Case Study 3: fMRI Prediction

Graphical LASSO

Machine Learning/Statistics for Big Data
CSE599C1/STAT592, University of Washington
Emily Fox
February 26th, 2013

Multivariate Normal Models

- So far, we looked at the univariate multiple regression
- If one has a multivariate response $y^i \in \mathbb{R}^d$
  - Assuming independence between dimensions

Multivariate Normal Models

- If one has a multivariate response $y^i \in \mathbb{R}^d$
  - Assuming correlation between the output dimensions
- Assume linear (or other mean regression) is removed and focus on the correlation structure
- Matrix valued parameter!

High-Dimensional Covariance

- What if *d* is large?
- A few common approaches:
  - Low-rank approximations
  - Sparsity assumptions

Low-Rank Approximations

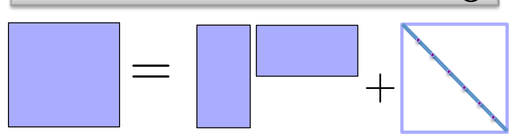
- In general, assume some matrix parameter
- Here, Σ must be a symmetric, positive definite matrix

Low-Rank Approximations

- In pictures…

$$\Sigma_0 = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$$

$$\Sigma = \Lambda\Lambda' + \Sigma_0$$



- Number of parameters:

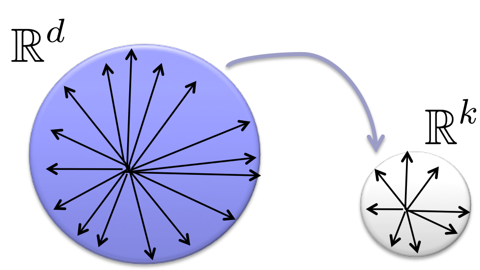
Latent Factor Models

- Low-rank approximation arises from a latent factor model

- Proof:

Lower-dim Embeddings

Sharing information in
low-dim subspace



Sparsity Assumptions

- What if we assume Σ is sparse?

- More often, we can reasonably make statements about *conditional independence*

Information Form

- Motivations for considering "information form" of multivariate normal
  - Easier to read off conditional densities
  - Has log-linear form in terms of "information parameters"

Conditional Densities

- Assume a model with

  and divide the dimensions into two sets

- Then,

Conditional Densities

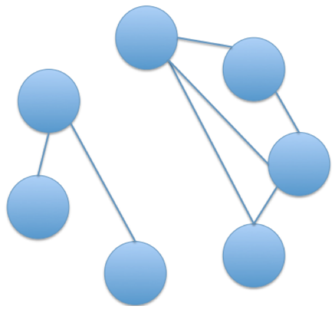
- Let $A = \{s, t\}$

$$p(y_A \mid y_{\bar{A}}) = \mathcal{N}^{-1}(\eta_A - \Omega_{A\bar{A}} y_{\bar{A}}, \Omega_{AA})$$

- Therefore,

Connection with Graphical Models

- Undirected graphical model or Markov random field (MRF)



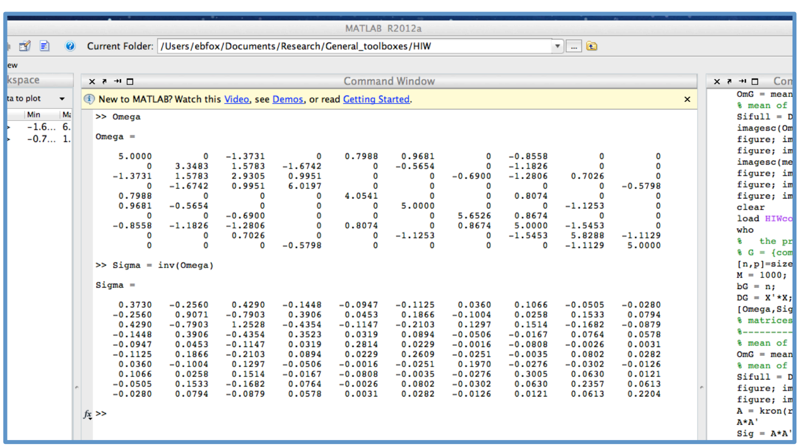
$$p(y \mid \eta, \Omega) \propto \prod_t \psi_t(y_t) \prod_{(s,t)\in E} \psi_{st}(y_s, y_t)$$

$$\psi_t(y_t) \propto e^{\eta_t y_t}$$

$$\psi_{st}(y_s, y_t) \propto e^{-\frac{1}{2} y_s \Omega_{st} y_t}$$

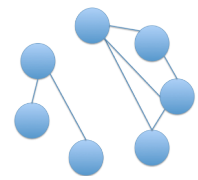
Sparse Precision vs. Covariance

- For a sparse precision matrix, the covariance need not be



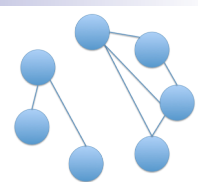
ML Estimation for Given Graph

- Assume a known graph *G = {V,E}*
- Rewrite log likelihood:



ML Estimation for Given Graph

$$L(\Omega) = \log|\Omega| - \mathrm{tr}(S\Omega)$$



- Take gradient:

- Many approaches to solving:
  - Barrier method – add penalty if Ω leaves the positive definite cone (Dahl et al. 2008)
  - Coordinate descent method (cf., Hastie et al. 2009)
  - …

ML Estimation for Given Graph

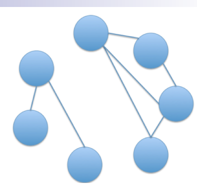
- Can show that the optimal solution satisfies

- Example:

$$G = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix} \qquad S = \begin{pmatrix} 10 & 1 & 5 & 4 \\ 1 & 10 & 2 & 6 \\ 5 & 2 & 10 & 3 \\ 4 & 6 & 3 & 10 \end{pmatrix}$$

$$\Omega = \begin{pmatrix} & & 0 & \\ & & & 0 \\ 0 & & & \\ & 0 & & \end{pmatrix} \qquad \Sigma = \begin{pmatrix} 10 & 1 & & 4 \\ 1 & 10 & 2 & \\ & 2 & 10 & 3 \\ 4 & & 3 & 10 \end{pmatrix}$$

Estimating Graph Structure



- To learn the structure of the Gaussian graphical model, we want to trade off fit and sparsity
  - Measure of fit:
  - Encouraging sparsity:
- Overall objective = “graphical LASSO” or “Glasso”

Solving the Graphical LASSO

- Objective is convex, but non-smooth as in LASSO
- Also, positive definite constraint!

- There are many approaches to optimizing the objective
  - Most common = coordinate descent akin to shooting algorithm (Friedman et al. 2008)

- Some issues…
  - Ballpark: several minutes for a 1000-variable problem
  - Algorithms scale as *O(d^3)*

Faster Computations

From Daniela Witten's talk at JSM 2012:

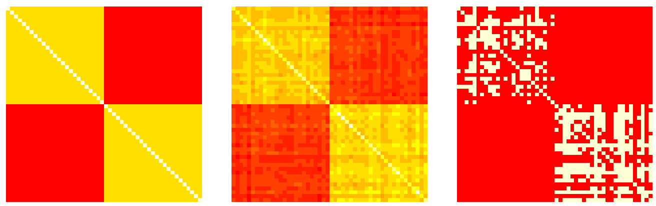
1. The jth variable is unconnected from all others in the graphical lasso solution if and only if $|S_{ij}| \leq \lambda$ for all $i = 1, \ldots, j-1, j+1, \ldots, p$.
2. Let $\mathbf{A}$ denote the $p \times p$ matrix whose elements take the form $A_{ii} = 1,\ A_{ij} = 1_{|S_{ij}|>\lambda}$. Then the connected components of $\mathbf{A}$ are the same as the connected components of the graphical lasso solution.

We can obtain the *exact* right answer by solving the graphical lasso on each connected component separately!

Citations: Witten et al. JCGS 2011, Mazumder and Hastie JMLR 2012

Covariance Screening for Glasso

From Daniela Witten's talk at JSM 2012:



- The solution to the graphical lasso problem with $\lambda = 0.7$ has five connected components (why 5?!)
- Perform graphical lasso on each component separately!
- Reduction in computational time: From $O(50^3)$ to $O(24^3)$.