Adaptive First-Order Methods for General Sparse Inverse Covariance Selection

Zhaosong Lu[*]

December 2, 2008
Revised: September 15, 2009; January 30, 2010

Abstract

In this paper we consider estimating sparse inverse covariance of a Gaussian graphical model whose conditional independence is assumed to be *partially* known. Similarly as in [7, 18], we formulate it as an l_1-norm penalized maximum likelihood estimation problem. Further, we propose an algorithm framework, and develop two first-order methods, that is, the adaptive spectral projected gradient (ASPG) method and the adaptive Nesterov's smooth (ANS) method, for solving this estimation problem. Finally, we compare the performance of these two methods with glasso [10, 11] on a set of randomly generated instances. Our computational results demonstrate that our methods are capable of solving problems of size at least a thousand and number of constraints of nearly a half million within a reasonable amount of time, and moreover the ASPG method generally outperforms the ANS method and glasso.

Key words: Sparse inverse covariance selection, adaptive spectral projected gradient method, adaptive Nesterov's smooth method

AMS 2000 subject classification: 90C22, 90C25, 90C47, 65K05, 62J10

1 Introduction

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 setting a certain number of entries in the inverse covariance matrix to zero (e.g., see [8]). Since zeros in the inverse of covariance matrix correspond to conditional independence among the variables, covariance

[*]Department of Mathematics, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada. (email: zhaosong@sfu.ca). This author was supported in part by SFU President's Research Grant and NSERC Discovery Grant.

selection also highlights the sparse structure in the underlying model. It has numerous real-word applications, for example, speech recognition [3] and gene network analysis [9].

In the recent years, a variety of approaches have been proposed to determine a robust estimate of the true variance matrix, and simultaneously discover the pattern of zeros in the inverse covariance matrix. (All notations used below are defined in Subsection 1.1.) Given a sample covariance matrix $\Sigma \in \mathcal{S}^n_+$, d'Aspremont et al. [7] formulated sparse inverse covariance selection as the following l_1-norm penalized maximum likelihood estimation problem:

$$\max_X \, \{\log\det X - \langle \Sigma, X\rangle - \rho e^T|X|e : \; X \succeq 0\}, \tag{1}$$

where $\rho > 0$ is a parameter controlling the trade-off between likelihood and sparsity of the solution. They also studied Nesterov's smooth approximation scheme [15] and block-coordinate descent (BCD) method for solving (1). Independently, Yuan and Lin [18] proposed a similar estimation problem to (1) as follows:

$$\max_X \, \{\log\det X - \langle \Sigma, X\rangle - \rho \sum_{i\neq j} |X_{ij}| : \; X \succeq 0\}. \tag{2}$$

They showed that problem (2) can be suitably solved by the interior point algorithm developed in Vandenberghe et al. [17]. As demonstrated in [7, 18], the estimation problems (1) and (2) are capable of discovering the sparse structure, that is, the conditional independence in the underlying graphical model. Recently, Lu [13] proposed a variant of Nesterov's smooth method [15] for problems (1) and (2) that substantially outperforms the existing methods in literature. In addition, Dahl et al. [6] studied the maximum likelihood estimation of a Gaussian graphical model whose conditional independence is known, which can be formulated as

$$\max_X \, \{\log\det X - \langle \Sigma, X\rangle : \; X \succeq 0, \; X_{ij} = 0 \; \forall (i,j) \in \bar{E}\}, \tag{3}$$

where $\bar{E}$ is a collection of all pairs of conditional independent nodes. They showed that when the underlying graph is nearly-chordal, Newton's method and preconditioned conjugate gradient method can be efficiently applied to solve (3).

In practice, the sparsity structure of a Gaussian graphical model is often partially known from the knowledge of its random variables. In this paper we consider estimating sparse inverse covariance of a Gaussian graphical model whose conditional independence is assumed to be *partially* known in advance (but it can be completely unknown). Given a sample covariance matrix $\Sigma \in \mathcal{S}^n_+$, we can naturally formulate it as the following constrained l_1-norm penalized maximum likelihood estimation problem:

$$\begin{array}{cl} \max\limits_X & \log\det X - \langle \Sigma, X\rangle - \sum\limits_{(i,j)\notin\Omega} \rho_{ij}|X_{ij}|, \\ \text{s.t.} & X \succeq 0, \; X_{ij} = 0 \; \forall (i,j) \in \Omega, \end{array} \tag{4}$$

where Ω consists of a set of pairs of conditionally independent nodes, and $\{\rho_{ij}\}_{(i,j)\notin\Omega}$ is a set of nonnegative parameters controlling the trade-off between likelihood and sparsity of the solution. It shall be mentioned that unlike in [6], we do not assume any specific structure on the sparsity of underlying graph for problem (4). Clearly, we can observe that (i) $(i,i) \notin \Omega$ for $1 \leq i \leq n$, and $(i,j) \in \Omega$ if and only if $(j,i) \in \Omega$; (ii) $\rho_{ij} = \rho_{ji}$ for any $(i,j) \notin \Omega$; and (iii) problems (1)-(3) can be viewed as special cases of problem (4) by choosing appropriate Ω and $\{\rho_{ij}\}_{(i,j)\notin\Omega}$. For example, if setting $\Omega = \emptyset$ and $\rho_{ij} = \rho$ for all (i,j), problem (4) becomes (1).

It is easy to observe that problem (4) can be reformulated as a constrained smooth convex problem that has an explicit $\mathcal{O}(|\bar{\Omega}|)$-logarithmically homogeneous self-concordant barrier function, where $|\bar{\Omega}|$ is the cardinality of the complement of Ω. Thus it can be suitably solved by interior point (IP) methods (see Nesterov and Nemirovski [16] and Vandenberghe et al. [17]). The worst-case iteration complexity of IP methods for finding an ϵ-optimal solution to (4) is $\mathcal{O}(|\bar{\Omega}|^{1/2}\log(\epsilon_0/\epsilon))$, where ϵ_0 is an initial gap. For a general graph, each iterate of IP methods requires $\mathcal{O}(n^6)$ arithmetic cost for assembling and solving a typically dense Newton system with $\mathcal{O}(n^2)$ variables. Therefore, the total worst-case arithmetic cost of IP methods for finding an ϵ-optimal solution to (4) is $\mathcal{O}(|\bar{\Omega}|^{1/2}n^6\log(\epsilon_0/\epsilon))$, which is prohibitive when n is relatively large. We shall mention that when the underlying graph is nearly-chordal, the worst-case arithmetic cost of IP methods can be dramatically reduced (see, for example, [6]).

Recently, Friedman et al. [10, 11] proposed a gradient type method, called *glasso*, for solving problem (4). They first converted (4) into the following penalization problem

$$\max_{X \succeq 0} \ \log\det X - \langle \Sigma, X\rangle - \sum_{i,j} \rho_{ij}|X_{ij}| \tag{5}$$

by setting ρ_{ij} to an extraordinary large number (say, 10^9) for all $(i,j) \in \Omega$. Then they applied a slight variant of the BCD method [7] to the dual of (5) in which each iteration solves a lasso (l_1-regularized) least-squares problem using a coordinate descent approach. We observe that glasso performs extremely well when ρ_{ij} is relatively large for all $(i,j) \notin \Omega$, the underlying graph is highly sparse and a large percentage of pairs of conditional independent nodes is known beforehand, but otherwise its performance may decline substantially (see Section 3 for details).

In this paper we propose adaptive first-order methods for problem (4). Instead of solving (5) only once with a set of huge penalty parameters $\{\rho_{ij}\}_{(i,j)\in\Omega}$, our methods consist of solving a sequence of problems (5) with a set of moderate penalty parameters $\{\rho_{ij}\}_{(i,j)\in\Omega}$ that are gradually adjusted until a desired approximate solution is found. We also derive an explicit upper bound on the number of updates on the penalty parameters. In addition, for each given ρ, problem (5) is solved by the adaptive spectral projected gradient (ASPG) method and the adaptive Nesterov's smooth (ANS) method that are proposed in this paper. The computational results demonstrate that our methods are capable of solving problems of size at least a thousand and number of constraints of nearly a half million within a reasonable amount of time, and moreover the ASPG method generally outperforms the ANS method and glasso [10, 11].

The rest of paper is organized as follows. In Subsection 1.1, we introduce the notations used in this paper. In Section 2, we propose an algorithm framework and develop two first-order methods, that is, the ASPG and ANS methods, for solving problem (4). The performance of these two methods are compared with glasso [10, 11] on a set of randomly generated instances in Section 3. Finally, we present some concluding remarks in Section 4.

1.1 Notation

In this paper, all vector spaces are assumed to be finite dimensional. The symbols $\Re^n$, $\Re^n_+$ and $\Re^n_{++}$ denote the n-dimensional Euclidean space, the nonnegative orthant of $\Re^n$ and the positive orthant of $\Re^n$, respectively. The set of all $m \times n$ matrices with real entries is denoted by $\Re^{m\times n}$. 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 $\Re^{m\times n}$, 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 $|X|$ 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, and $\lambda_i(X)$ denotes its ith largest eigenvalue. Given an $n \times n$ (partial) matrix ρ, $\mathrm{Diag}(\rho)$ denotes the diagonal matrix whose ith diagonal element is ρ_{ii} for $i = 1, \ldots, n$. Given matrices X and Y in $\Re^{m\times n}$, $X * Y$ denotes the pointwise product of X and Y, namely, $X * Y \in \Re^{m\times n}$ whose ijth entry is $X_{ij}Y_{ij}$ for all i, j. Finally, we denote by $\mathcal{Z}_+$ the set of all nonnegative integers.

2 Adaptive first-order methods

In this section, we discuss some suitable first-order methods for general sparse inverse covariance selection problem (4). In particular, we first provide an algorithm framework for it in Subsection 2.1. Then we specialize this framework by considering two first-order methods, namely, the adaptive spectral projected gradient method and the adaptive Nesterov's smooth method in Subsection 2.2.

2.1 Algorithm framework

In this subsection, we provide an algorithm framework for general sparse inverse covariance selection problem (4).

Throughout this paper, we make the following assumption for problem (4).

Assumption 1 *For all $(i, j) \notin \Omega$, $\rho_{ij} \geq 0$ is given and fixed, and moreover,* $\Sigma + \mathrm{Diag}(\rho) \succ 0$.

Note that Σ is a sample covariance matrix, and hence $\Sigma \succeq 0$. In addition, $\mathrm{Diag}(\rho) \succeq 0$. Thus $\Sigma + \mathrm{Diag}(\rho) \succeq 0$. It may not be, however, positive definite in general. But we can always perturb ρ_{ii} by adding a small positive number (say, 10^{-8}) when necessary to ensure the above assumption holds.

We first establish the existence of an optimal solution for problem (4) as follows.

Proposition 2.1 *Problem (4) has a unique optimal solution $X^* \in \mathcal{S}^n_{++}$.*

Proof. Let $f(X)$ denote the objective function of (4). Since $(i,i) \notin \Omega$ for $i = 1, \ldots, n$, it is easy to see that $X = I$ is a feasible solution of problem (4). We now show that the sup-level set $S_f(I) = \{X \succeq 0 : \ f(X) \geq f(I), \ X_{ij} = 0 \ \forall (i,j) \in \Omega\}$ is compact. Indeed, using the definition of $f(\cdot)$, we observe that for any $X \in S_f(I)$,

$$\begin{aligned} f(I) &\leq f(X) \leq \log\det X - \langle \Sigma + \mathrm{Diag}(\rho), X\rangle \leq \sum_{i=1}^{n} \left[\log\lambda_i(X) - \lambda_{\min}(\Sigma+\mathrm{Diag}(\rho))\lambda_i(X)\right], \\ &\leq (n-1)\left[-1 - \log\lambda_{\min}(\Sigma+\mathrm{Diag}(\rho))\right] + \log\lambda_{\max}(X) - \lambda_{\min}(\Sigma+\mathrm{Diag}(\rho))\lambda_{\max}(X), \end{aligned}$$

where the last inequality follows from the fact that for any $a > 0$,

$$\max_t \{\log t - at : \ t \geq 0\} \ = \ -1 - \log a. \tag{6}$$

Hence, we obtain that for any $X \in S_f(I)$,

$$\log\lambda_{\max}(X) - \lambda_{\min}(\Sigma+\mathrm{Diag}(\rho))\lambda_{\max}(X) \ \geq \ f(I) - (n-1)\left[-1 - \log\lambda_{\min}(\Sigma+\mathrm{Diag}(\rho))\right], \tag{7}$$

which implies that there exists some $\beta(\rho) > 0$ such that $\lambda_{\max}(X) \leq \beta(\rho)$ for all $X \in S_f(I)$. Thus, $S_f(I) \subseteq \{X \in \mathcal{S}^n : 0 \preceq X \preceq \beta(\rho)I\}$. Further, using this result along with $\Sigma \succeq 0$ and the definition of $f(\cdot)$, we easily observe that for any $X \in S_f(I)$,

$$\begin{aligned} \log\lambda_{\min}(X) &= f(X) - \sum_{i=1}^{n-1}\log\lambda_i(X) + \langle\Sigma, X\rangle + \sum_{(i,j)\notin\Omega}\rho_{ij}|X_{ij}|, \\ &\geq f(I) - (n-1)\log\beta(\rho). \end{aligned}$$

It follows that there exists some $\alpha(\rho) > 0$ such that $\lambda_{\min}(X) \geq \alpha(\rho)$ for all $X \in S_f(I)$. Hence, $S_f(I) \subseteq \{X \in \mathcal{S}^n : \alpha(\rho)I \preceq X \preceq \beta(\rho)I\}$ is bounded, which together with the fact that $f(\cdot)$ is continuous in the latter set, implies that $S_f(I)$ is compact. Therefore, problem (4) has at least an optimal solution. Further, observing that $f(\cdot)$ is strictly concave, we conclude that problem (4) has a unique optimal solution. ■

Similarly, we can show that the following result holds.

Proposition 2.2 *Under Assumption 1, problem (5) has a unique optimal solution in $\mathcal{S}^n_{++}$ for any $\rho_{ij} \geq 0 \ \forall (i,j) \in \Omega$, where Ω is given in (4).*

Before presenting an algorithm framework for problem (4), we introduce a terminology for (4) as follows.

Definition 1 *Let $\epsilon_o \geq 0$ and $\epsilon_c \geq 0$ be given. Let $f(\cdot)$ and f^* denote the objective function and the optimal value of (4), respectively. $X \in \mathcal{S}^n_+$ is an (ϵ_o, ϵ_c)-optimal solution of problem (4) if $f(X) \geq f^* - \epsilon_o$ and $\max\limits_{(i,j)\in\Omega} |X_{ij}| \leq \epsilon_c$.*

Analogously, we can define an ϵ_o-optimal solution for problem (5). Given that our ultimate aim is to estimate a sparse inverse covariance matrix $X^* \succeq 0$ that satisfies $X^*_{ij} = 0 \;\forall (i,j) \in \Omega$ and approximately maximizes the log-likelihood, we now briefly discuss how to obtain such an approximate solution X^* from an (ϵ_o, ϵ_c)-optimal solution $\bar{X}^*$ of (4). Let us define $\tilde{X}^* \in \mathcal{S}^n$ by letting $\tilde{X}^*_{ij} = \bar{X}^*_{ij} \;\forall (i,j) \notin \Omega$ and $\tilde{X}^*_{ij} = 0 \;\forall (i,j) \in \Omega$. We then set $X^* := \tilde{X}^* + t^* I$, where

$$t^* = \arg\max\{\log\det(\tilde{X}^* + tI) - \langle \Sigma, \tilde{X}^* + tI\rangle : \; t \geq -\lambda_{\min}(\tilde{X}^*)\}.$$

It is not hard to see that t^* can be easily found. We also observe that $X^* \in \mathcal{S}^n_{++}$, and moreover, it satisfies $X^*_{ij} = 0 \;\forall (i,j) \in \Omega$ and retains the same sparsity as $\tilde{X}^*$. In addition, by setting the log-likelihood value at $\tilde{X}^*$ to $-\infty$ if $\lambda_{\min}(\tilde{X}^*) \leq 0$, we can see that the log-likelihood value at X^* is at least as good as that at $\tilde{X}^*$. Thus, X^* is a desirable estimation of sparse inverse covariance, provided $\bar{X}^*$ is a good approximate solution to (4).

In the remainder of this paper, we focus on finding an (ϵ_o, ϵ_c)-optimal solution of problem (4) for any pair of positive (ϵ_o, ϵ_c). We next present an algorithm framework for it based on an adaptive l_1 penalty approach.

Algorithm framework for general sparse inverse covariance selection (GSICS):

Let $\epsilon_o > 0$, $\epsilon_c > 0$ and $r_\rho > 1$ be given. Let $\rho^0_{ij} > 0, \forall (i,j) \in \Omega$ be given such that $\rho^0_{ij} = \rho^0_{ji}, \forall (i,j) \in \Omega$. Set $\rho_{ij} = \rho^0_{ij}$ for all $(i,j) \in \Omega$.

1) Find an ϵ_o-optimal solution X^{ϵ_o} of problem (5).

2) If $\max\limits_{(i,j)\in\Omega} |X^{\epsilon_o}_{ij}| \leq \epsilon_c$, terminate. Otherwise, set $\rho_{ij} \leftarrow \rho_{ij} r_\rho$ for all $(i,j) \in \Omega$, and go to step 1).

end

Remark. To make the above framework complete, one needs to choose suitable methods for finding an approximate solution of (5) in step 1). We will propose two first-order methods for it in Subsection 2.2. In step 2), some other strategies can also be applied for updating the penalty parameters $\{\rho_{ij}\}_{(i,j)\in\Omega}$. For example, given any $(i,j) \in \Omega$, one can update ρ_{ij} only when $|X^{\epsilon_o}_{ij}| > \epsilon_c$. But we observed in our experimentation that such a strategy generally performs worse than the one described above. In addition, instead of using a common ratio r_ρ for all $(i,j) \in \Omega$, one can associate with each $(i,j) \in \Omega$ an individual ratio r_{ij}. Also, the ratio r_ρ does not need to be fixed for all iterations, and it can vary from iteration to iteration depending on the amount of violation incurred in $\max\limits_{(i,j)\in\Omega} |X^{\epsilon_o}_{ij}| \leq \epsilon_c$. ■

Before establishing the convergence for the framework GSICS, we first study the convergence of the l_1-penalty method for finding an approximate solution of a general nonlinear programming (NLP) problem.

Given a set $\emptyset \neq \mathcal{X} \subseteq \Re^n$ and functions $f : \mathcal{X} \to \Re$, $g : \mathcal{X} \to \Re^k$ and $h : \mathcal{X} \to \Re^l$, consider the NLP problem:

$$\begin{array}{rcl} f^* & = & \sup\limits_{x \in \mathcal{X}} \;\; f(x) \\ & & \text{s.t.} \;\; g(x) = 0, \;\; h(x) \leq 0. \end{array} \tag{8}$$

We associate with the NLP problem (8) the following l_1-penalized function:

$$P(x; \lambda, \mu) \;:=\; f(x) - \lambda^T |g(x)| - \mu^T h^+(x), \tag{9}$$

where $\lambda \in \Re^k_+$, $\mu \in \Re^l_+$ and $(h^+(x))_i = \max\{0, h_i(x)\}$ for $i = 1, \ldots, l$.

We now establish a convergence result for the l_1 penalty method for finding an approximate solution of the NLP problem (8) under some assumption on $f(x)$.

Proposition 2.3 *Let $\epsilon_o > 0$ and $\epsilon_c > 0$ be given. Assume that there exists some $\bar{f} \in \Re$ such that $f(x) \leq \bar{f}$ for all $x \in \mathcal{X}$. Let $x^{\epsilon_o}_{\lambda,\mu} \in \mathcal{X}$ be an ϵ_o-optimal solution of the problem*

$$\sup\{P(x; \lambda, \mu) : \; x \in \mathcal{X}\} \tag{10}$$

for $\lambda \in \Re^k_+$ and $\mu \in \Re^l_+$, and let $v_{\lambda,\mu} := \min\{\min\limits_i \lambda_i, \min\limits_i \mu_i\}$. Then $f(x^{\epsilon_o}_{\lambda,\mu}) \geq f^* - \epsilon_o$, and moreover, $\left\|\left(g(x^{\epsilon_o}_{\lambda,\mu}); h^+(x^{\epsilon_o}_{\lambda,\mu})\right)\right\|_\infty \leq \epsilon_c$ holds whenever $v_{\lambda,\mu} \geq (\bar{f} - f^* + \epsilon_o)/\epsilon_c$, where f^* is the optimal value of the NLP problem (8).

Proof. Let $f^*_{\lambda,\mu}$ denote the optimal value of problem (10). By the assumption that $f(x)$ is bounded above in $\mathcal{X}$, we see that both f^* and $f^*_{\lambda,\mu}$ are finite. Also, we observe that $f^*_{\lambda,\mu} \geq f^*$. Using this relation, (9) and the fact that $x^{\epsilon_o}_{\lambda,\mu}$ is an ϵ_o-optimal solution of (10), we have

$$f(x^{\epsilon_o}_{\lambda,\mu}) \;\geq\; P(x^{\epsilon_o}_{\lambda,\mu}; \lambda, \mu) \;\geq\; f^*_{\lambda,\mu} - \epsilon_o \;\geq\; f^* - \epsilon_o, \tag{11}$$

and hence, the first statement holds. We now prove the second statement. Using (9), (11) and the definition of $v_{\lambda,\mu}$, we have

$$\begin{array}{rcl} f(x^{\epsilon_o}_{\lambda,\mu}) - v_{\lambda,\mu} \left\|\left(g(x^{\epsilon_o}_{\lambda,\mu}); h^+(x^{\epsilon_o}_{\lambda,\mu})\right)\right\|_\infty & \geq & f(x^{\epsilon_o}_{\lambda,\mu}) - v_{\lambda,\mu} \left\|\left(g(x^{\epsilon_o}_{\lambda,\mu}); h^+(x^{\epsilon_o}_{\lambda,\mu})\right)\right\|_1 \\ & \geq & P(x^{\epsilon_o}_{\lambda,\mu}; \lambda, \mu) \;\geq\; f^* - \epsilon_o. \end{array} \tag{12}$$

From the assumption, we further know that $f(x^{\epsilon_o}_{\lambda,\mu}) \leq \bar{f}$ due to $x^{\epsilon_o}_{\lambda,\mu} \in \mathcal{X}$. This together with (12) immediately implies that the second statement holds. ■

We are now ready to establish a convergence result for the framework GSICS.

Theorem 2.4 *Let $\epsilon_o > 0$ and $\epsilon_c > 0$ be given. The framework GSICS generates an (ϵ_o, ϵ_c)-optimal solution to problem (4) in at most*

$$\left\lceil \frac{1}{\log r_\rho} \left\{ \log[\mathrm{Tr}(\Sigma + \mathrm{Diag}(\rho)) - \log\det(\Sigma + \mathrm{Diag}(\rho)) - n + \epsilon_o] - \log \epsilon_c - \log \min_{(i,j)\in\Omega} \rho^0_{ij} \right\} \right\rceil$$

number of outer iterations, or equivalently, updates on the penalty parameters $\{\rho_{ij}\}_{(i,j)\in\Omega}$.

Proof. Invoking that $\Sigma + \mathrm{Diag}(\rho) \succ 0$ (see Assumption 1), we obtain that for any $X \in \mathcal{S}^n_+$,

$$\begin{array}{rcl} \log\det X - \langle \Sigma, X\rangle - \sum\limits_{(i,j)\notin\Omega} \rho_{ij}|X_{ij}| & \leq & \log\det X - \langle \Sigma + \mathrm{Diag}(\rho), X\rangle \\ & \leq & \sup\{\log\det Y - \langle \Sigma + \mathrm{Diag}(\rho), Y\rangle : Y \succeq 0\} \\ & = & -\log\det(\Sigma + \mathrm{Diag}(\rho)) - n, \end{array}$$

where the last inequality follows from the fact that the supremum is achieved at $Y = (\Sigma + \mathrm{Diag}(\rho))^{-1} \succ 0$. Hence, the objective function of (4) is bounded above by $\bar{f} = -\log\det(\Sigma + \mathrm{Diag}(\rho)) - n$ in $\mathcal{S}^n_+$. In addition, we know that $X = I$ is a feasible solution of (4). Thus the optimal value f^* of (4) is bounded below by the objective function value at $X = I$, which is $-\mathrm{Tr}(\Sigma + \mathrm{Diag}(\rho))$. Using these observations and Proposition 2.3, we can easily see that the conclusion holds. ■

2.2 Adaptive first-order methods for problem (5)

In this subsection, we discuss some suitable methods for solving problem (5) that arises in step 1) of the framework GSICS. In particular, we first reformulate (5) into a smooth minimization problem. Then we propose two first-order methods for solving (5) and its smooth counterpart simultaneously. We also discuss how to efficiently incorporate them into the framework GSICS for solving problem (4).

2.2.1 Smooth reformulation

It is clear that (5) is generally a nonsmooth problem. In this subsection, we will reformulate it into a smooth minimization problem.

As stated in Proposition 2.2, problem (5) has a unique optimal solution. We next provide some bounds on it. Before proceeding, we define

$$\mathcal{U} := \{U \in \mathcal{S}^n : \ |U_{ij}| \leq 1 \ \forall ij\}, \tag{13}$$

and

$$\phi(X, U) := \log\det X - \langle \Sigma + \rho * U, X\rangle \quad \forall (X, U) \in \mathcal{S}^n_{++} \times \mathcal{U}. \tag{14}$$

Proposition 2.5 *Let $f_\rho(\cdot)$ and X^*_ρ denote the objective function and the unique optimal solution of problem (5), respectively. Let ϑ be defined as*

$$\vartheta := \max\left\{f_\rho((\Sigma + \mathrm{Diag}(\rho))^{-1}), \theta\right\} - (n-1)[-1 - \log\lambda_{\min}(\Sigma + \mathrm{Diag}(\rho))], \tag{15}$$

where $\theta := n(-1 - \log\mathrm{Tr}(\Sigma + \rho) + \log n)$. Then $\alpha_\rho I \preceq X^*_\rho \preceq \beta_\rho I$, where $\alpha_\rho := 1/(\|\Sigma\| + \|\rho\|)$ and β_ρ is the largest positive root of the following equation

$$\log t - \lambda_{\min}(\Sigma + \mathrm{Diag}(\rho))t - \vartheta = 0.$$

Proof. Let $\mathcal{U}$ and $\phi(\cdot,\cdot)$ be defined in (13) and (14), respectively. Since $X^*_\rho \in \mathcal{S}^n_{++}$ is the optimal solution of problem (5), it can be easily shown that there exists some $U^* \in \mathcal{U}$ such that (X^*_ρ, U^*) is a saddle point of $\phi(\cdot,\cdot)$ in $\mathcal{S}^n_{++} \times \mathcal{U}$, and hence

$$X^*_\rho = \arg \min_{X \in \mathcal{S}^n_{++}} \phi(X, U^*).$$

This relation along with (14) immediately yields $X^*_\rho(\Sigma + \rho * U^*) = I$. Hence, we have

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

which together with (13) and the fact that $U^* \in \mathcal{U}$, implies that $X^* \succeq \frac{1}{\|\Sigma\|+\|\rho\|} I = \alpha_\rho I$.

We next bound X^*_ρ from above. Let f^*_ρ denote the optimal value of problem (5). In view of the definition of $f_\rho(\cdot)$ and (6), we have

$$f^*_\rho \;\geq\; \max_{t>0} f_\rho(tI) \;=\; \max_{t>0}\; n \log t - t\mathrm{Tr}(\Sigma + \rho) \;=\; n(-1 - \log \mathrm{Tr}(\Sigma + \rho) + \log n) \;=:\; \theta.$$

Thus, $f^*_\rho \geq \max\{f_\rho((\Sigma + \mathrm{Diag}(\rho))^{-1}), \theta\}$. Using this result and following a similar procedure as for deriving (7), we can show that

$$\log \lambda_{\max}(X^*_\rho) - \lambda_{\min}(\Sigma + \mathrm{Diag}(\rho))\lambda_{\max}(X^*_\rho) \;\geq\; \vartheta,$$

where ϑ is given in (15), and hence the statement $X^*_\rho \preceq \beta_\rho I$ immediately follows. ∎

In view of Proposition 2.5, we see that problem (5) is equivalent to the following problem

$$\max_{\alpha_\rho I \preceq X \preceq \beta_\rho I} \log\det X - \langle \Sigma, X\rangle - \sum_{i,j} \rho_{ij}|X_{ij}|, \tag{16}$$

where α_ρ and β_ρ are defined in Proposition 2.5.

We further observe that problem (16) can be rewritten as

$$\max_{X \in \mathcal{X}_\rho} \{f_\rho(X) := \min_{U \in \mathcal{U}} \phi(X, U)\}, \tag{17}$$

where $\mathcal{U}$ and $\phi(\cdot,\cdot)$ are defined in (13) and (14), respectively, and $\mathcal{X}_\rho$ is given as follows:

$$\mathcal{X}_\rho := \{X \in \mathcal{S}^n : \; \alpha_\rho I \preceq X \preceq \beta_\rho I\}. \tag{18}$$

Observing that $\phi(X, U) : \mathcal{X}_\rho \times \mathcal{U} \to \Re$ is a smooth function which is *strictly* concave in X for every fixed $U \in \mathcal{U}$, and convex in U for every fixed $X \in \mathcal{X}_\rho$, we can conclude that (i) problem (17) and its dual, that is,

$$\min_{U \in \mathcal{U}} \{g_\rho(U) := \max_{X \in \mathcal{X}_\rho} \phi(X, U)\} \tag{19}$$

are both solvable and have the same optimal value; and (ii) the function $g_\rho(\cdot)$ is convex differentiable and its gradient is given by

$$\nabla g_\rho(U) = \nabla_U \phi(X(U), U) \quad \forall U \in \mathcal{U},$$

where

$$X(U) := \arg\max_{X \in \mathcal{X}_\rho} \phi(X, U). \tag{20}$$

The following result shows that the approximate solution of problem (17) (or equivalently, (5)) can be obtained by solving smooth convex minimization problem (19) .

Proposition 2.6 *Let X_ρ^* be the unique optimal solution of problem (17), and let f_ρ^* be the optimal value of problems (17) and (19). Suppose that the sequence $\{U_k\}_{k=0}^\infty \subseteq \mathcal{U}$ is such that $g_\rho(U_k) \to f_\rho^*$ as $k \to \infty$. Then, $X(U_k) \to X_\rho^*$ and $g_\rho(U_k) - f_\rho(X(U_k)) \to 0$ as $k \to \infty$, where $X(\cdot)$ is defined in (20).*

Proof. The proof is similar to that of Theorem 2.4 of Lu [13]. ■

From Proposition 2.6, we see that problem (5) can be solved simultaneously while solving problem (19). Indeed, suppose that $\{U_k\}_{k=0}^\infty \subseteq \mathcal{U}$ is a sequence of approximate solutions obtained by solving (19). It follows from Proposition 2.6 that given any $\epsilon_o > 0$, there exists some iterate U_k such that $g_\rho(U_k) - f_\rho(X(U_k)) \le \epsilon_o$. It is clear that $X(U_k)$ is an ϵ_o-optimal solution of (17) and hence (5). We next propose two first order methods, namely, the adaptive spectral projected gradient method and the adaptive Nesterov's smooth method for problems (19) and (17) (or equivalently, (5)).

2.2.2 Adaptive spectral gradient projection method

In this subsection, we propose an adaptive spectral projected gradient (ASPG) method for solving problems (19) and (17) (or equivalently, (5)). We also discuss how to efficiently incorporate this method into the framework GSICS for solving problem (4).

The spectral gradient projection (SPG) methods were developed by Birgin et al. [4] for minimizing a smooth function over a closed convex set, which modify the classical projected gradient methods (see [2]) by incorporating the nonmonotone line search technique proposed by Grippo et al. [12] and the Barzilai-Borwein's gradient method [1]. We next discuss one of them (namely, the SPG2 method [4]) for solving the problem

$$\min \{g_{\rho,\beta}(U) : \ U \in \mathcal{U}\}, \tag{21}$$

and its dual

$$\max \{f_\rho(X) : \ \alpha_\rho I \preceq X \preceq \beta I\} \tag{22}$$

for some $\beta \ge \alpha_\rho$, where

$$g_{\rho,\beta}(U) := \max_{\alpha_\rho I \preceq X \preceq \beta I} \phi(X, U), \tag{23}$$

$\mathcal{U}$, $\phi(\cdot,\cdot)$, $f_\rho(\cdot)$ and α_ρ are defined in (13), (14), (17) and Proposition 2.5, respectively. Simply speaking, given a current iterate U_k, the SPG method first computes a projected gradient direction $d_k = P_{\mathcal{U}}(U_k - \alpha_k \nabla g_{\rho,\beta}(U_k))$, where α_k is obtained by the Barzilai-Borwein's spectral scheme [1] and $P_{\mathcal{U}}$ is a projection mapping defined below. The nonmonotone line search technique [12] is then applied to determine a step length λ so that the objective function $g_{\rho,\beta}$ sufficiently decreases at least every fixed number of iterations, and the next iterate is defined by $U_{k+1} := U_k + \lambda d_k$. This scheme is repeated until a suitable stopping criterion is satisfied.

We denote by $X_\beta(U)$ the unique optimal solution of problem (23). In view of (14), it is not hard to observe that $g_{\rho,\beta}(U)$ is differentiable, and moreover $X_\beta(U)$ and $\nabla g_{\rho,\beta}(U)$ have closed-form expressions for any U (see (30) of [13]). In addition, since $\mathcal{U}$ is a simple set, the projection of a point to $\mathcal{U}$ can be cheaply computed. Thus the SPG method [4] can be suitably applied to solve problem (21). For ease of subsequent presentation, we now describe the SPG method [4] for (21) in detail. The following notation will be used throughout this subsection.

Given a sequence $\{U_k\}_{k=0}^\infty \subseteq \mathcal{U}$ and an integer $M \geq 1$, we define

$$g_k^M := \max\ \{g_{\rho,\beta}(U_{k-j}) :\ 0 \leq j \leq \min\{k, M-1\}\}.$$

Also, let $P_{\mathcal{U}} : \Re^{n\times n} \to \mathcal{U}$ be defined as

$$P_{\mathcal{U}}(U) := \arg\min\{\|\hat{U} - U\|_F :\ \hat{U} \in \mathcal{U}\} \quad \forall U \in \Re^{n\times n}.$$

The SPG method for problems (21) and (22):

Let $\epsilon_o > 0$, $\gamma \in (0,1)$, $0 < \sigma_1 < \sigma_2 < 1$ and $0 < \alpha_{\min} < \alpha_{\max} < \infty$ be given. Let $M \geq 1$ be an integer. Choose $U_0 \in \mathcal{U}$, $\alpha_0 \in [\alpha_{\min}, \alpha_{\max}]$ and set $k = 0$.

```
1) If $g_{\rho,\beta}(U_k) - f_\rho(X_\beta(U_k)) \leq \epsilon_o$, terminate.

2) Compute $d_k = P_{\mathcal{U}}(U_k - \alpha_k \nabla g_{\rho,\beta}(U_k)) - U_k$. Set $\lambda \leftarrow 1$.

   2a) Set $U_+ = U_k + \lambda d_k$.
   2b) If $g_{\rho,\beta}(U_+) \leq g_k^M + \gamma\lambda\langle d_k, \nabla g_{\rho,\beta}(U_k)\rangle$, set $U_{k+1} = U_+$, $s_k = U_{k+1} - U_k$, $y_k = \nabla g_{\rho,\beta}(U_{k+1}) - \nabla g_{\rho,\beta}(U_k)$. Otherwise, choose $\lambda_+ \in [\sigma_1\lambda, \sigma_2\lambda]$, set $\lambda \leftarrow \lambda_+$ and go to step 2a).
   2c) Compute $b_k = \langle s_k, y_k\rangle$. If $b_k \leq 0$, set $\alpha_{k+1} = \alpha_{\max}$. Otherwise, compute $a_k = \langle s_k, s_k\rangle$ and set $\alpha_{k+1} = \min\ \{\alpha_{\max}, \max\{\alpha_{\min}, a_k/b_k\}\}$.

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

end

Remark. In step 2b) of the above SPG method, λ_+ is usually chosen by a safeguarded quadratic interpolation scheme that is commonly used to determine a step length for line search methods (see, for example, [5]). ■

We next establish a convergence result for the SPG method when applied to solve problems (21) and (22).

Theorem 2.7 *Let $\epsilon_o > 0$ be given. The SPG method generates a pair of ϵ_o-optimal solutions $(U_k, X_\beta(U_k))$ to problems (21) and (22) in a finite number of iterations.*

Proof. Suppose by contradiction that the SPG method does not terminate. Then it generates a sequence $\{U_k\}_{k=0}^\infty \subseteq \mathcal{U}$ satisfying $g_{\rho,\beta}(U_k) - f_\rho(X_\beta(U_k)) > \epsilon_o$. Note that $g_{\rho,\beta}(\cdot)$ is convex, which together with Theorem 2.4 of [4] implies that any accumulation point of $\{U_k\}_{k=0}^\infty$ is an optimal solution of problem (21). By the continuity of $g_{\rho,\beta}(\cdot)$, it further implies that any accumulation point of $\{g_{\rho,\beta}(U_k)\}_{k=0}^\infty$ is the optimal value f_ρ^* of (21). Using this observation and the fact that $\{g_{\rho,\beta}(U_k)\}_{k=0}^\infty$ is bounded, we conclude that $g_{\rho,\beta}(U_k) \to f_\rho^*$ as $k \to \infty$. Further, in view of Proposition 2.6 by replacing β_ρ with β, and $g_\rho(\cdot)$ with $g_{\rho,\beta}(\cdot)$, we have $g_{\rho,\beta}(U_k) - f_\rho(X_\beta(U_k)) \to 0$ as $k \to \infty$, and arrive at a contradiction. Therefore, the conclusion of this theorem holds. ■

Based on the above discussion, we see that the SPG method can be directly applied to find a pair of ϵ_o-optimal solutions to problems (19) and (17) (or equivalently, (5)) by setting $\beta = \beta_\rho$, where β_ρ is given in Proposition 2.5. It may converge, however, very slowly when β_ρ is large. Indeed, similarly as in [13], one can show that $\nabla g_{\rho,\beta}(U)$ is Lipschitz continuous on $\mathcal{U}$ with constant $L = \beta^2(\max_{i,j} \rho_{ij})^2$ with respect to the Frobenius norm. Let α_k, b_k and d_k be defined as above. Since $g_{\rho,\beta}(\cdot)$ is convex, we have $b_k \geq 0$. Moreover, we observed that it is almost always positive. In addition, $\alpha_{\min}$ and $\alpha_{\max}$ are usually chosen to be small (e.g., 10^{-15}) and large (e.g., 10^{15}), respectively (see [4]). Thus for the SPG method, we generally have

$$\alpha_{k+1} = \frac{\|U_{k+1} - U_k\|_F^2}{\langle U_{k+1} - U_k, \nabla g_{\rho,\beta}(U_{k+1}) - \nabla g_{\rho,\beta}(U_k)\rangle} \;\geq\; \frac{1}{L} \;=\; \frac{1}{\beta^2(\max\limits_{i,j} \rho_{ij})^2}.$$

Recall that β_ρ is an upper bound of $\lambda_{\max}(X_\rho^*)$, and typically it is overly large, where X_ρ^* is the optimal solution of (5). When $\beta = \beta_\rho$, we see from above that α_k can be very small, and so is $U_{k+1} - U_k$ due to

$$\|U_{k+1} - U_k\|_F \;\leq\; \|d_k\|_F \;=\; \|P_{\mathcal{U}}(U_k - \alpha_k \nabla g_{\rho,\beta}(U_k)) - U_k\|_F \;\leq\; \alpha_k \|\nabla g_{\rho,\beta}(U_k)\|_F.$$

Therefore, the SPG method may converge very slowly when applied to problem (19) directly.

To alleviate the aforementioned computational difficulty, we next propose an adaptive SPG (ASPG) method for problems (19) and (17) (or equivalently, (5)) by solving a sequence of problems (21) with $\beta = \beta_0, \beta_1, \ldots, \beta_m$ for some $\{\beta_k\}_{k=0}^m$ approaching $\lambda_{\max}(X_\rho^*)$ monotonically from below.

The adaptive SPG (ASPG) method for problems (17) and (19):

Let $\epsilon_o > 0$, $\beta_0 \ll \beta_\rho$ and $r_\beta > 1$ be given. Choose $U_0 \in \mathcal{U}$ and set $k = 0$.

1) Set $\beta \leftarrow \beta_k$. Apply the SPG method to find a pair of ϵ_o-optimal solutions $(\hat{U}_k, X_\beta(\hat{U}_k))$ to problems (21) and (22) starting from U_0.

2) If $\beta = \beta_\rho$ or $\lambda_{\max}(X_\beta(\hat{U}_k)) < \beta$, terminate.

3) Set $U_0 \leftarrow \hat{U}_k$, $\beta_{k+1} = \min\{\beta r_\beta, \beta_\rho\}$, $k \leftarrow k+1$, and go to step 1).

end

We now establish a convergence result for the ASPG method for solving problems (19) and (17) (or equivalently, (5)).

Theorem 2.8 *Let $\epsilon_o > 0$ be given. The ASPG method generates a pair of ϵ_o-optimal solutions to problems (19) and (17) (or equivalently, (5)) in a finite number of total (inner) iterations.*

Proof. First, we clearly see that β is updated only for a finite number of times. Using this observation and Theorem 2.7, we conclude that the ASPG method terminates in a finite number of total (inner) iterations. Suppose that it terminates at $\beta = \beta_k$ for some k. We now claim that $(\hat{U}_k, X_\beta(\hat{U}_k))$ is a pair of ϵ_o-optimal solutions to problems (19) and (17). Indeed, we clearly have $\beta = \beta_\rho$ or $\lambda_{\max}(X_\beta(\hat{U}_k)) < \beta$, which together with the definition of $g_\rho(\cdot)$ and $g_{\rho,\beta}(\cdot)$ (see (19) and (21)), implies that $g_\rho(\hat{U}_k) = g_{\rho,\beta}(\hat{U}_k)$. Thus, we obtain that

$$g_\rho(\hat{U}_k) - f_\rho(X_\beta(\hat{U}_k)) \ = \ g_{\rho,\beta}(\hat{U}_k) - f_\rho(X_\beta(\hat{U}_k)) \ \leq \ \epsilon_o,$$

which along with the fact $X_\beta(\hat{U}_k) \in \mathcal{X}_\rho$, implies that $(\hat{U}_k, X_\beta(\hat{U}_k))$ is a pair of ϵ_o-optimal solutions to problems (19) and (17). ■

As discussed above, the ASPG method is able to find a pair of ϵ_o-optimal solutions to problems (5) and (19). We next discuss how to efficiently incorporate it into the framework GSICS for finding an (ϵ_o, ϵ_c)-optimal solution to problem (4).

Recall from the framework GSICS (see Subsection 2.1) that in order to obtain an (ϵ_o, ϵ_c)-optimal solution to problem (4), we need to find an ϵ_o-optimal solution of problem (5) for a sequence of penalty parameters $\{\rho^k\}_{k=1}^m$, which satisfy for $k = 1, \ldots, m$, $\rho^k_{ij} = \rho_{ij}$ $\forall (i,j) \notin \Omega$ and $\rho^k_{ij} = \rho^0_{ij} r_\rho^{k-1}$ $\forall (i,j) \in \Omega$ for some $r_\rho > 1$ and $\rho^0_{ij} > 0$ $\forall (i,j) \in \Omega$. Suppose that a pair of ϵ_o-optimal solutions $(X_{\hat{\beta}_k}(\hat{U}_k)), \hat{U}_k)$ of problems (5) and (19) with $\rho = \rho^k$ are already found by the ASPG method for some $\hat{\beta}_k \in [\alpha_{\rho^k}, \beta_{\rho^k}]$. Then, we choose the initial U_0 and β_0 for the ASPG method when applied to solve problems (5) and (19) with $\rho = \rho^{k+1}$ as follows:

$$(U_0)_{ij} = \begin{cases} (\hat{U}_k)_{ij}/r_\rho, & \text{if } (i,j) \in \Omega; \\ (\hat{U}_k)_{ij}, & \text{otherwise.} \end{cases}, \qquad \beta_0 = \lambda_{\max}(X_{\hat{\beta}_k}(\hat{U}_k)). \tag{24}$$

We now provide some interpretation on such a choice of U_0 and β_0. First, we claim that $U_0 \in \mathcal{U}$ and $\beta_0 \in [\alpha_{\rho^{k+1}}, \beta_{\rho^{k+1}}]$. Indeed, since $\hat{U}_k \in \mathcal{U}$ and $r_\rho > 1$, we easily see that $U_0 \in \mathcal{U}$. In addition, using the definition of β_ρ (see Proposition 2.5) and the fact that $\text{Diag}(\rho^{k+1}) = \text{Diag}(\rho^k)$, we observe that $\beta_{\rho^{k+1}} = \beta_{\rho^k}$. Also, notice that $\rho^{k+1}_{ij} \geq \rho^k_{ij} \geq 0$ for all (i,j). It follows that $\|\rho^{k+1}\| \geq \|\rho^k\|$, which together with the definition of α_ρ (see Proposition 2.5) implies that $\alpha_{\rho^{k+1}} \leq \alpha_{\rho^k}$. By the definition of $X_{\hat{\beta}_k}(\hat{U}_k)$, we also know that

$$\alpha_{\rho^k} \ \leq \ \lambda_{\max}(X_{\hat{\beta}_k}(\hat{U}_k) \ \leq \ \hat{\beta}_k \ \leq \ \beta_{\rho^k}.$$

Hence, $\beta_0 \in [\alpha_{\rho^{k+1}}, \beta_{\rho^{k+1}}]$. Let f^*_ρ denote the optimal value of problem (5) for any given ρ. Clearly, we know from the ASPG method that either $\lambda_{\max}(X_{\hat{\beta}_k}(\hat{U}_k)) < \hat{\beta}_k < \beta_{\rho^k}$ or $\lambda_{\max}(X_{\hat{\beta}_k}(\hat{U}_k)) \leq \hat{\beta}_k = \beta_{\rho^k}$ holds, which together with (19) and (23) implies that

$$g_{\rho^k,\hat{\beta}_k}(\hat{U}_k) = g_{\rho^k}(\hat{U}_k) \in [f^*_{\rho^k}, f^*_{\rho^k} + \epsilon_o]. \tag{25}$$

Usually, $\alpha_{\rho^{k+1}} \approx \alpha_{\rho^k} \approx 0$. Using these relations along with (25), (19) and (23), we further observe that

$$\begin{array}{ccccccc} f^*_{\rho^{k+1}} & \leq & g_{\rho^{k+1}}(U_0) & \approx & g_{\rho^k}(\hat{U}_k) & \leq & f^*_{\rho^k} + \epsilon_o, \\ g_{\rho^{k+1},\beta_0}(U_0) & \approx & g_{\rho^k,\beta_0}(\hat{U}_k) & = & g_{\rho^k}(\hat{U}_k). & & \end{array}$$

It follows that when $f^*_{\rho^{k+1}}$ is close to $f^*_{\rho^k}$, U_0 is nearly an ϵ_o-optimal solution for problems (19) and (21) with $\rho = \rho^{k+1}$ and $\beta = \beta_0$. Therefore, we expect that for the above choice of U_0 and β_0, the ASPG method can solve problems (5) and (19) with $\rho = \rho^{k+1}$ rapidly when $f^*_{\rho^{k+1}}$ is close to $f^*_{\rho^k}$.

2.2.3 Adaptive Nesterov's smooth method

In this subsection, we propose an adaptive Nesterov's smooth (ANS) method for solving problems (19) and (17) (or equivalently, (5)). We also briefly mention how to efficiently incorporate this method into the framework GSICS for solving problem (4).

Recently, Lu [13] studied Nesterov's smooth method [14, 15] for solving a special class of problems (19) and (17), where ρ is a positive multiple of ee^T. He showed that an ϵ_o-optimal solution to problems (19) and (17) can be found in at most $\sqrt{2}\beta_\rho(\max_{i,j} \rho_{ij}) \max_{U \in \mathcal{U}} \|U - U_0\|_F / \sqrt{\epsilon_o}$ iterations by Nesterov's smooth method for some initial point $U_0 \in \mathcal{U}$ (see pp. 12 of [13] for details). Given that β_ρ is an estimate and typically an overestimate of $\lambda_{\max}(X^*_\rho)$, where X^*_ρ is the unique optimal solution of problem (5), the above iteration complexity can be exceedingly large and Nesterov's smooth method generally converges extremely slowly. Lu [13] further proposed an adaptive Nesterov's smooth (ANS) method for solving problems (19) and (17) (see pp. 15 of [13]). In his method, $\lambda_{\max}(X^*_\rho)$ is estimated from $\lambda_{\max}(X(U_k))$, and its estimation is adaptively adjusted based on the change of $\lambda_{\max}(X(U_k))$ as the algorithm progresses, where U_k is an approximate solution of problem (19). As a result, his method can gradually provide a tight estimate of $\lambda_{\max}(X^*_\rho)$ and it has an asymptotically optimal iteration complexity.

We now extend the ANS method [13] to problems (19) and (17) for a general ρ. Recall from Subsection 2.2.2 that $\nabla g_\rho(U)$ is Lipschitz continuous on $\mathcal{U}$ with constant $L = \beta^2_\rho(\max_{i,j} \rho_{ij})^2$ with respect to the Frobenius norm. It is then straightforward to extend the ANS method [13] to problems (19) and (5) for a general ρ by replacing the corresponding Lipschitz constants by the ones computed from the above formula. For ease of reference, we provide the details of the ANS method for problems (19) and (17) below.

Throughout the remainder of this section, we assume that α_ρ, β_ρ, $g_{\rho,\beta}(\cdot)$ and $X_\beta(\cdot)$ are given in Proposition 2.5 and Subsection 2.2.2, respectively. We now introduce a definition that will be used subsequently.

Definition 2 *Given any $U \in \mathcal{U}$ and $\beta \in [\alpha_\rho, \beta_\rho]$, $X_\beta(U)$ is called "active" if $\lambda_{\max}(X_\beta(U)) = \beta$ and $\beta < \beta_\rho$; otherwise it is called "inactive".*

We are now ready to present the ANS method [13] for problems (19) and (17) (or equivalently, (5)).

The ANS method for problems (17) and (19)

```
Let $\epsilon > 0$, $\varsigma_1, \varsigma_2 > 1$, and let $\varsigma_3 \in (0,1)$ be given. Let $\rho_{\max} = \max_{i,j} \rho_{ij}$. Choose $U_0 \in \mathcal{U}$ and
$\beta \in [\alpha_\rho, \beta_\rho]$. Set $L = \beta^2 \rho_{\max}^2$, $\sigma = 1$, and $k = 0$.

  1) Compute $X_\beta(U_k)$.

      1a) If $X_\beta(U_k)$ is active, find the smallest $s \in \mathcal{Z}_+$ such that $X_{\bar\beta}(U_k)$ is inactive, where
          $\bar\beta = \min\{\varsigma_1^s \beta, \beta_\rho\}$. Set $k = 0$, $U_0 = U_k$, $\beta = \bar\beta$, $L = \beta^2 \rho_{\max}^2$ and go to step 2).
      1b) If $X_\beta(U_k)$ is inactive and $\lambda_{\max}(X_\beta(U_k)) \le \varsigma_3 \beta$, set $k = 0$, $U_0 = U_k$,
          $\beta = \max\{\min\{\varsigma_2 \lambda_{\max}(X_\beta(U_k)), \beta_\rho\}, \alpha_\rho\}$, and $L = \beta^2 \rho_{\max}^2$.

  2) If $g_{\rho,\beta}(U_k) - f_\rho(X_\beta(U_k)) \le \epsilon$, terminate. Otherwise, compute $\nabla g_{\rho,\beta}(U_k)$.

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

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

  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
```

Similarly as the ASPG method, we can easily incorporate the above ANS method into the framework GSICS to find an (ϵ_o, ϵ_c)-optimal solution to problem (4) by applying the same strategy for updating the initial U_0 and β_0 detailed at the end of Subsection 2.2.2. For convenience of subsequent reference, the resulting method is also called the ANS method, namely, the adaptive Nesterov's smooth method.

3 Computational results

In this section, we test sparsity recovery ability of the model (4) and compare the performance of the method glasso [10, 11] with two first-order methods, that is, the adaptive spectral projected gradient (ASPG) method and the adaptive Nesterov's smooth (ANS) method proposed in Section 2 for solving problem (4) on a set of randomly generated instances.

The codes for the ASPG and ANS methods are written in MATLAB while the main subroutines of glasso were written in Fortran 90 by Friedman et al. [11] running with R interface. For all experiments in this section, we run the codes of ASPG and ANS in MATLAB 7.8.0 (R2009a) and glasso in R 2.10.1 on a Dell Optiplex Gx280 PC with Intel core 2.40 GHz. We set $\gamma = 10^{-4}$, $M = 50$, $\sigma_1 = 0.1$, $\sigma_2 = 0.9$, $\alpha_{\min} = 10^{-15}$, $\alpha_{\max} = 10^{15}$ for the ASPG method, and set $\varsigma_1 = \varsigma_2 = 1.05$ and $\varsigma_3 = 0.95$ for the ANS method. In addition, we set $\beta_0 = 1$, $r_\beta = 10$, $r_\rho = 2$, and $\rho^0_{ij} = 0.5$ for all $(i,j) \in \Omega$ for these two methods. Moreover, the ASPG and ANS methods start from the initial point $U_0 = 0$ and terminate once an (ϵ_o, ϵ_c)-optimal solution of problem (4) is found, where $\epsilon_o = 0.1$ and $\epsilon_c = 10^{-4}$. For glasso, ρ_{ij} is set to 10^9 by default for all $(i,j) \in \Omega$, and its other parameters and termination criterion are also set by default.

All instances of problem (4) used in this section are randomly generated in a similar manner as described in d'Aspremont et al. [7] and Lu [13]. Indeed, we first generate a sparse matrix $A \in \mathcal{S}^n_{++}$, with a density prescribed by ϱ and set $\Omega = \{(i,j) : A_{ij} = 0, |i-j| \geq 2\}$. We then generate a matrix $B \in \mathcal{S}^n$ by

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

where $V \in \mathcal{S}^n$ contains pseudo-random values drawn from a uniform distribution on the interval $[-1, 1]$, 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 choose $\tau = 0.15$, $\vartheta = 1.0e-4$ for generating all instances.

In the first experiment we compare the performance of the ASPG and ANS methods that are proposed in Section 2 for problem (4). For all instances in this test, Ω and Σ are randomly generated by the above approach with $\varrho = 50\%$. In addition, for all $(i,j) \notin \Omega$, we set $\rho_{ij} = \rho_{\bar{\Omega}}$ for some $\rho_{\bar{\Omega}} > 0$. The performance of ASPG and ANS on these instances is presented in Table 1. The row size n of each sample covariance matrix Σ is given in column one. The size of the set Ω is given in column two. The CPU times (in seconds) of ANS and ASPG are given in the last six columns for $\rho_{\bar{\Omega}} = 0.5$, 0.05 and 0.005, respectively. From Table 1, we see that both methods are capable of solving all instances within a reasonable amount of time. Moreover, the ASPG method, namely, the adaptive spectral gradient method, generally outperforms the ANS method, that is, the adaptive Nesterov's smooth method.

As seen from above, the performance of ASPG is generally superior to ANS. We next compare the performance of the method glasso [10, 11] with the ASPG method for problem (4). All instances in this experiment are randomly generated by the above approach with $\varrho = 10\%$, 50%, 90% and 100%, respectively. It is clear that for $\varrho = 100\%$, Ω is an empty set. In addition, for all $(i,j) \notin \Omega$, we set $\rho_{ij} = \rho_{\bar{\Omega}}$ for some $\rho_{\bar{\Omega}} > 0$. The performance of ASPG and glasso on these instances is presented in Tables 2-5, respectively. In each table, the row size n of each sample covariance matrix Σ is given in column one. The size of the set Ω is given in column two. The CPU times (in seconds) of ASPG and glasso are given in the last six columns for $\rho_{\bar{\Omega}} = 0.5$, 0.05 and 0.005, respectively. We can observe that, for a given Σ and Ω, the performance of glasso declines as $\rho_{\bar{\Omega}}$ decreases, but the performance of ASPG

Table 1: Comparison of ANS and ASPG

Problem		$\rho_{\bar{\Omega}} = 0.5$		$\rho_{\bar{\Omega}} = 0.05$		$\rho_{\bar{\Omega}} = 0.005$	
n	size(Ω)	ANS	ASPG	ANS	ASPG	ANS	ASPG
100	4776	2.0	1.6	1.0	1.0	1.7	1.1
200	19438	17.3	9.8	4.3	3.2	8.3	4.2
300	44136	44.4	33.9	12.0	8.5	26.3	13.1
400	78738	99.4	65.2	27.9	20.0	59.5	30.4
500	123300	200.4	134.5	50.5	29.4	112.3	52.2
600	177614	362.1	291.6	78.9	44.6	191.0	121.3
700	241944	562.9	443.6	127.2	74.8	310.8	181.4
800	317184	798.4	628.7	189.7	113.6	487.7	258.4
900	400952	1343.1	1288.3	292.6	182.3	768.4	440.2
1000	494610	2397.5	2022.6	504.6	281.0	1192.5	618.2

Table 2: Comparison of ASPG and glasso for $\varrho = 10\%$

Problem		$\rho_{\bar{\Omega}} = 0.5$		$\rho_{\bar{\Omega}} = 0.05$		$\rho_{\bar{\Omega}} = 0.005$	
n	size(Ω)	ASPG	glasso	ASPG	glasso	ASPG	glasso
100	8792	12.4	0.2	5.7	0.4	5.6	0.5
200	35646	20.6	1.7	9.7	2.8	9.1	4.2
300	80604	104.0	4.6	28.1	8.5	27.3	14.3
400	143636	153.2	14.7	44.5	25.4	53.9	42.5
500	224788	267.1	25.8	103.3	50.7	122.3	92.4

stays reasonably stable. In addition, as the size of Ω (namely, the number of known pairs of conditionally independent nodes) decreases, or, equivalently, ϱ increases, the performance of glasso degrades while the performance of ASPG generally improves. We also see that, when $\rho_{\bar{\Omega}}$ is relatively large and ϱ is small (that is, the underlying graph is highly sparse and a large percentage of pairs of conditional independent nodes is known beforehand), glasso outperforms ASPG (see Table 2), but otherwise the performance of ASPG is generally superior to glasso (see Tables 3-5).

Our third experiment is similar to the one conducted in d'Aspremont et al. [7]. We intend to test sparsity recovery ability of the model (4). To this aim, we specialize $n = 30$ and the matrix $A \in \mathcal{S}^n_{++}$ to be the one with diagonal entries around one and a few randomly chosen, nonzero off-diagonal entries equal to $+1$ or -1. And the sample covariance matrix Σ is then generated by the aforementioned approach. In addition, we set $\Omega = \{(i,j) : \; A_{ij} = 0, |i-j| \geq 5\}$ and $\rho_{ij} = 0.1$ for all $(i,j) \notin \Omega$. For such an instance, the model (4) is solved by the ASPG method whose parameters, initial point and termination criterion are exactly the same as above. In Figure 1, we plot the sparsity patterns of the original inverse covariance matrix A, the approximate solution to problem (4) and the noisy inverse covariance matrix B^{-1} for such a randomly generated instance. We observe that the model (4) is capable of recovering the sparsity pattern of the original inverse covariance matrix.

Table 3: Comparison of ASPG and ANS for $\varrho = 50\%$

Problem		$\rho_{\bar{\Omega}} = 0.5$		$\rho_{\bar{\Omega}} = 0.05$		$\rho_{\bar{\Omega}} = 0.005$	
n	size(Ω)	ASPG	glasso	ASPG	glasso	ASPG	glasso
100	4776	1.6	2.4	1.0	5.6	1.1	17.3
200	19438	9.8	13.4	3.2	43.8	4.2	148.3
300	44136	33.9	63.4	8.5	186.3	13.1	779.4
400	78738	65.2	152.7	20.0	505.6	30.4	2288.3
500	123300	134.5	228.7	29.4	934.5	52.2	4321.1

Table 4: Comparison of ASPG and glasso for $\varrho = 90\%$

Problem		$\rho_{\bar{\Omega}} = 0.5$		$\rho_{\bar{\Omega}} = 0.05$		$\rho_{\bar{\Omega}} = 0.005$	
n	size(Ω)	ASPG	glasso	ASPG	glasso	ASPG	glasso
100	960	1.5	4.1	0.9	14.4	1.3	76.9
200	3738	8.2	34.7	3.1	127.9	7.6	765.0
300	8750	28.5	75.4	7.6	499.2	19.5	3041.8
400	15764	54.2	210.7	11.5	1413.4	37.5	8651.8
500	25072	141.0	442.8	22.8	3203.1	63.4	20178.6

4 Concluding remarks

In this paper we considered estimating sparse inverse covariance of a Gaussian graphical model whose conditional independence is assumed to be partially known. Naturally, we formulated it as a constrained l_1-norm penalized maximum likelihood estimation problem. Further, we proposed an algorithm framework, and developed two first-order methods, that is, the adaptive spectral projected gradient (ASPG) method and the adaptive Nesterov's smooth (ANS) method for solving it. Our computational results demonstrate that both methods are capable of solving problems of size at least a thousand and number of constraints of nearly a half million within a reasonable amount of time, and moreover the ASPG method generally outperforms the ANS method and glasso [10, 11].

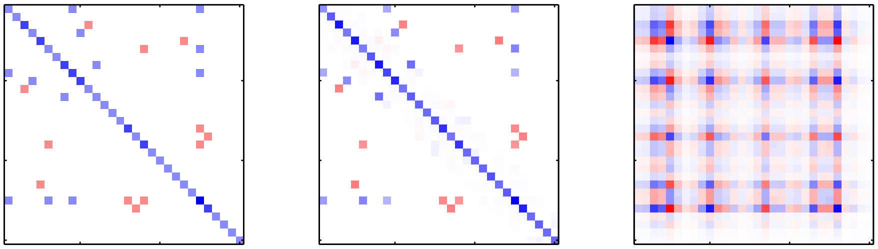
The source codes for the ASPG and ANS methods (written in MATLAB) are available online at www.math.sfu.ca/∼zhaosong. They can also be applied to problem (4) with $\Omega = \emptyset$, namely, the case where the underlying sparsity structure is completely unknown. It shall be mentioned that these codes can be extended straightforwardly to more general problems of the form

$$\begin{array}{ll} \max\limits_{X} & \log\det X - \langle \Sigma, X\rangle - \sum\limits_{(ij)\notin\Omega} \rho_{ij}|X_{ij}| \\ \text{s.t.} & \alpha I \preceq X \preceq \beta I, \\ & X_{ij} = 0 \quad \forall (i,j) \in \Omega, \end{array}$$

where $0 \le \alpha < \beta \le \infty$ are some fixed bounds on the eigenvalues of the solution.

Table 5: Comparison of ASPG and glasso for $\varrho = 100\%$

Problem		$\rho_{\bar\Omega} = 0.5$		$\rho_{\bar\Omega} = 0.05$		$\rho_{\bar\Omega} = 0.005$	
n	size(Ω)	ASPG	glasso	ASPG	glasso	ASPG	glasso
100	0	0.9	3.6	0.2	17.0	0.1	75.1
200	0	4.8	39.5	0.7	166.6	0.5	964.1
300	0	20.2	119.3	2.1	639.8	1.4	3964.0
400	0	50.8	257.7	4.7	1601.3	2.8	11406.0
500	0	92.6	415.6	7.8	3388.5	5.5	22779.6



Original inverse A

Approximate solution of (4)

Noisy inverse B^{-1}

Figure 1: Sparsity recovery.

Acknowledgements

The author is in debt to two anonymous referees for insightful comments and suggestions, which have greatly improved the paper. We also thank David Alexander Campbell for kind help in installing R package.

References

[1] J. Barzilai and J. M. Borwein, *Two point step size gradient methods*, IMA J. Numer. Anal., 8 (1988), pp. 141–148.

[2] D. P. Bertsekas, *Nonlinear Programming*, 2nd edition, Athena Scientific, Belmont, Massachusetts, 1999.

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

[4] E. G. Birgin, J. M. Martínez, and M. Raydan, *Nonmonotone spectral projected gradient methods on convex sets*, SIAM J. Optim., 10 (2000), pp. 1196–1211.

[5] E. G. Birgin, J. M. Martínez, and M. Raydan, *Spectral projected gradient methods*, Encyclopedia of Optimization (second ed.), Springer (2009), pp. 3652–3659.

[6] J. Dahl, L. Vandenberghe, and V. Roychowdhury, *Covariance selection for nonchordal graphs via chordal embedding*, Optim. Methods Softw., 23 (2008), pp. 501–520.

[7] A. d'Aspremont, O. Banerjee, and L. El Ghaoui, *First-order methods for sparse covariance selection*, SIAM J. Matrix Anal. Appl., 30 (2008), pp. 56–66.

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

[9] 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.

[10] J. Friedman, T. Hastie, and R. Tibshirani, *Sparse inverse covariance estimation with the graphical lasso*, Biostatistics, 9 (2008), pp. 432–441.

[11] J. Friedman, T. Hastie, and R. Tibshirani. glasso: Graphical lasso for R, November 2007.

[12] L. Grippo, F. Lampariello, and S. Lucidi, *A nonmonotone line search technique for Newton's method*, SIAM J. Numer. Anal., 23 (1986), pp. 707–716.

[13] Z. Lu, *Smooth optimization approach for sparse covariance selection*, SIAM J. Optim., 19 (2009), pp. 1807–1827.

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

[15] Y. E. Nesterov, *Smooth minimization of nonsmooth functions*, Math. Programming, 103 (2005), pp. 127–152.

[16] Y. E. Nesterov and A. S. Nemirovski, *Interior point Polynomial algorithms in Convex Programming: Theory and Applications*, SIAM, Philadelphia, 1994.

[17] L. Vandenberghe, S. Boyd, and S. Wu, *Determinant maximization with linear matrix inequality constraints*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 499–533.

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