Case Study 3: fMRI Prediction

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

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

©Emily Fox 2014

Case Study 3: fMRI Prediction

Multivariate Normal Models

- So far, we looked at univariate multiple regression
  \[ y^i = \beta_0 + \beta_1 x_1^i + \ldots + \beta_p x_p^i + \epsilon^i = \beta^T x^i + \epsilon^i \]
  \[ \epsilon^i \sim N(0, \sigma^2) \]
  \[ y^i \sim \mathcal{N}(\beta^T x^i, \sigma^2) \]
  \[ y^i \in \mathbb{R}^d \]
- If one has a multivariate response
  \[ y^i \sim N\left( \begin{bmatrix} \beta_0^{(1)} \\ \vdots \\ \beta_0^{(d)} \end{bmatrix} x^i, \begin{bmatrix} \sigma^2 & \ldots & 0 \\ \ldots & \sigma^2 & \ldots \\ 0 & \ldots & \sigma^2 \end{bmatrix} \right) \]
- Assuming independence between dimensions
- boils down to independent regression problems

©Emily Fox 2014
Multivariate Normal Models

- If one has a multivariate response $y^i \in \mathbb{R}^d$
  - Assuming correlation between the output dimensions
    $$y^i \sim N(B\theta^i, \Sigma)$$
    recall: $\text{cov}(y_s, y_t) = \Sigma_{st}$
- 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 of this in Case Study 4*

Low-Rank Approximations

- In pictures...
  $$\Sigma = \Lambda \Lambda' + \Sigma_0$$
  $\Sigma_0 = \text{diag}(\sigma_1^2, \ldots, \sigma_d^2)$
  $\Sigma$ is large

- Number of parameters:
  $$d \cdot k + d = d(k+1) \approx \frac{d(d+1)}{2}$$
  *Significant reduction in param. for $k << d$*
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

Proof:
\[
\begin{aligned}
\mathbb{E}[(y - \mathbb{E}[y])(y - \mathbb{E}[y])^T] &= \mathbb{E}[y y^T] \\
&= \mathbb{E}[\Lambda \eta^T + \sqrt{\Sigma_0}] \Lambda^T + 2 \mathbb{E}[\Lambda \eta^T \sqrt{\Sigma_0}] + \mathbb{E}[\sqrt{\Sigma_0} \sqrt{\Sigma_0}]
\end{aligned}
\]

Lower-dim Embeddings

Sharing information in low-dim subspace
Sparsity Assumptions

- What if we assume $\Sigma$ is sparse?
  - $\Sigma_{ij} = 0 \Rightarrow y_i \perp y_j$
  - $\text{cov}(y_i, y_j) = 0$
  - Could assume $\Sigma$ 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” $\perp$ “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

\[ y \sim N^{-1}(\eta, \Sigma) \]

and divide the dimensions into two sets \( A, \bar{A} \).

Then,

\[
\begin{bmatrix}
    Y_A \\
    \bar{Y}_A
\end{bmatrix} \sim N^{-1}
\begin{bmatrix}
    \begin{bmatrix}
        \eta_A \\
        \eta_{\bar{A}}
    \end{bmatrix},
    \begin{bmatrix}
        \Sigma_{AA} & \Sigma_{A\bar{A}} \\
        \Sigma_{\bar{A}A} & \Sigma_{\bar{A}\bar{A}}
    \end{bmatrix}
\end{bmatrix}
\]

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

\[ \text{cond. cov.} \]

Let \( A = \{s, t\} \), \( \bar{A} = \text{everything else} \).

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

\[ \text{cov}(y_s, y_t | y_s) = \Sigma_{aa} = \begin{bmatrix}
    \Sigma_{ss} & 0 \\
    0 & \Sigma_{tt}
\end{bmatrix} \]

Therefore,

\[ y_s \perp y_t | y_s \iff \Sigma_{st} = 0 \]
Connection with Graphical Models

- Undirected graphical model or Markov random field (MRF)

No edge $\iff y_s \perp \!\!\!\!\!\!\perp y_t | y_{\neq s,t}$

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

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

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

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

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

Sparse Precision vs. Covariance

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

node potentials

edge potentials

D’s spread a cond. ind. statement
does not imply sparsity of cov. (ind. assumption)

$\{ y_i \}$ in still fully correlated!
Assume a known graph \( G = \{V,E\} \)

Rewrite log likelihood:

\[
\begin{align*}
\log p(y|\Theta) &= \frac{N}{2} \log |\Omega| - \frac{1}{2} \sum_i (y_i - \mu)^T \Omega (y_i - \mu) \\
&= \frac{N}{2} \log |\Omega| - \frac{1}{2} \text{tr}[(y_i - \mu)(y_i - \mu)^T \Omega] \\
&= \frac{N}{2} \log |\Omega| - \frac{1}{2} \text{tr}(S_i \Omega) \\
&\approx \frac{1}{N} \sum_i (y_i - \mu)(y_i - \mu)^T
\end{align*}
\]

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

Take gradient:

\[
\nabla L(\Omega) = \Omega^{-1} S \\
\text{s.t. } \Omega_{ii} = 0 \quad \text{if } (i,i) \notin E \\
\Omega > 0 \quad \text{pos.def.} \\
\]

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)
- ...
Estimating Graph Structure

- To learn the structure of the Gaussian graphical model, we want to trade off fit and sparsity
  - Measure of fit: \( \log \text{likelihood} = \log |\Sigma| - \text{tr}(S\Sigma) + \text{const} \)
  - Encouraging sparsity: \( \|\Sigma\|_1 = \sum_i \|\Sigma_{ii}\| \), want to min

- Overall objective = “graphical LASSO” or “Glasso”

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

Just as in LASSO, but w/ a matrix parameter and s.t. \( \Omega \succeq 0 \)
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)

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

- Other approach = ADMM

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 $i = 1, \ldots, j - 1, j + 1, \ldots, d$.

2. Let $A$ denote the $\mathbb{R}^d \times \mathbb{R}^d$ 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 $O(24^3)$. 

©Emily Fox 2014