

Convex Optimization Techniques for Covariance Selection

Onureena Banerjee
Advisor: **Prof. Laurent El Ghaoui**

Joint work with **Prof. Alexandre d'Aspremont**, *Princeton University*
& **Dr. Georges Natsoulis**, *Iconix Pharmaceuticals*

Outline

- **Introduction**
- Algorithms
- Parameter Selection
- Numerical Results

Introduction

- Draw n independent samples $y_i \sim \mathcal{N}_p(0, \Sigma)$, where Σ is unknown.

We wish to estimate Σ from these samples.

- **Prior belief:** many conditional independencies among the variables in this distribution.
- Zeros in inverse covariance correspond to conditional independence properties among variables (Wermuth 1976).
- **Goal:** From y_1, \dots, y_n , try to recover the zero pattern of Σ^{-1} .
- **Covariance selection:** choosing which elements of our estimate $\hat{\Sigma}^{-1}$ to set to zero.

Introduction

- Let $S = \frac{1}{n} \sum_{i=1}^n y_i y_i^T$, the sample covariance matrix
- Log likelihood function:

$$\ell(y_1, \dots, y_n) = -(np/2) \log(2\pi) - (n/2) \log \det \Sigma - (n/2) \text{tr}(\Sigma^{-1} S)$$

- Let $X \succ 0$ be our estimate of the precision matrix Σ^{-1} .

Maximum likelihood:

$$\max_{X \succ 0} \log \det X - \text{tr}(XS) \tag{1}$$

- If $S \succ 0$, the solution is $X = S^{-1}$.

Introduction

- **Problem:** How do we estimate Σ in such a way that $\hat{\Sigma}^{-1}$ is sparse?
- **One solution:** penalize the log likelihood function.
- Minimize **Akaike Information Criterion:** $AIC = 2k - 2 \log L$ (no. of parameters k and likelihood function L).
- Attempts to find minimal model that correctly explains the data.

$$\max_{X \succ 0} \log \det X - \mathbf{tr}(SX) - \rho \mathbf{Card}(X)$$

where $\mathbf{Card}(X)$ is the number of nonzero elements in X .

- Set $\rho = \frac{2}{n}$ for the AIC.
- Alternative: Bayesian Information Criterion (BIC): set $\rho = \frac{\log(n)}{n}$.

But this is an NP-hard combinatorial problem.

Introduction

- Instead, replace $\mathbf{Card}(X)$ by $\|X\|_1 = \sum_{ij} |x_{ij}|$:

$$\max_{X \succ 0} \log \det X - \mathbf{tr}(SX) - \rho \|X\|_1 \quad (2)$$

- Convex, non-smooth, unbounded problem.
- Same idea used in l_1 -norm penalized regression (LASSO), for example.

Robustness and Regularization

We can write (2) as

$$\max_{X \succ 0} \min_{|u_{ij}| \leq \rho} \log \det X - \text{tr}(X(S + U))$$

Exchanging the min and max, we obtain the **dual problem**:

$$\min_U \{-\log \det(S + U) - p : |u_{ij}| \leq \rho, S + U \succ 0\} \quad (3)$$

- This can be interpreted as a **robust MLE** problem with componentwise noise of magnitude ρ on the elements of S .
- Using this, can show that adding the l_1 -norm penalty **regularizes** the solution \hat{X} , even for S singular (eg., $n < p$):

$$\frac{1}{\|S\|_2 + p\rho} I \preceq \hat{X} \preceq \frac{p}{\rho} I$$

Outline

- Introduction
- **Algorithms**
- Parameter Selection
- Numerical Results

Algorithms

Existing Interior Point Methods:

- Dual problem (3) is amenable to interior point methods (e.g. MAXDET).
- Complexity: $\mathcal{O}(p^6 \log(1/\epsilon))$.
- Involves storing a dense Hessian of size $\mathcal{O}(p^2)$.
- Standard software can solve this problem efficiently when the number of matrix entries is in the low hundreds.

Need **new algorithms** to solve problems where p is higher than the tens.

Note: We cannot possibly obtain a complexity better than $\mathcal{O}(p^3)$.

Algorithms

Contributions:

Two new algorithms aimed at solving large ($p = 1000$) problems where S , the sample covariance, is dense.

- **Block coordinate ascent algorithm:**
 - Solves the dual problem.
 - Good empirical performance: attains high accuracy (low primal-dual gap) quickly.
- **Nesterov's first order method for non-smooth minimization:**
 - Solves (a modified version of) the primal non-smooth problem.
 - Rigorous complexity results.
 - Trades off a worse dependence on accuracy for a better dependence on problem size.

Dual problem

Dual problem:

$$\min_U \{-\log \det(S + U) - p : |u_{ij}| \leq \rho, S + U \succ 0\}$$

- In the dual problem, diagonal elements of an optimal U are $U_{ii} = \rho$.
- Since dual problem has compact feasible set, primal and dual problems are equivalent.
- The dual solution $S + \hat{U}$ is an estimate of the covariance matrix, Σ
- Primal and dual solutions related by $\hat{X} = (S + \hat{U})^{-1}$.

Dual block-coordinate ascent

For simple notation, let $W := S + U$ to write the dual problem as:

$$\max\{\log \det W : |w_{ij} - s_{ij}| \leq \rho, W \succ 0\} \quad (4)$$

- **Initialize:** set $W^0 := S + \rho I$. Diagonal elements are fixed at optimal values.
- Optimize over **one column/row pair at a time:**

Partition the variable W and S as

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

Update rule for column w_{12} :

$$\hat{w}_{12} := \arg \min_y \{y^T W_{11}^{-1} y : \|y - s_{12}\|_\infty \leq \rho\} \quad (5)$$

Dual block-coordinate ascent

- Iterates produced by block coordinate ascent are **strictly positive definite**:

$$W^k \succ 0$$

- As a consequence, the QP to be solved at each step has a unique solution, guaranteeing convergence (e.g., see Bertsekas, 1998)
- Cycle through the columns in order. After each sweep through all columns, check the primal-dual gap:

$$\mathbf{tr}(SX) + \rho \|X\|_1 \leq p + \epsilon$$

where $X = W^{-1}$.

- Cost of method for K sweeps through all columns: $\mathcal{O}(Kp^4)$

Dual block-coordinate ascent

- The block coordinate method may be used to deduce a **property of the solution** to the original problem
- Suppose that, for column s_{12} in the sample covariance,

$$|s_{12}| \leq \rho \tag{6}$$

- Then the corresponding column in the solution will be zero: $\hat{w}_{12} = 0$
- This property can sometimes be used to **reduce the size** of the problem in advance, by setting to zero rows/columns of W corresponding to columns in the sample covariance S that satisfy (6).

Dual block-coordinate ascent

Connection to LASSO:

- At each iteration the BCA approach solves a box-constrained QP (5)
- The dual of this QP is

$$\min_x x^T W_{11}x - s_{12}^T x + \rho \|x\|_1 \quad (7)$$

- To clarify, let Q denote the (unique) positive definite square root of W_{11} , and let $y := \frac{1}{2}Q^{-1}s_{12}$. The problem can then be written

$$\min_x \|y - Qx\|_2^2 + \rho \|x\|_1 \quad (8)$$

- If W_{11} were a principal minor of S , then this would be a penalized regression of one variable against all others.

Dual block-coordinate ascent

Differences with the approach of Meinshausen and Bühlman (2005):

- **Their approach:** do l_1 norm penalized regression of each variable against all others, once.
- We begin with some regularization: $W^0 = S + \rho I$ so each LASSO-type problem has a unique solution
- We update the problem data after each iteration. In this sense, the block coordinate ascent method may be interpreted as a **recursive LASSO**.
- Instead of just one regression per variable, we continue until we converge to the solution of the original penalized maximum likelihood problem.

Nesterov's method

Background

Subgradient method: a simple method for minimizing nondifferentiable convex functions.

- Subgradient of f at x : any vector g s.t. $f(y) \geq f(x) + g^T(y - x)$ for all y .
- Subgradient method: $x^{(k+1)} = x^{(k)} - \alpha_k g$, where g is any subgradient of f at $x^{(k)}$.
- convergence rate: $O(\frac{1}{\epsilon^2})$, where ϵ is desired absolute accuracy
- slow rate of convergence, but low complexity per iteration
- black box model: complexity estimate cannot be improved

Nesterov's method

Background

- Nesterov (1983): for **smooth** problems, **optimal gradient scheme**:

$$O\left(\sqrt{\frac{L}{\epsilon}}\right)$$

- Nesterov (2005): for a special class of **non-smooth** problems:
 - **Approximate** non-smooth objective f by a smooth function \tilde{f}
 - Can choose \tilde{f} such that L is $O\left(\frac{1}{\epsilon}\right)$
 - Apply optimal gradient scheme to smooth problem.
 - Result: gradient scheme with efficiency $O\left(\frac{1}{\epsilon}\right)$

Nesterov's method

Nesterov starts by assuming that the problem has the following **min-max structure**:

$$\min_{x \in Q_1} f(x)$$

where

$$f(x) = \hat{f}(x) + \max_u \{ \langle Ax, u \rangle - \hat{\phi}(u) : u \in Q_2 \}$$

- Q_1 and Q_2 closed, bounded, convex sets.
- A is a linear operator.
- \hat{f} is a convex function with a Lipschitz continuous gradient.
- $\hat{\phi}(u)$ is a continuous convex function

Nesterov's method

If a problem can be written in this form, then the algorithm works as follows:

Regularization.

- Add a strongly convex penalty to f to produce \tilde{f} .
- \tilde{f} is a smooth uniform approximation to f everywhere, Lipschitz continuous gradient.

Optimal first order minimization.

- Use Nesterov's 1983 optimal first order scheme for smooth functions to solve the regularized problem.

Nesterov's method

Optimal first-order minimization. Nesterov's 1983 algorithm:

Set desired accuracy ϵ and initialize $X_0 = \beta I$.

For $k \geq 0$ do:

1. Compute $\nabla \tilde{f}(X_k) = -X_k^{-1} + S + U^*(X_k)$
2. Find $Y_k = \arg \min_Y \{ \mathbf{tr}(\nabla \tilde{f}(X_k))(Y - X_k) + \frac{1}{2}L\|Y - X_k\|_F^2 : Y \in Q_1 \}$
3. Find $Z_k = \arg \min_X \{ \frac{L}{\sigma_1}d_1(X) + \sum_{i=0}^k \frac{i+1}{2} \mathbf{tr}(\nabla \tilde{f}(X_k))(X - X_i) : X \in Q_2 \}$
4. Update $X_{k+1} = \frac{2}{k+3}Z_k + \frac{k+1}{k+3}Y_k$

Nesterov's method

Challenges

- Our non-smooth primal problem is unbounded.
- Not clear how best to modify problem to fit Nesterov's structure.
- Nesterov's method is only fast if the two subproblems can be solved explicitly or very efficiently. Not clear how to optimally choose sets Q_1 and Q_2 so that this is true.
- Not clear how to choose associated prox functions to yield the best complexity estimate.

Nesterov's method

One possible solution:

To apply Nesterov's results, we can replace our original problem:

$$\max_{X \succ 0} \log \det X - \mathbf{tr}(XS) - \rho \|X\|_1$$

by one where we impose bounds on the eigenvalues of the primal variable X :

$$Q_1 := \{X \in \mathbf{S}^p : \alpha I \preceq X \preceq \beta I\}$$

$$Q_2 := \{U \in \mathbf{S}^p : |u_{ij}| \leq \rho\}$$

We can use our previously calculated bounds if no a priori bounds are given.

Nesterov's method

Complexity estimate:

Cost per iteration: using best choices for Q_1 , Q_2 , related parameters:

- Step 1: involves a matrix inversion: $O(p^3)$.
- Steps 2 and 3: involve eigenvalue decomposition, reducing them to vector problems that can be solved analytically: $O(p^3)$.
- Cost per iteration: $O(p^3)$.

Nesterov's method

Complexity estimate:

- Maximum number of iterations required to achieve solution:

$$N(\epsilon) = \kappa \frac{\sqrt{p(\log \kappa)}}{\epsilon} (4p\alpha\rho + \sqrt{\epsilon})$$

- Computed using a result from Nesterov (2005).
- Here, κ is a bound on the condition number of the solution β/α .
- If κ is unknown, we can use the bounds calculated previously.
- If κ is fixed a priori, then the number of iterations is $O(p^{1.5}/\epsilon)$, making the total complexity $O(p^{4.5}/\epsilon)$. (Compare to $O(p^6 \log(1/\epsilon))$ for IPMs).

Outline

- Introduction
- Algorithms
- **Parameter Selection**
- Numerical Results

Parameter Selection

An unresolved problem is how best to select the penalty parameter ρ .

Possible avenues:

- Determine connection to Akaike's problem, use AIC or BIC ρ values.
 - Difficult to prove that replacing $\mathbf{Card}(X)$ with $\|X\|_1$ yields a bound on the solution to Akaike's problem.
- Follow an analysis similar Meinshausen and Bühlmann (2005) to derive a formula.
 - Meinshausen and Bühlmann are able to bound the probability of making an error in inferring the graphical model.
 - However, they are examining a simpler optimization problem, and are able to exploit this simplicity.

Parameter Selection

Independence Assumption Approach:

- Suppose that ρ is fixed. If $\rho < |s_{ij}|$ then the our estimate of the covariance matrix cannot have a zero in that location: $\hat{w}_{ij} \neq 0$.
- As soon as ρ is selected, before solving the optimization problem, we've already decided that certain pairs of variables cannot be independent.
- For any pair (i, j) , let

H_0^{ij} : Variables i and j are independent.

H_1^{ij} : Variables i and j are not independent.

- If $\rho < |s_{ij}|$, then, at some level of significance, we should reject H_0^{ij} .

Parameter Selection

Independence Assumption Approach:

- Fix a significance level α .

- Let

$$m := \max_{i,j} s_{ii}s_{jj}$$

$$\rho := \frac{t_{n-2}(\alpha)\sqrt{m}}{\sqrt{n-2+t_{n-2}^2(\alpha)}} \quad (9)$$

where $t_{n-2}(\alpha)$ = two tailed $100\alpha\%$ point of the t-distribution, for $n - 2$ degrees of freedom.

- Then $\rho < |s_{ij}|$ implies that we reject H_0^{ij} at the α level of significance.

Parameter Selection

Independence Assumption Approach:

Proof: The given formula (9) for ρ implies, for any pair i, j ,

$$(n - 2)\rho^2 \geq t_{n-2}^2(\alpha)(s_{ii}s_{jj} - \rho^2) \quad (10)$$

Now suppose that $\rho < |s_{ij}|$. This implies $\rho^2 < s_{ij}^2 \leq s_{ii}s_{jj}$, since $S \succeq 0$.

This implies, from (10), that

$$(n - 2)\frac{\rho^2}{s_{ii}s_{jj} - \rho^2} \geq t_{n-2}^2(\alpha) \quad (11)$$

and also that

$$\frac{s_{ij}}{s_{ii}s_{jj} - s_{ij}^2} > \frac{\rho^2}{s_{ii}s_{jj} - \rho^2} \quad (12)$$

Parameter Selection

Independence Assumption Approach:

Proof, cont.: Combining (11) and (12):

$$(n - 2) \frac{s_{ij}^2}{s_{ii}s_{jj} - s_{ij}^2} > t_{n-2}^2(\alpha)$$

Taking the square root of both sides and yields

$$\sqrt{n - 2} \frac{|r_{ij}|}{\sqrt{1 - r_{ij}^2}} > t_{n-2}(\alpha)$$

where $r_{ij} = s_{ij} / \sqrt{s_{ii}s_{jj}}$ is the sample correlation coefficient. This is the condition, under a likelihood ratio test of size α , for rejecting H_0^{ij} . \square

Parameter Selection

Independence Assumption Approach:

- If $\rho \geq |s_{ij}|$, then \hat{w}_{ij} may or may not be zero.
- This choice yields an asymptotically consistent estimator:
As $n \rightarrow \infty$, our estimate of the covariance matrix $\hat{W} \rightarrow S$, and $S \rightarrow \Sigma$.
- Allows a hypothesis testing interpretation of some of the decisions made by the estimator.
- **Drawbacks:** This is a two-dimensional result, dealing with independence instead of conditional independence. Also may not be a good choice for p large and $p \gg n$.

Outline

- Introduction
- Algorithms
- Parameter Selection
- **Numerical Results**

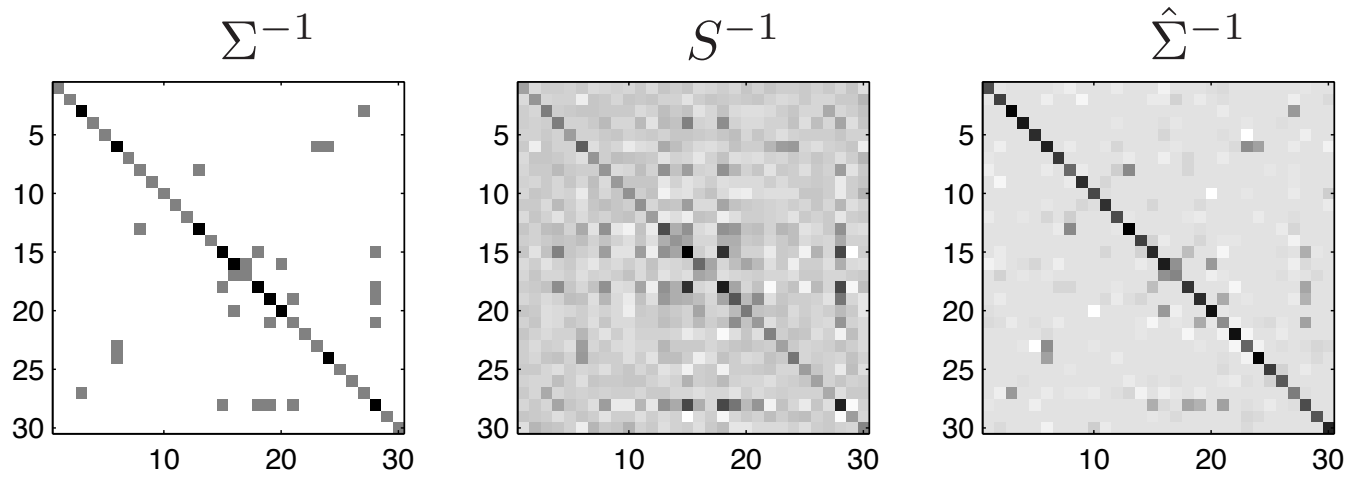
Numerical Examples

Generating random examples:

- Randomly select a sparse Σ^{-1} of size p .
- Generate n vectors according to $\mathcal{N}(0, \Sigma)$.
- Calculate sample covariance matrix S from these data.
- Solve penalized MLE problem.
- Compare zero pattern of solution $\hat{\Sigma}^{-1}$ to true Σ^{-1} .

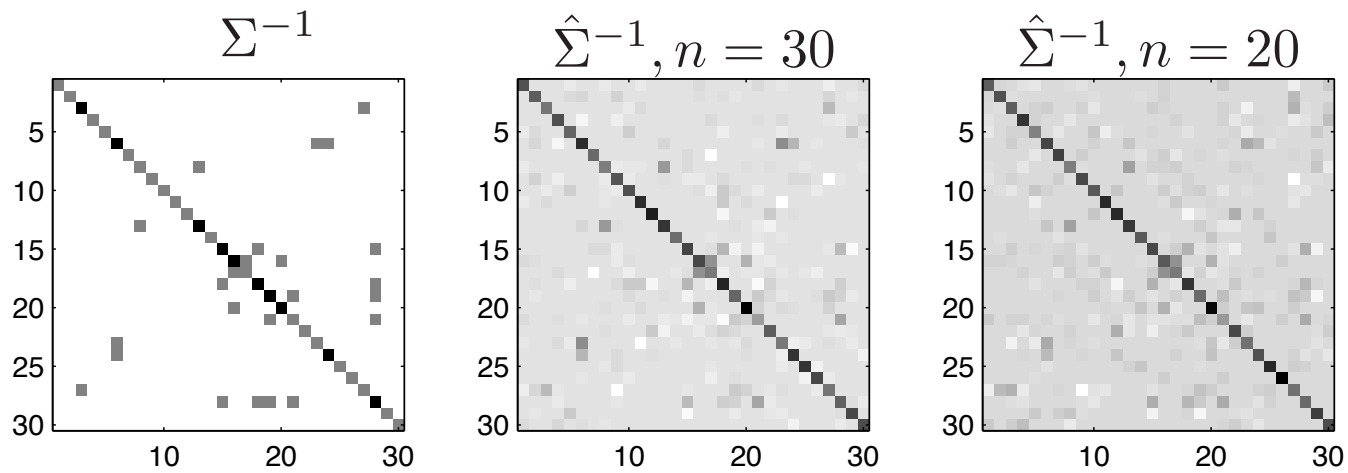
Numerical Examples

Example 1, $n = 60$:



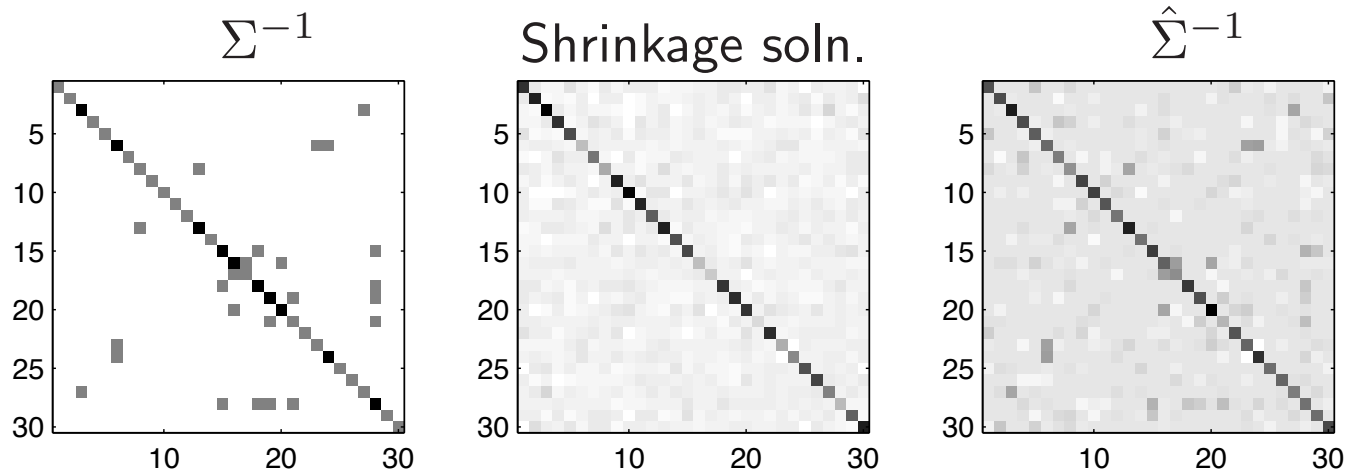
Numerical Examples

Example 2, $n = 30$, $n = 20$:



Numerical Examples

Example 3, comparison with shrinkage approach:



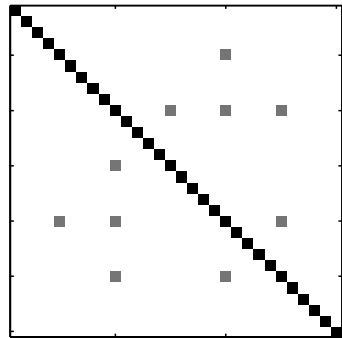
Numerical Examples

Recovering zero pattern masked by noise:

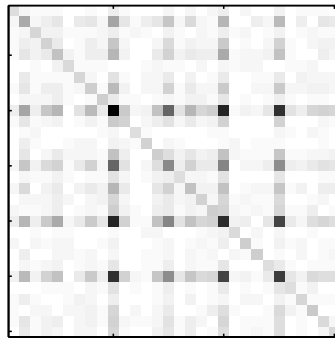
- Randomly select a sparse matrix A .
- Obtain S by adding a uniform noise of magnitude σ to A^{-1} .
- Solve sparse MLE problem using this S .
- Compare zero patterns of solution \hat{X} and A .

Numerical Examples

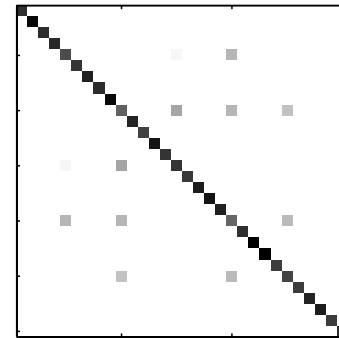
Recovering zero pattern masked by noise: $\rho = \sigma$.



A



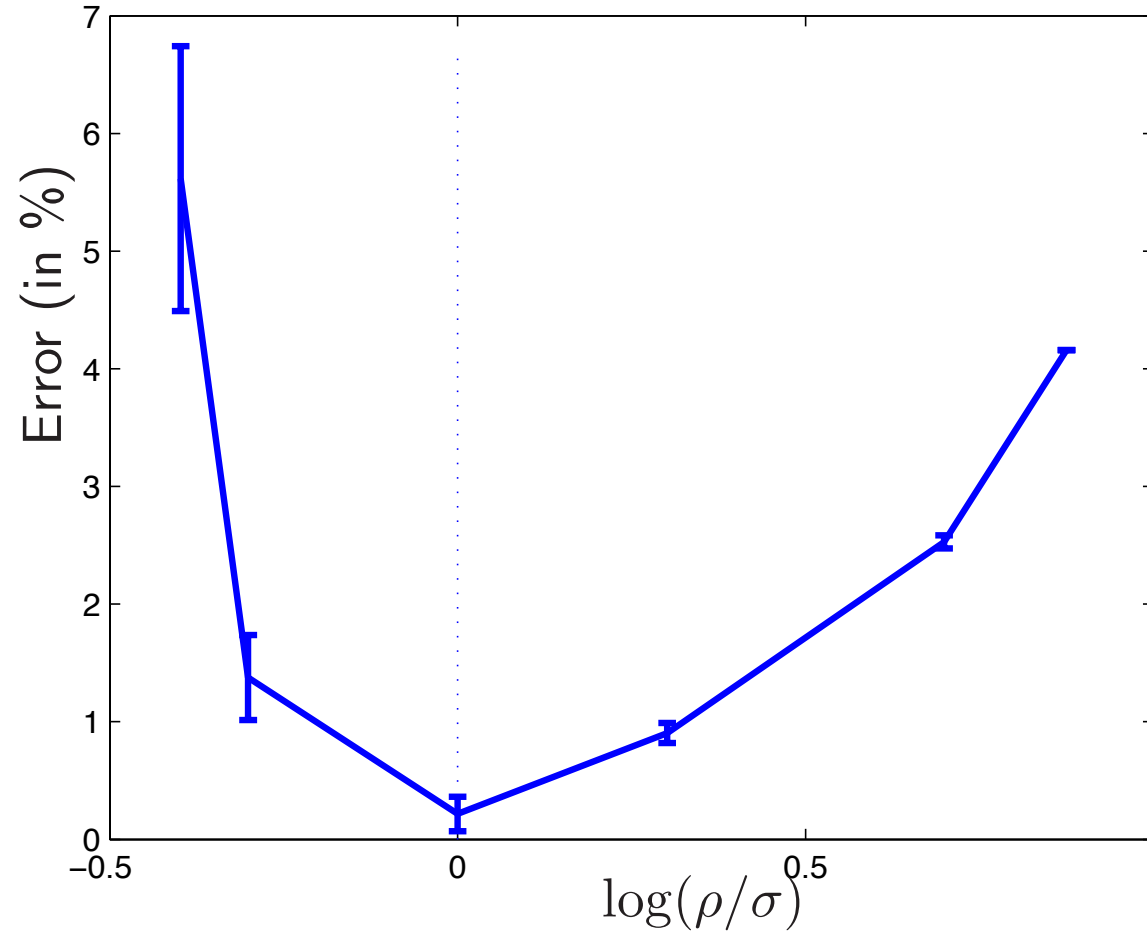
S^{-1}



Sparse MLE

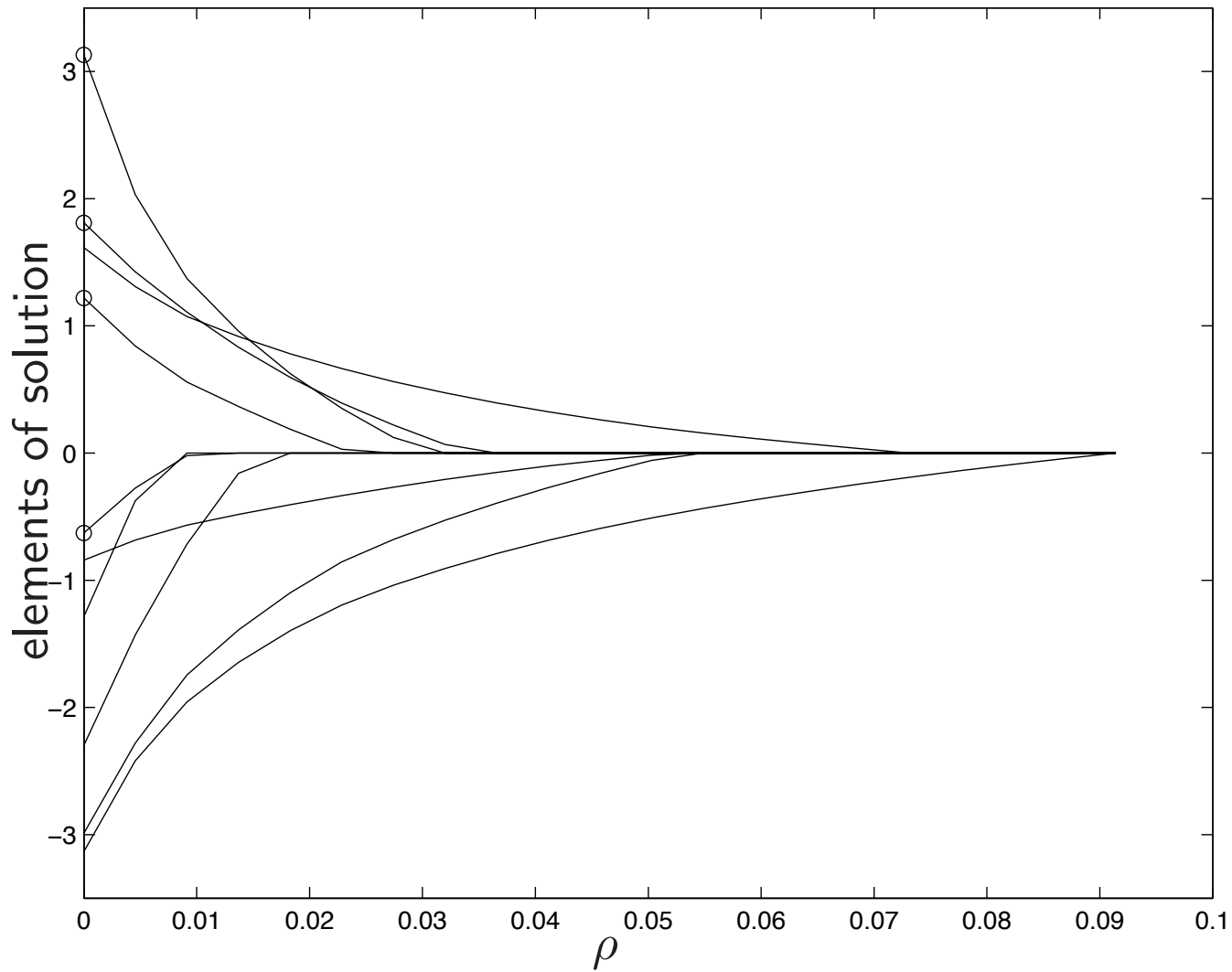
Numerical Examples

Recovering zero pattern masked by noise:



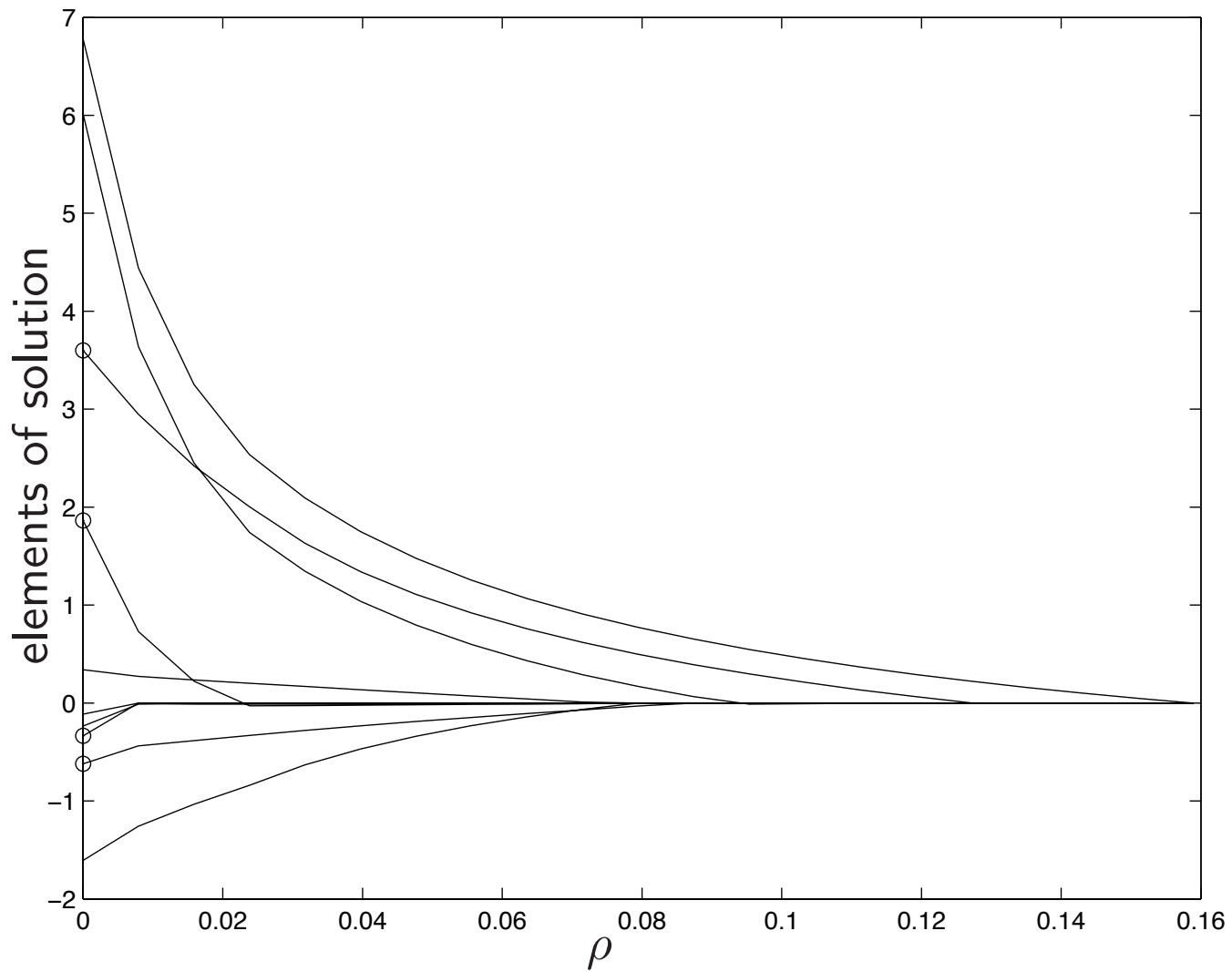
Numerical Examples

Path Following:



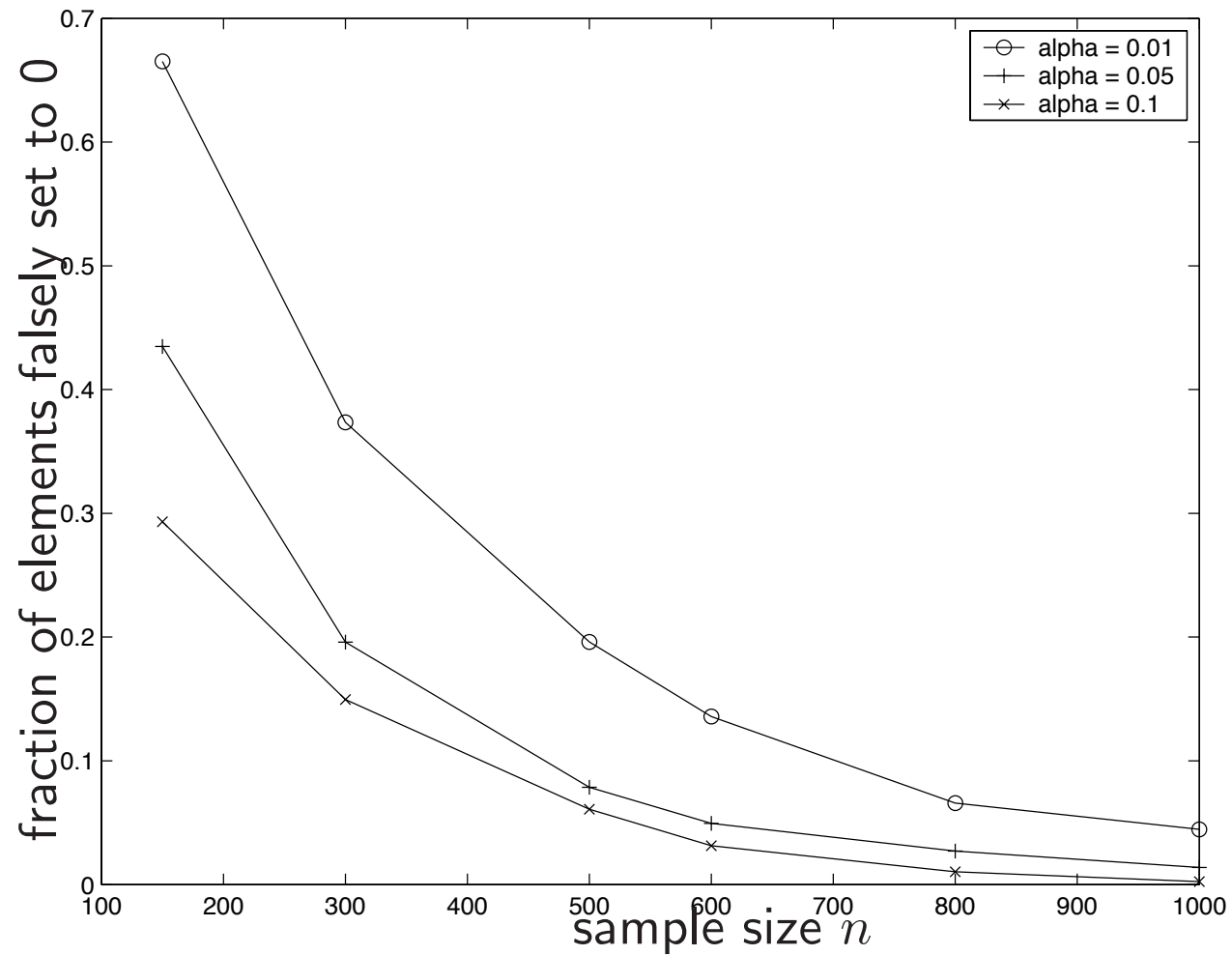
Numerical Examples

Path Following:



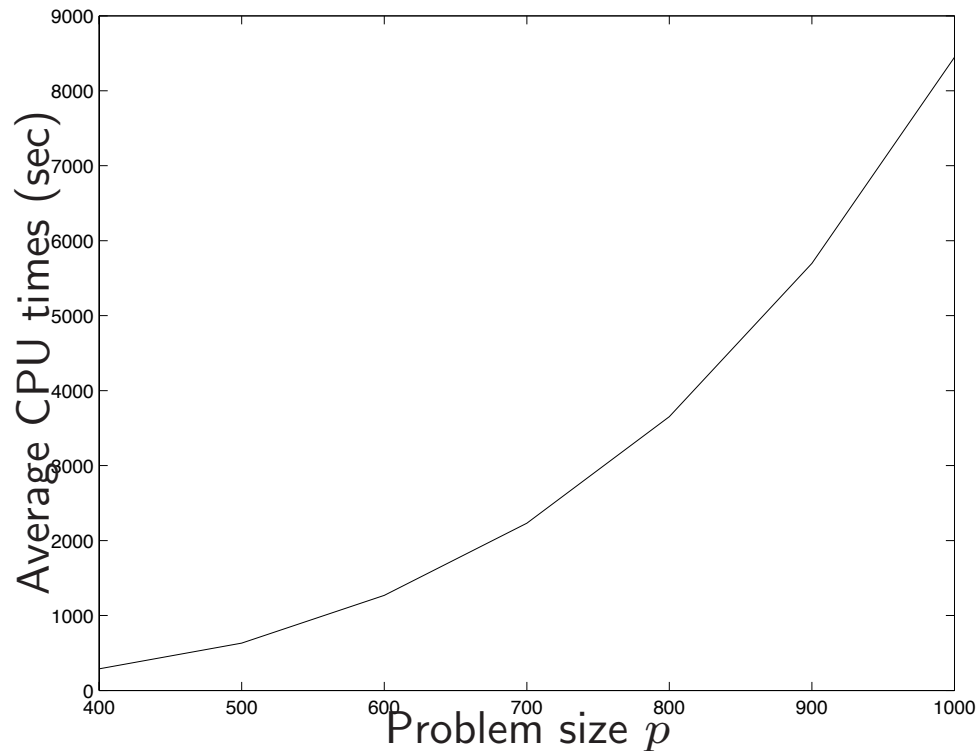
Numerical Examples

Simple formula: Elements of Σ^{-1} falsely set to 0 ($p = 500$)



Numerical Examples

CPU Times:



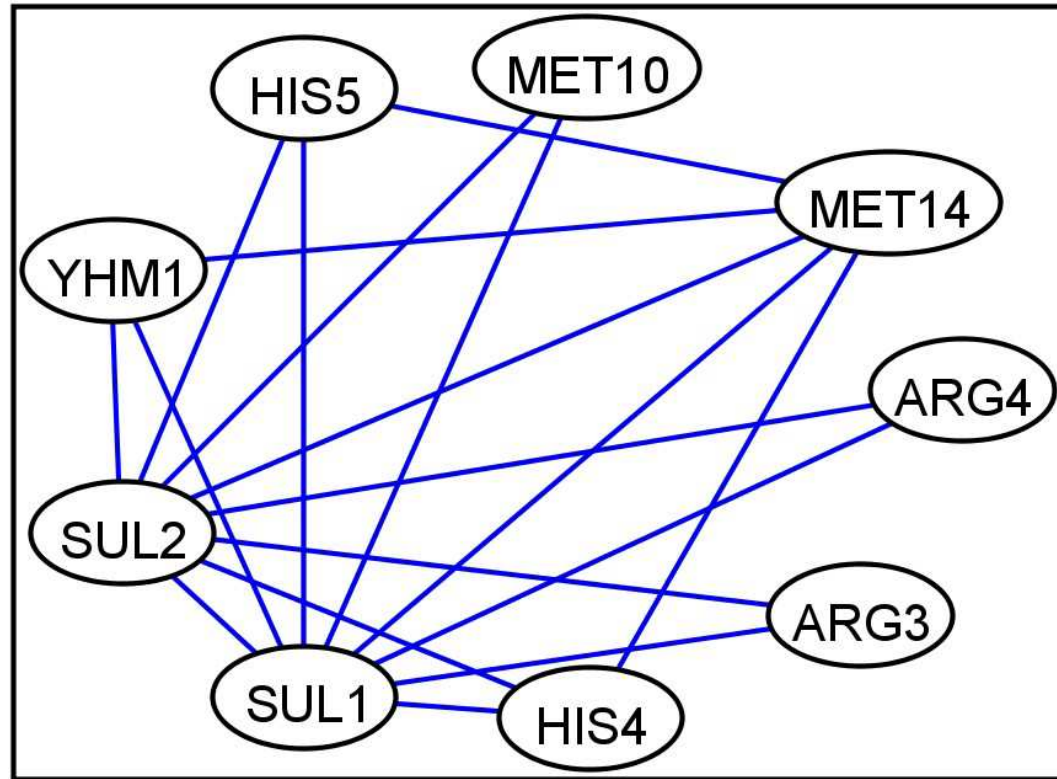
- Can solve a problem of size $p = 1000$ in 2.5 hours (2.20GHz processor, 1.96GB RAM).
- Can solve a problem of size $p = 3000$ in 4 days.

Gene expression data

Rosetta Inpharmatics Compendium:

- Public data compiled by Hughes et al. (2000).
- $n = 253$ samples and $p = 6136$ variables.
- Using $\alpha = 0.1$ in formula yields $\rho = 0.0313$.
- Applying property discussed earlier reduces size of problem to $\hat{p} = 537$.
- Problem solved in 18 minutes and is $\hat{\Sigma}^{-1}$ is 99% sparse.

Gene expression data



The algorithm correctly picked up a network of genes related to amino acid metabolism, as identified by Hughes et al. using biological reasoning.

Gene expression data

Iconix Pharmaceuticals Compendium:

- $p = 8500$ variables, $n = 1600$ samples
- Solved problem with a subset of $\hat{p} = 3000$ genes with highest variance.
- The first order neighbors of any node of a graphical model form the set of **predictors** for that variable.
- One possible test: look at the set of first order neighbors of a particular gene, compare to existing biological information.
- LDL receptor is believed to be one of the key mediators of the effect of both statins and estrogenic compounds on LDL cholesterol.

Gene expression data

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

- Several of these genes are directly involved in either lipid or steroid metabolism (K03249, AI411979, AI410548, NM_013200, Y00102).
- Genes such as Cbp/p300 are global transcriptional regulators.
- Finally, some are un-annotated ESTs, their connection to the LDL receptor in this analysis may provide clues to their function.

Conclusions

- Solving l_1 -norm penalized MLE problem potentially useful for recovering sparse Σ^{-1} from dense, singular S .
- Existing interior point methods can only solve problem efficiently for p in the tens.
- Nesterov: rigorous complexity estimate with substantially better dependence on problem size p than interior point methods.
- Dual BCA: solves problems of size $p = 1000$ in 2.5 hours.
- Some remaining questions . . .

Questions

- How should the penalty parameter ρ be selected?
- Thresholding solution using local fdr method of Efron et al. (2001)?
 - Schäfer & Strimmer (2005) apply this to their shrinkage approach.
- Comparison to stochastic algorithm of Dobra & West (2005)?
 - Like Meinshausen & Bühlmann, they do not compute any covariance matrices.
 - Work directly with Gaussian graphical models.
 - Method fits the entire distribution, even with $p = 12,000$ variables.