces an* ϵ*-optimal solution to problem (26) and its dual.*

Proof. For the final iterate $(X_{\hat{\beta}}(U_k), U_k)$, we clearly know that $f_{\hat{\beta}}(U_k) - g(X_{\hat{\beta}}(U_k)) \le \epsilon$, and $X_{\hat{\beta}}(U_k)$ is inactive. As shown in (35), $f(U_k) = f_{\hat{\beta}}(U_k)$. Hence, we have $f(U_k) - g(X_{\hat{\beta}}(U_k)) \le \epsilon$. We also know that $U_k \in \mathcal{U}$, and $X_{\hat{\beta}}(U_k) \in \mathcal{X}$ due to $\hat{\beta} \in [\alpha, \beta]$. Thus, statement 1) immediately follows. After the last update of $\hat{\beta}$, the algorithm VSMACS behaves exactly like the algorithm SMACS as applied to solve the problem $(P_{\hat{\beta}})$ except with the termination criterion $f(U_k) - g(X_{\hat{\beta}}(U_k)) \le \epsilon$. Thus, it follows from statement 1) and Theorem 2.4 that statement 2) holds. ■

4 Computational results

In this section, we compare the performance of the smooth minimization approach and its variant proposed in this paper with two other first-order methods studied in d'Aspremont et al. [9], that is, Nesterov's smooth approximation scheme, and block coordinate descent method for solving problem (15) (or equivalently, (26)) on a set of randomly generated instances.

All instances used in this section were randomly generated in the same manner as described in d'Aspremont et al. [9]. First, we generate a sparse invertible matrix $A \in \mathcal{S}^n$ with positive diagonal entries and a density prescribed by ϱ. We then generate the matrix $B \in \mathcal{S}^n$ by

$$B = A^{-1} + \tau V,$$

where $V \in \mathcal{S}^n$ is an independent and identically distributed uniform random matrix, and τ is a small positive number. Finally, we obtain the following randomly generated sample covariance matrix:

$$\Sigma = B - \min\{\lambda_{\min}(B) - \vartheta, 0\} I,$$

where ϑ is a small positive number. In particular, we set $\varrho = 0.01$, $\tau = 0.15$ and $\vartheta = 1.0e-4$ for generating all instances.

In the first experiment, we compare the performance of our smooth minimization approach and its variant with Nesterov's smooth approximation scheme studied in d'Aspremont et al. [9] for the problem (26) with $\alpha = 0.1$, $\beta = 10$ and $\rho = 0.5$. For convenience of presentation, we label these three first-order methods as SM, VSM, and NSA, respectively. The codes for them are written in Matlab. More specifically, the code for NSA follows the algorithm presented in d'Aspremont et al. [9], and the codes for SM and VSM are written in accordance with the algorithms SMACS and VSMACS, respectively. Moreover, we set $\varsigma_1 = \varsigma_2 = 1.05$ and $\varsigma_3 = 0.95$

Table 1: Comparison of NSA, SM and VSM

Problem	Iter			Obj			Time		
n	NSA	SM	VSM	NSA	SM	VSM	NSA	SM	VSM
50	3657	456	19	-76.399	-76.399	-76.393	51.2	5.1	0.2
100	7629	919	26	-186.717	-186.720	-186.714	902.9	57.3	1.0
150	20358	1454	48	-318.195	-318.194	-318.184	8172.3	256.8	4.0
200	27499	2293	101	-511.246	-511.245	-511.242	26151.0	952.0	17.0
250	45122	3059	127	-3793.255	-3793.256	-3793.257	86915.4	2334.0	36.1
300	54734	3880	160	-3187.163	-3187.171	-3187.172	183485.2	5015.9	73.3
350	64641	4633	181	-2756.717	-2756.734	-2756.734	350010.1	9264.2	126.6
400	74839	5307	175	-3490.640	-3490.667	-3490.667	612823.1	16244.1	173.0
450	85948	5937	177	-3631.003	-3631.044	-3631.043	1013189.4	26006.2	251.7

for the algorithm VSMACS. These three methods terminate once the duality gap is less than $\epsilon = 0.1$. All computations are performed on an Intel Xeon 2.66 GHz machine with Red Hat Linux version 8.

The performance of the methods NSA, SM and VSM for the randomly generated instances are presented in Table 1. The row size n of each sample covariance matrix Σ is given in column one. The numbers of iterations of NSA, SM and VSM are given in columns two to four, and the objective function values are given in columns five to seven, and the CPU times (in seconds) are given in the last three columns, respectively. From Table 1, we conclude that: i) the method SM, namely, the smooth minimization approach, outperforms substantially the method NSA, that is, Nesterov's smooth approximation scheme; and ii) the method VSM, namely, the variant of the smooth minimization approach, substantially outperforms the other two methods.

In the above experiment, we have already seen that the method VSM outperforms substantially two other first-order methods, namely, SM and NSA for solving problem (26). In the second experiment, we compare the performance of the method VSM with another first-order method studied in d'Aspremont et al. [9], that is, block coordinate descent method (*labeled as* BCD). It shall be reminded that the method BCD is only applicable for solving the problem (15) with $\tilde{\alpha} = 0$ and $\tilde{\beta} = \infty$. Thus, we only compare the performance of these two methods for the problem (15) with such $\tilde{\alpha}$ and $\tilde{\beta}$. As shown in Subsection 3.2, the problem (15) with $\tilde{\alpha} = 0$ and $\tilde{\beta} = \infty$ is equivalent to the problem (26) with α and β given in (17), and hence it can be solved by applying the method VSM to the latter problem instead. The code for the method BCD was written in Matlab by d'Aspremont and El Ghaoui [10]. These two methods terminate once the duality gap is less than $\epsilon = 0.1$. All computations are performed on an Intel Xeon 2.66 GHz machine with Red Hat Linux version 8.

The performance of the methods BCD and VSM for the randomly generated instances are presented in Table 2. The row size n of each sample covariance matrix Σ is given in column one. The numbers of iterations of BCD and VSM are given in columns two to three, and the objective function values are given in columns four to six, and the CPU times (in seconds) are given in the last two columns, respectively. It shall be mentioned that BCD and VSM are both

Table 2: Comparison of BCD and VSM

Problem	Iter		Obj		Time	
n	BCD	VSM	BCD	VSM	BCD	VSM
100	124	32	-186.522	-186.522	20.7	1.2
200	531	108	-449.210	-449.209	289.2	18.1
300	1530	145	-767.615	-767.614	2235.6	67.1
400	2259	153	-1082.679	-1082.677	8144.7	152.5
500	3050	153	-1402.503	-1402.502	21262.4	271.9
600	3705	164	-1728.628	-1728.627	45093.8	473.9
700	4492	162	-2057.894	-2057.892	85268.3	705.7
800	4958	168	-2392.713	-2392.712	136077.9	1052.9
900	5697	160	-2711.874	-2711.874	204797.1	1409.8
1000	6536	160	-3045.808	-3045.808	326414.0	1951.3

feasible methods, and moreover, (15) and (26) are maximization problems. Therefore for these two methods, the larger objective function value, the better. From Table 2, we conclude that the method VSM, namely, the variant of the smooth minimization approach, substantially outperforms the BCD, that is, block coordinate descent method, and it also produces better objective function values for most of the instances.

5 Concluding remarks

In this paper, we proposed a smooth optimization approach for solving a class of non-smooth strongly concave maximization problems. We also discussed the application of this approach to covariance selection, and proposed a variant of this approach. The computational results showed that the variant of the smooth optimization approach outperforms substantially the latter one, and two other first-order methods studied in d'Aspremont et al. [9].

Although some preliminary convergence properties of the variant of the smooth optimization approach are established in Subsection 3.4, we are currently unable to provide a thorough proof of its convergence. A possible direction leading to a complete proof would be to show that the update on $\hat{\beta}$ in the algorithm VSMACS can occur only for a finite number of times. This is left for the future research.

In addition, the current code for the variant of the smooth minimization approach is written in Matlab, which is available for free download online at www.math.sfu.ca/∼zhaosong. The C code for this method is under development, which would enable us to solve large-scale problems more efficiently. We will also plan to extend these codes for solving more general problems of the form

$$\begin{array}{ll} \max\limits_{X} & \log\det X - \langle \Sigma, X\rangle - \sum\limits_{ij} \omega_{ij}|X_{ij}| \\ \text{s.t.} & \tilde{\alpha} I \preceq X \preceq \tilde{\beta} I, \end{array}$$

where $\omega_{ij} = \omega_{ji} \geq 0$ for all $i, j = 1, \ldots, n$, and $0 \leq \tilde{\alpha} < \tilde{\beta} \leq \infty$ are the fixed bounds on the eigenvalues of the solution.

Acknowledgement

The author would like to thank Professor Alexandre d'Aspremont for a careful discussion on the iteration complexity of Nesterov's smooth approximation scheme for covariance selection given in [9].

References

[1] J. Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki, editors, *Second international symposium on information theory*, pages 267–281. Akedemiai Kiado, Budapest, 1973.

[2] A. Auslender and M. Teboulle. Interior gradient and proximal methods for convex and conic optimization. *SIAM Journal on Optimization*, 16:697–725, 2006.

[3] O. Banerjee, L. El Ghaoui, A. d'Aspremont, and G. Natsoulis. Convex optimization techniques for fitting sparse gaussian graphical models. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, pages 89–96, New York, NY, USA, 2006. ACM Press.

[4] J. A. Bilmes. Factored sparse inverse covariance matrices. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2:1009–1012, 2000.

[5] K. P. Burnham and R. D. Anderson. Multimodel inference. understanding aic or bic in model selection. *Sociological methods and research*, 33:261–304, 2004.

[6] S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20:33–61, 1998.

[7] J. Dahl, V. Roychowdhury, and L. Vandenberghe. Maximum likelihood estimation of gaussian graphical models: numerical implementation and topology selection. Manuscript, University of California Los Angeles, October 2004.

[8] J. Dahl, L. Vandenberghe, and V. Roychowdhury. Covariance selection for non-chordal graphs via chordal embedding. Manuscript, University of California Los Angeles, July 2006. Submitted to *Optimization Methods and Software*.

[9] A. d'Aspremont, O. Banerjee, and L. El Ghaoui. First-order methods for sparse covariance selection. Technical report, Princeton University, 2006.

[10] A. d'Aspremont and L. El Ghaoui. Covsel: First order methods for sparse covariance selection, March 2006.

[11] A. Dempster. Covariance selection. *Biometrics*, 28:157–175, 1972.

[12] A. Dobra, C. Hans, B. Jones, J. R. Nevins, G. Yao, and M. West. Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, 90:196–212, 2004.

[13] A. Dobra and M. West. Bayesian covariance selection. Isds working paper, Duke University, July 2004.

[14] D. L. Donoho and J. Tanner. Sparse nonnegative solutions of underdetermined linear equations by linear programming. *Proceedings of the National Academy of Sciences*, 102:9446–9451, 2005.

[15] J. Z. Huang, N. Liu, and M. Pourahmadi. Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1):85–98, 2006.

[16] B. Jones, C. Carvalho, C. Dobra, A. Hans, C. Carter, and M. West. Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science*, 20:388–400, 2005.

[17] Y. E. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. *Doklady AN SSSR*, 269:543–547, 1983. translated as Soviet Math. Docl.

[18] Y. E. Nesterov. Smooth minimization of nonsmooth functions. *Mathematical Programming*, 103:127–152, 2005.

[19] Y. E. Nesterov and A. S. Nemirovski. *Interior point Polynomial algorithms in Convex Programming: Theory and Applications*. Society for Industrial and Applied Mathematics, Philadelphia, 1994.

[20] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal statis tical society, series B*, 58:267–288, 1996.

[21] L. Vandenberghe, S. Boyd, and S. Wu. Determinant maximization with linear matrix inequality constraints. *SIAM Journal on Matrix Analysis and Applications*, 19(2):499–533, 1998.

[22] M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.