An undirected graph, each vertex represents a random variable.

The absence of an edge between two vertices means the corresponding random variables are conditionally independent, given other variables.

The Gaussian distribution is widely used for such graphical models, because of its convenient analytical properties.

Penalized regression methods for inducing sparsity in the precision matrix are central to the construction of Gaussian graphical models.
Denote the covariance matrix by $\Sigma$, then the inverse covariance matrix $\Theta = \Sigma^{-1}$ is called precision matrix. Let $\theta_{ij}$ be the $(i, j)$th element of $\Theta$.

$$\theta_{ij} = -\sigma_{ij; \text{rest}} \det(\Sigma^{(ij)}) \det(\Sigma)^{-1}.$$ 

- $\sigma_{ij; \text{rest}}$: conditional/partial covariance of variables $i$ and $j$, given the other variables.
- $\Sigma^{(ij)}$: matrix $\Sigma$ with $i$th row and $j$th column removed.
- If $\theta_{ij} = 0$, then variables $i$ and $j$ are conditionally independent, given other variables.
Suppose we partition $X^T = (X_1^T, X_2)$, where $X_1$ consists of the first $d - 1$ variables and $X_2$ is the last.

We have the partition of $\Sigma$ and $\Theta$:

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \sigma_{12} \\ \sigma_{12}^T & \sigma_{22} \end{pmatrix}, \quad \Theta = \begin{pmatrix} \Theta_{11} & \theta_{12} \\ \theta_{12}^T & \theta_{22} \end{pmatrix}.$$

Let $\beta = \Sigma_{11}^{-1}\sigma_{12}$ be the multiple linear regression coefficient of $X_2$ on $X_1$.

Since $\Sigma\Theta = I$,

$$\Sigma_{11}\theta_{12} + \sigma_{12}\theta_{22} = 0,$$

$$\beta = \Sigma_{11}^{-1}\sigma_{12} = -\theta_{12}/\theta_{22}.$$
Regression coefficient:

\[ \beta = -\frac{\theta_{12}}{\theta_{22}}. \]

We can learn about the dependence structure through multiple linear regression.

Meinshausen and Bhlmann (2006) try to estimate which components \( \theta_{ij} \) are zero, rather than fully estimate \( \Theta \). They fit a lasso regression using each variable as the response and the others as predictors.
Lasso

- Minimize

\[ Q(\beta) = \frac{1}{2} \| Y - X \beta \|^2 + \lambda \sum_j |\beta_j|. \]

- When \( n = p = 1 \) and \( X = 1 \),

\[ Q(\beta) = \frac{1}{2} (y - \beta)^2 + \lambda |\beta|. \]

\[ Q'(\beta) = -y + \beta + \lambda \cdot \text{sign}(\beta) = 0. \]

- Lasso solution

\[ \hat{\beta}(\lambda) = \text{sign}(y)(|y| - \lambda)_+ = S(y, \lambda), \]

where \( S(y, \lambda) \) is called the soft-thresholding operator.
A more systematic approach by Friedman, Hastie and Tibshirani (2008).

- Consider maximizing the penalized log-likelihood

\[
\log(\det(\Theta)) - \text{trace}(S\Theta) - \lambda\|\Theta\|_1.
\]

- \(S\): sample covariance matrix.
- \(\|\Theta\|_1\): element \(L_1\) norm, the sum of the absolute values of the elements of \(\Theta\).
- The gradient equation

\[
\Theta^{-1} - S - \lambda \cdot \text{Sign}(\Theta) = 0.
\]
The gradient equation

$$\Theta^{-1} - S - \lambda \cdot \text{Sign}(\Theta) = 0.$$ 

Let $W = \Theta^{-1}$ and

$$
\begin{pmatrix}
W_{11} & w_{12} \\
W_{12}^T & w_{22}
\end{pmatrix}
\begin{pmatrix}
\Theta_{11} & \theta_{12} \\
\theta_{12}^T & \theta_{22}
\end{pmatrix} =
\begin{pmatrix}
I & 0 \\
0^T & 1
\end{pmatrix}.
$$

$$w_{12} = -W_{11}\theta_{12}/\theta_{22} = W_{11}\beta,$$

where $\beta = -\theta_{12}/\theta_{22}$.

The upper right block of the gradient equation:

$$W_{11}\beta - s_{12} + \lambda \cdot \text{Sign}(\beta) = 0$$

which is recognized as the estimation equation for the Lasso regression.
Algorithm 17.2 Graphical Lasso.

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

2. Repeat for $j = 1, 2, \ldots p, 1, 2, \ldots p, \ldots$ until convergence:
   
   (a) Partition the matrix $\mathbf{W}$ 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 $\mathbf{W}_{11} \hat{\beta} - s_{12} + \lambda \cdot \text{Sign}(\beta) = 0$ using the cyclical coordinate-descent algorithm (17.26) for the modified lasso.

   (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 $1/\hat{\theta}_{22} = w_{22} - w_{12}^T \hat{\beta}$.
Coordinate descent: Let $V = W_{11}$,

$$\hat{\beta}_i \leftarrow S(s_{12i} - \sum_{k \neq j} V_{ki} \hat{\beta}_k, \lambda) / V_{ii},$$

where $S(y, \lambda)$ is the soft-thresholding operator.
We analyze a flow cytometry dataset on $d = 11$ proteins and $n = 7466$ cells. Several methods are compared:

- Graphical Lasso
- Bayesian Network
- Truncated Vine (Sequential MST)
- Factor Analysis
A common discrepancy measure in the psychometrics and structural equation modeling literatures is:

\[ D = \log(\det[R_{\text{model}}(\hat{\delta})]) - \log(\det[R_{\text{data}}]) + \text{tr}[R_{\text{model}}^{-1}(\hat{\delta})R_{\text{data}}] - d. \]

\(d\): number of variables.

\(R_{\text{data}}\): sample correlation matrix.

\(R_{\text{model}}(\hat{\delta})\): model-based correlation matrix based on the estimate of the parameter \(\delta\). If either model has some conditional independence relations, then the dimension of \(\delta\) is less than \(d(d - 1)/2\).
Other comparisons are the AIC/BIC based on a Gaussian log-likelihood.

Also useful are the average and max absolute deviations of the model-based correlation matrix from the empirical correlation matrix:

$$\max_{j<k} |R_{\text{data},jk} - R_{\text{model},jk}(\hat{\delta})|.$$
<table>
<thead>
<tr>
<th>Model</th>
<th>Dfit</th>
<th>MaxAbsDiff</th>
<th>AIC($\times 10^5$)</th>
<th>BIC($\times 10^5$)</th>
<th>#Par</th>
</tr>
</thead>
<tbody>
<tr>
<td>BN</td>
<td>0.013</td>
<td>0.019</td>
<td>1.969</td>
<td>1.972</td>
<td>36</td>
</tr>
<tr>
<td>glasso($\lambda = 0.13$)</td>
<td>1.232</td>
<td>0.200</td>
<td>2.060</td>
<td>2.062</td>
<td>33</td>
</tr>
<tr>
<td>glasso($\lambda = 0.10$)</td>
<td>0.930</td>
<td>0.159</td>
<td>2.038</td>
<td>2.040</td>
<td>37</td>
</tr>
<tr>
<td>glasso($\lambda = 0.08$)</td>
<td>0.700</td>
<td>0.126</td>
<td>2.020</td>
<td>2.023</td>
<td>41</td>
</tr>
<tr>
<td>1-truncated seq. MST</td>
<td>1.030</td>
<td>0.306</td>
<td>2.044</td>
<td>2.045</td>
<td>10</td>
</tr>
<tr>
<td>2-truncated seq. MST</td>
<td>0.568</td>
<td>0.242</td>
<td>2.010</td>
<td>2.012</td>
<td>19</td>
</tr>
<tr>
<td>3-truncated seq. MST</td>
<td>0.328</td>
<td>0.197</td>
<td>1.992</td>
<td>1.994</td>
<td>27</td>
</tr>
<tr>
<td>4-truncated seq. MST</td>
<td>0.224</td>
<td>0.229</td>
<td>1.985</td>
<td>1.987</td>
<td>34</td>
</tr>
<tr>
<td>5-truncated seq. MST</td>
<td>0.142</td>
<td>0.150</td>
<td>1.979</td>
<td>1.982</td>
<td>40</td>
</tr>
<tr>
<td>1-factor</td>
<td>2.682</td>
<td>0.571</td>
<td>2.168</td>
<td>2.169</td>
<td>11</td>
</tr>
<tr>
<td>2-factor</td>
<td>1.689</td>
<td>0.529</td>
<td>2.094</td>
<td>2.095</td>
<td>21</td>
</tr>
<tr>
<td>3-factor</td>
<td>0.832</td>
<td>0.456</td>
<td>2.030</td>
<td>2.032</td>
<td>30</td>
</tr>
<tr>
<td>4-factor</td>
<td>0.245</td>
<td>0.119</td>
<td>1.986</td>
<td>1.989</td>
<td>38</td>
</tr>
</tbody>
</table>


The End