Jñānābha, Vol. 41, 2011

CONVEX OPTIMIZATION TECHNIQUES DUE TO NESTROV AND COMPUTATIONAL COMPLEXITY

By

Paras Bhatnagar and Prashant Chauhan

Department of Mathematics, College of Engineering and Technology (IETM)
Moradabad, Uttar Pradesh, India

(Received : May 22, 2010*)*

ABSTRACT

A large body of literature is devoted to the estimation of covariance matrices in a large-scale setting. Recent work in this area includes the shrinkage approach proposed by Schäfer and Strimmer [16], where the authors analystically calculate the optimal shrinkage intensity, yielding a good, computationally inexpensive estimate. Our focus is an estimate with the property that the corresponding inverse covariance matrix is sparse.

Dempster [4] introduced the concept of covariance selection, where the number of parameters to be estimated is reduced by setting to zero some elements of the inverse covariance matix. Covariance selection can lead to a more robust estimate of S if enough entries of its inverse are set to zero. Traditionally, a greedy forward/backward search algorithm is employed to determine the zero patterns Lauritzen [9]. However, this method quickly becomes computationally infeasible as p grows.

2000 Mathematics Subject Classification : Primary 90C25; Secondary 90C47, 49K30

Keywords and Phrases : Applications of Mathematical Optimization, Robustness, Duality and Bounds, Convergence and Property of Solution, Convex Optimization, Nesterov's Method.

1. Introduction. In this paper we investigate the following related idea. Beginning with a dense empirical covariance matrix S, we compute a maximum likelihood estimate of S with an II-norm penalty added to encourage sparsity in the inverse. The authors Li and Gui [10] introduce a gradient descent algorithm in which they account for the sparsity of the inverse covariance matrix by defining a loss function that is the negative of the log likelihood function. Recently, Huang, Liu and Pourahmodi [7], and Dahl [2] considered penalized maximum liklihood estimation, and Dahl [2] in particular, proposed a set of large scale methods to solve problems where a sparse structure of S-1 is known a priori. Our contribution is threefold: we present a provably convergent algorithm that is efficient for large-

scale instances, yielding a sparse, invertible estimate of S-1, even for $n<p$; we obtain some basic complexity estimates for the problem; and finally we test our algorithm on synthetic data as well as gene expression data from two datasets.

Notations. For a $p \times p$ matrix X,$X \geq 0$ means X is symmetric and positive semi-definite; $\|x\|$ denotes the largest singular value norm, $\|x\|_1$ the sum of the absolute values of its elements, and $\|x\|_\infty$ their largest magnitude.

2. Preliminaries. In this section we set up the problem and discuss some of its properties.

2.1. Problem Setup. Let $S \geq 0$ be a given empirical matrix, for data drawn from a multivariate Gaussian distribution. Let the variable X be our estimate of the inverse covariance matrix. We consider the penalized maximum-liklihood problem

$$max \log det X - (X,X) - \rho \|x\|_1 \qquad ...(1.1)$$

where $\langle S,X \rangle = trace(SX)$ denotes the scalar product between two symmetric matrices S and X, and the term $\|x\|_1 = \sum_{i,j} |X_{ij}|$ penalizes nonzero elements of X.

Here, the scalar parameter $\rho > 0$ controls the size of the penalty, hence the sparsity of the solution. The penalty term involving the sum of absolute values of the entries of X is a proxy for the number of its nonzero elements, and is often used-albeit with vector, not matrix, variables-in regression techniques, such as *LASSO* in Tibshirani [17], when sparsity of the solution is a concern.

The classical maximum likelihood estimate of Σ is recovered for $\rho > 0$, and is simply S, the empirical covariance matrix. Furthermore, as noted above, for $p >> n$, the matrix S is likely to be singular. It is desirable for our estimate of S to be invertible. We shall show that our proposed estimator performs some regularization, so that our estimate is invertible for every $\rho > 0$.

2.2. Robustness, Duality and Bounds. By introducing a dual variable U, we can write (1) as $\max_{X>0} \min_{\|U\|_\infty \leq \rho} \log det X + \langle X, S+U \rangle$.

Here $\|U\|_\infty$ denotes the maximal absolute value of the entries of U. This corresponds to seeking an estimate with maximal worst-case likelihood, over all component wise bounded additive perturbations $S+U$ of the empirical covariance matrix S. Such a "robust optimization" interpretation can be given to a number of estimation problems, most notably support vector machines for classification.

We can obtain the dual problem by exchanging the *max* and the *min*:

$$\min_U \{ -log\, det(S+U) - p : \|U\|_\infty \leq \rho, S+U > 0 \} \qquad (2)$$

The diagonal elements of an optimal U are simply $\hat{U}_{ii} = \rho$. The

corresponding covariance matrix estimate is $\sum := S + \hat{U}$. Since the above dual problem has a compact feasible set, the primal and dual problems are equivalent. The optimality conditions relate the primal and dual solutions by $\sum X = 1.$

The following theorem shows that adding the l_1-norm penalty regularizes the solution.

Theorem-1 For every $\rho > 0$, the optimal solution to the penalized *ML* problem (1) is unique, and bounded as follows :

$$\alpha(p) I \le X \le \beta(p) I, \text{ where } \alpha(p) = \frac{1}{\|S\| + \rho p}, \beta(p) = \frac{p}{\rho}.$$

Proof. An optimal X satisfies $X=(S+U)^{-1}$, where $\|U\|_\infty \le \rho$. Thus, we can without loss of generality impose that $X \ge \alpha(p) I$, where $\alpha(p)$ is defined in the theorem. Likewise, we can show that X is bounded above. Indeed, at optimum, the primal-dual gap is zero:

$$\begin{aligned} 0 &= -\log\det(S+P) - p - \log\det X + \langle S, X\rangle + \rho\|X\|_1 \\ &= -p + \langle S, X\rangle + \rho\|X\|_1, \end{aligned}$$

where we have used $(S+U)X=1$. Since S,X are both positive semi-definite, we obtain $\|X\| \le \|X\|_F \le \|X\|_1 \le \beta(p) I$ as claimed. Problem (2) is smooth and convex. When $p(p+1)/2$ is in the low hundreds, the problem can be solved by existing software that uses an interior point method Vandenberghe [18], The complexity to compute an ε-suboptimal solution using such-second-order methods, however, is $O\left(p^6 \log(1/\varepsilon)\right)$ making them infeasible for even moderately large p.

The authors Dahl et al. [2] developed a set of algorithms to estimate the nonzero entries of $\sum^{-1}$ when the sparsity pattern is known a priori and corresponds to an undirected graphical model that is not chordal. Here our focus is on relatively large, dense problems, for which the sparsity pattern is not known a priori. Note that we cannot expect to do better than $O(p^3)$, which is the cost of solving the non-penalized problem $\rho=0$ for a dense sample covariance matrix S.

2.3 Choice of Regularization Parameter ρ. In this section we provide a simple heuristic for choosing the penality parameter ρ, based on hypothesis testing, We emphasize that while the choice of ρ is an important issue that deserves a thorough investigation, It is not the focus of this paper.

The heuristic is based on the observation that if $\rho < |S_{ij}|$ then there cannot be zero

in that element of our estimate of the covariance matrix $\Sigma_{ij} \neq 0$ suppose we choose ρ according to

$$\rho = \frac{t_{n-2}(\gamma) max_{i,j} S_{ii} S_{ji}}{\sqrt{n-2+t^2_{n-2}(\gamma)}}, \tag{3}$$

where $t_{n-2}(\gamma)$ denotes the two-tailed 100γ% point of the t-distribution, for n-2 degrees of freedom. With this choice, and using the fact that $S \geq 0$, it can be shown that $\rho < |S_{ij}|$ implies the condition for rejecting the null hypothesis that variables i and j are independent in the underlying distribution, under a likelihood ratio test of size γ Muirhead [13]. We note that this choice yields an asymtotically consistent estimator. As $n \to \infty$, we recover the sample covariance S as our estimate of the covariance matrix, and S converges to the true covariance Σ.

3. Block Coordinate Descent Methd. In this section we present an efficient algorithm for solving the dual problem (2) based on block coordinate descent.

3.1 Algorithm. We first describe a method for solving (2) by optimizing over one column and row of $S+U$ at a time. Let $W:=S+U$ be our estimate of the true covariance. The algorithm begins by initializing $W^0=S+\rho I$. The diagonal elements of W^0 are set to their optimal values, and are left unchanged in what follows.

We can permute rows and columns of W, so that we are optimizing over the last column and row. Partition W and S as

$$W = \begin{pmatrix} W_{11} & w_{12} \\ w_{12}^T & w_{22} \end{pmatrix} \qquad S = \begin{pmatrix} S_{11} & s_{12} \\ s_{12}^T & s_{22} \end{pmatrix}$$

where $w_{12}, s_{12} \in R^{p-1}$ the update rule is found by solving the dual problem (2), with U fixed except for its last column and row. This leads to a box-constrained quadratic program (QP):

$$\hat{w}_{12} = arg\,min \left\{ y^T W_{11}^{-1} y : \|y - s_{12}\|_\infty \leq \rho \right\} \tag{4}$$

We cycle through the columns in order, solving a QP at each step. After each sweep through all columns, we check to see if the primal-dual gap is less than ε, a given tolerance. The primal variable is related to W by $X=W^{-1}$. The duality gap condition is then

$$\langle S, X \rangle + \rho \|X\|_1 \leq p + \varepsilon .$$

3.2. Convergence and Property of Solution. Iterates produced by the coordinate descent algorithm are strictly positive definite. Indeed, since $S \geq 0$, we have that $W^0 > 0$ for any $\rho > 0$. Now suppose that, at iteration k, $W > 0$. This implies

that the following Schur complement is positive : $w_{22} - w_{12}^T W_{11}^{-1} w_{12} > 0$. By the update rule (4), we have

$w_{22} - \hat{w}_{12} W_{11}^{-1} \hat{w}_{12} > w_{22} - w_{12}^T W_{11}^{-1} w_{12} > 0$,

which, using Schur complements again, implies that the new iterate satisfies $\hat{W} > 0$.

Note that since the method generates a sequence of feasible primal and dual points, the stopping criterion is nonheuristic. As a consequence, the *QP* (4) to be solved at each iteration has a unique solution. This implies that the method converges to the true solution of (2), by virtue of general results on block-coordinate descent algorithms Bertsekas [1].

The above results shed some interesting light on the solution to problem (2). Suppose that the column s_{12} of the sample covariance satisfies $|s_{12}| \leq \rho$, where the inequalities hold component wise. Then the corresponding column of the solution is zero $\Sigma_{12} = 0$. Indeed, if the zero vector is in the constraint set of the *QP* (4), then it must be the solution to that *QP*. As the constraint set will not change no matter how many times we return to that column, the corresponding column of all iterates will be zero. Since the iterates converge to the solution, the solution must have zero for that column. This property can be used to reduce the size of the problem in advance, by setting to zero columns of *W* that correspond to columns in the sample covariance *S* that meet the above condition.

Using the work of Luo and Tseng [11], it is possible to show that the local convergence rate of this method is at least linear. In practice we have found that a small number of sweeps through all columns, independent of problem size *p,* is sufficient to achieve convergence. For a fixed number of *K* sweeps, the cost of the method is $O(Kp^4)$, since each iteration costs $O(p^3)$.

3.3 Connection to *LASSO*. The dual of (4) is

$$\min_x x^T W_{11} x - s_{12}^T + \rho \|x\|_1 \ . \tag{5}$$

Strong duality obtains so that problems (5) and (4) are equivalent. If we let Q denote the square root of $W_{11}, b := \frac{1}{2} Q^{-1}$, then we can write(5) as $\min_x \|Qx - b\|_2^2 + \rho \|x\|_1$

The above is a penalized least-squares problem, often referred to as *LASSO*. If W_{11} were a principal minor of the sample covariance *S,* then the above would be equivalent to a penalized regression of one variable against all others. Thus, the approach is reminiscent of the approach explored by Meinshausen and Buhlmann [12]. but there are two major differences. First, we begin with some regularization, and as a consequence, each penalized regression problem has a unique solution. Second, and more importantly, we update the problem data after each regression;

in particular, W_{11} is never a minor of S. In a sense, the coordinate descent method can be interpreted as a recursive *LASSO* method.

4. Nesterov's Method. In this section, we apply the recent results due to Nesterov [15] to obtain a first-order method for solving (1). Our main goal is not to obtain another algorithm, as we have found that the coordinate descent is already quite efficient; rather, we seek to use Nesterov's formalism to derive a rigorous complexity estimate for the problem, improved over that delivered by interior point methods.

As we shall see, Nesterov's framework allows us to obtain an algorithm that has a complexity of $O\left(p^{4.5}/\varepsilon\right)$, where $\varepsilon>0$ is the desired accuracy on the objective of problem (1). This is to be contrasted with the complexity of interior-point methods, $O\left(p^6\, log\left(1/\varepsilon\right)\right)$. Thus, Nesterov's method provides a much better dependence on problem size, at the expense of a degraded dependence on accuracy. In our opinion, obtaining an estimate that is accurate numerically up to dozens of digits has little practical value, as it is much more important to be able to solve larger problems with less accuracy. Note also that the memory requirements for Nesterov's methods are much better than those of interior-point methods.

4.1 Idea of Nesterov's Method. Nesterov's method [15] applies to a class of non-smooth, convex optimization problems, of the form

$$\min_x \left\{ f(x) : x \in Q_1 \right\} \tag{6}$$

where the objective function is described as

$$f(x) = \hat{f}(x) + \max_u \left\{ \langle Ax, u \rangle_2 : u \in Q_2 \right\}.$$

Here Q_1 and Q_2 are bounded, closed, convex sets, $\hat{f}(x)$ is differentiable (with Lipschitz continuous gradient) and convex on Q_1, and A is a linear operator. Observe that we can write (1) in this form if we impose bounds on the eigenvalues of the solution, X. To this end, we let

$$Q_1 = \left\{ X : \alpha I \le X \le \beta I \right\},$$

$$Q_2 = \left\{ U : \|U\|_\infty \le \rho \right\},$$

where $\alpha, \beta \left(0 < \alpha < \beta\right)$ are given. We also define

$$\hat{f}(X) = -log\, det\, X + \langle S, X \rangle, \text{ and } A = \rho I.$$

To Q_1 and Q_2, we associate norms and continuous, strongly convex functions, called prox-functions, $d_1(X)$ and $d_2(U)$. For Q_1 we choose the Frobenius norm, and a prox-function $d_1(X) = -log\, det\, X + log\, \beta$. For Q_2, we choose the Frobenius norm again, and a prox-function $d_2(U) = \|U\|_F^2 / 2$.

The method applies a smoothing technique to the non-smooth problem (6), which replaces the objective of the original problem, $f(X)$, by a penalized function involving the prox-function $d_2(U)$:

$$\tilde{f}(X) = \hat{f}(X) + \underset{U \in Q_2}{mix}\left\{\langle AX, U\rangle - \mu d_2(U)\right\}. \tag{7}$$

The above function turns out to be a smooth uniform approximation. It is differentiable, convex on Q_1, and has a Lipschitz-continuous gradient, with a constant L that can be computed as detailed below. A specific gradient scheme is then applied to this smooth approximation, with convergence rate $O(L/\varepsilon)$.

4.2 Algorithm and Complexity Estimate. To detail the algorithm and compute the complexity, we first calculate some parameters corresponding to our definitions above. First, the strong convexity parameter for $d_1(X)$, on Q_1 is $\sigma_1 = 1/\beta^2$, in the sense that $\nabla^2 d_1(X)[H,H] = trace\left(X^{-1}HX^{-1}H\right) \geq \beta^{-2}\|H\|_F^2$ for every symmetric H. Furthermore, the center of the set Q_1 is $X_0 = arg\,min_{X \in Q_1} d_1(X) = \beta I$, and satisfies $d_1(X_0) = 0$. Without choice, we have $D_1 = max_{x \in Q_1} d_1(X) = \rho log\,\beta/\alpha$.

Similary, the strong convexity parameter for $d_2(U)$ on Q_2 is $\sigma_2 = 1$ and we have $D_2 = max_{U \in Q_2} d_2(U) = p^2/2$. With this choice, the center of the set Q_2 is $U_0 = arg\,min_{U \in Q_2} d_2(U) = 0.$

For a desired accuracy ε, we set the smoothness parameter $\mu = \varepsilon/2D_2$, and start with the initial point $X_0 = \beta I$. The algorithm proceeds as follows.

For $k \geq 0$ do

1. Compute $\nabla\tilde{f}(X_k) = -X_k^{-1} + S + U^*(X_k)$ where $U^*(X)$ solves (7).

2. Find $Y_k = arg\,min_\gamma \left\{\left\langle \nabla\tilde{f}(X_k), Y - X_k\right\rangle + \frac{1}{2}L(\varepsilon)\|Y - X_k\|_F^2 : X \in Q_1\right\}$.

3. Find $Z_k = arg\,min_X \left\{\frac{L(\varepsilon)}{\sigma_1}d_1(X) + \sum_{i=0}^{k}\frac{i+1}{2}\left\langle \nabla\tilde{f}(X_i), X - X_i\right\rangle : X \in Q_1\right\}$.

4. Update $X_k = \frac{2}{k+3}Z_k + \frac{k+1}{k+3}Y_k$.

In our case, the Lipschitz constant for the gradient of our smooth approximation to the objective function is $L(\varepsilon) = M + D_2\|A\|^2/(2\sigma_2\varepsilon)$, where

$M = 1/\alpha^2$ is the Lipschitz constant for the gradient of $\tilde{f}$, and the norm $\|A\|$ is induced by the Frobenius norm, and is equal to ρ. The algorithm is guaranteed to produce an ε-suboptimal solution after a number of steps not exceeding

$$N(\varepsilon) = 4\|A\|\sqrt{\frac{D_1 D_2}{\sigma_1\sigma_2}}\frac{1}{\varepsilon} + \sqrt{\frac{MD_1}{\sigma_1\varepsilon}} \tag{8}$$

$$= \frac{k\sqrt{p(\log k)}}{\varepsilon}\left(4p\alpha + \sqrt{\varepsilon}\right),$$

where $k = \beta/\alpha$ is bound on the condition number of the solution.

Now we are ready to estimate the complexity of the algorithm. For step 1, the gradient of the smooth approximation is readily computed in closed form, via the computation of the inverse of *X*. Step 2 essentially amounts to projecting on Q_1, and requires an eigen value problem to be solved; likewise for step 3. In fact, each iteration costs $O(p^3)$. The number of iterations necessary to achieve an objective with absolute accuracy less than ε is given in (8) by, $N(\varepsilon) = O(p^{15}/\varepsilon)$, if the condition number *k* is fixed a priori. Thus, the Complexity of the algorithm is $O(p^{4.5}/\varepsilon)$.

5. Numerical Results. In this section we present some numerical results. We begin with a small synthetic example to test the ability of the method to recover a sparse structure from a noisy matrix. Starting with a sparse matrix *A,* we obtain *S* by adding a uniform noise of magnitude $\sigma = 0.1$ to A^{-1}.

In figure 1 we plot the sparsity patterns of *A,* S^{-1}, and the solution $\hat{X}$ to (1) using *S* and $\rho = \sigma$.

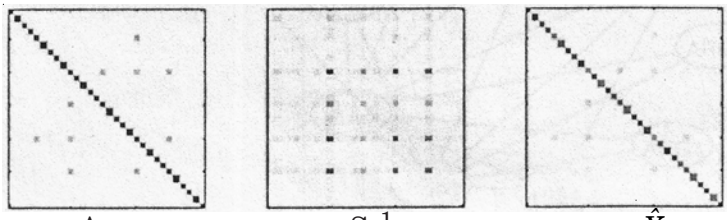

Figure 1. Recovering the Sparsity Pattern

We plot the underlying sparse matrix *A,* the inverse of the noisy version of A^{-1}, and the solution to problem (1) for ρ equal to the noise level.

We next perform the following experiment to see what happens to the solution of (1) as we vary the parameter ρ above and below the noise level σ. For each value of

ρ, we randomly generate 10 sparse matrices A of size n=50. We then obtain sample covariance matrices S as above, again using σ=0.1.

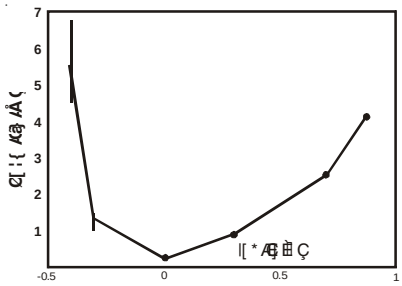

Figure 2. Recovering Structure

Gene Expression Properties

The 300 experiment compendium dataset contains n=253 samples with p=6136 variables. With a view towards obtaining a very sparse graph, we set γ=0.1 in the heuristic formula (3) of section (2.3) to obtain ρ=0.0313.

Applying the property of the solution discussed in section (3.2), the size of the problem was reduced to $\hat{p}$ =537. Three sweeps through all columns were required to achieve a duality gap of $\varepsilon = 0.146$, with a total computing time of 18-minutes, 34-seconds. The resulting estimate of the inverse covariance matrix $\hat{\Sigma}^{-1}$ is 99% sparse and has a condition number of 21.84. Figure (4) shows a sample subgraph obtained from $\hat{\Sigma}^{-1}$, generated using the Graph Explore program developed by Dobra and West [6]. The method has picked out a cluster of genes associated with amino acidmetabolism, as described by Hughes et al. [7].

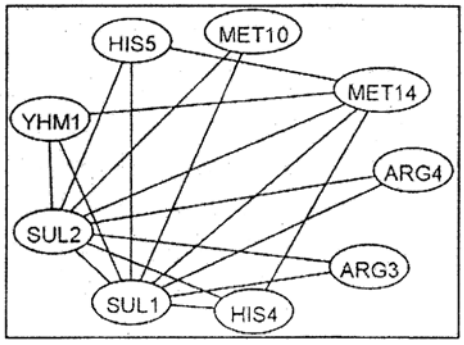

Figure 3. Application to Hughes Dataset Using ρ=0.0313.

As we have seen, the penalized maximum likelihood problem formulated here is useful for recovering a sparse underlying precision matrix Σ^{-1} from a dense sample covariance matrix S, even when the number of samples n is small relative to the number of variables p. In preliminary tests, the method appears to be a potentially valuable tool for analyzing gene expression data, although further testing is

required.

REFERENCES

[1] D. Bertsekas, Non linear programming, *Athena Scientific* (1998).

[2] J. Dahl., V. Roycho Wdhury and L. Vandenberghe, *Maximum Likelihood Estimation of Gaussian Graphical Models : Numerical Implementation and Topology Selection,* UCLA Preprint.

[3] A. D' Aspremount, L. El Ghaoui, M. Jordan and G.R.G. Lanckriet, A direct formulation for sparse PCA using semide. nite programming. *Advances in Neural Information Processing Systems,* **17** (2004).

[4] A.P. Dempster, Covariance selection, *Biometrics,* **28** (1972), 157-175.

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

[6] A. Dobra and M. West, Bayesian covariance selection, Working paper, *ISDS, Duke University*, 2004.

[7] I.Z., Huang, N., Liu, and M. Pourahmadi, Covariance selection and estimation via penalized normal likelihood, *Wharton Preprint*, 2005.

[8] S.H. Functional discovery via a compendium of expression *Pro. Les. Cell,* **102** (2000), 109-126.

[9] S. Laurizen, *Graphical Models*. Springer Verlag, 1996.

[10] H. Li and I. Gui, Gradient directed regularization for sparse Gaussian concentration graphs, with applications to inference of genetic networks, *University of Pennsylvania Technical Report,* 2005.

[11] 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) 7-35.

[12] N. Meinshausen and P. Buhlmann. High dimensional graphs and variable selection with the lasso, *Annals Statistics*, 2005 (in press)

[13] R.J. Muirhead, *Aspcts of Multivariate Statistical Theory,* John Wiley and Sons, Inc, 1982.

[14] G. Natsoulis, L., El Ghaoui, G. Lanckriet, A. Tolley, F. Leroy, S. Dunlea, B. Eynon, C. Pearson, S. Tugendreich and K. Jarnagin, Classification of a large microarray data set: algorithm comparison and analysis of drug signatures, *Genome Research,* **15** (2005), 724-736.

[15] Y., Nesterov, Smooth minimization of non-smooth functions, *Math. Prog. Ser. A,* **103** (2005), 127-152.

[16] J., Schafer, and K. Strimmer, A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics, *Statistical Applications in Genetics and Molecular Biology,* **4** (2005).

[17] R. Tibshirani, Regression shrinkage and selection via the lasso, *Journal Royal Statistical Society, Series B,* **58** (1996).

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