FIRST-ORDER METHODS FOR SPARSE COVARIANCE SELECTION

ALEXANDRE D'ASPREMONT*, ONUREENA BANERJEE†, AND LAURENT EL GHAOUI‡

Abstract. Given a sample covariance matrix, we solve a maximum likelihood problem penalized by the number of nonzero coefficients in the inverse covariance matrix. Our objective is to find a sparse representation of the sample data and to highlight conditional independence relationships between the sample variables. We first formulate a convex relaxation of this combinatorial problem, we then detail two efficient first-order algorithms with low memory requirements to solve large-scale, dense problem instances.

1. Introduction. We discuss a problem of model selection[1]. Given n variables drawn from a Gaussian distribution $\mathcal{N}(0, C)$, where the true covariance matrix C is unknown, we estimate C from a sample covariance matrix Σ by maximizing its log-likelihood. Following [7], setting a certain number of coefficients in the inverse covariance matrix Σ^{-1} to zero, a procedure known as *covariance selection*, improves the stability of this estimation procedure by reducing the number of parameters to estimate and highlights structure in the underlying model.

Here, we focus on the problem of *discovering* this pattern of zeroes in the inverse covariance matrix. We seek to trade-off the log-likelihood of the solution with the number of zeroes in its inverse, and solve the following estimation problem:

$$\begin{array}{ll} \text{maximize} & \log\det X - \langle \Sigma, X\rangle - \rho\,\mathbf{Card}(X) \\ \text{subject to} & \alpha\mathbf{I}_n \preceq X \preceq \beta\mathbf{I}_n \end{array} \tag{1.1}$$

in the variable $X \in \mathbf{S}_n$, where $\Sigma \in \mathbf{S}_n^+$ is the sample covariance matrix, $\mathbf{Card}(X)$ is the cardinality of X, i.e. the number of nonzero components in X, $\rho > 0$ is a parameter controlling the trade-off between log-likelihood and cardinality, finally $\alpha,\ \beta > 0$ fix bounds on the eigenvalues of the solution.

Zeroes in the inverse covariance matrix correspond to conditionally independent variables in the model and this approach can be used to simultaneously determine a robust estimate of the covariance matrix and, perhaps more importantly, discover *structure* in the underlying graphical model. In particular, we can view (1.1) as a model selection problem using Aikake (AIC, see [1]) or Bayes (BIC, see [5]) information criterions. Both these problems can be written as in (1.1) with $\rho = 2/N$ for the AIC problem and $\rho = 2\log(N/2)/N$ for the BIC problem, where N is the sample size. This has applications in speech recognition (see [2, 3]) or gene networks analysis (see [9, 8] for example).

The $\mathbf{Card}(X)$ penalty term makes the estimation problem (1.1) combinatorial (NP-Hard in fact), and our first objective here is to derive a convex relaxation to this problem which can be solved efficiently. We then derive two first-order algorithms geared towards memory efficiency and large-scale, *dense* problem instances.

In [3], Bilmes proposed a method for covariance selection based on choosing statistical dependencies according to conditional mutual information computed using

*ORFE Department, Princeton University, Princeton, NJ 08544. aspremon@princeton.edu
†EECS Department, UC Berkeley, Berkeley, CA 94720. onureena@eecs.berkeley.edu
‡EECS Department, UC Berkeley, Berkeley, CA 94720. elghaoui@eecs.berkeley.edu

[1] A subset of the results discussed here appeared in the proceedings of the International Conference on Machine Learning, Pittsburgh 2006.

training data. Other recent work involves identifying those Gaussian graphical models that are best supported by the data and any available prior information on the covariance matrix. This approach is used by [13, 9] on gene expression data. Recently, [6, 12] also considered penalized maximum likelihood estimation for covariance selection. In contrast to our results here, [12] work on the Cholesky decomposition of X using an iterative (heuristic) algorithm to minimize a nonconvex penalized likelihood problem, while [6] propose a set of large scale interior point algorithms to solve sparse problems, i.e. problems for which the conditional independence structure is already known.

The paper is organized as follows, in Section 2, we detail our convex relaxation of problem (1.1) and study the dual. In Section 3, we derive two efficient algorithms to solve it. Finally, in Section 4 we describe some numerical results.

2. Problem setup.

2.1. Convex relaxation. Given a sample covariance matrix $\Sigma \in \mathbf{S}_n^+$, we can write the following convex relaxation to the estimation problem (1.1):

$$\begin{array}{ll} \text{maximize} & \log\det X - \langle \Sigma, X\rangle - \rho \mathbf{1}^T |X| \mathbf{1} \\ \text{subject to} & \alpha \mathbf{I}_n \preceq X \preceq \beta \mathbf{I}_n, \end{array} \tag{2.1}$$

with variable $X \in \mathbf{S}^n$, where $\mathbf{1}$ is the n-vector of ones, so that $\mathbf{1}^T|X|\mathbf{1} = \sum_{i,j=1}^n |X_{ij}|$. The penalty term involving the sum of absolute values of the entries of X is a proxy for the number of its non-zero elements: the function $\mathbf{1}^T|X|\mathbf{1}$ can be seen as the largest convex lower bound on $\mathbf{Card}(X)$ on the hypercube, an argument used by [11] for rank minimization. It is also often used in regression techniques, such as the LASSO studied by [19], when sparsity of the solution is a concern. This relaxation is provably tight in certain cases (see [10]). In our model, the bounds (α, β) on the eigenvalues of X are fixed, and user-chosen. Although we allow $\alpha = 0$, $\beta = +\infty$, such bounds are useful in practice to control the condition number of the solution.

For $\rho = 0$, and provided $\Sigma \succ 0$, problems (1.1) and (2.1) have a unique solution $X^\star = \Sigma^{-1}$, and the corresponding maximum-likelihood estimate is Σ. Due to noise in the data, in practice, the sample estimate Σ may not have a sparse inverse, even if the underlying graphical model exhibits conditional independence properties. By striking a trade-off between maximality of the likelihood and number of non-zero elements in the inverse covariance matrix, our approach is potentially useful at *discovering structure*, precisely conditional independence properties in the data. This means that we have to focus on the case where the matrix X is *dense*. At the same time, it serves as a regularization technique: when Σ is rank-deficient, there is no well-defined maximum-likelihood estimate, whereas the solution to problem (2.1) is always unique and well-defined for $\rho > 0$, as seen below.

2.2. Dual problem, robustness. We can rewrite the relaxation (2.1) as the following min-max problem:

$$\max_{\{X:\ \alpha\mathbf{I}_n \preceq X \preceq \beta\mathbf{I}_n\}} \ \min_{\{U:\ |U_{ij}| \le \rho\}} \log\det X - \langle \Sigma + U, X\rangle \tag{2.2}$$

which gives a natural interpretation of problem (2.1) as a worst-case *robust maximum likelihood* problem with componentwise bounded, additive noise on the sample covariance matrix Σ. The corresponding Lagrangian is given by:

$$L(X, U, P, Q) = \log\det X - \mathbf{Tr}((\Sigma + U + Q - P)X) - \alpha\,\mathbf{Tr}\,P + \beta\,\mathbf{Tr}\,Q$$

and we get the following dual to (2.1):

$$\begin{array}{ll} \text{minimize} & -\log\det(\Sigma+U+Q-P)-n+\alpha\,\mathbf{Tr}\,P-\beta\,\mathbf{Tr}\,Q \\ \text{subject to} & P,Q\succeq 0,\ |U_{ij}|\leq\rho,\quad i,j=1,\ldots,n, \end{array} \tag{2.3}$$

in the variables $U, P, Q \in \mathbf{S}_n$.

When $\alpha = 0$ and $\beta = +\infty$, the first-order optimality conditions impose $X(\Sigma + U) = \mathbf{I}_n$, hence we always have:

$$X \succeq \alpha(n)\mathbf{I}_n \quad \text{with} \quad \alpha(n) := \frac{1}{\|\Sigma\| + n\rho},$$

zero duality gap also means $\mathbf{Tr}(\Sigma X) = n - \rho\mathbf{1}^T|X|\mathbf{1}$. Because X and Σ are both positive semidefinite, we get:

$$\|X\|_2 \leq \|X\|_F \leq \mathbf{1}^T|X|\mathbf{1} \leq \frac{n}{\rho},$$

which, together with $\mathbf{Tr}(\Sigma X) \geq \lambda_{\min}(\Sigma)\|X\|_2$, means $\|X\|_2 \leq n/\lambda_{\min}(\Sigma)$. Finally then, we must always have:

$$X \preceq \beta(n)\mathbf{I}_n \quad \text{with} \quad \beta(n) := n\min\left(\frac{1}{\rho}, \|\Sigma^{-1}\|_2\right).$$

and $0 < \alpha(n) \leq \lambda(X) \leq \beta(n) < +\infty$ at the optimum. Setting $\alpha = 0$ and $\beta = +\infty$ in problem (2.1) is then equivalent to setting $\alpha = \alpha(n)$ and $\beta = \beta(n)$. Since the objective function of problem (2.1) is strictly convex when $0 < \alpha(n) \leq \lambda(X) \leq \beta(n) < +\infty$, this shows that (2.1) always has a unique solution.

3. Algorithms. In this section, we present two algorithms for solving problem (2.1), one based on an optimal first-order method developed in [18], the other based on a block-coordinate gradient method.

Of course, problem (2.1) is convex and can readily be solved using interior point methods (see [4] for example). However, such second-order methods become quickly impractical for solving (1.1), since the corresponding complexity to compute an ϵ-suboptimal solution is $O(n^6\log(1/\epsilon))$. Note however that 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.

3.1. Smooth optimization. The recently-developed first-order algorithms due to [18] trade-off a better dependence on problem size against a worst dependence on accuracy, usually $1/\epsilon$ instead of its logarithm and the method we describe next has a complexity of $O(n^{4.5}/\epsilon)$. In addition, the memory space requirement of these first-order methods is much lower than that of interior-point methods, which involve forming a dense Hessian, and hence become quickly prohibitive with a problem having $O(n^2)$ variables.

Nesterov's format. The algorithm in [18] supposes that the function to minimize conforms to a certain representation. This is the case for our problem here, so we first write (2.2) in the saddle-function format described in [18]:

$$\min_{X\in\mathcal{Q}_1} -\log\det X + \langle\Sigma, X\rangle + \rho\mathbf{1}^T|X|\mathbf{1} \equiv \min_{X\in\mathcal{Q}_1}\max_{U\in\mathcal{Q}_2} \hat{f}(X) + \langle A(X), U\rangle$$

where we define $\hat{f}(X) = -\log\det X + \langle\Sigma, X\rangle$, $A = \rho I_{n^2}$, and

$$\mathcal{Q}_1 := \{X\in\mathcal{S}^n \ :\ \alpha\mathbf{I}_n \preceq X \preceq \beta\mathbf{I}_n\}, \quad \mathcal{Q}_2 := \{U\in\mathcal{S}^n \ :\ \|U\|_\infty \leq 1\}.$$

The adjoint of this problem, corresponding to the dual problem (2.3), is then written:

$$\max_{U\in\mathcal{Q}_2} \phi(U) \quad \text{where} \quad \phi(U) := \min_{X\in\mathcal{Q}_1} -\log\det X + \langle \Sigma + U, X\rangle. \tag{3.1}$$

When a function can be represented in this saddle function format, the method described in [18] combines two steps. *Regularization*: by adding a strongly convex penalty to the saddle function representation of f, the algorithm first computes a smooth ϵ-approximation of f with Lipschitz continuous gradient. This can be seen as a generalized Moreau-Yosida regularization step (see [14] for example). *Optimal first order minimization*: the algorithm then applies the optimal first-order scheme for functions with Lipschitz continuous gradients detailed in [16] to the regularized function. Each iteration requires efficiently computing the regularized function value and its gradient. In all the semidefinite programming applications detailed here, this can be done extremely efficiently, with a complexity of $O(n^3)$ and memory requirements in $O(n^2)$. The method is only efficient if all these steps can be performed explicitly or at least very efficiently. As we will see below, this is the case here.

Prox-functions and related parameters. To $\mathcal{Q}_1$ and $\mathcal{Q}_2$ we now associate norms and so-called prox-functions. For $\mathcal{Q}_1$, we use the Frobenius norm, and a prox-function:

$$d_1(X) = -\log\det X + \log\beta.$$

The function d_1 is strongly convex on $\mathcal{Q}_1$, with a convexity parameter of $\sigma_1 = 1/\beta^2$, in the sense that $\nabla^2 d_1(X)[H,H] = \mathbf{Tr}(X^{-1}HX^{-1}H) \geq \beta^{-2}\|H\|_F^2$ for every H. Furthermore, the center of the set, $X_0 := \arg\min_{X\in\mathcal{Q}_1} d_1(X)$ is $X_0 = \beta\mathbf{I}_n$, and satisfies $d_1(X_0) = 0$. With our choice, we have $D_1 := \max_{X\in\mathcal{Q}_1} d_1(X) = n\log(\beta/\alpha)$.

To $\mathcal{Q}_2$, we also associate the Frobenius norm, and the prox-function $d_2(U) = \|U\|_F^2/2$. With this choice, the center U_0 of $\mathcal{Q}_2$ is $U_0 = 0$. Furthermore, the function d_2 is strongly convex on its domain, with convexity parameter with respect to the 1-norm $\sigma_1 = 1$, and we have $D_2 := \max_{U\in\mathcal{Q}_2} d_2(U) = n^2/2$.

The function $\hat{f}$ has a gradient that is Lipschitz-continuous with respect to the Frobenius norm on the set $\mathcal{Q}_1$, with Lipschitz constant $M = 1/\alpha^2$. Finally, the norm (induced by the Frobenius norm) of the operator A is $\|A\| = \rho$.

Smooth minimization. The method is based on replacing the objective of the original problem, $f(X)$, with $f_\epsilon(X)$, where $\epsilon > 0$ is the desired accuracy, and f_ϵ is a penalized function involving the prox-function d_2, defined as:

$$f_\epsilon(X) := \hat{f}(X) + \max_{U\in\mathcal{Q}_2} \langle X, U\rangle - (\epsilon/2D_2)d_2(U). \tag{3.2}$$

The above function turns out to be a smooth uniform approximation to f everywhere, with maximal error $\epsilon/2$. Furthermore, the function f_ϵ is has a Lipschitz-continuous gradient, with Lipschitz constant given by $L(\epsilon) := M + D_2\|A\|^2/(2\sigma_2\epsilon)$. A specific first-order algorithm detailed in [16] for smooth, constrained convex minimization is then applied to the function f_ϵ, with convergence rate in $O(\sqrt{L(\epsilon)/\epsilon})$.

Nesterov's algorithm. Choose $\epsilon > 0$ and set $X_0 = \beta\mathbf{I}_n$, the algorithm then updates primal and dual iterates Y_k and $\hat{U}_k$ using the following steps:

1. Compute $\nabla f_\epsilon(X_k) = -X^{-1} + \Sigma + U^*(X_k)$, where $U^*(X)$ solves (3.2).
2. Find $Y_k = \mathrm{argmin}_{Y\in\mathcal{Q}_1} \{\langle \nabla f_\epsilon(X_k), Y - X_k\rangle + \frac{1}{2}L(\epsilon)\|Y - X_k\|_F^2\}$
3. Find $Z_k = \mathrm{argmin}_{Z\in\mathcal{Q}_1} \left\{\frac{L(\epsilon)d_1(Z)}{\sigma_1} + \sum_{i=0}^k \frac{i+1}{2}(f_\epsilon(X_i) + \langle \nabla f_\epsilon(X_i), Z - X_i\rangle)\right\}$
4. Update $X_k = \frac{2}{k+3}Z_k + \frac{k+1}{k+3}Y_k$ and $\hat{U}_k = \frac{k\hat{U}_{k-1} + 2U^*(X_k)}{(k+2)}$

5. Repeat until the duality gap is less than the target precision:

$$-\log\det Y_k + \langle \Sigma, Y_k\rangle + \rho \mathbf{1}^T |Y_k| \mathbf{1} - \phi(\hat{U}_k) \leq \epsilon.$$

The key to the method's success is that step 1-3 and 5 can be performed explicitly and only involve an eigenvalue decomposition. Step one above computes the (smooth) function value and gradient. The second step computes the *gradient mapping*, which matches the gradient step for unconstrained problems (see [17, p.86]). Step three and four update an *estimate sequence* [17, p.72] of f_μ whose minimum can be computed explicitly and gives an increasingly tight upper bound on the minimum of f_μ. We now present these steps in detail for our problem.

Step 1. The first step requires computing the gradient of the function

$$f_\epsilon(X) = \hat{f}(X) + \max_{u\in\mathcal{Q}_2} \langle X, U\rangle - (\epsilon/2D_2)d_2(U).$$

This function can be expressed in closed form as $f_\epsilon(X) = \hat{f}(X) + \sum_{i,j}\psi_\mu(X_{ij})$, where

$$\psi_\epsilon(x) := \begin{cases} |x| - (\epsilon/4D_2) & \text{if } |x| \geq (\epsilon/2\rho D_2), \\ D_2 x^2/\epsilon & \text{otherwise,} \end{cases}$$

which is simply the Moreau-Yosida regularization of the absolute value and the gradient of the function at X is

$$\nabla f_\mu(X) = -X^{-1} + \Sigma + U^*(X),$$

with

$$U^*(X) := \max(\min(X/\mu, \rho), -\rho),$$

with min. and max. understood componentwise. The cost of this step is dominated by that of computing the inverse of X, which is $O(n^3)$.

Step 2. This step involves a problem of the form

$$T_{\mathcal{Q}_1}(X) = \arg\min_{Y\in\mathcal{Q}_1} \langle \nabla f_\epsilon(X), Y - X\rangle + \frac{1}{2}L\|Y - X\|_F^2,$$

where $X \in \mathcal{Q}_1$ is given. This problem can be reduced to one of projection on $\mathcal{Q}_1$, namely

$$\min_{Y\in\mathcal{Q}_1} \|Y - G\|_F^2,$$

where $G := X - L^{-1}\nabla f_\epsilon(X)$. Using the rotational invariance of this problem, we reduce it to a vector problem:

$$\min_\lambda \textstyle\sum_i (\lambda_i - \gamma_i)^2 \ : \ \alpha \leq \lambda_i \leq \beta, \quad i = 1, \ldots, n,$$

where γ is the vector of eigenvalues of G. This problem admits a simple explicit solution:

$$\lambda_i = \min(\max(\gamma_i, \alpha), \beta), \quad i = 1, \ldots, n.$$

The corresponding solution is then $Y = V^T\mathbf{diag}(\lambda)V$, where $G = V^T\mathbf{diag}(\gamma)V$ is the eigenvalue decomposition of G. The cost of this step is dominated by the cost of forming the eigenvalue decomposition of G, which is $O(n^3)$.

Step 3. The third step involves solving a problem of the form

$$Z := \arg\max_{X \in \mathcal{Q}_1} \; d_1(X) + \langle S, X \rangle, \tag{3.3}$$

where S is given. Again, due to the rotational invariance of the objective and feasible set, we can reduce the problem to a one-dimensional problem:

$$\min_\lambda \; \textstyle\sum_i \sigma_i \lambda_i - \log \lambda_i \;\; : \;\; \alpha \leq \lambda_i \leq \beta,$$

where σ contains the eigenvalues of S. This problem has a simple, explicit solution:

$$\lambda_i = \min(\max(1/\sigma_i, \alpha), \beta), \quad i = 1, \ldots, n.$$

The corresponding solution is then $Y = V^T \mathbf{diag}(\lambda) V$, where $S = V^T \mathbf{diag}(\sigma) V$ is the eigenvalue decomposition of S. Again, the cost of this step is dominated by the cost of forming the eigenvalue decomposition of S, which is $O(n^3)$.

Computing $\phi(\hat{U}_k)$. For a given matrix $\hat{U}_k$ the function ϕ is computed as in (3.1):

$$\phi(\hat{U}_k) = \min_{X \in \mathcal{Q}_1} - \log\det X + \langle \Sigma + \hat{U}_k, X \rangle.$$

This means projecting $(\Sigma + \hat{U}_k)^{-1}$ on $\mathcal{Q}_1$, i.e. only involves an eigenvalue decomposition.

Complexity estimate. To summarize, for step 1, the gradient of f_ϵ is readily computed in closed form, via the computation of the inverse of X. Step 2 essentially amounts to projecting on $\mathcal{Q}_1$, and requires an eigenvalue problem to be solved; likewise for step 3. In fact, each iteration costs $O(n^3)$. The number of iterations necessary to achieve an objective with absolute accuracy less than ϵ is then given by:

$$N(\epsilon) := 4\|A\| \frac{1}{\epsilon} \sqrt{\frac{D_1 D_2}{\sigma_1 \sigma_2}} + \sqrt{\frac{M D_1}{\sigma_1 \epsilon}} = \frac{\kappa \sqrt{n(\log \kappa)}}{\epsilon} (4n\alpha\rho + \sqrt{\epsilon}), \tag{3.4}$$

where $\kappa = \log(\beta/\alpha)$ bounds the solution's condition number. Thus, the overall complexity when $\rho > 0$ is in $O(n^{4.5}/\epsilon)$, as claimed.

3.2. Block-coordinate gradient methods. In this section, we focus on the particular case where $\alpha = 0$ and $\beta = +\infty$ (hence implicitly $\alpha = \alpha(n)$, $\beta = \beta(n)$) and derive gradient minimization algorithms that take advantage of the problem structure. We consider the following problem:

$$\max_X \;\; \log\det X - \langle \Sigma, X \rangle - \rho \mathbf{1}^T |X| \mathbf{1} \tag{3.5}$$

in the variable $X \in \mathbf{S}_n$, where $\rho > 0$ again controls the trade-off between log-likelihood and sparsity of the inverse covariance matrix. Its dual is given by:

$$\begin{array}{ll} \text{minimize} & -\log\det(\Sigma + U) - n \\ \text{subject to} & |U_{ij}| \leq \rho, \quad i, j = 1, \ldots, n. \end{array} \tag{3.6}$$

in the variable $U \in \mathbf{S}_n$. We partition the matrices X and U in block format:

$$X = \begin{pmatrix} Z & x \\ x^T & y \end{pmatrix} \quad \text{and} \quad U = \begin{pmatrix} V & u \\ u^T & w \end{pmatrix},$$

where $Z \succ 0$ and U are fixed and $x,\ u \in \mathbf{R}^{(n-1)},\ y,\ w \in \mathbf{R}$ are the variables (row and column) we are updating. We also partition the sample matrix according to the same block structure:

$$\Sigma = \begin{pmatrix} A & b \\ b^T & c \end{pmatrix},$$

where $A \in \mathbf{S}_{(n-1)},\ b \in \mathbf{R}^{(n-1)},\ c \in \mathbf{R}$. In the methods that follow, we will update only one column (and corresponding row) at a time and without loss of generality we can always assume that we are updating the last one.

Block-coordinate descent. The dual problem (3.6):

$$\begin{array}{ll} \text{minimize} & -\log\det(\Sigma + U) - n \\ \text{subject to} & |U_{ij}| \leq \rho, \quad i,j = 1,\ldots,n. \end{array}$$

in the variable $U \in \mathbf{S}_n$, can be written in block format as:

$$\begin{array}{ll} \text{minimize} & -\log\det(A+V) - \log\big((w+c) - (b+u)^T(A+V)^{-1}(b+u)\big) - n \\ \text{subject to} & |w| \leq \rho,\ |u_i| \leq \rho, \quad i = 1,\ldots,n. \end{array}$$

in the variables $u \in \mathbf{R}^{(n-1)}$ and $w \in \mathbf{R}$ (V is fixed at each iteration). We directly get $w = \rho$ so the diagonal of the optimal solution must be $\rho\mathbf{1}$. The main step at each iteration is then a box constrained quadratic program (QP):

$$\begin{array}{ll} \text{minimize} & (b+u)^T(A+V)^{-1}(b+u) \\ \text{subject to} & |u_i| \leq \rho, \quad i = 1,\ldots,n, \end{array} \tag{3.7}$$

in the variable $u \in \mathbf{R}^{(n-1)}$. To summarize, the block coordinate descent algorithm proceeds as follows:

1. Pick the row and column to update.
2. Compute $(A+V)^{-1}$.
3. Solve the box constrained QP in (3.7).
4. Repeat until duality gap is less than precision: $\langle \Sigma, X\rangle - n + \rho\mathbf{1}^T|X|\mathbf{1} \leq \epsilon$.

At each iteration, we need to compute the inverse of the submatrix $(A+V) \in \mathbf{S}_{(n-1)}$, but we can update this inverse using the Sherman-Woodbury-Morrison formula on two rank-two updates, hence it is only necessary to compute a full inverse at the first iteration.

Block-coordinate ascent. For a fixed Z, problem (3.5) is equivalent to:

$$\begin{array}{ll} \text{maximize} & \log\big(y - x^TZ^{-1}x\big) - 2b^Tx - y(c+\rho) - 2\rho\|x\|_1 \\ \text{subject to} & y - x^TZ^{-1}x > 0,\ y > 0, \end{array}$$

in the variables $x \in \mathbf{R}^{(n-1)},\ y \in \mathbf{R}$, where $Z \succ 0$ (given) and the Schur complement constraints imply $X \succ 0$. We can solve for the optimal y explicitly and the problem in x becomes:

$$\max_x -x^TQx - 2b^Tx - 2\rho\|x\|_1,$$

where $Q := (c+\rho)Z^{-1}$. Its dual is also box-constrained QP:

$$\begin{array}{ll} \text{minimize} & (b+u)^TZ(b+u) \\ \text{subject to} & \|u\|_\infty \leq \rho, \end{array}$$

in the variable $u \in \mathbf{R}^{(n-1)}$. At the optimum for this QP, we must have:

$$x = -\frac{1}{(c+\rho)} Z(b+u), \quad \text{and} \quad y = \frac{1}{(c+\rho)} + \frac{1}{(c+\rho)^2}(b+u)^T Z(b+u).$$

and we iterate as above.

Smooth optimization for box-constrained QPs. The two block-coordinate methods detailed in this section both amount to solving a sequence of box-constrained quadratic program of the form:

$$\begin{array}{ll} \text{minimize} & x^T A x + b^T x \\ \text{subject to} & \|x\|_\infty \leq \rho, \end{array} \tag{3.8}$$

in the variable $x \in \mathbf{R}^n$. The objective function has a Lipschitz continuous gradient with constant $L = 4c\lambda^{\max}(A)\sqrt{n}$ on the box $\mathcal{B} = \{x \in \mathbf{R}^n : \|x\|_\infty \leq \rho\}$, where we can define a prox function $(1/2)\|x\|^2$ which is strongly convex with constant 1 and bounded above by $(1/2)n\rho^2$ on $\mathcal{B}$. From [16] or [18], we know that solving (3.8) up to a precision ϵ will require at most $2n^{0.75}\sqrt{2\rho^3\lambda^{\max}(A)}/\sqrt{\epsilon}$ iterations of the first-order method detailed in [16], with each iteration equivalent to a matrix-vector product and a projection on the box $\mathcal{B}$. This means that the total complexity of solving (3.8) is given by:

$$O\left(n^{2.75}\sqrt{\frac{\rho^3\lambda^{\max}(A)}{\epsilon}}\right)$$

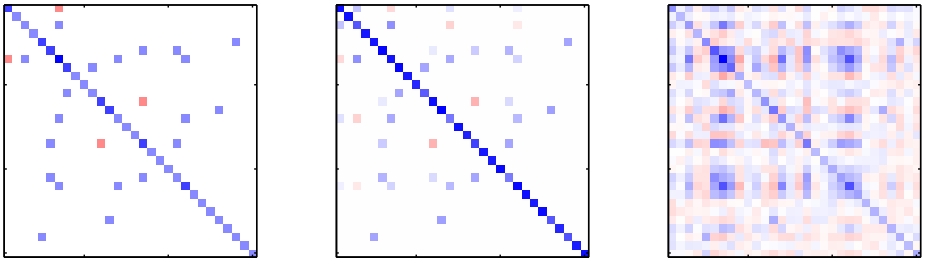
Complexity estimate. Following [15], with block coordinate descent corresponding to coordinate descent with the almost cyclic rule and using the fact that $\log\det(X)$ satisfies the strict convexity assumptions in [15, A2], we can show that the convergence rate of the block coordinate descent method is at least linear. Each iteration requires solving a box-constrained QP and takes $O(n^3 log(1/\epsilon))$ operations using an interior point solver or $O(n^{2.75}/\sqrt{\epsilon})$ using the optimal first-order scheme in [16]. We cannot use the same argument to show convergence of block coordinate ascent but empirical performance is comparable. In practice we have found that a small number of sweeps through all columns, independent of problem size n, is sufficient for convergence.

Implementation. The block coordinate descent methods implemented here correspond to coordinate descent using the almost cyclic rule, alternative row/column selection rules could improve the convergence speed. Also, each iteration of the block coordinate descent method corresponds to two rank-two updates of the inverse matrix, hence the cost of maintaining the inverse submatrix using the Sherman Woodbury Morrison formula is only $O(n^2)$.

4. Numerical results. In this section we test the performance of the methods detailed above on some randomly generated examples. We first form a sparse matrix A with a diagonal equal to one and a few randomly chosen, nonzero off-diagonal terms equal to +1 or -1. We then form the matrix:

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

where $V \in \mathbf{S}_n$ is a symmetric, i.i.d uniform random matrix. Finally, we make B positive definite by shifting its eigenvalues, and use this noisy, random matrix to test our covariance selection methods.



Original inverse A | Solution for $\rho = 0.5$ | Noisy inverse B^{-1}

FIG. 4.1. *Recovering the sparsity pattern. We plot the original inverse covariance matrix A, the solution to problem (2.1) and the noisy inverse B^{-1}.*

In Figure 4.1, we plot the sparsity patterns of the original inverse covariance matrix A, the solution to problem (2.1) and the noisy inverse B^{-1} in a randomly generated example with $n = 30$, $\sigma = 0.15$ and $\rho = 0.5$. In Figure 4.2 we represent the dependence structure of interest rates (sampled over a year) inferred from the inverse covariance matrix. Each node represents a particular interest rate maturity and the nodes are linked if the corresponding coefficient in the inverse covariance matrix is nonzero (i.e. they are conditionally dependent). We compare the solution to problem (2.1) on this matrix for $\rho = 0$ and $\rho = 0.1$ and notice that in the sparse solution the rates appear clearly clustered by maturity.

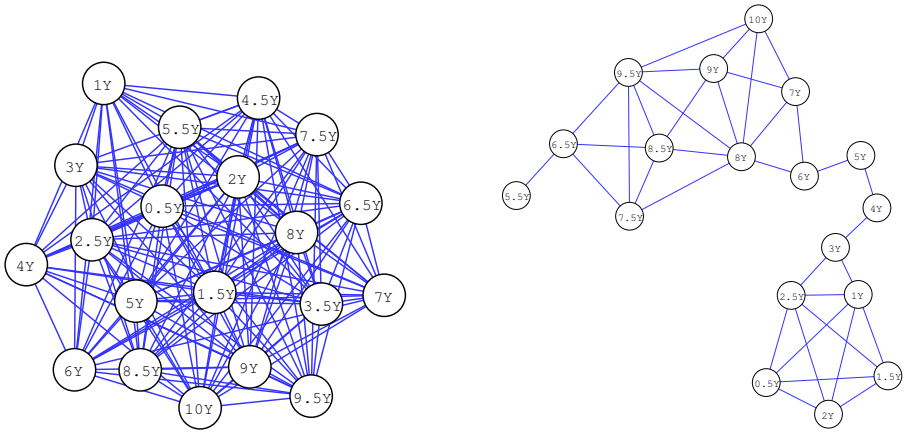


FIG. 4.2. *We plot the network formed using the solution to problem (2.1) on an interest rate covariance matrix for $\rho = 0$ (left) and $\rho = 0.1$ (right). In the sparse solution the rates appear clearly clustered by maturity.*

In Figure 4.3, we study computing times for various choices of algorithms and problem sizes. On the *left*, we plot CPU time to reduce the duality gap by a factor 10^{-2} versus problem size n, on randomly generated problems, using the coordinate descent code and the optimal first-order for solving box QPs. On the *right*, we plot duality gap versus CPU time for both smooth minimization and block-coordinate algorithms for a randomly generated problem of size $n = 250$. For the smooth minimization

code, we set $\alpha = 1/\lambda^{\max}(B)$ and we plot computing time for both $\beta = 1/(2\lambda^{\min}(B))$ (Smooth. Opt. 1/2) and $\beta = 2/\lambda^{\min}(B)$ (Smooth. Opt. 2).

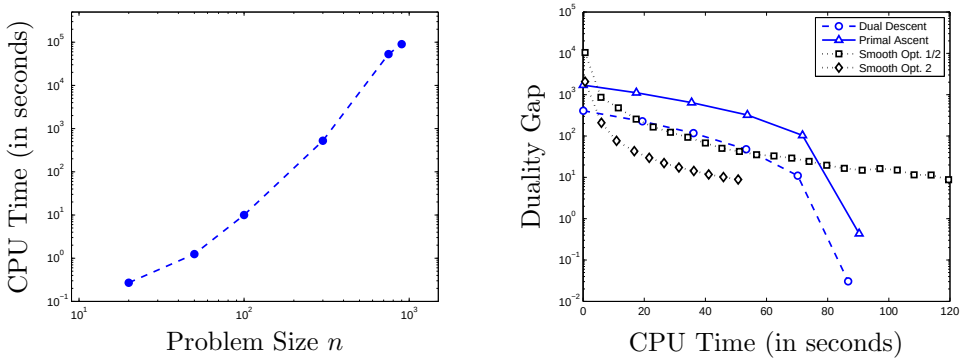


FIG. 4.3. *Computing time.* Left*: We plot CPU time to reduce the duality gap by a factor* 10^{-2} *versus problem size* n*, on randomly generated problems, using the coordinate descent code and the optimal first-order algorithm for solving box QPs.* Right*: We plot duality gap versus CPU time for both smooth minimization and block-coordinate algorithms, for a problem of size* $n = 250$*.*

Acknowledgments.. The authors would like to thank Francis Bach, Peter Bartlett and Martin Wainwright for enlightening discussions on the topic. We would also like to acknowledge financial support from NSF grant DMS-0625352, EUROCONTROL grant C20083E/BM/05 and a gift from Google, Inc.

REFERENCES

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

[2] J. A. Bilmes, *Natural statistic models for automatic speech recognition*, Ph.D. thesis, UC Berkeley, Dept. of EECS, CS Division, (1999).

[3] ———, *Factored sparse inverse covariance matrices*, IEEE International Conference on Acoustics, Speech, and Signal Processing, (2000).

[4] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[5] 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.

[6] J. Dahl, V. Roychowdhury, and L. Vandenberghe, *Maximum likelihood estimation of gaussian graphical models: numerical implementation and topology selection*, UCLA preprint, (2005).

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

[8] 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.

[9] A. Dobra and M. West, *Bayesian covariance selection*, ISDS Working paper, (2004).

[10] 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.

[11] 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.

[12] J. Z. Huang, N. Liu, and M. Pourahmadi, *Covariance selection and estimattion via penalized normal likelihood*, Wharton Preprint, (2005).

[13] B. Jones, C. Carvalho, C. Dobra, A. Hans, C. Carter, and M. West, *Experiments in stochastic computation for high-dimensional graphical models*, ISDS Discussion Paper 04-01, (2004).

[14] C. Lemaréchal and C. Sagastizábal, *Practical aspects of the Moreau-Yosida regularization: theoretical preliminaries*, SIAM Journal on Optimization, 7 (1997), pp. 367–385.

[15] Z. Q. Luo and P. Tseng, *On the convergence of the coordinate descent method for convex differentiable minimization*, Journal of Optimization Theory and Applications, 72 (1992), pp. 7–35.

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

[17] ———, *Introductory Lectures on Convex Optimization*, Springer, 2003.

[18] ———, *Smooth minimization of nonsmooth functions*, Mathematical Programming, Series A, 103 (2005), pp. 127–152.

[19] R. Tibshirani, *Regression shrinkage and selection via the LASSO*, Journal of the Royal statistical society, series B, 58 (1996), pp. 267–288.