Chapter 5
Theory of the Multinormal

In the preceding chapter we saw how the multivariate normal distribution comes into play in many applications. It is useful to know more about this distribution, since it is often a good approximate distribution in many situations. Another reason for considering the multinormal distribution relies on the fact that it has many appealing properties: it is stable under linear transforms, zero correlation corresponds to independence, the marginals and all the conditionals are also multivariate normal variates, etc. The mathematical properties of the multinormal make analyses much simpler.

In this chapter we will first concentrate on the probabilistic properties of the multinormal, then we will introduce two "companion" distributions of the multinormal which naturally appear when sampling from a multivariate normal population: the Wishart and the Hotelling distributions. The latter is particularly important for most of the testing procedures proposed in Chap. 7.

5.1 Elementary Properties of the Multinormal

Let us first summarise some properties which were already derived in the previous chapter.

- The pdf of $X \sim N_p(\mu, \Sigma)$ is

$$f(x) = |2\pi\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(x-\mu)^{\top}\Sigma^{-1}(x-\mu)\right\}. \tag{5.1}$$

© Springer-Verlag Berlin Heidelberg 2015

W.K. Härdle, L. Simar, *Applied Multivariate Statistical Analysis*,
DOI 10.1007/978-3-662-45171-7_5

The expectation is $\mathsf{E}(X) = \mu$, the covariance can be calculated as $\mathsf{Var}(X) = \mathsf{E}(X-\mu)(X-\mu)^\top = \Sigma$.

- Linear transformations turn normal random variables into normal random variables. If $X \sim N_p(\mu, \Sigma)$ and $\mathcal{A}(p \times p), c \in \mathbb{R}^p$, then $Y = \mathcal{A}X + c$ is p-variate Normal, i.e.

$$Y \sim N_p(\mathcal{A}\mu + c, \mathcal{A}\Sigma\mathcal{A}^\top). \tag{5.2}$$

- If $X \sim N_p(\mu, \Sigma)$, then the Mahalanobis transformation is

$$Y = \Sigma^{-1/2}(X-\mu) \sim N_p(0, \mathcal{I}_p) \tag{5.3}$$

and it holds that

$$Y^\top Y = (X-\mu)^\top \Sigma^{-1}(X-\mu) \sim \chi^2_p. \tag{5.4}$$

Often it is interesting to partition X into sub-vectors X_1 and X_2. The following theorem tells us how to correct X_2 to obtain a vector which is independent of X_1.

Theorem 5.1 *Let* $X = \binom{X_1}{X_2} \sim N_p(\mu, \Sigma)$, $X_1 \in \mathbb{R}^r$, $X_2 \in \mathbb{R}^{p-r}$. *Define* $X_{2.1} = X_2 - \Sigma_{21}\Sigma_{11}^{-1}X_1$ *from the partitioned covariance matrix*

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Then

$$X_1 \sim N_r(\mu_1, \Sigma_{11}), \tag{5.5}$$

$$X_{2.1} \sim N_{p-r}(\mu_{2.1}, \Sigma_{22.1}) \tag{5.6}$$

are independent with

$$\mu_{2.1} = \mu_2 - \Sigma_{21}\Sigma_{11}^{-1}\mu_1, \quad \Sigma_{22.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}. \tag{5.7}$$

Proof

$$X_1 = \mathcal{A}X \quad \text{with} \quad \mathcal{A} = (\, \mathcal{I}_r \,,\, 0 \,)$$

$$X_{2.1} = \mathcal{B}X \quad \text{with} \quad \mathcal{B} = (\, -\Sigma_{21}\Sigma_{11}^{-1} \,,\, \mathcal{I}_{p-r} \,).$$

Then, by (5.2) X_1 and $X_{2.1}$ are both normal. Note that

$$\mathsf{Cov}(X_1, X_{2.1}) = \mathcal{A}\Sigma\mathcal{B}^\top = \left(\boxed{\begin{matrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{matrix}} \quad \mathbf{0} \right) \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \begin{pmatrix} (-\Sigma_{21}\Sigma_{11}^{-1})^\top \\ \\ \boxed{\begin{matrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{matrix}} \end{pmatrix},$$

$$\mathcal{A}\Sigma = (\mathcal{I}_r\ 0) \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} = (\Sigma_{11} \quad \Sigma_{12}),$$

$$\begin{aligned} \text{hence, } \mathcal{A}\Sigma\mathcal{B}^\top &= (\Sigma_{11} \quad \Sigma_{12}) \begin{pmatrix} (-\Sigma_{21}\Sigma_{11}^{-1})^\top \\ \mathcal{I}_{p-r} \end{pmatrix} \\ &= \left(-\Sigma_{11}\left(\Sigma_{21}\Sigma_{11}^{-1}\right)^\top + \Sigma_{12}\right). \end{aligned}$$

Recall that $\Sigma_{21} = (\Sigma_{12})^\top$. Hence $\mathcal{A}\Sigma\mathcal{B}^\top = -\Sigma_{11}\Sigma_{11}^{-1}\Sigma_{12} + \Sigma_{12} \equiv 0$.

Using (5.2) again we also have the joint distribution of $(X_1, X_{2.1})$, namely

$$\begin{pmatrix} X_1 \\ X_{2.1} \end{pmatrix} = \begin{pmatrix} \mathcal{A} \\ \mathcal{B} \end{pmatrix} X \sim N_p\left(\begin{pmatrix} \mu_1 \\ \mu_{2.1} \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22.1} \end{pmatrix} \right).$$

With this block diagonal structure of the covariance matrix, the joint pdf of $(X_1, X_{2.1})$ can easily be factorised into

$$\begin{aligned} f(x_1, x_{2.1}) = {} & |2\pi\Sigma_{11}|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(x_1 - \mu_1)^\top \Sigma_{11}^{-1}(x_1 - \mu_1) \right\} \\ & \times |2\pi\Sigma_{22.1}|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(x_{2.1} - \mu_{2.1})^\top \Sigma_{22.1}^{-1}(x_{2.1} - \mu_{2.1}) \right\} \end{aligned}$$

from which the independence between X_1 and $X_{2.1}$ follows. □

The next two corollaries are direct consequences of Theorem 5.1.

Corollary 5.1 *Let* $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_p(\mu, \Sigma)$, $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$. $\Sigma_{12} = 0$ *if and only if* X_1 *is independent of* X_2.

The independence of two linear transforms of a multinormal X can be shown via the following corollary.

Corollary 5.2 *If* $X \sim N_p(\mu, \Sigma)$ *and given some matrices* $\mathcal{A}$ *and* $\mathcal{B}$, *then* $\mathcal{A}X$ *and* $\mathcal{B}X$ *are independent if and only if* $\mathcal{A}\Sigma\mathcal{B}^\top = 0$.

The following theorem is also useful. It generalises Theorem 4.6. The proof is left as an exercise.

Theorem 5.2 *If* $X \sim N_p(\mu, \Sigma)$, $\mathcal{A}(q \times p)$, $c \in \mathbb{R}^q$ *and* $q \leq p$, *then* $Y = \mathcal{A}X + c$ *is a* q*-variate Normal, i.e.*

$$Y \sim N_q(\mathcal{A}\mu + c, \mathcal{A}\Sigma\mathcal{A}^\top).$$

The conditional distribution of X_2 given X_1 is given by the next theorem.

Theorem 5.3 *The conditional distribution of* X_2 *given* $X_1 = x_1$ *is normal with mean* $\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1)$ *and covariance* $\Sigma_{22.1}$, *i.e.*

$$(X_2 \mid X_1 = x_1) \sim N_{p-r}(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1), \Sigma_{22.1}). \tag{5.8}$$

Proof Since $X_2 = X_{2.1} + \Sigma_{21}\Sigma_{11}^{-1}X_1$, for a fixed value of $X_1 = x_1$, X_2 is equivalent to $X_{2.1}$ plus a constant term:

$$(X_2|X_1 = x_1) = (X_{2.1} + \Sigma_{21}\Sigma_{11}^{-1}x_1),$$

which has the normal distribution $N(\mu_{2.1} + \Sigma_{21}\Sigma_{11}^{-1}x_1, \Sigma_{22.1})$. □

Note that the conditional mean of $(X_2 \mid X_1)$ is a linear function of X_1 and that the conditional variance does not depend on the particular value of X_1. In the following example we consider a specific distribution.

Example 5.1 Suppose that $p = 2$, $r = 1$, $\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and $\Sigma = \begin{pmatrix} 1 & -0.8 \\ -0.8 & 2 \end{pmatrix}$. Then $\Sigma_{11} = 1$, $\Sigma_{21} = -0.8$ and $\Sigma_{22.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} = 2 - (0.8)^2 = 1.36$. Hence the marginal pdf of X_1 is

$$f_{X_1}(x_1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x_1^2}{2}\right)$$

and the conditional pdf of $(X_2 \mid X_1 = x_1)$ is given by

$$f(x_2 \mid x_1) = \frac{1}{\sqrt{2\pi(1.36)}} \exp\left\{-\frac{(x_2 + 0.8x_1)^2}{2 \times (1.36)}\right\}.$$

As mentioned above, the conditional mean of $(X_2 \mid X_1)$ is linear in X_1. The shift in the density of $(X_2 \mid X_1)$ can be seen in Fig. 5.1.

Sometimes it will be useful to reconstruct a joint distribution from the marginal distribution of X_1 and the conditional distribution $(X_2|X_1)$. The following theorem shows under which conditions this can be easily done in the multinormal framework.

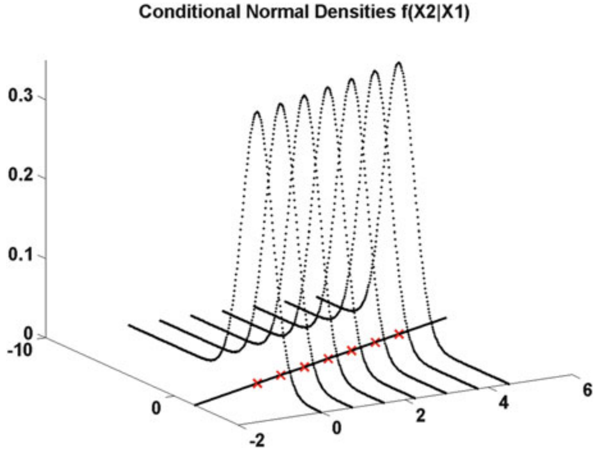


Fig. 5.1 Shifts in the conditional density `MVAcondnorm`

Theorem 5.4 *If* $X_1 \sim N_r(\mu_1, \Sigma_{11})$ *and* $(X_2|X_1 = x_1) \sim N_{p-r}(\mathcal{A}x_1 + b, \Omega)$ *where* Ω *does not depend on* x_1*, then* $X = \binom{X_1}{X_2} \sim N_p(\mu, \Sigma)$*, where*

$$\mu = \begin{pmatrix} \mu_1 \\ \mathcal{A}\mu_1 + b \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{11}\mathcal{A}^\top \\ \mathcal{A}\Sigma_{11} & \Omega + \mathcal{A}\Sigma_{11}\mathcal{A}^\top \end{pmatrix}.$$

Example 5.2 Consider the following random variables

$$X_1 \sim N_1(0, 1),$$

$$X_2|X_1 = x_1 \sim N_2\left(\begin{pmatrix} 2x_1 \\ x_1 + 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right).$$

Using Theorem (5.4), where $\mathcal{A} = (2 \quad 1)^\top$, $b = (0 \quad 1)^\top$ and $\Omega = \mathcal{I}_2$, we easily obtain the following result:

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_3\left(\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 1 \\ 2 & 5 & 2 \\ 1 & 2 & 2 \end{pmatrix}\right).$$

In particular, the marginal distribution of X_2 is

$$X_2 \sim N_2 \left(\begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 5 & 2 \\ 2 & 2 \end{pmatrix} \right),$$

thus conditional on X_1, the two components of X_2 are independent but marginally they are not.

Note that the marginal mean vector and covariance matrix of X_2 could have also been computed directly by using (4.28)–(4.29). Using the derivation above, however, provides us with useful properties: we have multinormality.

Conditional Approximations

As we saw in Chap. 4 (Theorem 4.3), the conditional expectation $\mathsf{E}(X_2|X_1)$ is the mean squared error (MSE) best approximation of X_2 by a function of X_1. We have in this case

$$X_2 = \mathsf{E}(X_2|X_1) + U = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(X_1 - \mu_1) + U. \tag{5.9}$$

Hence, the best approximation of $X_2 \in \mathbb{R}^{p-r}$ by $X_1 \in \mathbb{R}^r$ is the linear approximation that can be written as:

$$X_2 = \beta_0 + \mathcal{B}\, X_1 + U \tag{5.10}$$

with $\mathcal{B} = \Sigma_{21}\Sigma_{11}^{-1}$, $\beta_0 = \mu_2 - B\mu_1$ and $U \sim N(0, \Sigma_{22.1})$.

Consider now the particular case where $r = p - 1$. Now $X_2 \in \mathbb{R}$ and $\mathcal{B}$ is a row vector $\beta^\top$ of dimension $(1 \times r)$

$$X_2 = \beta_0 + \beta^\top X_1 + U. \tag{5.11}$$

This means, geometrically speaking, that the best MSE approximation of X_2 by a function of X_1 is a hyperplane. The marginal variance of X_2 can be decomposed via (5.11):

$$\sigma_{22} = \beta^\top \Sigma_{11}\beta + \sigma_{22.1} = \sigma_{21}\Sigma_{11}^{-1}\sigma_{12} + \sigma_{22.1}. \tag{5.12}$$

The ratio

$$\rho^2_{2.1\ldots r} = \frac{\sigma_{21}\Sigma_{11}^{-1}\sigma_{12}}{\sigma_{22}} \tag{5.13}$$

is known as the square of the multiple correlation between X_2 and the r variables X_1. It is the percentage of the variance of X_2 which is explained by the linear approximation $\beta_0 + \beta^\top X_1$. The last term in (5.12) is the residual variance of X_2. The square of the multiple correlation corresponds to the coefficient of determination introduced in Sect. 3.4, see (3.39), but here it is defined in terms of the r.v. X_1 and X_2. It can be shown that $\rho_{2.1\ldots r}$ is also the maximum correlation attainable between X_2 and a linear combination of the elements of X_1, the optimal linear combination being precisely given by $\beta^\top X_1$. Note that when $r = 1$, the multiple correlation $\rho_{2.1}$ coincides with the usual simple correlation $\rho_{X_2 X_1}$ between X_2 and X_1.

Example 5.3 Consider the "classic blue" pullover example (Example 3.15) and suppose that X_1 (sales), X_2 (price), X_3 (advertisement) and X_4 (sales assistants) are normally distributed with

$$\mu = \begin{pmatrix} 172.7 \\ 104.6 \\ 104.0 \\ 93.8 \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} 1037.21 & & & \\ -80.02 & 219.84 & & \\ 1430.70 & 92.10 & 2624.00 & \\ 271.44 & -91.58 & 210.30 & 177.36 \end{pmatrix}.$$

(These are in fact the sample mean and the sample covariance matrix but in this example we pretend that they are the true parameter values.)

The conditional distribution of X_1 given (X_2, X_3, X_4) is thus an univariate normal with mean

$$\mu_1 + \sigma_{12}\Sigma_{22}^{-1} \begin{pmatrix} X_2 - \mu_2 \\ X_3 - \mu_3 \\ X_4 - \mu_4 \end{pmatrix} = 65.670 - 0.216 X_2 + 0.485 X_3 + 0.844 X_4$$

and variance

$$\sigma_{11.2} = \sigma_{11} - \sigma_{12}\Sigma_{22}^{-1}\sigma_{21} = 96.761$$

The linear approximation of the sales (X_1) by the price (X_2), advertisement (X_3) and sales assistants (X_4) is provided by the conditional mean above. (Note that this coincides with the results of Example 3.15 due to the particular choice of μ and Σ.) The quality of the approximation is given by the multiple correlation $\rho^2_{1.234} = \frac{\sigma_{12}\Sigma_{22}^{-1}\sigma_{21}}{\sigma_{11}} = 0.907$. (Note again that this coincides with the coefficient of determination r^2 found in Example 3.15.)

This example also illustrates the concept of partial correlation. The correlation matrix between the four variables is given by

$$P = \begin{pmatrix} 1 & -0.168 & 0.867 & 0.633 \\ -0.168 & 1 & 0.121 & -0.464 \\ 0.867 & 0.121 & 1 & 0.308 \\ 0.633 & -0.464 & 0.308 & 1 \end{pmatrix},$$

so that the correlation between X_1 (sales) and X_2 (price) is -0.168. We can compute the conditional distribution of (X_1, X_2) given (X_3, X_4), which is a bivariate normal with mean:

$$\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} \sigma_{13} & \sigma_{14} \\ \sigma_{23} & \sigma_{24} \end{pmatrix} \begin{pmatrix} \sigma_{33} & \sigma_{34} \\ \sigma_{43} & \sigma_{44} \end{pmatrix}^{-1} \begin{pmatrix} X_3 - \mu_3 \\ X_4 - \mu_4 \end{pmatrix} = \begin{pmatrix} 32.516 + 0.467X_3 + 0.977X_4 \\ 153.644 + 0.085X_3 - 0.617X_4 \end{pmatrix}$$

and covariance matrix:

$$\begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} - \begin{pmatrix} \sigma_{13} & \sigma_{14} \\ \sigma_{23} & \sigma_{24} \end{pmatrix} \begin{pmatrix} \sigma_{33} & \sigma_{34} \\ \sigma_{43} & \sigma_{44} \end{pmatrix}^{-1} \begin{pmatrix} \sigma_{31} & \sigma_{32} \\ \sigma_{41} & \sigma_{42} \end{pmatrix} = \begin{pmatrix} 104.006 & \\ -33.574 & 155.592 \end{pmatrix}.$$

In particular, the last covariance matrix allows the partial correlation between X_1 and X_2 to be computed for a fixed level of X_3 and X_4:

$$\rho_{X_1 X_2 | X_3 X_4} = \frac{-33.574}{\sqrt{104.006 \cdot 155.592}} = -0.264,$$

so that in this particular example with a fixed level of advertisement and sales assistance, the negative correlation between price and sales is more important than the marginal one.
MVAbluepullover

Summary

↪	If $X \sim N_p(\mu, \Sigma)$, then a linear transformation $\mathcal{A}X + c$, $\mathcal{A}(q \times p)$, where $c \in \mathbb{R}^q$, has distribution $N_q(\mathcal{A}\mu + c, \mathcal{A}\Sigma\mathcal{A}^\top)$.
↪	Two linear transformations $\mathcal{A}X$ and $\mathcal{B}X$ with $X \sim N_p(\mu, \Sigma)$ are independent if and only if $\mathcal{A}\Sigma\mathcal{B}^\top = 0$.
↪	If X_1 and X_2 are partitions of $X \sim N_p(\mu, \Sigma)$, then the conditional distribution of X_2 given $X_1 = x_1$ is again normal.
↪	In the multivariate normal case, X_1 is independent of X_2 if and only if $\Sigma_{12} = 0$.
↪	The conditional expectation of $(X_2 \| X_1)$ is a linear function if $\binom{X_1}{X_2} \sim N_p(\mu, \Sigma)$.
↪	The multiple correlation coefficient is defined as $\rho^2_{2.1\ldots r} = \frac{\sigma_{21}\Sigma_{11}^{-1}\sigma_{12}}{\sigma_{22}}$.
↪	The multiple correlation coefficient is the percentage of the variance of X_2 explained by the linear approximation $\beta_0 + \beta^\top X_1$.

5.2 The Wishart Distribution

The Wishart distribution (named after its discoverer) plays a prominent role in the analysis of estimated covariance matrices. If the mean of $X \sim N_p(\mu, \Sigma)$ is known to be $\mu = 0$, then for a data matrix $\mathcal{X}(n \times p)$ the estimated covariance matrix is proportional to $\mathcal{X}^\top\mathcal{X}$. This is the point where the Wishart distribution comes in, because $\mathcal{M}(p \times p) = \mathcal{X}^\top\mathcal{X} = \sum_{i=1}^n x_i x_i^\top$ has a Wishart distribution $W_p(\Sigma, n)$.

Example 5.4 Set $p = 1$, then for $X \sim N_1(0, \sigma^2)$ the data matrix of the observations

$$\mathcal{X} = (x_1, \dots, x_n)^\top \quad \text{with} \quad \mathcal{M} = \mathcal{X}^\top\mathcal{X} = \sum_{i=1}^n x_i x_i$$

leads to the Wishart distribution $W_1(\sigma^2, n) = \sigma^2\chi_n^2$. The one-dimensional Wishart distribution is thus in fact a χ^2 distribution.

When we talk about the distribution of a matrix, we mean of course the joint distribution of all its elements. More exactly: since $\mathcal{M} = \mathcal{X}^\top\mathcal{X}$ is symmetric we only need to consider the elements of the lower triangular matrix

$$\mathcal{M} = \begin{pmatrix} m_{11} & & & \\ m_{21} & m_{22} & & \\ \vdots & \vdots & \ddots & \\ m_{p1} & m_{p2} & \dots & m_{pp} \end{pmatrix}. \tag{5.14}$$

Hence the Wishart distribution is defined by the distribution of the vector

$$(m_{11}, \dots, m_{p1}, m_{22}, \dots, m_{p2}, \dots, m_{pp})^\top. \tag{5.15}$$

Linear transformations of the data matrix $\mathcal{X}$ also lead to Wishart matrices.

Theorem 5.5 *If $\mathcal{M} \sim W_p(\Sigma, n)$ and $\mathcal{B}(p \times q)$, then the distribution of $\mathcal{B}^\top\mathcal{M}\mathcal{B}$ is Wishart $W_q(\mathcal{B}^\top\Sigma\mathcal{B}, n)$.*

With this theorem we can standardise Wishart matrices since with $\mathcal{B} = \Sigma^{-1/2}$ the distribution of $\Sigma^{-1/2}\mathcal{M}\Sigma^{-1/2}$ is $W_p(\mathcal{I}, n)$. Another connection to the χ^2-distribution is given by the following theorem.

Theorem 5.6 *If $\mathcal{M} \sim W_p(\Sigma, m)$, and $a \in \mathbb{R}^p$ with $a^\top\Sigma a \neq 0$, then the distribution of $\dfrac{a^\top\mathcal{M}a}{a^\top\Sigma a}$ is χ_m^2.*

This theorem is an immediate consequence of Theorem 5.5 if we apply the linear transformation $x \mapsto a^\top x$. Central to the analysis of covariance matrices is the next theorem.

Theorem 5.7 (Cochran) *Let $\mathcal{X}(n \times p)$ be a data matrix from a $N_p(0, \Sigma)$ distribution and let $\mathcal{C}(n \times n)$ be a symmetric matrix.*

(a) *$\mathcal{X}^\top \mathcal{C} \mathcal{X}$ has the distribution of weighted Wishart random variables, i.e.*

$$\mathcal{X}^\top \mathcal{C} \mathcal{X} = \sum_{i=1}^{n} \lambda_i W_p(\Sigma, 1),$$

where λ_i, $i = 1, \ldots, n$, are the eigenvalues of $\mathcal{C}$.

(b) *$\mathcal{X}^\top \mathcal{C} \mathcal{X}$ is Wishart if and only if $\mathcal{C}^2 = \mathcal{C}$. In this case*

$$\mathcal{X}^\top \mathcal{C} \mathcal{X} \sim W_p(\Sigma, r),$$

and $r = \operatorname{rank}(\mathcal{C}) = \operatorname{tr}(\mathcal{C})$.

(c) *$n\mathcal{S} = \mathcal{X}^\top \mathcal{H} \mathcal{X}$ is distributed as $W_p(\Sigma, n-1)$ (note that $\mathcal{S}$ is the sample covariance matrix).*

(d) *$\bar{x}$ and $\mathcal{S}$ are independent.*

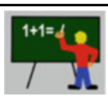
The following properties are useful:

1. If $\mathcal{M} \sim W_p(\Sigma, n)$, then $\mathsf{E}(\mathcal{M}) = n\Sigma$.
2. If $\mathcal{M}_i$ are independent Wishart $W_p(\Sigma, n_i)$ $i = 1, \ldots, k$, then $\mathcal{M} = \sum_{i=1}^{k} \mathcal{M}_i \sim W_p(\Sigma, n)$ where $n = \sum_{i=1}^{k} n_i$.
3. The density of $W_p(\Sigma, n-1)$ for a positive definite $\mathcal{M}$ is given by:

$$f_{\Sigma, n-1}(\mathcal{M}) = \frac{|\mathcal{M}|^{\frac{1}{2}(n-p-2)} e^{-\frac{1}{2}\operatorname{tr}(\mathcal{M}\Sigma^{-1})}}{2^{\frac{1}{2}p(n-1)} \pi^{\frac{1}{4}p(p-1)} |\Sigma|^{\frac{1}{2}(n-1)} \prod_{i=1}^{p} \Gamma\{\frac{n-i}{2}\}}, \tag{5.16}$$

where Γ is the gamma function: $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$.

For further details on the Wishart distribution, see Mardia, Kent, and Bibby (1979).



Summary

- ↪ The Wishart distribution is a generalisation of the χ^2-distribution. In particular $W_1(\sigma^2, n) = \sigma^2 \chi^2_n$.
- ↪ The empirical covariance matrix $\mathcal{S}$ has a $\frac{1}{n} W_p(\Sigma, n-1)$ distribution.
- ↪ In the normal case, $\bar{x}$ and $\mathcal{S}$ are independent.
- ↪ For $\mathcal{M} \sim W_p(\Sigma, m)$, $a^\top \mathcal{M} a / a^\top \Sigma a \sim \chi^2_m$.

5.3 Hotelling's T^2-Distribution

Suppose that $Y \in \mathbb{R}^p$ is a standard normal random vector, i.e. $Y \sim N_p(0, \mathcal{I})$, independent of the random matrix $\mathcal{M} \sim W_p(\mathcal{I}, n)$. What is the distribution of $Y^\top \mathcal{M}^{-1} Y$? The answer is provided by the Hotelling T^2-distribution: $n\, Y^\top \mathcal{M}^{-1} Y$ is Hotelling $T^2_{p,n}$ distributed.

The Hotelling T^2-distribution is a generalisation of the Student t-distribution. The general multinormal distribution $N(\mu, \Sigma)$ is considered in Theorem 5.8. The Hotelling T^2-distribution will play a central role in hypothesis testing in Chap. 7.

Theorem 5.8 *If $X \sim N_p(\mu, \Sigma)$ is independent of $\mathcal{M} \sim W_p(\Sigma, n)$, then*

$$n(X - \mu)^\top \mathcal{M}^{-1}(X - \mu) \sim T^2_{p,n}.$$

Corollary 5.3 *If $\overline{x}$ is the mean of a sample drawn from a normal population $N_p(\mu, \Sigma)$ and $\mathcal{S}$ is the sample covariance matrix, then*

$$(n-1)(\overline{x} - \mu)^\top \mathcal{S}^{-1}(\overline{x} - \mu) = n(\overline{x} - \mu)^\top \mathcal{S}_u^{-1}(\overline{x} - \mu) \sim T^2_{p,n-1}. \tag{5.17}$$

Recall that $\mathcal{S}_u = \frac{n}{n-1}\mathcal{S}$ is an unbiased estimator of the covariance matrix. A connection between the Hotelling T^2- and the F-distribution is given by the next theorem.

Theorem 5.9

$$T^2_{p,n} = \frac{np}{n-p+1}\, F_{p,n-p+1}.$$

Example 5.5 In the univariate case ($p = 1$), this theorem boils down to the well-known result:

$$\left(\frac{\bar{x} - \mu}{\sqrt{\mathcal{S}_u}/\sqrt{n}}\right)^2 \sim T^2_{1,n-1} = F_{1,n-1} = t^2_{n-1}$$

For further details on Hotelling T^2-distribution see Mardia et al. (1979). The next corollary follows immediately from (3.23), (3.24) and from Theorem 5.8. It will be useful for testing linear restrictions in multinormal populations.

Corollary 5.4 *Consider a linear transform of $X \sim N_p(\mu, \Sigma)$, $Y = \mathcal{A}X$ where $\mathcal{A}(q \times p)$ with $(q \le p)$. If $\overline{x}$ and $\mathcal{S}_X$ are the sample mean and the covariance matrix, we have*

$$\overline{y} = \mathcal{A}\overline{x} \sim N_q\left(\mathcal{A}\mu, \frac{1}{n}\mathcal{A}\Sigma\mathcal{A}^\top\right)$$

$$n\mathcal{S}_Y = n\mathcal{A}\mathcal{S}_X\mathcal{A}^\top \sim W_q(\mathcal{A}\Sigma\mathcal{A}^\top, n-1)$$

$$(n-1)(\mathcal{A}\overline{x} - \mathcal{A}\mu)^\top(\mathcal{A}\mathcal{S}_X\mathcal{A}^\top)^{-1}(\mathcal{A}\overline{x} - \mathcal{A}\mu) \sim T^2_{q,n-1}$$

The T^2 distribution is closely connected to the univariate t-statistic. In Example 5.4 we described the manner in which the Wishart distribution generalises the χ^2-distribution. We can write (5.17) as:

$$T^2 = \sqrt{n}(\overline{x} - \mu)^\top \left(\frac{\sum_{j=1}^n (x_j - \overline{x})(x_j - \overline{x})^\top}{n-1}\right)^{-1} \sqrt{n}(\overline{x} - \mu)$$

which is of the form

$$\begin{pmatrix}\text{multivariate normal}\\ \text{random vector}\end{pmatrix}^\top \left(\frac{\begin{matrix}\text{Wishart random}\\ \text{matrix}\end{matrix}}{\text{degrees of freedom}}\right)^{-1} \begin{pmatrix}\text{multivariate normal}\\ \text{random vector}\end{pmatrix}.$$

This is analogous to

$$t^2 = \sqrt{n}(\overline{x} - \mu)(s^2)^{-1}\sqrt{n}(\overline{x} - \mu)$$

or

$$\begin{pmatrix}\text{normal}\\ \text{random variable}\end{pmatrix} \left(\frac{\begin{matrix}\chi^2\text{-random}\\ \text{variable}\end{matrix}}{\text{degrees of freedom}}\right)^{-1} \begin{pmatrix}\text{normal}\\ \text{random variable}\end{pmatrix}$$

for the univariate case. Since the multivariate normal and Wishart random variables are independently distributed, their joint distribution is the product of the marginal normal and Wishart distributions. Using calculus, the distribution of T^2 as given above can be derived from this joint distribution.

Summary

↪ Hotelling's T^2-distribution is a generalisation of the t-distribution. In particular $T^2_{1,n} = t_n$.

↪ $(n-1)(\overline{x}-\mu)^{\top}\mathcal{S}^{-1}(\overline{x}-\mu)$ has a $T^2_{p,n-1}$ distribution.

↪ The relation between Hotelling's T^2- and Fisher's F-distribution is given by $T^2_{p,n} = \frac{np}{n-p+1}\ F_{p,n-p+1}$.

5.4 Spherical and Elliptical Distributions

The multinormal distribution belongs to the large family of elliptical distributions which has recently gained a lot of attention in financial mathematics. Elliptical distributions are often used, particularly in risk management.

Definition 5.1 A $(p \times 1)$ random vector Y is said to have a spherical distribution $S_p(\phi)$ if its characteristic function $\psi_Y(t)$ satisfies: $\psi_Y(t) = \phi(t^{\top}t)$ for some scalar function $\phi(.)$ which is then called the characteristic generator of the spherical distribution $S_p(\phi)$. We will write $Y \sim S_p(\phi)$.

This is only one of several possible ways to define spherical distributions. We can see spherical distributions as an extension of the standard multinormal distribution $N_p(0, \mathcal{I}_p)$.

Theorem 5.10 *Spherical random variables have the following properties:*

1. *All marginal distributions of a spherically distributed random vector are spherical.*
2. *All the marginal characteristic functions have the same generator.*
3. *Let $X \sim S_p(\phi)$, then X has the same distribution as $ru^{(p)}$ where $u^{(p)}$ is a random vector distributed uniformly on the unit sphere surface in $\mathbb{R}^p$ and $r \geq 0$ is a random variable independent of $u^{(p)}$. If $\mathsf{E}(r^2) < \infty$, then*

$$\mathsf{E}(X) = 0\,, \quad \mathsf{Cov}(X) = \frac{\mathsf{E}(r^2)}{p}\mathcal{I}_p.$$

The random radius r is related to the generator ϕ by a relation described in Fang, Kotz, and Ng (1990, p. 29). The moments of $X \sim S_p(\phi)$, provided that they exist, can be expressed in terms of one-dimensional integral.

A spherically distributed random vector does not, in general, necessarily possess a density. However, if it does, the marginal densities of dimension smaller than $p-1$ are continuous and the marginal densities of dimension smaller than $p-2$

are differentiable (except possibly at the origin in both cases). Univariate marginal densities for p greater than 2 are non-decreasing on $(-\infty, 0)$ and non-increasing on $(0, \infty)$.

Definition 5.2 A $(p \times 1)$ random vector X is said to have an elliptical distribution with parameters $\mu(p \times 1)$ and $\Sigma(p \times p)$ if X has the same distribution as $\mu + \mathcal{A}^\top Y$, where $Y \sim S_k(\phi)$ and $\mathcal{A}$ is a $(k \times p)$ matrix such that $\mathcal{A}^\top \mathcal{A} = \Sigma$ with $\operatorname{rank}(\Sigma) = k$. We shall write $X \sim EC_p(\mu, \Sigma, \phi)$.

Remark 5.1 The elliptical distribution can be seen as an extension of $N_p(\mu, \Sigma)$.

Example 5.6 The multivariate t-distribution. Let $Z \sim N_p(0, \mathcal{I}_p)$ and $s \sim \chi^2_m$ be independent. The random vector

$$Y = \sqrt{m}\, \frac{Z}{s}$$

has a multivariate t-distribution with m degrees of freedom. Moreover the t-distribution belongs to the family of p-dimensional spherical distributions.

Example 5.7 The multinormal distribution. Let $X \sim N_p(\mu, \Sigma)$. Then $X \sim EC_p(\mu, \Sigma, \phi)$ and $\phi(u) = \exp(-u/2)$. Figure 4.3 shows a density surface of the multivariate normal distribution: $f(x) = \det(2\pi\Sigma)^{-\frac{1}{2}} \exp\{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1} (x-\mu)\}$ with $\Sigma = \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}$ and $\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ Notice that the density is constant on ellipses. This is the reason for calling this family of distributions "elliptical".

Theorem 5.11 *Elliptical random vectors X have the following properties:*

1. *Any linear combination of elliptically distributed variables are elliptical.*
2. *Marginal distributions of elliptically distributed variables are elliptical.*
3. *A scalar function $\phi(.)$ can determine an elliptical distribution $EC_p(\mu, \Sigma, \phi)$ for every $\mu \in \mathbb{R}^p$ and $\Sigma \geq 0$ with* $\operatorname{rank}(\Sigma) = k$ iff *$\phi(t^\top t)$ is a p-dimensional characteristic function.*
4. *Assume that X is non-degenerate. If $X \sim EC_p(\mu, \Sigma, \phi)$ and $X \sim EC_p(\mu^*, \Sigma^*, \phi^*)$, then a constant $c > 0$ exists that*

$$\mu = \mu^*, \quad \Sigma = c\Sigma^*, \quad \phi^*(.) = \phi(c^{-1}.).$$

   In other words $\Sigma, \phi, \mathcal{A}$ are not unique, unless we impose the condition that $\det(\Sigma) = 1$.
5. *The characteristic function of X, $\psi(t) = \mathsf{E}(e^{\mathbf{i}t^\top X})$ is of the form*

$$\psi(t) = e^{\mathbf{i}t^\top \mu} \phi(t^\top \Sigma t)$$

   for a scalar function ϕ.

6. $X \sim EC_p(\mu, \Sigma, \phi)$ *with* $\text{rank}(\Sigma) = k$ iff X *has the same distribution as:*

$$\mu + r\mathcal{A}^\top u^{(k)} \tag{5.18}$$

where $r \geq 0$ *is independent of* $u^{(k)}$ *which is a random vector distributed uniformly on the unit sphere surface in* $\mathbb{R}^k$ *and* $\mathcal{A}$ *is a* $(k \times p)$ *matrix such that* $\mathcal{A}^\top \mathcal{A} = \Sigma$.

7. *Assume that* $X \sim EC_p(\mu, \Sigma, \phi)$ *and* $\mathsf{E}(r^2) < \infty$. *Then*

$$\mathsf{E}(X) = \mu \quad \mathsf{Cov}(X) = \frac{\mathsf{E}(r^2)}{\text{rank}(\Sigma)}\Sigma = -2\phi^\top(0)\Sigma.$$

8. *Assume that* $X \sim EC_p(\mu, \Sigma, \phi)$ *with* $\text{rank}(\Sigma) = k$. *Then*

$$Q(X) = (X - \mu)^\top \Sigma^{-1}(X - \mu)$$

has the same distribution as r^2 *in Eq. (5.18).*

5.5 Exercises

Exercise 5.1 *Consider* $X \sim N_2(\mu, \Sigma)$ *with* $\mu = (2, 2)^\top$ *and* $\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ *and the matrices* $\mathcal{A} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}^\top$, $\mathcal{B} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}^\top$. *Show that* $\mathcal{A}X$ *and* $\mathcal{B}X$ *are independent.*

Exercise 5.2 *Prove Theorem 5.4.*

Exercise 5.3 *Prove proposition (c) of Theorem 5.7.*

Exercise 5.4 *Let*

$$X \sim N_2\left(\begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}\right)$$

and

$$Y \mid X \sim N_2\left(\begin{pmatrix} X_1 \\ X_1 + X_2 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right).$$

(a) Determine the distribution of $Y_2 \mid Y_1$.
(b) Determine the distribution of $W = X - Y$.

Exercise 5.5 *Consider* $\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \sim N_3(\mu, \Sigma)$. *Compute* μ *and* Σ *knowing that*

$$\begin{aligned} Y \mid Z &\sim N_1(-Z, 1) \\ \mu_{Z|Y} &= -\frac{1}{3} - \frac{1}{3}Y \\ X \mid Y, Z &\sim N_1(2 + 2Y + 3Z, 1). \end{aligned}$$

Determine the distributions of $X \mid Y$ *and of* $X \mid Y + Z$.

Exercise 5.6 *Knowing that*

$$\begin{aligned} Z &\sim N_1(0, 1) \\ Y \mid Z &\sim N_1(1 + Z, 1) \\ X \mid Y, Z &\sim N_1(1 - Y, 1) \end{aligned}$$

(a) find the distribution of $\begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$ *and of* $Y \mid X, Z$.

(b) find the distribution of

$$\begin{pmatrix} U \\ V \end{pmatrix} = \begin{pmatrix} 1 + Z \\ 1 - Y \end{pmatrix}.$$

(c) compute $\mathsf{E}(Y \mid U = 2)$.

Exercise 5.7 *Suppose* $\begin{pmatrix} X \\ Y \end{pmatrix} \sim N_2(\mu, \Sigma)$ *with* Σ *positive definite. Is it possible that*

(a) $\mu_{X|Y} = 3Y^2$,
(b) $\sigma_{XX|Y} = 2 + Y^2$,
(c) $\mu_{X|Y} = 3 - Y$, *and*
(d) $\sigma_{XX|Y} = 5$?

Exercise 5.8 *Let* $X \sim N_3\left(\begin{pmatrix}1\\2\\3\end{pmatrix}, \begin{pmatrix}11 & -6 & 2\\ -6 & 10 & -4\\ 2 & -4 & 6\end{pmatrix}\right)$.

(a) Find the best linear approximation of X_3 *by a linear function of* X_1 *and* X_2 *and compute the multiple correlation between* X_3 *and* (X_1, X_2).

(b) Let $Z_1 = X_2 - X_3,\ Z_2 = X_2 + X_3$ *and* $(Z_3 \mid Z_1, Z_2) \sim N_1(Z_1 + Z_2, 10)$. *Compute the distribution of* $\begin{pmatrix}Z_1\\Z_2\\Z_3\end{pmatrix}$.

Exercise 5.9 *Let* $(X, Y, Z)^\top$ *be a trivariate normal r.v. with*

$$\begin{aligned} Y \mid Z &\sim N_1(2Z, 24)\\ Z \mid X &\sim N_1(2X+3, 14)\\ X &\sim N_1(1, 4)\\ \text{and } \rho_{XY} &= 0.5. \end{aligned}$$

Find the distribution of $(X, Y, Z)^\top$ *and compute the partial correlation between* X *and* Y *for fixed* Z. *Do you think it is reasonable to approximate* X *by a linear function of* Y *and* Z*?*

Exercise 5.10 *Let* $X \sim N_4\left(\begin{pmatrix}1\\2\\3\\4\end{pmatrix}, \begin{pmatrix}4 & 1 & 2 & 4\\ 1 & 4 & 2 & 1\\ 2 & 2 & 16 & 1\\ 4 & 1 & 1 & 9\end{pmatrix}\right)$.

(a) Give the best linear approximation of X_2 *as a function of* (X_1, X_4) *and evaluate the quality of the approximation.*

(b) Give the best linear approximation of X_2 *as a function of* (X_1, X_3, X_4) *and compare your answer with part (a).*

Exercise 5.11 *Prove Theorem 5.2.*
(Hint: complete the linear transformation $Z = \begin{pmatrix}\mathcal{A}\\ \mathcal{I}_{p-q}\end{pmatrix} X + \begin{pmatrix}c\\ 0_{p-q}\end{pmatrix}$ *and then use Theorem 5.1 to get the marginal of the first* q *components of* Z.*)*

Exercise 5.12 *Prove Corollaries 5.1 and 5.2.*