**Case Study 3: fMRI Prediction**

# Coping with Large Covariances: Graphical Models, Graphical LASSO

today

Machine Learning for Big Data
CSE547/STAT548, University of Washington
Emily Fox
May 5th, 2015

# Multivariate Normal Models

- So far, we looked at univariate multiple regression

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

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

so far

$$\underbrace{y^i}_{\in \mathbb{R}^d} \sim N\left( \underbrace{\begin{bmatrix} — \beta^{(1)T} — \\ — \beta^{(2)T} — \\ \vdots \\ — \beta^{(d)T} — \end{bmatrix}}_{B^T} x^i, \begin{bmatrix} \sigma^2 & & 0 \\ & \ddots & \\ 0 & & \sigma^2 \end{bmatrix} \right)$$

leads to $d$ ind. problems

$\beta^{(l)}$ are coeff. for the $l^{th}$ semantic feature

# Multivariate Normal Models

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

"dog" and "furry"

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

non-diagonal

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

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

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

$\Sigma$ sym. pos. def.

- Matrix valued parameter!

see more on matrix valued params in Case Study 4

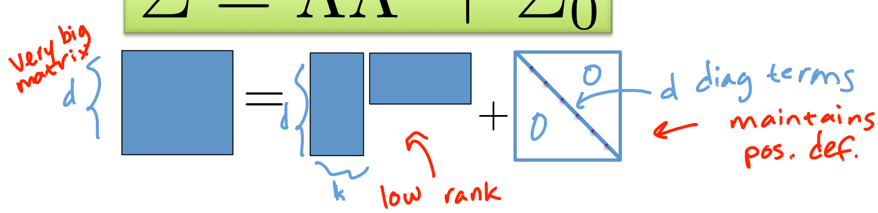
 3

# Low-Rank Approximations

- In pictures…

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

$$\underset{d \times d}{\Sigma} = \underset{d \times k}{\Lambda}\Lambda' + \Sigma_0$$




- Number of parameters:

$$\boxed{dk + d = d(k+1) << \frac{d(d+1)}{2}}$$

sig. reduction in param. for $k << d$

 4

# Latent Factor Models

here: assume linear term is removed

- Original multivariate regression

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

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

$$\boxed{\underset{\text{obs.}}{y^i} = \underset{\text{"factor loadings"}}{\underset{d\times k}{\Lambda}} \underset{\text{"latent factors"}}{\underset{k\times 1}{\eta^i}} + \tilde{\epsilon}^i \qquad \eta^i \sim N_k(0, I) \qquad \tilde{\epsilon}^i \sim N_d(0, \underset{\text{diag.}}{\Sigma_0})}$$

ind. across dims

- Proof:

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

$$= E[(\Lambda\eta + \tilde{\epsilon})(\Lambda\eta + \tilde{\epsilon})^T] = \Lambda \underbrace{E[\eta\eta^T]}_{I} \Lambda^T$$

$$+ 2E[\eta^T]\Lambda^T \cancel{E[\tilde{\epsilon}]}^{0}$$

$$+ E[\tilde{\epsilon}\tilde{\epsilon}^T] \rightarrow \Sigma_0$$

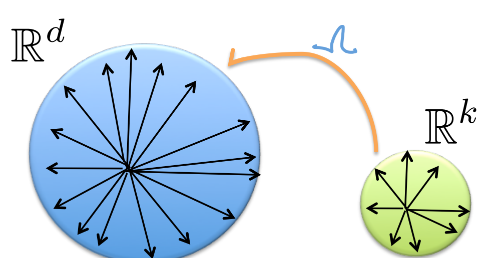
$$= \Lambda I \Lambda^T + \Sigma_0$$

# Lower-dim Embeddings

**Sharing information in**

***low-dim subspace***




obs. $y^i$

"can you hold it?" and "is it bigger than a bread box" } redundant ← latent factor "is it big?"

$\eta^i$ latent factor

# Sparsity Assumptions

- What if we assume $\Sigma$ is sparse?

$(i \neq j)\ \Sigma_{ij} = 0 \overset{\text{Gaussian}}{\Longrightarrow} y_i \perp\!\!\!\perp y_j$ ← ind.

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

Could assume $\Sigma$ sparse to reduce # params, but each 0 encodes an indep. statement

→ often too strong of an assumption

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

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

# 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"

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

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

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

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

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

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

# Conditional Densities

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

- Assume a model with

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

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

- Then,

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

Submatrix of $\Omega$ with rows indexed by indices in $A$ and columns by $\bar{A}$

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



# Conditional Densities

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

- Let $A = \{s, t\}$ "cat" "dog" $\bar{A}$ = 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 if $\Omega_{st} = 0$? $\Rightarrow \begin{bmatrix} \Omega_{ss} & 0 \\ 0 & \Omega_{tt} \end{bmatrix}$

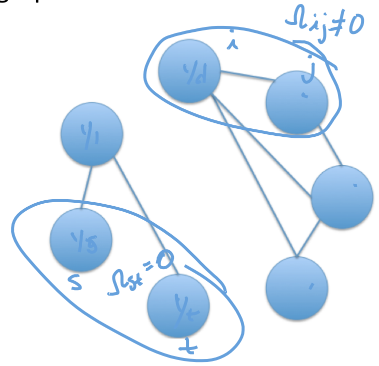
$$\mathrm{cov}(y_s, y_t \mid y_{\setminus st}) = \Omega_{AA}^{-1} = \begin{bmatrix} \Omega_{ss}^{-1} & 0 \\ 0 & \Omega_{tt}^{-1} \end{bmatrix}$$

- Therefore, $\Leftrightarrow$ $\boxed{y_s \perp\!\!\!\perp y_t \mid y_{\setminus st} \Leftrightarrow \Omega_{st} = 0}$

# Connection with Graphical Models

- Undirected graphical model or Markov random field (MRF)

No edge $(s,t)$
$\Rightarrow$
$y_s \perp y_t \mid y_{\setminus st}$




In Gaussian graphical model case

$\Omega_{st}=0$ define 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

$$\left\{ \begin{aligned} \psi_t(y_t) &\propto e^{\eta_t y_t} \\ \psi_{st}(y_s, y_t) &\propto e^{-\frac{1}{2} y_s \Omega_{st} y_t} \end{aligned} \right.$$



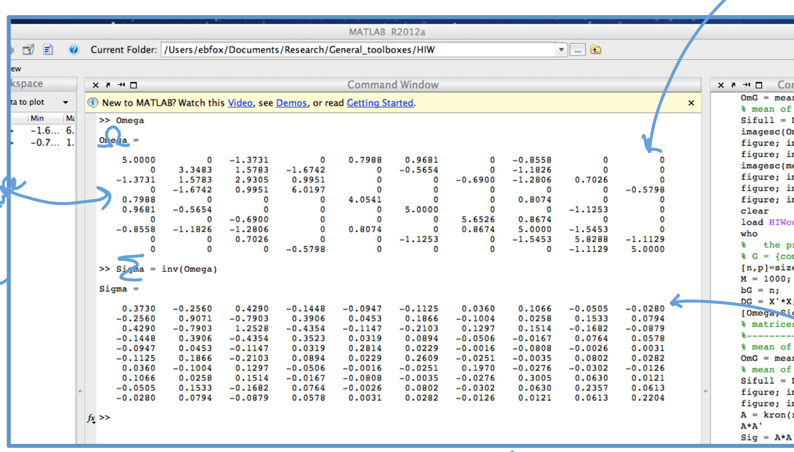


# Sparse Precision vs. Covariance

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

zeros encode statements of cond. ind.

read graph structure from here




does not imply sparsity in cov. (ind. statements)

$\Rightarrow y$ is still fully correlated

# ML Estimation for Given Graph

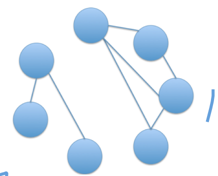
- Assume a known graph $G = \{V, E\}$

nodes

edges

- Rewrite log likelihood:

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



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

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

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

$$S_\mu = \sum_i (y_i-\mu)(y_i-\mu)^T$$

trace trick:

$$x^T A x = tr(x^T A x) = tr(x x^T A)$$

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

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

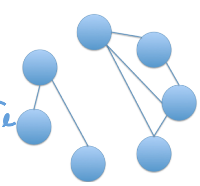
In our case, $\mu = 0$





# ML Estimation for Given Graph

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



- Take gradient:

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

constraint on our ML est. from graph structure

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

$\Omega$ pos. def., sym. matrix ← hard :(

- Many approaches to solving:
  - Barrier method – add penalty discouraging $\Omega$ from leaving 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

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

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

- Example:

adj. matrix
1 = edge

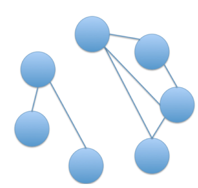
$$G = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix} \qquad S = \begin{pmatrix} 10 & 1 & 5 & 4 \\ 1 & 10 & 2 & 6 \\ 5 & 2 & 10 & 3 \\ 4 & 6 & 3 & 10 \end{pmatrix}$$

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

 15

# 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"
    $\|\Omega\|_1 = \sum_{s,t} |\Omega_{s,t}|$ ← want to min



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

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

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

 16

# 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) See HW 3
- Some issues…
  - Ballpark: several minutes for a 1000-variable problem
  - Algorithms scale as *O(d^3)*
- Other approach = ADMM also HW 3

# Faster Computations

From Daniela Witten's talk at JSM 2012:

1. The $j$th variable is unconnected from all others in the graphical lasso solution if and only if $|S_{ij}| \leq \lambda$ for all $i = 1, \ldots, j-1, j+1, \ldots, p$. sample cov is small relative to chosen penalty
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.

ind. on the thresholded values

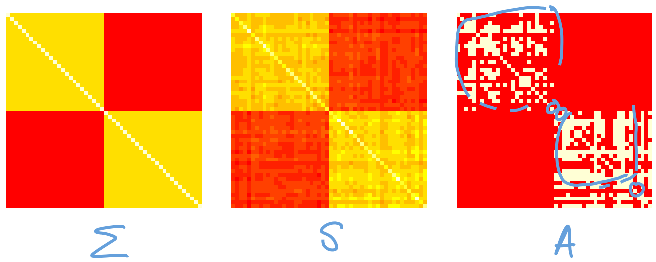
We can obtain the *exact* right answer by solving the graphical lasso on each connected component separately!

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 $O(24^3)$.