The Wishart distribution
Gaussian graphical models
Decomposable Gaussian graphical models
Linear structural equation systems

Decomposable Graphical Gaussian Models

Steffen Lauritzen, University of Oxford

Graphical Models and Inference, Lecture 9, Michaelmas Term 2010

November 8, 2010

The Wishart distribution
Gaussian graphical models
Decomposable Gaussian graphical models
Linear structural equation systems

Definition
Basic properties
Wishart density

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 *d-dimensional Wishart distribution* with parameter Σ and *n degrees of freedom* if

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

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 χ^2:

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

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

The Wishart distribution
Gaussian graphical models
Decomposable Gaussian graphical models
Linear structural equation systems

Definition
Basic properties
Wishart density

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 $\lambda \in R^d$,

$$\lambda^\top W \lambda \sim \sigma_\lambda^2 \chi^2(n),$$

where $\sigma_\lambda^2 = \lambda^\top \Sigma \lambda$.

The Wishart distribution
Gaussian graphical models
Decomposable Gaussian graphical models
Linear structural equation systems

Definition
Basic properties
Wishart density

If $W \sim \mathcal{W}_d(n, \Sigma)$, where Σ 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

$$
\begin{aligned}
&f_d(w \mid n, \Sigma) \\
&= c(d, n)^{-1}(\det \Sigma)^{-n/2}(\det w)^{(n-d-1)/2} e^{-\operatorname{tr}(\Sigma^{-1} w)/2}
\end{aligned}
$$

for w positive definite, and 0 otherwise.

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

$$
c(d, n) = 2^{nd/2}(2\pi)^{d(d-1)/4} \prod_{i=1}^{d} \Gamma\{(n+1-i)/2\}.
$$

The Wishart distribution
Gaussian graphical models
Decomposable Gaussian graphical models
Linear structural equation systems

Definition and likelihood function
Iterative Proportional Scaling

Consider $X = (X_v, v \in V) \sim \mathcal{N}_V(0, \Sigma)$ with Σ regular and $K = \Sigma^{-1}$.

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

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

implying

$$Cov(X_\alpha, X_\beta | X_{V \setminus \{\alpha, \beta\}}) = (K^{-1})_{\alpha\beta} = -k_{\alpha\beta}/(k_{\alpha\alpha}k_{\beta\beta} - k_{\alpha\beta}^2).$$

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*

$$\alpha \not\sim \beta \iff k_{\alpha\beta} = 0.$$

The Wishart distribution
Gaussian graphical models
Decomposable Gaussian graphical models
Linear structural equation systems

Definition and likelihood function
Iterative Proportional Scaling

$\mathcal{S}(\mathcal{G})$ denotes the symmetric matrices A with $a_{\alpha\beta} = 0$ unless $\alpha \sim \beta$ and $\mathcal{S}^+(\mathcal{G})$ their positive definite elements.

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

The likelihood function based on a sample of size n is

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

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

The Wishart distribution
Gaussian graphical models
Decomposable Gaussian graphical models
Linear structural equation systems

Definition and likelihood function
Iterative Proportional Scaling

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

$$a_{ij}^u = \left\{ \begin{array}{ll} 1 & \text{if } u \in V \text{ and } i = j = u \\ 1 & \text{if } u \in E \text{ and } u = \{i, j\} \\ 0 & \text{otherwise.} \end{array} \right).$$

Then, as $K \in \mathcal{S}(\mathcal{G})$,

$$K = \sum_{v \in V} k_v A^v + \sum_{e \in E} k_e A^e \qquad (1)$$

and hence

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

The Wishart distribution
Gaussian graphical models
Decomposable Gaussian graphical models
Linear structural equation systems

Definition and likelihood function
Iterative Proportional Scaling

Hence we can identify the family as a (regular and canonical) *exponential family with* $-\operatorname{tr}(A^u W)/2, u \in V \cup E$ *as canonical sufficient statistics.*

This yields the likelihood equations

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

which can also be expressed as

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

or, equivalently

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

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

The Wishart distribution
Gaussian graphical models
Decomposable Gaussian graphical models
Linear structural equation systems

Definition and likelihood function
Iterative Proportional Scaling

For $K \in \mathcal{S}^+(\mathcal{G})$ and $c \in \mathcal{C}$, define the operation of 'adjusting the c-marginal' as follows. Let $a = V \setminus c$ and

$$T_c K = \begin{pmatrix} n(w_{cc})^{-1} + K_{ca}(K_{aa})^{-1}K_{ac} & K_{ca} \\ K_{ac} & K_{aa} \end{pmatrix}. \qquad (2)$$

The C-marginal covariance $\tilde{\Sigma}_{cc}$ corresponding to the adjusted concentration matrix becomes

$$
\begin{aligned}
\tilde{\Sigma}_{cc} &= \{(T_c K)^{-1}\}_{cc} \\
&= \left\{ n(w_{cc})^{-1} + K_{ca}(K_{aa})^{-1}K_{ac} - K_{ca}(K_{aa})^{-1}K_{ac} \right\}^{-1} \\
&= w_{cc}/n,
\end{aligned}
$$

hence $T_c K$ *does indeed adjust the marginals*. From (2) 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.

The Wishart distribution
Gaussian graphical models
Decomposable Gaussian graphical models
Linear structural equation systems

Definition and likelihood function
Iterative Proportional Scaling

Next we choose any ordering $(c_1, \ldots, c_k)$ of the cliques in $\mathcal{G}$.
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 $\mathcal{G}$. Then*

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

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

The Wishart distribution
Gaussian graphical models
Decomposable Gaussian graphical models
Linear structural equation systems

Basic factorizations
Maximum likelihood estimates
An example

If the graph $\mathcal{G}$ is chordal, we say that the graphical model is
decomposable.

In this case, *the IPS-algorithm converges in a finite number of
steps*, as in the discrete case.

We also have the familiar *factorization of densities*

$$f(x \mid \Sigma) = \frac{\prod_{C \in \mathcal{C}} f(x_C \mid \Sigma_C)}{\prod_{S \in \mathcal{S}} f(x_S \mid \Sigma_S)^{\nu(S)}} \tag{3}$$

where $\nu(S)$ is the number of times S appear as intersection
between neighbouring cliques of a junction tree for $\mathcal{C}$.

The Wishart distribution
Gaussian graphical models
Decomposable Gaussian graphical models
Linear structural equation systems

Basic factorizations
Maximum likelihood estimates
An example

Relations for trace and determinant

Using the factorization (3) we can for example match the
expressions for the trace and determinant of Σ

$$\text{tr}(KW) = \sum_{C \in \mathcal{C}} \text{tr}(K_C W_C) - \sum_{S \in \mathcal{S}} \nu(S) \, \text{tr}(K_S W_S)$$

and further

$$\det \Sigma = \{\det(K)\}^{-1} = \frac{\prod_{C \in \mathcal{C}} \det\{\Sigma_C\}}{\prod_{S \in \mathcal{S}} \{\det(\Sigma_S)\}^{\nu(S)}}$$

These are some of many relations that can be derived using the
decomposition property of chordal graphs.

The Wishart distribution
Gaussian graphical models
Decomposable Gaussian graphical models
Linear structural equation systems

Basic factorizations
Maximum likelihood estimates
An example

The same factorization clearly holds for the maximum likelihood estimates:

$$f(x \mid \hat{\Sigma}) = \frac{\prod_{C \in \mathcal{C}} f(x_C \mid \hat{\Sigma}_C)}{\prod_{S \in \mathcal{S}} f(x_S \mid \hat{\Sigma}_S)^{\nu(S)}} \tag{4}$$

Moreover, it follows from the general likelihood equations that

$$\hat{\Sigma}_A = W_A/n \text{ whenever } A \text{ is complete.}$$

Exploiting this, we can obtain an explicit formula for the maximum likelihood estimate in the case of a chordal graph.
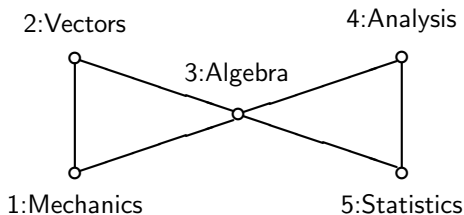
The Wishart distribution
Gaussian graphical models
Decomposable Gaussian graphical models
Linear structural equation systems

Basic factorizations
Maximum likelihood estimates
An example

For a $|d| \times |e|$ matrix $A = \{a_{\gamma\mu}\}_{\gamma \in d, \mu \in e}$ we let $[A]^V$ denote the matrix obtained from A by filling up with zero entries to obtain full dimension $|V| \times |V|$, i.e.

$$\left([A]^V\right)_{\gamma\mu} = \begin{cases} a_{\gamma\mu} & \text{if } \gamma \in d, \mu \in e \\ 0 & \text{otherwise}. \end{cases}$$

The maximum likelihood estimates exists if and only if $n \geq C$ for all $C \in \mathcal{C}$. Then the following simple formula holds for the maximum likelihood estimate of K:

$$\hat{K} = n \left\{ \sum_{C \in \mathcal{C}} \left[(w_C)^{-1} \right]^V - \sum_{S \in \mathcal{S}} \nu(S) \left[(w_S)^{-1} \right]^V \right\}.$$

The Wishart distribution
Gaussian graphical models
Decomposable Gaussian graphical models
Linear structural equation systems

Basic factorizations
Maximum likelihood estimates
An example

Mathematics marks

This graph is chordal with cliques $\{1, 2, 3\}$, $\{3, 4, 5\}$ with separator $S = \{3\}$ having $\nu(\{3\}) = 1$.

The Wishart distribution
Gaussian graphical models
Decomposable Gaussian graphical models
Linear structural equation systems

Basic factorizations
Maximum likelihood estimates
An example

Since one degree of freedom is lost by subtracting the average, we get in this example

$$\hat{K} = 87 \begin{pmatrix} w^{11}_{[123]} & w^{12}_{[123]} & w^{13}_{[123]} & 0 & 0 \\ w^{21}_{[123]} & w^{22}_{[123]} & w^{23}_{[123]} & 0 & 0 \\ w^{31}_{[123]} & w^{32}_{[123]} & w^{33}_{[123]} + w^{33}_{[345]} - 1/w_{33} & w^{34}_{[345]} & w^{35}_{[345]} \\ 0 & 0 & w^{43}_{[345]} & w^{44}_{[345]} & w^{45}_{[345]} \\ 0 & 0 & w^{53}_{[345]} & w^{54}_{[345]} & w^{55}_{[345]} \end{pmatrix}$$

where $w^{ij}_{[123]}$ is the ijth element of the inverse of

$$W_{[123]} = \begin{pmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \end{pmatrix}$$

and so on.

The Wishart distribution
Gaussian graphical models
Decomposable Gaussian graphical models
Linear structural equation systems

Definition
A simple example
More general systems

Consider a *directed acyclic graph* $\mathcal{D}$ and associate for every vertex a random variable X_v. Consider now the equation system

$$X_v \leftarrow \alpha_v^\top X_{\mathsf{pa}(v)} + \beta_v + U_v, v \in V \tag{5}$$

where $U_v, v \in V$ are independent random disturbances with $U_v \sim \mathcal{N}(0, \sigma_v^2)$.

Such an equation system is known as a *recursive structural equation system*.

Structural equation systems are used heavily in social sciences and in economics. The term *structural* refers to the fact that the equations are assumed to be *stable under intervention* so that fixing a value of x_v^* would change the system only by removing the line in the equation system (5) defining x_v^*.

The Wishart distribution
Gaussian graphical models
Decomposable Gaussian graphical models
Linear structural equation systems

Definition
A simple example
More general systems

A recursive structural equation system defines a multivariate Gaussian distribution with joint density

$$
\begin{aligned}
f(x \mid \alpha, \sigma) &= \prod_v (2\pi)^{-1/2} \sigma_v^{-1} e^{-\frac{(x_v - \alpha_v^\top x_{\mathrm{pa}(v)} - \beta_v)^2}{2\sigma_v^2}} \\
&= (2\pi)^{-|V|/2} \left(\prod_v \sigma_v^{-1} \right) \\
&\quad \times e^{-\sum_v \frac{(x_v - \alpha_v^\top x_{\mathrm{pa}(v)} - \beta_v)^2}{2\sigma_v^2}},
\end{aligned}
$$

from which the joint concentration matrix K can easily be derived.

The Wishart distribution
Gaussian graphical models
Decomposable Gaussian graphical models
Linear structural equation systems

Definition
A simple example
More general systems

Consider the system

$$
\begin{aligned}
X_1 &\leftarrow U_1 \\
X_2 &\leftarrow U_2 \\
X_3 &\leftarrow \alpha_{31} X_1 + U_3 \\
X_4 &\leftarrow \alpha_{42} X_2 + \alpha_{43} X_3 + U_4.
\end{aligned}
$$

The quadratic expression in the exponent becomes

$$
\frac{x_1^2}{\sigma_1^2} + \frac{x_2^2}{\sigma_2^2} + \frac{(x_3 - \alpha_{31} x_1)^2}{\sigma_3^2} + \frac{(x_4 - \alpha_{42} x_2 - \alpha_{43} x_3)^2}{\sigma_4^2}.
$$

The Wishart distribution
Gaussian graphical models
Decomposable Gaussian graphical models
Linear structural equation systems

Definition
A simple example
More general systems

Expanding the squares and identifying terms yields the concentration matrix

$$K = \begin{pmatrix} \frac{1}{\sigma_1^2} + \frac{\alpha_{31}^2}{\sigma_3^2} & 0 & \frac{-\alpha_{31}}{\sigma_3^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} + \frac{\alpha_{42}^2}{\sigma_4^2} & \frac{\alpha_{42}\alpha_{43}}{\sigma_4^2} & \frac{-\alpha_{42}}{\sigma_4^2} \\ \frac{-\alpha_{31}}{\sigma_3^2} & \frac{\alpha_{42}\alpha_{43}}{\sigma_4^2} & \frac{1}{\sigma_3^2} + \frac{\alpha_{43}^2}{\sigma_4^2} & \frac{-\alpha_{43}}{\sigma_4^2} \\ 0 & \frac{-\alpha_{42}}{\sigma_4^2} & \frac{-\alpha_{43}}{\sigma_4^2} & \frac{1}{\sigma_4^2} \end{pmatrix}.$$

The Wishart distribution
Gaussian graphical models
Decomposable Gaussian graphical models
Linear structural equation systems

Definition
A simple example
More general systems

The covariance matrix can in principle be found by inverting the above. However, it is easier to express X in terms of the Us as

$$
\begin{aligned}
X_1 &= U_1 \\
X_2 &= U_2 \\
X_3 &= \alpha_{31} U_1 + U_3 \\
X_4 &= \alpha_{43}\alpha_{31} U_1 + \alpha_{42} U_2 + \alpha_{43} U_3 + U_4
\end{aligned}
$$

and then calculate the covariances directly to obtain

$$
\begin{pmatrix}
\sigma_1^2 & 0 & \alpha_{31}\sigma_1^2 & \alpha_{43}\alpha_{31}\sigma_1^2 \\
0 & \sigma_2^2 & 0 & \alpha_{42}\sigma_2^2 \\
\alpha_{31}\sigma_1^2 & 0 & \sigma_3^2 + \alpha_{31}^2\sigma_1^2 & \alpha_{43}\alpha_{31}^2\sigma_1^2 + \alpha_{43}\sigma_3^2 \\
\alpha_{43}\alpha_{31}\sigma_1^2 & \alpha_{42}\sigma_2^2 & \alpha_{43}\alpha_{31}^2\sigma_1^2 + \alpha_{43}\sigma_3^2 & \omega_4^2
\end{pmatrix},
$$

where

$$
\omega_4^2 = \alpha_{43}^2\alpha_{31}^2\sigma_1^2 + \alpha_{42}^2\sigma_2^2 + \alpha_{43}^2\sigma_3^2 + \sigma_4^2.
$$

The Wishart distribution
Gaussian graphical models
Decomposable Gaussian graphical models
Linear structural equation systems

Definition
A simple example
More general systems

Systems of structural equations of the type considered are called *recursive* in contrast to *feedback systems* of equations where directed cycles in the corresponding graph are allowed.

It is sometimes customary to allow correlations between the disturbance terms which makes conditional independence relations more complex.

This type of model and analysis goes back to the geneticist Sewall Wright who coined the term *path analysis* to the calculus of effects based on this kind of models.

This is one of the early precursors for modern graphical modelling.

Markov properties for *directed* graphs are necessary to understand these models.