Undirected graphs and the graphical lasso

- Undirected graphical models
- Gaussian models; known structure
- Gaussian models with unknown structure — graphical lasso
- Binary graphical models
- Further details chapter 17
Undirected graphs and the graphical lasso

- We have $p$ variables measured on $N$ observations—eg $p$ proteins measured in $N$ cells
- Our goal is to estimate the best undirected graph on the variables. In a such a graph, an edge between two nodes implies that the variables at those nodes have non-zero partial dependence. For Gaussian variables, this is non-zero partial correlation. That is, the two variables are correlated given the other variables.
- A missing edge between two nodes means that the variables at those nodes are conditionally independent, given the other variables.
Data for fitting graphical models

Example: $p = 5$ proteins measured in $N = 10,000$ cells from FACS (Fluorescence-Activated Cell Sorting)

<table>
<thead>
<tr>
<th></th>
<th>protein1</th>
<th>protein2</th>
<th>protein3</th>
<th>protein4</th>
<th>protein5</th>
</tr>
</thead>
<tbody>
<tr>
<td>cell 1</td>
<td>2.5</td>
<td>2.6</td>
<td>-1.3</td>
<td>2.8</td>
<td>4.7</td>
</tr>
<tr>
<td>cell 2</td>
<td>4.7</td>
<td>-3.3</td>
<td>1.8</td>
<td>3.3</td>
<td>3.4</td>
</tr>
<tr>
<td>cell 3</td>
<td>-1.2</td>
<td>-1.4</td>
<td>2.1</td>
<td>4.4</td>
<td>2.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
An undirected graph with 5 nodes. There is no edge, for example, between $x_2$ and $x_3$ and hence $x_2$ and $x_3$ are independent given $x_1, x_4,$ and $x_5$. 
Other types of graphs

- **Correlation graphs** display the correlation (rather than partial correlation) between the variables.

- **Directed acyclic graphs (DAGs)** have directed edges (arrows). These represent joint probability distributions that can be factored into products of conditional distributions and they try to model causality. They are much more difficult to estimate than undirected graphs.
Naive procedure for constructing graphs

(Meinshausen & Bühlmann 2006).

• The coefficient of the $j$th predictor in a linear regression is proportional to the partial correlation between the response variable and the $j$th predictor.

• Idea: apply the lasso to the graph problem by treating each node in turn as the response variable.

• include an edge $i \leftrightarrow j$ in the graph if either the coefficient of $j$th predictor in regression for $x_i$ is non-zero, or the $i$th predictor in regression for $x_j$ is non-zero.
Some background

- Recall that if $X = (X_1, X_2, \ldots, X_p)$ has a multivariate Gaussian distribution with mean-vector 0 (for convenience) and covariance $\Sigma$, then $\Theta = \Sigma^{-1}$ captures the conditional distributions of each $X_j$ given the rest.

- Hence we can think of the edges in a graph as representing the non-zeroes in the inverse covariance $\Sigma^{-1}$. A missing edge between nodes $i$ and $j$ means that $(\Sigma^{-1})_{ij} = 0$. This means that $X_i$ and $X_j$ are conditionally independent given the other variables.
More background

• Now

\[ X_j | X_{-j} \sim N(\sum_{i \neq j} X_i \beta_{ij}, \sigma_{jj}), \]

where

\[ \beta_{ij} = -\frac{\theta_{ij}}{\theta_{jj}}; \quad \sigma_{jj} = \frac{1}{\theta_{jj}}. \]

• Thus the regression coefficient \( \beta_{ij} \) is proportional to the partial covariance \( \theta_{ij} \); it is also proportional to the partial correlation \( \rho_{ij} = \theta_{ij}/(\theta_{ii}\theta_{jj})^{1/2} \). In particular \( \beta_{ij} = 0 \iff \theta_{ij} = 0 \), \( \beta_{ij} = 0 \iff \rho_{ij} = 0 \)

• The naive procedure from above ignores the symmetric relationship \( \theta_{ij} = \theta_{ji} \).
A simpler problem

Fitting graphs with fixed missing edges

How can we fit a model like the one below, forcing the missing edges to be omitted?
A simpler problem — continued

- Formally, \( X = (X_1, X_2, \ldots, X_p) \sim N(0, \Sigma) \) and we want to fit the model constraining some elements of \( \Sigma^{-1} \) to be zero.

- Setting \( \Theta = \Sigma^{-1} \), the log-likelihood can be written as
  \[
  \ell(\Theta) = \log \det \Theta - \text{tr}(S\Theta)
  \]
  where \( S \) is the observed covariance matrix of the data.

- One idea: fit a linear regression model for each node as a function of the others, leaving out a predictor if the edge is missing. Eg leave out \( X_3, X_4, X_5 \) in the model for predicting \( X_1 \). Then we use the regression coefficients to fill in the corresponding elements of \( \hat{\Sigma}^{-1} \).

- This doesn’t quite seem right, as it ignores the fact the 5 models should somehow be linked together.
An exact solution

- Consider a standard LS regression of the outcome $y = X_j$ on a predictor matrix $Z = X_{-j}$. The estimates for this fit can be written as $(Z^T Z)^{-1} (Z^T y)$.

- Now $Z^T Z$ is the sample covariance of the predictors. The algorithm for exact maximization of $\ell(\Theta)$ replaces $Z^T Z$ with $\hat{\Sigma}_{-j,-j}$, the model’s current estimate of the covariance among the $X_{-j}$; a modified linear regression,

- We apply this modified linear regression to each node, leaving out variables that represent missing edges; and cycling through the nodes until convergence

- this is a simple, apparently new algorithm for this problem
A Modified Regression Algorithm for Estimation of an Undirected Gaussian Graphical Model with Known Structure.

1. Initialize $\mathbf{W} = \mathbf{S}$.

2. Repeat for $j = 1, 2, \ldots, p, 1, 2, \ldots, p, \ldots$ until convergence:
   
   (a) Partition the matrix $\mathbf{W}$ into part 1: all but the $j$th row and column, and part 2: the $j$th row and column.

   (b) Solve $\mathbf{W}_{11}^* \beta^* - s_{12}^* = 0$ for the unconstrained edge parameters $\beta^*$, using the reduced system of equations that omits excluded variables. Obtain $\hat{\beta}$ by padding $\hat{\beta}^*$ with zeros in the appropriate positions.

   (c) Update $w_{12} = \mathbf{W}_{11} \hat{\beta}$

3. In the final cycle (for each $j$) solve for $\hat{\theta}_{12} = -\hat{\beta} \cdot \hat{\theta}_{22}$, with $1/\hat{\theta}_{22} = s_{22} - w_{12}^T \hat{\beta}$. 
Where does update $w_{12} = \mathbf{W}_{11}\hat{\beta}$ come from?

\[
\begin{pmatrix}
\mathbf{W}_{11} & w_{12} \\
w_{12}^T & w_{22}
\end{pmatrix}
\begin{pmatrix}
\Theta_{11} & \theta_{12} \\
\theta_{12}^T & \theta_{22}
\end{pmatrix}
= 
\begin{pmatrix}
\mathbf{I} & 0 \\
0^T & 1
\end{pmatrix}.
\]  

(1)

This implies

\[
w_{12} = -\mathbf{W}_{11} \theta_{12} / \theta_{22}
\]  

(2)

\[
= \mathbf{W}_{11} \beta
\]  

(3)

This as part of the formula for the inverse of a partitioned inverse of a matrix.
An example

\[ S = \begin{pmatrix} 10 & 1 & 5 & 4 \\ 1 & 10 & 2 & 6 \\ 5 & 2 & 10 & 3 \\ 4 & 6 & 3 & 10 \end{pmatrix} \]
An example — continued

We apply the preceding algorithm to this problem; for example, in the modified regression for variable 1 variable 3 is left out. The procedure quickly converged to the solutions:

\[
\Sigma = \begin{pmatrix}
10.00 & 1.00 & 1.31 & 4.00 \\
1.00 & 10.00 & 2.00 & 0.87 \\
1.31 & 2.00 & 10.00 & 3.00 \\
4.00 & 0.87 & 3.00 & 10.00 \\
\end{pmatrix}
\]

\[
\Sigma^{-1} = \begin{pmatrix}
0.12 & -0.01 & 0.00 & -0.05 \\
-0.01 & 0.11 & -0.02 & 0.00 \\
0.00 & -0.02 & 0.11 & -0.03 \\
-0.05 & 0.00 & -0.03 & 0.13 \\
\end{pmatrix}
\]

Note the zeroes in \(\hat{\Sigma}^{-1}\), corresponding to the missing edges (1,3) and (2,4). Note also that the corresponding elements in \(\hat{\Sigma}\) are the only elements different from \(S\). The estimation of \(\hat{\Sigma}\) is an example of what is sometimes called the positive definite “completion” of \(S\).
Back to estimation of the graph structure

A more exact formulation

- Assume $x \sim N(0, \Sigma)$ and let $\ell(\Sigma; X)$ be the log-likelihood. $X_j, X_k$ are conditionally independent if $(\Sigma^{-1})_{jk} = 0$

- Let $\Theta = \Sigma^{-1}$, and let $S$ be the empirical covariance matrix, the problem is to maximize the penalized log-likelihood

\[
\log \det \Theta - \text{tr}(S\Theta) \quad \text{subject to } ||\Theta||_1 \leq t,
\]

- Proposed in Yuan and Lin (2007)

- Convex problem! How to maximize? Consider generic lasso problem

\[
\min \left( \frac{1}{2} (z - X\beta)^T (z - X\beta) \right) \quad \text{s.t. } |\beta|_1 \leq t.
\]

- we can write this as in terms of inner products as

\[
\min J(\beta, X^TX, z^TX) \equiv \beta^T X^TX\beta / 2 - z^TX\beta
\]

s.t. $|\beta|_1 \leq t$
• the Naive method for the graphic structure problem applies the lasso to the data \((X_{[\cdot,j]}, X_j)\). This is same as minimizing 
\[ J(\beta, (X^T X)_{[\cdot,j], [\cdot,j]}, X^T_{[j,j]} X_{[\cdot,j], [\cdot,j]}). \]

• the exact method –The graphical lasso– substitutes \(\hat{\Sigma}_{[\cdot,j,j]}\) for 
\((X^T X)_{[\cdot,j,j]}\) in the above.
Graphical Lasso

1. Initialize $W = S + \lambda I$. The diagonal of $W$ remains unchanged in what follows.

2. Repeat for $j = 1, 2, \ldots p, 1, 2, \ldots p, \ldots$ until convergence:
   
   (a) Partition the matrix $W$ into part 1: all but the $j$th row and column, and part 2: the $j$th row and column.
   
   (b) Solve the estimating equations
   
   \[ W_{11}\beta - s_{12} + \lambda \cdot \text{Sign}(\beta) = 0 \]
   
   using the cyclical coordinate-descent algorithm
   
   (c) Update $w_{12} = W_{11}\hat{\beta}$

3. In the final cycle (for each $j$) solve for $\hat{\theta}_{12} = -\hat{\beta} \cdot \hat{\theta}_{22}$, with

\[
1/\hat{\theta}_{22} = w_{22} - w_{12}^T\hat{\beta}.
\]
Why does this work?

• Considered one row and column at a time, the log-likelihood and the lasso problem have the same subgradient

• Banerjee, Ghaoui & d’Aspremont (2008) found this connection, but didn’t fully exploit it

• Resulting algorithm is a blockwise coordinate descent procedure
### Speed comparisons

*Timings for $p = 400$ variables*

<table>
<thead>
<tr>
<th>Method</th>
<th>CPU time</th>
</tr>
</thead>
<tbody>
<tr>
<td>State-of-the-art convex optimizer</td>
<td>27 min</td>
</tr>
<tr>
<td>Graphical lasso</td>
<td>6.2 sec</td>
</tr>
</tbody>
</table>
Cell signalling proteins
Estimates from the graphical lasso procedure, for different values of the regularization parameter: starting at the top left graph (no regularization) to the bottom right (full regularization).
Binary graphical models

- Observations at nodes are binary valued: 0 or 1.
- Standard model is the Ising model

\[
p(X, \Theta) = \exp \left[ \sum_{(j,k) \in E} \theta_{jk} X_j X_k - \Phi(\Theta) \right]
\]

where \( E \) is the set of edges in the graph

- It implies a logistic form for each node conditional on the others:

\[
\Pr(X_j = 1 | X_{-j} = x_{-j}) = \frac{1}{1 + \exp(-\theta_{j0} - \sum_{(j,k) \in E} \theta_{jk} x_k)}
\]
Binary graphical models — continued

- difficulty is computation of the normalizing constant ("partition function") \( \Phi(\Theta) \), which requires enumeration over \( 2^N \) possible data values. Gibbs sampling is often used to approximate the resulting expectations.

- generalizations include Boltzmann machines for image modelling: see Chapter 17 and Geoff Hinton’s website at University of Toronto.

- Graphical lasso can be adapted, but not is not very effective since the bottleneck is the partition function.

• glasso—graphical lasso
References

