Multivariate Gaussian Analysis

Steffen Lauritzen, University of Oxford

BS2 Statistical Inference, Lecture 7, Hilary Term 2009

February 13, 2009
For a positive definite covariance matrix $\Sigma$, the multivariate Gaussian distribution has density on $\mathcal{R}^d$

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

where $K = \Sigma^{-1}$ is the concentration matrix of the distribution.

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

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

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

$$Y = AX + b \sim \mathcal{N}_r(A\xi + b, A\Sigma A^\top).$$
Partition $X$ into $X_1$ and $X_2$, where $X_1 \in \mathbb{R}^r$ and $X_2 \in \mathbb{R}^s$ with $r + s = d$ and partition mean vector, concentration and covariance matrix accordingly.

*Then, if $X \sim \mathcal{N}_d(\xi, \Sigma)$

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

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

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

where

$$\xi_{1|2} = \xi_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \xi_2) \quad \text{and} \quad \Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

In particular, *if $\Sigma_{12} = 0$ if and only if $X_1$ and $X_2$ are independent.*
From the matrix identities

\[ K_{11}^{-1} = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} = \Sigma_{1|2} \] (2)

and

\[ K_{11}^{-1} K_{12} = -\Sigma_{12} \Sigma_{22}^{-1}, \] (3)

it follows 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 \( \Sigma \) where as the conditional concentration is simply expressed in terms of \( K \).
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 $O AO^\top = \text{diag}(\lambda_1, \ldots, \lambda_d)$ and using

$$\text{tr}(O AO^\top) = \text{tr}(AO^\top O) = \text{tr}(A).$$

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

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

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

$$\lambda^\top A \lambda = \text{tr}(\lambda^\top A \lambda) = \text{tr}(A \lambda \lambda^\top)$$

since $\lambda^\top A \lambda$ is a scalar.
Consider first 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 (1), we get the likelihood function

\[
L(K) = (2\pi)^{-nd/2}(\det K)^{n/2}e^{-\sum_{\nu=1}^n x_\nu^\top Kx_\nu / 2}
\]

\[
\propto (\det K)^{n/2}e^{-\sum_{\nu=1}^n \text{tr}\{Kx_\nu x_\nu^\top\} / 2}
\]

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

\[
= (\det K)^{n/2}e^{-\text{tr}(Kw) / 2}.
\]

where

\[
W = \sum_{\nu=1}^n X_\nu X_\nu^\top = X^\top X,
\]

is the matrix of sums of squares and products. Here we have let \( X \) be the \( n \times d \) matrix with rows equal to \( X_\nu^\top \).
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

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

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

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

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 \( \mathbb{E}(W) = n\Sigma \). 

Steffen Lauritzen, University of Oxford

Multivariate Gaussian Analysis
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

$$c(d, n) = 2^{nd/2}(2\pi)^{d(d-1)/4} \prod_{i=1}^{d} \Gamma\{(n + 1 - i)/2\}.$$
Let $X_1, \ldots, X_n$ be independent and identically distributed as $\mathcal{N}_d(\xi, \Sigma)$. Let $X$ be the $n \times d$ matrix with rows equal to $X_i^\top$ and assume that $\Pi_1, \ldots, \Pi_k$ are $n \times n$ matrices for orthogonal projections onto subspaces $L_1, \ldots, L_k$ of $\mathbb{R}^n$, that is,

$$\Pi_u \Pi_v = \delta_{uv} \Pi_u \quad \text{and} \quad \Pi_u^\top = \Pi_u.$$ 

Then, if $\Pi_i \xi = 0$ we have

$$W_u = X^\top \Pi_u X \sim \mathcal{W}_d(f_i, \Sigma),$$

where $f_u = \dim L_u = \text{rank} \Pi_u = \text{tr} \Pi_u$. Further, $W_1, \ldots, W_k$ are independent.
Let $W \sim \mathcal{W}_d(n, \Sigma)$ with $\Sigma$ regular and $n > d$. Then $W_{22}$ is regular with probability one and

(i) $W_{1|2}$ is independent of $(W_{12}, W_{22})$;
Let $W \sim \mathcal{W}_d(n, \Sigma)$ with $\Sigma$ regular and $n > d$. Then $W_{22}$ is regular with probability one and

(i) $W_{1|2}$ is independent of $(W_{12}, W_{22})$;
(ii) $W_{1|2} \sim \mathcal{W}_r(n - s, \Sigma_{1|2})$;
Let \( W \sim \mathcal{W}_d(n, \Sigma) \) with \( \Sigma \) regular and \( n > d \). Then \( W_{22} \) is regular with probability one and

(i) \( W_{1|2} \) is independent of \((W_{12}, W_{22})\);

(ii) \( W_{1|2} \sim \mathcal{W}_r(n - s, \Sigma_{1|2}) \);

(iii) \( W_{22} \sim \mathcal{W}_s(n, \Sigma_{22}) \);
Let $W \sim \mathcal{W}_d(n, \Sigma)$ with $\Sigma$ regular and $n > d$. Then $W_{22}$ is regular with probability one and

(i) $W_{1|2}$ is independent of $(W_{12}, W_{22})$;
(ii) $W_{1|2} \sim \mathcal{W}_r(n - s, \Sigma_{1|2})$;
(iii) $W_{22} \sim \mathcal{W}_s(n, \Sigma_{22})$;
(iv) The conditional distribution of $W_{12}$ given $W_{22} = w_{22}$ is multivariate Gaussian $\mathcal{N}_{r \times s}(\Sigma_{12} \Sigma^{-1}_{22} w_{22}, \Lambda)$ where

$$\Lambda_{ij,kl} = \text{Cov}(W_{ij}, W_{kl} \mid W_{22} = w_{22}) = w_{jl} \sigma_{ik}^{1|2} w_{jl}.$$
In the special case with $\Sigma_{12} = 0$ this can be simplified to $W_{1|2} \sim \mathcal{W}_r(n - s, \Sigma_{11})$ and

$$W_{12} \mid W_{22} = w_{22} \sim \mathcal{N}_{r \times s}(0, \Lambda)$$

with $\Lambda_{ij,kl} = \sigma_{ik}w_{j,l}$.

It follows that in this case, i.e. when $\Sigma_{12} = 0$, it holds that

$$W_{12} W_{22}^{-1} W_{21} \sim \mathcal{W}_r(s, \Sigma_{11}),$$