Graphical Gaussian Models

Steffen Lauritzen, University of Oxford

Graphical Models and Inference, Lecture 12, Michaelmas Term 2009

November 26, 2009
A $d$-dimensional random vector $X = (X_1, \ldots, X_d)$ is has a multivariate Gaussian distribution or normal distribution on $\mathbb{R}^d$ if there is a vector $\xi \in \mathbb{R}^d$ and a $d \times d$ matrix $\Sigma$ such that

$$\lambda^T X \sim \mathcal{N}(\lambda^T \xi, \lambda^T \Sigma \lambda) \quad \text{for all } \lambda \in \mathbb{R}^d. \quad (1)$$

We then write $X \sim \mathcal{N}_d(\xi, \Sigma)$.

It holds that

$$X_i \sim \mathcal{N}(\xi_i, \sigma_{ii}), \quad \text{Cov}(X_i, X_j) = \sigma_{ij}.$$  

Hence $\xi$ is the mean vector and $\Sigma$ the covariance matrix of the distribution.
Density of multivariate Gaussian

If $\Sigma$ is *positive definite*, i.e. if $\lambda^\top \Sigma \lambda > 0$ for $\lambda \neq 0$, the distribution has density on $\mathcal{R}^d$

$$f(x \mid \xi, \Sigma) = (2\pi)^{-d/2} (\det K)^{1/2} e^{-(x-\xi)^\top K (x-\xi)/2}, \quad (2)$$

where $K = \Sigma^{-1}$ is the *concentration matrix* of the distribution. We then also say that $\Sigma$ is *regular*. 
Adding independent Gaussians yields a Gaussian

If \( X \sim \mathcal{N}_d(\xi_1, \Sigma_1) \) and \( X_2 \sim \mathcal{N}_d(\xi_2, \Sigma_2) \) and \( X_1 \perp \perp X_2 \)

\[
X_1 + X_2 \sim \mathcal{N}_d(\xi_1 + \xi_2, \Sigma_1 + \Sigma_2).
\]

Linear transformations preserve Gaussianity:

\[
Y = AX + b \sim \mathcal{N}_r(A\xi + b, A\Sigma A^\top).
\]
Partition $X$, $\xi$, $K$ and $\Sigma$ as

$$\xi = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}, \quad K = \begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

Then, if $X \sim \mathcal{N}_d(\xi, \Sigma)$ it holds that $X_2 \sim \mathcal{N}_s(\xi_2, \Sigma_{22})$.

If $\Sigma_{22}$ is regular, it further holds that

$$X_1 \mid X_2 = x_2 \sim \mathcal{N}_r(\xi_{1\mid 2}, \Sigma_{1\mid 2}),$$

where

$$\xi_{1\mid 2} = \xi_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \xi_2) = \xi_1 - K_{11}^{-1}K_{12}(x_2 - \xi_2)$$

and

$$\Sigma_{1\mid 2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = (K_{11})^{-1}.$$
Consider the case where $\xi = 0$ and a sample $X^1 = x^1, \ldots, X^n = x^n$ from a multivariate Gaussian distribution $\mathcal{N}_d(0, \Sigma)$ with $\Sigma$ regular. Using (2), we get the likelihood function

$$L(K) = (2\pi)^{-nd/2} (\det K)^{n/2} e^{-\sum_{\nu=1}^n (x^\nu)^\top K x^\nu / 2}$$

$$\propto (\det K)^{n/2} e^{-\sum_{\nu=1}^n \text{tr}\{K x^\nu (x^\nu)^\top\} / 2}$$

$$= (\det K)^{n/2} e^{-\text{tr}\{K \sum_{\nu=1}^n x^\nu (x^\nu)^\top\} / 2}$$

$$= (\det K)^{n/2} e^{-\text{tr}(K W) / 2}. \quad (3)$$

where

$$W = \sum_{\nu=1}^n X^\nu (X^\nu)^\top$$

is the matrix of *sums of squares and products*. 

Steffen Lauritzen, University of Oxford
Writing the trace out

\[ \text{tr}(KW) = \sum_i \sum_j k_{ij} W_{ji} \]

emphasizes that it is linear in both \( K \) and \( W \) and we can recognize this as a linear and canonical exponential family with \( K \) as the canonical parameter and \(-W/2\) as the canonical sufficient statistic. Thus, the likelihood equation becomes

\[ E(-W/2) = -n\Sigma/2 = -W/2 \]

since \( E(W) = n\Sigma \). Solving, we get

\[ \hat{K}^{-1} = \hat{\Sigma} = W/n \]

in analogy with the univariate case.
Rewriting the likelihood function as

\[ \log L(K) = \frac{n}{2} \log(\det K) - \text{tr}(KW)/2 \]

we can of course also differentiate to find the maximum, leading to the equation

\[ \frac{\partial}{\partial k_{ij}} \log(\det K) = w_{ij}/n, \]

which in combination with the previous result yields

\[ \frac{\partial}{\partial K} \log(\det K) = K^{-1}. \]

The latter can also be derived directly by writing out the determinant, and it holds for any non-singular square matrix, i.e. one which is not necessarily positive definite.
The Wishart distribution is the sampling distribution of the matrix of sums of squares and products. More precisely:

A random $d \times d$ matrix $W$ has a \textit{d-dimensional Wishart distribution} with parameter $\Sigma$ and $n$ \textit{degrees of freedom} if

$$W \overset{D}{=} \sum_{i=1}^{n} X^\nu (X^\nu)^\top$$

where $X^\nu \sim \mathcal{N}_d(0, \Sigma)$. We then write

$$W \sim \mathcal{W}_d(n, \Sigma).$$

The Wishart is the multivariate analogue to the $\chi^2$:

$$\mathcal{W}_1(n, \sigma^2) = \sigma^2 \chi^2(n).$$

If $W \sim \mathcal{W}_d(n, \Sigma)$ its mean is $\mathbf{E}(W) = n\Sigma$. 
If $W_1$ and $W_2$ are independent with $W_i \sim \mathcal{W}_d(n_i, \Sigma)$, then
\[W_1 + W_2 \sim \mathcal{W}_d(n_1 + n_2, \Sigma).\]

If $A$ is an $r \times d$ matrix and $W \sim \mathcal{W}_d(n, \Sigma)$, then
\[AWA^\top \sim \mathcal{W}_r(n, A\Sigma A^\top).\]

For $r = 1$ we get that when $W \sim \mathcal{W}_d(n, \Sigma)$ and $\lambda \in \mathbb{R}^d$,
\[\lambda^\top W \lambda \sim \sigma_{\lambda}^2 \chi^2(n),\]
where $\sigma_{\lambda}^2 = \lambda^\top \Sigma \lambda$. 

The multivariate Gaussian Distribution
The Wishart distribution
Gaussian graphical models

Definition
Basic properties
Wishart density

If $W \sim \mathcal{W}_d(n, \Sigma)$, where $\Sigma$ is regular, then $W$ is regular with probability one if and only if $n \geq d$.

When $n \geq d$ the Wishart distribution has density

$$f_d(w \mid n, \Sigma) = c(d, n)^{-1} (\det \Sigma)^{-n/2} (\det w)^{(n-d-1)/2} e^{-\text{tr}(\Sigma^{-1}w)/2}$$

for $w$ positive definite, and 0 otherwise.

The Wishart constant $c(d, n)$ is

$$c(d, n) = 2^{nd/2} (2\pi)^{d(d-1)/4} \prod_{i=1}^{d} \Gamma\{(n + 1 - i)/2\}.$$
Consider $X = (X_v, v \in V) \sim \mathcal{N}_V(0, \Sigma)$ with $\Sigma$ regular and $K = \Sigma^{-1}$.

The concentration matrix of the conditional distribution of $(X_\alpha, X_\beta)$ given $X_{V \setminus \{\alpha, \beta\}}$ is

$$K_{\{\alpha, \beta\}} = \begin{pmatrix}
    k_{\alpha\alpha} & k_{\alpha\beta} \\
    k_{\beta\alpha} & k_{\beta\beta}
\end{pmatrix}.$$

Hence

$$\alpha \perp \perp \beta \mid V \setminus \{\alpha, \beta\} \iff k_{\alpha\beta} = 0.$$

Thus the dependence graph $\mathcal{G}(K)$ of a regular Gaussian distribution is given by

$$\alpha \not\sim \beta \iff k_{\alpha\beta} = 0.$$
$S(G)$ denotes the symmetric matrices $A$ with $a_{\alpha \beta} = 0$ unless $\alpha \sim \beta$ and $S^+(G)$ their positive definite elements.

A **Gaussian graphical model** for $X$ specifies $X$ as multivariate normal with $K \in S^+(G)$ and otherwise unknown.

Note that the density then factorizes as

$$\log f(x) = \text{constant} - \frac{1}{2} \sum_{\alpha \in V} k_{\alpha \alpha} x_{\alpha}^2 - \sum_{\{\alpha, \beta\} \in E} k_{\alpha \beta} x_{\alpha} x_{\beta},$$

hence *no interaction terms involve more than pairs*.

This is different from the discrete case and generally makes things easier.
Examination marks of 88 students in 5 different mathematical subjects. The empirical concentration matrix is

<table>
<thead>
<tr>
<th></th>
<th>Mechanics</th>
<th>Vectors</th>
<th>Algebra</th>
<th>Analysis</th>
<th>Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mech</td>
<td>5.24</td>
<td>-2.44</td>
<td>-2.74</td>
<td>0.01</td>
<td>-0.14</td>
</tr>
<tr>
<td>Vec</td>
<td>-2.44</td>
<td>10.43</td>
<td>-4.71</td>
<td>-0.79</td>
<td>-0.17</td>
</tr>
<tr>
<td>Alg</td>
<td>-2.74</td>
<td>-4.71</td>
<td>26.95</td>
<td>-7.05</td>
<td>-4.70</td>
</tr>
<tr>
<td>An</td>
<td>0.01</td>
<td>-0.79</td>
<td>-7.05</td>
<td>9.88</td>
<td>-2.02</td>
</tr>
<tr>
<td>Stats</td>
<td>-0.14</td>
<td>-0.17</td>
<td>-4.70</td>
<td>-2.02</td>
<td>6.45</td>
</tr>
</tbody>
</table>
The multivariate Gaussian Distribution
The Wishart distribution
Gaussian graphical models

Graphical model for mathmarks

Vectors

Analysis

Algebra

Mechanics

Statistics

This analysis is from Whittaker (1990).
We have An, Stats \perp\!
| Mech, Vec | Alg.
Frets’ heads

This example is concerned with a study of heredity of head dimensions (Frets 1921). Lengths $L_i$ and breadths $B_i$ of the heads of 25 pairs of first and second sons are measured. Previous analyses by Whittaker (1990) support the graphical model:

![Graphical model](image-url)
The likelihood function based on a sample of size $n$ is

$$L(K) \propto (\det K)^{n/2} e^{-\text{tr}(KW)/2},$$

where $W$ is the Wishart matrix of sums of squares and products, $W \sim \mathcal{W}|V|(n, \Sigma)$ with $\Sigma^{-1} = K \in S^+(\mathcal{G})$.

Define the matrices $A^u$, $u \in V \cup E$ as those with elements

$$a_{ij}^u = \begin{cases} 1 & \text{if } u \in V \text{ and } i = j = u \\ 1 & \text{if } u \in E \text{ and } u = \{i, j\} \\ 0 & \text{otherwise.} \end{cases}$$
Then, as $K \in \mathcal{S}(\mathcal{G})$, 

$$K = \sum_{v \in V} k_v A^v + \sum_{e \in E} k_e A^e$$  

(4)

and hence

$$\text{tr}(KW) = \sum_{v \in V} k_v \text{tr}(A^v W) + \sum_{e \in E} k_e \text{tr}(A^e W)$$

leading to the log-likelihood function

$$l(K) = \log L(K) \sim \frac{n}{2} \log(\det K) - \frac{1}{2} \text{tr}(KW)$$

$$= \frac{n}{2} \log(\det K) - \sum_{v \in V} k_v \text{tr}(A^v W)/2 + \sum_{e \in E} k_e \text{tr}(A^e W)/2.$$
Hence we can identify the family as a (regular and canonical) exponential family with 
\[ -\text{tr}(A^u W)/2, \ u \in V \cup E \] 
as canonical sufficient statistics.

The likelihood equations can be obtained from this fact or by differentiation, combining the fact that

\[ \frac{\partial}{\partial k_u} \log \det(K) = \text{tr}(A^u \Sigma) \]

with (4). This eventually yields the likelihood equations

\[ \text{tr}(A^u W) = n \text{tr}(A^u \Sigma), \quad u \in V \cup E. \]
The likelihood equations

\[ \text{tr}(A^u W) = n \text{tr}(A^u \Sigma), \quad u \in V \cup E. \]

can also be expressed as

\[ n\hat{\sigma}_{vv} = w_{vv}, \quad n\hat{\sigma}_{\alpha\beta} = w_{\alpha\beta}, \quad v \in V, \{\alpha, \beta\} \in E. \]

We should remember the model restriction \( \Sigma^{-1} \in S^+(G) \).

This ‘fits variances and covariances along nodes and edges in \( G \)’ so we can write the equations as

\[ n\hat{\Sigma}_{cc} = w_{cc} \text{ for all cliques } c \in C(G), \]

hence making the equations analogous to the discrete case.

General theory of exponential families ensure the solution to be unique, provided it exists.
For $K \in S^+(G)$ and $c \in C$, define the operation of ‘adjusting the $c$-marginal’ as follows. Let $a = V \setminus c$ and

$$T_cK = \begin{pmatrix} n(w_{cc})^{-1} + K_{ca}(K_{aa})^{-1}K_{ac} & K_{ca} \\ K_{ac} & K_{aa} \end{pmatrix}.$$ (5)

This operation is clearly well defined if $w_{cc}$ is positive definite.

Recall the identity

$$(K_{11})^{-1} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$ 

Switching the role of $K$ and $\Sigma$ yields

$$\Sigma_{11} = (K^{-1})_{11} = (K_{11} - K_{12}K_{22}^{-1}K_{21})^{-1}$$

and hence

$$\Sigma_{cc} = (K^{-1})_{cc} = \left\{ K_{cc} - K_{ca}(K_{aa})^{-1}K_{ac} \right\}^{-1}.$$
Thus the $C$-marginal covariance $\tilde{\Sigma}_{cc}$ corresponding to the adjusted concentration matrix becomes

\[
\tilde{\Sigma}_{cc} = \{ (T_c K)^{-1} \}_{cc} \\
= \{ n(w_{cc})^{-1} + K_{ca}(K_{aa})^{-1} K_{ac} - K_{ca}(K_{aa})^{-1} K_{ac} \}^{-1}
\]

\[
= w_{cc}/n,
\]

hence $T_c K$ does indeed adjust the marginals. From (5) it is seen that the pattern of zeros in $K$ is preserved under the operation $T_c$, and it can also be seen to stay positive definite.

In fact, $T_b$ scales proportionally in the sense that

\[
f\{ x \mid (T_c K)^{-1} \} = f(x \mid K^{-1}) \frac{f(x_c \mid w_{cc}/n)}{f(x_c \mid \Sigma_{cc})}.
\]

This clearly demonstrates the analogy to the discrete case.
Next we choose any ordering \((c_1, \ldots, c_k)\) of the cliques in \(G\). Choose further \(K_0 = I\) and define for \(r = 0, 1, \ldots\)

\[
K_{r+1} = (T_{c_1} \cdots T_{c_k})K_r.
\]

Then we have: Consider a sample from a covariance selection model with graph \(G\). Then

\[
\hat{K} = \lim_{r \to \infty} K_r,
\]

provided the maximum likelihood estimate \(\hat{K}\) of \(K\) exists.