Case Study 3: fMRI Prediction

Coping with Large Covariances: Graphical Models, Graphical LASSO

Multivariate Normal Models

- So far, we looked at univariate multiple regression
  \[ y^i = \beta_0 + \beta_1 x^i + \ldots + \beta_p x^i_p + \epsilon^i \sim \mathcal{N}(0, \sigma^2) \quad y^i \in \mathbb{R}^p \]
  \[ = \beta^T x^i + \epsilon^i \]
  \[ \Rightarrow y^i \sim \mathcal{N}(\beta^T x^i, \sigma^2) \]

- If one has a multivariate response \( y^f \in \mathbb{R}^d \)
  
  So far, independence between dimensions
  \[ y^f \sim \mathcal{N}
  \begin{pmatrix}
  \beta^T \\
  \vdots \\
  \beta^{(d)T}
  \end{pmatrix} x^i, 
  \begin{pmatrix}
  \sigma^2 & 0 & \ldots & 0 \\
  0 & \sigma^2 & \ldots & 0 \\
  \vdots & \vdots & \ddots & \vdots \\
  0 & 0 & \ldots & \sigma^2
  \end{pmatrix}
  \]

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

leads to \( d \) ind. problems
Multivariate Normal Models

- If one has a multivariate response $y_i \in \mathbb{R}^d$
  - Assuming correlation between the output dimensions
    
    $y_i \sim N(B^T x_i, \Sigma)$
    
    (non-diagonal)
    
    recall: $\text{cov}(y_i, y_j) = \Sigma_{ij}$

- 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

Low-Rank Approximations

- In pictures...

  \[
  \Sigma = \Lambda \Lambda' + \Sigma_0
  \]

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

  \[
  \begin{array}{c}
  \text{dxd} \\
  \text{dcd}
  \end{array}
  \]

- Number of parameters:

  \[
  dk + d = d(k+1) < \frac{d(d+1)}{2}
  \]

  \(\text{sig. reduction in param. for } k \ll d\)
Latent Factor Models

- Original multivariate regression
  \[ y^i = B^T x^i + e^i, \quad e^i \sim N(0, \Sigma) \]
- Latent factor model assumption:
  \[ \Sigma = \Lambda \Lambda' + \Sigma_0 \]
- Low-rank approximation arises from a latent factor model

\[ y^i = \Lambda \eta^i + \varepsilon^i \]

\[ \etahat \sim N_k(0, I) \]
\[ \varepsilonhat \sim N_d(0, \Sigma_0) \]

Proof:
\[ \text{Cov}(y, \eta, \xi_0) = \mathbb{E}[(y - \mathbb{E}y)(y - \mathbb{E}y)^T] = \mathbb{E}[yy^T] \]
\[ = \mathbb{E}[(\Lambda \eta + \varepsilon)(\Lambda \eta + \varepsilon)^T] = \Lambda \mathbb{E}[\eta \eta^T] \Lambda' + 2 \mathbb{E}[\eta \varepsilon^T] + \mathbb{E}[\varepsilon \varepsilon^T] \]
\[ = \Lambda I \Lambda'^T + \Sigma_0 + \mathbb{E}[\varepsilon \varepsilon^T] \]

Lower-dim Embeddings

Sharing information in low-dim subspace

\[ \mathbb{R}^d \rightarrow \mathbb{R}^k \]

"can you hold it?"
"is it bigger than a bread box"
"is it big?"

©Emily Fox 2015
**Sparsity Assumptions**

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

\[
(i \neq j) \Sigma_{ij} = 0 \rightarrow \text{Gaussian } y_i \perp \! \! \! \! \perp y_j \\
\text{cov}(y_i, y_j) = 0
\]

Could assume $\Sigma$ sparse to reduce # params, but each 0 encodes an independence statement, often too strong of an assumption.

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

  "cat" $\perp$ "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”
Conditional Densities

- Assume a model with
  \[ y \sim N^{-1}(\eta, \Omega) \]
  and divide the dimensions into two sets
- Then,
  \[ \begin{bmatrix} y_A^1 \\ y_A^2 \end{bmatrix} \sim N \left( \begin{bmatrix} \eta_A \\ \Omega_{AA} \end{bmatrix}, \begin{bmatrix} \Omega_{AA} & \Omega_A \Omega_{\bar{A}} \\ \Omega_A & \Omega_{\bar{A}} \Omega_{\bar{A}} \end{bmatrix} \right) \]

- Let \( A = \{ s, t \} \)
  \[ p(y_A \mid y_{\bar{A}}) = N^{-1}(\eta_A - \Omega_{A\bar{A}}y_{\bar{A}}, \Omega_{AA}) \]
  what if \( \Omega_{st} = 0 \) ?
  \[ \begin{bmatrix} \Omega_{ss} & \Omega_{st} \\ \Omega_{ts} & \Omega_{tt} \end{bmatrix} \]
  \[ \text{cov}(y_s, y_t \mid y_{\bar{st}}) = \Omega_{AA}^{-1} = \begin{bmatrix} \Omega_{ss}^{-1} & 0 \\ 0 & \Omega_{tt}^{-1} \end{bmatrix} \]

- Therefore,
  \[ y_s \perp y_t \mid y_{\bar{st}} \iff \Omega_{st} = 0 \]
Connection with Graphical Models

- Undirected graphical model or Markov random field (MRF)

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

\[ \psi_t(y_t) \propto e^{\eta_t y_t} \]

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

Sparse Precision vs. Covariance

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

$\Rightarrow \text{y is still fully correlated}$

©Emily Fox 2015
ML Estimation for Given Graph

- Assume a known graph $G = (V,E)$
- Rewrite log likelihood:

$$
\log p(y|\theta) = \frac{N}{2}\log|\Omega| - \frac{1}{2} \sum_{i} (y_i - \mu)^T \Omega (y_i - \mu) 
$$

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

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

In our case, $\mu = 0$

**ML Estimation for Given Graph**

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

- Take gradient:

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

s.t. $S_{st} = 0$ if $(s,t) \notin E$  \hspace{1cm} linear constraint

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)
- ...

©Emily Fox 2015
ML Estimation for Given Graph

- Can show that the optimal solution satisfies
  \[ \hat{\Sigma}_{ML,G} = S_{\text{tr}} \quad \text{if} \quad (s,t) \in E \]
  \[ \hat{\Sigma}_{ML,G} = 0 \quad \text{if} \quad s = t \quad \text{or} \quad (s,t) \notin E \]

Example:

Example adjacency matrix:
\[ G = \begin{pmatrix}
  0 & 1 & 0 & 1 \\
  1 & 0 & 1 & 0 \\
  0 & 1 & 0 & 1 \\
  1 & 0 & 1 & 0
\end{pmatrix} \]

Example precision matrix:
\[ S = \begin{pmatrix}
  10 & 1 & 5 & 4 \\
  1 & 10 & 2 & 6 \\
  5 & 2 & 10 & 3 \\
  4 & 6 & 3 & 10
\end{pmatrix} \]

Example precision matrix:
\[ S = \begin{pmatrix}
  10 & 1 & 13 & 4 \\
  1 & 10 & 2 & 0.87 \\
  13 & 2 & 10 & 3 \\
  4 & 0.87 & 3 & 10
\end{pmatrix} \]

Estimating Graph Structure

- To learn the structure of the Gaussian graphical model, we want to trade off fit and sparsity
  - Measure of fit: \( \log-likelihood \quad \log |S\|_1 - \text{tr} (S \hat{\Sigma}) + \text{const.} \)
  - Encouraging sparsity: \( \|\hat{\Sigma}\|_1 = \sum_{s,t} |\hat{\Sigma}_{s,t}| \quad \text{want to min} \)
  - Overall objective = “graphical LASSO” or “Glasso”

\[ F(S) = -\log |S\|_1 + \text{tr} (S \hat{\Sigma}) + \lambda \|\hat{\Sigma}\|_1 \]

Just as in LASSO, but with a matrix parameter and \( S \geq 0 \)
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)
    - 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 $i = 1, \ldots, j - 1, j + 1, \ldots, p$.

2. Let $A$ denote the $p \times p$ matrix whose elements take the form $A_{ij} = 1, A_{ij} = 1_{|S_{ij}| > \lambda}$. 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:

- 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 $O(24^3)$.