Graphical Model Selection

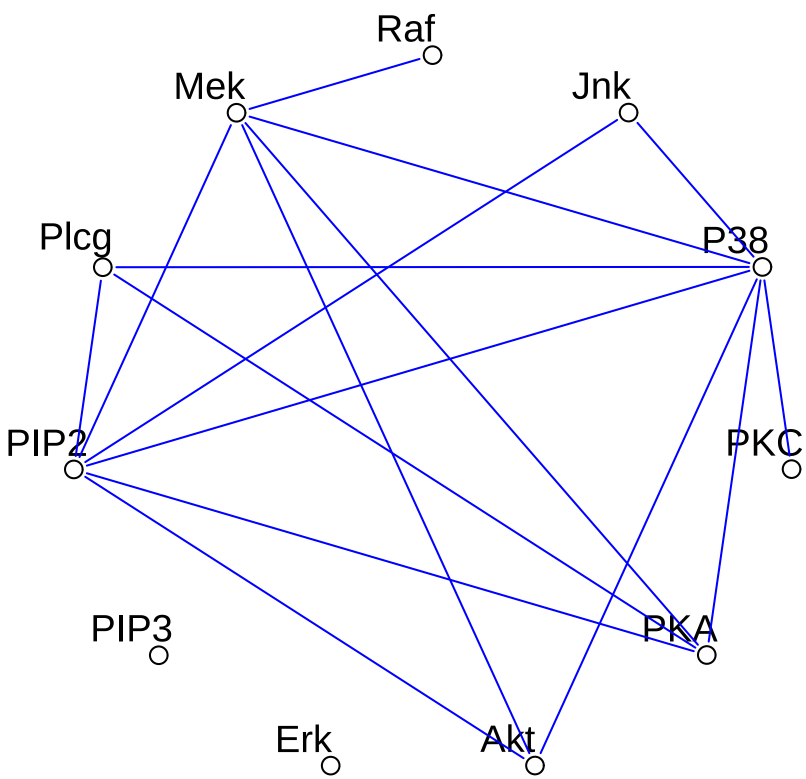
*Trevor Hastie**

Stanford University

** joint work with Jerome Friedman, Rob Tibshirani, Rahul Mazumder and Jason Lee*

Outline

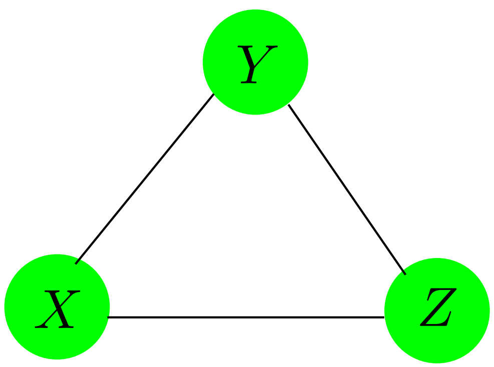
- Undirected Graphical Models
- Gaussian models for quantitative variables
- Estimation with known structure
- Estimating the structure via L_1 regularization
- Log-linear models for qualitative variables.



Flow Cytometry

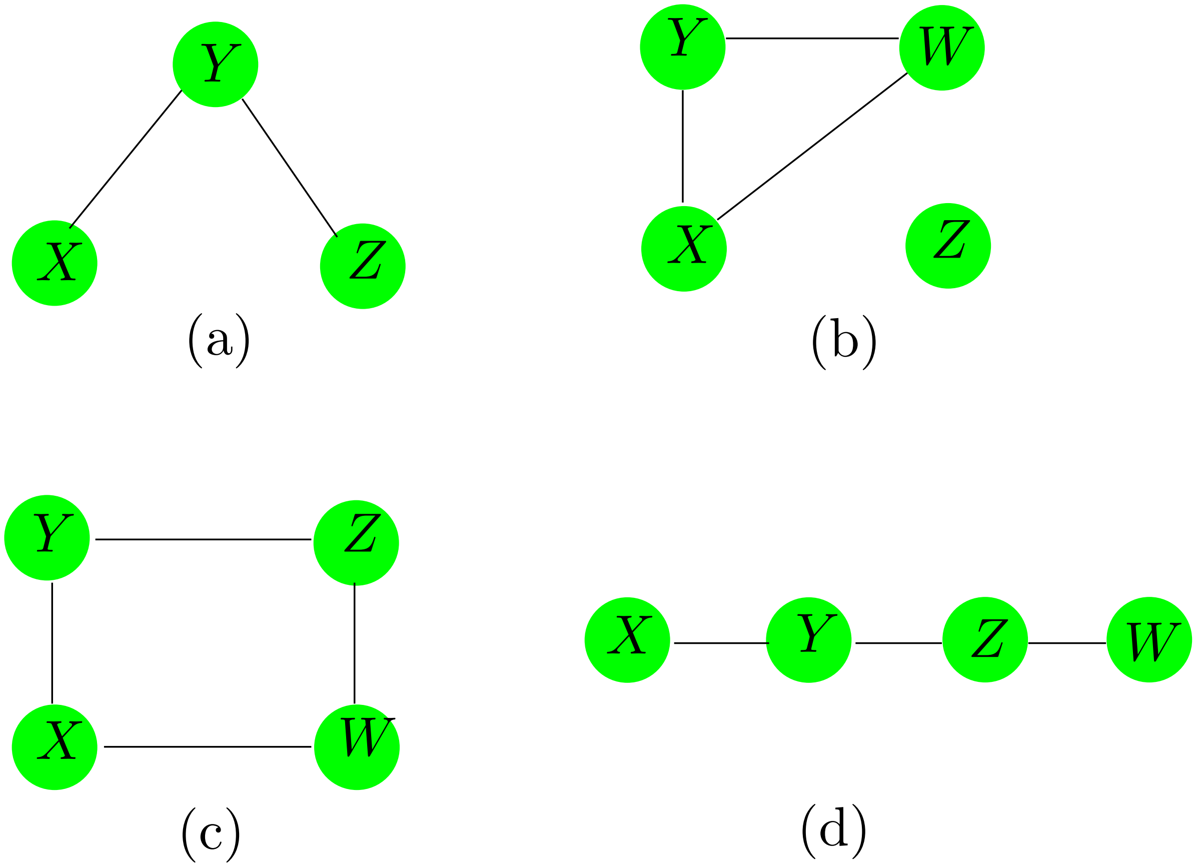
11 proteins measured on 7466 cells. Shown is an estimated *undirected* graphical model or *Markov Network*. `Raf` and `Jnk` are conditionally independent, given the rest. `PIP3` is independent of everything else, as is `Erk`. *(Sachs et al, 2003)*. The model was estimated using the *graphical lasso*.

Undirected graphical models



- Represent the joint distribution of a set of variables.
- Dependence structure is represented by the presence or absence of edges.
- Pairwise Markov graphs represent densities having no higher than second-order dependencies (e.g. Gaussian)

Conditional Independence in Undirected Graphical Models



No edge joining X and $Z \iff X \perp Z|\text{rest}$

E.g. in (a), X and Z are conditionally independent give Y.

Gaussian graphical models

Suppose all the variables $X = X_1, \ldots, X_p$ in a graph are Gaussian, with joint density $X \sim N(\mu, \mathbf{\Sigma})$

Let $X = (Z, Y)$ where $Y = X_p$ and $Z = (X_1, \ldots, X_{p-1})$. Then with

$$\mu = \begin{pmatrix} \mu_Z \\ \mu_Y \end{pmatrix}, \quad \mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{ZZ} & \sigma_{ZY} \\ \sigma_{ZY}^T & \sigma_{YY} \end{pmatrix},$$

we can write the conditional distribution of Y given Z (the rest) as

$$Y|Z = z \sim N\left(\mu_Y + (z - \mu_Z)^T \mathbf{\Sigma}_{ZZ}^{-1} \sigma_{ZY},\ \sigma_{YY} - \sigma_{ZY}^T \mathbf{\Sigma}_{ZZ}^{-1} \sigma_{ZY}\right)$$

- The regression coefficients $\beta = \mathbf{\Sigma}_{ZZ}^{-1} \sigma_{ZY}$ determine the conditional (in)dependence structure.
- In particular, if $\beta_j = 0$, then Y and Z_j are conditionally independent, given the rest.

Inference through regression

- Fit regressions of each variable X_j on the rest.
- Do variable selection to decide which coefficients should be zero.
- Meinshausen and Bühlmann (2006) use *lasso* regressions to achieve this (more later).

Problem:

- in Gaussian model, if X_j is conditionally independent of X_i, given the rest, then $\beta_{ji} = 0$.
- But then X_i is conditionally independent of X_j, given the rest, and $\beta_{ij} = 0$ as well.
- Regression methods don't honor this symmetry.

$\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$ and conditional dependence structure

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{ZZ} & \sigma_{ZY} \\ \sigma_{ZY}^T & \sigma_{YY} \end{pmatrix} \qquad \mathbf{\Theta} = \begin{pmatrix} \mathbf{\Theta}_{ZZ} & \theta_{ZY} \\ \theta_{ZY}^T & \theta_{YY} \end{pmatrix}$$

Since $\mathbf{\Sigma}\mathbf{\Theta} = \mathbf{I}$, using partitioned inverses we get

$$\begin{aligned} \theta_{ZY} &= -\theta_{YY} \cdot \mathbf{\Sigma}_{ZZ}^{-1}\sigma_{ZY} \\ &= -\theta_{YY}\beta_{Y|Z}. \end{aligned}$$

Hence $\mathbf{\Theta}$ contains all the conditional dependence information for the multivariate Gaussian model.

In particular, any $\theta_{ij} = 0$ implies conditional independence of X_i and X_j, given the rest.

Estimating Θ by Gaussian Maximum Likelihood

Given a sample $x_i, \ i = 1, \ldots, N$ we can write down the Gaussian log-likelihood of the data:

$$\ell(\mu, \mathbf{\Sigma}; \{x_i\}) = -\frac{N}{2} \log \det(\mathbf{\Sigma}) - \frac{1}{2} \sum_{i=1}^{N} (x_i - \mu)^T \mathbf{\Sigma}^{-1} (x_i - \mu)$$

Partially maximizing w.r.t μ we get $\hat{\mu} = \bar{x}$. Setting $\mathbf{S} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})^T$, we get (up to constants)

$$\ell(\mathbf{\Theta}; \mathbf{S}) = \log \det \mathbf{\Theta} - \mathrm{trace}(\mathbf{S}\mathbf{\Theta})$$

and (by some miracle)

$$\frac{d\ell(\mathbf{\Theta}; \mathbf{S})}{d\mathbf{\Theta}} = \mathbf{\Theta}^{-1} - \mathbf{S}.$$

Hence $\hat{\mathbf{\Theta}} = \mathbf{S}^{-1}$ and of course $\hat{\mathbf{\Sigma}} = \mathbf{S}$.

Solving for $\hat{\Theta}$ through regression

We can solve for $\hat{\mathbf{\Theta}} = \mathbf{S}^{-1}$ one column at a time in the score equations

$$\mathbf{\Theta}^{-1} - \mathbf{S} = \mathbf{0}.$$

Let $\mathbf{W} = \hat{\mathbf{\Theta}}^{-1}$. Suppose we solve for the last column of $\mathbf{\Theta}$. Using the partitioning as before, we can write

$$\begin{aligned} w_{12} &= -\mathbf{W}_{11}\theta_{12}/\theta_{22} \\ &= \mathbf{W}_{11}\beta, \end{aligned}$$

with $\beta = -\theta_{12}/\theta_{22}$ ($p-1$ vector). Hence the score equation says

$$\mathbf{W}_{11}\beta - s_{12} = 0$$

This looks like an OLS estimating equation $\mathbf{Z}^T\mathbf{Z}\beta = \mathbf{Z}^T\mathbf{y}$.

- Since $\mathbf{W} = \hat{\mathbf{\Theta}}^{-1} = \mathbf{S}$, then $\mathbf{W}_{11} = \mathbf{S}_{11}$ and $\hat{\beta} = \mathbf{S}_{11}^{-1} s_{12}$, the OLS regression coefficient of X_p on the rest.
- Again through partitioned inverses, we have that

$$\hat{\theta}_{22} = 1/\left(s_{22} - w_{12}^T \hat{\beta}\right),$$

  (the inverse MSR). Hence we get $\hat{\theta}_{12}$ from $\hat{\beta}$.
- So with p regressions we construct $\hat{\mathbf{\Theta}}$.
  This does not seem like such a big deal, because each of the regressions requires inverting a $(p-1) \times (p-1)$ matrix. The payoff comes when we restrict the regressions (next).

Solving for $\boldsymbol{\Theta}$ when zero structure is known

We add Lagrange terms to the log-likelihood corresponding to the missing edges

$$\max_{\Theta}\left[\log\det\boldsymbol{\Theta} - \mathrm{trace}(\mathbf{S}\boldsymbol{\Theta}) - \sum_{(j,k)\notin E}\gamma_{jk}\theta_{jk}\right]$$

Score equations: $\boldsymbol{\Theta}^{-1} - \mathbf{S} - \boldsymbol{\Gamma} = \mathbf{0}$

$\boldsymbol{\Gamma}$ is a matrix of Lagrange parameters with nonzero values for all pairs with edges absent.

Can solve by regression as before, except now iteration is needed.

Graphical Regression Algorithm

With partitioning as before, given $\mathbf{W}_{11}$ we need to solve

$$w_{12} - s_{12} - \gamma_{12} = 0.$$

With $w_{12} = \mathbf{W}_{11}\beta$, this is

$$\mathbf{W}_{11}\beta - s_{12} - \gamma_{12} = 0$$

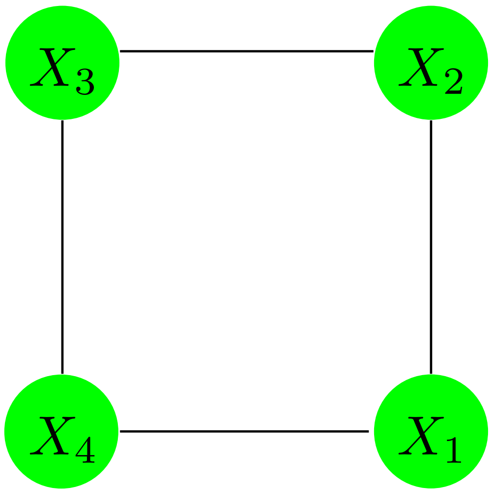
These are the score equations for a constrained regression where

1. We use the current estimate of $\mathbf{W}_{11}$ rather than $\mathbf{S}_{12}$ for the predictor covariance matrix
2. We confine ourselves to the sub-system obtained by omitting the variables constrained to be zero:

$$\mathbf{W}_{11}^{*}\beta^{*} - s_{12}^{*} = 0$$

- We then fill in $\hat{\beta}$ with $\hat{\beta}^*$ (and zeros), replace $w_{12} \leftarrow \mathbf{W}_{11}\hat{\beta}$, and proceed to the next column.
- As we cycle around the columns, the $\mathbf{W}$ matrix changes, as do the regressions, until the system converges.
- We retain all the $\hat{\beta}$s for each column in a matrix $\mathbf{B}$.
- Only at convergence do we need to estimate the $\hat{\theta}_{22} = 1/\left(s_{22} - w_{12}^T\hat{\beta}\right)$ for each column, to recover the entire matrix $\hat{\mathbf{\Theta}}$.

Simple four-variable example with known structure



$$\mathbf{S} = \begin{pmatrix} 10 & 1 & 5 & 4 \\ 1 & 10 & 2 & 6 \\ 5 & 2 & 10 & 3 \\ 4 & 6 & 3 & 10 \end{pmatrix}$$

$$\hat{\mathbf{\Sigma}} = \begin{pmatrix} 10.00 & 1.00 & 1.31 & 4.00 \\ 1.00 & 10.00 & 2.00 & 0.87 \\ 1.31 & 2.00 & 10.00 & 3.00 \\ 4.00 & 0.87 & 3.00 & 10.00 \end{pmatrix}, \quad \hat{\mathbf{\Theta}} = \begin{pmatrix} 0.12 & -0.01 & 0.00 & -0.05 \\ -0.01 & 0.11 & -0.02 & 0.00 \\ 0.00 & -0.02 & 0.11 & -0.03 \\ -0.05 & 0.00 & -0.03 & 0.13 \end{pmatrix}$$

Estimating the graph structure using the lasso penalty

Use *lasso* regularized log-likelihood

$$\max_{\mathbf{\Theta}} \left[\log \det \mathbf{\Theta} - \text{trace}(\mathbf{S}\mathbf{\Theta}) - \lambda \cdot \|\mathbf{\Theta}\|_1\right]$$

with score equations $\mathbf{\Theta}^{-1} - \mathbf{S} - \lambda \cdot \text{Sign}(\mathbf{\Theta}) = \mathbf{0}$.

Solving column-wise leads as before to

$$\mathbf{W}_{11}\beta - s_{12} + \lambda \cdot \text{Sign}(\beta) = 0$$

Compare with solution to lasso problem

$$\min_{\beta} \tfrac{1}{2}\|\mathbf{y} - \mathbf{Z}\beta)\|_2^2 + \lambda \cdot \|\beta\|_1$$

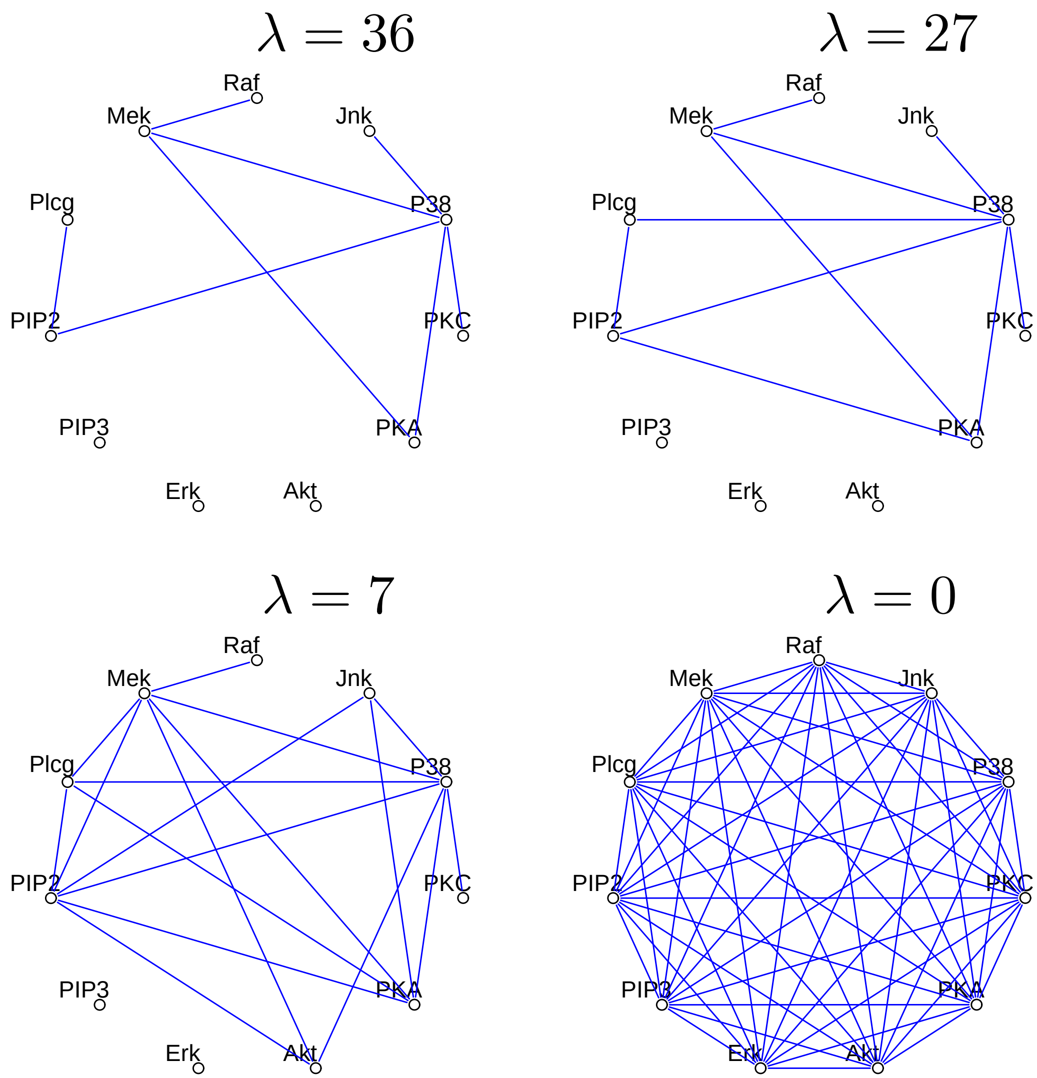
with solution

$$\mathbf{Z}^T\mathbf{Z}\beta - \mathbf{Z}^T\mathbf{y} + \lambda \cdot \text{Sign}(\beta) = 0$$

This leads to the *graphical lasso* algorithm.

Graphical Lasso Algorithm

1. Initialize $\mathbf{W} = \mathbf{S} + \lambda\mathbf{I}$. The diagonal of $\mathbf{W}$ remains unchanged in what follows.
2. Repeat for $j = 1, 2, \ldots p, 1, 2, \ldots p, \ldots$ until convergence:
   (a) Partition the matrix $\mathbf{W}$ into part 1: all but the jth row and column, and part 2: the jth row and column.
   (b) Solve the estimating equations $\mathbf{W}_{11}\beta - s_{12} + \lambda \cdot \text{Sign}(\beta) = 0$ using cyclical coordinate-descent.
   (c) Update $w_{12} = \mathbf{W}_{11}\hat{\beta}$
3. In the final cycle (for each j) solve for $\hat{\theta}_{12} = -\hat{\beta} \cdot \hat{\theta}_{22}$, with $1/\hat{\theta}_{22} = w_{22} - w_{12}^T\hat{\beta}$.



Fit using the `glasso` package in R (on CRAN).

Cyclical coordinate descent

This is Gauss-Seidel algorithm for solving system

$$\mathbf{W}\beta - s + \lambda \cdot \text{Sign}(\beta) = 0$$

where $\text{Sign}(\beta) = \pm 1$ if $\beta \neq 0$, else $\in [0, 1]$ if $\beta = 0$.

ith row: $w_{ii}\beta_i - (s_i - \sum_{j \neq i} w_{ij}\beta_j) + \lambda \cdot \text{Sign}(\beta_i) = 0$

$$\Rightarrow \beta_i \leftarrow \text{Soft}(s_i - \sum_{j \neq i} w_{ij}\beta_j,\ \lambda)/w_{ii}$$

and $\text{Soft}(z, \lambda) = \text{sign}(z) \cdot (|z| - \lambda)_+$.

Other approaches

The graphical lasso scales as $O(p^2k)$, where k is the number of non-zero values of $\mathbf{\Theta}$; can thus be $O(p^4)$ for dense problems.

- Meinshausen and Bühlmann (2006) run lasso regression of each X_j on all the rest. Have different strategies for removing an edge (j,k). For example, if both $\hat\beta_{jk}$ and $\hat\beta_{kj}$ are zero, remove the edge. Useful for $p \gg N$ problems, since $O(p^2N)$.

- Can do as above, except constrain β_{jk} and β_{kj} jointly, using a *group lasso* penalty on the pairs:

$$\min_{\beta_1,\ldots,\beta_p}\left[\sum_{j=1}^{p}\sum_{i=1}^{N}(x_{ij}-\sum_{k\neq j}x_{ik}\beta_{jk})^2+\lambda\cdot\sum_{k<j}\sqrt{\beta_{jk}^2+\beta_{kj}^2}\right]$$

- Same as above, except respect the symmetry between $\beta_{jk} = -\theta_{jk}/\theta_{jj}$ and $\beta_{kj} = -\theta_{kj}/\theta_{kk}$ with $\theta_{jk} = \theta_{kj}$:

$$\min_{\boldsymbol{\Theta}} \frac{1}{2} \sum_{j=1}^{p} \left[N \log \delta_j - \frac{1}{\delta_j} \sum_{i=1}^{N} (x_{ij} - \sum_{k \neq j} x_{ik} \beta_{jk})_2^2 \right] + \lambda \sum_{k<j} |\theta_{kj}|$$

with $\beta_{jk} = -\theta_{jk}/\theta_{jj}$, $\beta_{kj} = -\theta_{kj}/\theta_{kk}$, $\theta_{jk} = \theta_{kj}$, and $\delta_j = 1/\theta_{jj}$.

Small simulation: $p = 500$, $N = 500$, λ chosen so 25% of θ_{ij} nonzero. Timing in seconds.

Graphical Lasso	184
Meinshausen-Bühlmann	12
Grouped paired lasso	3
Symmetric paired lasso	10

Qualitative Variables

- With binary variables, second order *Ising* model. Correspond to first-order interaction models in log-linear models.
- Conditional distributions are logistic regressions.
- Exact maximum-likelihood inference difficult for large p; computations grow exponentially due to computation of partition function.
- Approximations based on lasso-penalized logistic regression (Wainwright et al 2007). Symmetric version in Hoeffling and Tibshirani (2008), using *pseudo-likelihood*

Mixed variables

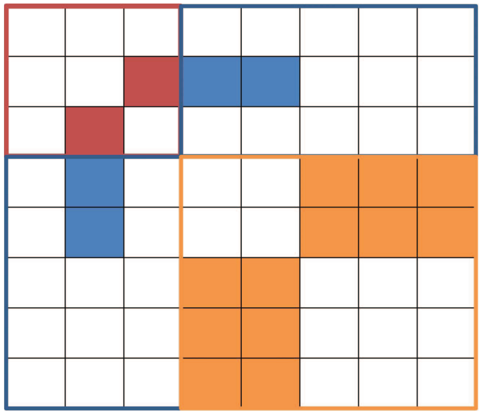
General Markov random field representation, with edge and node potentials. Work with PhD student Jason Lee.

$$p(x, y; \Theta) \propto \exp\left(\sum_{s=1}^{p}\sum_{t=1}^{p} -\frac{1}{2}\beta_{st}x_s x_t + \sum_{s=1}^{p}\alpha_s x_s + \sum_{s=1}^{p}\sum_{j=1}^{q}\rho_{sj}(y_j)x_s + \sum_{j=1}^{q}\sum_{r=1}^{q}\phi_{rj}(y_r, y_j)\right)$$

- Pseudo likelihood allows simple inference with mixed variables. Conditionals for continuous are Gaussian linear regression models, for categorical are binomial or multinomial logistic regressions.
- Parameters come in symmetric blocks, and the inference should respect this symmetry (next slide)

Group-lasso penalties

Parameters in blocks. Here we have an interaction between a pair of quantitative variables (red), a 2-level qualitative with a quantitative (blue), and an interaction between the 2 level and a 3 level qualitative.



Minimize a pseudo-likelihood with lasso and group-lasso penalties on parameter blocks.

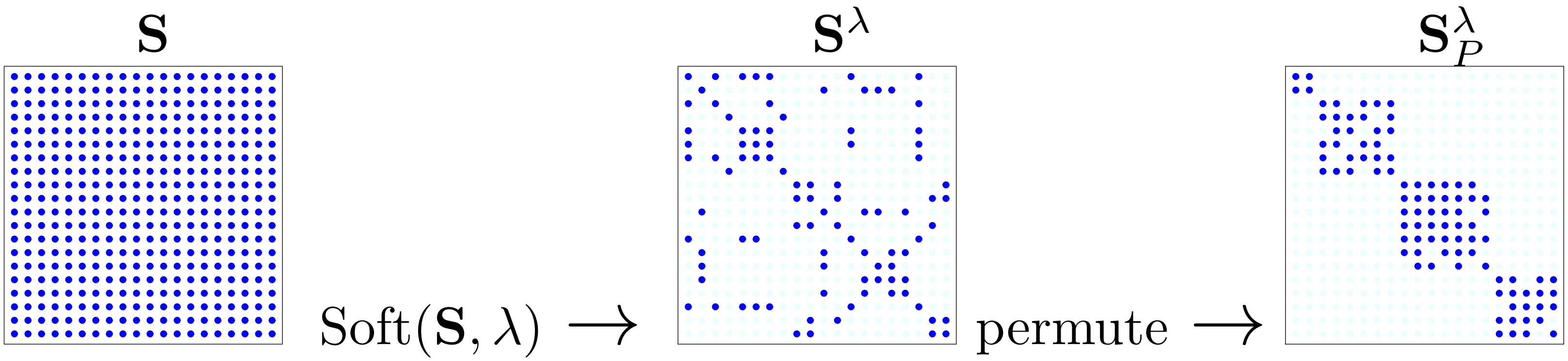
$$\min_{\Theta} \ell(\Theta) + \lambda \left(\sum_{s=1}^{p} \sum_{t=1}^{s-1} |\beta_{st}| + \sum_{s=1}^{p} \sum_{j=1}^{q} \|\rho_{sj}\|_2 + \sum_{j=1}^{q} \sum_{r=1}^{j-1} \|\phi_{rj}\|_F \right)$$

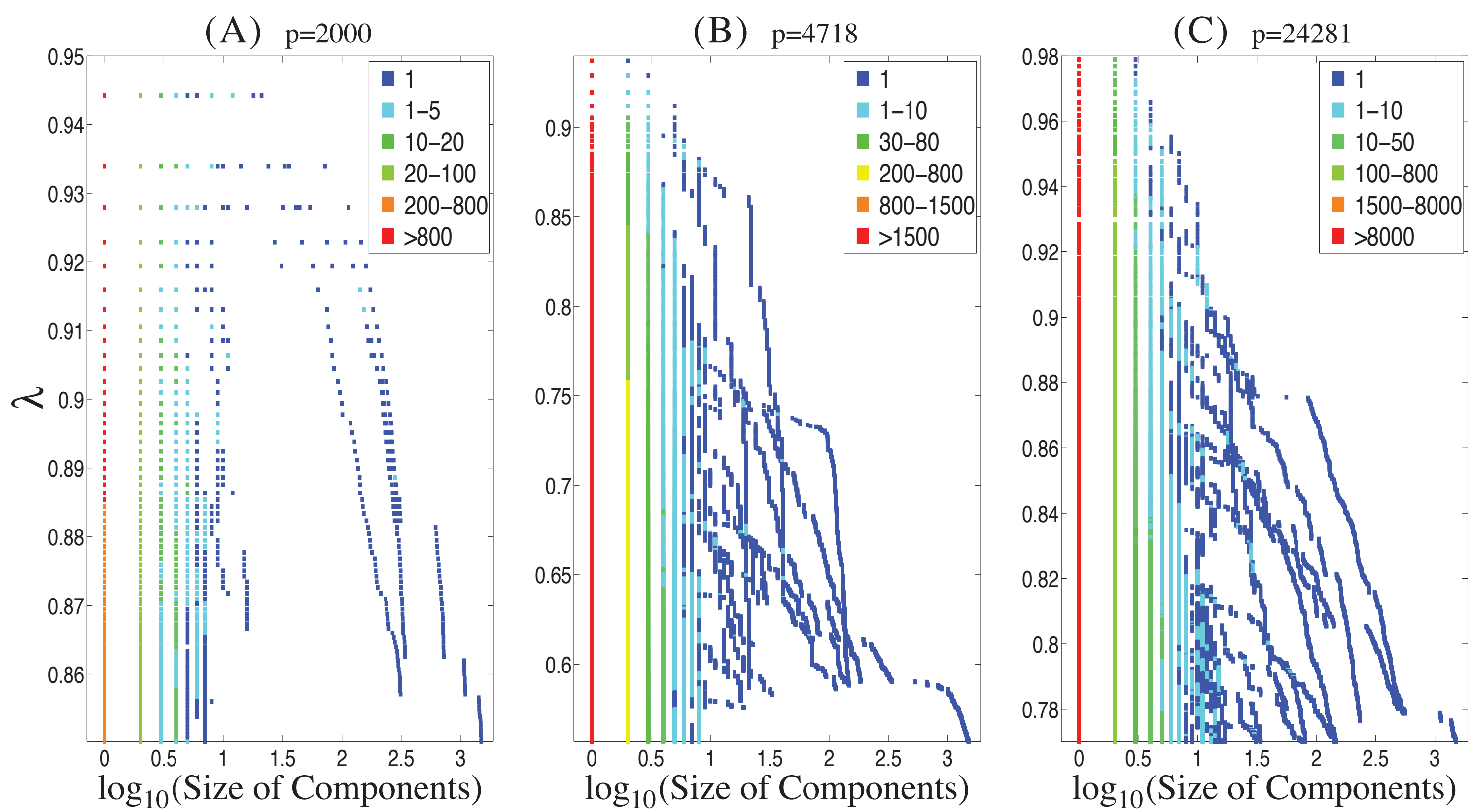
Solved using proximal Newton algorithm for a decreasing sequence of values for λ [Lee and Hastie, 2013].

Large scale graphical lasso

- The cost of `glasso` is $O(np^2 + p^{3+\Delta})$ where $\Delta \in [0, 1]$; prohibitive for genomic scale p.
- For many of these problems, $n \ll p$, so we can only fit very sparse solutions anyway.
- Simple idea [Mazumder and Hastie, 2011]:
  - Compute **S** and soft-threshold: $\mathrm{S}^{\lambda}_{ij} = \mathrm{sign}(\mathrm{S}_{ij})(|\mathrm{S}_{ij}| - \lambda)_+$.
  - Reorder rows and columns to achieve block-diagonal pattern [Tarjan, 1972]
  - Run `glasso` on each corresponding block of **S** with parameter λ, and then reconstruct.
  - Solution solves original glasso problem!

similar result found independently by Witten et al, 2011


S
S^λ
S_P^λ
Soft(S, λ) →
permute →


(A) p=2000
(B) p=4718
(C) p=24281
λ
log10(Size of Components)
1
1–5
10–20
20–100
200–800
>800
1
1–10
30–80
200–800
800–1500
>1500
1
1–10
10–50
100–800
1500–8000
>8000