ALTERNATING DIRECTION METHODS FOR SPARSE COVARIANCE SELECTION *

XIAOMING YUAN†

Abstract. The mathematical model of the widely-used sparse covariance selection problem (SCSP) is an NP-hard combinatorial problem, whereas it can be well approximately by a convex relaxation problem whose maximum likelihood estimation is penalized by the L_1 norm. This convex relaxation problem, however, is still numerically challenging, especially for large-scale cases. Recently, some efficient first-order methods inspired by Nesterov's work have been proposed to solve the convex relaxation problem of SCSP. This paper is to apply the well-known alternating direction method (ADM), which is also a first-order method, to solve the convex relaxation of SCSP. Due to the full exploitation to the separable structure of a simple reformulation of the convex relaxation problem, the ADM approach is very efficient for solving large-scale SCSP. Our preliminary numerical results show that the ADM approach substantially outperforms existing first-order methods for SCSP.

Key words. Alternating direction method, Sparse covariance selection, First-order methods, Large-scale, Separable structure.

AMS subject classifications. 90C06, 90C22,90C25,62K05,62J10

1. Introduction.

Statistician are often interested in estimating the true covariance matrix C from a sample covariance matrix Σ by maximizing its log-likelihood, for n given variables drawn from a Gaussian distribution $\mathcal{N}(0, C)$, see, e.g., [6, 12, 13, 20, 29, 30]. Since zero entries in the inverse covariance matrix correspond to the conditional independence of the corresponding variables, it is of desire to set a certain number of entries of the estimated covariance matrix to zero in order to improve the stability of the estimation and highlight conditional independence relationships among sample variables. This procedure is well known as the sparse covariance selection (see, e.g., [14]), and it captures a broad spectrum of applications in various fields such as the speech recognition, gene networks analysis, machine learning and so on, see, e.g., [1, 5, 6, 9, 15].

Mathematically, the sparse covariance selection problem (SCSP) is to maximize the log-likelihood which is penalized by the number of nonzero entries over a set of some constraints on the eigenvalues:

$$\begin{array}{ll} \max & \log(\det(X)) - \langle \Sigma, X\rangle - \rho \mathbf{Card}(X) \\ s.t. & \lambda_{\min} I \preceq X \preceq \lambda_{\max} I, \end{array} \tag{1.1}$$

where $X \in S^n$, the space of symmetric $n \times n$ matrices; $\Sigma \in S^n$ is known; **Card** (X) is the cardinality of X (i.e., the number of nonzero entries of X); I is the identity matrix in $R^{n\times n}$; $\rho > 0$ is a given scalar controlling the trade-off between the maximality of log-likelihood and cardinality; λ_{min} and λ_{max} are given bounds on the eigenvalues of X; $\langle\cdot\rangle$ is the standard trace inner product of S^n; and $X \preceq (\prec) Y$ means that $Y - X$ is positive semidefinite (definite). Despite that it reflecting both the maximum likelihood and sparsity well, the model (1.1) is an NP-hard combinatorial problem due to the presence of **Card** (X), see [14].

---

*This work has been presented on the 20th International Symposium of Mathematical Programming which was held at Chicago in August 2009.

†Department of Mathematics, Hong Kong Baptist University, Hong Kong, China (xmyuan@hkbu.edu.hk).This author was supported by the RGC grant 203009 and the NSFC grant 10701055.

A plausible approach to tackle the difficult problem (1.1) is to relax (1.1) to a convex optimization problem, by replacing the penalized term **Card** (X) with some proxies. This idea motivated the authors of [14] (see [44] for a slightly different model) to consider the following convex relaxation of (1.1):

$$\begin{array}{ll} \max & \log(\det(X)) - \langle \Sigma, X\rangle - e^T|X|e \\ s.t. & \lambda_{\min} I \preceq X \preceq \lambda_{\max} I, \end{array} \tag{1.2}$$

where $e \in R^n$ denotes the vector whose elements are all 1. Thus, $e^T|X|e = \sum_{i,j=1}^n |X_{ij}|$. As argued in [19], the term $e^T|X|e$ can be viewed as the largest convex lower bound on **Card**(X). We refer to [8, 11, 16, 40] for the rational of using $e^T|X|e$ as the proxy of **Card**(X) in regression. In addition to the easier solvability, as analyzed in [14], the convex relaxation (1.2) has many nice advantages such that it is more potential to discover the underlying distribution's structure; it can be viewed as the robust maximum likelihood estimation with noise on the sample covariance matrix Σ, and it serves as a regularization technique when the sample covariance matrix Σ is rank-deficient. For these reasons, this paper focuses on solving the convex relaxation (1.2) of SCSP.

The convex problem (1.2), however, is still not easy to solve, especially when the dimension of variables is large-scale. It is easy to verify that (1.2) can be reformulated as a constrained smooth convex problem that has an explicit $O(n^2)$-logarithmically homogeneous self-concordant barrier function, see, e.g., [32, 37]. Hence, popular solvers of interior point methods such as SeDuMe [39] and SDPT3 [42] are implementable to solve (1.2), see, e.g., [7]. Interior point methods, however, are not numerically practical for solving large-scale cases of (1.2), as analyzed in [14]. This difficulty is clear if we notice the fact that the complexity of each iteration for solving the resulted subproblem with ϵ accuracy is $O(n^6 \log(1/\epsilon))$ when interior point methods are applied to solve (1.2). Recently, the influential work of Nesterov [34, 35] has inspired some remarkable efforts to develop first-order methods for solving (1.2). More specifically, authors of [14] applied the smoothing techniques in [35] to solve (1.2) and thus developed an efficient first-order method with the complexity $O(1/\epsilon)$. This work immediately motivated the author in [32] to derive an improved first-order method with the complexity $O(1/\sqrt{\epsilon})$, by applying Nesterov's smoothing techniques to solve the dual counterpart of (1.2).

The success application of these efficient work [14, 32] enhances the promising role of first-order methods, and it urges us to focus on the approach of first-order methods for solving (1.2). On the other hand, as pointed out by Nesterov in [36]: "It was becoming more and more clear that the proper use of the problem's structure can lead to very efficient optimization methods.....". Hence, this paper is devoted to the effort of developing first-order methods for solving (1.2) by fully exploiting its intrinsic structure. More specifically, we shall show that a simple reformulation of (1.2) (see (2.1)) is readily implementable by the classical alternating direction method (ADM), which is also a first-order method and has been widely used in many areas such as convex programming, variational inequalities, image processing [4, 10, 17, 18, 21, 22, 23, 24, 25, 27, 31, 33, 41]. By taking full advantage of its high-level separable structure of the ADM-oriented reformulation, the ADM approach will be verified to be very efficient for solving large-scale (1.2). In particular, when the ADM approach is applied to solve (1.2), the computational load of each iteration is dominated by the computation of only one eigenvalue decomposition of an $n \times n$ matrix whose complexity is $O(n^3)$. Therefore, the complexity for one single iteration of the ADM

approach is the same as the variant of smooth minimization method in [32], and it is much lower than that of the Nesterov's method in [14] which requires two eigenvalue decompositions and one inverse matrix computation. Indeed, as pointed out in [14]: "we cannot expect to do better than $O(n^3)$, which is the cost of solving the non-penalized problem for dense covariance matrices Σ". At the same time, we will show, by numerical comparison for solving large-scale cases of (1.2), that the total number of iterations of the ADM approach is significantly smaller than those of [14, 32]. We hence believe that the ADM approach is a simple, yet powerful, approach for solving well-structured problems such as (1.2).

The rest of the paper is as follows. In Section 2, we first provide a reformulation of problem (1.2) with high-level separable structure. Then, we propose the ADM approach for solving this reformulation, and elaborate the procedure of solving the resulted subproblems. An appropriate stopping criterion to implement the ADM approach to solve (1.2) is finally presented in this section. In Section 3, we analyze some properties of the sequence generated by the ADM approach for solving (1.2). In Section 4, we propose two concrete ADM type algorithms for solving (1.2) and prove the convergence. In Section 5, we report some numerical results of the ADM approach for solving some large-scale cases of (1.2) and the comparison with some existing methods. Finally, some conclusions are drawn in Section 6.

2. The ADM approach. In this section, we first present a reformulation of (1.2) which exhibits nice separable structure in both the objective function and the constraints. Thus, the ADM approach becomes implementable. Then, we elaborate the procedure of solving the subproblems emerging in the implementation of ADM, and analyze its complexity. At the end, we present a stopping criterion in order to implement ADM.

2.1. An ADM-oriented reformulation. By introducing an auxiliary variable y, the convex relaxation of SCSP (1.2) is obviously equivalent to the following problem:

$$\begin{array}{ll} \min & \langle \Sigma, X\rangle - \log(\det(X)) + \rho e^T|Y|e \\ \text{s.t} & X - Y = 0, \\ & X \in S_\lambda^n := \{X \succeq 0 \mid \lambda_{\min} I \preceq X \preceq \lambda_{\max} I\}. \end{array} \tag{2.1}$$

2.2. The ADM approach for (2.1). Then, the Augmented Lagrangian function of (2.1) is

$$L(X, Y, Z) := \langle \Sigma, X\rangle - \log(\det(X)) + \rho e^T|Y|e - \langle Z, X - Y\rangle + \frac{\beta}{2}\|X - Y\|^2,$$

where $Z \in \mathcal{R}^{n\times n}$ is the multiplier of the linear constraint $X - Y = 0$ and $\beta > 0$ is the penalty parameter for the violation of the linear constraint. Obviously, the classical Augmented Lagrangian method (see, e.g., [3, 38]) is applicable for solving (2.1), and its iterative scheme is:

$$\begin{cases} (X^{k+1}, Y^{k+1}) \in \operatorname{argmin}_{X\in S_\lambda^n, Y\in\mathcal{R}^{n\times n}}\{L(X, Y, Z^k)\}, \\ Z^{k+1} = Z^k - \beta(X^{k+1} - Y^{k+1}), \end{cases} \tag{2.2}$$

where (X^k, Y^k, Z^k) is the given triple of iterate. The direct application of the Augmented Lagrangian method, however, treats (2.1) as a generic minimization problem with linear constraints, and ignores its favorable separable structure emerging in both

the constraints and the objective function. Hence, the variables X and Y are minimized simultaneously in (2.2). This ignorance of Augmented Lagrangian method, however, can be made up by the well-known ADM method (see [22, 23, 24, 25]) which minimizes the variables X and Y serially. More specifically, ADM solves the following problems to generate the new iterate:

$$\begin{cases} X^{k+1} \in \operatorname{argmin}_{X\in S^n_\lambda}\{L(X, Y^k, Z^k)\}, & \text{(2.3a)}\\ Y^{k+1} \in \operatorname{argmin}_{Y\in\mathcal{R}^{n\times n}}\{L(X^{k+1}, Y, Z^k)\}, & \text{(2.3b)}\\ Z^{k+1} = Z^k - \beta(X^{k+1} - Y^{k+1}). & \text{(2.3c)}\end{cases}$$

Therefore, the ADM method (2.3) is virtually a practical version of the Augmented Lagrangian method (2.2) by taking advantage of the high-level separable structure of the problem (2.1) to the full extent.

2.3. Solving subproblems of ADM. According to (2.3), when the ADM approach applied to solve (2.1), the main computation of each iteration is to solve two minimization problems. We now elaborate the strategies of solving these sub problems, and derive the computational complexity of ADM for solving (2.1).

First, we consider the first minimization problem (2.3a). It is easy to verify that it is equivalent to the following minimization problem:

$$X^{k+1} = \operatorname{argmin}_{X\in S^n_\lambda}\big\{\frac{1}{2}\|X - [Y^k - \frac{1}{\beta}(\Sigma - Z^k)]\|^2 - \frac{1}{\beta}\log(\det(X))\big\}. \tag{2.4}$$

For the case of that $\lambda_{\min} = 0$ and $\lambda_{\max} = +\infty$ (in this case, S^n_λ reduces to the S^n_+, the cone of positive semidefinite matrices), we have actually $X^{k+1} \succ 0$ due to the log term in the objective function of (2.1). Hence, solving (2.4) reduces to solving the following matrix equation:

$$X - \Big(Y^k - \frac{1}{\beta}(\Sigma - Z^k)\Big) - \frac{1}{\beta}X^{-1} = 0. \tag{2.5}$$

For convenience, we denote by

$$A = Y^k - \frac{1}{\beta}(\Sigma - Z^k) \tag{2.6}$$

and let

$$A = V\Lambda V^T \quad \text{with} \quad \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) \tag{2.7}$$

be the the symmetric Schur decomposition of A; $V = (v_1, \ldots, v_n)$ be an orthogonal matrix whose column vector $v_i (i = 1, \ldots, n)$ are eigenvectors of A; and $\lambda_i (i = 1, \ldots, n)$ be the corresponding eigenvalues. Then, the solution of (2.5) should have the same eigenvectors as A. Let

$$X = V\tilde{\Lambda}V^T \quad \text{with} \quad \tilde{\Lambda} = \operatorname{diag}(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_n). \tag{2.8}$$

be the symmetric Schur decomposition of X where V is identical with that in (2.7) and $\tilde{\lambda}_i$ $(i = 1, 2, \cdot, n)$ are to be determined. In fact, substitute (2.7) and (2.8) into (2.5), it turns out that (2.5) reduces to the following easier equation:

$$\tilde{\Lambda} - \Lambda - \frac{1}{\beta}\tilde{\Lambda}^{-1} = 0.$$

Recall that both Λ and $\tilde{\Lambda}$ are diagonal matrices. Hence, we have

$$\tilde{\lambda}_j = \frac{\lambda_j + \sqrt{\lambda_j^2 + (4/\beta)}}{2}, \quad j = 1, \ldots, n. \tag{2.9}$$

Obviously, $\tilde{\lambda}_j > 0$. Therefore, the solution of (2.3a) is: $X^{k+1} = V\tilde{\Lambda}V^T$ where V is obtained by (2.7) and $\tilde{\Lambda}$ is obtained by (2.9).

For the general case that $\lambda_{\max} \geq \lambda_{\min} \geq 0$, it is analogous that the solution of (2.3a) is given by

$$\tilde{X} = V\tilde{\Lambda}V^T \quad \text{with} \quad \tilde{\Lambda} = \mathrm{diag}(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_n),$$

where

$$\tilde{\lambda}_j = \min\Big(\max\Big(\lambda_{\min}, \ \frac{\lambda_j + \sqrt{\lambda_j^2 + (4/\beta)}}{2}\Big), \ \lambda_{\max}\Big), \quad j = 1, \ldots, n;$$

V and λ_j are obtained by (2.7).

For the second subproblem (2.3b), it is easy to verify that it is actually a shrinkage problem which usually arises in image processing (see, e.g., [8]). In fact, (2.3b) is characterized by the following inclusion:

$$0 \in \rho\partial(|Y|) + [Z^k - \beta(X^{k+1} - Y)].$$

Here, $\partial(|Y|) := (s_{ij}) \in \mathcal{R}^{n\times n}$ with $s_{ij} \in \partial(y_{ij})$ where $\partial(\cdot)$ denotes the subgradient operator of the nondifferentiable convex function $|\cdot|$. Hence, we conclude that (2.3b) can be solved easily with explicit solution:

$$Y^{k+1} = \frac{1}{\beta}\big\{(\beta X^{k+1} - Z^k) - P_{B_\infty^\rho}[\beta X^{k+1} - Z^k]\big\},$$

where

$$B_\infty^\rho = \{X \in \mathbf{R}^{n\times n} \mid -\rho \leq X_{ij} \leq \rho\}.$$

Therefore, the computation load of each iteration of the ADM approach is dominated by the eigenvalue decomposition of A (see (2.6)-(2.7)). Recall that the complexity of implementing the symmetric QR Algorithm (Algorithm 8.3.3 in [26]) to compute an approximate symmetric Schur decomposition is about $9n^3$ flops.

2.4. Stopping criterion of ADM. Note that the reformulation (2.1) is equivalent to the following problem:

$$\begin{cases} X \in S_\Lambda^n, \quad \langle X' - X, \Sigma - X^{-1} - Z\rangle \geq 0, \quad \forall X' \in S_\Lambda^n, \\ 0 \in \rho\partial(|Y|) + Z, \\ X - Y = 0. \end{cases} \tag{2.10}$$

Therefore, the problem (2.1) has the following variational inequality characterization: Find $u \in \Omega := S_\Lambda^n \times \mathcal{R}^{n\times n} \times \mathcal{R}^{n\times n}$ such that

$$u \in \Omega, \qquad \langle u' - u, F(u)\rangle \geq 0, \qquad \forall\ u' \in \Omega, \tag{2.11}$$

where

$$u = \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \qquad \text{and} \qquad F(u) = \begin{pmatrix} \Sigma - X^{-1} - Z \\ \rho\partial(|Y|) + Z \\ X - Y \end{pmatrix}. \tag{2.12}$$

Let $(X^{k+1}, Y^{k+1}, Z^{k+1}) \in \Omega$ be generated by the ADM approach (2.3). Note that (2.3a) is characterized by

$$\langle X' - X^{k+1}, \Sigma - (X^{k+1})^{-1} - [Z^k - \beta(X^{k+1} - Y^k)]\rangle \geq 0, \quad \forall X' \in S^n_\Lambda.$$

Then, we have

$$\langle \begin{pmatrix} X' - X^{k+1} \\ Y' - Y^{k+1} \\ Z' - Z^{k+1} \end{pmatrix}, \begin{pmatrix} \Sigma - (X^{k+1})^{-1} - Z^{k+1} \\ \rho\partial(|Y^{k+1}|) + Z^{k+1} \\ X^{k+1} - Y^{k+1} \end{pmatrix} - \begin{pmatrix} \beta(Y^k - Y^{k+1}) \\ 0 \\ \frac{1}{\beta}(Z^k - Z^{k+1}) \end{pmatrix} \rangle \geq 0, \forall (X', Y', Z') \in \Omega. \tag{2.13}$$

Therefore, it is clear that $(X^{k+1}, Y^{k+1}, Z^{k+1})$ is a solution of (2.10) if and only if $Y^k = Y^{k+1}$ and $Z^k = Z^{k+1}$. This observation motivates us to develop the stopping criterion for implementing ADM in the following manner:

$$\max\{ey, ez\} \leq \epsilon, \tag{2.14}$$

where $\epsilon > 0$ and

$$ey := \max_{i,j}\{|(Y^k - Y^{k+1})_{ij}|\} \qquad \text{and} \qquad ez := \max_{i,j}\{|(Z^k - Z^{k+1})_{ij}|\}. \tag{2.15}$$

3. Contractive Properties of the ADM approach. In this section, we prove a contractive property of the sequence generated by the ADM approach (2.3), which ensures convergence for the ADM type algorithms to be developed.

LEMMA 3.1. *Let $(X^{k+1}, Y^{k+1}, Z^{k+1}) \in \Omega$ be the iterate generate by the ADM approach (2.3) and (X^*, Y^*, Z^*) be a solution of (2.1). Let*

$$v^i = \begin{pmatrix} Y^i \\ Z^i \end{pmatrix}, \qquad v^* = \begin{pmatrix} Y^* \\ Z^* \end{pmatrix} \qquad \text{and} \qquad G = \begin{pmatrix} \beta I_n & \\ & \frac{1}{\beta} I_n \end{pmatrix}.$$

Then, we have

$$\langle v^k - v^*, G(v^k - v^{k+1})\rangle \geq \varphi(v^k, v^{k+1}) := \|v^k - v^{k+1}\|^2_G - \langle Z^k - Z^{k+1}, Y^k - Y^{k+1}\rangle. \tag{3.1}$$

Proof. Recall (2.13). For any $u^* = (X^*, Y^*, Z^*)$, we have

$$\langle \begin{pmatrix} X^* - X^{k+1} \\ Y^* - Y^{k+1} \\ Z^* - Z^{k+1} \end{pmatrix}, \begin{pmatrix} \Sigma - (X^{k+1})^{-1} - Z^{k+1} \\ \rho\partial(|Y^{k+1}|) + Z^{k+1} \\ X^{k+1} - Y^{k+1} \end{pmatrix} - \begin{pmatrix} \beta(Y^k - Y^{k+1}) \\ 0 \\ \frac{1}{\beta}(Z^k - Z^{k+1}) \end{pmatrix} \rangle \geq 0,$$

which can be rewritten as

$$\langle v^{k+1} - v^*, G(v^k - v^{k+1})\rangle \geq \langle u^{k+1} - u^*, F(u^{k+1})\rangle + \langle \begin{pmatrix} X^{k+1} - X^* \\ Y^{k+1} - Y^* \end{pmatrix}, \begin{pmatrix} -\beta(Y^k - Y^{k+1}) \\ \beta(Y^k - Y^{k+1}) \end{pmatrix} \rangle. \tag{3.2}$$

Now, we observe the right-hand-side of (3.2). First, since $\langle u^{k+1}-u^*, F(u^*)\rangle \geq 0$ and $F(u)$ is monotone (which is an immediate fact from (2.11)-(2.12)), we have $\langle u^{k+1}-u^*, F(u^{k+1})\rangle \geq 0$. For the second term of the right-hand-side of (3.2), using $X^*-Y^*=0$ and $\beta(X^{k+1}-Y^{k+1})=Z^k-Z^{k+1}$, we have

$$\langle \begin{pmatrix} X^{k+1}-X^* \\ Y^{k+1}-Y^* \end{pmatrix}, \begin{pmatrix} -\beta(Y^k-Y^{k+1}) \\ \beta(Y^k-Y^{k+1}) \end{pmatrix} \rangle = -\langle Z^k-Z^{k+1}, Y^k-Y^{k+1}\rangle.$$

Therefore, the inequality (3.1) is derived immediately from the above inequality and (3.2) □

4. ADM-based Algorithms. Based on the previous analysis, we propose two concrete ADM type algorithms for solving (2.1), i.e., the original ADM and the extended ADM. For notational convenience, in this section, we denote by $\tilde{W}^k := (\tilde{X}^k, \tilde{Y}^k, \tilde{Z}^k)$ the triple generated by the iterative scheme (2.3) from the given $w^k = (X^k, Y^k, Z^k) \subset S^n_+ \times S_B \times S^n$.

4.1. The original ADM. The original ADM takes the following schemes to generate the new triple $W^{k+1}=(X^{k+1}, Y^{k+1}, Z^{k+1})$:

The original ADM in [22, 23]:

$$\begin{cases} X^{k+1} = \tilde{X}^k; \\ Y^{k+1} = \tilde{Y}^k; \\ Z^{k+1} = \tilde{Z}^k. \end{cases} \tag{4.1}$$

Convergence of the original ADM is referred to [4, 22, 23, 24]. In fact, the convergence is alternatively clear via the following analysis.

THEOREM 4.1. *The sequence generated by the original ADM method (4.1) is Féjer monotone with respect to the solution set.*

Proof. . It follows from (2.3) that

$$0 \in \rho\partial(|Y^{k+1}|) + Z^{k+1}.$$

Thus, we have

$$\langle Y^k - Y^{k+1}, \rho\partial(|Y^{k+1}|) + Z^{k+1}\rangle = 0. \tag{4.2}$$

Analogously, we get

$$\langle Y^{k+1} - Y^k, \rho\partial(|Y^k|) + Z^k\rangle = 0. \tag{4.3}$$

Adding (4.2) and (4.3), we obtain

$$\langle Y^k - Y^{k+1}, Z^{k+1} - Z^k\rangle \geq \rho\langle Y^k - Y^{k+1}, \partial(|Y^k|) - \partial(|Y^{k+1}|)\rangle \geq 0,$$

where the second inequality comes from the fact that $|Y|$ is a convex function. Hence, from (3.1), we have

$$\varphi(v^k, v^{k+1}) \geq \|v^k - v^{k+1}\|^2_G. \tag{4.4}$$

By using (4.4), we obtain

$$\begin{aligned}\|v^{k+1}-v^*\|_G^2 &= \|(v^k-v^*)-(v^k-v^{k+1})\|_G^2\\ &= \|v^k-v^*\|_G^2-2\langle v^k-v^*,v^k-v^{k+1}\rangle+\|v^k-v^{k+1}\|_G^2\\ &\le \|v^k-v^*\|_G^2-2\varphi(v^k,v^{k+1})+\|v^k-v^{k+1}\|_G^2\\ &= \|v^k-v^*\|_G^2-\|v^k-v^{k+1}\|_G^2.\end{aligned}$$

Therefore, the sequence $\{v^k\}$ generated by the original ADM (4.1) is Féjer monotone, which implies the convergence of OADM immediately, see, e.g., [2].□

4.2. The extended ADM. To solve a class of variational inequalities with separable structures, the original ADM (2.3) was extended in [43] and thus an ADM based descent method was developed therein. This technique can be readily used to solve (2.1). Accordingly, we present the following extended ADM, based on the idea of [43].

The extended ADM in [43]:

$$\begin{pmatrix} Y^{k+1}\\ Z^{k+1}\end{pmatrix}=\begin{pmatrix} Y^{k}\\ Z^{k}\end{pmatrix}-\gamma\alpha_k^*\begin{pmatrix} Y^{k}-\tilde{Y}^{k}\\ Z^{k}-\tilde{Z}^{k}\end{pmatrix}, \tag{4.5}$$

where

$$\alpha_k^*=\frac{\varphi(v^k,\tilde{v}^k)}{\|v^k-\tilde{v}^k\|_G^2}; \tag{4.6}$$

$\varphi(v^k,\tilde{v}^k)$ is defined in (3.1) and $\gamma\in(0,\ 2)$.

Based on (3.1) and (4.4), it is implied that $(v^k-\tilde{v}^k)$ is a descent direction of $\|v-v^*\|^2$ at $v=v^k$. Hence, it is reasonable to derive the descent step (4.5) with the step size (4.6). Convergence of the extended ADM method (4.5) is implied immediately from the following result.

THEOREM 4.2. *The sequence generated by the extended ADM method (4.5) is Féjer monotone with respect to the solution set.*

Proof. Note that the iterative scheme (4.5) can be written as

$$v^{k+1}=v^k-\gamma\alpha_k^*(v^k-\tilde{v}^k).$$

It follows from (3.1) and (4.6) that

$$\begin{aligned}\|v^{k+1}-v^*\|_G^2 &= \|(v^k-v^*)-\gamma\alpha_k^*(v^k-\tilde{v}^k)\|_G^2\\ &= \|v^k-v^*\|_G^2-2\gamma\alpha_k^*\langle v^k-v^*,G(v^k-\tilde{v}^k)\rangle+\gamma^2(\alpha_k^*)^2\|v^k-\tilde{v}^k\|_G^2\\ &\le \|v^k-v^*\|_G^2-2\gamma\alpha_k^*\varphi(v^k,\tilde{v}^k)+\gamma^2(\alpha_k^*)^2\|v^k-\tilde{v}^k\|_G^2\\ &= \|v^k-v^*\|_G^2-\gamma(2-\gamma)\alpha_k^*\varphi(v^k,\tilde{v}^k).\end{aligned}$$

On the other hand, it is very easy to derive that

$$\varphi(v^k,\tilde{v}^k)\ge\frac{1}{2}\|v^k-\tilde{v}^k\|_G^2, \tag{4.7}$$

which obviously implies that $\alpha_k^*\ge\frac{1}{2}$. Thus, we obtain

$$\|v^{k+1}-v^*\|_G^2\le\|v^k-v^*\|_G^2-\frac{\gamma(2-\gamma)}{4}\|v^k-\tilde{v}^k\|_G^2.$$

Therefore, the sequence $\{v^k\}$ generated by the extended ADM (4.5) is Féjer monotone, which implies the convergence immediately, see, e.g., [2].□

5. Numerical Results. In this section, we focus on the numerical performance of the proposed ADM approach and its comparison with the efficient Variant Smoothing Minimization (VSM) method proposed in [32]. For the purpose of comparison, we shall test the identical case of (1.2) as that tested by existing first-order methods in [14, 32]. Note that in [32], the VSM method was shown to be more efficient than the Nesterov's smooth approximation scheme and the block-coordinate descent method developed in [14].

In particular, the sample covariance matrix Σ in (1.2) is generated by

$$\Sigma = A^{-1} + \tau V - \min\{\lambda_{min}(A^{-1} + \tau V) - \vartheta, 0\}\mathcal{I},$$

where $V \in S^n$ is an independent and identically distributed uniform random matrix; $A \in S^n$ is a sparse invertible matrix with positive diagonal entries and the density is prescribed by the constant ϱ; $\mathcal{I}$ is the identity matrix in $R^{n\times n}$; τ and ϑ are both small positive constants. For comparison, we take the same values as in [32], i.e., $\varrho = 0.01$, $\tau = 0.15$, $\vartheta = 0.0001$; and in (1.2) we take $\lambda_{min} = 0$, $\lambda_{max} = +\infty$ and $\rho = 0.5$.

To implement the ADM approach, we take the initial iterate as $(Y^0 = I_n, Z^0 = \mathbf{0}_n)$, and $\gamma \equiv 1.5$ for the extend ADM (4.5). Note that ADM are theoretically convergent for any constant $\beta > 0$. In numerical experiments, the initial value of β is taken as $\max(0.15n, 50)$, and we reduce the value of β by multiplying the constant $2/3$ if $ez \le 0.8\varepsilon$ (see (2.15) for the definitions of ey and ez). We refer to, e.g., [28], for the convergence of ADM with variable β.

All the codes of the ADM approach were written by Matlab 7.1. On the other hand, we run the matlab codes which were downed from the author's website of [32] to implement the smoothing minimization approach in [32]. All the codes were run on on a Dell Poweredge 1950 dual processors server equipped with Quad Core Xeon 3.0GHz CPU, 16GB RAM running Fedora 8 Linux.

Note that we adopt (2.14) as the stopping criterion, while VSM in [32] terminates iterations when the duality gap of (1.2) is less than 0.1. For the purpose of comparison, we also measure the duality gap of (1.2) when the new iterate is generated by the ADM approach. More specifically, it is easy to know that the solution of (2.10) (X^*, Y^*, Z^*) should satisfy

$$X^* \succeq 0, \qquad \Sigma - (X^*)^{-1} - Z^* \succeq 0, \qquad \langle X^*,\ \Sigma - (X^*)^{-1} - Z^*\rangle = 0.$$

For $u^k = (X^k, Y^k, Z^k)$ generated by the ADM approach, the duality gap between (1.2) and its dual counterpart is $\langle X^k, \Sigma - (X^k)^{-1} - Z^k\rangle$. We denote

$$F_X(u^k) = \Sigma - (X^k)^{-1} - Z^k; \tag{5.1}$$

$$F_X^+(u^k) = P_{S_+^n}[F_X(u^k)];$$

and

$$F_X^-(u^k) = F_X^+(u^k) - F_X(u^k).$$

Then, both $F_X^+(\tilde{u}^k)$ and $F_X^-(\tilde{u}^k)$ are positive semi-definite. Note that it is reasonable to measure the duality gap for the iterate u^k generated by the ADM approach by the following:

$$\text{Dgap1} := \langle x^k,\ F_X^+(u^k) + F_X^-(u^k)\rangle$$

and

$$\text{Dgap2} := \text{The maximal eigenvalue of } F_X^-(u^k).$$

In the numerical experiment, we take $\varepsilon = 10^{-3}$ in (2.14). We compare the original ADM (OADM) and the extended ADM (EADM) with VSM in [32]. In the following table, we report the respective iteration number (It.) and the computing time in seconds (CPU.) for VSM, OADM and EADM. For VSM, we report the duality gap (Gap); while for OADM and EADM, we report both "Dgap1" and "Dgap2".

Table 1. Numerical Comparison of VSM and ADM

	VSM			OADM				EADM			
n	It.	CPU.	Gap	It.	CPU.	Dgap1	Dgap2	It.	CPU.	Dgap1	Dgap 2
100	32	0.6	9.61e-02	21	0.11	1.37e-02	2.09e-05	21	0.11	3.15e-03	2.24 e-05
200	108	9.7	9.89e-02	36	0.76	1.13e-02	9.25e-05	28	0.62	2.57e-03	9.61e-04
300	145	34.6	9.97e-02	34	1.76	4.95e-02	4.63e-04	26	1.40	3.91e-03	9.95e-04
400	153	81.6	9.82e-02	33	3.56	1.59e-02	6.84e-04	26	1.87	2.73e-03	8.02e-04
500	153	153.2	9.46e-02	36	6.96	3.48e-03	1.13e-03	27	5.24	3.57e-03	1.29e-03
600	164	279.8	9.89e-02	35	10.29	5.23e-02	2.59e-03	29	9.04	3.07e-03	1.42e-03
700	162	437.5	9.80e-02	48	21.25	5.80e-04	5.15e-04	34	16.11	8.72e-04	7.74e-04
800	168	671.5	9.95e-02	55	34.91	7.81e-04	1.07e-03	41	27.58	7.74e-04	4.86e-04
900	160	918.3	9.99e-02	48	41.97	2.14e-04	3.65e-04	35	32.23	2.87e-03	1.62e-03
1,000	160	1,271.0	9.99e-02	60	70.19	2.00e-04	1.58e-04	43	52.66	3.57e-03	2.42e-03
2,000	171	11,017.2	9.99e-02	62	637.30	4.66e-04	2.43e-04	55	566.26	6.29e-04	1.94e-04

The data in Table 1 shows that for solving SCSP, the ADM approach substantially outperforms VSM in terms of both iterative numbers and computing time.

6. Conclusions. In this paper, the well known alternating direction method (ADM) is applied to solve the sparse covariance selection problem (SCSP), and its numerical performance significantly outperforms existing first-order methods. The ADM approach is eligible for solving large-scale SCSP, because of its numerical efficiency and easy implementation. The efficiency of the ADM approach for SCSP, which is mainly due to the full exploitation to the favorable separable structure of a reformulation of SCSP, emphasizes the rationale of developing attractive first-order methods for some particular problems by taking advantage of their inherent structures.

REFERENCES

[1] J. Akaike, *Information theory and an extension of the maximum likelihood principle In B. N. Petrov and F. Csaki, editors*, Second international symposium on information theory (1973), pp, 267-281. Akedemiai Kiado, Budapest.
[2] H. H. Bauschke and P. L. Combettes, *Aweak-to-strong convergence principle for Fejr-monotone methods in Hilbert spaces*, Mathematics of Operations Research, 26(2) (2001), pp. 248-264.
[3] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier methods*, Academic Press, 1982.
[4] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and distributed computation: Numerical methods*, Prentice-Hall, Englewood Cliffs, NJ, 1989.
[5] J. A. Bilmes, *Natural statistic models for automatic speech recognition*, Ph.D. thesis, UC Berkeley, 1999.
[6] J. A. Bilmes, *Factored sparse inverse covariance matrices*, IEEE International Conference on Acoustics, Speech, and Signal Processing, 2000.
[7] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge University Press, 2004.
[8] A. M. Bruckstein, D. L Donoho and M. Elad, *From Sparse Solutions of Systems of Equations to Sparse Modeling of Signals and Images*, SIAM Review, 51(1) (2009), pp. 34-81.

[9] K. P. Burnham and R. D. Anderson, *Multimodel inference. understanding aic or bic in model selection*, Sociological methods and research, 33 (2004), pp. 261-304.
[10] G. Chen, and M. Teboulle, *A proximal-based decomposition method for convex minimization problems*, Mathematical Programming, 64 (1994), pp. 81-101.
[11] S. Chen, D. Donoho, and M. Saunders, *Atomic decomposition by basis pursuit*, SIAM Journal on Scientific Computing, 20 (1998), pp. 33-61.
[12] J. Dahl, L. Vandenberghe and V. Roychowdhury, *Covariance selection for non-chordal graphs via chordal embedding*, Optimization Methods and Software 23 (4) (2008), pp. 501-520.
[13] A. Dempster, *Covariance selection*, Biometrics, 28 (1972), pp. 157-175.
[14] A. dAspremont, O. Banerjee, and L. El Ghaoui *First-Order Methods for Sparse Covariance Selection*, SIAM Journal on Matrix Analysis and its Applications, 30(1) (2008), pp. 56-66.
[15] A. Dobra, C. Hans, B. Jones, J.R. J. R. Nevins, G. Yao, and M. West, *Sparse graphical models for exploring gene expression data*, Journal of Multivariate Analysis, 90 (2004), pp. 196-212.
[16] D. L. Donoho and J. Tanner, *Sparse nonnegative solutions of underdetermined linear equations by linear programming*, Proceedings of the National Academy of Sciences, 102 (2005), pp. 9446-9451.
[17] J. Eckstein and M. Fukushima, *Some reformulation and applications of the alternating directions method of multipliers*, In: Hager, W. W. *et al.* eds., Large Scale Optimization: State of the Art, Kluwer Academic Publishers, pp. 115-134, 1994.
[18] E. Esser, *Applications of Lagrangian-based alternating direction methods and connections to split Bregman*, Manuscript, http://www.math.ucla.edu/applied/cam/.
[19] M. Fazel, H. Hindi, and S. Boyd, *A rank minimization heuristic with application to minimum order system approximation, Proceedings American Control Conference*, 6 (2001), pp. 4734-4739.
[20] J. Friedman, T. Hastie and R. Tibshirani, *Sparse inverse covariance estimation with the graphical lasso*, Biostatistics, 9 (2008), pp. 432-441.
[21] M. Fukushima, *Application of the alternating direction method of multipliers to separable convex programming problems*, Computational Optimization and Applications, 1(1992), pp. 93-111.
[22] D. Gabay, *Application of the method of multipliers to varuational inequalities*, In: Fortin, M., Glowinski, R., eds., Augmented Lagrangian methods: Application to the numerical solution of Boundary-Value Problem, North-Holland, Amsterdam, The Netherlands, pp. 299-331, 1983.
[23] D. Gabay and B. Mercier, *A dual algorithm for the solution of nonlinear variational problems via finite element approximations*, Computational Mathematics with Applications, 2(1976), pp. 17-40.
[24] R. Glowinski, *Numerical methods for nonlinear variational problems*, Springer-Verlag, New York, Berlin, Heidelberg, Tokyo, 1984.
[25] R. Glowinski and P. Le Tallec, Augmented Lagrangian and Operator Splitting Methods in Nonlinear Mechanics, SIAM Studies in Applied Mathematics, Philadelphia, PA, 1989.
[26] G. H. Golub and C. F. van Loan, Matrix Computation, third edition, The Johns Hopkins University Press, 1996.
[27] B. S. He, L. Z. Liao, D. Han and H. Yang, *A new inexact alternating directions method for monontone variational inequalities*, Mathematical Programming, 92(2002), pp. 103-118.
[28] B.S. He and H. Yang, Some convergence properties of a method of multipliers for linearly constrained monotone variational inequalities, Operations Research Letters, 23 (1998), pp. 151-161.
[29] J. Z. Huang, N. Liu, and M. Pourahmadi, *Covariance matrix selection and estimation via penalised normal likelihood*, Biometrika, 93(1) (2006), pp. 85-98.
[30] B. Jones, C. Carvalho, C. Dobra, A. Hans, C. Carter, and M. West, *Experiments in stochastic computation for high-dimensional graphical models*, Statistical Science, 20 (2005), pp. 388-400.
[31] S. Kontogiorgis and R. R. Meyer, *A variable-penalty alternating directions method for convex optimization*, Mathematical Programming, 83(1998), pp. 29-53.
[32] Z. Lu, *Smooth Optimization Approach for Sparse Covariance Selection*, SIAM Journal on Optimization, 19(4) (2009), pp. 1807-1827.
[33] M. Ng, P. A. Weiss and X. M. Yuan, *Solving constrained total-Variation problems via alternating direction methods*, Manuscript, 2009.
[34] 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.

[35] Y. E. Nesterov, *Smooth minimization of nonsmooth functions*, Mathematical Programming, 103 (2005), pp. 127-152.
[36] Y. Nesterov, *Gradient methods for minimizing composite objective function*, CORE Discussion Paper 2007/76, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain, Belgium, 2007.
[37] Y. E. Nesterov and A. S. Nemirovski, *Interior point Polynomial algorithms in Convex Programming: Theory and Applications*, Society for Industrial and Applied Mathematics, Philadelphia, 1994.
[38] J. Nocedal and S. J. Wright, Numerical Optimization. Second Edition, Spriger Verlag, 2006.
[39] J. F. Sturm, *Using SeDuMi 1.02, a Matlab toolbox for optimization over symmetric cones*, Optimization Methods and Software 11 & 12 (1999), pp. 625-653.
[40] R. Tibshirani, *Regression shrinkage and selection via the LASSO*, Journal of the Royal statistical society, series B, 58 (1996), pp. 267-288.
[41] P. Tseng, *Alternating projection-proximal methods for convex programming and variational inequalities*, SIAM Journal on Optimization, 7(1997), pp. 951-965.
[42] R. H. Tütüncü, K. C. Toh and M. J. Todd, *Solving semidefinite-quadrtic-linear programs using SDPT3*, Mathematical Programming, 95(2003), pp. 189-217.
[43] C. H. Ye and X. M. Yuan, *A Descent Method for stuctured monotone variational inequalities*, Optimization Methods and Software, 22(2007), pp. 329-338.
[44] M. Yuan and Y. Lin, Model selection and estimation in the Gaussian graphical model. Biometrika, 94(1) (2007), pp. 19-35.