Case Study 3: fMRI Prediction

Coping with Large Covariances: Latent Factor Models, Graphical Models, Graphical LASSO

last lecture

this lecture

Machine Learning for Big Data
CSE547/STAT548, University of Washington
Emily Fox
February 11th, 2014

$p >> n$

Multivariate Normal Models

- So far, we looked at univariate multiple regression

$$y^i = \beta_0 + \beta_1 x_1^i + \cdots + \beta_p x_p^i + \epsilon^i \qquad \epsilon^i \sim N(0,\sigma^2) \qquad y^i \in \mathbb{R}$$
$$= \beta^T x^i + \epsilon^i$$
$$\Rightarrow y^i \sim N(\beta^T x^i, \sigma^2)$$

- If one has a multivariate response $y^i \in \mathbb{R}^d$ ← # of semantic feature
  - Assuming independence between dimensions

$$y^i \sim N\left(\overbrace{\begin{bmatrix} - \beta^{(1)T} - \\ - \beta^{(2)T} - \\ \vdots \\ - \beta^{(d)T} - \end{bmatrix}}^{B^T} x^i, \begin{bmatrix} \sigma^2 & & 0 \\ & \ddots & \\ 0 & & \sigma^2 \end{bmatrix} \right)$$

boils down to d ind. reg. problems

$\beta^{(l)}$ are reg coeff. for the lth dim

Multivariate Normal Models

- If one has a multivariate response $y^i \in \mathbb{R}^d$
  - Assuming correlation between the output dimensions

"dog" and "furry"

$$y^i \sim N(B^T x^i, \Sigma)$$

recall: $\mathrm{cov}(y_s, y_t) = \Sigma_{st}$

- Assume linear (or other mean regression) is removed and focus on the correlation structure

$$y^i \sim N(0, \Sigma)$$

focus

Σ sym., pos. def.

- Matrix valued parameter!

See more of this in Case Study 4

Low-Rank Approximations

- In pictures…

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

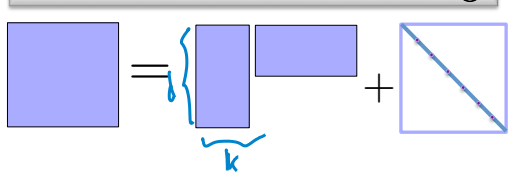
d x 1

d is large

d x d d x k

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

k << d



- Number of parameters:

$$d \cdot k + d = d(k+1) \ll \frac{d(d+1)}{2}$$

sig. reduction in param. for k<<d

Latent Factor Models

- Original multivariate regression

here: assume linear term is removed

$$\mathbf{y}^i = \cancel{B^T x^i} + \epsilon^i, \qquad \epsilon^i \sim N(0, \Sigma)$$

- Latent factor model assumption: $\Sigma = \Lambda\Lambda' + \Sigma_0$
- Low-rank approximation arises from a latent factor model

$$\boxed{\begin{array}{l} \underset{\text{obs}}{y^i} = \overset{d\times k}{\underset{\text{"factor loadings"}}{\Lambda}} \underset{\text{latent factors}}{\eta^i} + \tilde{\epsilon}^i \quad \leftarrow \text{ind. across } d \text{ dims} \qquad \eta^i \overset{iid}{\sim} N_k(0, I) \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \tilde{\epsilon}^i \overset{iid}{\sim} N_d(0, \Sigma_0) \end{array}}$$

- Proof:

$$\mathrm{cov}(y; \Lambda, \Sigma_0) = E[(y - \cancelto{0}{E[y]})(y - \cancelto{0}{E[y]})^T] = E[yy^T]$$

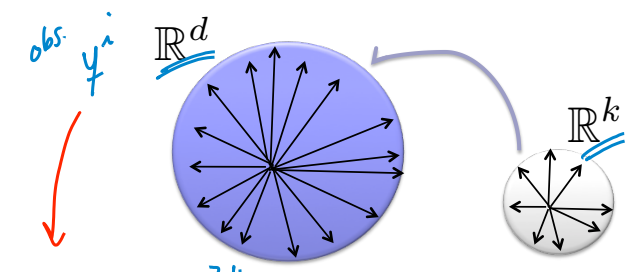
$$= E[(\Lambda\eta + \tilde{\epsilon})(\Lambda\eta + \tilde{\epsilon})^T] = \Lambda \cancelto{I}{E[\eta\eta^T]} \Lambda^T + 2\cancelto{0}{E[\eta^T]}\Lambda E[\tilde{\epsilon}] + \cancelto{\Sigma_0}{E[\tilde{\epsilon}\tilde{\epsilon}^T]}$$

$$= \Lambda\Lambda^T + \Sigma_0$$

Lower-dim Embeddings

Sharing information in

low-dim subspace



Sparsity Assumptions

- What if we assume Σ is sparse?

$(i \neq j)$ $\boxed{\Sigma_{ij} = 0 \Rightarrow y_i \perp\!\!\!\perp y_j}$ ← assuming $y^i \sim N$

$\mathrm{cov}(y_i, y_j) = 0$

Could assume Σ sparse to reduce # params, but each 0 encodes an indep. assumpt... often too strong

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

"cat" $\perp\!\!\!\perp$ "dog" | "animal", "furry", "pet", ...

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"

$$\frac{1}{\sqrt{2\pi|\Sigma|}} e^{-\frac{1}{2}(y-\mu)^T \Sigma^{-1}(y-\mu)}$$

$y \sim N(\mu, \Sigma)$

$y^T \Sigma^{-1} y - 2y^T \Sigma^{-1}\mu + \underbrace{\mu^T \Sigma^{-1} \mu}_{\text{const wrt } y}$

$\Omega = \Sigma^{-1}$

$\eta = \Sigma^{-1}\mu$

$$\propto e^{\eta^T y - \frac{1}{2} y^T \Omega y}$$

$y \sim N^{-1}(\eta, \Omega)$

Conditional Densities

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_d \end{bmatrix} \quad \bar{A} \quad A$$

- Assume a model with

$$y \sim N^{-1}(\eta, \Omega)$$

and divide the dimensions into two sets $A, \bar{A}$

- Then,

$$\begin{bmatrix} y_A \\ y_{\bar{A}} \end{bmatrix} \sim N^{-1}\left(\begin{bmatrix} \eta_A \\ \eta_{\bar{A}} \end{bmatrix}, \begin{bmatrix} \Omega_{AA} & \Omega_{A\bar{A}} \\ \Omega_{\bar{A}A} & \Omega_{\bar{A}\bar{A}} \end{bmatrix} \right)$$

submatrix of Ω with row indices in A and column indices in $\bar{A}$

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

cond. cov.

Conditional Densities

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_s \\ \vdots \\ y_t \\ \vdots \\ y_d \end{bmatrix} \quad A = \{s, t\}$$

- Let $A = \{s, t\}$, $\bar{A}$ = everything else

y_s, y_t $\quad$ $y_{\setminus st}$

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

$$\begin{bmatrix} \Omega_{ss} & \Omega_{st} \\ \Omega_{ts} & \Omega_{tt} \end{bmatrix}$$

what is $\Omega_{st} = 0$? $\Rightarrow \begin{bmatrix} \Omega_{ss} & 0 \\ 0 & \Omega_{tt} \end{bmatrix}$

$$\mathrm{cov}(y_s, y_t \mid y_{\setminus st}) = \Omega_{AA}^{-1} = \begin{bmatrix} \Omega_{ss}^{-1} & 0 \\ 0 & \Omega_{tt}^{-1} \end{bmatrix}$$

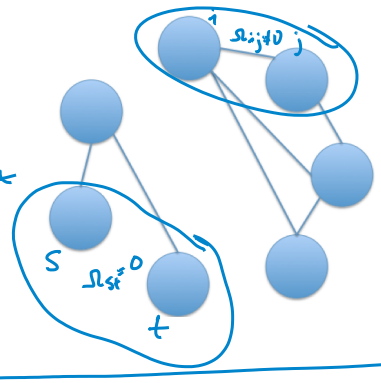
- Therefore,

$$\boxed{y_s \perp\!\!\!\perp y_t \mid y_{\setminus st} \iff \Omega_{st} = 0}$$

Connection with Graphical Models

- Undirected graphical model or Markov random field (MRF)

No edge
$\Rightarrow$
$y_s \perp\!\!\!\perp y_t \mid y_{\setminus st}$



In Gaussian graphical model case,
$\Omega_{st} = 0$ defines the edge set

In particular,
$E = \{(s,t) : \Omega_{st} \neq 0\}$

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

node potentials

edge potentials

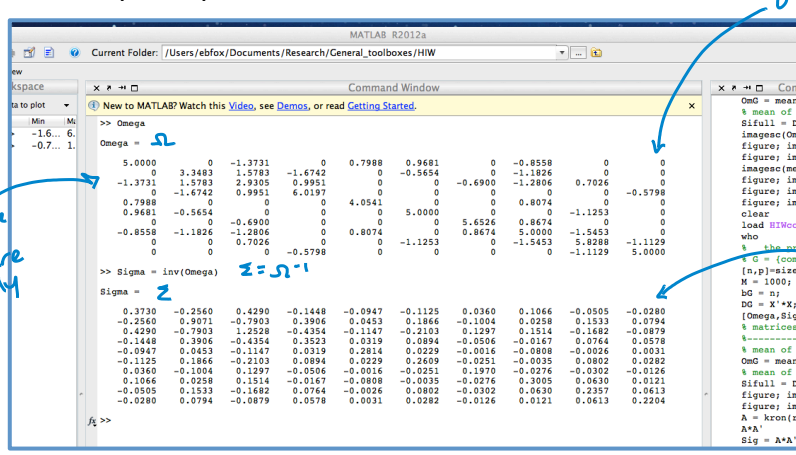
$$\psi_t(y_t) \propto e^{\eta_t y_t}$$

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

Sparse Precision vs. Covariance

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

0's encode cond. ind statements



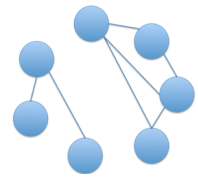
read the graph structure directly from this

does not imply sparsity of cov. (ind assumption)

$\Rightarrow$ y is still fully correlated!

ML Estimation for Given Graph

$$\frac{1}{\sqrt{2\pi|\Sigma|}} e^{-\frac{1}{2}(y-\mu)^T\Sigma^{-1}(y-\mu)}$$



- Assume a known graph $G = \{V,E\}$
- Rewrite log likelihood: $y^1, \dots, y^N$ N obs.

$$\log p(y|\theta) = \frac{N}{2}\log|\Omega| - \frac{1}{2}\sum_i \underbrace{(y^i-\mu)^T}_{x^T} \underset{A}{\Omega} \underbrace{(y^i-\mu)}_{x}$$

$$= \frac{N}{2}\log|\Omega| - \frac{1}{2}\sum_i tr\left[(y^i-\mu)(y^i-\mu)^T\Omega\right]$$

$$= \frac{N}{2}\log|\Omega| - \frac{1}{2}tr(S_\mu\Omega)$$

$$S_\mu = \sum_i (y^i-\mu)(y^i-\mu)^T$$

Trace Trick

$$v^TAx = tr(x^TAv) = tr(xx^TA)$$

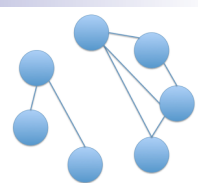
$$L(\Omega) = \log|\Omega| - tr(S\Omega)$$

$\frac{1}{N}S_\mu$ = sample cov. matrix

In our case, we're assuming $\mu = 0$

ML Estimation for Given Graph

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



- Take gradient:

$$\nabla L(\Omega) = \Omega^{-1} - S$$

s.t. $\Omega_{st} = 0$ if $(s,t) \notin E$ ← linear constraint

$\Omega \succ 0$ pos. def ← hard!!

- 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)
  - …

$\hat{\Omega}^{ML} \xrightarrow{inv.} \hat{\Sigma}^{ML}$

ML Estimation for Given Graph

- Can show that the optimal solution satisfies

$\hat{\Sigma}^{ML,G} = S_{st}$ if $(s,t) \in E$ } match to sample cov.
if $s = t$

$\Omega_{st} = 0$ if $(s,t) \notin E$

- Example:

adj. matrix
1 = edge

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

$$\Omega = \begin{pmatrix} \cdot & \cdot & 0 & \cdot \\ \cdot & \cdot & \cdot & 0 \\ 0 & \cdot & \cdot & \cdot \\ \cdot & 0 & \cdot & \cdot \end{pmatrix} \qquad \hat{\Sigma}^{ML,G} = \begin{pmatrix} 10 & 1 & \underline{1.31} & 4 \\ 1 & 10 & 2 & \underline{0.87} \\ \underline{1.31} & 2 & 10 & 3 \\ 4 & \underline{0.87} & 3 & 10 \end{pmatrix}$$

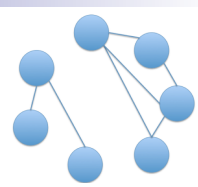
Estimating Graph Structure

- To learn the structure of the Gaussian graphical model, we want to trade off fit and sparsity
  - Measure of fit: log likelihood

$$\log|\Omega| - tr(S\Omega) + const$$

  - Encouraging sparsity: $\Omega_{st} = 0 \Rightarrow$ no edge "sparsity"

$$\|\Omega\|_1 = \sum_{s,t} |\Omega_{st}| \leftarrow \text{want to min}$$



- Overall objective = "graphical LASSO" or "Glasso"

$$F(\Omega) = -\log|\Omega| + tr(S\Omega) + \underline{\underline{\lambda}}\|\Omega\|_1$$

Just as in LASSO, but w/ a matrix parameter and s.t. $\Omega \succ 0$

Solving the Graphical LASSO

- Objective is convex, but non-smooth as in LASSO ... subgrad.
- 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)* $(y^i \in \mathbb{R}^d)$
- Other approach = ADMM

Faster Computations

From Daniela Witten's talk at JSM 2012:

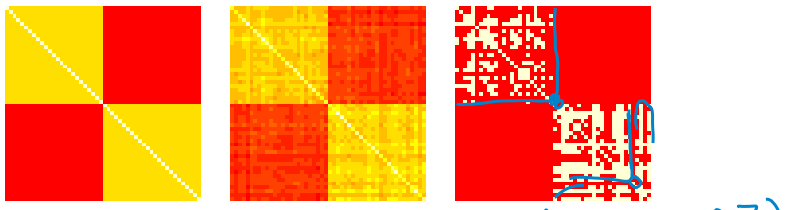
1. The jth variable is unconnected from all others in the graphical lasso solution if and only if $|S_{ij}| \leq \lambda$ for all $i = 1, \ldots, j-1, j+1, \ldots, d$. strength of our penalty
2. Let **A** denote the $d \times d$ matrix whose elements take the form $A_{ii} = 1$, $A_{ij} = 1_{|S_{ij}|>\lambda}$.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



Σ S A (for $\lambda = 0.7$)

- 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 $O(24^3)$.