A Block Coordinate Gradient Descent Method for Regularized Convex Separable Optimization and Covariance Selection

Sangwoon Yun * Paul Tseng † Kim-Chuan Toh ‡

March 29, 2010

In honor of Professor Paul Tseng, who went missing while on a kayak trip in Jinsha river, China, on August 13, 2009, for his contributions to the theory and algorithms for large-scale optimization

Abstract

We consider a class of unconstrained nonsmooth convex optimization problems, in which the objective function is the sum of a convex smooth function on an open subset of matrices and a separable convex function on a set of matrices. This problem includes the covariance selection estimation problem that can be expressed as an ℓ_1-penalized maximum likelihood estimation problem. In this paper, we propose a block coordinate gradient descent method (abbreviated as BCGD) for solving this class of nonsmooth separable problems with the coordinate block chosen by a Gauss-Seidel rule. The method is simple, highly parallelizable, and suited for large-scale problems. We establish global convergence and, under a local Lipschizian error bound assumption, linear rate of convergence for this method. For the covariance selection estimation problem, the method can terminate in $O(n^3/\epsilon)$ iterations with an ϵ-optimal solution. We compare the performance of the BCGD method with first-order methods studied in [11, 12] for solving the covariance selection problem on randomly generated instances. Our numerical experience suggests that the BCGD method can be efficient for large-scale covariance selection problems with constraints.

Key words. Block coordinate gradient descent, complexity, convex optimization, covariance selection, global convergence, linear rate convergence, ℓ_1-penalization, maximum likelihood estimation

*Korea Institute for Advanced Study, 207-43 Cheongnyangni 2-dong, Dongdaemun-gu, Seoul 130-722, Korea(yswcs@kias.re.kr).

†Department of Mathematics, University of Washington, Seattle, WA 98195, U.S.A. (tseng@math.washington.edu).

‡Department of Mathematics, National University of Singapore, 2 Science Drive 2, Singapore 117543 (mattohkc@nus.edu.sg); and Singapore-MIT Alliance, 4 Engineering Drive 3, Singapore 117576.

1 Introduction

It is well-known that undirected graphical models offer a way to describe and explain the relationships among a set of variables, a central element of multivariate data analysis. Given n variables drawn from a Gaussian distribution $\mathcal{N}(0, C)$ for which the true covariance matrix $C \succ 0_n$ is unknown, we wish to estimate C from a sample covariance matrix $S \succeq 0_n$ by maximizing its log-likelihood. Of particular interest is imposing a certain sparsity on the inverse covariance matrix C^{-1}. This is known as the covariance selection (estimation) problem. Many authors have studied a variety of approaches to estimate sparse inverse covariance matrix. d'Aspremont, Banerjee, and El Ghaoui [5] formulated this covariance selection problem as following:

$$\begin{array}{ll} \max\limits_{X \in \mathcal{S}^n} & \log\det X - \langle S, X\rangle - \rho \mathbf{Card}(X) \\ \text{s.t.} & \alpha I \preceq X \preceq \beta I, \end{array}$$

where $\mathbf{Card}(X)$ is the cardinality of X, i.e., the number of nonzero components in X, $\rho > 0$ is a given parameter controlling the trade-off between log-likelihood and cardinality, $0 \leq \alpha < \beta \leq \infty$ are bounds on the eigenvalues of the solution. The cardinality penalty term makes this problem a NP-hard combinatorial problem. Hence they used an argument that is often used in regression techniques, such as the Lasso [23], when sparsity of the solution is a concern, to relax $\mathbf{Card}(X)$ to $\sum_{i,j=1}^n |X_{ij}|$. Hence the above problem can be relaxed as the following problem:

$$\begin{array}{ll} \max\limits_{X \in \mathcal{S}^n} & \log\det X - \langle S, X\rangle - \rho \sum_{i,j=1}^n |X_{ij}| \\ \text{s.t.} & \alpha I \preceq X \preceq \beta I. \end{array} \tag{1}$$

They proposed a first order method developed in [15] for solving (1) and a block coordinate descent method, in which a solution of a box constrained quadratic program with $n-1$ variables is required at each iteration, for solving (1) and its dual problem when $\alpha = 0$ and $\beta = +\infty$:

$$\begin{array}{ll} \min\limits_{X \in \mathcal{S}^n} & -\log\det X - n \\ \text{s.t.} & |(X - S)_{ij}| \leq \rho_{ij},\ i, j = 1, ..., n, \end{array} \tag{2}$$

where $\rho_{ij} = \rho$ for all $i, j = 1, ..., n$. The box constrained quadratic program arises because only one column (and corresponding row) is updated at each iteration, and is solved by using either interior point method or the optimal first-order method in [14]. Yuan and Lin [29] proposed a covariance selection estimation problem (1) with $\alpha = 0$, $\beta = +\infty$, and the term $\sum_{i,j=1}^n |X_{ij}|$ replaced by $\sum_{i \neq j} |X_{ij}|$. They proposed the interior point method developed in [26]. Recently, Lu [11] proposed to solve (1) by applying a variant of Nesterov's smooth method [15] to the following generalization of (2):

$$\begin{array}{ll} \min\limits_{Z \in \mathcal{S}^n} & h^*(Z) \\ \text{s.t.} & |(Z - S)_{ij}| \leq \rho,\ i, j = 1, ..., n, \end{array} \tag{3}$$

where h^* denotes the conjugate function of $h(X) = \begin{cases} -\log\det X & \text{if } \alpha I \preceq X \preceq \beta I; \\ \infty & \text{else} \end{cases}$. Since h is strictly convex, h^* is essentially smooth [19, Section 26] with $\mathrm{dom} h^* = \mathcal{S}^n$ if $\beta < \infty$ and $\mathrm{dom} h^* = \mathcal{S}^n_{++}$ if $\beta = \infty$. Moreover, $h^*(Z)$ and $\nabla h^*(Z)$ can be easily calculated from the spectral decomposition of Z. The saddle function format for the problem (1) is different from the one considered in [5]. Friedman, Hastie, and Tibshirani [8] considered the dual problem (2) and proposed a block coordinate descent method, in which a solution of an ℓ_1-penalized linear least squares problem with $n-1$ variables, known as Lasso, is required at each iteration. This ℓ_1-penalized linear least squares problem is a dual problem of the box constrained quadratic program in [5] and is solved by using coordinate descent method proposed in [7]. Dahl, Vandenberghe, and Roychowdhury [4] studied the maximum likelihood estimation of a Gaussian graphical model whose conditional independence is known and it can be formulated as follows:

$$\begin{array}{rl} \max\limits_{X \in \mathcal{S}^n} & \log\det X - \langle S, X \rangle \\ \text{s.t.} & X_{ij} = 0 \ \forall (i,j) \in V, \end{array} \tag{4}$$

where V is a collection of all pairs of conditional independent nodes. We note that $(i,i) \notin V$ for $1 \le i \le n$ and $(i,j) \in V$ if and only if $(j,i) \in V$. They proposed Newton's method and the preconditioned conjugate gradient method by using efficient methods for evaluating the gradient and Hessian of the log-likelihood function when the underlying graph is nearly-chordal.

During the writing of this paper, Scheinberg and Rish [22] proposed a greedy algorithm based on a coordinate ascent method for solving (1) with $\alpha = 0$ and $\beta = +\infty$. Yuan [30] proposed an alternating direction method for solving (1). Lu [12] considered estimating sparse inverse covariance of a Gaussian graphical model whose conditional independence is assumed to be partially known in advance. This can be formulated as the following constrained ℓ_1-penalized maximum likelihood estimation problem:

$$\begin{array}{rl} \max\limits_{X \in \mathcal{S}^n} & \log\det X - \langle S, X \rangle - \sum_{(i,j) \notin V} \rho_{ij} |X_{ij}| \\ \text{s.t.} & X_{ij} = 0 \ \forall (i,j) \in V, \end{array} \tag{5}$$

where V is as in (4) and $\{\rho_{ij}\}_{(i,j) \notin V}$ is a set of nonnegative parameters. He proposed adaptive first-order (adaptive spectral projected gradient and adaptive Nesterov's smooth) methods for the problem (5). His methods consist of solving a sequence of the following unconstrained penalization problem:

$$\max_{X \in \mathcal{S}^n} \log\det X - \langle S, X \rangle - \sum_{i,j} \rho_{ij} |X_{ij}| \tag{6}$$

with a set of moderate penalty parameters $\{\rho_{ij}\}_{(i,j) \in V}$ that are adaptively adjusted until a desired approximate solution is found. The dual problem of (5) is given as follows:

$$\begin{array}{rl} \min\limits_{X \in \mathcal{S}^n} & -\log\det X - n \\ \text{s.t.} & |(X - S)_{ij}| \le \upsilon_{ij}, \ i,j = 1, ..., n, \end{array} \tag{7}$$

where $v_{ij} = \rho_{ij}$ for all $(i,j) \notin V$ and $v_{ij} = \infty$ for all $(i,j) \in V$. Recently, Wang, Sun, and Toh [27] proposed a Newton-CG primal proximal point algorithm (which employs the essential ideas of the proximal point algorithm, the Newton method, and the preconditioned conjugate gradient solver) for solving log-determinant optimization problems that include the problems (1) with $\alpha = 0$ and $\beta = +\infty$, (2), (4), (5), and (7).

In this paper, we consider a more general unconstrained nonsmooth convex optimization problem of the form:

$$\min_{X \in \Re^{m\times n}} \quad F(X) := f(X) + P(X), \tag{8}$$

where f is real-valued and convex smooth (i.e., continuously differentiable) on its domain $\mathrm{dom} f = \{X \in \Re^{m\times n} \mid f(X) < \infty\}$, which is assumed to be open, and $P : \Re^{m\times n} \to (-\infty, \infty]$ is a proper, convex, lower semicontinuous (lsc) [19] function. Of particular interest is when P is separable, i.e.,

$$P(X) = \sum_{i,j=1}^{n} P_{ij}(X_{ij}), \tag{9}$$

for some proper, convex, lsc functions $P_{ij} : \Re \to (-\infty, \infty]$. We assume that $\mathrm{dom} F \neq \emptyset$. More generally, P can be block-separable; see (28). The problem (1) with $\alpha = 0$ and $\beta = +\infty$ is a special case of (8) and (9) with

$$f(X) = -\log\det X + \langle S, X\rangle, \quad P_{ij}(X_{ij}) = \rho|X_{ij}|, \quad X \in \mathcal{S}^n. \tag{10}$$

The problem (2) is a special case of (8) and (9) with

$$f(X) = -\log\det X, \quad P_{ij}(X_{ij}) = \begin{cases} 0 & \text{if } L_{ij} \le (X-S)_{ij} \le U_{ij}; \\ \infty & \text{else,} \end{cases}, \quad X \in \mathcal{S}^n, \tag{11}$$

where $L_{ij} = -\rho$ and $U_{ij} = \rho$ for $i, j = 1, \ldots, n$. The problem (4) is a special case of (8) and (9) with

$$f(X) = -\log\det X + \langle S, X\rangle, \quad P_{ij} \equiv 0 \ \forall (i,j) \notin V, \quad P_{ij} \equiv \delta_{\{0\}} \ \forall (i,j) \in V, \quad X \in \mathcal{S}^n,$$

where δ_Q denotes the indicator function for a set Q. The problem (5) is a special case of (8) and (9) with

$$f(X) = -\log\det X + \langle S, X\rangle, \quad P_{ij} = \rho_{ij}|X_{ij}| \ \forall (i,j) \notin V, \quad P_{ij} \equiv \delta_{\{0\}} \ \forall (i,j) \in V, \quad X \in \mathcal{S}^n. \tag{12}$$

The problem (7) is a special case of (8) and (9) with

$$f(X) = -\log\det X, \quad P_{ij}(X_{ij}) = \begin{cases} 0 & \text{if } L_{ij} \le (X-S)_{ij} \le U_{ij}; \\ \infty & \text{else,} \end{cases}, \quad X \in \mathcal{S}^n, \tag{13}$$

where $L_{ij} = -\rho_{ij}$, $U_{ij} = \rho_{ij}$ for all $(i,j) \notin V$ and $L_{ij} = -\infty$, $U_{ij} = \infty$ for all $(i,j) \in V$. In all five examples, $\mathrm{dom} f = \mathcal{S}^n_{++}$. The problem (3) is a special case of (8) and (9) with $f \equiv h^*$, P_{ij} given by (11), $Z \in \mathcal{S}^n$.

Recently, Tseng and Yun [25] proposed a block coordinate gradient descent method for minimizing the sum of a smooth function and a (block) separable convex function on $\Re^n$. This method approximates the smooth function by a quadratic function at the current iterate, applies block coordinate descent to generate a feasible descent direction, and then updates the current iterate by performing an inexact line search along the descent direction. Numerical performances in [25, 31] suggest that the BCGD method can be effective in practice. We extend this method to solve (8) and, in particular, the covariance selection problems (1) with $\alpha = 0$ and $\beta = +\infty$, (5) with $\rho_{ij} > 0$ for $(i,j) \notin V$, and the dual problems (2), (7) with $\rho_{ij} > 0$ for $(i,j) \notin V$. As in [25], we choose the coordinate block according to a Gauss-Seidel rule and choose the stepsize according to an Armijo-like rule; see (17) and (19). We show that each cluster point of the iterates generated by this method is a stationary point of F; see Theorem 3.1. We next show that if a local Lipschizian error bound on the distance to the set of stationary points holds, then the iterates generated by the BCGD method converge at least linearly to a stationary point of F; see Theorem 3.2. We also give the general iteration complexity bound for the BCGD method; see Theorem 4.1. When the BCGD method is applied to solve (5), we do not need to solve a sequence of the unconstrained penalization problem (6) as did in [12]. This is the advantage of our method over the adaptive first-order methods in [12].

The paper is organized as follows. In Section 2, we describe the BCGD method and discuss how to choose the coordinate block that ensures the convergence. In Section 3, we establish the global convergence and the asymptotic convergence rate of the BCGD method. In Section 4, we analyze the complexity of the BCGD method. In Section 5, for the regularized negative log-likelihood problem, i.e., $f(X) = -\log\det(X) + \langle S, X\rangle$, we study the boundedness of the level set associated with the initial point and the Lipschitz continuity of ∇f on this level set. Also we analyze the complexity of the BCGD method when it is applied to solve the regularized negative log-likelihood problem. When specialized to the problem (1) with $\alpha = 0$ and $\beta = +\infty$, (2), (5) with $\rho_{ij} > 0$ for $(i,j) \notin V$, and (7) with $\rho_{ij} > 0$ for $(i,j) \notin V$, the resulting worst-case complexity bound for achieving ϵ-optimality is $O(n^5/\epsilon)$ operations. In Section 6, we describe an implementation of the BCGD method and report our numerical experience in solving large-scale covariance selection problems with or without constraints on randomly generated instances. Also, we compare our method with first-order methods studied in [11, 12]. In Section 7, we discuss conclusions and extensions.

In our notation, $\Re^n$ denotes the space of n-dimensional real column vectors, ^{T} denotes transpose. $\Re^{m\times n}$ denotes the space of $m \times n$ matrices. $\mathcal{S}^n$ denotes the space of $n \times n$ real symmetric matrices, $\mathcal{S}^n_{++}$ denotes the space of $n\times n$ real symmetric positive definite matrices. $\mathcal{N} = \{1,\ldots,n\}\times\{1,\ldots,n\}$. For any $m \times n$ real matrices X and Y, $\langle X, Y\rangle = \mathrm{trace}(XY^T)$ and $\|X\|_F = \sqrt{\langle X, X\rangle}$. For simplicity, we write $\|X\| = \|X\|_F$. For any $\mathcal{J} \subseteq \mathcal{N}$, $X_{\mathcal{J}}$ denotes the submatrix of X comprising X_{ij}, $(i,j) \in \mathcal{J}$ (we note that we can not use $X_{\mathcal{J}}$ for all subset $\mathcal{J}$ of $\mathcal{N}$) and $\Pi_{\mathcal{J}}(X)$ denotes the element of $\Re^{m\times n}$ obtained by setting the entries of X not indexed by $\mathcal{J}$ to zero. A linear mapping $\mathcal{H} : \Re^{m\times n} \to \Re^{m\times n}$ is self-adjoint if $\mathcal{H} = \mathcal{H}^*$, where $\mathcal{H}^*$ denotes the adjoint of H, and is positive definite if $\langle X, \mathcal{H}(X)\rangle > 0$ whenever

$X \neq 0$. The identity linear mapping is denoted by $\mathcal{I}$. We let $\lambda_{\min}(\mathcal{H}) = \min_{\|X\|=1}\langle X, \mathcal{H}(X)\rangle$, $\lambda_{\max}(\mathcal{H}) = \max_{\|X\|=1}\langle X, \mathcal{H}(X)\rangle$. The identity matrix is denoted by I and the matrix of zero entries is denoted by 0. For any $n \times n$ real symmetric matrices A and B, we write $A \succeq B$ (respectively, $A \succ B$) to mean that $A - B$ is positive semidefinite (respectively, positive definite). Unless otherwise specified, $\{X^k\}$ denotes the sequence $X^0, X^1, ...$ and, for any integer $\ell \geq 0$, $\{X^{k+\ell}\}_{\mathcal{K}}$ denotes a subsequence $\{X^{k+\ell}\}_{k\in\mathcal{K}}$ with $\mathcal{K} \subseteq \{0, 1, ...\}$.

2 Block Coordinate Gradient Descent Method

In this section, we describe the block coordinate gradient descent method for solving (8). In our method, we use the gradient ∇f to build a quadratic approximation of f and apply coordinate descent to generate an improving direction D at X. More precisely, we choose a nonempty index subset $\mathcal{J} \subseteq \mathcal{N}$ and a self-adjoint positive definite linear mapping $\mathcal{H} : \Re^{m\times n} \to \Re^{m\times n}$ (approximating the Hessian $\nabla^2 f(X)$), and move X along the direction $D = D_{\mathcal{H}}(X; \mathcal{J})$, where

$$D_{\mathcal{H}}(X;\mathcal{J}) := \arg\min_{D\in\Re^{m\times n}} \Big\{ \langle \nabla f(X), D\rangle + \frac{1}{2}\langle D, \mathcal{H}(D)\rangle + P(X+D) \;\mid\; D_{ij} = 0,\ \forall (i,j) \notin \mathcal{J} \Big\}. \tag{14}$$

By the separability of P, if $X \in \mathcal{S}^n$ and $\mathcal{H}(D) = (H_{ij}D_{ij})_{ij}$ where $H \in \mathcal{S}^n$ with $H_{ij} > 0$ for all $i,\ j \in \{1, ..., n\}$, then $(D_{\mathcal{H}}(X;\mathcal{N}))_{ij}$, the (i,j)th entry of $D_{\mathcal{H}}(X;\mathcal{N})$, is easily computable.

- If $P \equiv 0$, then $(D_{\mathcal{H}}(X;\mathcal{N}))_{ij} = -\frac{(\nabla f(X))_{ij}}{H_{ij}}$.
- If P is given by (9) with P_{ij} in (11), then $(D_{\mathcal{H}}(X;\mathcal{N}))_{ij} = \text{mid}\Big\{L_{ij} - X_{ij}, -\frac{(\nabla f(X))_{ij}}{H_{ij}}, U_{ij} - X_{ij}\Big\}$.
- If $P(X) = \sum_{i,j} \rho_{ij}|X_{ij}|$, then $(D_{\mathcal{H}}(X;\mathcal{N}))_{ij} = -\text{mid}\Big\{\frac{(\nabla f(X))_{ij}-\rho_{ij}}{H_{ij}}, X_{ij}, \frac{(\nabla f(X))_{ij}+\rho_{ij}}{H_{ij}}\Big\}$.

[mid$\{a, b, c\}$ denotes the median (mid-point) of a, b, c.] Hence (14) decomposes into subproblems that can be solved in parallel.

Using the convexity of P and a similar argument as in the proof of [25, Lemma 1], we have the following lemma showing that $D_{\mathcal{H}}(X;\mathcal{J})$ is a descent direction at X whenever $D_{\mathcal{H}}(X;\mathcal{J}) \neq 0$.

Lemma 2.1 *For any $X \in \text{dom}F$, nonempty $\mathcal{J} \subseteq \mathcal{N}$ and self-adjoint positive definite $\mathcal{H}$, let $D = D_{\mathcal{H}}(X;\mathcal{J})$. Then*

$$\langle \nabla f(X), D\rangle + P(X+D) - P(X) \leq -\langle D, \mathcal{H}(D)\rangle. \tag{15}$$

$$F(X+\alpha D) \leq F(X) + \alpha\Big(\langle \nabla f(X), D\rangle + P(X+D) - P(X)\Big) + o(\alpha) \quad \forall \alpha \in (0, \bar{\alpha}], \tag{16}$$

for any $\bar{\alpha} \in (0, 1]$ satisfying $X + \bar{\alpha}D \in \text{dom}f$.

We now describe formally the block coordinate gradient descent method.

BCGD method:

Choose $X^0 \in \mathrm{dom}F$. For $k = 0, 1, 2, ...$, generate X^{k+1} from X^k according to the iteration:

Step 1. Choose a nonempty $\mathcal{J}^k \subseteq \mathcal{N}$ and a self-adjoint positive definite $\mathcal{H}^k : \Re^{m\times n} \to \Re^{m\times n}$.

Step 2. Solve (14) with $X = X^k$, $\mathcal{J} = \mathcal{J}^k$ and $\mathcal{H} = \mathcal{H}^k$ to obtain $D^k = D_{\mathcal{H}^k}(X^k; \mathcal{J}^k)$.

Step 3. Choose a stepsize $\alpha^k > 0$ and set $X^{k+1} = X^k + \alpha^k D^k$.

For the stepsize rule, we use the following adaptation of the Armijo rule based on Lemma 2.1. This rule is simple and requires only function evaluation.

Armijo rule:

Choose $\alpha^k_{\text{init}} > 0$ such that $X^k + \alpha^k_{\text{init}} D^k \in \mathrm{dom}f$ and let α^k be the largest element of $\{\alpha^k_{\text{init}}\beta^j\}_{j=0,1,...}$ satisfying

$$F(X^k + \alpha^k D^k) \le F(X^k) + \alpha^k \sigma \Delta^k, \tag{17}$$

where $0 < \beta < 1$, $0 < \sigma < 1$, and

$$\Delta^k := \langle \nabla f(X^k), D^k \rangle + P(X^k + D^k) - P(X^k). \tag{18}$$

Here Δ^k may be interpreted as the predicted descent when moving from X^k to $X^k + D^k$. Since $\mathcal{H}^k$ is positive definite, we see from Lemma 2.1 that, for some $\bar{\alpha}^k \in (0, 1]$ satisfying $X^k + \bar{\alpha}^k D^k \in \mathrm{dom}f$,

$$F(X^k + \alpha^k D^k) \le F(X^k) + \alpha\Delta^k + o(\alpha) \quad \forall \alpha \in (0, \bar{\alpha}^k],$$

and $\Delta^k \le -\langle D^k, \mathcal{H}^k(D^k)\rangle < 0$ whenever $D^k \neq 0$. Since $0 < \sigma < 1$, this shows that α^k given by the Armijo rule is well defined and positive. And, by choosing α^k_{init} based on the previous stepsize α^{k-1}, the number of function evaluations can be kept small in practice.

For convergence of the BCGD method, the index subset $\mathcal{J}^k$ must be chosen judiciously. As in [25], global convergence can be ensured by choosing $\mathcal{J}^k$ in a *Gauss-Seidel* manner, i.e., $\mathcal{J}^0, \mathcal{J}^1, ...$ collectively cover $\mathcal{N}$ for every T consecutive iterations, where $T \ge 1$ (also see [24] and references therein), i.e.,

$$\mathcal{J}^k \cup \mathcal{J}^{k+1} \cup \cdots \cup \mathcal{J}^{k+T-1} = \mathcal{N}, \quad k = 0, 1, ... \tag{19}$$

For our convergence rate analysis, we need a more restrictive choice of $\mathcal{J}^k$, specifically, there exists a subsequence $\mathcal{T} \subseteq \{0, 1, ...\}$ such that

$$0 \in \mathcal{T}, \qquad \mathcal{N} = \Big(\text{disjoint union of } \mathcal{J}^k, \mathcal{J}^{k+1}, ..., \mathcal{J}^{\tau(k)-1}\Big) \quad \forall k \in \mathcal{T}, \tag{20}$$

where $\tau(k) := \min\{k' \in \mathcal{T} \mid k' > k\}$. For example, for $\mathcal{T} = \{0, n, 2n, 3n, \ldots\}$, we have $\tau(0) = n$.

When we apply the BCGD method to solve the covariance selection problems (1), (2), (4), (5), and (7), $f(X)$ has the form $-\log\det X + \langle S, X\rangle$ with $X \in \mathcal{S}^n$ and $S \succeq 0_n$. The specific choices of $\mathcal{J}^k$ are guided by the need to efficiently compute $\langle \nabla f(X^k), D^k\rangle$ and $F(X^k + \alpha^k D^k)$ in the Armijo stepsize rule. Hence we want to avoid computing $\det(X^k + \alpha^k D^k)$ and $\nabla f(X^k) = (X^k)^{-1}$ from scratch since it would require $O(n^3)$ operations. Following the proposal in [5], we will choose $\mathcal{J}^k = \{(i,j), (j,i) \mid i = 1, ..., n\}$ where $j = k + 1 (\text{mod } n)$ for the Gauss-Seidel rule (19) and $\mathcal{J}^k = \{(i,j), (j,i) \mid i = 1, ..., j\}$ where $j = k + 1 (\text{mod } n)$ for the restricted Gauss-Seidel rule (20). By these choices, we update only one column (and corresponding row) of X at each iteration. Then $\det(X^k + \alpha^k D^k)$ can be computed in $O(n^2)$ operations by using the Schur complement of $(X^k + \alpha^k D^k)_{/j/j}$, where $(X^k + \alpha^k D^k)_{/j/j}$ denotes the matrix produced by removing the jth column and row, in $X^k + \alpha^k D^k$. Similarly, $(X^{k+1})^{-1}$ can be updated in $O(n^2)$ operations from $(X^k)^{-1}$ using the Sherman-Woodbury-Morrison formula; see Section 6 for details.

3 Global Convergence and Asymptotic Convergence Rate

In this section we analyze the global convergence and asymptotic convergence rate of the BCGD method, analogous to those obtained in [25], when the (restricted) Gauss-Seidel rule is used.

We say that $X \in \Re^{m\times n}$ is *a stationary point* of F if $X \in \text{dom}F$ and $F'(X; D) \geq 0$ for all $D \in \Re^{m\times n}$. The following lemma gives an alternative characterization of stationarity. Its proof is nearly identical to [25, Lemma 2] and is omitted.

Lemma 3.1 *For any self-adjoint positive definite $\mathcal{H}$. An $X \in \text{dom}F$ is a stationary point of $F(X)$ if and only if $D_{\mathcal{H}}(X; \mathcal{N}) = 0$.*

Lemma 3.1 suggests that $\|D_{\mathcal{H}}(X; \mathcal{N})\|$ acts as a "residual" function, measuring how close X comes to being stationary for F. In what follows, we denote

$$\mathcal{X}^0 := \{X \mid F(X) \leq F(X^0)\}, \qquad \mathcal{X}^0_\varrho := \Big(\mathcal{X}^0 + \varrho B\Big) \cap \text{dom}P, \tag{21}$$

for some $\varrho > 0$, where $B := \{X \in \Re^{m\times n} \mid \|X\| \leq 1\}$ is the unit ball in $\Re^{m\times n}$. For our convergence rate analysis and complexity analysis (Theorems 3.2 and 4.1), we will make the following assumption on f.

Assumption 1 *There exist $L \geq 0$ and $\varrho > 0$ such that $\mathcal{X}^0_\varrho \subseteq \text{dom}f$ and*

$$\|\nabla f(Y) - \nabla f(Z)\| \leq L\|Y - Z\| \quad \forall Y, Z \in \mathcal{X}^0_\varrho. \tag{22}$$

Assumption 1 is satisfied by (3) with $L = \beta^2$ and $\varrho = \infty$ when h^* is replaced by f, P is given by (9) with P_{ij} as in (11), and $\beta < \infty$. This is because h is strongly convex with modulus $1/\beta^2$ so ∇h^* is Lipschitz continuous on $\mathcal{S}^n$ with constant β^2; see [9]. It is also satisfied by (10), (11), (12), (13) with $L = 1/((1-\omega)^2\underline{\zeta}^2)$, $\varrho = \omega\underline{\zeta}$ for $0 < \omega < 1$ when there exists a positive constant $\underline{\zeta}$ such that $\underline{\zeta} I \preceq X$ for all $X \in \mathcal{X}^0$; see Section 5.

Since $\Re^{m\times n}$ can be identified with $\Re^{mn}$, by using Assumption 2, we will extend the global convergence and the linear convergence rate analysis of the BCGD method in [25, Theorem 1 and 2] to the problem (8). However, unlike in [25], ∇f need not be Lipschitz continuous on dom$P$, and so some care is needed to choose the stepsize α so that $X + \alpha D_{\mathcal{H}}(X; \mathcal{J})$ is in $\mathcal{X}^0_\varrho \subseteq$ dom$f$. Next, we have a lemma concerning stepsizes satisfying the Armijo descent condition (17). Its proof is omitted since it is almost identical to that of [25, Lemma 5(b)].

Lemma 3.2 *Under Assumption 1, for any $X \in \mathcal{X}^0$, nonempty $\mathcal{J} \subseteq \mathcal{N}$, and self-adjoint positive definite $\mathcal{H}$ with $\lambda_{\min}(\mathcal{H}) \geq \underline{\lambda} > 0$, we have that $D = D_{\mathcal{H}}(X; \mathcal{J})$ satisfies the descent condition*

$$F(X + \alpha D) - F(X) \leq \sigma\alpha\Delta, \tag{23}$$

for any $\sigma \in (0, 1)$ and any $0 \leq \alpha \leq \min\{1, \varrho/\|D\|, 2\underline{\lambda}(1-\sigma)/L\}$, where $\Delta = \langle \nabla f(X), D\rangle + P(X + D) - P(X)$,

We say that P is block-separable with respect to nonempty $\mathcal{J}$ if

$$P(X) = P_{\mathcal{J}}(X_{\mathcal{J}}) + P_{\mathcal{J}^c}(X_{\mathcal{J}^c}) \quad \forall X \in \Re^{m\times n}$$

for some proper, convex, lsc function $P_{\mathcal{J}}$ and $P_{\mathcal{J}^c}$.

The next theorem establishes, under the following reasonable assumption on our choice of $\mathcal{H}^k$, the global convergence of the BCGD method using the generalized Guass-Seidel rule (19) to choose $\{\mathcal{J}^k\}$.

Assumption 2 *$\underline{\lambda} \leq \lambda_{\min}(\mathcal{H}^k)$ and $\lambda_{\max}(\mathcal{H}^k) \leq \bar{\lambda}$ for all k, where $0 < \underline{\lambda} \leq \bar{\lambda}$.*

The next assumption is a bound assumption on the level set $\mathcal{X}^0$. It will be needed to prove part (d) of the next theorem.

Assumption 3 *There exists a positive constant Λ such that $\|X\| \leq \Lambda$ for all $X \in \mathcal{X}^0$.*

Note that Assumption 3 is satisfied for the problems (10), (11), (12) with $\rho_{ii} > 0$ for $i = 1, ..., n$, and (13); see Section 5.

Theorem 3.1 *Let $\{X^k\}$, $\{\mathcal{H}^k\}$, $\{D^k\}$ be sequences generated by the BCGD method under Assumption 2, where $\{\alpha^k\}$ is chosen by the Armijo rule with $\inf_k \alpha^k_{\text{init}} > 0$. Then the following results hold.*

(a) $\{F(X^k)\}$ *is nonincreasing and* Δ^k *given by* (18) *satisfies*

$$-\Delta^k \geq \langle D^k, \mathcal{H}^k(D^k)\rangle \geq \underline{\lambda}\|D^k\|^2 \quad \forall k, \tag{24}$$

$$F(X^{k+1}) - F(X^k) \leq \sigma\alpha^k\Delta^k \leq 0 \quad \forall k. \tag{25}$$

(b) *If* $\{X^k\}_{\mathcal{K}}$ *is a convergent subsequence of* $\{X^k\}$*, then* $\{\alpha^k\Delta^k\} \to 0$ *and* $\{D^k\}_{\mathcal{K}} \to 0$*.*

(c) *If* $\{\mathcal{J}^k\}$ *is chosen by the Gauss-Seidel rule* (19)*,* P *is block-separable with respect to* $\mathcal{J}^k$ *for all* k*, and* $\sup_k \alpha^k < \infty$*, then every cluster point of* $\{X^k\}$ *is a stationary point of* F*.*

(d) *If Assumptions 1 and 3 are satisfied, then we have* $\inf_k \alpha^k > 0$*. Furthermore, if* $\lim_{k\to\infty} F(X^k) > -\infty$*, then* $\{\Delta^k\} \to 0$ *and* $\{D^k\} \to 0$*.*

Proof. (a), (b), and (c) can be proved by using nearly identical argument as used in the proof of [25, Theorem 1 (a), (b), (e)].

(d) Since α^k is chosen by the Armijo rule (17), either $\alpha^k = \alpha^k_{\text{init}}$ or else, by Lemma 3.2, $\alpha^k/\beta > \min\{1, \varrho/\|D^k\|, 2\underline{\lambda}(1-\sigma)/L\}$. Since P is convex, there is a constant M_p such that

$$\frac{\underline{\lambda}}{4}\|X\|^2 + P(X) \geq M_p, \quad \forall X \in \text{dom}P. \tag{26}$$

By (14),

$$\begin{aligned} 0 &\geq \langle \nabla f(X^k), D^k\rangle + \frac{1}{2}\langle D^k, \mathcal{H}^k(D^k)\rangle + P(X^k + D^k) - P(X^k) \\ &\geq \langle \nabla f(X^k), D^k\rangle + \frac{\underline{\lambda}}{2}\|D^k\|^2 - \frac{\underline{\lambda}}{4}\|X^k + D^k\|^2 - P(X^k) + M_p \\ &\geq \frac{\underline{\lambda}}{4}\|D^k\|^2 - (\|\nabla f(X^k)\| + \frac{\underline{\lambda}}{2}\|X^k\|)\|D^k\| - \frac{\underline{\lambda}}{4}\|X^k\|^2 + f(X^k) - F(X^0) + M_p, \end{aligned} \tag{27}$$

where the second inequality uses Assumption 2, (26) and the third inequality uses the Cauchy-Schwarz inequality and (25). By the continuity of ∇f and f, and Assumption 3, there are constants M_1, M_2, and M_3 such that $\|\nabla f(X^k)\| \leq M_1$, $\|X^k\| \leq M_2$, $f(X^k) \geq M_3$ for all $k \geq 0$. Then, by (27),

$$\frac{\underline{\lambda}}{4}\|D^k\|^2 - (M_1 + \frac{\underline{\lambda}}{2}M_2)\|D^k\| - \frac{\underline{\lambda}}{4}M_2^2 + M_3 + M_p - F(X^0) \leq 0$$

Hence there is a positive constant ϑ such that $\{\|D^k\|\} \leq \vartheta$. This together with $\inf_k \alpha^k_{\text{init}} > 0$ implies $\inf_k \alpha^k > 0$. If $\lim_{k\to\infty} F(X^k) > -\infty$ also, then this and (25) imply $\{\Delta^k\} \to 0$, which together with (24) imply $\{D^k\} \to 0$. ∎

Theorem 3.1 readily extends to any stepsize rule that yields a larger descent than the Armijo rule at each iteration.

Corollary 3.1 *Theorem 3.1 still holds if in the BCGD method the iterates are instead updated by* $X^{k+1} = X^k + \tilde{\alpha}^k D^k$, *where* $\tilde{\alpha}^k \geq 0$ *satisfies* $F(X^k + \tilde{\alpha}^k D^k) \leq F(X^k + \alpha^k D^k)$ *for* $k = 0, 1, \ldots$ *and* $\{\alpha^k\}$ *is chosen by the Armijo rule with* $\inf_k \alpha^k_{\rm init} > 0$.

Proof. It is readily seen using $F(X^{k+1}) \leq F(X^k + \alpha^k D^k)$ that Theorem 3.1(a) holds. The proofs of Theorem 3.1(b), (c) and (d) remain unchanged. ■

For example, $\tilde{\alpha}^k$ may be generated by the minimization rule:

$$\tilde{\alpha}^k \in \arg\min_{\alpha \geq 0} \{F(X^k + \alpha D^k) \mid X^k + \alpha D^k \in {\rm dom} F\}$$

or by the limited minimization rule:

$$\tilde{\alpha}^k \in \arg\min_{0 \leq \alpha \leq s} \{F(X^k + \alpha D^k) \mid X^k + \alpha D^k \in {\rm dom} F\},$$

where $0 < s < \infty$. The latter stepsize rule yields a larger descent than the Armijo rule with $\alpha^k_{\rm init} = s$. We will use the limited minimization rule in our numerical test on the covariance selection problem of the dual form (2) and (7); see Section 6.

The next theorem establishes the convergence rate of the BCGD method under Assumption 2 and the following assumption that is analogous to [25, Assumption 2 (a)]. (Since f and P are convex, the assumption that is analogous to [25, Assumption 2 (b)] is not required.) In what follows, $\bar{\mathcal{X}}$ denotes the set of stationary points of F, and

$$\mathrm{dist}(X, \bar{\mathcal{X}}) = \min_{\bar{X} \in \bar{\mathcal{X}}} \|X - \bar{X}\| \qquad \forall X \in \Re^{m \times n}.$$

Assumption 4 $\bar{\mathcal{X}} \neq \emptyset$ *and, for any* $\xi \geq \min_X F(X)$, *there exist scalars* $\tau > 0$ *and* $\bar{\epsilon} > 0$ *such that*

$$\mathrm{dist}(X, \bar{\mathcal{X}}) \leq \tau \|D_{\mathcal{I}}(X; \mathcal{N})\| \quad \textit{whenever } F(X) \leq \xi,\ \|D_{\mathcal{I}}(X; \mathcal{N})\| \leq \bar{\epsilon}.$$

Assumption 4 is a local Lipschitzian error bound assumption, saying that the distance from X to $\bar{\mathcal{X}}$ is locally in the order of the norm of the residual at X; see [13, 25] and references therein. By applying similar argument used in the proof of [25, Theorem 4] to the problem (8), we obtain the following sufficient conditions for Assumption 4 to hold.

Proposition 3.1 *Suppose that* $\bar{\mathcal{X}} \neq \emptyset$ *and any of the following conditions hold.*

C1 *f is strongly convex and Assumption 1 is satisfied.*

C2 *$f(X) = g(\mathcal{E}(X)) + \langle Q, X \rangle$ for all $X \in \Re^{m \times n}$, where $\mathcal{E} : \Re^{m \times n} \to \Re^{m \times n}$ is a linear mapping, $Q \in \Re^{m \times n}$, and g is differentiable on* dom$g$, *g is strongly convex and Assumption 1 is satisfied with replacing f by g and $\mathcal{X}^0$ by any level set of g. P is polyhedral.*

C3 $f(X) = \max_{Y\in\mathcal{Y}}\{\langle(\mathcal{E}(X)), Y\rangle - g(Y)\} + \langle Q, X\rangle$ *for all* $X \in \Re^{m\times n}$, *where* $\mathcal{Y}$ *is a polyhedral set in* $\Re^{m\times n}$, $\mathcal{E} : \Re^{m\times n} \to \Re^{m\times n}$ *is a linear mapping,* $Q \in \Re^{m\times n}$, *and* g *is differentiable on* $\mathrm{dom} g$, *Assumption 1 is satisfied with replacing* f *by* g *and* $\mathcal{X}^0$ *by any level set of* g. P *is polyhedral.*

Then Assumption 4 holds.

The next theorem establishes, under Assumptions 1, 2, 3 and 4, the linear rate of convergence of the BCGD method using the restricted Gauss-Seidel rule (20) to choose $\mathcal{J}^k$, assuming that P is block-separable, i.e.,

$$P(X) = \sum_{\ell=k}^{\tau(k)-1} P_{\mathcal{J}^\ell}(X_{\mathcal{J}^\ell}) \quad \forall X \in \Re^{m\times n},\ k \in \mathcal{T} \tag{28}$$

for some proper, convex, lsc function $P_{\mathcal{J}^\ell}$. This assumption is satisfied when $P(X) = \sum_{\mathcal{J}} \omega_{\mathcal{J}} \|X_{\mathcal{J}}\|$, with $\omega_{\mathcal{J}} > 0$ and $\mathcal{J} = \{(i,j) \in \mathcal{N} \mid i \in \mathcal{N}_1, j \in \mathcal{N}_2\}$ for some $\mathcal{N}_1, \mathcal{N}_2 \subseteq \{1, ..., n\}$, or $P(X) = \sum_{\mathcal{J}} \omega_{\mathcal{J}} \|X_{\mathcal{J}}\|_*$, with $\omega_{\mathcal{J}} > 0$, $\|\cdot\|_*$ denote the nuclear norm, i.e., the sum of singular values [18] and $\mathcal{J} = \{(i,j) \in \mathcal{N} \mid i \in \mathcal{N}_1, j \in \mathcal{N}_2\}$ for some $\mathcal{N}_1, \mathcal{N}_2 \subseteq \{1, ..., n\}$. We omit the proof of the theorem since it is nearly identical to that of [25, Theorem 2]. In what follows, by Q-linear and R-linear convergence, we mean linear convergence in the quotient and the root sense, respectively [17, Chapter 9].

Theorem 3.2 *Under Assumption 1, let* $\{X^k\}$, $\{\mathcal{H}^k\}$, $\{D^k\}$ *be sequences generated by the BCGD method satisfying Assumption 2, where* $\{\mathcal{J}^k\}$ *is chosen by the restricted Gauss-Seidel rule* (20) *with* $\mathcal{T} \subseteq \{0, 1, ...\}$ *and assuming* (28). *Then the following results hold.*

(a) $\|D_{\mathcal{I}}(X^k; \mathcal{N})\| \le \sup_j \alpha^j C r^k$ *for all* $k \in \mathcal{T}$, *where* $r^k = \sum_{\ell=k}^{\tau(k)-1} \|D^\ell\|$ *and* $C > 0$ *depends on* $n, L, \underline{\lambda}, \bar{\lambda}$.

(b) *If Assumption 3 is satisfied and* F *satisfies Assumption 4,* $\{\alpha^k\}$ *is chosen by the Armijo rule with* $\sup_k \alpha^k_{\mathrm{init}} \le 1$ *and* $\inf_k \alpha^k_{\mathrm{init}} > 0$, *then either* $\{F(X^k)\} \downarrow -\infty$ *or* $\{F(X^k)\}_{\mathcal{T}}$ *converges at least Q-linearly and* $\{X^k\}_{\mathcal{T}}$ *converges at least R-linearly.*

4 Iteration Complexity

In this section we give an upper bound on the number of iterations for the BCGD method to achieve ϵ-optimality. For simplicity, we give the proof for the case when $\{\mathcal{J}^k\}$ is chosen by the restricted Gauss-Seidel rule (20). Extension of this result to the generalized Gauss-Seidel rule (19) can be achieved; see Remark 1. In what follows, $\lceil\cdot\rceil$ denotes the ceiling function and $\tau^i(0) = \tau(\tau^{i-1}(0))$ for a positive integer i and $\tau^0(0) = 0$.

Theorem 4.1 *Under Assumptions 1, 3 and assuming also* $\inf_X F(X) > -\infty$, *let* $\{X^k\}$, $\{\mathcal{H}^k\}$, $\{D^k\}$ *be sequences generated by the BCGD method under Assumption 2, where* $\{\mathcal{J}^k\}$ *is chosen by the restricted Gauss-Seidel rule* (20) *with* $\mathcal{T} \subseteq \{0, 1, ...\}$ *and assuming* (28), *and* $\{\alpha^k\}$ *is chosen by the Armijo rule with* $\inf_k \alpha^k_{\rm init} > 0$ *and* $\sup_k \alpha^k_{\rm init} < \infty$. *Suppose* $X^k + \min\{\alpha^k_{\rm init}, \frac{\alpha^k}{\beta}\}D^k \in \mathcal{X}^0_\varrho$ *for all* k *and we define* $e^k := F(X^k) - \inf_X F(X)$. *Let* $t_k = \max\{i \mid \tau^i(0) \le k\}$, $C_1 = \frac{2C_0}{\sigma}$, $C_2 = \sqrt{2\bar\lambda C_1}r^0$, $C_0 = \frac{1}{\underline\alpha} + \frac{\sup_\ell \alpha^\ell \max_\ell \{\sum_{i=\ell}^{\tau(\ell)-1} L_i\}}{\underline\lambda^2}$ *with the Lipschitz constant* L_i *of* $\nabla f(\cdot)_{\mathcal{J}^i}$ *over* $\mathcal{X}^0$, $r^0 := \max_X \{\mathrm{dist}(X, \bar{\mathcal{X}}) \mid X \in \mathcal{X}^0\}$, $\underline\alpha := \min\{\inf_k \alpha^k_{\rm init}, \beta\min\{1, 2\underline\lambda(1-\sigma)/L, \varrho/\sup_k \|D^k\|\}\}$. *If* $\sqrt{e^k - e^{\tau(k)}} \ge C_2/C_1$, $\forall\ k > 0$, *then* $e^k \le \epsilon$ *whenever*

$$t_k \ \ge\ \left\lceil \ln\left(\frac{e^0}{\epsilon}\right) / \ln\left(1 - \frac{1}{2C_1}\right)^{-1} \right\rceil;$$

otherwise, $e^k \le \epsilon$ *whenever*

$$t_k \ge \left\lceil \frac{4C_2^2}{\epsilon} \right\rceil.$$

Proof. By (20), $\mathcal{N}$ is the disjoint union of $\mathcal{J}^k, \mathcal{J}^{k+1}, \ldots, \mathcal{J}^{\tau(k)-1}$. This together with (18) and (28) yields that

$$\sum_{i=k}^{\tau(k)-1} \Delta^i \ \le\ \min_{D\in\Re^{m\times n}} \left\{ \langle \tilde G^k, D\rangle + \frac{1}{2}\langle D, \tilde{\mathcal{H}}^k(D)\rangle + P(X^k + D) - P(X^k) \right\}, \tag{29}$$

where $\tilde G^k := \sum_{i=k}^{\tau(k)-1} \Pi_{\mathcal{J}^i}(G^i)$, $G^i := \nabla f(X^i)$, and $\tilde{\mathcal{H}}^k : \Re^{m\times n} \to \Re^{m\times n}$ is the self-adjoint positive definite mapping satisfying

$$\langle D, \tilde{\mathcal{H}}^k(D)\rangle = \sum_{i=k}^{\tau(k)-1} \langle \Pi_{\mathcal{J}^i}(D), \mathcal{H}^i(\Pi_{\mathcal{J}^i}(D))\rangle \quad \forall D \in \Re^{m\times n}. \tag{30}$$

Letting

$$\tilde D^k := D_{\tilde{\mathcal{H}}^k}(X^k; \mathcal{N}), \quad \tilde\Delta^k := \langle G^k, \tilde D^k\rangle + P(X^k + \tilde D^k) - P(X^k),$$

we then have from (29) that

$$\begin{aligned}
\sum_{i=k}^{\tau(k)-1} \Delta^i \ &\le\ \langle \tilde G^k, \tilde D^k\rangle + \frac{1}{2}\langle \tilde D^k, \tilde{\mathcal{H}}^k(\tilde D^k)\rangle + P(X^k + \tilde D^k) - P(X^k) \\
&=\ \tilde\Delta^k + \langle \tilde G^k - G^k, \tilde D^k\rangle + \frac{1}{2}\langle \tilde D^k, \tilde{\mathcal{H}}^k(\tilde D^k)\rangle \\
&\le\ \frac{\tilde\Delta^k}{2} + \|\tilde G^k - G^k\|\|\tilde D^k\| \ \le\ \frac{\tilde\Delta^k}{2} + \frac{1}{\underline\lambda}\|\tilde G^k - G^k\|^2 + \frac{\underline\lambda}{4}\|\tilde D^k\|^2,
\end{aligned}$$

where the second inequality uses Lemma 2.1 and the third inequality uses $ab \le (a^2 + b^2)/2$ for any $a, b \in \Re$. Also, letting L_i be the Lipschitz constant of $\nabla f(\cdot)_{\mathcal{J}^i}$ over $\mathcal{X}^0$, we have

$$\|\tilde G^k - G^k\|^2 \ =\ \sum_{i=k}^{\tau(k)-1} \|\Pi_{\mathcal{J}^i}(\nabla f(X^i)) - \Pi_{\mathcal{J}^i}(\nabla f(X^k))\|^2$$

$$
\begin{aligned}
&\leq \sum_{i=k}^{\tau(k)-1} L_i \|X^i - X^k\|^2 \leq \sum_{i=k}^{\tau(k)-1} L_i \sum_{j=k}^{i-1} (\alpha^j)^2 \|D^j\|^2 \leq \sum_{i=k}^{\tau(k)-1} L_i \sum_{j=k}^{i-1} (\alpha^j)^2 \frac{|\Delta^j|}{\underline{\lambda}} \\
&= \sum_{j=k}^{\tau(k)-2} \left(\sum_{i=j+1}^{\tau(k)-1} L_i \right) (\alpha^j)^2 \frac{|\Delta^j|}{\underline{\lambda}} \leq \left(\sup_\ell \alpha^\ell \sum_{i=k}^{\tau(k)-1} L_i \right) \sum_{j=k}^{\tau(k)-1} \alpha^j \frac{|\Delta^j|}{\underline{\lambda}} \\
&\leq \left(\sup_\ell \alpha^\ell \max_\ell \left\{ \sum_{i=\ell}^{\tau(\ell)-1} L_i \right\} \right) \sum_{j=k}^{\tau(k)-1} \alpha^j \frac{|\Delta^\ell|}{\underline{\lambda}}.
\end{aligned}
$$

Using the above two inequalities and $-|\Delta^i| = \Delta^i \geq \alpha^i \Delta^i / \underline{\alpha}$ (since $\alpha^i \geq \underline{\alpha}$ by Lemma 3.2), we have

$$
0 \leq \frac{\tilde{\Delta}^k}{2} + \frac{\underline{\lambda}}{4} \|\tilde{D}^k\|^2 - C_0 \sum_{j=k}^{\tau(k)-1} \alpha^j \Delta^j \leq \frac{1}{4} \tilde{\Delta}^k - C_0 \sum_{j=k}^{\tau(k)-1} \alpha^j \Delta^j, \tag{31}
$$

where $C_0 := \frac{1}{\underline{\alpha}} + \frac{\sup_\ell \alpha^\ell \max_\ell \left\{ \sum_{i=\ell}^{\tau(\ell)-1} L_i \right\}}{\underline{\lambda}^2}$, and the second inequality uses $-\tilde{\Delta}^k \geq \langle \tilde{D}^k, \tilde{\mathcal{H}}^k(\tilde{D}^k) \rangle \geq \underline{\lambda} \|\tilde{D}^k\|^2$ (it can be seen using Assumption 2 that $\underline{\lambda} \leq \lambda_{\min}(\tilde{\mathcal{H}}^k) \leq \lambda_{\max}(\tilde{\mathcal{H}}^k) \leq \bar{\lambda}$).

By Fermat's rule [20, Theorem 10.1],

$$
\tilde{D}^k \in \arg\min_D \ \langle G^k + \tilde{\mathcal{H}}^k(\tilde{D}^k), D \rangle + P(X^k + D).
$$

Let $\bar{X}^k \in \bar{\mathcal{X}}$ satisfy $\|X^k - \bar{X}^k\| = \mathrm{dist}(X^k, \bar{\mathcal{X}})$. Then

$$
\begin{aligned}
\tilde{\Delta}^k &= \langle G^k + \tilde{\mathcal{H}}^k(\tilde{D}^k), \tilde{D}^k \rangle + P(X^k + \tilde{D}^k) - P(X^k) - \langle \tilde{\mathcal{H}}^k(\tilde{D}^k), \tilde{D}^k \rangle \\
&\leq \langle G^k + \tilde{\mathcal{H}}^k(\tilde{D}^k), \bar{X}^k - X^k \rangle + P(\bar{X}^k) - P(X^k) - \lambda_{\min}(\tilde{\mathcal{H}}^k) \|\tilde{D}^k\|^2 \\
&\leq f(\bar{X}^k) - f(X^k) + \langle \tilde{\mathcal{H}}^k(\tilde{D}^k), \bar{X}^k - X^k \rangle + P(\bar{X}^k) - P(X^k) - \underline{\lambda} \|\tilde{D}^k\|^2 \\
&= F(\bar{X}^k) - F(X^k) + \langle \tilde{\mathcal{H}}^k(\tilde{D}^k), \bar{X}^k - X^k \rangle - \underline{\lambda} \|\tilde{D}^k\|^2,
\end{aligned}
$$

where the second inequality uses the convexity of f, Assumption 2, and (30). Thus

$$
\begin{aligned}
F(X^k) - F(\bar{X}^k) &\leq \langle \tilde{\mathcal{H}}^k(\tilde{D}^k), \bar{X}^k - X^k \rangle - \tilde{\Delta}^k - \underline{\lambda} \|\tilde{D}^k\|^2 \\
&\leq \|\tilde{\mathcal{H}}^k(\tilde{D}^k)\| \|\bar{X}^k - X^k\| - \tilde{\Delta}^k - \frac{\underline{\lambda}}{2} \|\tilde{D}^k\|^2 \\
&\leq \sqrt{\bar{\lambda} \langle \tilde{\mathcal{H}}^k(\tilde{D}^k), \tilde{D}^k \rangle} r^0 - \tilde{\Delta}^k - \frac{\underline{\lambda}}{2} \|\tilde{D}^k\|^2 \\
&\leq \sqrt{\bar{\lambda} |\tilde{\Delta}^k|}\ r^0 - \tilde{\Delta}^k - \frac{\underline{\lambda}}{2} \|\tilde{D}^k\|^2 \\
&\leq \sqrt{-4\bar{\lambda} C_0 \sum_{j=k}^{\tau(k)-1} \alpha^j \Delta^j}\ r^0 - 2C_0 \sum_{j=k}^{\tau(k)-1} \alpha^j \Delta^j,
\end{aligned} \tag{32}
$$

where the third inequality uses the self-adjoint positive definite property of $\tilde{\mathcal{H}}^k$ and the last inequality uses (31).

Finally, we have from the Armijo rule (17) that

$$F(X^{j+1}) - F(X^j) \;\leq\; \sigma\alpha^j\Delta^j, \quad j = k, k+1, \ldots, \tau(k)-1.$$

Summing this over $j = k, k+1, ..., \tau(k)-1$ yields that

$$F(X^{\tau(k)}) - F(X^k) \leq \sigma \sum_{j=k}^{\tau(k)-1} \alpha^j\Delta^j. \tag{33}$$

Combining (33) with (32) yields

$$e^k \leq C_2\sqrt{e^k - e^{\tau(k)}} + C_1(e^k - e^{\tau(k)}),$$

where $C_1 := \frac{2C_0}{\sigma}$ and $C_2 := \sqrt{2\bar{\lambda}C_1 r^0}$. Consider the two cases: (i) $\sqrt{e^k - e^{\tau(k)}} \geq C_2/C_1$ (ii) $\sqrt{e^k - e^{\tau(k)}} \leq C_2/C_1$. In case (i), we have $e^k \leq 2C_1(e^k - e^{\tau(k)})$ and rearranging terms yields

$$e^{\tau(k)} \leq \left(1 - \frac{1}{2C_1}\right) e^k.$$

This implies that if $\sqrt{e^k - e^{\tau(k)}} \geq C_2/C_1,\ \forall k \geq 0$ then $e^k \leq \epsilon$ whenever

$$e^0 \left(1 - \frac{1}{2C_1}\right)^{t_k} \leq \epsilon$$

or, equivalently,

$$t_k \;\geq\; \left\lceil \ln\left(\frac{e^0}{\epsilon}\right) / \ln\left(1 - \frac{1}{2C_1}\right)^{-1} \right\rceil.$$

In case (ii), we have $e^k \leq 2C_2\sqrt{e^k - e^{\tau(k)}}$ and rearranging terms yields

$$e^{\tau(k)} \leq e^k - \frac{(e^k)^2}{4C_2^2}. \tag{34}$$

We may assume $e^k > 0,\ \forall k \geq 0$ (otherwise, $e^k \leq \epsilon$). Then we consider the reciprocals $\xi^j = 1/e^j$. By (34) and $e^k > 0$, we have $0 \leq e^k/(4C_2^2) < 1$. Thus (34) yields

$$\xi^{\tau(k)} - \xi^k \geq \frac{1}{e^k(1 - e^k/(4C_2^2))} - \frac{1}{e^k} = \frac{1}{4C_2^2 - e^k} \geq \frac{1}{4C_2^2}.$$

This implies that if $\sqrt{e^k - e^{\tau(k)}} \leq C_2/C_1,\ \forall k \geq 0$ then $\xi^{\tau^{t_k}(0)} = \xi^0 + \sum_{i=0}^{t_k-1}(\xi^{\tau^{i+1}(0)} - \xi^{\tau^i(0)}) \geq \frac{t_k}{4C_2^2}$ and consequently

$$e^{\tau^{t_k}(0)} = \frac{1}{\xi^{\tau^{t_k}(0)}} \leq \frac{4C_2^2}{t_k}.$$

It follows that $e^k \leq \epsilon$ whenever

$$t_k \geq \left\lceil \frac{4C_2^2}{\epsilon} \right\rceil.$$

■

Remark 1 *By (19),* $\mathcal{J}^k \cup \mathcal{J}^{k+1} \cup \cdots \cup \mathcal{J}^{k+T-1} = \mathcal{N}$ *for* $k = 0, 1, \ldots$. *Let* $\mathcal{J}^{k_1} = \mathcal{J}^k$ *and* $\mathcal{J}^{k_i} = \mathcal{J}^{k+i-1} - (\mathcal{J}^k \cup \mathcal{J}^{k+1} \cup \cdots \cup \mathcal{J}^{k+i-2})$ *for* $i = 2, \ldots, T$. *By excluding empty sets and renumbering if necessary, there is a set* $\{\mathcal{J}^{k_1}, \ldots, \mathcal{J}^{k_{r(k)}}\}$ *such that* $r(k) \leq T$, $\mathcal{J}^{k_1} \cup \mathcal{J}^{k_2} \cup \cdots \cup \mathcal{J}^{k_{r(k)}} = \mathcal{N}$, $\mathcal{J}^{k_i} \cap \mathcal{J}^{k_j} = \emptyset$ *for* $i \neq j$, *and* $\mathcal{J}^{k_i} \neq \emptyset$ *for all* $i = 1, \ldots, r(k)$. *Also, for each* $i = 1, \ldots, r(k)$, *there is the smallest integer* q_i^k *such that* $\mathcal{J}^{k_i} \subseteq \mathcal{J}^{q_i^k}$ *and* $k \leq q_i^k < k + T$. *Then* $q_i^k \neq q_j^k$ *if* $i \neq j$. *Hence if we assume that*

$$P(X) = \sum_{\ell=1}^{r(k)} P_{\mathcal{J}^{k_\ell}}(X_{\mathcal{J}^{k_\ell}}) \quad \forall X \in \Re^{m\times n},\ k \geq 0. \tag{35}$$

instead of assuming (28). Then, since $\Delta^k \leq 0$ *for all* $k \geq 0$,

$$\sum_{i=k}^{k+T-1} \Delta^i \leq \sum_{i=1}^{r(k)} \Delta^{q_i^k}.$$

By the choice of $\mathcal{J}^{k_i}$ *and (35),*

$$\begin{aligned}
& \Delta^{q_i^k} \\
\leq\ & \langle G^{q_i^k}, D^{q_i^k}\rangle + \frac{1}{2}\langle D^{q_i^k}, \mathcal{H}^{q_i^k}(D^{q_i^k})\rangle + P(X^{q_i^k} + D^{q_i^k}) - P(X^{q_i^k}) \\
\leq\ & \langle G^{q_i^k}, \tilde{D}^{k_i}\rangle + \frac{1}{2}\langle \tilde{D}^{k_i}, \mathcal{H}^{q_i^k}(\tilde{D}^{k_i})\rangle + P(X^{q_i^k} + \tilde{D}^{k_i}) - P(X^{q_i^k}) \\
=\ & \min_{D\in\Re^{m\times n}} \left\{ \langle \tilde{G}^{k_i}, D\rangle + \frac{1}{2}\langle D, \mathcal{H}^{q_i^k}(D)\rangle + P(X^k + D) - P(X^k) \mid D_{ij} = 0,\ \forall (i,j) \notin \mathcal{J}^{k_i} \right\},
\end{aligned}$$

where $\tilde{G}^{k_i} := \Pi_{\mathcal{J}^{k_i}}(G^{q_i^k})$, $G^{q_i^k} := \nabla f(X^{q_i^k})$, $\tilde{D}^{k_i} := D_{\mathcal{H}^{q_i^k}}(X^{q_i^k}; \mathcal{J}^{k_i})$, *and the second inequality uses* $\mathcal{J}^{k_i} \subseteq \mathcal{J}^{q_i^k}$, *the equality uses* $\Pi_{\mathcal{J}^{k_i}}(X^{q_i^k}) = \Pi_{\mathcal{J}^{k_i}}(X^k)$. *Using the above two inequalities, we have*

$$\sum_{i=k}^{k+T-1} \Delta^i \leq \min_{D\in\Re^{m\times n}} \left\{ \langle \tilde{G}^k, D\rangle + \frac{1}{2}\langle D, \tilde{\mathcal{H}}^k(D)\rangle + P(X^k + D) - P(X^k) \right\},$$

where $\tilde{G}^k := \sum_{i=1}^{r(k)} \tilde{G}^{k_i}$ *and* $\tilde{\mathcal{H}}^k : \Re^{m\times n} \to \Re^{m\times n}$ *is the self-adjoint positive definite mapping satisfying*

$$\langle D, \tilde{\mathcal{H}}^k(D)\rangle = \sum_{i=1}^{r(k)} \langle \Pi_{\mathcal{J}^{k_i}}(D), \mathcal{H}^{q_i^k}(\Pi_{\mathcal{J}^{k_i}}(D))\rangle \quad \forall D \in \Re^{m\times n}.$$

Then proceeding as in the proof of Theorem 4.1 and using $C_0 := \frac{1}{\underline{\alpha}} + \frac{\sup_\ell \alpha^\ell \max_\ell \left\{ \sum_{i=1}^{r(\ell)} L_{q_i^\ell} \right\}}{\underline{\lambda}^2}$, *we can obtain the similar results of Theorem 4.1.*

5 Regularized Log-likelihood Problem: Covariance Selection

In this section, we study the boundedness of level set

$$\mathcal{X}^0 = \{X \in \mathcal{S}^n \mid F(X) \leq F(X^0)\}$$

and the Lipschitz continuity of ∇f on $\mathcal{X}^0$ for the regularized log-likelihood problem, i.e.,

$$f(X) = -\log\det X + \langle S, X\rangle$$

with $S \succeq 0_n$, which includes the covariance selection problems (1), (2), (5), and (7). The former ensures the existence of cluster point of $\{X^k\}$ while the latter will be used for the convergence rate analysis and the complexity analysis of the BCGD method when applied to the regularized log-likelihood problem. In what follows, $f(X) = -\log\det X + \langle S, X\rangle$ with $S \succeq 0_n$. Also we assume $\bar{\mathcal{X}} \neq \emptyset$. Then, since the objective function is strictly convex, the optimal solution is unique and is denoted by X^*.

Assumption 5 *There exist positive constants $\underline{\zeta}$ and $\bar{\zeta}$ such that $\underline{\zeta} I \preceq X \preceq \bar{\zeta} I$ for all $X \in \mathcal{X}^0$, where $\mathcal{X}^0 = \{X \in \mathcal{S}^n \mid F(X) \leq F(X^0)\}$.*

If Assumption 5 is satisfied, then the function f is strongly convex on $\mathcal{X}^0$, with a convexity parameter of $1/\bar{\zeta}^2$, in the sense that $\langle Y, \nabla^2 f(X)Y\rangle = \langle Y, X^{-1}YX^{-1}\rangle \geq \bar{\zeta}^{-2}\|Y\|^2$ for every $Y \in \mathcal{S}^n$. Also, f has a gradient that is Lipschitz continuous with respect to the Frobenius norm on $\mathcal{X}^0$, with Lipshitz constant $L = 1/\underline{\zeta}^2$.

The following lemma shows that Assumption 5 is satisfied when P has linear growth asymptotically.

Lemma 5.1 *Suppose $P(X) \geq \varpi\mathrm{tr}(X)$ whenever $X \succ 0$ and $\mathrm{tr}(X) \geq \zeta$, for some scalars $\varpi > 0$, $\zeta > 0$. Then Assumption 5 is satisfied.*

Proof. For any $X \in \mathcal{X}^0$ with $\mathrm{tr}(X) \geq \zeta$, we have

$$\begin{aligned} F(X^0) \;&\geq\; -\log\det X + \langle S, X\rangle + P(X) \;\geq\; -\log\det X + \varpi\mathrm{tr}(X) \\ &= \sum_{i=1}^{n} \left(-\log(\lambda_i(X)) + \varpi\lambda_i(X)\right) \;\geq\; -\log(\lambda_i(X)) + \varpi\lambda_i(X) + (n-1)\min_{t>0}\left(-\log(t) + \varpi t\right) \\ &= -\log(\lambda_i(X)) + \varpi\lambda_i(X) + (n-1)(\log(\varpi) + 1), \end{aligned} \tag{36}$$

for any $i = 1, \ldots, n$, where the first inequality uses $S \succeq 0_n$, $X \succ 0_n$.

Using $-\log(t) \geq -\log(\bar{t}) - \frac{1}{\bar{t}}(t - \bar{t})$, with $\bar{t} = \frac{2}{\varpi}$, we obtain from (36) that

$$F(X^0) \geq \log(\varpi/2) + \frac{\varpi}{2}\lambda_i(X) + 1 + (n-1)(\log(\varpi) + 1).$$

Rearranging terms yields

$$\lambda_i(X) \leq \frac{2}{\varpi}\left(F(X^0) - n(\log(\varpi) + 1) + \log(2)\right), \quad i = 1, \ldots, n.$$

This shows that $X \preceq \bar{\zeta} I$ for all $X \in \mathcal{X}^0$, where $\bar{\zeta}$ is the constant on the right-hand side.

Since $X \in \mathcal{X}^0$ so that $X \succ 0$, we also have $\lambda_i(X) > 0$ so (36) implies

$$F(X^0) > -\log(\lambda_i(X)) + (n-1)(\log(\varpi)+1).$$

Rearranging terms and taking exponential yields

$$\lambda_i(X) > \exp((n-1)(\log(\varpi)+1) - F(X^0)), \quad i = 1, \ldots, n.$$

Therefore there is a positive constant $\underline{\zeta}$ such that $X \succeq \underline{\zeta} I$ for all $X \in \mathcal{X}^0$. ∎

The assumption of Lemma 5.1 is satisfied by $P(X) = \rho \sum_{i,j=1}^n |X_{ij}|$ with $\varpi = \rho$ and ζ any positive scalar. This is also satisfied with P given by (9) and P_{ij} as in (11) or (13), with ϖ any positive scalar and ζ any scalar greater than $\mathrm{tr}(S+U)$, when $U_{ii} < \infty$ for $i = 1, ..., n$, and P given by (9) with P_{ij} as in (12), with $\varpi = \min_i\{\rho_{ii}\}$ and ζ any positive scalar, when $\rho_{ii} > 0$ for $i = 1, ..., n$.

The following lemma shows that, for any $X \in \mathcal{X}^0$, $\|X - X^*\|$ can be bounded from above by the Frobenius norm of the solution of the subproblem (14) under Assumption 5. It can be proved by proceeding as in the proof of [25, Theorem 4].

Lemma 5.2 *Suppose Assumption* 5 *holds. Then, for any* $X \in \mathcal{X}^0$,

$$\|X - X^*\| \leq (\bar{\zeta}^2 + (\bar{\zeta}/\underline{\zeta})^2)\|D_{\mathcal{I}}(X; \mathcal{N})\|. \tag{37}$$

The following theorem establishes, under Assumptions 1, 2, and 5, the linear rate of convergence of the BCGD method and gives an upper bound on the number of iterations for the BCGD method to achieve ϵ-optimality when $\mathcal{J}^k$ is chosen by the restricted Gauss-Seidel rule (20).

Theorem 5.1 *Under Assumptions 1 and* 5, *let* $\{X^k\}$, $\{\mathcal{H}^k\}$, $\{D^k\}$ *be sequences generated by the BCGD method under Assumption* 2, *where* $\{\mathcal{J}^k\}$ *is chosen by the restricted Gauss-Seidel rule* (20) *with* $\mathcal{T} \subseteq \{0, 1, ...\}$ *and assuming* (28) *and* $\{\alpha^k\}$ *is chosen by the Armijo rule with* $\inf_k \alpha^k_{\mathrm{init}} > 0$ *and* $\sup_k \alpha^k_{\mathrm{init}} < \infty$. *Then the following results hold.*

(a) *If we further assume that* $\{\alpha^k\}$ *is chosen by the Armijo rule with* $\sup_k \alpha^k_{\mathrm{init}} \leq 1$, *then either* $\{F(X^k)\} \downarrow -\infty$ *or* $\{F(X^k)\}_{\mathcal{T}}$ *converges at least Q-linearly and* $\{X^k\}_{\mathcal{T}}$ *converges* X^* *at least R-linearly.*

(b) *Suppose* $X^k + \min\{\alpha^k_{\mathrm{init}}, \frac{\alpha^k}{\beta}\}D^k \in \mathcal{X}^0_\varrho$ *for all* k *and we define* $e^k := F(X^k) - F(X^*)$ *for all* $k \geq 0$. *Let* t_k, C_1, C_2, r^0 *be defined as in Theorem 4.1 with* $C_0 = \frac{1}{\underline{\alpha}} + \frac{\sup_\ell \alpha^\ell \max_\ell\{\tau(\ell)-\ell\}}{(\underline{\lambda}\underline{\zeta})^2}$ *and* $\underline{\alpha} = \min\{\inf_k \alpha^k_{\mathrm{init}}, \beta \min\{1, 2\underline{\lambda}(1-\sigma)(1-\omega)^2\underline{\zeta}^2, \varrho/\sup_k \|D^k\|\}\}$. *If* $\sqrt{e^k - e^{\tau(k)}} \geq C_2/C_1$, $\forall\ k > 0$, *then* $e^k \leq \epsilon$ *whenever*

$$t_k \ \geq \ \left\lceil \ln\left(\frac{e^0}{\epsilon}\right) / \ln\left(1 - \frac{1}{2C_1}\right)^{-1} \right\rceil;$$

otherwise, $e^k \leq \epsilon$ *whenever*

$$t_k \geq \left\lceil \frac{4C_2^2}{\epsilon} \right\rceil.$$

Proof. (a) By Assumption 5 and Lemma 5.2, Assumption 4 holds with $\tau = (\bar{\zeta}^2 + (\bar{\zeta}/\underline{\zeta})^2)$ and $\bar{\epsilon} = \infty$. Then, by Theorem 3.2 (b), either $\{F(X^k)\} \downarrow -\infty$ or $\{F(X^k)\}_{\mathcal{T}}$ converges at least Q-linearly and $\{X^k\}_{\mathcal{T}}$ converges X^* at least R-linearly.

(b) The results are obtained by using Theorem 4.1 with $L_i = 1/\underline{\zeta}^2$ and $L = 1/(1-\omega)^2\underline{\zeta}^2$. ■

The following lemma gives the upper bound of $\langle A, B\rangle$ in terms of the maximum eigenvalue of A, the rank and the Frobenius norm of B when $A \succeq 0_n$ and $B \in \mathcal{S}^n$. It will be used in Lemma 5.4.

Lemma 5.3 *Suppose* $A \succeq 0_n$ *and* $B \in \mathcal{S}^n$. *Then* $\langle A, B\rangle \leq \lambda_{\max}(A)\sqrt{\mathrm{rank}(B)}\|B\|$.

Proof. By Fan's inequality [3, Theorem 1.2.1], we have

$$\begin{aligned}\langle A, B\rangle &\leq \sum_{i=1}^{n} \lambda_i(A)\lambda_i(B) \leq \sum_{i=1}^{n} \lambda_{\max}(A)|\lambda_i(B)| = \lambda_{\max}(A) \sum_{i=1}^{\mathrm{rank}(B)} |\lambda_i(B)| \\ &\leq \lambda_{\max}(A)\sqrt{\mathrm{rank}(B) \sum_{i=1}^{\mathrm{rank}(B)} (\lambda_i(B))^2} = \lambda_{\max}(A)\sqrt{\mathrm{rank}(B)}\|B\|\end{aligned}$$

where $\lambda_i(A)$ and $\lambda_i(B)$ are the i-th eigenvalue of A and B respectively. ■

The next lemma gives an explicit bound on $\|D_{\mathcal{H}}(X; \mathcal{J})\|$ if Assumption 5 is satisfied and P is Lipschitz continuous on $\mathrm{dom}P$.

Lemma 5.4 *Suppose Assumption* 5 *holds. For any* $X \in \mathcal{X}^0$, *nonempty* $\mathcal{J} \subseteq \mathcal{N}$, *and self-adjoint positive definite* $\mathcal{H}$ *with* $0 < \underline{\lambda} \leq \lambda_{\min}(\mathcal{H})$, *let* $D = D_{\mathcal{H}}(X; \mathcal{J})$ *and* $G = \nabla f(X)$. *If* P *is Lipschitz continuous on* $\mathrm{dom}P$ *with Lipshitz constant* L_p, *then* $\|D\| \leq 2(\|S\| + \sqrt{\mathrm{rank}(D)}/\underline{\zeta} + L_p)/\underline{\lambda}$, *where* $D = D_{\mathcal{H}}(X; \mathcal{J})$. *Moreover, the descent condition (23) is satisfied for* $\sigma \in (0, 1)$ *whenever* $0 \leq \alpha \leq \min\{\bar{\alpha}, 2\underline{\lambda}(1-\sigma)(1-\omega)^2\underline{\zeta}^2\}$. *with* $\bar{\alpha} = \min\{\omega\underline{\zeta}\underline{\lambda}/2(\|S\| + \sqrt{\mathrm{rank}(D)}/\underline{\zeta} + L_p)\}$ *and* $\omega \in (0, 1)$.

Proof. By (14),

$$\langle G, D\rangle + \frac{1}{2}\langle D, \mathcal{H}(D)\rangle + P(X + D) \leq P(X).$$

Since $G = -X^{-1} + S$ and $\underline{\zeta} I \preceq X$, by Lemma 5.3 with $A = X^{-1}$ and $B = D$, we have

$$\langle G, D\rangle = \langle S, D\rangle - \langle X^{-1}, D\rangle \geq -(\|S\| + \sqrt{\mathrm{rank}(D)}/\underline{\zeta})\|D\|.$$

This together with the Lipschitz continuity of P and $\langle D, \mathcal{H}(D)\rangle \geq \underline{\lambda}\|D\|^2$ implies that

$$\begin{aligned} & -(\|S\| + \sqrt{\mathrm{rank}(D)}/\underline{\zeta})\|D\| + \frac{\underline{\lambda}}{2}\|D\|^2 - L_p\|D\| \\ \leq \ & \langle G, D\rangle + \frac{1}{2}\langle D, \mathcal{H}(D)\rangle + P(X+D) - P(X) \leq 0 \end{aligned}$$

Hence

$$\|D\| \leq C_0, \tag{38}$$

where $C_0 = 2(\|S\| + \sqrt{\mathrm{rank}(D)}/\underline{\zeta} + L_p)/\underline{\lambda}$. If $X \in \mathcal{X}^0$, then, for any $\alpha \in [0, \bar{\alpha}]$ with $\bar{\alpha} = \min\{1, \omega\underline{\zeta}\underline{\lambda}/2(\|S\| + \sqrt{\mathrm{rank}(D)}/\underline{\zeta} + L_p)\}$ and $\omega \in (0,1)$,

$$0 \prec (1-\omega)\underline{\zeta} I = (\underline{\zeta} - \omega\underline{\zeta})I \preceq (\underline{\zeta} - \bar{\alpha}C_0)I \preceq X - \bar{\alpha}\|D\|I \preceq X + \alpha D.$$

Hence $\|(X+\alpha D)^{-1} - X^{-1}\| \leq \frac{1}{(1-\omega)^2\underline{\zeta}^2}\|\alpha D\|$. Therefore there is some $\varrho > 0$ such that f satisfies (22) with $L = \frac{1}{(1-\omega)^2\underline{\zeta}^2}$. By Lemma 3.2, the descent condition (23) is satisfied for $\sigma \in (0,1)$ whenever $0 \leq \alpha \leq \min\{\bar{\alpha}, 2\underline{\lambda}(1-\sigma)(1-\omega)^2\underline{\zeta}^2\}$. ∎

For the problem (2) (or (11)) and (7) (or (13)), $L_p = 0$ and Assumption 5 is satisfied. Hence if, for all k, we take $\mathcal{H}^k = I$, $L_i = 1/\underline{\zeta}^2$, $L = 4/\underline{\zeta}^2$, $\alpha^k_{\mathrm{init}} = \min\{1, \bar{\alpha}\}$ that is given in Lemma 5.4, and we choose $\mathcal{J}^k = \{(i,j), (j,i) \mid i = 1, ..., n\}$ where $j = k+1 (\mathrm{mod}\ n)$ or $\mathcal{J}^k = \{(i,j), (j,i) \mid i = 1, ..., j\}$ where $j = k+1 (\mathrm{mod}\ n)$, then $\mathrm{rank}(D^k) = 2$ and so the iteration bounds in Theorem 5.1 reduced to

$$O\left(\frac{n}{\underline{\zeta}^2}\ln\left(\frac{e^0}{\epsilon}\right)\right) \quad \text{or} \quad O\left(\frac{n^2\bar{\zeta}^2}{\epsilon\underline{\zeta}^2}\right).$$

Since $\mathcal{N}$ is the union of $\mathcal{J}^k, \mathcal{J}^{k+1}, \ldots, \mathcal{J}^{k+n-1}$, the resulting complexity bounds on the number of iterations for achieving ϵ-optimality can be

$$O\left(\frac{n^2}{\underline{\zeta}^2}\ln\left(\frac{e^0}{\epsilon}\right)\right) \quad \text{or} \quad O\left(\frac{n^3\bar{\zeta}^2}{\epsilon\underline{\zeta}^2}\right).$$

For the problems (1) with $\alpha = 0$, $\beta = +\infty$ (or (10)) and (5) with $\rho_{ij} > 0$ for $(i,j) \notin V$ (or (12)), we have $L_p = n\tilde{\rho}$ with $\tilde{\rho} = \rho$ for (1) and $\tilde{\rho} = \max_{(i,j)\notin V}\{\rho_{ij}\}$ for (5). Hence if, for all k, we take $\mathcal{H}^k = I$, $L_i = 1/\underline{\zeta}^2$, $L = 4/\underline{\zeta}^2$, $\alpha^k_{\mathrm{init}} = \min\{1, \bar{\alpha}\}$ that is given in Lemma 5.4, and we choose $\mathcal{J}^k = \{(i,j), (j,i) \mid i = 1, ..., n\}$ where $j = k+1 (\mathrm{mod}\ n)$ or $\mathcal{J}^k = \{(i,j), (j,i) \mid i = 1, ..., j\}$ where $j = k+1 (\mathrm{mod}\ n)$, then the resulting complexity bounds on the number of iterations for achieving ϵ-optimality can be

$$O\left(\frac{n^2(1+\underline{\zeta})}{\underline{\zeta}^2}\ln\left(\frac{e^0}{\epsilon}\right)\right) \quad \text{or} \quad O\left(\frac{n^3\bar{\zeta}^2(1+\underline{\zeta})}{\epsilon\underline{\zeta}^2}\right).$$

The computational cost per each iteration of the BCGD method is $O(n^2)$ operations except the first iteration when it is applied to solve (1) with $\alpha = 0$, $\beta = +\infty$, (2), (5) with $\rho_{ij} > 0$ for $(i,j) \notin V$, (7), and we choose $\mathcal{J}^k = \{(i,j),(j,i) \mid i = 1,...,n\}$ where $j = k+1 (\mathrm{mod}\ n)$; see Section 6. Hence the BCGD method can be implemented to achieve ϵ-optimality in

$$O\left(\frac{n^5 \bar{\zeta}^2 (1+\zeta)}{\epsilon \underline{\zeta}^2}\right)$$

operations. In contrast, The worst-case iteration complexity of interior point methods for finding an ϵ-optimal solution to (1) is $O(n \ln(e^0/\epsilon))$, where e^0 is an initial duality gap. But each iterate of interior point methods requires $O(n^6)$ arithmetic cost for assembling and solving a typically dense Newton system with $O(n^2)$ variables. Thus, the total worst-case arithmetic cost of interior point methods for finding an ϵ-optimal solution to (1) is $O(n^7 \ln(e^0/\epsilon))$ operations. And the first-order method proposed in [11] requires $O(n^3)$ operations per iteration dominated by eigenvalue decomposition and matrix multiplication of $n \times n$ matrices. As indicated in [11], the overall worst-case arithmetic cost of this first-order method for finding an ϵ-optimal solution to (1) is $O(\beta n^4/\sqrt{\epsilon})$ operations.

6 Numerical Experiments on Covariance Selection Problems

In this section, we describe the implementation of the BCGD method and report our numerical results for solving the covariance selection problems (2) and (7) on randomly generated instances. In particular, we report the comparison of the BCGD method with a first-order method (called ANS, the MATLAB code is available) [11, 12] for solving the covariance selection problem of the form (1) with $\alpha = 0$ and $\beta = +\infty$ and the covariance selection problem of the form (5). We have implemented the BCGD method in MATLAB. All runs are performed on an Intel Xeon 3.20GHz, running Linux and MATLAB (Version 7.6)

6.1 Implementation of the BCGD method

In our implementation of the BCGD method, we choose a self-adjoint positive definite linear mapping as follows:

$$\mathcal{H}^k(D) = (H^k_{ij} D_{ij})_{ij}, \tag{39}$$

where $H^k = h^k (h^k)^T$ with $h^k_j = \min\{\max\{((X^k)^{-1})_{jj}, 10^{-10}\}, 10^{10}\}\ \forall j = 1,...,n$. If $10^{-10} \le ((X^k)^{-1})_{jj} \le 10^{10}$ for all $j = 1,...,n$, then $\langle D, \mathcal{H}^k(D)\rangle = \langle \mathrm{diag}((X^k)^{-1}) D \mathrm{diag}((X^k)^{-1}), D\rangle$. By (2), $f(X) = -\log\det X - n$. Hence $\nabla^2 f(X^k)[D,D] = \langle (X^k)^{-1} D (X^k)^{-1}, D\rangle$. Then the above choice $\mathcal{H}^k$ can be viewed as a diagonal approximation to the Hessian. Also this choice has the advantage that $(X^k)^{-1}$ is already evaluated for the gradient and D^k has a closed form and can be computed efficiently in MATLAB using vector operations. We

tested the alternative choice of $\mathcal{H}^k = \eta^k \mathcal{I}$ for some fixed constant η^k including 1, but its overall performance was much worse. Also we tested another alternative choice of (39) with $H^k = h^k (h^k)^T + (W_{ij}^k W_{ij}^k)_{ij}$ where $W_{ij}^k = ((X^k)^{-1})_{ij}$ for $i \neq j$ and $W_{ij}^k = 0$ for $i = j$. This choice is a somewhat better approximation of Hessian than (39), but its overall performance was similar to that of (39). We choose the index subset $\mathcal{J}^k$ by the Gauss-Seidel (cyclic) rule, $\mathcal{J}^k = \{(i,j), (j,i) \mid i = 1, ..., n\}$ where $j = k + 1 (\text{mod } n)$. Hence we update only one column (and corresponding row) at each iteration. In order to satisfy the Armijo descent condition (17), α^k should be chosen to satisfy $X^k + \alpha^k D^k \succ 0$. Since $X^k \succ 0$, there is a positive scalar $\bar{\alpha}^k$ that satisfies $X^k + \bar{\alpha}^k D^k \succ 0$. Hence if $\alpha^k \in (0, \bar{\alpha}^k]$, then $X^k + \alpha^k D^k \succ 0$. We describe below how to compute $\bar{\alpha}^k > 0$ that satisfies $X^k + \bar{\alpha}^k D^k \succ 0$. By permutation and the choice of $\mathcal{J}^k$, we can always assume that we are updating the last column (and last row). Suppose we partition the matrices X^k and D^k in block format:

$$X^k = \begin{pmatrix} V^k & u^k \\ (u^k)^T & w^k \end{pmatrix} \text{ and } D^k = \begin{pmatrix} 0_{n-1} & d^k \\ (d^k)^T & r^k \end{pmatrix}, \tag{40}$$

where $V^k \in \mathcal{S}^{n-1}$, u^k, $d^k \in \Re^{n-1}$, and w^k, $r^k \in \Re$. Then, by [10, Theorem 7.7.6], $X^k + \alpha D^k \succ 0$ if and only of $V^k \succ 0$ and $w^k + \alpha r^k - (u^k + \alpha d^k)^T (V^k)^{-1} (u^k + \alpha d^k) > 0$. Since $X^k \succ 0$, $V^k \succ 0$. This implies that $X^k + \alpha D^k \succ 0$ if and only if

$$a_1^k \alpha^2 + 2 a_2^k \alpha - a_3^k < 0, \tag{41}$$

where $a_1^k = (d^k)^T (V^k)^{-1} (d^k)$, $a_2^k = (u^k)^T (V^k)^{-1} (d^k) - 0.5 r^k$ and $a_3^k = w^k - (u^k)^T (V^k)^{-1} (u^k)$. Since $X^k \succ 0$, $a_3^k > 0$. If $D^k \neq 0$, then either $d^k \neq 0$ or $d^k = 0$ and $r^k \neq 0$.

Case (1): If $d^k \neq 0$, then $a_1^k > 0$. This together with $a_3^k > 0$ implies that (41) is satisfied for all $\alpha \in (0, \bar{\alpha})$ where $\bar{\alpha} = \frac{-a_2^k + \sqrt{(a_2^k)^2 + a_1^k a_3^k}}{a_1^k}$. Therefore $X^k + \alpha D^k \succ 0$ for all $\alpha \in (0, \bar{\alpha})$. We set $\bar{\alpha}^k = \max\{0.9\bar{\alpha}, -a_2^k / a_1^k\}$.

Case (2): If $d^k = 0$ and $r^k > 0$, then $a_1^k = 0$ and $a_2^k = -0.5 r^k < 0$. Hence the inequality (41) is satisfied for all $\alpha > 0$. Therefore $X^k + \alpha D^k \succ 0$ for all $\alpha \in (0, \infty)$. We set $\bar{\alpha}^k = 10$.

Case (3): If $d^k = 0$ and $r^k < 0$, then $a_1^k = 0$ and $a_2^k = -0.5 r^k > 0$. Hence the inequality (41) is satisfied for all $\alpha \in (0, \bar{\alpha})$, where $\bar{\alpha} = a_3^k / (2 a_2^k)$. Therefore $X^k + \alpha D^k \succ 0$ for all $\alpha \in (0, \bar{\alpha})$. We set $\bar{\alpha}^k = 0.9\bar{\alpha}$.

The stepsize α^k can be chosen by the Armijo rule (17) by setting $\alpha_{\text{init}}^k = \bar{\alpha}^k$. But, for (2) and (7) (see (11) and (13)) with $0 < \alpha \leq 1$, by using (40),

$$\begin{aligned} & f(X^k + \alpha D^k) + P(X^k + \alpha D^k) - f(X^k) - P(X^k) \;=\; f(X^k + \alpha D^k) - f(X^k) \\ = \;& -\log\det V^k - \log\left(w^k + \alpha r^k - (u^k + \alpha d^k)^T (V^k)^{-1} (u^k + \alpha d^k)\right) - n \\ & + \log\det V^k + \log\left(w^k - (u^k)^T (V^k)^{-1} (u^k)\right) + n \\ = \;& -\log\left(a_3^k - a_1^k \alpha^2 - 2 a_2^k \alpha\right) + \log a_3^k. \end{aligned} \tag{42}$$

Hence if $-\log(a_3^k - a_1^k\alpha^2 - 2a_2^k\alpha) + \log a_3^k < 0$, then $a_3^k - a_1^k\alpha^2 - 2a_2^k\alpha > a_3^k$. And so $a_1^k\alpha^2 + 2a_2^k\alpha < 0$. Since $\alpha > 0$, $a_2^k < 0$ and so $0 < \alpha < -2a_2^k/a_1^k$ if $a_1^k \neq 0$ or $\alpha > 0$ if $a_1^k = 0$. Also $a_1^k = 0$ if and only if $d^k = 0$. This together with $a_2^k < 0$ implies that, for $0 < \alpha \leq 1$, the quantity $f(X^k + \alpha D^k) + P(X^k + \alpha D^k) - f(X^k) - P(X^k)$ is minimized when

$$\alpha = \begin{cases} \min\{1, -a_2^k/a_1^k\} & \text{if } d^k \neq 0\ ; \\ 1 & \text{else.} \end{cases}$$

Hence we would better use the limited minimization rule:

$$\min_{0<\alpha\leq\Gamma} -\log(a_3^k - a_1^k\alpha^2 - 2a_2^k\alpha) + \log a_3^k,$$

where $\Gamma = \min\{1, -a_2^k/a_1^k\}$ if $d^k \neq 0$ or $\Gamma = 1$ if $d^k = 0$.

Remark 2 *We can directly apply the BCGD method to the primal covariance selection problem (1) with $\alpha = 0$, $\beta = \infty$. Suppose that the Gauss-Seidel (cyclic) rule is employed to choose $\{\mathcal{J}^k\}$. Then only one column and corresponding row are updated at each iteration. Since $P(X) = \rho\sum_{i,j=1}^n |X_{ij}|$, by using (40),*

$$\begin{aligned} & f(X^k + \alpha D^k) + P(X^k + \alpha D^k) - f(X^k) - P(X^k) \\ = & -\log\det V^k - \log(w^k + \alpha r^k - (u^k + \alpha d^k)^T (V^k)^{-1}(u^k + \alpha d^k)) \\ & + \log\det V^k + \log(w^k - (u^k)^T (V^k)^{-1}(u^k)) + 2(\|u^k + \alpha d^k\|_1 - \|u^k\|_1) + |w^k + \alpha r^k| - |w^k| \\ = & -\log(a_3^k - a_1^k(\alpha)^2 - 2a_2^k\alpha) + \log a_3^k + 2(\|u^k + \alpha d^k\|_1 - \|u^k\|_1) + |w^k + \alpha r^k| - |w^k|, \end{aligned}$$

where a_1^k, a_2^k, and a_3^k are as defined previously. Hence the limited minimization rule can be used for finding a stepsize, but finding such a stepsize is not as simple as that of the dual formulation (2). But the method can still be implementable in $O(n^2)$ operations per iteration. Similarly, we can directly apply the BCGD method to the more general covariance selection problem (5).

Next, we give a comment on the computational complexity of the BCGD method when it is applied to the problems (2) and (7). We analyze the cost of each iteration of the BCGD method. The main computational cost per iteration is a number of inner products, vector-scalar multiplications and vector additions, each requiring $O(n)$ floating-point operations, plus a number of matrix-vector multiplications and matrix additions, each requiring $O(n^2)$ floating-point operations. In addition, each iteration requires 1 gradient (the inverse of the $n \times n$ matrix X) evaluation to find the direction, and 1 evaluation for the inverse of the $(n-1) \times (n-1)$ submatrix of X to find the stepsize. These generally requires $O(n^3)$ floating-point operations. But we only updates one column and corresponding row of the current matrix, and so the inverse of X can be updated by using the Sherman-Woodbury-Morrison formula [16, Appendix A.2] and the inverse of $(n-1)\times(n-1)$ submatrix of X can be evaluated by using the Schur complement [21, Subsection 13.2.2]. Both computations

require $O(n^2)$ operations . Therefore we only compute the full inverse of the initial matrix at the first iteration. For the evaluation of the current function value, the computation of the determinant of $n \times n$ matrix is needed and generally requires (n^3) floating-point operations. But, by using (42), it only requires $O(n^2)$ operations.

As indicated at the end of Section 5, the iteration bound of the BCGD method for achieving ϵ-optimality, when it is applied to (2) and (7), reduces to $O\left(\frac{n^3\bar{\zeta}^2}{\epsilon\underline{\zeta}^2}\right)$, where $\bar{\zeta}, \underline{\zeta}$ are two constants that satisfy Assumption 5. Hence the BCGD method can be implemented to achieve ϵ-optimality in $O\left(\frac{n^5\bar{\zeta}^2}{\epsilon\underline{\zeta}^2}\right)$ operations.

Throughout our numerical experiments, we terminate the BCGD method when the following conditions are satisfied:

$$\sqrt{\langle D_{H^k}(X^k;\mathcal{N}), \mathcal{H}^k(D_{H^k}(X^k;\mathcal{N}))\rangle} \le \varepsilon_1, \tag{43}$$

$$\frac{\left|\langle S,(X^k)^{-1}\rangle + \sum_{(i,j)\notin V}\rho_{ij}|((X^k)^{-1})_{ij}| - n\right|}{1+\left|\log\det(X^k) + \langle S,(X^k)^{-1}\rangle + \sum_{(i,j)\notin V}\rho_{ij}|((X^k)^{-1})_{ij}|\right|} \le \varepsilon_2, \tag{44}$$

where ε_1 and ε_2 are a small positive numbers. (In our numerical runs, we set $\varepsilon_1 = 5 \times 10^{-3}$ and $\varepsilon_2 = 10^{-4}$.) Here we scale $D_{H^k}(X^k;\mathcal{N})$ by H^k to reduce its sensitivity to H^k. The criterion (43) is motivated by Lemma 3.1 and is an extension of a criterion that has been suggested previously in [25]. The criterion (44) is based on the relative duality gap between the problems (5) and (7). We should note that the sequence of iterates $\{X^k\}$ generated by the BCGD method is for the dual problem (7), and we use $(X^k)^{-1}$ to estimate an approximate primal optimal solution for (5).

The stopping conditions for the ANS method are set as suggested in [12]. But we modified the "absolute gap" condition in the ANS method to the "relative gap" condition (44), with the same tolerance $\varepsilon_2 = 10^{-4}$, since the latter is more commonly used in optimization algorithms. Note that each iteration of the ANS method requires the computation of the full eigenvalue decomposition of a symmetric matrix. In its original implementation, the ANS method use the routine `eig.m` in MATLAB to carry out this task. Here we replaced `eig.m` by the LAPACK routine `dsyevd.f` (via mex-interface to MATLAB), which is based on a divide-and-conquer strategy. On our machine, the latter routine is about 5-10 times faster than the former.

6.2 Numerical Experiments on Covariance Selection Problems of the form (1) and (5)

In this subsection, we report the performance of the BCGD method and compare it to the ANS method [12] for solving the covariance selection problem of the form (1) (with $\alpha = 0$ and $\beta = +\infty$) and (5) on randomly generated instances. We note that the BCGD method is applied to solve the dual problems (2) and (7).

The BCGD method can be directly applied to solve (1) (with $\alpha = 0$ and $\beta = +\infty$) and (5). In this case, the stepsize can also be chosen by the Armijo rule or the limited minimization rule; see Remark 2. But its overall performance was worse than that of using the dual formulation. Hence we only present the numerical results of the BCGD method being applied to the dual formulation.

All the test instances used in this subsection were generated randomly in a similar manner as described in [5, 11, 12]. First, we generate a random sparse matrix $\mathtt{A} \in \mathcal{S}^n$ whose nonzero elements are set randomly to be ± 1. Then we generate a sparse inverse covariance matrix Σ^{-1} from $\mathtt{A}$ as follows:

$$\mathtt{A} = \mathtt{A} * \mathtt{A}'; \ \mathtt{d} = \mathtt{diag}(\mathtt{A}); \ T = \mathtt{diag}(\mathtt{d}) + \mathtt{max}(\mathtt{min}(\mathtt{A} - \mathtt{diag}(\mathtt{d}), 1), -1);$$
$$\Sigma^{-1} = T - \min\{1.2\lambda_{\min}(T) - \vartheta, 0\}I,$$

where ϑ is a small positive number. Then we generate a matrix $B \in \mathcal{S}^n$ by

$$B = \Sigma + \tau \frac{\|\Sigma\|_F}{\|\Xi\|_F}\Xi,$$

where $\Xi \in \mathcal{S}^n$ is a random matrix whose elements are drawn from the uniform distribution on the interval $[-1, 1]$, and τ is a small positive number. Finally, we obtain the following randomly generated sample covariance matrix:

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

In our experiments, we set $\tau = 0.15$ and $\vartheta = 10^{-4}$ for generating all instances. Let $\Omega = \{(i, j) \mid (\Sigma^{-1})_{ij} = 0, |i - j| \geq 2\}$. For the problem (5), we set V to be a random subset of Ω such that $\mathbf{card}(V)$ is about 50% of $\mathbf{card}(\Omega)$.

The regularization parameters ρ_{ij} are set to the constant values of $5/n$ for all the instances. The parameter values are chosen empirically to achieve a reasonable recovery of the true inverse covariance matrix Σ^{-1}. We should note that in general, it is impossible to recover Σ^{-1} accurately based on S by solving (1) or (5). Thus the purpose of solving (1) or (5) is not to recover the true matrix Σ^{-1} accurately but to detect the sparsity pattern of Σ^{-1} while maintaining a reasonable approximation to the true matrix. We evaluate the recovery success of Σ^{-1} by a matrix X based on the following criteria:

$$\mathrm{L}_Q := \tfrac{1}{n}\|\Sigma X - I\|_F, \quad \mathrm{L}_E := \tfrac{1}{n}(\langle \Sigma, X\rangle - \log\det(\Sigma X) - n) \tag{45}$$

$$\text{Specificity} = \tfrac{\mathrm{TN}}{\mathrm{TN+FP}}, \quad \text{Sensitivity} = \tfrac{\mathrm{TP}}{\mathrm{TP+FN}}, \tag{46}$$

where TP, TN, FP, and FN denotes the number of true positives, true negatives, false positives, and false negatives, respectively, with respect to the sparsity pattern of Σ^{-1}. In our situation, L_E and L_Q (which were considered in [28] without the normalization factor $1/n$), measure the quality of the approximation of Σ^{-1} by X, and Specificity measures the quality of zero entries while Sensitivity measures the quality of nonzero entries. We should

mention that the estimated matrix X would not be sparse in general but have many small entries. Thus we postprocess the matrix X by setting all entries which are smaller than 5×10^{-2} in absolute value to 0.

Table 1: Comparison of the BCGD and ANS methods in solving the problems (1) and (5) with the data matrix S generated randomly. The regularization parameters ρ_{ij} are set to $\rho_{ij} = 5/n$ for all the problems. The numbers in each parenthesis are: L_Q, L_E, Specificity and Sensitivity, respectively.

problem	n | density(%) | **card**(V)	iteration count		primal objective value		time (secs)	
		BCDG	ANS	BCDG	ANS	BCDG	ANS
random	500 | 2.74 | 0	1662 (2.9-2| 5.2-1| 0.99| 0.72)	46	-8.18195357 2	2.42-2	5.5	11.5
random	1000 | 4.15 | 0	8701 (1.5-2| 2.2-1| 0.99| 0.99)	87	-4.32170724 2	8.60-5	145.7	117.7
random	1500 | 4.63 | 0	12661 (1.4-2| 2.9-1| 0.99| 0.98)	84	-4.35887269 2	1.65-3	491.9	371.1
random	2000 | 5.14 | 0	18781 (1.2-2| 3.2-1| 0.98| 0.97)	93	-2.80176287 2	6.03-4	1285.1	953.9
random	500 | 1.97 | 60702	3601 (2.9-2| 5.5-1| 1.00| 0.76)	619	-8.42619444 2	-1.54-1	12.5	146.9
random	1000 | 3.33 | 241887	11341 (1.6-2| 2.2-1| 1.00| 0.99)	807	-4.45131714 2	-3.53-3	195.8	1053.4
random	1500 | 3.71 | 542496	13321 (1.4-2| 2.9-1| 1.00| 0.99)	969	-4.63013088 2	-1.21-1	528.2	3839.5
random	2000 | 4.13 | 961274	20681 (1.3-2| 3.2-1| 1.00| 0.99)	1256	-3.19691367 2	-7.90-2	1448.8	10845.3

Table 1 reports the performance of the BCGD and ANS methods. The dimension n and density (percentage of nonzero entries) of the inverse covariance matrix and the number of constraints **card**(V) are given in the second major column. In the third major column, we report the number of iterations taken by the BCGD and ANS methods. The numbers in each parenthesis correspond to $\mathrm{L_Q}$, L_E, Specificity, and Sensitivity, respectively. From the Specificity and Sensitivity values, we see that the solution of (1) or (5) can estimate the sparsity pattern of Σ^{-1} very accurately, but it is less accurate in the actual approximation of Σ^{-1}. The fourth major column gives the primal objective values. Note that the sub-column under "ANS" gives the difference in the primal objective values between the ANS and BCGD methods. As we are reporting the primal objective value for (1) or (5), a positive difference means that the ANS method achieved a better objective value than the BCGD method. Conversely, a negative difference would indicate that the BCGD method has achieved a better objective value. The last major column reports the CPU times.

From Table 1, we see that the BCGD and ANS methods are comparable in their performance when solving the problem (1) with $\alpha = 0$ and $\beta = +\infty$. But for the problem (5), the BCGD method substantially outperforms the ANS method in terms of the CPU time taken and the quality of the primal objective values attained.

7 Conclusions

In this paper we have proposed a block coordinate gradient descent method for solving convex nonsmooth optimization problems on a set of matrices. We also analyzed its computational complexity. Important applications of such convex nonsmooth optimization problems

include the covariance selection problem. On the both primal and dual formulation of the covariance selection problems (1) with $\alpha = 0$, $\beta = +\infty$, (2), (5), and (7), this method achieves linear convergence, and terminates in $O(n^5/\epsilon)$ operations with an ϵ-optimal solution. The BCGD method can be applied to solve (5) without solving a sequence of the unconstrained penalization problem (6) as did in [12]. Hence our preliminary numerical experience suggests that our method is efficient to solve the dual formulation of large-scale covariance selection problems especially with a lot of constraints.

There are some topics that can be considered as a future study. Can the complexity bound in Section 4 and 5 be sharpened? The Gauss-Southwell type rules for choosing $\mathcal{J}^k$, studied in [25], can also be extended to our general problem. We did not consider it here because we do not have a better choice for updating a coordinate block than the choice of one column and corresponding row. Can we still efficiently find $\mathcal{J}^k$ satisfying the Gauss-Southwell type rules?

References

[1] Banerjee, O., Ghaoui, L. E., and D'Aspremont, A., Model selection through sparse maximum likelihood estimation, J. Mach. Learn. Res. 9 (2008), 485–516.

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

[3] Borwein, J. M. and Lewis, A. S., Convex Analysis and Nonlinear Optimization: Theory and Examples, Springer-Verlag, New York, 2000.

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

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

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

[7] Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R., Pathwise coordinate optimization, Ann. Appl. Stat. 1 (2007), 302–332.

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

[9] Goebel, R. and Rockafellar, R. T., Local strong convexity and local Lipschitz continuity of the gradients of convex functions, J. Convex Anal. 15 (2008), 263–270.

[10] Horn, R. and Johnson C., Matrix Analysis, Cambridge University Press, Cambridge, UK, 1999.

[11] Zhaosong, L., Smooth optimization approach for covariance selection, SIAM J. Optim. 19 (2009), 1807–1827.

[12] Zhaosong, L., Adaptive first-order methods for general sparse inverse covariance selection, December 2008 (revised Jan 2010), to appear in SIAM J. Matrix Anal. Appl.

[13] Luo, Z.-Q. and Tseng, P., Error bounds and convergence analysis of feasible descent methods: a general approach, Ann. Oper. Res. 46 (1993), 157–178.

[14] Nesterov, Y., A method of solving a convex programming problem with convergence rate $O(1/k^2)$, Soviet Mathematics Doklady 27 (1983), 372–376.

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

[16] Nocedal, J. and Wright S. J., Numerical Optimization, Springer-Verlag, New York, 1999.

[17] Ortega, J. M. and Rheinboldt, W. C., Iterative Solution of Nonlinear Equations in Several Variables, reprinted by SIAM, Philadelphia, 2000.

[18] Recht, B., Fazel, M., and Parrilo, P., Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization, August 2008, to appear in SIAM Review.

[19] Rockafellar, R. T., Convex Analysis, Princeton University Press, Princeton, 1970.

[20] Rockafellar, R. T. and Wets R. J.-B., Variational Analysis, Springer-Verlag, New York, 1998.

[21] Saad, Y., Iterative Methods for Sparse Linear Systems, PWS Publishing Company, Boston, 1996.

[22] Scheinberg, K. and Rish, I., SINCO - a greedy coordinate ascent method for sparse inverse covariance selection problem, Preprint, July 2009.

[23] Tibshirani, R., Regression shrinkage and selection via the lasso, J. Royal Statist. Soc. B. 58 (1996), 267–288.

[24] Tseng, P., Convergence of block coordinate descent method for nondifferentiable minimization, J. Optim. Theory Appl. 109 (2001), 473–492.

[25] Tseng, P. and Yun, S., A coordinate gradient descent method for nonsmooth separable minimization, Math. Prog. B. 117 (2009), 387–423.

[26] Vandenberghe, L., Boyd, S., and Wu, S.-P., Determinant maximization with linear matrix iequality constrains, SIAM J. Matrix Anal. Appl. 19 (1998), 499–533.

[27] Wang, C., Sun, D., and Toh, K.-C., Solving log-determinant optimization problems by a Newton-CG primal proximal point algorithm, Preprint, September 2009.

[28] Wong, F., Carter, C. K. and Kohn, R., Efficient estimation of covariance selection models, Biometrika, 90 (2003), pp. 809–830.

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

[30] Yuan, X., Alternating direction methods for sparse covariance selection, Preprint, August 2009.

[31] Yun, S. and Toh, K.-C., A coordinate gradient descent method for ℓ_1-regularized convex minimization, April 2008 (revised Jan 2009), to appear in Comput. Optim. Appl.