Exploring patterns of dependence in financial data.

Alexandre d’Aspremont
CNRS – CMAP, Ecole Polytechnique

Joint work with O. Banerjee, L. El Ghaoui, U.C. Berkeley.

Support from NSF, ERC SIPA and Google.
We estimate a sample covariance matrix $\Sigma$ from empirical data. . .

- Objective: infer dependence relationships between variables.
- We only want to isolate a few key links.

Elementary solution: look at the magnitude of the covariance coefficients:

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

then simply threshold smaller coefficients to zero (not always psd).
Covariance Selection

Before

After
Covariance Selection

Following Dempster [1972], look for zeros in the inverse covariance matrix:

Parsimony. Suppose that we are estimating a Gaussian density:

\[
f(x, \Sigma) = \frac{1}{\sqrt{2\pi}} \frac{1}{\text{det } \Sigma} \exp \left(-\frac{1}{2} x^T \Sigma^{-1} x \right),
\]

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

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

with here \(t_{ij}(x) = x_i x_j\) and \(\alpha_{ij} = \Sigma_{ij}^{-1}\). Dempster [1972] calls \(\Sigma_{ij}^{-1}\) a concentration coefficient.
Conditional independence.

- Suppose $X, Y, Z$ have are jointly normal with covariance matrix $\Sigma$, with

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

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

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

\[
C = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} = \Sigma_{11}^{-1}
\]

- So $X$ and $Y$ are conditionally independent iff $\Sigma_{11}^{-1}$ is diagonal, which is also

\[
\Sigma_{xy}^{-1} = 0
\]
Suppose we have iid noise $\epsilon_i \sim \mathcal{N}(0, 1)$ and the following linear model

\[
\begin{align*}
x &= z + \epsilon_1 \\
y &= z + \epsilon_2 \\
z &= \epsilon_3
\end{align*}
\]

Graphically, this is
Covariance Selection

- The covariance matrix and inverse covariance are given by

\[
\Sigma = \begin{pmatrix}
2 & 1 & 1 \\
1 & 2 & 1 \\
1 & 1 & 1 \\
\end{pmatrix} \quad \Sigma^{-1} = \begin{pmatrix}
1 & 0 & -1 \\
0 & 1 & -1 \\
-1 & -1 & 3 \\
\end{pmatrix}
\]

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

- Graphically, this is again
Let $I^L J = [1, n]^2$, Dempster [1972] shows:

- **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**.

- **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 $I$. 

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 et al. [2004], Dobra and West [2004] for example.

- **Speech Recognition.** See Bilmes [1999], Bilmes [2000] or Chen and Gopinath [1999].

- Related work by Dahl et al. [2005]: interior point methods for sparse MLE.
Financial data

Estimating covariance matrices from financial data.

- Asset returns are given by (schematically)

\[ \Delta S_t = \Delta M_t + \epsilon_t \]

where

- \( M_t \) is the market return
- \( \epsilon_t \) is an idiosyncratic component

- All assets are usually highly correlated: \( M_t \) dominates the picture. We are only interested in the correlation between \( \epsilon_t \) for various assets.

- The inverse matrix is also used to compute portfolios on the efficient frontier for \( \text{CAPM} \).
Outline

- Introduction
- Penalized maximum likelihood estimation
- Algorithms & complexity
- Consistency
- Graph layout
- Numerical experiments
Penalized Maximum Likelihood Estimation
Akaike [1973]: penalize the likelihood function:

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

where $\text{Card}(X)$ 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...
We can form a convex relaxation of AIC or BIC penalized MLE

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

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

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

Classic $l_1$ heuristic: $\|X\|_1$ is a convex lower bound on $\text{Card}(X)$.

Outline

- Introduction
- Penalized maximum likelihood estimation
- Algorithms & complexity
- Consistency
- Graph layout
- Numerical experiments
The problem

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

is convex in the variable $X \in S_n$. This means that we can get explicit complexity bounds and efficient algorithms.

- Standard convex optimization algorithms easily solve small instances. (see Boyd and Vandenberghe [2004])
- Specialized techniques solve larger problems with complexity $O(n^{4.5})$. We can exploit the block structure of the dual. Cost per iteration comparable to that of a penalized regression (LASSO).
- In practice, we can get a good solution with complexity $O(n^{3.5})$. A bit harder than computing a matrix inverse. . .
Complexity options...
The convex relaxation of the covariance selection problem has a particular min-max structure

\[
\max_{X \in S^n} \min_{|U_{ij}| \leq \rho} \log \det X - \text{Tr}((S + U)X)
\]

This min-max representation 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 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 and 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.
Nesterov’s method

Regularization. The objective is first smoothed by penalization. We solve the following (modified) problem

\[
\max \{X \in S^n: \alpha I_n \preceq X \preceq \beta I_n\} \quad \min \{U \in S^n: |U_{ij}| \leq \rho\}
\]

\[
\log \det X - \text{Tr}((S - U)X) - (\epsilon/2D_2)d_2(U)
\]

an \(\epsilon\) approximation of the original problem if \(\alpha \leq 1/(\|S\| + n\rho)\) and \(\beta \geq n/\rho\).

- Prox on \(Q_2 := \{U \in S^n: \|U\|_\infty \leq 1\}\) is \(d_2(U) = \frac{1}{2} \text{Tr}(U^T U) = \frac{1}{2}\|U\|^2\)
- Prox \(d_1(X)\) for the set \(\{\alpha I_n \preceq X \preceq \beta I_n\}\) given by

\[
d_1(X) = -\log \det X + \log \beta
\]

This corresponds to a classic Moreau-Yosida regularization of the penalty \(\|X\|_1\) and the function \(f\) has a Lipschitz continuous gradient with constant

\[
L := M + D_2\rho^2/(2\epsilon)
\]
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 $k = 0, \ldots, N(\epsilon)$ do

1. Compute $\nabla f(X_k) = -X^{-1} + \Sigma + U^*(X_k)$

2. Find $Y_k = \arg\min_Y \{\text{Tr}(\nabla f(X_k)(Y - X_k)) + \frac{1}{2}L \|Y - X_k\|_F^2 : Y \in Q_1\}$.

3. Find $Z_k = \arg\min_X \sum_{i=0}^{k} \frac{i+1}{2} \text{Tr}(\nabla f(X_i)(X - X_i)) : X \in Q_1$.

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

At each iteration

■ **Step 1:** only amounts 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 $Q_1 = \{\alpha I_n \preceq X \preceq \beta I_n\}$ and require an eigenvalue decomposition.

This means that the total complexity estimate of the method is

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

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

\[
\begin{align*}
\text{maximize} & \quad \log \det (S + U) \\
\text{subject to} & \quad \|U\|_\infty \leq \rho \\
& \quad S + U \succeq 0
\end{align*}
\]

- Let \( C = S + U \) be the current iterate, after permutation we can always assume that we optimize over the last column

\[
\begin{align*}
\text{maximize} & \quad \log \det \begin{pmatrix} C^{11} & C^{12} + u \\ C^{21} + u^T & C^{22} \end{pmatrix} \\
\text{subject to} & \quad \|u\|_\infty \leq \rho
\end{align*}
\]

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

- Each iteration reduces to a simple box-constrained QP

\[
\begin{align*}
\text{minimize} & \quad u^T (C^{11})^{-1} u \\
\text{subject to} & \quad \|u\|_\infty \leq \rho
\end{align*}
\]
Outline

- Introduction
- Penalized maximum likelihood estimation
- Algorithms & complexity
- Consistency
- Graph layout
- Numerical experiments
Consistency

Proposition 1

**Consistency.** Let $\hat{C}_k^\lambda$ denote our estimate of the connectivity component of node $k$. Let $\alpha$ be a given level in $[0,1]$. Consider the following choice for the penalty parameter

$$
\lambda(\alpha) := \left( \max_{i > j} \hat{\sigma}_i \hat{\sigma}_j \right) q \frac{t_{n-2}(\alpha/2p^2)}{n - 2 + t_{n-2}^2(\alpha/2p^2)}
$$

where $t_{n-2}(\alpha)$ denotes the $(100 - \alpha)$% point of the Student’s t-distribution for $n - 2$ degrees of freedom, and $\hat{\sigma}_i$ is the empirical variance of variable $i$. Then

$$
\Pr(\exists k \in \{1, \ldots, p\} : \hat{C}_k^\lambda \not\subseteq C_k) \leq \alpha.
$$

**Proof.** Argument similar to Meinshausen and Buhlmann [2006].
In practice, we can use **cross-validation**

- Remove a random subset of the variables and compute the inverse covariance matrix.
- Compute the pattern of zeros.
- Repeat the procedure for various variable subsets and various values of the penalty $\rho$.

**How do we pick the value of the penalty parameter $\rho$?**

- We pick the $\rho$ minimizing the variability of these dependence relationships across samples.
- Also, dependence relationships which show up in most subsampled networks are considered more reliable.
Outline

- Introduction
- Penalized maximum likelihood estimation
- Algorithms & complexity
- Consistency
- Graph layout
- Numerical experiments
How do we represent these results?

- Turn the pattern of zeros in the inverse covariance into a graph.
- Use graph visualization algorithms to layout this graph.

Trickier than it sounds. . .

- Graph layout problems are usually very hard. Again, good approximation algorithms exist.
- Many possible representations.
- Some coefficients are close to zero (numerical noise): threshold.
Many characteristics of the graph have a statistical interpretation.

- if the graph is **chordal**, then there is a linear/Gaussian model with the same sparsity pattern (see Wermuth [1980] for an early reference on linear recursive models and path analysis).

*Left:* a chordal graphical model: no cycles of length greater than three.  
*Right:* a non-chordal graphical model of U.S. swap rates.
Network Interpretation

- If there is a path between two nodes on a graph, then the corresponding variables have nonzero covariance (see Gilbert [1994] for a survey of graph theory/sparse linear algebra).

*Left:* connected model of U.S. swap rates, with dense covariance matrix.  
*Right:* disconnected model, the covariance matrix is block-diagonal.
Outline

- Introduction
- Penalized maximum likelihood estimation
- Algorithms & complexity
- Consistency
- Graph layout
- Numerical experiments
Sparse covariance model. *Left:* ROC curves for both thresholding and covariance selection using 20 samples to compute the covariance. *Right:* Binary dependence classification performance of inverse sample covariance thresholding (THRES) and covariance selection (COVSEL) for various sample sizes, measured by area under ROC curve.
Forward rates covariance matrix for maturities ranging from 0.5 to 10 years.

\[
\rho = 0 \quad \rho = .01
\]
Graph of conditional covariance among a cluster of U.S. dollar exchange rates. Positive dependencies are plotted as green links, negative ones in red, thickness reflects the magnitude of the covariance.
Hedge fund returns

- We track 116 hedge funds between January 1995 and December 2005.
- Monthly hedge fund returns from the Center for International Securities and Derivatives Markets hedge fund database, via WRDS.
- Hedge fund nodes are colored to represent their primary strategy.
Hedge fund returns: strategies
Hedge fund returns: markets

A. d’Aspremont, TSE, November 2011.
Covariance selection highlights key dependence structure.

Very good statistical performance compared to thresholding techniques.

Results are often intuitive.

Slides, papers and MATLAB software available at:

http://www.cmap.polytechnique.fr/~aspremon

R package using a pathwise algorithm at

http://cran.r-project.org/web/packages/Covpath/index.html

A free network layout software called cytoscape:

http://www.cytoscape.org
References


A. d’Aspremont, TSE, November 2011.