
Convex Optimization Techniques for Fitting Sparse Gaussian Graphical Models

Onureena Banerjee

Laurent El Ghaoui

EECS Dept., UC Berkeley, Berkeley, CA 94720

Alexandre d’Aspremont

ORFE Dept., Princeton University, Princeton, NJ 08544

Georges Natsoulis

Iconix Pharmaceuticals, Mountain View, CA 94043

ONUREENA@EECS.BERKELEY.EDU

ELGHAOUI@EECS.BERKELEY.EDU

ASPREMON@PRINCETON.EDU

GNATSOULIS@ICONIXPHARM.COM

Abstract

We consider the problem of fitting a large-scale covariance matrix to multivariate Gaussian data in such a way that the inverse is sparse, thus providing model selection. Beginning with a dense empirical covariance matrix, we solve a maximum likelihood problem with an l_1 -norm penalty term added to encourage sparsity in the inverse. For models with tens of nodes, the resulting problem can be solved using standard interior-point algorithms for convex optimization, but these methods scale poorly with problem size. We present two new algorithms aimed at solving problems with a thousand nodes. The first, based on Nesterov’s first-order algorithm, yields a rigorous complexity estimate for the problem, with a much better dependence on problem size than interior-point methods. Our second algorithm uses block coordinate descent, updating row/columns of the covariance matrix sequentially. Experiments with genomic data show that our method is able to uncover biologically interpretable connections among genes.

1. Introduction

The estimation of large-scale covariance matrices from data is a common problem, with applications in many fields, ranging from bioinformatics to finance. For

jointly Gaussian data, this problem is equivalent to model selection among undirected Gaussian graphical models. Such models, sometimes called concentration graphs (or gene relevance networks in bioinformatics), have been shown to be valuable for evaluating patterns of association among variables (see (Dobra et al., 2004), for example). Zeros in the inverse covariance matrix correspond to conditional independence properties among the variables, as well as to missing edges in the associated graphical model. In this setting, a sparse inverse covariance matrix, if it fits the data well, is very useful to practitioners, as it simplifies the understanding of the data. Sparsity is also often justified from a statistical viewpoint, as it results in a more parsimonious, and also more robust, model.

Estimating an underlying $p \times p$ covariance matrix Σ becomes a non-trivial task when p is large. In analyzing multivariate data, the empirical covariance matrix S , the maximum likelihood estimate, is often used. However, when the number of samples n is small relative to p , this matrix cannot be considered a good estimate of the true covariance. Furthermore, for $n \ll p$, the empirical covariance S is singular so that we cannot even access information about all conditional independencies.

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 & Strimmer, 2005), where the authors analytically 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.

Appearing in *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, 2006. Copyright 2006 by the author(s)/owner(s).

Dempster (Dempster, 1972) 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 matrix. Covariance selection can lead to a more robust estimate of Σ if enough entries of its inverse are set to zero. Traditionally, a greedy forward/backward search algorithm is employed to determine the zero pattern (Lauritzen, 1996). However, this method quickly becomes computationally infeasible as p grows.

Alternatively, the set of neighbors of any particular node in the graph may be found by regressing that variable against the remaining $p-1$ variables. This has been explored successfully by (Dobra & West, 2004; Dobra et al., 2004), who use a stochastic algorithm to manage tens of thousands of variables. In (Meinshausen & Bühlmann, 2005), the authors have studied a GGM inference technique using the LASSO of (Tibshirani, 1996), in which an l_1 -norm penalty is added to each regression problem to make the graph as sparse as possible.

In this paper we investigate the following related idea. Beginning with a dense empirical covariance matrix S , we compute a maximum likelihood estimate of Σ with an l_1 -norm penalty added to encourage sparsity in the inverse. The authors of (Li & Gui, 2005) 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 et al., 2005; Dahl et al., 2005) considered penalized maximum likelihood estimation, and (Dahl et al., 2005) in particular, proposed a set of large scale methods to solve problems where a sparse structure of Σ^{-1} is known a priori. Here, we will not make this assumption, and instead try to discover structure (the zero pattern) as we search for a regularized estimate.

Our contribution is threefold: we present a provably convergent algorithm that is efficient for large-scale instances, yielding a sparse, invertible estimate of Σ^{-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.

The paper is organized as follows: we specify the problem and outline some of its basic properties (section 2); we present a convergent algorithm based on block coordinate descent (section 3) and make a connection to the LASSO. We describe how one can apply a recent methodology for convex optimization due to Nesterov (Nesterov, 2005), and obtain as a result a computational complexity estimate that has a much better

dependence on problem size than interior-point algorithms (section 4). In section 5 we present the results of some numerical experiments comparing these two algorithms, involving in particular gene expression data. Finally in section 6 we briefly state our conclusions.

Notation

For a $p \times p$ matrix X , $X \succeq 0$ means X is symmetric and positive semi-definite; $\|X\|$ denotes the largest singular value norm, $\|X\|_1$ the sum 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 $S \succeq 0$ be a given empirical covariance 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-likelihood problem

$$\max_{X \succ 0} \log \det X - \langle S, X \rangle - \rho \|X\|_1 \quad (1)$$

where $\langle S, X \rangle = \text{trace}(SX)$ 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, 1996), when sparsity of the solution is a concern. The authors of (d’Aspremont et al., 2004), have used a similar penalization approach for sparse principal component analysis.

The classical maximum likelihood estimate of Σ is recovered for $\rho = 0$, and is simply S , the empirical covariance matrix. Due to noise in the data, however, S may not have a sparse inverse, even if there are many conditional independence properties in the underlying distribution. Since we strike a trade-off between maximality of the likelihood and the number of non-zero elements in the inverse covariance matrix, our approach is potentially useful for discovering conditional independence properties. Furthermore, as noted above, for $p \gg n$, the matrix S is likely to be singular. It is de-

sirable for our estimate of Σ 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

$$\max_{X \succ 0} \min_{\|U\|_\infty \leq \rho} \log \det X + \langle X, S + U \rangle,$$

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 componentwise 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:

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

The diagonal elements of an optimal U are simply $\hat{U}_{ii} = \rho$. The corresponding covariance matrix estimate is $\hat{\Sigma} := S + \hat{U}$. 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 $\hat{\Sigma}X = I$.

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: $\alpha(p)I \preceq X \preceq \beta(p)I$, where*

$$\alpha(p) := \frac{1}{\|S\| + \rho p}, \quad \beta(p) := \frac{p}{\rho}.$$

Proof: An optimal X satisfies $X = (S + U)^{-1}$, where $\|U\|_\infty \leq \rho$. Thus, we can without loss of generality impose that $X \succeq \alpha(p)I$, 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:

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

where we have used $(S + U)X = I$. Since S, X are both positive semi-definite, we obtain

$$\|X\| \leq \|X\|_F \leq \|X\|_1 \leq \beta(p)I,$$

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 (see (Vandenberghe et al., 1998) for example). The complexity to compute an ϵ -suboptimal solution using such second-order methods, however, is $O(p^6 \log(1/\epsilon))$, making them infeasible for even moderately large p .

The authors of (Dahl et al., 2005) developed a set of algorithms to estimate the nonzero entries of Σ^{-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 penalty 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. We include this here to clarify our numerical experiments of section 5.3.

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: $\hat{\Sigma}_{ij} \neq 0$. Suppose we choose ρ according to

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

where $t_{n-2}(\gamma)$ denotes the two-tailed 100 γ % point of the t-distribution, for $n-2$ degrees of freedom. With this choice, and using the fact that $S \succeq 0$, 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 (see (Muirhead, 1982) for example).

We note that this choice yields an asymptotically consistent estimator. As $n \rightarrow \infty$, 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 Method

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

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

where $w_{12}, s_{12} \in \mathbf{R}^{p-1}$. 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):

$$\hat{w}_{12} := \arg \min_y \{y^T W_{11}^{-1} y : \|y - s_{12}\|_\infty \leq \rho\} \quad (4)$$

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

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

3.2. Convergence and Property of Solution

The iterates produced by the coordinate descent algorithm are strictly positive definite. Indeed, since $S \succeq 0$, we have that $W^0 \succ 0$ for any $\rho > 0$. Now suppose that, at iteration k , $W \succ 0$. 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

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

which, using Schur complements again, implies that the new iterate satisfies $\hat{W} \succ 0$. 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, 1998).

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 componentwise. Then the corresponding column of the solution is zero: $\hat{\Sigma}_{12} = 0$. 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 & Tseng, 1992), 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

$$\min_x x^T W_{11} x - s_{12}^T x + \rho \|x\|_1 \quad (5)$$

Strong duality obtains so that problems (5) and (4) are equivalent. If we let Q denote the square root of W_{11} , and $b := \frac{1}{2} Q^{-1} s_{12}$, 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 & Bühlmann, 2005), 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, 2005) 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 will see, Nesterov's framework allows us to obtain an algorithm that has a complexity of $O(p^{4.5}/\epsilon)$, where $\epsilon > 0$ is the desired accuracy on the objective of problem (1). This is to be contrasted with the complexity of interior-point methods, $O(p^6 \log(1/\epsilon))$. 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 applies to a class of non-smooth, convex optimization problems, of the form

$$\min_x \{f(x) : x \in Q_1\} \quad (6)$$

where the objective function is described as

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

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

$$Q_1 := \{X : \alpha I \preceq X \preceq \beta I\},$$

$$Q_2 := \{U : \|U\|_\infty \leq \rho\},$$

where α, β ($0 < \alpha < \beta$) are given. (Note that Theorem 1 allows us to set α and β if no such a priori bounds are given.) We also define $\hat{f}(X) := -\log \det X + \langle S, X \rangle$, 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 $d_2(U)$:

$$\tilde{f}(X) = \hat{f}(X) + \max_{U \in Q_2} \{\langle AX, U \rangle - \mu d_2(U)\}. \quad (7)$$

The above function turns out to be a smooth uniform approximation to f everywhere. 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 $O(L/\epsilon)$.

4.2. Algorithm and Complexity Estimate

To detail the algorithm and compute the complexity, we must 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 $\nabla^2 d_1(X)[H, H] = \text{trace}(X^{-1}HX^{-1}H) \geq \beta^{-2}\|H\|_F^2$ 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 $d_1(X_0) = 0$. With our choice, we have $D_1 := \max_{X \in Q_1} d_1(X) = p \log(\beta/\alpha)$.

Similarly, the strong convexity parameter for $d_2(U)$ on Q_2 is $\sigma_2 := 1$, and we have $D_2 := \max_{U \in Q_2} d_2(U) = \rho^2/2$. 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 $\mu := \epsilon/2D_2$, 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 $Y_k = \arg \min_Y \{\langle \nabla \tilde{f}(X_k), Y - X_k \rangle + \frac{1}{2}L(\epsilon)\|Y - X_k\|_F^2 : Y \in Q_1\}$.
3. Find $Z_k = \arg \min_X \{\frac{L(\epsilon)}{\sigma_1}d_1(X) + \sum_{i=0}^k \frac{i+1}{2} \langle \nabla \tilde{f}(X_i), X - X_i \rangle : X \in Q_1\}$.
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 $L(\epsilon) := M + D_2\|A\|^2/(2\sigma_2\epsilon)$, where $M := 1/\alpha^2$ is the Lipschitz constant for the gradient of \hat{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

$$\begin{aligned} N(\epsilon) &:= 4\|A\| \sqrt{\frac{D_1 D_2}{\sigma_1 \sigma_2}} \cdot \frac{1}{\epsilon} + \sqrt{\frac{M D_1}{\sigma_1 \epsilon}} \\ &= \frac{\kappa \sqrt{p(\log \kappa)}}{\epsilon} (4p\alpha\rho + \sqrt{\epsilon}). \end{aligned} \quad (8)$$

where $\kappa = \beta/\alpha$ is a 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 eigenvalue 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 $N(\epsilon) = O(p^{1.5}/\epsilon)$, if the condition number κ is fixed a priori. Thus, the complexity of the algorithm is $O(p^{4.5}/\epsilon)$.

5. Numerical Results

In this section we present some numerical results.

5.1. Recovering Structure

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 $\rho = \sigma$.

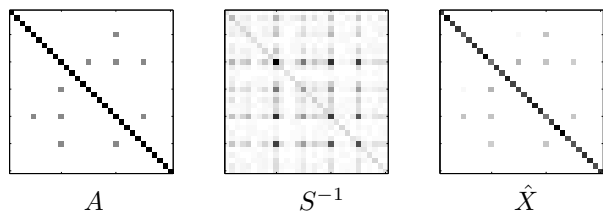


Figure 1. Recovering the sparsity pattern. We plot the underlying sparse matrix A , the inverse of the noisy version of A^{-1} , and the solution to problem (1) for ρ equal to the noise level.

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 $\sigma = 0.1$. Next, we count the number of misclassified zero and nonzero elements in the solution to (1). In figure 2, we plot the percentage of errors versus $\log(\rho/\sigma)$, as well as error bars corresponding to one standard deviation. As shown, for $\rho = \sigma$, we can almost exactly recover the underlying sparsity pattern, but even for a wide range of values of ρ above and below σ , the percentage of errors is small.

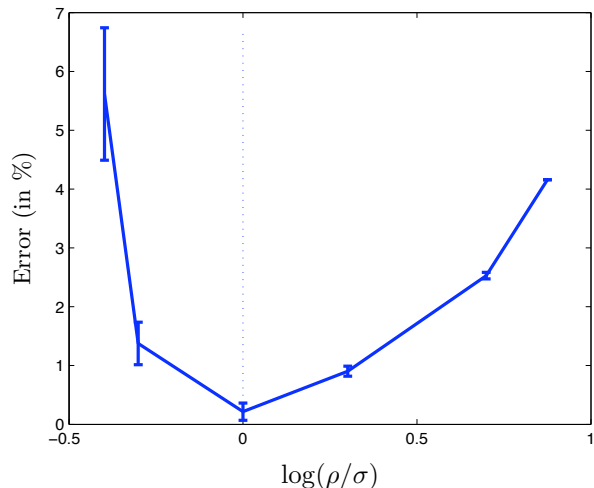


Figure 2. Recovering structure: Average and standard deviation of the percentage of errors (false positives + false negatives) versus ρ on random problems.

5.2. CPU Times Versus Problem Size

For a sense of the practical performance of the Nesterov method and the block coordinate descent method, we randomly selected 10 sample covariance matrices S for problem sizes p ranging from 400 to 1000. In each case, the number of samples n was chosen to be about a third of p . In figure 3 we plot the average CPU time to achieve a duality gap of $\epsilon = 0.1$. CPU times were computed using an AMD Athlon 64 2.20GHz processor with 1.96GB of RAM.

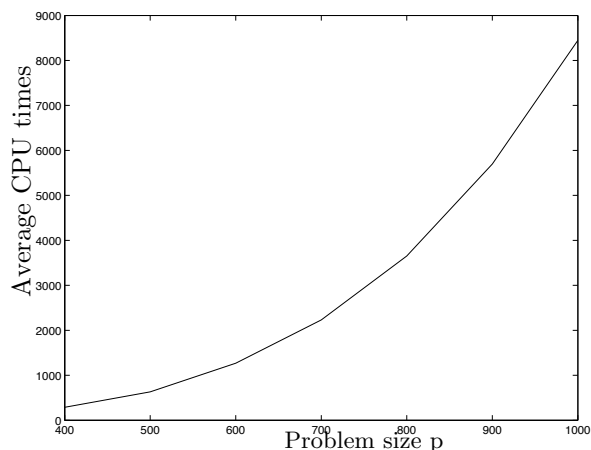


Figure 3. Average CPU times vs. problem size using block coordinate descent. We plot the average CPU time (in seconds) to reach a gap of $\epsilon = 0.1$ versus problem size p .

As shown, we are typically able to solve a problem of

size $p = 1000$ in about two and half hours.

5.3. Trial on Gene Expression Profiles

For illustration, we tested our method on two genomic data sets.

Rosetta Inpharmatics compendium dataset.

We first applied the block coordinate descent method to the Rosetta Inpharmatics Compendium (Hughes et al., 2000). The 300 experiment compendium dataset contains $n = 253$ samples with $p = 6136$ variables. With a view towards obtaining a very sparse graph, we set $\gamma = 0.1$ in the heuristic formula (3) of section (2.3) to obtain $\rho = 0.0313$.

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 $\epsilon = 0.146$, 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 GraphExplore program developed by (Dobra & West, 2004). The method has picked out a cluster of genes associated with amino acid metabolism, as described by (Hughes et al., 2000).

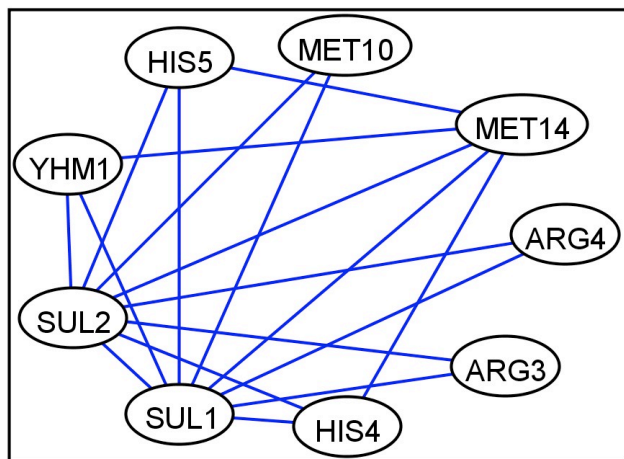


Figure 4. Application to Hughes dataset. We applied our method to the Rosetta Inpharmatics compendium, using $\rho = 0.0313$. Shown above is a sample subgraph containing some genes associated with amino acid metabolism.

Iconix microarray dataset. Next we analyzed a subset of a 10,000 gene microarray dataset from 160 drug treated rat livers (Natsoulis et al., 2005). In this study, rats were treated with a variety of fibrates, sta-

Table 1. Predictor genes for LDL receptor.

ACCESSION	GENE
BF553500	CBP/P300-INTERACTING TRANSACTIVATOR
BF387347	EST
BF405996	CALCIUM CHANNEL, VOLTAGE DEPENDENT
NM_017158	CYTOCHROME P450, 2C39
K03249	ENOYL-CoA, HYDRATASE/3-HYDROXYACYL Co A DEHYDROG.
BE100965	EST
AI411979	CARNITINE O-ACETYLTRANSFERASE
AI410548	3-HYDROXYISOBUTYRYL-Co A HYDROLASE
NM_017288	SODIUM CHANNEL, VOLTAGE-GATED
Y00102	ESTROGEN RECEPTOR 1
NM_013200	CARNITINE PALMITOYLTRANSFERASE 1B

tin, or estrogen receptor agonist compounds. The 500 most variable genes were submitted to the block coordinate descent approach. Again setting $\gamma = 0.1$ in the heuristic formula (3), we obtained $\rho = 0.0853$.

The sample covariance for the data has $\text{rank}(S) = 159$. By applying the property of the solution discussed in section (3.2), the size of the problem was reduced to $\hat{p} = 339$. Six sweeps through all the columns were required to achieve a duality gap of $\epsilon = 0.01$, with a total computing time of about 10 minutes. The solution has a condition number of 41.55.

The first order neighbors of any node in a Gaussian graphical model form the set of predictors for that variable. Using this method, we found that LDL receptor had one of the largest number of first-order neighbors in the Gaussian graphical model. The LDL receptor is believed to be one of the key mediators of the effect of both statins and estrogenic compounds on LDL cholesterol. Table 1 lists some of the first order neighbors of LDL receptor.

It is perhaps not surprising that several of these genes are directly involved in either lipid or steroid metabolism (K03249, AI411979, AI410548, NM_013200, Y00102). Other genes such as Cbp/p300 are known to be global transcriptional regulators. Finally, some are un-annotated ESTs. Their connection to the LDL receptor in this analysis may provide clues to their function.

6. Conclusions

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.

By imposing a priori bounds on the condition number of the solution we were able to improve the dependence of the computational complexity estimate on problem size p from $O(p^6 \log(1/\epsilon))$ to $O(p^{4.5}/\epsilon)$, where ϵ is the desired accuracy. This is a substantial improvement given that we cannot expect to do better than $O(p^3)$. The block coordinate descent method performs well in practice, typically solving problems with $p = 1000$ variables in about two and half hours on a desktop PC.

Acknowledgements

The authors would like to thank Francis Bach, Peter Bartlett, and Martin Wainwright for enlightening discussions on this topic, as well as the reviewers for very valuable suggestions and comments.

References

- Bertsekas, D. (1998). *Nonlinear programming*. Athena Scientific.
- Dahl, J., Roychowdhury, V., & Vandenberghe, L. (2005). Maximum likelihood estimation of gaussian graphical models: numerical implementation and topology selection. *UCLA preprint*.
- d’Aspremont, A., El Ghaoui, L., Jordan, M., & Lanckriet, G. R. G. (2004). A direct formulation for sparse PCA using semidefinite programming. *Advances in Neural Information Processing Systems*, 17.
- Dempster, A. P. (1972). Covariance selection. *Biometrics*, 28, 157–75.
- Dobra, A., Hans, C., Jones, B., Nevins, J. J. R., Yao, G., & West, M. (2004). Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, 90, 196–212.
- Dobra, A., & West, M. (2004). Bayesian covariance selection. *Working paper, ISDS, Duke University*.
- Huang, J. Z., Liu, N., & Pourahmadi, M. (2005). Covariance selection and estimation via penalized normal likelihood. *Wharton Preprint*.
- Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H., He, Y. D., Kidd, M. J., King, A. M., Meyer, M. R., Slade, D., Lum, P. Y., Stepaniants, S. B., Shoemaker, D. D., Gachotte, D., Chakraburty, K., Simon, J., Bard, M., & Friend, S. H. (2000). Functional discovery via a compendium of expression profiles. *Cell*, 102, 109–126.
- Lauritzen, S. (1996). *Graphical models*. Springer Verlag.
- Li, H., & Gui, J. (2005). Gradient directed regularization for sparse gaussian concentration graphs, with applications to inference of genetic networks. *University of Pennsylvania Technical Report*.
- Luo, Z. Q., & Tseng, P. (1992). On the convergence of the coordinate descent method for convex differentiable minimization. *Journal of Optimization Theory and Applications*, 72, 7–35.
- Meinshausen, N., & Bühlmann, P. (2005). High dimensional graphs and variable selection with the lasso. *Annals of statistics, in press*.
- Muirhead, R. J. (1982). *Aspects of multivariate statistical theory*. John Wiley and Sons, Inc.
- Natsoulis, G., El Ghaoui, L., Lanckriet, G., Tolley, A., Leroy, F., Dunlea, S., Eynon, B., Pearson, C., Tugendreich, S., & Jarnagin, K. (2005). Classification of a large microarray data set: algorithm comparison and analysis of drug signatures. *Genome Research*, 15, 724–736.
- Nesterov, Y. (2005). Smooth minimization of non-smooth functions. *Math. Prog., Ser. A*, 103, 127–152.
- Schäfer, J., & Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal statistical society, series B*, 58.
- Vandenberghe, L., Boyd, S., & Wu, S.-P. (1998). Determinant maximization with linear matrix inequality constraints. *SIAM Journal on Matrix Analysis and Applications*, 19, 499 – 533.