Case Study 3: fMRI Prediction

Graphical LASSO

Multivariate Normal Models

- So far, we looked at the univariate multiple regression
  \[ y^i = \beta_0 + \beta_1 x_1^i + \ldots + \beta_p x_p^i + \epsilon^i \quad \epsilon^i \sim N(0, \sigma^2) \]
  \[ \Rightarrow y^i \sim N(\beta^T x^i, \sigma^2) \]

- If one has a multivariate response \( y^j \in \mathbb{R}^d \)
  \[ y^j \sim N \left( \begin{bmatrix} \beta_0^T \\ \beta_1^T \\ \vdots \\ \beta_p^T \\ \beta_1^T \end{bmatrix} x^j , \begin{bmatrix} \sigma^2 \\ \sigma^2 & 0 \\ \sigma^2 & \sigma^2 \\ \sigma^2 & \sigma^2 \\ \sigma^2 & \sigma^2 \end{bmatrix} \right) \]

\( a^{(l)} \) are reg. coeff. for the \( l^{th} \) dim
Multivariate Normal Models

- If one has a multivariate response $y^i \in \mathbb{R}^d$
  - Assuming correlation between the output dimensions
    $$y^i \sim N(B^i x^i, \Sigma)$$
    recall: $\text{cov}(y_5, y_7) = \Sigma_{57}$
- Assume linear (or other mean regression) is removed and focus on the correlation structure
  $$y^i \sim N(0, \Sigma)$$ sym., pos. def.

- Matrix valued parameter!
  See more of this in Case Study 4

High-Dimensional Covariance

- What if $d$ is large?
  - many semantic features
    $$\# \text{ params } (\Sigma) = \frac{d(d+1)}{2}$$ sym.
    Again, consider $d \gg N$ but $O(d^2)$ params to est.

- A few common approaches:
  - Low-rank approximations
  - Sparsity assumptions
Low-Rank Approximations

- In general, assume some matrix parameter
  \[ \Theta = AB, \quad k \ll d, m \]
  will see this in case Study 4
- Here, \( \Sigma \) must be a symmetric, positive definite matrix
  \[ \Sigma = \sum_{i=1}^{d} \sigma_i^2 + \sum_0 \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_d^2 \end{bmatrix} \]

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

\[ d < k \ll d \]

Number of parameters:
\[ d \cdot k \cdot d = d(k+1) \]

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

- Low-rank approximation arises from a latent factor model
  \[ y = \Lambda \eta + \varepsilon \]
  \[ \eta \sim N_{k}(0, I) \]
  \[ \varepsilon \sim N_{d}(0, \Sigma_\varepsilon) \]

- Proof:
  \[ \text{Cov}(y^T, \Lambda, \Sigma_\varepsilon) = E((y^T - E[y^T])(y^T - E[y^T])^T) = E[yy^T] - \Lambda E[\eta^T\eta] \Lambda^T + 2E[\eta^T\varepsilon\varepsilon^T] + E[\varepsilon\varepsilon^T] \]
  \[ = \Lambda \Sigma_\varepsilon \Lambda^T + \Sigma_\varepsilon \]

Lower-dim Embeddings

Sharing information in low-dim subspace
Sparsity Assumptions

- What if we assume $\Sigma$ is sparse?
  
  $\forall i \neq j \quad \Sigma_{ij} = 0 \quad \Rightarrow \quad y_i \perp \!\!\!\!\perp y_j$

  $\text{cor}(y_i, y_j) = 0$
  
  Could assume $\Sigma$ sparse to reduce params, but each 0 encodes an indep. assumption ... often too strong

- More often, we can reasonably make statements about conditional independence
  
  "cat" $\perp \!\!\!\!\perp$ "dog" $\perp \!\!\!\!\perp$ "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
  \[ y \sim N^{-1}(\eta, \Sigma) \]
  and divide the dimensions into two sets

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

  \[
  p(y_A | y_{\bar{A}}) = N^{-1}(\eta_A - \Sigma_{A\bar{A}}y_{\bar{A}}, \Sigma_{A\bar{A}})
  \]

- Let \( A = \{s, t\} \)

- Therefore,

Conditional Densities

- Let \( A = \{s, t\} \)

- \( \bar{A} \) = everything \( \notin \{s, t\} \)

- \[
  p(y_A | y_{\bar{A}}) = N^{-1}(\eta_A - \Sigma_{A\bar{A}}y_{\bar{A}}, \Sigma_{A\bar{A}})
  \]

- what if \( \Sigma_{st} = 0 \) ?

- \[
  \text{cov}(y_s, y_t | y_{\bar{A}}) = \Sigma_{AA}^{-1} = \begin{bmatrix}
  \Sigma_{ss} & 0 \\
  0 & \Sigma_{tt}
  \end{bmatrix}
  \]

- \( y_s \perp y_t | y_{\bar{A}} \iff \Sigma_{st} = 0 \)

- Therefore,
Connection with Graphical Models

- Undirected graphical model or Markov random field (MRF)

In Gaussian graphical model case, $\mathcal{N}(0,\Sigma)$ defines the edge set

In particular

$E = \{ (s,t) : \mathcal{N}(s,t) \neq 0 \}$

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

$\psi_t(y_t) \propto e^{\eta^t y_t}$

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

Sparse Precision vs. Covariance

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

$Z = \Sigma^{-1}$

$Y$ is still fully correlated!
ML Estimation for Given Graph

- Assume a known graph $G = (V,E)$

Rewrite log likelihood:

$$
\log p(y | \theta) = \frac{N}{2} \log |\Sigma| - \frac{1}{2} \sum \left( y_{i} - m_{i} \right)^{T} \Sigma^{-1} \left( y_{i} - m_{i} \right) \\
= \frac{N}{2} \log |\Sigma| - \frac{1}{2} \text{tr} \left[ \left( y_{i} - m_{i} \right) \left( y_{i} - m_{i} \right)^{T} \Sigma^{-1} \right] \\
= \frac{N}{2} \log |\Sigma| - \frac{1}{2} \text{tr} \left( S \Omega \right) \\
L(\Omega) = \log |\Sigma| - \text{tr} \left( S \Omega \right) \\
\text{In our case, } M = 0
$$

Trace trick:

$\text{Trace trick:} \quad x^{T} A x = \text{tr} \left( x^{T} A x \right) = \text{tr} \left( A x x^{T} \right)$

Matrix reference manual

Many approaches to solving:

- Barrier method – add penalty if $\Omega$ 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
  \[ \sum_{s,t} G_{st} = S_{st} \quad \text{if } (s,t) \in E \]
  \[ \Omega_{st} = 0 \quad \text{if } (s,t) \notin E \]

Example:

\[ G = \begin{pmatrix}
  0 & 1 & 0 & 1 \\
  1 & 0 & 1 & 0 \\
  0 & 1 & 0 & 1 \\
  1 & 0 & 1 & 0
\end{pmatrix} \quad S = \begin{pmatrix}
  10 & 1 & 5 & 4 \\
  1 & 10 & 2 & 6 \\
  5 & 2 & 10 & 3 \\
  4 & 6 & 3 & 10
\end{pmatrix} \]

\[ \Omega = \begin{pmatrix}
  0 & 0 & 0 & 0 \\
  0 & 0 & 0 & 0 \\
  0 & 0 & 0 & 0 \\
  0 & 0 & 0 & 0
\end{pmatrix} \quad \Sigma = \begin{pmatrix}
  10 & 1 & 5 & 4 \\
  1 & 10 & 2 & 6 \\
  5 & 2 & 10 & 3 \\
  4 & 6 & 3 & 10
\end{pmatrix} \]

Estimating Graph Structure

- To learn the structure of the Gaussian graphical model, we want to trade off fit and sparsity
  - Measure of fit: \[ \log \text{likelihood} = \log |\Sigma| - \text{tr}(SS') + \text{const.} \]
  - Encouraging sparsity: \[ \lVert S \rVert_1 = \sum_{s,t} \Omega_{st} \]
  - Overall objective = “graphical LASSO” or “Glasso”
    \[ F(S) = -\log |\Sigma| + \text{tr}(SS') + \lambda \lVert S \rVert_1 \]
    \[ \text{Just as in LASSO, but w/ a matrix parameter and s.t. } S \succeq 0 \]
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 $O(d^3)$

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 $A_{ii} = 1$, $A_{ij} = 1_{|S_{ij}| > \lambda}$. 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:

- 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 $O(24^3)$. 

©Emily Fox 2013