Sparse Covariance Selection using Semidefinite Programming

A. d'Aspremont
ORFE, Princeton University

Joint work with **O. Banerjee, L. El Ghaoui & G. Natsoulis**,
U.C. Berkeley & Iconix Pharmaceuticals

Support from EUROCONTROL and Google.

Introduction

We estimate a **sample covariance matrix** Σ from empirical data. . .

- Objective: infer **dependence** relationships between variables.
- We want this information to be as **sparse** as possible.
- Basic solution: look at the magnitude of the covariance coefficients:

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

and simply threshold smaller coefficients to zero. (not always psd.)

We can do better. . .

Covariance Selection

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

- **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 Σ^{-1} corresponds to a **sparse representation** of the density f as a member of an exponential family of distributions:

$$f(x,\Sigma) = \exp(\alpha_0 + t(x) + \alpha_{11}t_{11}(x) + \ldots + \alpha_{rs}t_{rs}(x))$$

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

- Dempster (1972) calls Σ^{-1}_{ij} a **concentration** coefficient.

There is more. . .

Covariance Selection

Covariance selection:

- With $m+1$ observations $x_i \in \mathbf{R}^n$ on n random variables, we estimate a sample covariance matrix S such that $S = \frac{1}{m}\sum_{i=1}^{m+1}(x_i - \bar{x})(x_i - \bar{x})^T$
- Choose a symmetric **subset** I of matrix coefficients and denote by J the remaining coefficients.
- Choose a covariance matrix estimator $\hat{\Sigma}$ such that:
  - $\hat{\Sigma}_{ij} = S_{ij}$ for all indices (i,j) in J
  - $\hat{\mathbf{\Sigma}}_{\mathbf{ij}}^{\mathbf{-1}} = \mathbf{0}$ for all indices (i,j) in I

We simply select a topology of zeroes in the inverse covariance matrix. . .

Covariance Selection

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

- **Maximum Entropy**. Among all Gaussian models Σ 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 Σ such that $\Sigma_{ij}^{-1} = 0$ on I, the choice $\hat{\Sigma}_{ij} = S_{ij}$ on J has **maximum likelihood**.

- **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.

Covariance Selection

Conditional independence:

- Suppose X, Y, Z have are jointly normal with covariance matrix Σ, with

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

  where $\Sigma_{11} \in \mathbf{R}^{2\times 2}$ and $\Sigma_{22} \in \mathbf{R}$.

- Conditioned on Z, X, Y are still normally distributed with covariance matrix C satisfying:

$$C = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \left(\Sigma^{-1}\right)_{11}^{-1}$$

- So X and Y are **conditionally independent** iff $\left(\Sigma^{-1}\right)_{11}$ is diagonal, which is also:

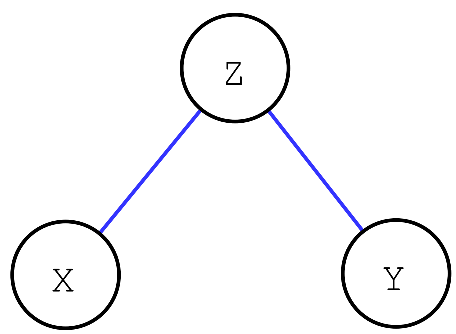
$$\Sigma_{xy}^{-1} = 0$$

Covariance Selection

- Suppose we have iid noise $\epsilon_i \sim \mathcal{N}(0, 1)$ and the following linear model:

$$\begin{aligned} x &= z + \epsilon_1 \\ y &= z + \epsilon_2 \\ z &= \epsilon_3 \end{aligned}$$

- Graphically, this is:



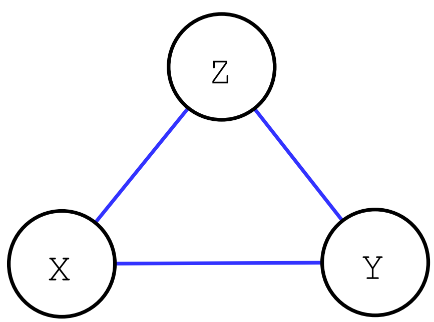
Covariance Selection

- The covariance matrix and inverse covariance are given by:

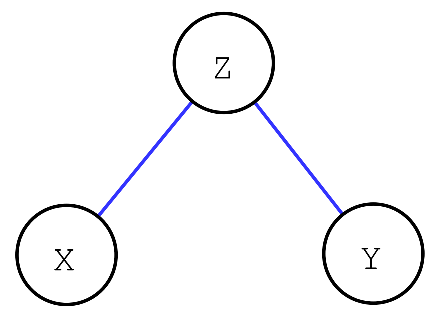
$$\Sigma = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{pmatrix} \qquad \Sigma^{-1} = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \\ -1 & -1 & 3 \end{pmatrix}$$

- The inverse covariance matrix has Σ_{12}^{-1} clearly showing that the variables x and y are independent conditioned on z.

- Graphically, this is again:

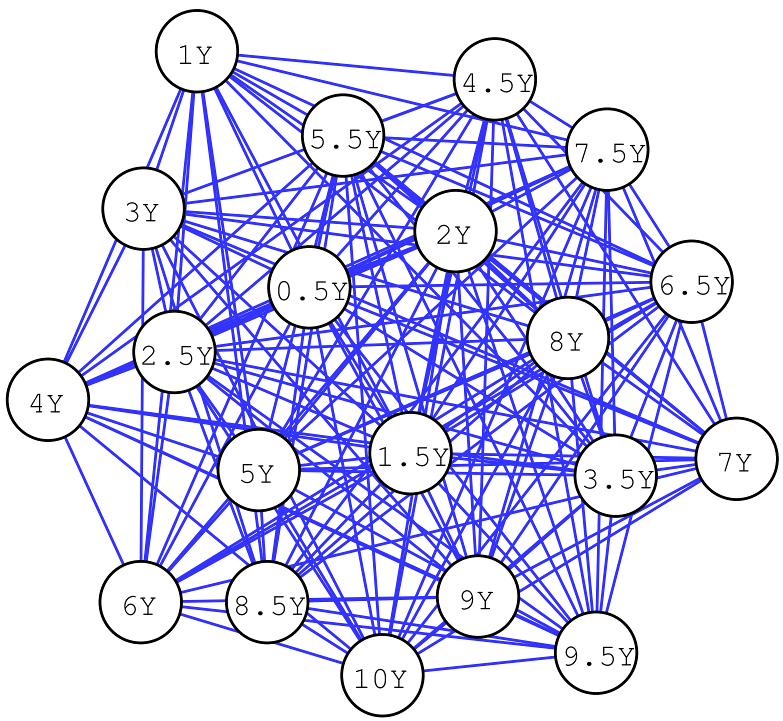


versus

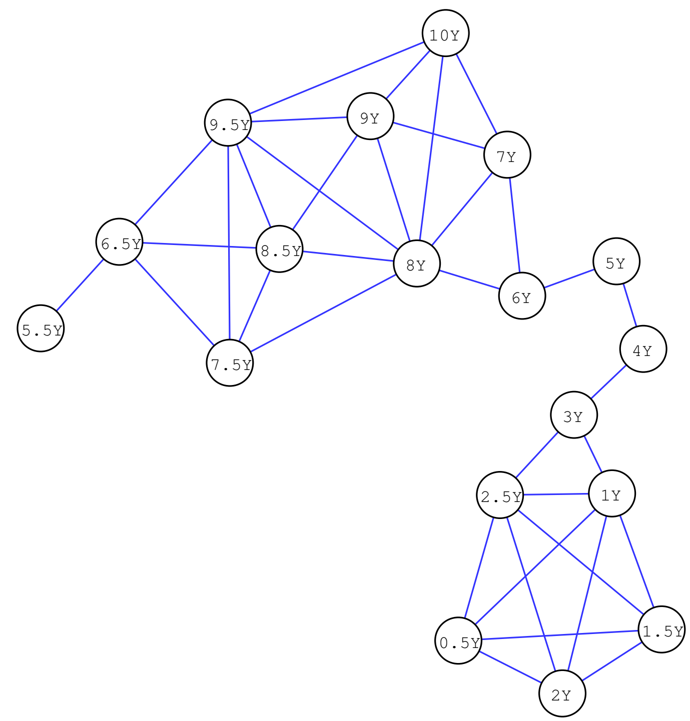


Covariance Selection

On a slightly larger scale. . .



Before



After

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**. Covariance estimation.

- Related work by Dahl, Roychowdhury & Vandenberghe (2005): interior point methods for large, sparse MLE.

- See also d'Aspremont, El Ghaoui, Jordan & Lanckriet (2005) on sparse principal component analysis (PCA).

Outline

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

Maximum Likelihood Estimation

- We can estimate Σ by solving the following maximum likelihood problem:

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

- This problem is convex, has an explicit answer $\Sigma = S^{-1}$ if $S \succ 0$.
- Problem here: how do we make Σ^{-1} **sparse**?
- 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.

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

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

- Classic l_1 heuristic: $\|X\|_1$ is a **convex lower bound** on $\mathbf{Card}(X)$.
- See Fazel, Hindi & Boyd (2001) for related applications.

l_1 relaxation

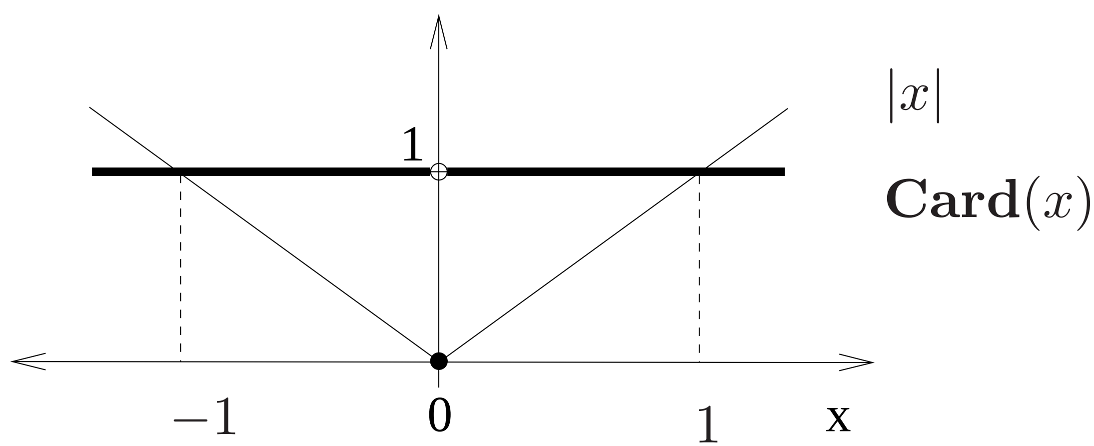
Assuming $|x| \leq 1$, this relaxation replaces:

$$\mathbf{Card}(x) = \sum_{i=1}^{n} 1_{\{x_i \neq 0\}}$$

with

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|$$

Graphically, this is:



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 ρ on the elements of S.

- The relaxed **sparsity** requirement is equivalent to a **robustification**.

- See d'Aspremont et al. (2005) 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**, **dense** instances. . .
- The $\|X\|_1$ penalty implicitly introduces $O(n^2)$ linear constraints and makes interior point methods too expensive.

Algorithms

Complexity options. . .

$O(n)$ $O(n)$ $O(n^2)$

Memory

$O(1/\epsilon^2)$ $O(1/\epsilon)$ $O(\log(1/\epsilon))$

First-order Smooth Newton IP **Complexity**

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

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

- **Regularization**. Add strongly convex penalty inside the min-max representation to produce an ϵ-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 the solve the regularized problem.

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

Nesterov's method

- Numerical steps: computing the **inverse** of X and two eigenvalue decompositions.

- Total complexity estimate of the method is:

$$O\left(\frac{\kappa\sqrt{(\log\kappa)}}{\epsilon}n^{4.5}\alpha\rho\right)$$

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:

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

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

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

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

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

$$\begin{array}{ll} \text{minimize} & u^T (C^{11})^{-1} u \\ \text{subject to} & \|u\|_\infty \leq \rho \end{array}$$

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

Dual block-coordinate descent

Complexity?

- Luo & Tseng (1992): block coordinate descent has linear convergence in this case.

Smooth first-order methods to solve the inner QP problem:

- The hardest numerical step at each iteration is computing an inverse.
- The matrix to invert is only updated by a low rank matrix at each iteration: use Sherman-Woodbury-Morrisson formula.

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 σ to its inverse

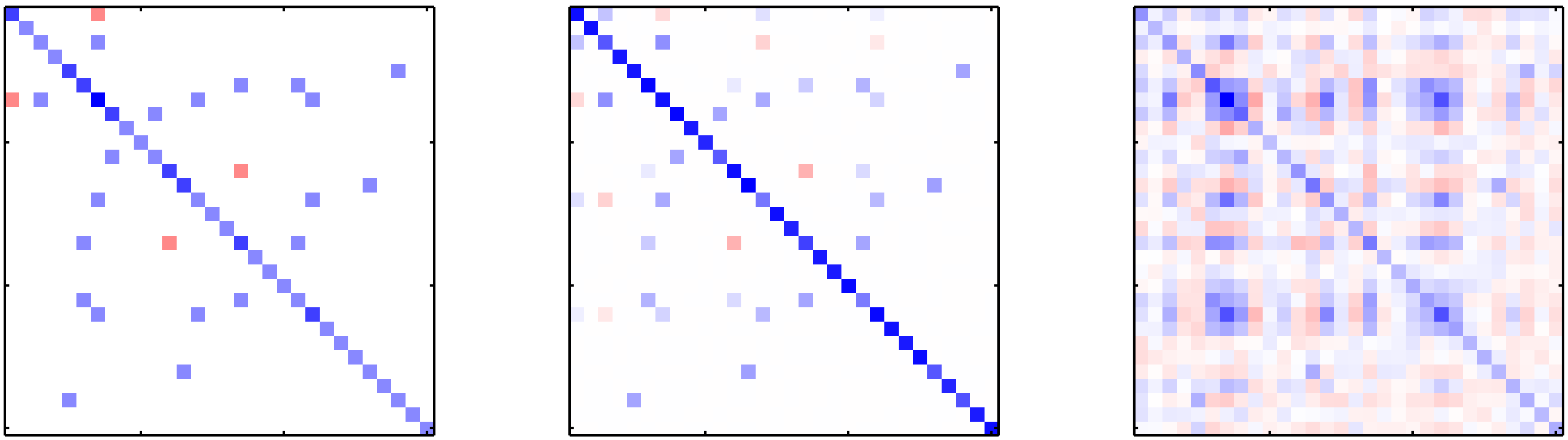
We then solve the penalized MLE problem (or the modified one):

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

and compare the solution with the original matrix A.

Numerical Examples

A basic example. . .

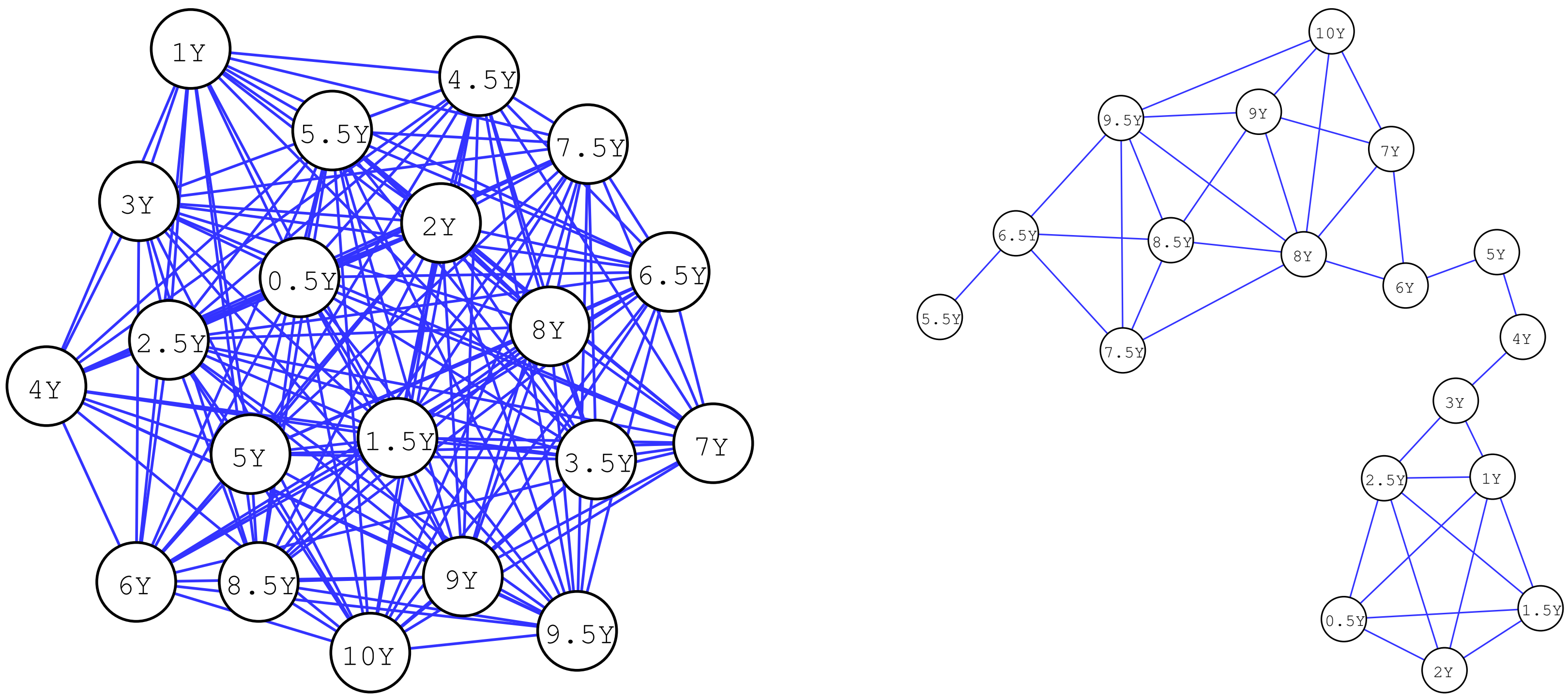


Original inverse A | Solution for $\rho = \sigma$ | Noisy inverse Σ^{-1}

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

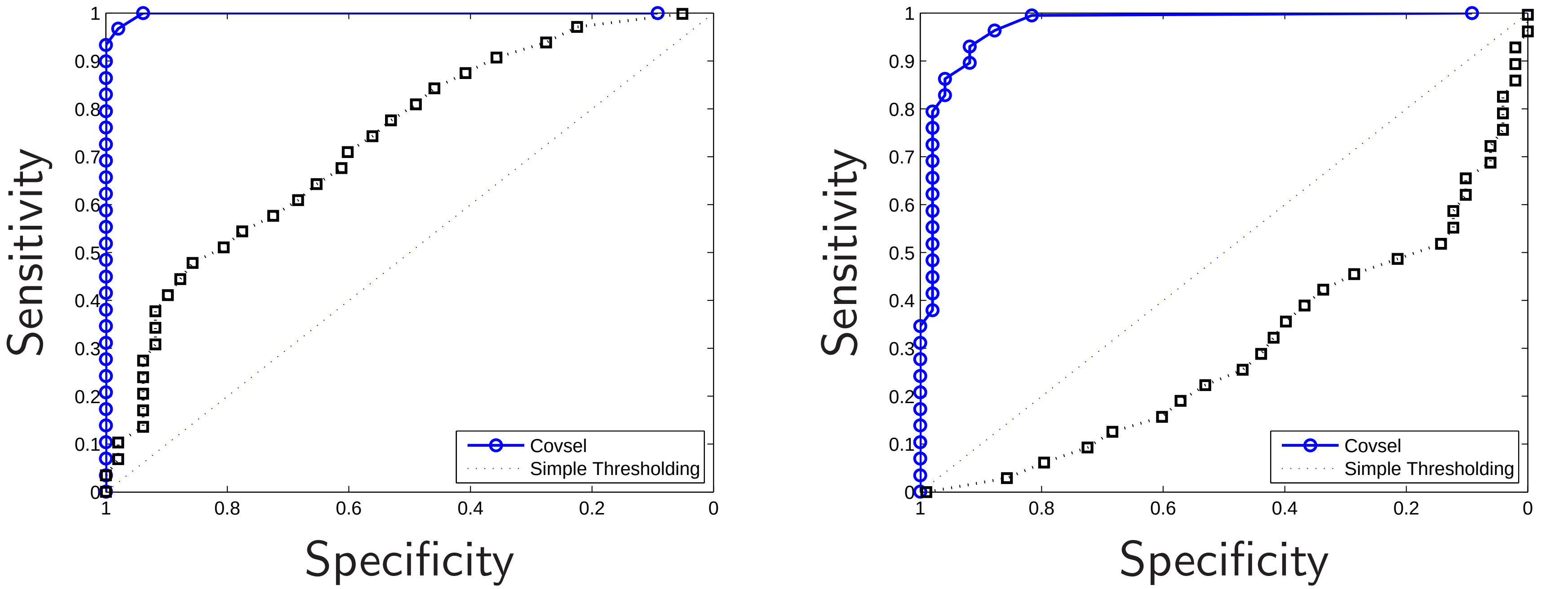
Covariance Selection

Forward rates covariance matrix for maturities ranging from 0.5 to 10 years.

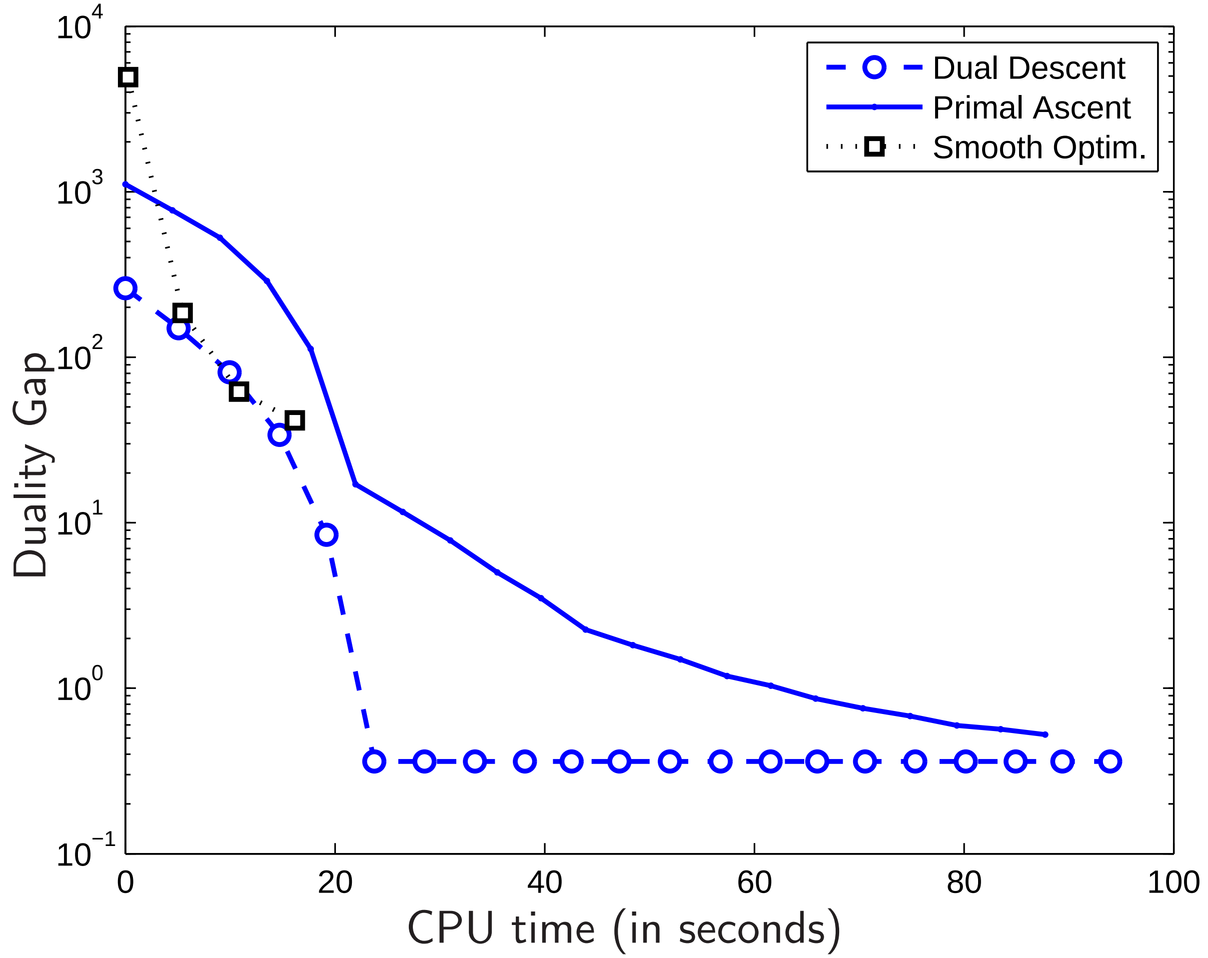


$\rho = 0$ $\rho = .01$

ROC curves



Classification Error. ROC curves for the solution to the covariance selection problem compared with a simple thresholding of B^{-1}, for various levels of noise: $\sigma = 0.3$ (left) and $\sigma = 0.5$ (right). Here $n = 50$.



Computing time. Duality gap versus CPU time (in seconds) on a random problem, solved using Nesterov's method (squares) and the coordinate descent algorithms (circles and solid line).

Conclusion

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

If you have financial applications in mind. . .

Network graphs generated using Cytoscape.

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', Akedemiai 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. (2005), 'A direct formulation for sparse PCA using semidefinite programming', *Advances in Neural Information Processing Systems* **17**, 41–48.

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. (2001), 'A rank minimization heuristic with application to minimum order system approximation', *Proceedings American Control Conference* **6**, 4734–4739.

Lemaréchal, C. & Sagastizábal, C. (1997), 'Practical aspects of the Moreau-Yosida regularization: theoretical preliminaries', *SIAM Journal on Optimization* **7**(2), 367–385.

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**(1), 7–35.

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 Mathematics Doklady* **27**(2), 372–376.

Nesterov, Y. (2005), 'Smooth minimization of nonsmooth functions', *Mathematical Programming, Series A* **103**, 127–152.