Gaussian Graphical Models

Steffen Lauritzen
University of Oxford

CIMPA Summerschool, Hammamet 2011, Tunisia

September 8, 2011
A $d$-dimensional random vector $X = (X_1, \ldots, X_d)$ 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

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

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

Taking $\lambda = e_i$ or $\lambda = e_i + e_j$ where $e_i$ is the unit vector with $i$-th coordinate 1 and the remaining equal to zero yields:

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

Hence $\xi$ is the **mean vector** and $\Sigma$ the **covariance matrix** of the distribution.
The definition (1) makes sense if and only if $\lambda^\top \Sigma \lambda \geq 0$, i.e. if $\Sigma$ is positive semidefinite. Note that we have allowed distributions with variance zero.

The multivariate moment generating function of $X$ can be calculated using the relation (1) as

$$m_d(\lambda) = E\{e^{\lambda^\top X}\} = e^{\lambda^\top \xi + \lambda^\top \Sigma \lambda / 2}$$

where we have used that the univariate moment generating function for $\mathcal{N}(\mu, \sigma^2)$ is

$$m_1(t) = e^{t\mu + \sigma^2 t^2 / 2}$$

and let $t = 1$, $\mu = \lambda^\top \xi$, and $\sigma^2 = \lambda^\top \Sigma \lambda$.

In particular this means that a multivariate Gaussian distribution is determined by its mean vector and covariance matrix.
Assume $X^\top = (X_1, X_2, X_3)$ with $X_i$ independent and $X_i \sim \mathcal{N}(\xi_i, \sigma_i^2)$. Then

$$\lambda^\top X = \lambda_1 X_1 + \lambda_2 X_2 + \lambda_3 X_3 \sim \mathcal{N}(\mu, \tau^2)$$

with

$$\mu = \lambda^\top \xi = \lambda_1 \xi_1 + \lambda_2 \xi_2 + \lambda_3 \xi_3, \quad \tau^2 = \lambda_1^2 \sigma_1^2 + \lambda_2^2 \sigma_2^2 + \lambda_3^2 \sigma_3^2.$$ 

Hence $X \sim \mathcal{N}_3(\xi, \Sigma)$ with $\xi^\top = (\xi_1, \xi_2, \xi_3)$ and

$$\Sigma = \begin{pmatrix}
\sigma_1^2 & 0 & 0 \\
0 & \sigma_2^2 & 0 \\
0 & 0 & \sigma_3^2
\end{pmatrix}.$$
If $\Sigma$ is \textit{positive definite}, i.e. if $\lambda^\top \Sigma \lambda > 0$ for $\lambda \neq 0$, the distribution has density on $\mathcal{R}^d$

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

where $K = \Sigma^{-1}$ is the \textit{concentration matrix} of the distribution. Since a positive semidefinite matrix is positive definite if and only if it is invertible, we then also say that $\Sigma$ is \textit{regular}.

If $X_1, \ldots, X_d$ are independent and $X_i \sim \mathcal{N}(\xi_i, \sigma_i^2)$ their joint density has the form (2) with $\Sigma = \text{diag}(\sigma_i^2)$ and $K = \Sigma^{-1} = \text{diag}(1/\sigma_i^2)$.

Hence \textit{vectors of independent Gaussians are multivariate Gaussian}. 

In the bivariate case it is traditional to write

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_1 \sigma_2 \rho \\ \sigma_1 \sigma_2 \rho & \sigma_2^2 \end{pmatrix},$$

with $\rho$ being the *correlation* between $X_1$ and $X_2$. Then

$$\det(\Sigma) = \sigma_1^2 \sigma_2^2 (1 - \rho^2) = \det(K)^{-1}$$

and

$$K = \frac{1}{\sigma_1^2 \sigma_2^2 (1 - \rho^2)} \begin{pmatrix} \sigma_2^2 & -\sigma_1 \sigma_2 \rho \\ -\sigma_1 \sigma_2 \rho & \sigma_1^2 \end{pmatrix}.$$
Thus the density becomes

\[
f(x \mid \xi, \Sigma) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{(1 - \rho^2)}} e^{-\frac{1}{2(1 - \rho^2)} \left\{ \frac{(x_1 - \xi_1)^2}{\sigma_1^2} - 2\rho \frac{(x_1 - \xi_1)(x_2 - \xi_2)}{\sigma_1 \sigma_2} + \frac{(x_2 - \xi_2)^2}{\sigma_2^2} \right\}}.
\]

The contours of this density are ellipses and the corresponding density is bell-shaped with maximum in \((\xi_1, \xi_2)\).
The marginal distributions of a vector $X$ can all be Gaussian without the joint being multivariate Gaussian:

For example, let $X_1 \sim \mathcal{N}(0, 1)$, and define $X_2$ as

$$X_2 = \begin{cases} X_1 & \text{if } |X_1| > c \\ -X_1 & \text{otherwise.} \end{cases}$$

Then, using the symmetry of the univariate Gaussian distribution, $X_2$ is also distributed as $\mathcal{N}(0, 1)$. 
However, the joint distribution is not Gaussian unless $c = 0$ since, for example, $Y = X_1 + X_2$ satisfies

$$
P(Y = 0) = P(X_2 = -X_1) = P(|X_1| \leq c) = \Phi(c) - \Phi(-c).
$$

Note that for $c = 0$, the correlation $\rho$ between $X_1$ and $X_2$ is 1 whereas for $c = \infty$, $\rho = -1$.

It follows that \textit{there is a value of $c$ so that $X_1$ and $X_2$ are uncorrelated}, and still not jointly Gaussian.
Adding two 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 \( X_1 \perp \!\!\!\perp X_2 \)

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

To see this, just note that

\[
\lambda^\top (X_1 + X_2) = \lambda^\top X_1 + \lambda^\top X_2
\]

and use the univariate addition property.
Linear transformations preserve multivariate normality:

If $A$ is an $r \times d$ matrix, $b \in \mathbb{R}^r$ and $X \sim \mathcal{N}_d(\xi, \Sigma)$, then

$$Y = AX + b \sim \mathcal{N}_r(A\xi + b, A\Sigma A^\top).$$

Again, just write

$$\gamma^\top Y = \gamma^\top(AX + b) = (A^\top \gamma)^\top X + \gamma^\top b$$

and use the corresponding univariate result.
Partition $X$ into into $X_1$ and $X_2$, where $X_1 \in \mathcal{R}^r$ and $X_2 \in \mathcal{R}^s$ with $r + s = d$.

Partition mean vector, concentration and covariance matrix accordingly as

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

so that $\Sigma_{11}$ is $r \times r$ and so on. \textit{Then, if } $X \sim \mathcal{N}_d(\xi, \Sigma)$ \textit{ then}

$$
X_2 \sim \mathcal{N}_s(\xi_2, \Sigma_{22}).
$$

This follows simply from the previous fact using the matrix

$$
A = (0_{sr} \ I_s).
$$

where $0_{sr}$ is an $s \times r$ matrix of zeros and $I_s$ is the $s \times s$ identity matrix.
If $\Sigma_{22}$ is regular, it further holds that

$$X_1 \mid X_2 = x_2 \sim \mathcal{N}_r(\xi_{1\mid2}, \Sigma_{1\mid2}),$$

where

$$\xi_{1\mid2} = \xi_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \xi_2) \quad \text{and} \quad \Sigma_{1\mid2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

In particular, $\Sigma_{12} = 0$ if and only if $X_1$ and $X_2$ are independent.
To see this, we simply calculate the conditional density.

\[
f(x_1 | x_2) \propto f_{\xi, \Sigma}(x_1, x_2)
\]

\[
\propto \exp \left\{ -(x_1 - \xi_1)^\top K_{11} (x_1 - \xi_1) / 2 - (x_1 - \xi_1)^\top K_{12} (x_2 - \xi_2) \right\}.
\]

The linear term involving \( x_1 \) has coefficient equal to

\[
K_{11} \xi_1 - K_{12} (x_2 - \xi_2) = K_{11} \{ \xi_1 - K_{11}^{-1} K_{12} (x_2 - \xi_2) \}.
\]

Using the matrix identities

\[
K_{11}^{-1} = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \tag{3}
\]

and

\[
K_{11}^{-1} K_{12} = -\Sigma_{12} \Sigma_{22}^{-1}, \tag{4}
\]
we find

\[ f(x_1 \mid x_2) \propto \exp \left\{-(x_1 - \xi_{1|2})^\top K_{11}(x_1 - \xi_{1|2})/2 \right\} \]

and the result follows.

From the identities (3) and (4) it follows in particular that then the conditional expectation and concentrations also can be calculated as

\[ \xi_{1|2} = \xi_1 - K_{11}^{-1}K_{12}(x_2 - \xi_2) \quad \text{and} \quad K_{1|2} = K_{11}. \]

Note that the *marginal covariance is simply expressed in terms of \( \Sigma \) whereas the *conditional concentration is simply expressed in terms of \( K \). Further, \( X_1 \) and \( X_2 \) are independent if and only if \( K_{12} = 0 \), giving \( K_{12} = 0 \) if and only if \( \Sigma_{12} = 0 \).
Consider $\mathcal{N}_3(0, \Sigma)$ with covariance matrix

$$
\Sigma = \begin{pmatrix}
1 & 1 & 1 \\
1 & 2 & 1 \\
1 & 1 & 2
\end{pmatrix}.
$$

The concentration matrix is

$$
K = \Sigma^{-1} = \begin{pmatrix}
3 & -1 & -1 \\
-1 & 1 & 0 \\
-1 & 0 & 1
\end{pmatrix}.
$$
The marginal distribution of \((X_2, X_3)\) has covariance and concentration matrix

\[
\Sigma_{23} = \begin{pmatrix}
2 & 1 \\
1 & 2
\end{pmatrix}, \quad (\Sigma_{23})^{-1} = \frac{1}{3} \begin{pmatrix}
2 & -1 \\
-1 & 2
\end{pmatrix}.
\]

The conditional distribution of \((X_1, X_2)\) given \(X_3\) has concentration and covariance matrix

\[
K_{12} = \begin{pmatrix}
3 & -1 \\
-1 & 1
\end{pmatrix}, \quad \Sigma_{12|3} = (K_{12})^{-1} = \frac{1}{2} \begin{pmatrix}
1 & 1 \\
1 & 3
\end{pmatrix}.
\]

Similarly, \(V(X_1 \mid X_2, X_3) = 1/k_{11} = 1/3\), etc.
A square matrix $A$ has *trace*

$$\text{tr}(A) = \sum_i a_{ii}.$$ 

The trace has a number of properties:

1. $\text{tr}(\gamma A + \mu B) = \gamma \text{tr}(A) + \mu \text{tr}(B)$ for $\gamma, \mu$ being scalars;
2. $\text{tr}(A) = \text{tr}(A^\top)$;
3. $\text{tr}(AB) = \text{tr}(BA)$
4. $\text{tr}(A) = \sum_i \lambda_i$ where $\lambda_i$ are the *eigenvalues* of $A$. 


For symmetric matrices the last statement follows from taking an orthogonal matrix $O$ so that $OAO^\top = \text{diag}(\lambda_1, \ldots, \lambda_d)$ and using

$$tr(OAO^\top) = tr( AO^\top O) = tr(A).$$

The trace is thus \textit{orthogonally invariant}, as is the determinant:

$$\det(OAO^\top) = \det(O) \det(A) \det(O^\top) = 1 \det(A) 1 = \det(A).$$

There is an important trick that we shall use again and again: For $\lambda \in \mathbb{R}^d$

$$\lambda^\top A\lambda = tr(\lambda^\top A\lambda) = tr(A\lambda\lambda^\top)$$

since $\lambda^\top A\lambda$ is a scalar.
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

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

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

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

$$= (\det K)^{n/2} e^{-\text{tr}(Kw) / 2}. \quad (5)$$

where

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

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

\[ \mathbf{E}(-W/2) = -n\Sigma/2 = -w/2 \]

since \( \mathbf{E}(W) = n\Sigma \). 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

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

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

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

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

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

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

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

If $\mathbf{W} \sim \mathcal{W}_d(n, \Sigma)$ its mean is $\mathbf{E}(\mathbf{W}) = n\Sigma$. 

Steffen Lauritzen University of Oxford

Gaussian Graphical Models
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 $\sigma_\lambda^2 = \lambda^\top \Sigma \lambda$. 
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 \(X_V \backslash \{\alpha, \beta\}\) is

\[
K_{\{\alpha, \beta\}} = \begin{pmatrix}
k_{\alpha\alpha} & k_{\alpha\beta} \\
k_{\beta\alpha} & k_{\beta\beta}
\end{pmatrix},
\]

Hence

\[ \alpha \perp \perp \beta \mid V \backslash \{\alpha, \beta\} \iff k_{\alpha\beta} = 0. \]

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

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

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 concentrations (on or above diagonal) and partial correlations (below diagonal) are

<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>Mechanics</td>
<td>5.24</td>
<td>-2.44</td>
<td>-2.74</td>
<td>0.01</td>
<td>-0.14</td>
</tr>
<tr>
<td>Vectors</td>
<td>0.33</td>
<td>10.43</td>
<td>-4.71</td>
<td>-0.79</td>
<td>-0.17</td>
</tr>
<tr>
<td>Algebra</td>
<td>0.23</td>
<td>0.28</td>
<td>26.95</td>
<td>-7.05</td>
<td>-4.70</td>
</tr>
<tr>
<td>Analysis</td>
<td>-0.00</td>
<td>0.08</td>
<td>0.43</td>
<td>9.88</td>
<td>-2.02</td>
</tr>
<tr>
<td>Statistics</td>
<td>0.02</td>
<td>0.02</td>
<td>0.36</td>
<td>0.25</td>
<td>6.45</td>
</tr>
</tbody>
</table>
This analysis is from Whittaker (1990).
We have An, Stats $\perp \perp$ Mech,Vec $|\text{Alg.}$.
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: