Sparse Graphical Models and the US Senate

Laurent El Ghaoui
<elghaoui@eecs.berkeley.edu>

ORFE Colloquium, Princeton University

October 11, 2007

Outline

- Fitting Gaussian graphical models:
  - Sparse covariance selection
  - Connection to LASSO regression
- Fitting binary graphical models
- Examples

Sparse Covariance Selection

- Draw n independent samples $y_i \sim \mathcal{N}_p(0, \Sigma)$, where Σ is unknown.
- *Prior belief*: many conditional independencies among the variables in this distribution.
- Zeros in inverse covariance correspond to conditional independence properties among variables.
- *Covariance estimation:*: From $y_1, \ldots, y_n$, recover the covariance matrix Σ.
- *Covariance selection*: choosing which elements of our estimate $\hat{\Sigma}^{-1}$ to set to zero.

Penalized Maximum-Likelihood Approach

Penalized ML problem:

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

- $\rho > 0$ is regularization parameter, and $\|X\|_1 := \sum_{i,j} |X_{ij}|$.
- Convex, non-smooth problem.
- Same idea used in l_1-norm penalized regression (LASSO), for example.

Robustness and Regularization

We can write our problem as

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

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

$$\min_{U} \{ -\log\det(S+U) - p : |U_{ij}| \leq \rho, S+U \succ 0 \} \tag{1}$$

Robustness and Regularization

- The dual problem 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., $n < p$):

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

Algorithms: Interior-Point Methods

- Problem is amenable to interior point methods (e.g. MAXDET).
- *Complexity* is very high: $\mathcal{O}(p^6 \log(1/\epsilon))$.
- 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.

Algorithms: First-Order Methods

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

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

- *Block-Coordinate Ascent Algorithm*
  - Efficient algorithm with good empirical perfomance.
  - Attains high accuracy quickly.
- *Nesterov's first order method for non-smooth minimization*:
  - Purpose: use Nesterov's formalism to derive a rigorous complexity estimate, better than that offered by interior point methods.

Dual Block-Coordinate Ascent

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

$$\max\{\log\det W : |w_{ij} - s_{ij}| \leq \rho, W \succ 0\} \tag{2}$$

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

Column/Row Update Rule

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}:

$$\hat{w}_{12} := \arg\min_y \{ y^T W_{11}^{-1} y : \|y - s_{12}\|_\infty \leq \rho \} \tag{3}$$

Above is a QP (complexity: $O(p^3)$)

Convergence & Complexity

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

$$\mathbf{Tr}(SX) + \rho\|X\|_1 \leq p + \epsilon \;\; (X = W^{-1})$$

- *Complexity* of K sweeps through all columns: $\mathcal{O}(Kp^4)$

Connection to LASSO

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

$$\min_x x^T W_{11} x - s_{12}^T x + \rho \|x\|_1 \tag{4}$$

- 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

$$\min_x \|y - Qx\|_2^2 + \rho \|x\|_1 \tag{5}$$

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

Related Approach

Differences with the approach of Meinshaussen and Bühlmann (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

- Maximum number of iterations required to achieve solution:

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

- 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).

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.

Recovering Zero Pattern Masked by Noise

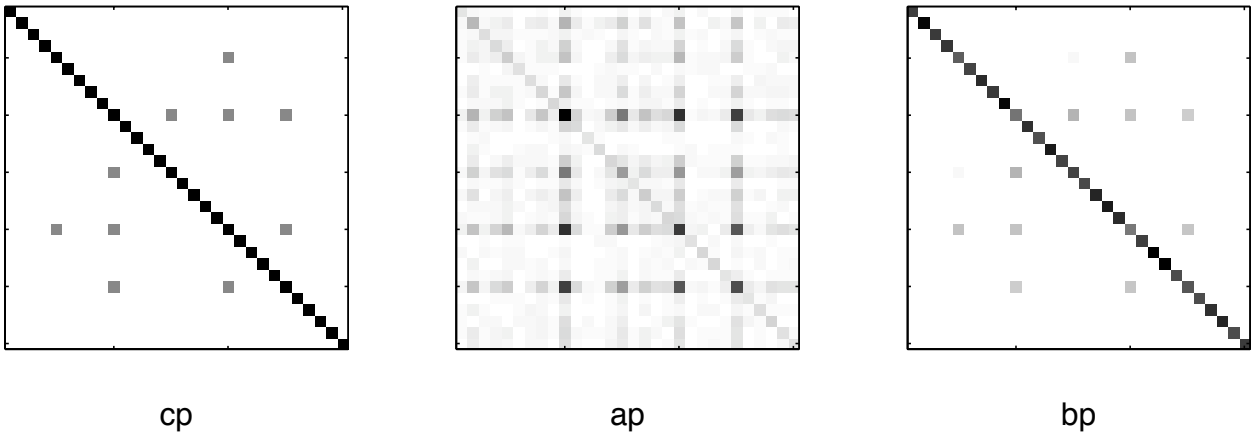
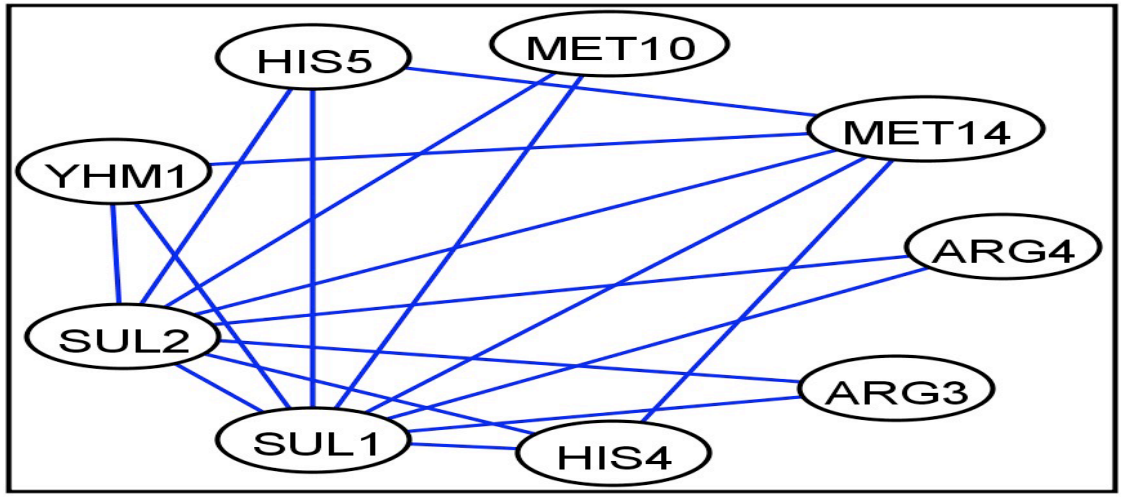


Figure 1: Recovering the sparsity pattern. We plot the original inverse covariance matrix A, the noisy inverse Σ^{-1} and the solution to the sparse ML problem for $\rho = .13$.

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.

Iconix 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.

Iconix 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

Iconix Data

- 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.

Connection between binary ASML and Gaussian SML

- Let S be the empirical covariance matrix, and let ρ be the parameter that controls the size of the penalty.

- $\hat{X}_{SML}(S;\rho) :=$ covariance estimate obtained using Gaussian SML
  $\hat{\Gamma}_{ASML}(S;\rho) :=$ covariance estimate obtained using binary ASML.

$$\hat{X}_{SML}(S;\rho) = \hat{\Gamma}_{ASML}(S + (\rho - \tfrac{1}{3})I;\rho) \quad \text{for } \rho > \tfrac{1}{3}$$

$$\hat{X}_{SML}(S - (\rho - \tfrac{1}{3})I;\rho) = \hat{\Gamma}_{ASML}(S;\rho) \quad \text{for } \rho \leq \tfrac{1}{3}$$

- We can therefore reuse the algorithms developed for the Gaussian SML problem.

Choice of Regularization Parameter ρ

- For any node k in the true undirected graphical model, let C_k denote its connectivity component (the set of nodes that are connected, through some chain of edges, to node k).

- Let $\hat{C}_k^\rho$ denote the estimate of C_k we obtain from our method using penalty parameter ρ.

- We can adapt the work of Meinshausen and Bühlmann (2005) to derive a formula for $\rho(\alpha)$ such that

$$P(\exists k : \hat{C}_k^\rho \subseteq C_k) \leq \alpha$$

Related Approach: Logistic Regression

- Just as LASSO can be used in the Gaussian case to estimate the set of neighbors of a particular node, we can use ℓ_1-norm penalized *logistic regression* to estimate the set of neighbors in the discrete case.

- Wainwright, Ravikumar and Lafferty (2006) have shown that penalized logistic regression yields an asymptotically consistent estimate of the set of neighbors of each node (discrete analog of Meinshausen and Bühlmann (2005)).

- Further work includes comparison of penalized logistic regression and sparse maximum likelihood methods.

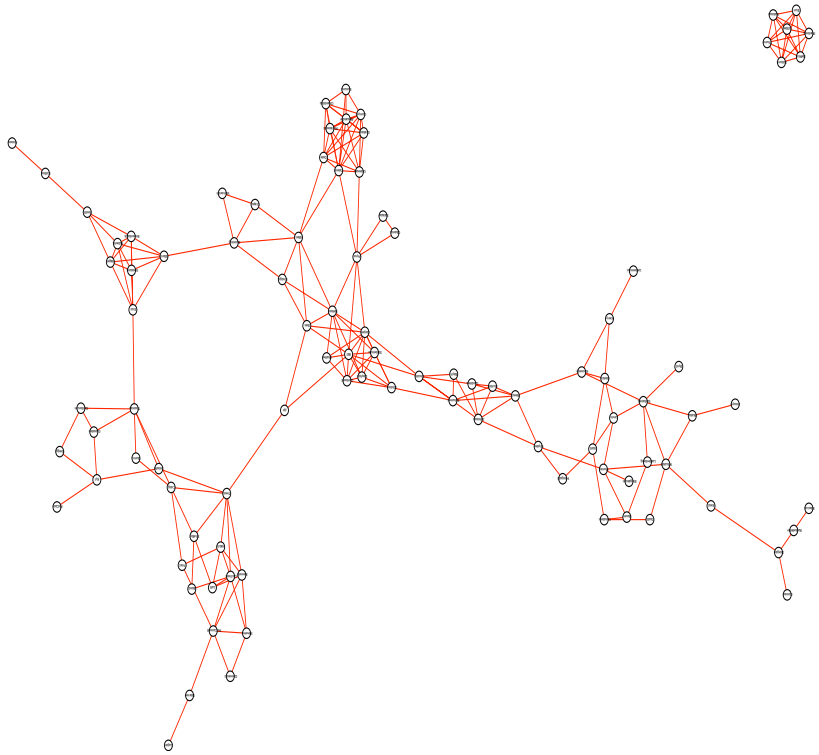
US Senate Voting Example

- *US Senate Voting Records Dataset*: US Senate voting records from the 109th Congress (2004 - 2006).
- 100 variables, 542 samples. Each sample is a bill that was put to a vote.
- Records 1 for yes, -1 for no on each bill.
- Goal: Treating each senator as a random variable, obtain a sparse undirected graphical model by applying the Approximate Sparse Maximum Likelihood (ASML) method.

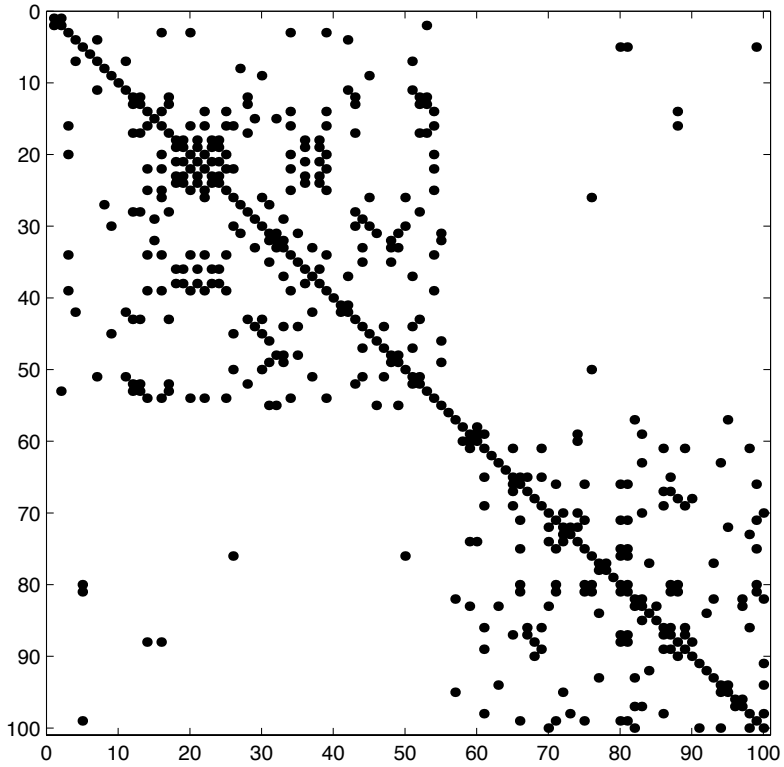
Obtaining a graph of senators

- Replaced missing data values with no votes (-1) for this application.
- Chose $\rho(\alpha)$ according to formula described above, with $\alpha = 0.05$.
- Sorted senators by party (Rep., Ind., Dem.) Two permutations of rows and columns of the solution are shown in the next few slides:
- (1) Sorted by region (Northeast to West) within each party. The ordering of the states by region is according to the US Census.
- (2) Sorted to reduce the spasity bandwidth within each party.

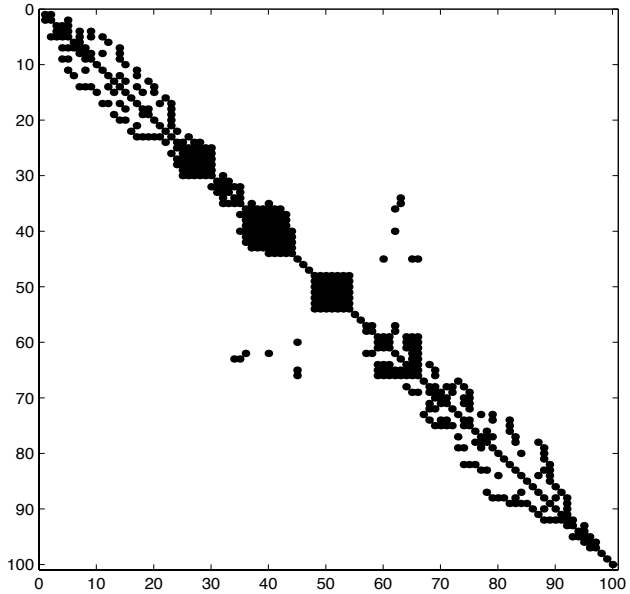
Sparsity pattern: a graph of senators



Sparsity pattern: senators sorted by region



Sparsity pattern: senators sorted for low bandwidth



Some tentative observations about the solution

- If we sort by region within each party, there is no obvious pattern except that senators from the same state and party are often (but not always) predictors for each other.

- Among Republicans, there are at least three large cliques, identified by solid squares along the diagonal of the low bandwidth plot. There is one Republican clique that is not connected to any other senators: Martinez, DeMint, Thune, Vitter, Burr, Coburn, and Isakson.

- There seem to be fewer similarly large cliques among Democrats.

Some tentative observations about the solution

- The neighbors for Chafee (R) are Democrats Cantwell, Bill Nelson, Carper, and Lincoln, and no Republicans. This matches statements made by and about Chafee in newspapers from 2004 - 2006.

- Democrat Pryor has among his neighbors Republicans Talent and Coleman. Conservative Democrat Ben Nelson has among his neighbors Republicans Allen and Ensign. These, along with the example of Chafee are the only direct interparty connections.

- The set of Republicans, excluding Chafee, Talent, Coleman, Allen and Ensign, is independent of the set of Democrats.

Some tentative observations about the solution

- At least four senators from the 109th congress have announced they are, or may be, running for president in 2008:
- Given the vote of his fellow Arizona Republican Kyl, the vote of McCain is independent of the votes of all other senators, Republican or Democrat.
- Obama is part of a small clique consisting of Democrats Salazar and Pryor.
- Clinton is part of the one large Democratic clique consisting of Carper, Bill Nelson, Stabenow, Dayton, and Cantwell.

- Brownback is part of a large Republican clique consisting of Sessions, Allard, Hagel, Enzi, and Roberts.

Final Remarks

- Connections between large-scale, numerical convex optimization and statistics are now well known and acted upon
- We explored connections between sparsity, regularization and robustness
- l_1-norm regularization useful in unsupervised (matrix) problems, for better *interpretability of the results*
- These problems can be formulated as robust optimization problems with interval data