

# **Sparse Covariance Selection via Robust Maximum Likelihood Estimation**

**O. Banerjee, A. d'Aspremont and L. El Ghaoui**

*ORFE, Princeton University & EECS, U.C. Berkeley*

Available *online* at [www.princeton.edu/~aspremon](http://www.princeton.edu/~aspremon)

# Introduction

- We estimate a *sample covariance matrix*  $\Sigma$  from empirical data.
- Objective: infer *dependence* relationships between variables.
- We also want this information to be as *sparse* as possible.
- Basic solution: look at the magnitude of the *covariance* coefficients:

$$|\Sigma_{ij}| > \beta \quad \Leftrightarrow \quad \text{variables } i \text{ and } j \text{ are related.}$$

We can do better. . .

# Covariance Selection

Following Dempster (1972), look for zeros in the *inverse* matrix instead:

- *Parsimony*. Suppose that we are estimating a Gaussian density:

$$f(x, \Sigma) = \left(\frac{1}{2\pi}\right)^{\frac{p}{2}} \left(\frac{1}{\det \Sigma}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2}x^T \Sigma^{-1}x\right),$$

a sparse inverse matrix  $\Sigma^{-1}$  corresponds to a *sparse representation* of the density  $f$  as a member of an exponential family of distributions:

$$\exp(\alpha_0 + t(x) + \alpha_{11}t_{11}(x) + \dots + \alpha_{rs}t_{rs}(x))$$

with here  $t_{ij}(x) = x_i x_j$  and  $\alpha_{ij} = \Sigma_{ij}^{-1}$ .

- Dempster (1972) calls  $\Sigma_{ij}^{-1}$  a *concentration* coefficient.

# Covariance Selection

- We have  $m + 1$  observations  $x_i \in \mathbf{R}^n$  on  $n$  random variables.
- We can estimate a sample covariance matrix  $S$  such that

$$S = \frac{1}{m} \sum_{i=1}^{m+1} (x_i - \bar{x})(x_i - \bar{x})^T$$

- We choose a (symmetric) *subset*  $I$  of index pairs and denote by  $J$  the remaining indices so that  $I \cup J = \mathbf{N}^2$ .
- We *choose* a matrix  $\hat{\Sigma}$  such that:
  - $\hat{\Sigma}_{ij} = S_{ij}$  for all indices  $(i, j)$  in  $J$
  - $\hat{\Sigma}_{ij}^{-1} = 0$  for all indices  $(i, j)$  in  $I$

# Covariance Selection

Why is this a better choice? Dempster (1972) shows:

- *Existence and Uniqueness.* If there is a positive semidefinite matrix  $\hat{\Sigma}_{ij}$  satisfying  $\hat{\Sigma}_{ij} = S_{ij}$  on  $J$ , then there is *only one such matrix* satisfying  $\hat{\Sigma}_{ij}^{-1} = 0$  on  $I$ .
- *Maximum Entropy.* Among all Gaussian models  $\Sigma$  such that  $\Sigma_{ij} = S_{ij}$  on  $J$ , the choice  $\hat{\Sigma}_{ij}^{-1} = 0$  on  $I$  has *maximum entropy*.
- *Maximum Likelihood.* Among all Gaussian models  $\Sigma$  such that  $\Sigma_{ij}^{-1} = 0$  on  $I$ , the choice  $\hat{\Sigma}_{ij} = S_{ij}$  on  $J$  has *maximum likelihood*.

## Applications & Related Work

- *Gene expression data*. The sample data is composed of gene expression vectors and we want to isolate links in the expression of various genes. See Dobra, Hans, Jones, Nevins, Yao & West (2004), Dobra & West (2004) for example.
- *Speech Recognition*. See Bilmes (1999), Bilmes (2000) or Chen & Gopinath (1999).
- *Finance*. Identify links between sectors, etc.
- Related work by Dahl, Roychowdhury & Vandenberghe (2005): interior point methods for large, sparse MLE.

# Outline

- Introduction
- **Robust Maximum Likelihood Estimation**
- Algorithms
- Numerical Results

# Maximum Likelihood Estimation

- We can estimate  $\Sigma$  by solving the following maximum likelihood problem:

$$\max_{X \in \mathbf{S}^n} \log \det X - \mathbf{Tr}(SX)$$

- Problem here: How do we make  $\Sigma^{-1}$  *sparse*?
- Or, in other words, how do we efficiently choose  $I$  and  $J$ ?
- Solution: penalize the MLE.

# AIC and BIC

Original solution in Akaike (1973), *penalize* the likelihood function:

$$\max_{X \in \mathbf{S}^n} \log \det X - \mathbf{Tr}(SX) - \rho \mathbf{Card}(X)$$

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

- $\rho = 2/(m + 1)$  for the Akaike Information Criterion (**AIC**).
- $\rho = \frac{\log(m+1)}{(m+1)}$  for the Bayesian Information Criterion (**BIC**).

Of course, this is a (NP-Hard) combinatorial problem. . .

# Convex Relaxation

- We can form a *convex relaxation* of AIC or BIC penalized MLE

$$\max_{X \in \mathbf{S}^n} \log \det X - \mathbf{Tr}(SX) - \rho \mathbf{Card}(X)$$

replacing  $\mathbf{Card}(X)$  by  $\|X\|_1 = \sum_{ij} |X_{ij}|$  to solve

$$\max_{X \in \mathbf{S}^n} \log \det X - \mathbf{Tr}(SX) - \rho \|X\|_1$$

- This is the classic  $l_1$  heuristic,  $\|X\|_1$  is a *convex lower bound* on  $\mathbf{Card}(X)$ .
- See Fazel, Hindi & Boyd (2000) or d'Aspremont, El Ghaoui, Jordan & Lanckriet (2004) for related applications.

# Robustness

- This penalized MLE problem can be rewritten:

$$\max_{X \in \mathbf{S}^n} \min_{|U_{ij}| \leq \rho} \log \det X - \mathbf{Tr}((S - U)X)$$

- This can be interpreted as a *robust MLE* problem with componentwise noise of magnitude  $\rho$  on the elements of  $S$ .
- The relaxed sparsity requirement is equivalent to a robustification.
- See d'Aspremont et al. (2004) for similar results on sparse PCA.

# Outline

- Introduction
- Robust Maximum Likelihood Estimation
- **Algorithms**
- Numerical Results

# Algorithms

- We need to solve:

$$\max_{X \in \mathbf{S}^n} \log \det X - \mathbf{Tr}(SX) - \rho \|X\|_1$$

- For medium size problems, this can be done using interior point methods.
- In practice, we need to solve *very large* instances. . .
- The  $\|X\|_1$  penalty implicitly introduce  $O(n^2)$  linear constraints.

# Algorithms

Here, we can exploit problem structure

- Our problem here has a particular *min-max* structure:

$$\max_{X \in \mathbf{S}^n} \min_{|U_{ij}| \leq \rho} \log \det X - \mathbf{Tr}((S - U)X)$$

- This min-max structure means that we use prox function algorithms by Nesterov (2005) (see also Nemirovski (2004)) to solve large, dense problem instances.
- We also detail a “greedy” block-coordinate descent method with good empirical performance.

# Nesterov's method

Solve

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

- Starts from a particular *min-max model* on the problem:

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

- assuming that:
  - $f$  is defined over a compact convex set  $Q_1 \subset \mathbf{R}^n$
  - $\hat{f}(x)$  is convex, differentiable and has a Lipschitz continuous gradient with constant  $M \geq 0$
  - $T \in \mathbf{R}^{n \times n}$
  - $\hat{\phi}(u)$  is a continuous convex function over some closed compact set  $Q_2 \subset \mathbf{R}^n$ .

# Nesterov's method

Assuming that a problem can be written according to this min-max model, the algorithm works as follows. . .

- *Regularization*. Add strongly convex penalty inside the min-max representation to produce an  $\epsilon$ -approximation of  $f$  with Lipschitz continuous gradient (generalized Moreau-Yosida regularization step, see Lemaréchal & Sagastizábal (1997) for example).
- *Optimal first order minimization*. Use optimal first order scheme for Lipschitz continuous functions detailed in Nesterov (1983) to solve the regularized problem.

Caveat: Only efficient if the subproblems involved in these steps can be solved explicitly or very efficiently. . .

# Nesterov's method

- The min-max model makes this an ideal candidate for *robust optimization*
- For fixed problem size, the number of iterations required to get an  $\epsilon$  solution is given by

$$O\left(\frac{1}{\epsilon}\right)$$

compared to  $O\left(\frac{1}{\epsilon^2}\right)$  for generic first-order methods.

- Each iteration has low memory requirements.
- Change in *granularity* of the solver: larger number of cheaper iterations.

# Nesterov's method

- We solve the following (modified) problem:

$$\max_{\{X \in \mathbf{S}^n : \alpha I_n \preceq X \preceq \beta I_n\}} \min_{\{U \in \mathbf{S}^n : |U_{ij}| \leq \rho\}} \log \det X - \mathbf{Tr}((S - U)X)$$

equivalent to the original problem if  $\alpha \leq 1/(\|S\| + n\rho)$  and  $\beta \geq n/\rho$ .

- We can write this in the *min-max model*'s format:

$$\min_{X \in Q_1} \max_{U \in Q_2} \hat{f}(X) + \mathbf{Tr}((TX)U) = \min_{X \in Q_1} f(X),$$

where:

- $\hat{f}(X) = -\log \det X + \mathbf{Tr}(SX)$
- $T = \rho I_{n^2}$
- $Q_1 := \{X \in \mathcal{S}^n : \alpha I_n \preceq X \preceq \beta I_n\}$
- $Q_2 := \{U \in \mathcal{S}^n : \|U\|_\infty \leq 1\}$ .

# Nesterov's method

*Regularization.* The objective is first smoothed by penalization, we define:

$$f_\epsilon(X) := \hat{f}(X) + \max_{U \in Q_2} \mathbf{Tr}(XU) - (\epsilon/2D_2)d_2(U)$$

which is an  $\epsilon$  approximation of  $f$  where

- the prox function on  $Q_2$  is  $d_2(U) = \frac{1}{2} \mathbf{Tr}(U^T U)$
- the constant  $D_2$  is given by  $D_2 := \max_{Q_2} d_2(U) = n^2/2$ .

In this case, this corresponds to a classic Moreau-Yosida regularization of the penalty  $\|X\|_1$  and the function  $f_\epsilon$  has a Lipschitz continuous gradient with constant:

$$L_\epsilon := M + D_2 \rho^2 / (2\epsilon)$$

# Nesterov's method

*Optimal first-order minimization.* The minimization algorithm in Nesterov (1983) then involves the following steps:

Choose  $\epsilon > 0$  and set  $X_0 = \beta I_n$ , **For**  $k = 0, \dots, N(\epsilon)$  **do**

1. Compute  $\nabla f_\epsilon(X_k) = -X^{-1} + \Sigma + U^*(X_k)$
2. Find 
$$Y_k = \arg \min_Y \left\{ \mathbf{Tr}(\nabla f_\epsilon(X_k)(Y - X_k)) + \frac{1}{2} L_\epsilon \|Y - X_k\|_F^2 : Y \in \mathcal{Q}_1 \right\}.$$
3. Find  $Z_k = \arg \min_X \left\{ L_\epsilon \beta^2 d_1(X) + \sum_{i=0}^k \frac{i+1}{2} \mathbf{Tr}(\nabla f_\epsilon(X_i)(X - X_i)) : X \in \mathcal{Q}_1 \right\}.$
4. Update  $X_k = \frac{2}{k+3} Z_k + \frac{k+1}{k+3} Y_k.$

## Nesterov's method

- We have chosen a prox function  $d_1(X)$  for the set  $\{\alpha I_n \preceq X \preceq \beta I_n\}$ :

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

- Step 1 only amount to computing the *inverse* of  $X$  and the (explicit) solution to the regularized subproblem on  $Q_2$ .
- Steps 2 and 3 are both projections on  $Q_1$  and require an *eigenvalue decomposition*.
- This means that the total complexity estimate of the method is:

$$O\left(\frac{\kappa \sqrt{n(\log \kappa)}}{\epsilon} (4n^4 \alpha \rho + n^3 \sqrt{\epsilon})\right)$$

where  $\log \kappa = \log(\beta/\alpha)$  bounds the solution's condition number.

# Dual block-coordinate descent

- Here we consider the dual of the original problem:

$$\begin{array}{ll} \text{maximize} & \log \det(S + U) \\ \text{subject to} & \|U\|_{\infty} \leq \rho \\ & S + U \succeq 0 \end{array}$$

- The diagonal entries of an optimal  $U$  are  $U_{ij} = \rho$ .
- We will solve for  $U$  column by column, sweeping all the columns.

## Dual block-coordinate descent

- Let  $C$  be the current iterate, after permutation we can always assume that we optimize over the last column:

$$\begin{aligned} & \text{maximize} && \log \det \begin{pmatrix} C^{11} & C^{12} + u \\ C^{21} + u^T & C^{22} \end{pmatrix} \\ & \text{subject to} && \|u\|_\infty \leq \rho \end{aligned}$$

where  $C^{12}$  is the last column of  $C$  (off-diag.).

- Each iteration reduces to a simple box-constrained QP:

$$\begin{aligned} & \text{minimize} && u^T C^{-1} u \\ & \text{subject to} && \|u\|_\infty \leq \rho \end{aligned}$$

- We stop when  $\text{Tr}(SX) + \rho \|X\|_1 - n \leq \epsilon$  where  $X = C^{-1}$ .

# Outline

- Introduction
- Robust Maximum Likelihood Estimation
- Algorithms
- **Numerical Results**

# Numerical Examples

Generate random examples:

- Take a sparse, random p.s.d. matrix  $A \in \mathbf{S}^n$
- We add a uniform noise with magnitude  $\sigma$  to its inverse

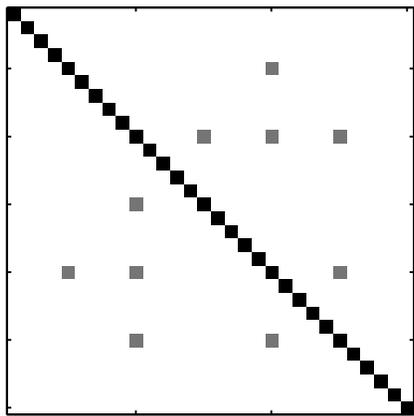
We then solve the penalized MLE problem (or the modified one):

$$\max_{X \in \mathbf{S}^n} \log \det X - \mathbf{Tr}(SX) - \rho \|X\|_1$$

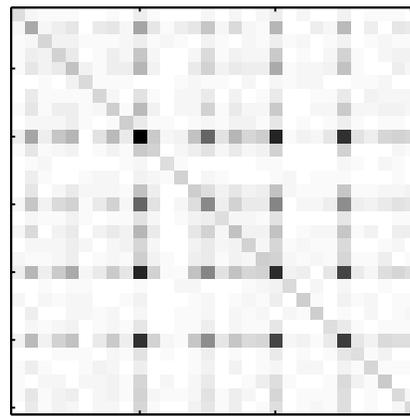
and compare the solution with the original matrix  $A$ .

# Numerical Examples

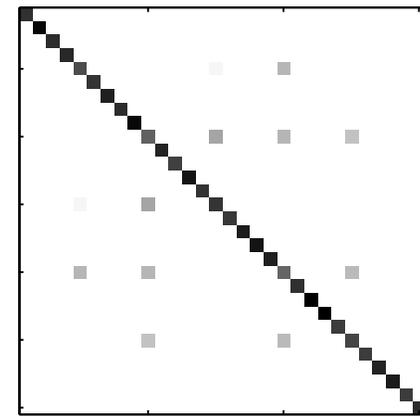
A basic example. . .



Original inverse  $A$

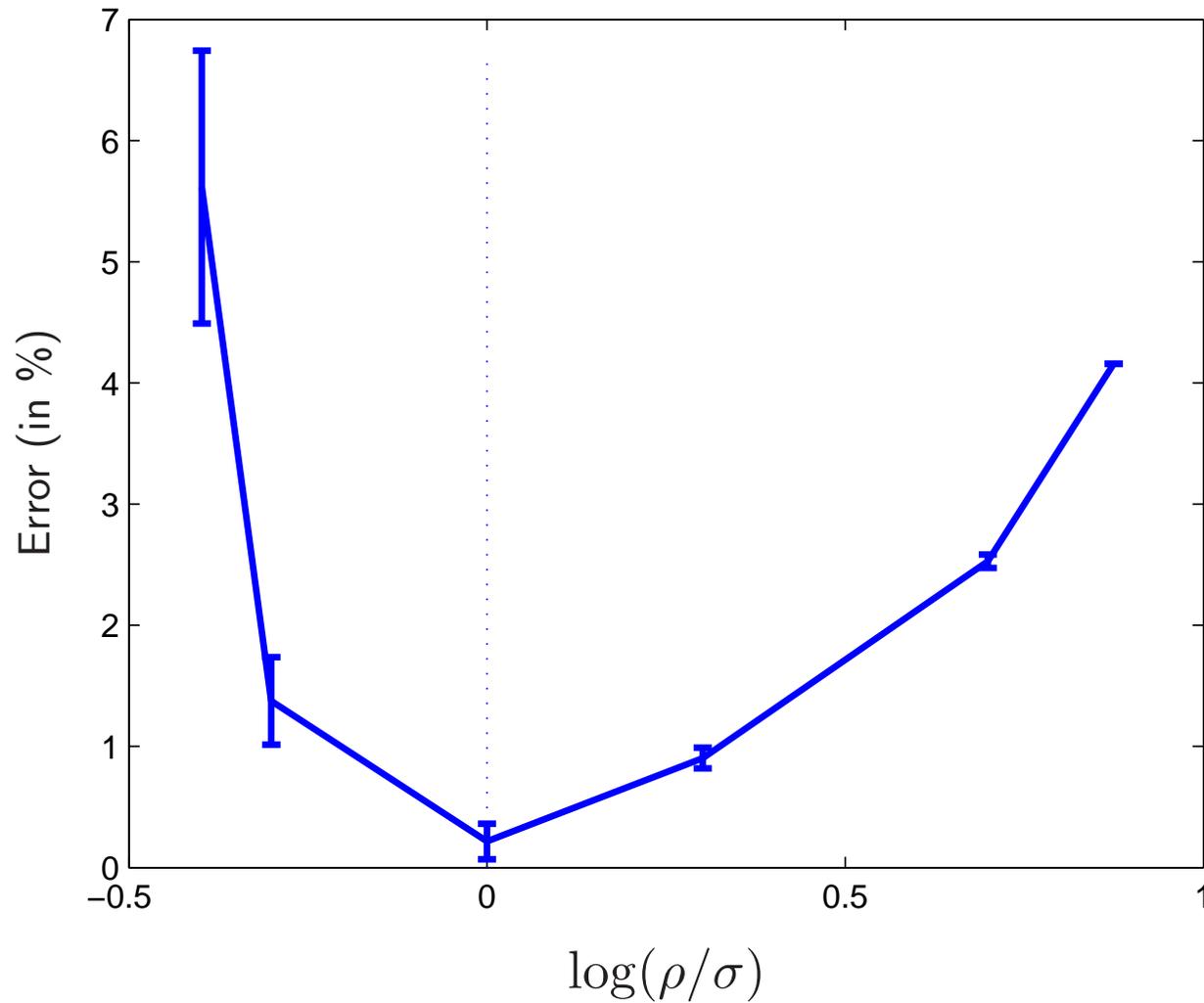


Noisy inverse  $\Sigma^{-1}$

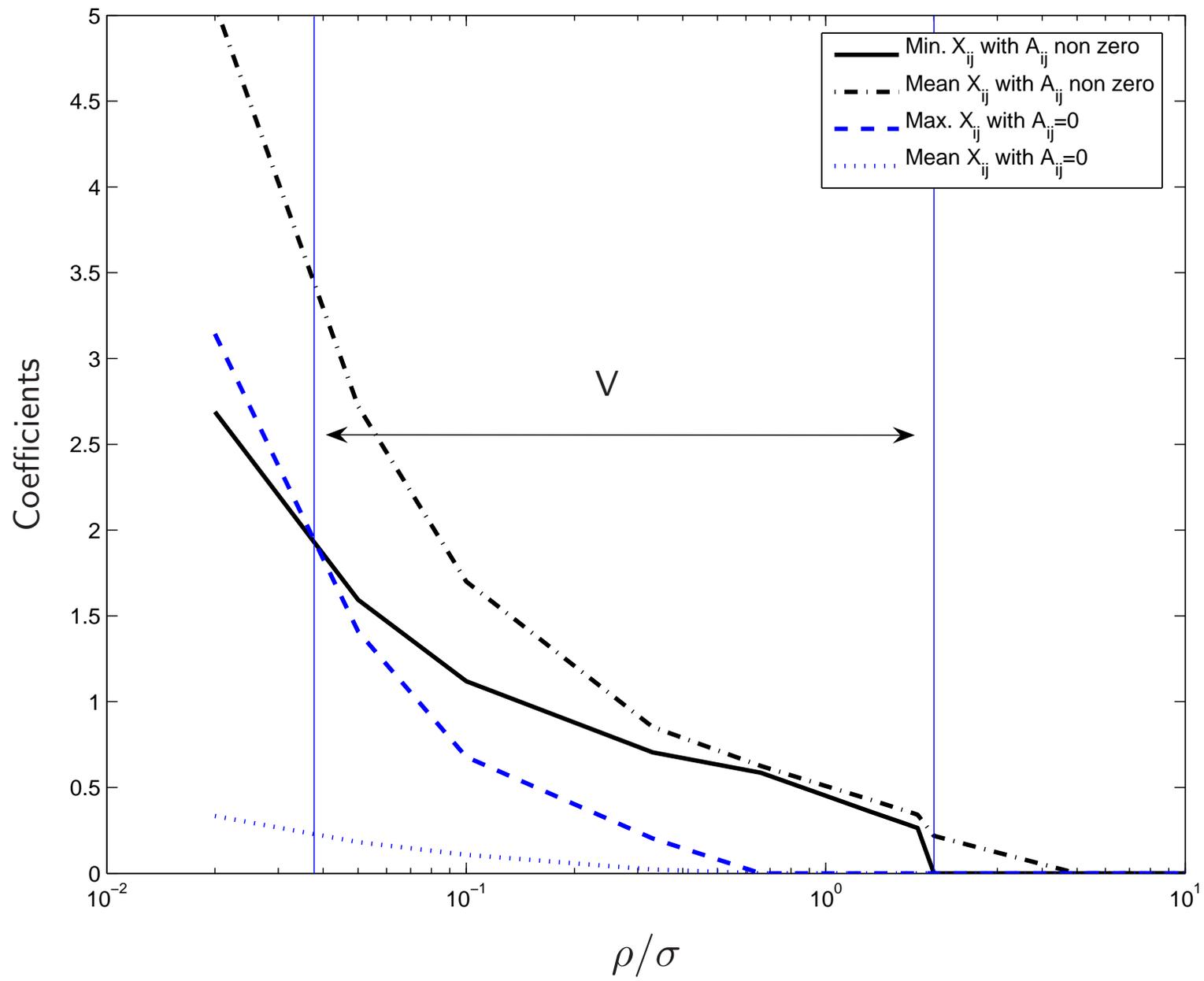


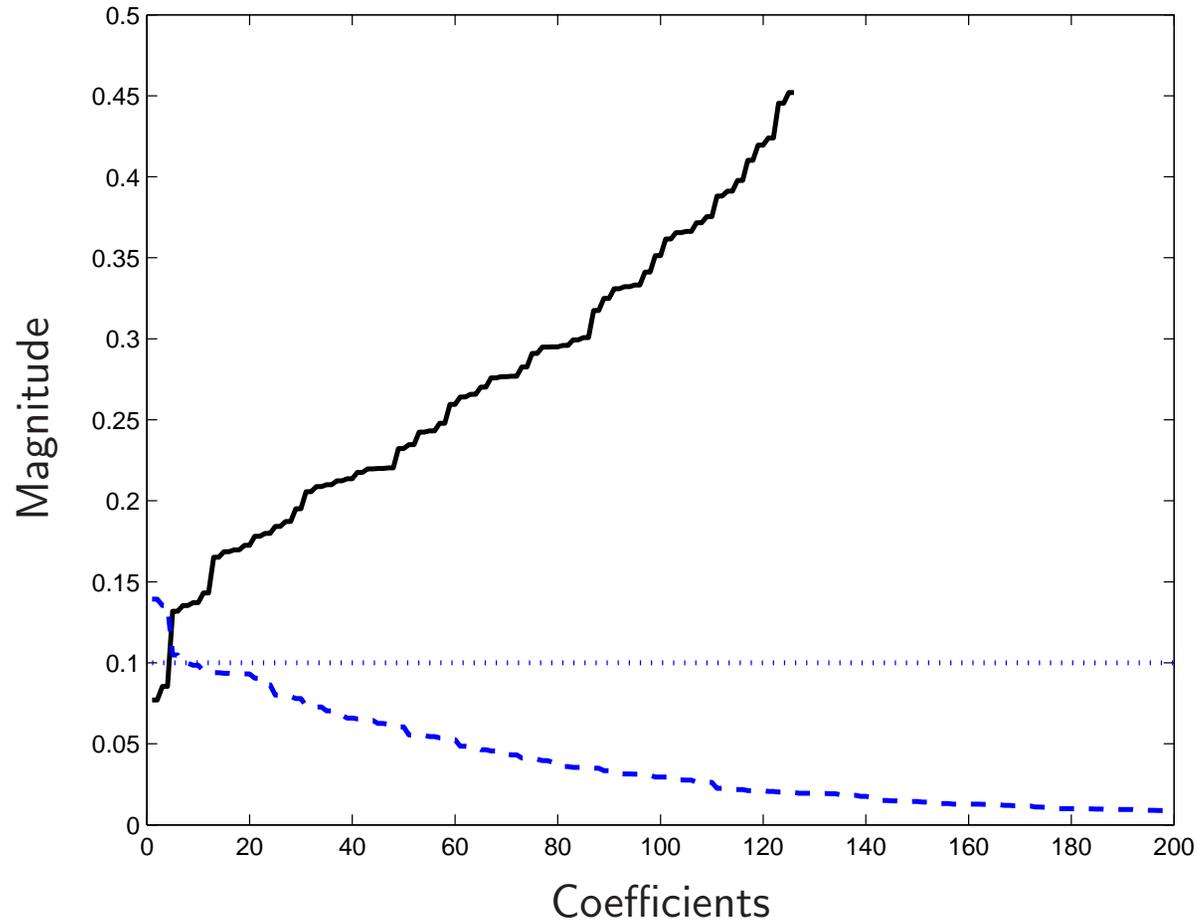
Solution for  $\rho = \sigma$

The original inverse covariance matrix  $A$ , the noisy inverse  $\Sigma^{-1}$  and the solution.

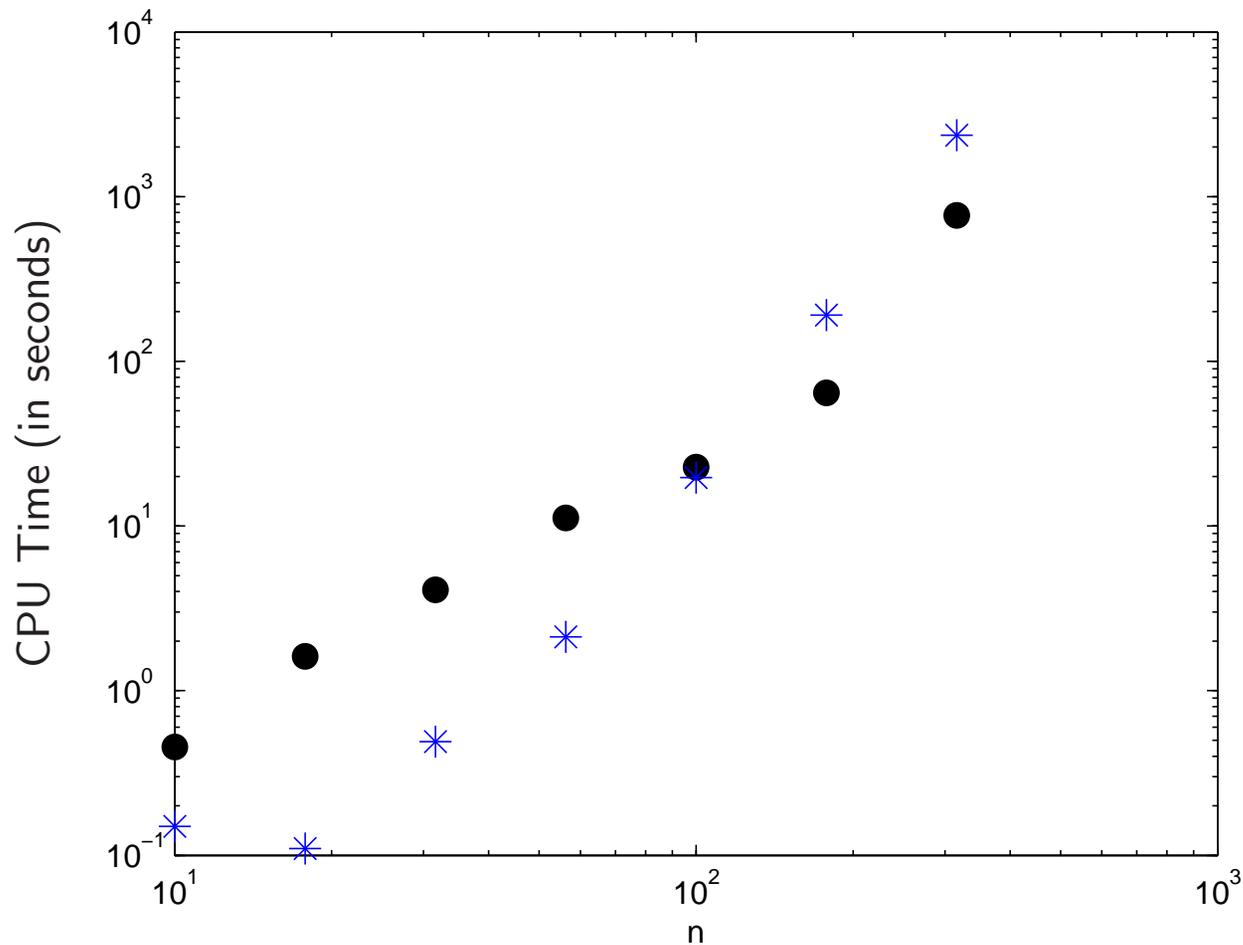


Average and standard deviation of the percentage of errors (false positives + false negatives) versus  $\log(\rho/\sigma)$  on random problems.

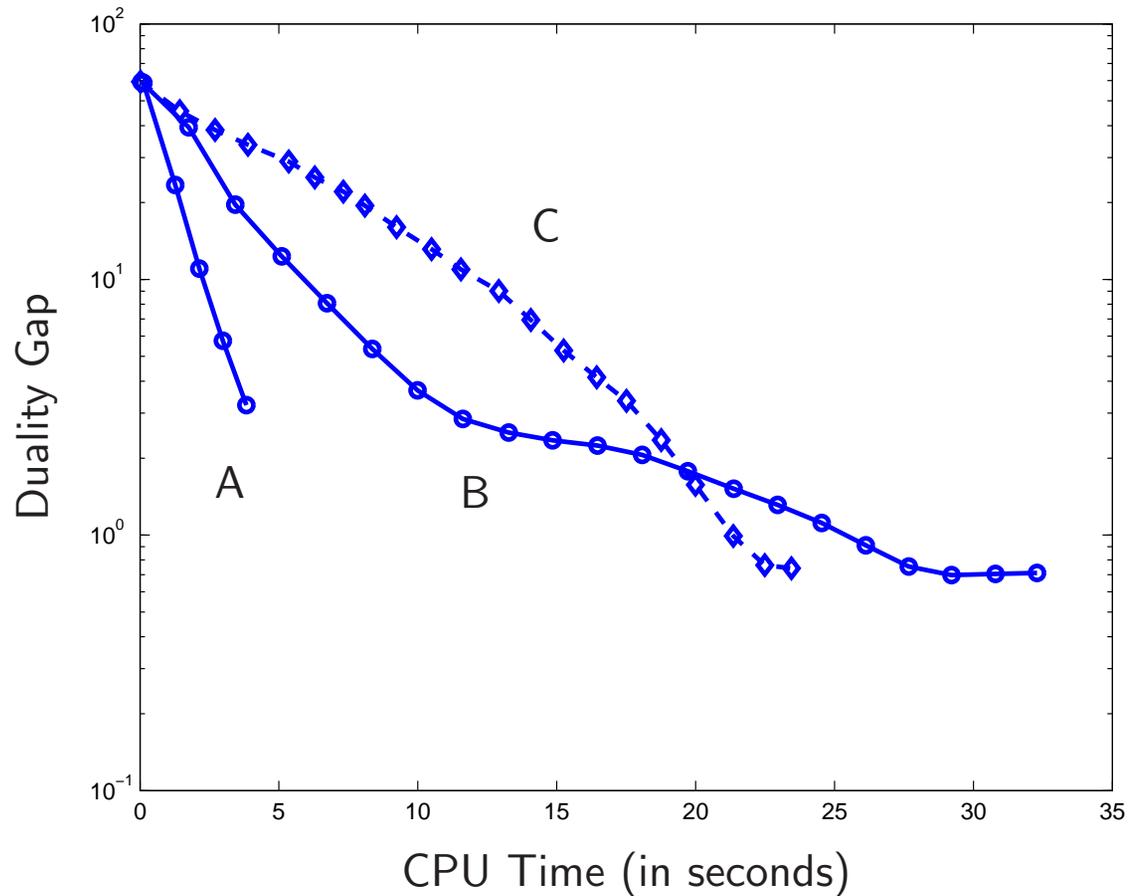




*Classification Error.* Magnitude of solution coefs associated with nonzero coefs in  $A$  (solid line) and magnitude of solution coefs associated with zero coefs in  $A$  (dashed line). Here  $n = 100$ .



*Computing time.* CPU time (in seconds) to reach gap of  $\epsilon = 1$  versus problem size  $n$  on random problems, solved using Nesterov's method (stars) and the coordinate descent algorithm (dots).



*Computing time.* Convergence plot a random problem of size  $n = 100$ , this time comparing Nesterov's method where  $\epsilon = 5$  (solid line A) and  $\epsilon = 1$  (solid line B) with one sweep of the block-coordinate descent method (dashed line C).

# Conclusion

- A convex relaxation for sparse covariance selection.
- Robustness interpretation.
- Two algorithms for dense large-scale instances.
- Precision requirements? Thresholding? . . .

Slides and software available *online* at [www.princeton.edu/~aspremon](http://www.princeton.edu/~aspremon)

# References

- Akaike, J. (1973), Information theory and an extension of the maximum likelihood principle, *in* B. N. Petrov & F. Csaki, eds, 'Second international symposium on information theory', Akademiai Kiado, Budapest, pp. 267–281.
- Bilmes, J. A. (1999), 'Natural statistic models for automatic speech recognition', *Ph.D. thesis, UC Berkeley, Dept. of EECS, CS Division* .
- Bilmes, J. A. (2000), 'Factored sparse inverse covariance matrices', *IEEE International Conference on Acoustics, Speech, and Signal Processing* .
- Chen, S. S. & Gopinath, R. A. (1999), 'Model selection in acoustic modeling', *EUROSPEECH* .
- 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. (1972), 'Covariance selection', *Biometrics* **28**, 157–175.
- 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**(1), 196–212.

- Dobra, A. & West, M. (2004), 'Bayesian covariance selection', *working paper* .
- Fazel, M., Hindi, H. & Boyd, S. (2000), 'A rank minimization heuristic with application to minimum order system approximation', *American Control Conference, September 2000* .
- Lemaréchal, C. & Sagastizábal, C. (1997), 'Practical aspects of the Moreau-Yosida regularization: theoretical preliminaries', *SIAM Journal on Optimization* **7**(2), 367–385.
- Nemirovski, A. (2004), 'Prox-method with rate of convergence  $o(1/t)$  for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle-point problems', *SIAM Journal on Optimization* **15**(1), 229–251.
- Nesterov, Y. (1983), 'A method of solving a convex programming problem with convergence rate  $O(1/k^2)$ ', *Soviet Math. Dokl.* **27**(2), 372–376.
- Nesterov, Y. (2005), 'Smooth minimization of nonsmooth functions', *Mathematical Programming, Series A* **103**, 127–152.