Chapter 2

Multivariate Statistics

The univariate normal distribution takes a central position in statistics. This situation is even more apparent in multivariate analysis, because the multivariate normal distribution is one of the few analytically tractable multivariate distributions (Eaton, 1983). Absolutely continuous multivariate distributions on $\mathbb{R}^m$ can be defined by specifying the density function of the distribution with respect to the Lebesgue measure on $\mathbb{R}^m$. The probability mass of such distributions cannot be concentrated in a null set. However, if the random vector takes values in, for example, a proper affine subspace, then this approach needs some modification. We will call these distributions singular distributions. Degenerated distributions give all their mass to just one point. Therefore, we call such distributions singular distributions. To incorporate singular distributions, it is convenient to define a multivariate distribution by specifying the distribution of each linear function of the random vector (see, e.g., Muirhead, 1982, theorem 1.2.2). We briefly repeat this discussion in subsection 2.1.1. It is often stated, however, that the density function of a normally distributed $m \times 1$ vector $X$ with singular covariance matrix ‘does not exist’. Whereas, indeed, the density does not exist with respect to the Lebesgue measure on $\mathbb{R}^m$, the density function does exist on an affine subspace, a topic we will discuss in subsection 2.1.2. To be able to work with the density function of singular normally distributed random vectors in practice, the problem is often redefined on this affine subspace (see, e.g., definition 3.2.2 of Tong, 1990). One of the contributions of this chapter is that we stay close to the original problem and do not have to consider transformations to find the probability density function.

---

Section 2.5 of this chapter is a revised version of Steerneman and Van Perlo-Ten Kleij (forthcoming, b).
In section 2.3 we study the Wishart distribution, the multivariate analogue of the chi-square distribution. The Wishart distribution occurs in a natural way if one considers the probability distribution of $X'X$, where $X$ is an $n \times k$ random matrix whose elements are normally distributed. Therefore, multivariate normally distributed random matrices are the topic of section 2.2. In this chapter we also consider singular Wishart distributions. Section 2.3 gives a very short introduction to the Wishart distribution; yet its main focus will be on a very famous theorem on partitioned Wishart distributed matrices. We show that this theorem also holds for the singular case. Although Srivastava and Khatri (1979) remarked this in their paper, it has not been proven completely to our knowledge. We show what happens if we allow singularity and replace the usual inverse of a matrix by the Moore-Penrose inverse. Subsequently, as a second extension of the theorem which only applies to the central Wishart distribution, we attempt to generalize it to the noncentral Wishart distribution. We discuss the problems we encounter and explore the boundaries of what can and what cannot be done.

Dropping the normality assumption, but retaining part of its features, leads to the class of spherical and elliptical distributions, concisely introduced in section 2.4. In section 2.5, we study a famous theorem by Schoenberg (1938) in the theory on spherical and elliptical distributions. The class of spherical and elliptical distributions contains distributions whose density contours have the same elliptical shape as some normal distributions, but this class also contains heavier-tailed distributions. The proof of the above-mentioned theorem was originally given by Schoenberg in 1938, alternative proofs were given by Kingman (1972), Donoghue (1969), and Berg et al. (1984). These proofs are rather complicated. We found a shorter, more elegant, and less complicated proof. The normal distribution discussed in the first part of this chapter takes a central position in this theorem. We also include Kingman’s proof of Schoenberg’s theorem, which exploits the property of exchangeability of a sequence of random variables. After a brief introduction to the concept of exchangeability and Kingman’s proof, we conclude this chapter with some remarks regarding the relationship between sphericity and exchangeability.

Sections 2.1–2.3 are partly based on Steerneman (1998), who was inspired by Ferguson (1967) and Eaton (1972).
2.1 The multivariate normal distribution

2.1.1 Definition

Consider a random variable \( X \) that has a normal distribution with mean \( \mu \) and variance \( \sigma^2 \), that is \( X \sim \mathcal{N}(\mu, \sigma^2) \) with characteristic function

\[
\phi_X(t) = e^{it\mu - \frac{1}{2}t^2\sigma^2}.
\]

It is usually assumed that \( \sigma^2 > 0 \). However, we shall allow \( \sigma^2 = 0 \), which implies that \( X \) equals \( \mu \) with probability one. If \( X \) is an \( m \)-dimensional random vector, its distribution is uniquely determined by the distributions of linear functions \( \alpha'X \), where \( \alpha \in \mathbb{R}^m \). This is sometimes called ‘Cramér’s device’ (theorem 1.2.2 from Muirhead, 1982). To see this, note that the characteristic function (c.f.) of \( \alpha'X \) equals

\[
\phi_{\alpha'X}(t) = E e^{it\alpha'X},
\]

so that

\[
\phi_{\alpha'X}(1) = E e^{i\alpha'X},
\]

which is the c.f. of \( X \) if considered as a function of \( \alpha \). Since we know that a distribution is uniquely determined by its c.f. the result follows. Encouraged by this fact, an elegant and well-known way to define the multivariate normal distribution is by means of linear combinations of univariate normal distributions.

Definition 2.1.1. The random vector \( X \in \mathbb{R}^m \) has an \( m \)-variate normal distribution if, for all \( \alpha \in \mathbb{R}^m \), the distribution of \( \alpha'X \) is univariate normal.

From this definition it immediately results that \( X_i \) follows a univariate normal distribution, say \( \mathcal{N}(\mu_i, \sigma_{ij}) \), for \( i = 1, \ldots, m \). This also implies that \( \text{Cov}(X_i, X_j) = \sigma_{ij} \) is defined. If we take \( \mu = (\mu_1, \ldots, \mu_m)' \) and \( \Sigma = (\sigma_{ij}) \), then it follows that \( E X = \mu \) and \( \text{Var}X = \Sigma \). For any \( \alpha \in \mathbb{R}^m \) we have \( E \alpha'X = \alpha'\mu \) and \( \text{Var}(\alpha'X) = \alpha'\Sigma\alpha \). We obtain the c.f. of \( X \) by observing

\[
\phi_X(t) = E e^{it'X} = \phi_{t'X}(1) = e^{it'\mu - \frac{1}{2}t'\Sigma t}, \tag{2.1.1}
\]

since \( t'X \sim \mathcal{N}(t'\mu, t'\Sigma t) \). Therefore, it makes sense to use the notation \( X \sim \mathcal{N}^m(\mu, \Sigma) \). An immediate consequence of definition 2.1.1 is the following proposition.

Proposition 2.1.1. If \( X \) has an \( m \)-variate normal distribution, then for any \( n \times m \) matrix \( A \) and \( b \in \mathbb{R}^n \), the random vector \( Y = AX + b \) follows an \( n \)-variate normal distribution.
As a consequence, we have for $Y$ as given in proposition 2.1.1 that

$$E Y = A\mu + b$$
$$\text{Var}(Y) = A\Sigma A'.$$

Note, however, that we have not proved the existence of the multivariate normal distribution yet. To do this, we have to show that $\varphi_X(t)$ is indeed a c.f. of a random vector. Suppose that $Z_1, \ldots, Z_k$ are independently distributed as $\mathcal{N}(0, 1)$, and $\Sigma$ is an $m \times m$ positive semi-definite matrix of rank $k$. The c.f. of $Z = (Z_1, \ldots, Z_k)'$ then equals

$$\varphi_Z(t) = E e^{it'Z} = \prod_{j=1}^{k} E e^{it_jZ_j} = \prod_{j=1}^{k} e^{-\frac{1}{2}t_j^2} = e^{-\frac{1}{2}t't}. \quad (2.1.2)$$

Define $X = \mu + AZ$, where $A$ is an $m \times k$ matrix of rank $k$ such that $\Sigma = AA'$. Then we have

$$\varphi_X(t) = E e^{it'X} = e^{it'\mu} \varphi_Z(A't) = e^{it'\mu - \frac{1}{2}t'\Sigma t},$$

which is the c.f. as given in (2.1.1).

In definition 2.1.1 we do not restrict ourselves to the situation where $\Sigma$ is positive definite, but also allow for singularity of the covariance matrix (we require $\Sigma$ to be positive semi-definite). If $\Sigma$ is not positive definite, then there exists a vector $\alpha \in \mathbb{R}^m, \alpha \neq 0$, such that $\text{Var}(\alpha'X) = \alpha'\Sigma\alpha = 0$, so that $\alpha'X = c$ with probability one, where $c = \alpha'E(X)$. Thus, the components $X_i$ of $X$ are linearly related, which means that $X$ lies in an affine subspace.

In the above notation, the distribution of $X$ is singular if $k < m$. Although the density of $X$ with respect to the Lebesgue measure on $\mathbb{R}^m$ does not exist in this case, it is important to be able to work with such a singular normal distribution, since situations in which the covariance matrix is singular can arise naturally, as illustrated by the following examples. We will show in the next subsection that the density does exist on an affine subspace.

**Example 2.1.1.** Consider the case where $X \in \mathbb{R}^k$ has a multinomial distribution with parameters $n$ and $\pi = (\pi_1, \ldots, \pi_k)'.$ The Central Limit Theorem tells us that, asymptotically,

$$\sqrt{n}(\hat{\pi} - \pi) \sim \mathcal{N}_k(0, \text{diag}(\pi) - \pi\pi'),$$

where $\hat{\pi} = \frac{1}{n}X$. Since we know that $\sum_{i=1}^{k} \pi_i = 1$, $\text{diag}(\pi) - \pi\pi'$ is singular.
Example 2.1.2. Consider a standard linear regression model \( y = X\beta + \varepsilon \), \( y \in \mathbb{R}^n \), \( X \) a nonstochastic \( n \times k \) matrix of full column rank, and \( \varepsilon \sim \mathcal{N}_n(0, \sigma^2 I_n) \). The distribution of the residuals \( e = y - X\hat{\beta} = (I_n - X(X'X)^{-1}X')\varepsilon \), where \( \hat{\beta} = (X'X)^{-1}X'y \), is straightforward to derive:

\[
e \sim \mathcal{N}_n \left( 0, \sigma^2(I_n - X(X'X)^{-1}X') \right),
\]

where the covariance matrix \( \sigma^2(I_n - X(X'X)^{-1}X') \) is of rank \( n - k \).

Example 2.1.3. A property of economic data is that there are many logical constraints on the data items, such as the fact that company profits must equal turnover minus expenses. If, for example, the data vector \( X = (X_1, \ldots, X_k)' \) must satisfy the restriction \( X_1 + \cdots + X_{k-1} = X_k \), then the covariance matrix \( \Sigma \) of \( X \) will be singular.

Other applications of singular covariance matrices will be discussed in chapter 4, which explicitly deals with (parsimoniously specified) singular covariance matrices.

2.1.2 The probability density function

It is well-known that if \( X \sim \mathcal{N}_m(\mu, \Sigma) \), with rank(\( \Sigma \)) = \( k < m \), then the probability density function of \( X \) with respect to the Lebesgue measure on \( \mathbb{R}^m \) does not exist. However, the density function exists on a subspace, a result which is due to Khatri (1968).

Assume that \( X \sim \mathcal{N}_m(\mu, \Sigma) \) with rank(\( \Sigma \)) = \( k \leq m \), so that \( \Sigma \) is allowed to be singular. In this situation \( \mathcal{K}(\Sigma) \), the kernel of \( \Sigma \), has dimension \( m - k \). The random vector \( X \) will have outcomes in an affine subspace of \( \mathbb{R}^m \).

Proposition 2.1.2. If \( X \sim \mathcal{N}_m(\mu, \Sigma) \), with rank(\( \Sigma \)) = \( k \leq m \), then \( X \) takes values in \( V = \mu + \mathcal{K}(\Sigma)^{\perp} \) with probability one.

Clearly \( V \) is an affine subspace of \( \mathbb{R}^m \) of dimension \( k \). Before we give the proof, a couple of definitions will be given. Let \( \Sigma = CA'C \) be a spectral decomposition of \( \Sigma \), where

\[
\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_k, 0, \ldots, 0) = \begin{pmatrix} \Lambda_1 & 0 \\ 0 & 0 \end{pmatrix}, \tag{2.1.3}
\]

\( \Lambda_1 = \text{diag}(\lambda_1, \ldots, \lambda_k) \), \( \lambda_1 \geq \cdots \geq \lambda_k > 0 \), and \( C \) is an orthogonal \( m \times m \) matrix. Moreover, let

\[
C = (C_1, C_2),
\]
where $C_1$ is an $m \times k$ matrix

\[
C_1 = (c_1, \ldots, c_k) \\
C_2 = (c_{k+1}, \ldots, c_m).
\]

Therefore, \{\(c_1, \ldots, c_m\)\} is an orthonormal basis for \(\mathbb{R}^m\) such that \{\(c_1, \ldots, c_k\)\} is an orthonormal basis for \(\mathcal{K}(\Sigma)\) and \{\(c_{k+1}, \ldots, c_m\)\} is an orthonormal basis for \(\mathcal{K}(\Sigma)\). Furthermore, note that \(\mathcal{R}(\Sigma) = \mathcal{K}(\Sigma)^\perp\), where \(\mathcal{R}(\Sigma)\) denotes the columns space of \(\Sigma\). This follows from the fact that \(\mathcal{R}(\Sigma) = \mathcal{K}(\Sigma)^\perp\Rightarrow \mathcal{R}(\Sigma)^\perp = \mathcal{K}(\Sigma)\); now \(x \in \mathcal{R}(\Sigma)^\perp \Leftrightarrow x'y = 0\) for all \(y \in \mathcal{R}(\Sigma)\Leftrightarrow x'\Sigma z = 0\) for all \(z \in \mathbb{R}^k\Leftrightarrow \Sigma x = 0 \Leftrightarrow x \in \mathcal{K}(\Sigma)\).

**Proof of proposition 2.1.2.** We observe that

\[
P\{X \in V\} = P\{X - \mu \in \mathcal{K}(\Sigma)^\perp\} = P\{y'(X - \mu) = 0\text{ for all } y \in \mathcal{K}(\Sigma)\} = P\{y'(X - \mu) = 0\text{ for all } y \text{ with } \Sigma y = 0\} = P\{c_j'(X - \mu) = 0\text{ for } j = k + 1, \ldots, m\}.
\]

Now we have for \(j = k + 1, \ldots, m\)

\[
c_j'(X - \mu) \sim \mathcal{N}(0, c_j'\Sigma c_j) = \delta_{\{0\}}.
\]

Therefore,

\[
P\{X \notin V\} = P\{c_j'(X - \mu) \neq 0\text{ for some } j = k + 1, \ldots, m\} \leq \sum_{j=k+1}^{m} P\{c_j'(X - \mu) \neq 0\} = 0.
\]

Hence \(P\{X \in V\} = 1\). \qed

Now we want to derive the probability density function of \(X\) on \(V\) with respect to the Lebesgue measure \(\lambda_V\) on \(V\). Consider the affine transformation \(T: \mathbb{R}^k \to V\) defined by

\[
x = T(y) = \sum_{i=1}^{k} y_i c_i + \mu = C_1 y + \mu.
\]

Note that \(y = T^{-1}(x) = C_1'(x - \mu)\). Let \(\lambda_k\) denote the Lebesgue measure on \(\mathbb{R}^k\). The following properties hold for \(T\):
2.1. The multivariate normal distribution

(i) $T$ is onto and one-to-one;
(ii) $T$ is bicontinuous;
(iii) $\lambda_V = \lambda_k T^{-1}$.

Property (iii) holds in fact for all choices $c_1, \ldots, c_k$ and $\mu$ such that $\mu \in V$ and $c_1, \ldots, c_k$ is an orthonormal basis for $V - \mu$. Before we derive the probability density function of $X$, we first introduce the definition

$$
\det_r A = \lim_{\alpha \to 0} \frac{\det(A + \alpha I_n)}{\alpha^{n-r}}
$$

for an $n \times n$ symmetric matrix $A$ of rank $r$. It is well-known that

$$
\det_r A = \prod_{i=1}^{r} \lambda_i(A),
$$

that is, $\det_r A$ equals the product of the nonzero eigenvalues of $A$ (see, e.g., Neudecker, 1995).

**Theorem 2.1.1.** If $X \sim \mathcal{N}_m(\mu, \Sigma)$ with $\text{rank}(\Sigma) = k \leq m$, then $X$ has probability density function

$$
f(x) = (2\pi)^{-\frac{k}{2}}[\det_k \Sigma]^{-\frac{1}{2}} \exp \left( -\frac{1}{2} (x - \mu)' \Sigma^+ (x - \mu) \right)
$$

for $x \in V = \mu + \mathcal{K}(\Sigma)^\perp$ with respect to the Lebesgue measure $\lambda_V$ on $V$, where $\det_k \Sigma$ denotes the product of the positive eigenvalues of $\Sigma$.

**Proof.** Let $Y = C_1^T (X - \mu)$, where $Y \sim \mathcal{N}_k(0, \Lambda_1)$. From the characteristic function of $Y$, we know that $Y_1, \ldots, Y_k$ are independently distributed with $Y_i \sim \mathcal{N}(0, \lambda_i)$ for $i = 1, \ldots, k$. This implies that $Y$ has probability density function

$$
h(y) = \prod_{i=1}^{k} (2\pi)^{-\frac{1}{2}} \lambda_i^{-\frac{1}{2}} \exp \left( -\frac{1}{2} y_i^2 / \lambda_i \right)
$$

$$
= (2\pi)^{-\frac{k}{2}} [\det \Lambda_1]^{-\frac{1}{2}} \exp \left( -\frac{1}{2} y' \Lambda_1^{-1} y \right)
$$

with respect to the Lebesgue measure $\lambda_k$ on $\mathbb{R}^k$. We have $Y = T^{-1}(X)$. Let $B$ be any Lebesgue measurable set in $V$, then

$$
P \{ X \in B \} = P \left\{ T^{-1}(X) \in T^{-1}(B) \right\}
$$

$$
= P \left\{ Y \in T^{-1}(B) \right\}
$$

$$
= \int_{T^{-1}(B)} h \, d\lambda_k
$$

$$
= \int_{T^{-1}(B)} h \, d\lambda_V T
$$

$$
= \int_B h(T^{-1}(x)) \, d\lambda_V(x).
$$
Hence
\[
f(x) = h(T^{-1}(x))
= (2\pi)^{-\frac{k}{2}} |\det \Lambda_1|^{-\frac{1}{2}} \exp \left( -\frac{1}{2} (x - \mu)' C_1 \Lambda_1^{-1} C_1' (x - \mu) \right)
= (2\pi)^{-\frac{k}{2}} |\det \Sigma|^{-\frac{1}{2}} \exp \left( -\frac{1}{2} (x - \mu)' \Sigma^+ (x - \mu) \right)
\]
is a density function of \(X\) on \(V\) with respect to \(\lambda_V\).

Note that, if we take \(k = m\) in theorem 2.1.1, equation (2.1.4) reduces to
\[
f(x) = (2\pi)^{-\frac{m}{2}} |\Sigma|^{-\frac{1}{2}} \exp \left( -\frac{1}{2} (x - \mu)' \Sigma^{-1} (x - \mu) \right),
\]
that is, the well-known expression for the probability density function of a nonsingular multivariate normal distribution with respect to the Lebesgue measure on \(\mathbb{R}^m\).

### 2.1.3 Some well-known properties of the multivariate normal distribution

In this subsection, we summarize some well-known results on the multivariate normal distribution. There is a huge amount of literature on the normal distribution (e.g., Anderson, 1984, Muirhead, 1982, Eaton, 1983, Tong, 1990), and therefore most of the results here will be stated without a proof. This subsection is certainly not intended to be exhaustive. The results stated in this subsection will be used further on in this chapter.

From proposition 2.1.1 and the characteristic function of the normal distribution, we can easily derive the following proposition.

**Proposition 2.1.3.** Let \(X \sim \mathcal{N}_m(\mu, \Sigma), b \in \mathbb{R}^m\), and \(A\) be an \(n \times m\) matrix, then \(\mathcal{L}(AX + b) = \mathcal{N}_n(A\mu + b, A\Sigma A')\).

It follows from proposition 2.1.3 that the marginal distribution of any subset of \(p\) components of \(X\) is \(p\)-variate normal \((p < m)\). Note, however, that the converse is not true in general (e.g., Muirhead, 1982), that is, if each component of a random vector is normal, this does not imply that the vector has a multivariate normal distribution, as illustrated by the following example\(^1\).

\(^1\)The converse is true if all components of \(X\) are independent and normal.
Example 2.1.4. Consider the case where \( U_1, U_2, \) and \( U_3 \) are i.i.d. \( \mathcal{N}(0,1) \) and \( Z \) is an arbitrary random variable. Define

\[
X_1 = \frac{U_1 + ZU_3}{\sqrt{1 + Z^2}} \quad \text{and} \quad X_2 = \frac{U_2 + ZU_3}{\sqrt{1 + Z^2}}.
\]

It is easy to show that in this case \( \mathcal{L}(X_1 | Z = z) = \mathcal{L}(X_2 | Z = z) = \mathcal{N}(0,1) \) which implies that \( \mathcal{L}(X_1) = \mathcal{L}(X_2) = \mathcal{N}(0,1) \). Conditional on \( Z \), the joint distribution of \( X_1 \) and \( X_2 \) is also normal,

\[
\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} | Z \sim \mathcal{N}_2 \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \frac{1}{1 + Z^2} \begin{pmatrix} 1 + Z^2 & Z^2 \\ Z^2 & 1 + Z^2 \end{pmatrix} \right).
\]

Obviously, the unconditional distribution of \( X_1 \) and \( X_2 \) need not be bivariate normal.

To study partitioned random vectors, define

\[
X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},
\]

where \( X_1, \mu_1 \in \mathbb{R}^p \) and \( \Sigma_{11} \) is \( p \times p \), \( p < m \). Let \( \Sigma_{11.2} = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^+ \Sigma_{21} \), the so-called generalized Schur-complement of \( \Sigma_{22} \) in \( \Sigma \). These (generalized) Schur-complements appear, for example, in expressions for the inverse of a partitioned matrix, in conditional distributions and partial correlation coefficients (Ouellette, 1981). A famous theorem on Wishart distributed partitioned matrices, which will be discussed in section 2.3, also exploits Schur-complements.

Proposition 2.1.4. \( X_1 \) and \( X_2 \) are independently distributed if and only if \( \Sigma_{12} = 0 \).

This proposition follows immediately by considering the characteristic function of \( X_1 \) and \( X_2 \).

Proposition 2.1.5.

\[
\mathcal{L} \left( \begin{pmatrix} X_1 - \Sigma_{12} \Sigma_{22}^+ X_2 \\ X_2 \end{pmatrix} \right) = \mathcal{N}_m \left( \begin{pmatrix} \mu_1 - \Sigma_{12} \Sigma_{22}^+ \mu_2 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11.2} & 0 \\ 0 & \Sigma_{22} \end{pmatrix} \right).
\]

Proof. Define the matrix

\[
A = \begin{pmatrix} I_p & -\Sigma_{12} \Sigma_{22}^+ \\ 0 & I_{m-p} \end{pmatrix},
\]
then the proposition follows by considering $L(AX)$. Note that $\Sigma_{12} = \Sigma_{12}^{+} \Sigma_{22}$. This can be established as follows. Consider $L(BX)$, where

$$B = \begin{pmatrix} I_p & 0 \\ 0 & I_{m-p} - \Sigma_{22}^{+} \Sigma_{22} \end{pmatrix}.$$ 

Then

$$\text{Var}(BX) = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} - \Sigma_{12}^{+} \Sigma_{22} \\ \Sigma_{21}^{+} \Sigma_{22}^{+} - \Sigma_{22}^{+} \Sigma_{22} & \Sigma_{21} - \Sigma_{22}^{+} \Sigma_{22} \end{pmatrix},$$

and hence $\Sigma_{12} - \Sigma_{12}^{+} \Sigma_{22} = 0$. □

As a consequence of proposition 2.1.5 we have

**Proposition 2.1.6.** The conditional distribution of $X_1$ given $X_2$ is given by

$$L(X_1|X_2) = \mathcal{N}_p(\mu_1 + \Sigma_{12}^{+} \Sigma_{22} (X_2 - \mu_2), \Sigma_{11}).$$

**Proof.** Since $X_1 - \Sigma_{12}^{+} \Sigma_{22} X_2$ and $X_2$ are independently distributed according to proposition 2.1.5, we have

$$L(X_1 - \Sigma_{12}^{+} \Sigma_{22} X_2|X_2) = \mathcal{N}_p(\mu_1 - \Sigma_{12}^{+} \Sigma_{22} \mu_2, \Sigma_{11}),$$

and hence

$$L(X_1|X_2) = L(X_1 - \Sigma_{12}^{+} \Sigma_{22} X_2 + \Sigma_{12}^{+} \Sigma_{22} (X_2 - \mu_2)|X_2) = \mathcal{N}_p(\mu_1 + \Sigma_{12}^{+} \Sigma_{22} (X_2 - \mu_2), \Sigma_{11}).$$

□

The entropy of a probability distribution is a measure for the information content of this distribution. The concept of entropy originates from the area of information theory and was introduced by Shannon (1948). A high entropy value indicates that a large amount of uncertainty or randomness is associated with the random variable. It is commonly accepted that if one is asked to select a distribution satisfying some constraints, and if these constraints do not determine a unique distribution, then one is best off picking the distribution having maximum entropy. The idea is that this distribution incorporates the least possible information. We will exploit this idea in chapter 4.

**Proposition 2.1.7.** The entropy of $\mathcal{N}_m(\mu, \Sigma)$ with respect to $\lambda_V$, with $V = \mu + \mathcal{K}(\Sigma)^{+}$ and $\text{rank}(\Sigma) = k \leq m$, equals

$$H = \frac{1}{2} k (1 + \log 2\pi) + \frac{1}{2} \log(\det_k \Sigma). \quad (2.1.5)$$
2.2. Multivariate normally distributed random matrices

Proof.

\[ H = - \int_V f \log f \, d\lambda_V \]
\[ = E \left[ \frac{1}{2} k \log 2\pi + \frac{1}{2} \log |\det \Sigma| + \frac{1}{2} (X - \mu) \Sigma^+ (X - \mu) \right] \]
\[ = \frac{1}{2} k (1 + \log 2\pi) + \frac{1}{2} \log |\det \Sigma|. \]

In the last step we used

\[ E (X - \mu) \Sigma^+ (X - \mu) = \text{tr} \Sigma^+ (X - \mu) (X - \mu)' = \text{tr} \Sigma^+ = \text{rank} (\Sigma^+) = k. \]

\[ \square \]

If we take \( k = m \) in proposition 2.1.7, then we obtain the well-known expression

\[ H = \frac{1}{2} m (1 + \log 2\pi) + \frac{1}{2} \log |\Sigma|, \quad (2.1.6) \]

for the entropy value of \( \mathcal{N}_m(\mu, \Sigma) \) with \( \Sigma \) nonsingular.

For the univariate case, Kagan et al. (1973) gave a general characterization of distributions with maximum entropy in their theorem 13.2.1. Their result can be extended to the multivariate case, and an important multivariate application is the following (theorem 13.2.2 of Kagan et al., 1973).

**Theorem 2.1.2.** Let \( X \) be an \( m \)-variate random vector with expectation vector \( \mu \) and covariance matrix \( \Sigma \). The maximum entropy distribution of \( X \) with respect to \( \lambda_V \) is the \( m \)-variate normal \( \mathcal{N}_m(\mu, \Sigma) \), where \( V = \mu + \mathcal{K}(\Sigma)^\perp \).

2.2 Multivariate normally distributed random matrices

Just as we generalized a random variable to a random vector, we now generalize a random vector to a random matrix. In section 2.3, where we study the Wishart distribution, it turns out that notational convenience is provided by using random matrices whose elements are multivariate normally distributed.

**Definition 2.2.1.** Let \( X \) be an \( n \times k \) matrix of random elements, such that

\[ \text{vec} X' \sim \mathcal{N}_{nk} (\text{vec} M', \Omega), \]

where \( M = \text{EX} \) is an \( n \times k \) matrix and \( \Omega \) is the \( nk \times nk \) covariance matrix of \( \text{vec} X' \). It is then written

\[ X \sim \mathcal{N}_{n \times k} (M, \Omega). \]
As a special case, we now consider the situation in which the rows of \( X \) are identically distributed.

**Proposition 2.2.1.** If \( X = (X_1, \ldots, X_n)' \) is a random \( n \times k \) matrix such that \( X_1, \ldots, X_n \) are independently distributed as \( N_k(\mu_i, \Sigma) \), then

\[
X \sim N_{n \times k}(M, I_n \otimes \Sigma),
\]

where \( M = (\mu_1, \ldots, \mu_n)' \). For the special case that the rows of \( X \) are independently and identically distributed as \( N_k(\mu, \Sigma) \),

\[
X \sim N_{n \times k}(\iota_n \mu', I_n \otimes \Sigma).
\]

**Proof.**

\[
L(\text{vec } X') = L(K_{nk}\text{vec } X) = N_{kn} (K_{nk}\text{vec } (\mu_1, \ldots, \mu_k), K_{nk}(I_k \otimes \Sigma)K_{nk}')
\]

and the result follows from definition 2.2.1. If \( \mu_1 = \ldots = \mu_n = \mu \), then \( M = \iota_n \mu' \). \qed

A similar result also holds in case the columns of \( X \) are independently and identically normally distributed.

**Proposition 2.2.2.** If \( X = (X_1, \ldots, X_k) \) is a random \( n \times k \) matrix such that \( X_1, \ldots, X_k \) are independently distributed as \( N_n(\mu_i, \Sigma) \), then

\[
X \sim N_{n \times k}(M, \Sigma \otimes I_k),
\]

where \( M = (\mu_1, \ldots, \mu_k) \). Furthermore, if the columns of \( X \) are independent and identically distributed as \( N_n(\mu, \Sigma) \), then

\[
X \sim N_{n \times k}(\mu_1, \Sigma \otimes I_k).
\]

**Proof.** Let \( K_{mn} \) denote the \( mn \times mn \) commutation matrix which, for arbitrary \( m \times n \) matrices \( A \), transforms \( \text{vec } A \) into \( \text{vec } A' \) (Magnus and Neudecker, 1979). Then we have

\[
L(\text{vec } X') = L(K_{nk}\text{vec } X) = N_{kn} (K_{nk}\text{vec } (\mu_1, \ldots, \mu_k), K_{nk}(I_k \otimes \Sigma)K_{nk}')
\]

and

\[
N_{kn} (K_{nk}\text{vec } M, K_{nk}(I_k \otimes \Sigma)K_{nk}').
\]
2.3. The central and noncentral Wishart distribution

Since $K_{nk} \text{vec } M = \text{vec } M'$ and $K_{nk}(I_k \otimes \Sigma)K_{nk}' = K_{nk}(I_k \otimes \Sigma)K_{kn} = \Sigma \otimes I_k$,

$$L(\text{vec } X') = \mathcal{N}_{nk}(\text{vec } M', \Sigma \otimes I_k),$$

and the desired result follows. If $\mu_1 = \ldots = \mu_k = \mu$, then $M = \mu I_k$. \hfill $\Box$

Analogous to what we did for multivariate normally distributed vectors, we are now interested in the distribution of a linear transformation of a normally distributed matrix.

**Proposition 2.2.3.** If $X \sim \mathcal{N}_{n \times k}(M, \Omega)$, $A$ is an $m \times n$ matrix, and $B$ is a $k \times h$ matrix, then

$$AXB \sim \mathcal{N}_{m \times h}(AMB, (A \otimes B')\Omega(A' \otimes B)).$$

**Proof.** It is straightforward to derive that

$$L(\text{vec } (AXB)') = L(\text{vec } B'X'A')$$

$$= L((A \otimes B')\text{vec } X')$$

$$= \mathcal{N}_{mh}((A \otimes B')\text{vec } M', (A \otimes B')\Omega(A \otimes B')')$$

$$= \mathcal{N}_{mh}(\text{vec } (AMB)', (A \otimes B')\Omega(A' \otimes B)).$$

\hfill $\Box$

It frequently occurs that we have $n$ independent $k$-variate normal observations assembled in a matrix $X$, which means that $X \sim \mathcal{N}_{n \times k}(M, I_n \otimes \Sigma)$. Note once more that we do not restrict ourselves to the situation that $\Sigma$ is nonsingular, so that it is possible that the observations lie in a hyperplane. If the rank of $\Sigma$ is known, we may wonder whether we are able to say something about the rank of $X$. The answer is given in the following theorem. The proof of this theorem is quite long. For continuity of this chapter, we therefore include the proof in appendix 2.A.

**Theorem 2.2.1.** If $X \sim \mathcal{N}_{n \times k}(M, I_n \otimes \Sigma)$, then

$$\text{rank}(X) = \min(n, \text{rank}(\Sigma + M'M))$$

with probability one.

### 2.3 The central and noncentral Wishart distribution

#### 2.3.1 Definition and some properties

The Wishart distribution is the multivariate analogue of the chi-square distribution. We know that, if $Z_1, \ldots, Z_n$ are independent with $L(Z_i) = \mathcal{N}(\mu_i, 1)$,
then \( \sum_{i=1}^{n} Z_i^2 \) has a noncentral chi-square distribution with \( n \) degrees of freedom and noncentrality parameter \( \lambda = \sum_{i=1}^{n} \mu_i^2 \). If \( \lambda = 0 \), or equivalently, \( \mu_1 = \ldots = \mu_n = 0 \), then it has the chi-square distribution. If the \( Z_i \) are independent \( k \)-dimensional random vectors with \( \mathcal{L}(Z_i) = \mathcal{N}_k(0, \Sigma) \), then the Wishart distribution arises from the distribution of the \( k \times k \) positive semi-definite matrix \( S = \sum_i Z_iZ_i' \). We know from proposition 2.2.1 that if \( Z \) has rows \( Z'_1, \ldots, Z'_n \), then \( \mathcal{L}(Z) = \mathcal{N}_{n \times k}(0, I_n \otimes \Sigma) \). This is how we arrive at the following definition of the Wishart distribution.

**Definition 2.3.1.** Let \( Z \sim \mathcal{N}_{n \times k}(M, I_n \otimes \Sigma) \), then \( S = Z'Z \) has the noncentral Wishart distribution of dimension \( k \), degrees of freedom \( n \), covariance matrix \( \Sigma \) and noncentrality matrix \( \Lambda = M'M \). This distribution is denoted by \( W_k'(n, \Sigma; \Lambda) \). If \( M = 0 \), then \( S \) has the Wishart distribution, denoted by \( W_k(n, \Sigma) \).

Note that, as with the definition of the multivariate normal distribution, definition 2.3.1 of the Wishart distribution is not in terms of a density function, and therefore also allows for singularity of this distribution. According to Eaton (1983, page 317), if \( \mathcal{L}(S) = W_k'(n, \Sigma; \Lambda) \), then \( S \) is positive definite with probability one if and only if \( n \geq k \) and \( \Sigma > 0 \). If \( \Lambda = 0 \), so that \( S \) follows a central Wishart distribution, then this result is true. For the case of the noncentral Wishart distribution, however, the rank-condition on \( \Sigma \) is more subtle, as we show in the following theorem.

**Theorem 2.3.1.** Let \( S \sim W_k'(n, \Sigma; \Lambda) \), then \( S \) is positive definite with probability one if and only if \( n \geq k \) and \( \Sigma + \Lambda > 0 \).

**Proof.** Let \( S = Z'Z \) according to definition 2.3.1. On account of theorem 2.2.1, it follows that

\[
1 = P \left\{ x' S x > 0 \text{ for all } 0 \neq x \in \mathbb{R}^k \right\} \\
= P \left\{ Z x \neq 0 \text{ for all } 0 \neq x \in \mathbb{R}^k \right\} \\
= P \{ \text{rank}(Z) = k \}
\]

if and only if \( n \geq k \) and \( \Sigma + M'M = \Sigma + \Lambda > 0 \). \( \square \)

We call the distribution of \( S \sim W_k'(n, \Sigma; \Lambda) \) singular if \( n < k \) and/or \( \Sigma + \Lambda \) is not positive definite.

In the remainder of this subsection, we summarize some well-known properties of the Wishart distribution. Once more, we do not intend to be exhaustive, yet we state some properties we need in subsection 2.3.2 to prove
an important theorem on partitioned Wishart distributed matrices. For the basic properties such as $E\mathbf{S}$, $\text{Var}(\mathbf{S})$, $E\mathbf{S}^{-1}$, and the characteristic function of $\mathbf{S}$, we refer to Muirhead (1982). A proof of the following lemma for the central case was given by Eaton (1983). This proof is easily generalized to the noncentral case. Whenever possible, we generalize known results for the case of the central Wishart distribution to the noncentral Wishart distribution.

**Proposition 2.3.1.** Let $\mathbf{S} \sim W_p^\prime(n, \Sigma)$ and $A$ be a $p \times k$ matrix, then $ASA' \sim W_p^\prime(n, A\Sigma A'; AA')$.

*Proof.* From definition 2.3.1, we know that $L(\mathbf{S}) = W_p^\prime(n, \Sigma)$, where $Z \sim \mathcal{N}_{n \times k}(M, I_n \otimes \Sigma)$ and $\Lambda = M'M$. Proposition 2.2.3 then tells us that $ZA' \sim \mathcal{N}_{n \times p}(MA', I_n \otimes A)$. Therefore, $L(ASA') = L(ZA')(ZA') = W_p^\prime(n, A\Sigma A'; AA')$.

A consequence of this lemma is that a Wishart distribution $W_p^\prime(n, \Sigma)$ can be generated from the $W_k^\prime(n, I_k; \Omega)$ distribution and $k \times k$ matrices. Another consequence is that the marginal distribution of any principal square submatrix of $\mathbf{S}$ is also Wishart.

**Proposition 2.3.2.** If $\mathbf{S}$ is $W_p^\prime(n, \Sigma)$ and $\mathbf{S}$, $\Sigma$ and $\Lambda$ are partitioned as

$$
\mathbf{S} = \begin{pmatrix}
S_{11} & S_{12} \\
S_{21} & S_{22}
\end{pmatrix}, \quad \Sigma = \begin{pmatrix}
\Sigma_{11} & 0 \\
0 & \Sigma_{22}
\end{pmatrix}, \quad \text{and} \quad \Lambda = \begin{pmatrix}
\Lambda_{11} & \Lambda_{12} \\
\Lambda_{21} & \Lambda_{22}
\end{pmatrix}
$$

where $S_{11}$ and $\Sigma_{11}$ are $p \times p$ matrices, then $S_{11}$ and $S_{22}$ are independent and their distributions are, respectively, $W_p^\prime(n, \Sigma_{11}; \Lambda_{11})$ and $W_{k-p}^\prime(n, \Sigma_{22}; \Lambda_{22})$.

*Proof.* Let $\mathbf{S} = X'X$, where $X \sim \mathcal{N}_{n \times k}(M, I_n \otimes \Sigma)$ and $\Lambda = M'M$. Define $X = (X_1, X_2)$, where $X_1$ is an $n \times p$ matrix. Partition $M$ accordingly as $M = (M_1, M_2)$. Because $\Sigma_{12} = 0$, we know that $X_1$ and $X_2$ are independent. But this implies that $S_{11} = X_1'X_1$ and $S_{22} = X_2'X_2$ are also independent. The fact that $S_{11} \sim W_p^\prime(n, \Sigma_{11}; \Lambda_{11})$ and $S_{22} \sim W_{k-p}^\prime(n, \Sigma_{22}; \Lambda_{22})$ follows immediately from proposition 2.3.1.

**Proposition 2.3.3.** If $S_1, \ldots, S_k$ are independent with $L(S_i) = W_p^\prime(n_i, \Sigma_i)$ for $i = 1, \ldots, k$, then

$$
L\left(\sum_{i=1}^k S_i\right) = W_p^\prime\left(\sum_{i=1}^k n_i, \Sigma; \sum_{i=1}^k \Lambda_i\right).
$$
Proof. Let $Z_1, \ldots, Z_k$ be independent, with $\mathcal{L}(Z_i) = \mathcal{N}_{n_i \times p}(M_i, I_{n_i} \otimes \Sigma)$ and $\Lambda_i = M_i' M_i$. Define 

$$Z = \begin{pmatrix} Z_1 \\ \vdots \\ Z_k \end{pmatrix}, \quad \text{and} \quad M = \begin{pmatrix} M_1 \\ \vdots \\ M_k \end{pmatrix}$$

so that $\mathcal{L}(Z) = \mathcal{N}_{n \times p}(M, I_n \otimes \Sigma)$, where $n = \sum_{i=1}^{k} n_i$. Since $\sum_{i=1}^{k} \Lambda_i = M'M$, the result easily follows by observing that $\mathcal{L} \left( \sum_{i=1}^{k} S_i \right) = \mathcal{L}(Z'Z)$. \qed

**Proposition 2.3.4.** Let $X \sim \mathcal{N}_{n \times k}(M, \Sigma \otimes \Sigma)$, where $P$ is an idempotent, symmetric $n \times n$ matrix of rank $m > 0$, then

$$\mathcal{L}(X'X - M'(I_n - P)M) = W'_k(m, \Sigma; M'PM).$$

Proof. Since $P$ is an idempotent, symmetric matrix of rank $m$, we know that there exists a matrix $C \in O(n)$, such that 

$$CPC' = \begin{pmatrix} I_m & 0 \\ 0 & 0 \end{pmatrix}.$$ 

On account of proposition 2.2.3, 

$$Y = CX \sim \mathcal{N}_{n \times k}(CM, \begin{pmatrix} I_m & 0 \\ 0 & 0 \end{pmatrix} \otimes \Sigma).$$

Partition 

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}, \quad C = \begin{pmatrix} C_1 \\ C_2 \end{pmatrix},$$

where $Y_1$ is a random $m \times k$ matrix and $C$ is an $m \times n$ matrix. Note that 

$$\mathcal{L}(Y_1) = \mathcal{N}_{m \times k}(C_1 M, I_m \otimes \Sigma),$$

and that $P\{Y_2 = C_2 M\} = 1$. Because $P = C_1' C_1$ and $I_n - P = C_2' C_2$, it follows that

$$X'X = X'PX + X'(I_n - P)X = X' C_1' X + X'C_2'X = Y_1' Y_1 + Y_2' Y_2.$$ 

Thus $\mathcal{L}(X'X) = \mathcal{L}(Y_1' Y_1 + M' C_2' C_2 M)$. From the definition of a Wishart distribution, we know that $\mathcal{L}(Y_1' Y_1) = W'_k(m, \Sigma; M'PM)$, so that the result immediately follows. \qed
Proposition 2.3.5. Let \( X \sim \mathcal{N}_{n \times k}(M, I_n \otimes \Sigma) \), and \( P \) be an idempotent, symmetric \( n \times n \) matrix of rank \( m > 0 \), then
\[
\mathcal{L}(X'PX) = W_k'(m, \Sigma; PM).
\]

Proof. According to proposition 2.2.3, \( PX \sim \mathcal{N}_{n \times k}(PM, P \otimes \Sigma) \). Since \( X'PX = X'P'PX \), it follows from proposition 2.3.4 that
\[
\mathcal{L}(X'PX - M'P(I_n - P)PM) = W_k'(m, \Sigma; M'PM).
\]
Since \( P(I_n - P) = 0 \), the proof is accomplished. \( \Box \)

2.3.2 Partitioned Wishart distributed matrices

In this subsection, we study partitioned Wishart distributed matrices. We derive a well-known theorem for the case of the central Wishart distribution, a theorem which is one of the most useful results for deriving the distribution of functions of Wishart matrices (Eaton, 1983). The distributions of many statistics, such as Hotelling’s \( T^2 \) and the generalized variance \( |S| \), follow from it. The theorem can also be applied to obtain the expectation of the inverse of a Wishart distributed matrix, after some invariance considerations. A proof for the case of the central Wishart distribution with non-singular covariance matrix \( \Sigma \) can be found in, for example, Muirhead (1982) and Eaton (1983). Whereas this theorem deals with the central Wishart distribution, we will try to obtain similar results for the noncentral Wishart distribution. In fact, in deriving the theorem, we start with the most general case of the noncentral Wishart distribution. These derivations very well show the complications that occur when considering the noncentral case. We show that parts of the theorem can be generalized to the noncentral case, and parts of the theorem cannot. As a second extension of the theorem, we will also allow the Wishart distribution to be singular. Although Srivastava and Khatri (1979) stated that the theorem is true if \( \Sigma \) is singular, they did not actually prove so. To our knowledge, no such extension of the proof has been given yet. We show what happens if we allow singularity and replace the usual inverse of a matrix by the Moore-Penrose inverse.

Consider \( S \sim W_k'(n, \Sigma; \Lambda) \) and partition \( S \) and \( \Sigma \):
\[
S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}, \quad \text{and} \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},
\]
where \( S_{11} \) and \( \Sigma_{11} \) are \( p \times p \) matrices, \( p < k \). Let \( S = X'X \), where \( X \sim \mathcal{N}_{n \times k}(M, I_n \otimes \Sigma) \), and \( M = (\mu_1, \ldots, \mu_n)' \). From proposition 2.2.1, we know
that this implies that the rows of the matrix $X$ are independently distributed as $\mathcal{N}_k(\mu_i, \Sigma)$. Partition $X = (X_1, X_2)$, where $X_1$ is a random $n \times p$ matrix, and partition $M$ accordingly as $M = (M_1, M_2)$. Let $r = \text{rank}(\Sigma + M_2')M_2) < n$. Since
\[
\mathcal{L}(X_2) = \mathcal{N}_{h \times (k-p)}(M_2, I_n \otimes \Sigma),
\] (2.3.1)
we observe that
\[
\mathcal{L}(S_{22}) = \mathcal{L}(X_2'X_2) = W_{k-p}(n, \Sigma_{22}; M_2'M_2).
\] (2.3.2)

According to proposition 2.1.6 and proposition 2.2.1,
\[
\mathcal{L}(X_1 | X_2) = \mathcal{N}_{n \times p}(M_1 + (X_2 - M_2)\Sigma_{22}^+\Sigma_{21}, I_n \otimes \Sigma_{11.2}).
\]

Define $P = X_2'X_2 + X_2'X_2$ and $Q = I_n - P$. These projection matrices are symmetric and idempotent $n \times n$ matrices with $PQ = 0$. Furthermore, we have that $PX_2 = X_2, QX_2 = 0$, and $X_1QX_1 = S_{11.2}$. From proposition 2.2.3 we derive that
\[
\mathcal{L}(PX_1 | X_2) = \mathcal{N}_{n \times p}(PM_1 + (X_2 - PM_2)\Sigma_{22}^+S_{21}, P \otimes \Sigma_{11.2}),
\] (2.3.3)

implying that
\[
\mathcal{L}(X_2'X_1 | X_2) = \mathcal{N}_{(k-p) \times p}(X_2'PM_1 + X_2'(X_2 - PM_2)\Sigma_{22}^+S_{21}, (X_2'PX_2) \otimes \Sigma_{11.2})
= \mathcal{N}_{(k-p) \times p}(X_2'M_1 + (S_{22} - X_2'M_2)\Sigma_{22}^+S_{21}, S_{22} \otimes \Sigma_{11.2})
= \mathcal{N}_{(k-p) \times p}(X_2'(M_1 - M_2\Sigma_{22}^+S_{21}) + S_{22}^{-1}S_{22}^{-1}S_{22}^+S_{21}, S_{22} \otimes \Sigma_{11.2}).
\] (2.3.4)

From (2.3.3) we also see that
\[
\mathcal{L}(QX_1 | X_2) = \mathcal{N}_{n \times p}(Q(M_1 - M_2\Sigma_{22}^+S_{21}), Q \otimes \Sigma_{11.2}).
\]

Since $\text{rank}(\Sigma + M_2) = r < n$, it follows from (2.3.1) and theorem 2.2.1 that $P\{\text{rank}(X_2) = r\} = 1$. Let $B$ denote the set of all $n \times (k - p)$ matrices of
of freedom corresponding to the distribution in (2.3.5) equals $n$.

Note that we restrict ourselves to matrices $N \in B$ to assure that the degrees of freedom corresponding to the distribution in (2.3.5) equals $n - r$. After all, if rank($N$) = $q < r$, then $\mathcal{L}(X'QX_1|X_2 = N)$ would be as in (2.3.5), yet with degrees of freedom equal to $n - q$. Let $W$ be a random $p \times p$ matrix, independent of $X$, with

$$W \sim W_p'(n - r, \Sigma_{11,2}, (M_1 - M_2^\Sigma_{22}^+ \Sigma_{21})'Q(M_1 - M_2^\Sigma_{22}^+ \Sigma_{21})).$$

(2.3.6)

This matrix $W$ turns out to be useful further on. We can now prove the following theorem.

**Theorem 2.3.2.** Consider $S \sim W_k(n, \Sigma)$ and partition $S$ and $\Sigma$:

$$S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}, \text{and } \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

where $S_{11}$ and $\Sigma_{11}$ are $p \times p$ matrices. Let $r = \text{rank}(\Sigma_{22}) < n$ and $p < k$. We then have

(i) $S_{11,2}$ and $(S_{21}, S_{22})$ are independently distributed,

(ii) $\mathcal{L}(S_{11,2}) = W_p'(n - r, \Sigma_{11,2}),$

(iii) $\mathcal{L}(S_{22}) = W_{k-p}(n, \Sigma_{22}),$

(iv) $\mathcal{L}(S_{21}|S_{22}) = \mathcal{N}_{(k-p)\times p}(S_{22}^\Sigma_{22}^+ \Sigma_{21}, S_{22} \otimes \Sigma_{11,2}).$

**Proof.** Because we now consider the central Wishart distribution, it follows from (2.3.2) with $M = 0$ that $\mathcal{L}(S_{22}) = \mathcal{L}(X'X_2) = W_{k-p}(n, \Sigma_{22})$, thereby proving part (iii) of the theorem. Equation (2.3.4) shows that

$$\mathcal{L}(X'X_1|X_2) = \mathcal{N}_{(k-p)\times p}(S_{22}^\Sigma_{22}^+ \Sigma_{21}, S_{22} \otimes \Sigma_{11,2}),$$

so that the distribution of $X'X_1|X_2$ depends on $X_2$ only through $S_{22}$. Consequently,

$$\mathcal{L}(S_{21}|S_{22}) = \mathcal{L}(X'X_1|X_2) = \mathcal{N}_{(k-p)\times p}(S_{22}^\Sigma_{22}^+ \Sigma_{21}, S_{22} \otimes \Sigma_{11,2}),$$

proving the last assertion of the theorem. The matrix $W$ as defined in (2.3.6) now follows a central Wishart distribution, that is, $W \sim W_p'(n - r, \Sigma_{11,2})$, and equation (2.3.5) gives

$$\mathcal{L}(X'QX_1|X_2 = N) = W_p'(n - r, \Sigma_{11,2}).$$

(2.3.7)
Let $\hat{S}_p^+$ denote the set of all symmetric positive semi-definite $p \times p$ matrices. Then, for any measurable set $A$ of $\hat{S}_p^+$, we have

\[
P\{X_1'QX_1 \in A\} = \mathbb{E}P\{X_1'QX_1 \in A|X_2\}
\]

\[
= \mathbb{E}P\{X_1'QX_1 \in A|X_2\}I_B(X_2)
\]

\[
= \mathbb{E}P\{W \in A\}I_B(X_2)
\]

\[
= P\{W \in A\}P\{X_2 \in B\}
\]

\[
= P\{W \in A\}.
\]

Thus, $L(X_1'QX_1) = LW = W_p(n-r, \Sigma_{11.2})$, and since $X_1'QX_1 = S_{11.2}$, we also proved (ii) of the theorem. Finally, to establish the first statement, let $A$ be any measurable subset of $\hat{S}_p^+$ and $F$ be any measurable subset of the set of all $n \times k$ matrices. From (2.3.3) and proposition 2.1.4 we know that $PX_1$ and $QX_1$ are conditionally independent given $X_2$, so $X_1'QX_1$ and $PX_1$ are conditionally independent given $X_2$. Furthermore, we have just seen that the distribution of $X_1'QX_1$ is $W_p(n-r, \Sigma_{11.2})$, independently distributed of $X_2$. Therefore,

\[
P\{X_1'QX_1 \in A, (PX_1, X_2) \in F\} = \mathbb{E}P\{X_1'QX_1 \in A, (PX_1, X_2) \in F|X_2\}
\]

\[
= \mathbb{E}P\{X_1'QX_1 \in A|X_2\} \times
\]

\[
\times P\{(PX_1, X_2) \in F|X_2\}I_B(X_2)
\]

\[
= P\{W \in A\} \mathbb{E}P\{(PX_1, X_2) \in F|X_2\}I_B(X_2)
\]

\[
= P\{X_1'QX_1 \in A\}P\{(PX_1, X_2) \in F\},
\]

showing that $X_1'QX_1$ and $(PX_1, X_2)$ are independently distributed. But then we are finished, since this implies that $S_{11.2}$ and $(S_{21}'PX_1, S_{22}'X_2) = (S_{21}, S_{22})$ are independent.

The above proof is similar to the proof of Eaton (1983), yet contains some subtleties because of the possibility of singularity of the covariance matrix. By considering the proof of theorem 2.3.2 step by step, it is easy to show where things go wrong if we consider the noncentral Wishart distribution. Equation (2.3.2) is a straightforward generalization of part (iii) of theorem 2.3.2. As for part (iv) of the theorem, it is clear from (2.3.4) that, as opposed to the case of the central Wishart distribution, $L(X_2'X_1|X_2)$ depends both on $X_2$ and $X_2'X_2$. Therefore, it is not straightforward to say something about the conditional distribution of $S_{21}$ given $S_{22}$. Finally, it follows from (2.3.5) that the distribution of $S_{11.2}$ does depend on $X_2$ via $Q$, again in contrast with the central case, complicating the generalization of (i) and (ii).
of theorem 2.3.2. These arguments show that we are not able to give a complete generalization of theorem 2.3.2 to the noncentral case. In the following, we will consider a special case where this is possible.

Suppose that \( S \sim W_k^\prime(n, \Sigma; \Lambda) \), where \( \Lambda = M'M \) with \( M = (M_1, M_2) \) and \( M_1 = M_2 \Sigma_{22}^{-1} \Sigma_{21} \). That is, \( M_1 \) is equal to its regression on \( M_2 \). Moreover, since the regression function of \( X_1 \) on \( X_2 \) is given by

\[
E(X_1|X_2) = M_1 + (X_2 - M_2) \Sigma_{22}^{-1} \Sigma_{21},
\]

which equals \( X_2 \Sigma_{22}^{-1} \Sigma_{21} \) if \( M_1 = M_2 \Sigma_{22}^{-1} \Sigma_{21} \), this relationship between \( M_1 \) and \( M_2 \) means that the regression of \( X_1 \) on \( X_2 \) contains no intercept term. Let \( r = \text{rank}(\Sigma_{22} + M_2' M_2) \). Equation (2.3.2) shows that

\[
L(S_{22}) = W_{k-p}(n, \Sigma_{22}; M_2' M_2).
\]

It follows from equation (2.3.4) that

\[
L(X_2'X_1|X_2) = N_{(k-p)\times p}(S_{22} \Sigma_{22}^{-1} \Sigma_{21}, S_{22} \otimes \Sigma_{11,2}),
\]

so that \( L(X_2'X_1|X_2) \) depends on \( X_2 \) only through \( S_{22} = X_2'X_2 \), and therefore \( L(S_{21}|S_{22}) = L(X_2'X_1|X_2) \). As for the distribution of \( S_{11,2} \), we know from (2.3.5) that

\[
L(X_1'QX_1|X_2 = N) = W_p(n-r, \Sigma_{11,2}).
\]

Analogous to the proof of theorem 2.3.2, we can show that the unconditional distribution of \( X_1'QX_1 \) is also \( W_p(n-r, \Sigma_{11,2}) \). The independence of \( S_{11,2} \) and \( (S_{21}, S_{22}) \) follows in the same way as in the proof of theorem 2.3.2. This proves the following theorem.

**Theorem 2.3.3.** Consider \( S \sim W_k^\prime(n, \Sigma; \Lambda) \) where \( \Lambda = M'M \) with \( M = (M_1, M_2) \) and \( M_1 = M_2 \Sigma_{22}^{-1} \Sigma_{21} \). Partition \( S \) and \( \Sigma \):

\[
S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}, \text{ and } \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},
\]

where \( S_{11} \) and \( \Sigma_{11} \) are \( p \times p \) matrices. Let \( r = \text{rank}(\Sigma_{22} + M_2' M_2) < n \) and \( p < k \). We then have

(i) \( S_{11,2} \) and \( (S_{21}, S_{22}) \) are independently distributed,
(ii) \( L(S_{11,2}) = W_p(n-r, \Sigma_{11,2}) \),
(iii) \( L(S_{22}) = W_{k-p}(n, \Sigma_{22}; M_2' M_2) \),
(iv) \( L(S_{21}|S_{22}) = N_{(k-p)\times p}(S_{22} \Sigma_{22}^{-1} \Sigma_{21}, S_{22} \otimes \Sigma_{11,2}) \).
We now give two examples of the case described by theorem 2.3.3.

Example 2.3.1. A special case of the noncentral Wishart distribution which is known in the literature is the case where \( \Lambda \) is of rank one (e.g., exercise 10.11 of Muirhead, 1982, p. 519). Consider the random matrix \( X \sim \mathcal{N}_{n \times k}(M, I_n \otimes I_k) \), so that \( S = X'X \sim W_k'(n, I_k; \Lambda) \), with \( \Lambda = \text{diag}(0, \ldots, 0, m') \). Partition \( S \) as

\[
S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix},
\]

where \( s_{22} \neq 0 \) is real-valued, \( S_{12} = S_{21}' \) is a random \((k - 1) \times 1\) vector, and \( S_{11} \) is a random \((k - 1) \times (k - 1)\) matrix. Put \( S_{11.2} = S_{11} - S_{12}S_{22}^{-1}S_{21} \). It immediately follows from theorem 2.3.3 that

- \( S_{11.2} \sim W_{k-1}(n - 1, I_{k-1}) \) and \( S_{11.2} \) is independent of \( S_{21} \) and \( s_{22} \),
- \( L(S_{21}|s_{22}) = \mathcal{N}_{k-1}(0, s_{22}I_{k-1}) \),
- \( L(s_{22}) = W_1'(n, 1, m'm) = \chi^2_n(\delta) \), with \( \delta = m'm \).

Example 2.3.2. Consider \( X \sim \mathcal{N}_{n \times k}(M, I_n \otimes \Sigma) \), where \( M = (M_1, M_2) = (0, M_2), M_2 \) is an \( n \times (k - p) \) matrix, and \( \Sigma_{21} = \Sigma_{12}' = 0 \), so that \( S \sim W_k'(n, \Sigma; \Lambda) \), with

\[
\Sigma = \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix} \quad \text{and} \quad \Lambda = \begin{pmatrix} 0 & 0 \\ 0 & M_2'M_2 \end{pmatrix}.
\]

Let \( r = \text{rank}(\Sigma_{22} + M_2'M_2) \). This case is slightly more general than the rank-one case of example 2.3.1. If we apply theorem 2.3.3, then (i) and (iii) hold for this case, while (ii) and (iv) simplify to

\[
L(S_{11,2}) = W_p(n - r, \Sigma_{11})
\]

respectively

\[
L(S_{21}|S_{22}) = \mathcal{N}_{(k-p) \times p}(0, S_{22} \otimes \Sigma_{11}).
\]

2.4 Towards the next section

Up to now, we have been mainly concerned with the normal distribution. Whereas normality is a very popular assumption when modeling or analyzing data, situations arise in which this is not appropriate. In financial data, for example, it often happens that there are some outliers in the data, or that the tails of the distribution are much heavier, so that normality is not realistic anymore. So, within the framework of robustness, it is interesting to
examine what happens if heavier-tailed distributions are allowed. The normality assumption is rather strong. In statistical applications with data that appear to have heavier tails, a researcher may look for a distribution that reflects certain characteristics of the normal distribution, but has heavier tails. Therefore, distributions that have elliptically shaped contours like a normal distribution are of interest. This leads to studying the class of spherical and elliptical distributions, see, for example, the textbooks by Fang and Zhang (1990) and Fang et al. (1990), and the survey by Chmielewski (1981).

A spherical distribution is an extension of the multivariate normal distribution with zero mean and identity covariance matrix, \( \mathcal{N}_m(0, I_m) \), whereas an elliptical distribution is an extension of the multivariate normal distribution with arbitrary mean and covariance matrix, \( \mathcal{N}_m(\mu, \Sigma) \). Many properties of the multivariate normal distribution are generated by those of \( \mathcal{N}_m(0, I_m) \). Similarly, characteristics of elliptical distributions can be derived from the properties of spherical distributions. Therefore, we only study the class of spherical distributions in the next section. In section 2.5, we first discuss some elementary concepts and findings in the theory of spherical distributions. We became especially interested in the work by Schoenberg (1938) who gave some fundamental results in the theory on spherical distributions. This theorem is our main focus in section 2.5. A new and shorter proof of his main theorem in this field will be given using elementary techniques from probability theory. The normal distribution occupies an important place in this theorem.

2.5 Schoenberg (1938) revisited

The central topic of this section is the work by Schoenberg (1938) on the class of spherical distributions. Subsection 2.5.1 gives some well-known preliminaries about the uniform distribution on the unit sphere in \( \mathbb{R}^m \). This distribution plays an essential role in the analysis of spherical distributions. We focus on the characteristic function of such a distribution, because it is useful in subsection 2.5.2. Subsection 2.5.2 considers the class of spherical distributions. A well-known result is that every spherical random vector can be decomposed into the product of two independent components: a random vector uniformly distributed on the unit sphere, and the length of this random vector. We consider this stochastic representation as a byproduct of the characterization of the class of characteristic functions belonging to spherical distributions. Subsection 2.5.2 concludes with a very
important theorem by Schoenberg (1938), which is of major importance in the literature on spherical distributions. In subsection 2.5.3, we give an alternative proof of this theorem we think is more elegant and less complicated. The subsections 2.5.1–2.5.3 are based upon Steerneman and Van Perlo-Ten Kleij (forthcoming, b).

Although exchangeability and sphericity are topics that can be studied individually, they are also naturally related. Subsection 2.5.4 gives a brief introduction to the concept of exchangeability, followed by an alternative proof of Schoenberg’s famous theorem in the context of exchangeability in subsection 2.5.5. This alternative proof given by Kingman (1972) suggests that there is some sort of relation between exchangeability and sphericity. This relationship will be addressed in subsection 2.5.6.

2.5.1 Uniform distribution on the unit sphere

In this subsection, our main interest is to characterize spherical distributions by means of their characteristic functions (c.f.). We first present some well-known results which will prove their usefulness in subsection 2.5.2, see, e.g., Müller (1966).

Suppose \( f \) is integrable on \( \mathbb{R}^m \) and we want to calculate \( \int_{\mathbb{R}^m} f(x) \, dx \). It is sometimes convenient to make a transformation to polar coordinates:

\[
\begin{align*}
    x &= ru, \quad r > 0, \quad u \in S^{m-1} = \{ u \in \mathbb{R}^m \mid \| u \| = 1 \}, \\
    u_1 &= \sin \theta_1 \sin \theta_2 \cdots \sin \theta_{m-2} \sin \theta_{m-1}, \\
    u_2 &= \sin \theta_1 \sin \theta_2 \cdots \sin \theta_{m-1} \cos \theta_{m-1}, \\
    u_3 &= \sin \theta_1 \sin \theta_2 \cdots \cos \theta_{m-2}, \\
    &\vdots \\
    u_{m-1} &= \sin \theta_1 \cos \theta_2, \\
    u_m &= \cos \theta_1,
\end{align*}
\]

where \( 0 < \theta_i \leq \pi, i = 1, \ldots, m - 2 \) and \( 0 < \theta_{m-1} \leq 2\pi \). The Jacobian of this transformation is

\[
dx = r^{m-1} \sin^{m-2} \theta_1 \sin^{m-3} \theta_2 \cdots \sin \theta_{m-2} \, dr \, d\theta_1 \cdots d\theta_{m-1}.
\]

Let \( \omega \) denote the unique measure on \( S^{m-1} \) which is invariant under orthogonal transformations, such that \( \omega(S^{m-1}) \) equals the area of \( S^{m-1} \) (see, e.g.,
Then
\[ d \mathbf{x} = r^{m-1} \, dr \, d\omega(u), \]
\[ d\omega(u) = \sin^{m-2} \theta_1 \sin^{m-3} \theta_2 \cdots \sin \theta_{m-2} \, d\theta_1 \cdots d\theta_{m-1}. \]

The area of \( S^{m-1} \) is
\[
\omega_m = \omega(S^{m-1})
= \int_{S^{m-1}} d\omega(u)
= \int_0^\pi \cdots \int_0^\pi \int_0^{2\pi} \sin^{m-2} \theta_1 \cdots \sin \theta_{m-2} \, d\theta_1 \cdots d\theta_{m-1}
= 2\pi^{m/2} \prod_{k=1}^{m-2} \left[ \int_0^\pi \sin^k \theta \, d\theta \right]. \tag{2.5.1}
\]

Define \( I_k = \int_0^\pi \sin^k \theta \, d\theta \). Substitution of \( t = \sin \theta \) gives
\[
I_k = 2 \int_0^1 t^k (1 - t^2)^{-1/2} \, dt
\]
and setting \( t^2 = x \) we find
\[
I_k = \int_0^1 x^{(k-1)/2} (1 - x)^{-1/2} \, dx
= \frac{\sqrt{\pi} \, \Gamma \left( \frac{k+1}{2} \right)}{\Gamma \left( \frac{k+2}{2} \right)}.
\]

Equation (2.5.1) now gives
\[
\omega_m = 2\pi^{m/2} \prod_{k=1}^{m-2} \frac{\Gamma \left( \frac{k+1}{2} \right)}{\Gamma \left( \frac{k+2}{2} \right)} = \frac{2\pi^{m/2}}{\Gamma \left( \frac{m}{2} \right)}. \tag{2.5.2}
\]

Let \( U \) be uniformly distributed on the \( m \)-dimensional unit sphere \( S^{m-1} \), which we will denote by \( U \sim U(S^{m-1}) \). Suppose \( f \) is integrable on \( S^{m-1} \) with respect to \( \omega \), then we have
\[
E f(U) = \frac{1}{\omega_m} \int_{S^{m-1}} f(u) \, d\omega(u).
\]

We will now derive the characteristic function of \( U \). Because the distribution of \( U \) is invariant under orthogonal transformations, we know that \( E e^{it' \cdot U} = E e^{it' \cdot C U} \) for all \( C \in O(m) \), \( t' \in \mathbb{R}^m \). Therefore, this c.f. is invariant under the
action of the group of all orthogonal $m \times m$ matrices, and hence it depends on $t$ through $\|t\|$, where $\|\cdot\|$ denotes the usual Euclidean norm. This implies that the c.f. of $U$ must be a function of $\|t\|$, so that we can write $Ee^{it'U} = \Omega_m(\|t\|)$ (see, e.g., Fang and Zhang, 1990). Moreover,

$$\Omega_m(\|t\|) = \frac{1}{\omega_m} \int_{S_{m-1}} e^{it'u} d\omega(u) = \frac{1}{\omega_m} \int_{S_{m-1}} e^{i\|t\|u} d\omega(u),$$

where we have used the fact that $\Omega_m(\|t\|) = \Omega_m(\|C't\|)$ for all $C \in \Theta(m)$ and in particular for $C = (c_1, c_2, \ldots, c_{m-1}, \|t\|^{-1}t)$, where the columns of this matrix are chosen to be mutually orthogonal and have unit length. Deriving the characteristic function of $U$ therefore comes down to calculating

$$\Omega_m(r) = \frac{1}{\omega_m} \int_{S_{m-1}} e^{iru} d\omega(u) = \frac{1}{\omega_m} \int_{0}^{\pi} \cdots \int_{0}^{\pi} \int_{0}^{2\pi} e^{ir\cos \vartheta_1 \sin^{m-2} \vartheta_1 \cdots \sin \theta_{m-2}} \cos \vartheta_1 \cdots d\vartheta_{m-2} d\theta_{m-1}$$

$$= \frac{\int_{0}^{\pi} e^{ir\cos \vartheta_1 \sin^{m-2} \vartheta_1} d\vartheta_1}{\int_{0}^{\pi} \sin^{m-2} \vartheta_1 d\vartheta_1} = \frac{\int_{0}^{\pi} \cos(r \cos \theta) \sin^{m-2} \theta d\theta}{\sqrt{\pi} \Gamma \left( \frac{m-1}{2} \right) / \Gamma \left( \frac{m}{2} \right)}.$$  (2.5.3)

The Bessel function of the first kind is defined by (see, e.g., Abramowitz and Stegun, 1972)

$$J_\nu(z) = \frac{(z/2)^\nu}{\sqrt{\pi} \Gamma \left( \nu + \frac{1}{2} \right)} \int_{0}^{\pi} \cos(z \cos \theta) \sin^{2\nu} \theta d\theta.$$

Furthermore,

$$J_\nu(z) = \frac{(z/2)^\nu}{\Gamma(\nu + 1)} \left[ \nu + 1, -z^2/4 \right]_0,$$

where the function $\left[ \nu + 1, -z^2/4 \right]_0$ is defined as follows:

$$\left[ \nu + 1, -z^2/4 \right]_0 = \sum_{j=0}^\infty \frac{1}{(\alpha)_j} \frac{x^j}{j!} \quad (\alpha \neq 0, -1, -2, \ldots),$$

with $(\alpha)_j = \alpha(\alpha+1) \cdots (\alpha+j-1)$. This yields the following expression:

$$\int_{0}^{\pi} \cos(z \cos \theta) \sin^{2\nu} \theta d\theta = \frac{\sqrt{\pi} \Gamma \left( \nu + \frac{1}{2} \right)}{\Gamma(\nu + 1)} \left[ \nu + 1, -z^2/4 \right]_0.$$  (2.5.4)
The characteristic function of $U$ can now be expressed as

$$\Omega_m(r) = {}_0F_1\left(m/2, -r^2/4\right) \quad (m \geq 1).$$

So in the end, we obtain the following relatively easy expression for $\Omega_m$:

$$\Omega_m(r) = 1 - \frac{r^2}{2m} + \frac{r^4}{2 \cdot 4 \cdot m(m + 2)} - \frac{r^6}{2 \cdot 4 \cdot 6 \cdot m(m + 2)(m + 4)} + \cdots$$

for $m \geq 1$. With regard to the history of the expression of $\Omega_m$ in terms of Bessel functions, we refer to footnote 5 of Hartman and Wintner (1940). Schoenberg (1938) also obtained these results. For an alternative derivation we also refer to Fang et al. (1990).

2.5.2 Spherical distributions and Schoenberg's theorem

If $X \sim \mathcal{N}_m(0, I_m)$, then the density function of $X$ is constant on unit spheres with the origin as the center. The distribution of $X$ is invariant under the action of the orthogonal group: $\mathcal{L}(X) = \mathcal{L}(CX)$ for all $C \in O(m)$. Inspired by this notion, we give the following definition of a spherical distribution.\(^2\)

In the literature, these distributions are also called spherically symmetric or spherically contoured.

**Definition 2.5.1.** The set of spherically contoured probability distributions on $\mathbb{R}^m$ consists of those probability distributions that are invariant under orthogonal transformations and is denoted by $SC_m$.

Assume that $\mathcal{L}(X) \in SC_m$ and let $\psi$ denote the characteristic function of $X$. Note that

$$\mathcal{L}(X) \in SC_m \iff \mathcal{L}(X) = \mathcal{L}(CX) \text{ for all } C \in O(m)$$
$$\iff E e^{itX} = E e^{itCX} \text{ for all } C \in O(m), t \in \mathbb{R}^m$$
$$\iff \psi(t) = \psi(Ct) \text{ for all } C \in O(m), t \in \mathbb{R}^m.$$

So $X$ has a spherical probability distribution if and only if its characteristic function is invariant under orthogonal transformations. If $m = 1$, for example, then we have $\psi(t) = E e^{itX} = E e^{-itX}$, hence $\psi(t) = \psi(-t) = E \cos(tX)$ and $\mathcal{L}(X) = \mathcal{L}(-X)$. Note that, because of invariance, we can write $\psi(t) = \phi(||t||)$, that is, the c.f. of $X$ depends on $t$ only through $||t||$ and the function $\phi$ is a function of a scalar variable. By taking $\phi(-s) = \phi(s)$ for $s \geq 0$ we see

---

\(^2\)The class of spherical distributions can be defined in a number of equivalent ways, see, for example, the discussion in Fang et al. (1990).
that \( \psi \) defines a characteristic function for a probability distribution on \( \mathbb{R} \). The converse will not be true in general: if \( \phi \) is a characteristic function on \( \mathbb{R} \), then we do not necessarily have that \( \psi(t) = \phi(\|t\|) \) for \( t \in \mathbb{R}^m \) is a c.f. on \( \mathbb{R}^m \). However, if \( \psi \) is the characteristic function of a spherically distributed random variable on \( \mathbb{R}^{m+k} \), then \( \psi(t,0) \) with \( t \in \mathbb{R}^m \) is the characteristic function of a spherically distributed random variable on \( \mathbb{R}^m \).

**Definition 2.5.2.** Let \( \phi \) be a symmetric characteristic function on \( \mathbb{R} \) such that \( \phi(\|t\|) \) for \( t \in \mathbb{R}^m \) is the characteristic function of an \( SC_m \) distribution, then we say that \( \phi \in \Phi_m \).

Obviously, \( \phi \in \Phi_{m+1} \) implies \( \phi \in \Phi_m \). If \( \phi \in \Phi_k \) for all \( k \), we say that \( \phi \in \Phi_\infty \). Therefore, \( \Phi_1 \supset \Phi_2 \supset \cdots \supset \Phi_\infty \). Schoenberg (1938) gave the following representation for characteristic functions in \( \Phi_m \).

**Theorem 2.5.1.** The random \( m \times 1 \) vector \( X \) has a spherical distribution with c.f. \( \psi(t) = \phi(\|t\|) \) if and only if

\[
\phi(\|t\|) = \int_0^\infty \Omega_m(\|t\|y) dQ(y)
\]

(2.5.6)

for some probability measure \( Q \) on \([0, \infty) \) (\( Q \) is in fact the probability distribution of \( \|X\| \)).

**Proof.** Assume \( \mathcal{L}(X) \in SC_m \) with c.f. \( \psi(t) = \phi(\|t\|) \). Let \( U \) be uniformly distributed on the unit sphere \( S^{m-1} \), independent of \( X \). First note that

\[
\phi(\|t\|) = \psi(t) = E \psi(\|t\|U) = E e^{it't'X} = E e^{it't'U} = E e^{it'R}\Omega_m(\|t\|R) = E e^{it'R} \int_0^\infty \Omega_m(\|t\|y) dQ(y),
\]

where we have used the fact that \( \psi(t) \) is constant on unit spheres, that is \( \psi(t) = \psi(\|t\|u) \) for all \( u \) with \( \|u\| = 1 \).

Conversely, the right-hand side of (2.5.6) is the c.f. of an \( SC_m \) distribution. This is easily seen if we define \( X = RU \) where \( R \geq 0 \) has probability distribution \( Q \) and \( U \sim \mathcal{U}(S^{m-1}) \) is independent of \( R \), because

\[
\int_0^\infty \Omega_m(\|t\|y) dQ(y) = E \Omega_m(\|t\|R) = E e^{it'R} = E e^{it'R}.
\]

which is the c.f. of \( X = RU \). Its distribution is obviously invariant under orthogonal transformations. \( \square \)
This theorem completely characterizes the family of spherical distributions on $\mathbb{R}^m$. Moreover, the theorem shows that $\mathcal{L}(X) \in SC_m$ if and only if $\mathcal{L}(X) = \mathcal{L}(RU)$, where $U \sim U(S^{m-1})$ and $R$ is a random variable on $[0, \infty)$ independent of $U$.

**Example 2.5.1.** If $X \sim \mathcal{N}_m(0, I_m)$, then it can be shown that $\mathcal{L}(X) = \mathcal{L}(RU)$, where $U \sim U(S^{m-1})$, and $\mathcal{L}(R) = \mathcal{L}(\|X\|) = \chi_n$ (the chi-distribution with $n$ degrees of freedom is the distribution followed by the square root of a chi-squared random variable).

The following theorem is also due to Schoenberg (1938) and shows that the c.f. $\phi$ of $X$ belongs to $\Phi_\infty$ if and only if $\mathcal{L}(X) \in SC_m$ is a scale mixture of normal distributions. An alternative proof in the context of exchangeability has been given by Kingman (1972) and a slightly adapted version of this proof can be found in Fang et al. (1990), also see subsection 2.5.5. We give an alternative proof of this theorem in subsection 2.5.3 we think is more elegant and less complicated. Donoghue (1969, page 201-206) already presented a simplified proof of the required global convergence, but it is still rather complicated and technical. Chapter 5 of Berg et al. (1984) is dedicated to an abstract form of Schoenberg’s theorem and generalizations. They used the concept of a Schoenberg triple. Their approach also leads to a simplification.

**Theorem 2.5.2 (Schoenberg, 1938).** The elements $\phi : \mathbb{R} \to [-1, 1]$ of $\Phi_\infty$ can be represented as

$$\phi(t) = \int_0^\infty e^{-t^2y^2/2}dQ(y) \quad (2.5.7)$$

where $Q$ is a probability measure on $[0, \infty)$.

### 2.5.3 An alternative proof of Schoenberg’s theorem

An important element in Schoenberg’s proof of theorem 2.5.2 is the behavior of $\Omega_m(r\sqrt{m})$ as $m \to \infty$. Pointwise convergence of $\Omega_m(r\sqrt{m})$ has been pointed out to Schoenberg by J. von Neumann (see footnote 12 of Schoenberg, 1938):

$$\lim_{m \to \infty} \Omega_m(r\sqrt{m}) = e^{-r^2/2}. \quad (2.5.8)$$

Hartman and Wintner (1940) attributed this result to Laplace, see their footnote 13. However, global uniform convergence of $\Omega_m(r\sqrt{m})$ is needed, that is, (2.5.8) must hold uniformly for all real values of $r$. This global uniform convergence is a key element in the proof, yet it is not easy to establish.
Schoenberg’s proof is quite complex and is organized in the form of three lemmas (Schoenberg, 1938, lemma 1 – lemma 3). Our proof of the theorem uses the basic ideas of Schoenberg, however, the crucial step of global uniform convergence of $\Omega_m(r/\sqrt{m})$ is proved by applying more modern probabilistic tools. While it took Schoenberg quite some effort to establish global uniform convergence of $\Omega_m(r/\sqrt{m})$, our argument is relatively short and, as we think, more transparent.

**Proof of theorem 2.5.2.** First, we will show that the c.f. $\phi \in \Phi_m$ for all $m$. Let $X_1, \ldots, X_m | Y$ be independently distributed with $X_i | Y \sim \mathcal{N}(0, Y^2)$ for $i = 1, \ldots, m$, and $\mathcal{L}(Y) = Q$. Define $X = (X_1, \ldots, X_m)'$, then $\mathcal{L}(X) \in SC_m$ and for its c.f. $\psi$ we obtain, letting $t \in \mathbb{R}^m$,

$$
\psi(t) = E e^{it'X} = E E(e^{it'X} | Y) = E e^{-\|t\|^2 Y^2/2} = \int_0^\infty e^{-\|t\|^2 y^2/2}dQ(y) = \phi(\|t\|).
$$

Hence, $\phi \in \Phi_m$ for all $m$.

Second, we have to prove that we can find such a representation for any $\phi \in \Phi_\infty$. Suppose $\phi \in \Phi_m$ for all $m$, then we can write

$$
\phi(t) = \int_0^\infty \Omega_m(ty/\sqrt{m})dF_m(y) \tag{2.5.9}
$$

for some probability distribution function $F_m$ on $[0, \infty)$. Now, if we let $m \to \infty$ in (2.5.9), it is tempting to exploit (2.5.8) to arrive at the representation (2.5.7). However, to apply Helly’s convergence theorem for distribution functions (e.g., Ash, 1972, theorem 8.2.1), pointwise convergence of $\Omega_m(r/\sqrt{m})$ is not sufficient. Because the interval of integration in (2.5.9) is infinite, we need global uniform convergence of $\Omega_m(r/\sqrt{m})$. As mentioned above, Schoenberg (1938) proved that $\Omega_m(r/\sqrt{m}) \to e^{-r^2/2}$ uniformly on $\mathbb{R}$ as $m \to \infty$. We shall give an alternative proof for this fact. We use the representation (2.5.3) for $\Omega_m$, that is

$$
\Omega_m(r) = \frac{\int_0^\pi e^{ir \cos \theta} \sin^{m-2} \theta \, d\theta}{\int_0^\pi \sin^{m-2} \theta \, d\theta}.
$$

On substituting $t = \cos \theta$ we obtain

$$
\Omega_m(r) = \frac{\Gamma \left( \frac{m}{2} \right)}{\sqrt{\pi} \Gamma \left( \frac{m-1}{2} \right)} \int_{-1}^1 e^{irt} (1 - t^2)^{(m-3)/2} \, dt.
$$
By making the transformation \( t = x / \sqrt{m} \), it is easy to see that \( \Omega_m (r \sqrt{m}) \) as a function of \( r \) is the c.f. corresponding to the probability density function

\[
f_m(x) = \frac{\Gamma \left( \frac{m}{2} \right)}{\sqrt{\pi m} \Gamma \left( \frac{m-1}{2} \right)} \left( 1 - \frac{x^2}{m} \right)^{(m-3)/2} I_{(-\sqrt{m}, \sqrt{m})}(x).
\]

On account of Stirling’s formula we know that

\[
\lim_{m \to \infty} \frac{\Gamma \left( \frac{m}{2} \right)}{\Gamma \left( \frac{m-1}{2} \right) \sqrt{m}} = 1.
\]

Moreover,

\[
\lim_{m \to \infty} \left( 1 - \frac{x^2}{m} \right)^{(m-3)/2} = e^{-x^2/2} \quad \text{for all } x.
\]

Therefore, \( f_m(x) \to f_\infty(x) = (2\pi)^{-1/2} e^{-x^2/2} \) for all \( x \), as \( m \to \infty \), where we immediately recognize the standard normal distribution. From Scheffé’s lemma, see, e.g., Billingsley (1968, page 224), it now follows that

\[
\lim_{m \to \infty} \int |f_m(x) - f_\infty(x)| \, dx = 0.
\]

However, convergence in \( L_1 \) of a sequence of probability density functions implies global uniform convergence of the corresponding sequence of characteristic functions. To see this, consider probability density functions \( g, g_1, g_2, \ldots \) with respect to a certain \( \sigma \)-finite measure \( \mu \), and let \( \phi, \phi_1, \phi_2, \ldots \) denote the associated c.f.’s. Let \( g_n \to g \) a.e. \( [\mu] \). Then

\[
|\phi_n(t) - \phi(t)| = \left| \int e^{itx} (g_n(x) - g(x)) \, d\mu(x) \right| \leq \int |g_n(x) - g(x)| \, d\mu(x).
\]

Thus, if we define \( \Omega_\infty(r) = e^{-r^2/2} \), the c.f. corresponding to \( f_\infty(x) \), then we have

\[
|\Omega_m (r \sqrt{m}) - \Omega_\infty(r)| \leq \int |f_m(x) - f_\infty(x)| \, dx. \tag{2.5.10}
\]

It now follows that \( \Omega_m (r \sqrt{m}) \to \Omega_\infty(r) \) uniformly in \( r \). Following the same reasoning as Schoenberg (1938), we know from Helly’s theorem that there exists a subsequence \( F_{m_k} \) of \( F_m \) such that \( F_{m_k} \to F \) weakly, where \( F \) is a probability distribution function. Now we write

\[
\phi(t) = \int_0^\infty \left( \Omega_m (ty \sqrt{m}) - e^{-t^2 y^2/2} \right) \, dF_m(y) + \int_0^\infty e^{-t^2 y^2/2} \, dF_m(y).
\]
As $m \to \infty$, we obtain from (2.5.10) that

\[
\left| \int_0^\infty \left( \Omega_m(ty \sqrt{m}) - e^{-t^2y^2/2} \right) dF_m(y) \right| \leq \int_{-\infty}^\infty \left| f_m(x) - f_\infty(x) \right| dx \to 0,
\]

and, from $F_{mk} \to F$ weakly, that

\[
\int_0^\infty e^{-t^2y^2/2} dF_{mk}(y) \to \int_0^\infty e^{-t^2y^2/2} dF(y).
\]

Putting things together, we find

\[
\phi(t) = \int_0^\infty e^{-t^2y^2/2} dF(y).
\]

Theorem 2.5.2 implies that a random vector $X$ has a c.f. $\in \Phi_\infty$ if and only if $X$ has the same distribution as $RZ$, where $Z \sim \mathcal{N}_n(0, I_n)$ is independent of $R$ having outcomes in $[0, \infty)$, that is, the distribution of $X$ is a scale mixture of normal distributions. Theorem 2.5.2 therefore also shows that a c.f. $\phi$ of a scalar variable defines a spherically contoured probability distribution for a random vector in every dimension if and only if this probability distribution is a scale mixture of normal distributions. Therefore, distributions with a c.f. in $\Phi_\infty$ inherit properties from the multivariate normal distribution so that results in $\Phi_\infty$ are more easily obtained than in $\Phi_m, m < \infty$. Once more, we experience the exceptional position of the normal distribution in the class of probability distributions.

### 2.5.4 Exchangeability

As already noticed, an alternative proof of theorem 2.5.2 has been given by Kingman (1972). The concept of exchangeability is essential in this proof. In this subsection, we give a short introduction to this topic. In subsection 2.5.5 we repeat Kingman’s proof of theorem 2.5.2 in somewhat more detail. This proof points out to the relationship between exchangeability and sphericity, a relationship we elaborate on in subsection 2.5.6.

One of the properties of a random vector $X = (X_1, \ldots, X_m)'$ which follows a spherical distribution is that it is permutation-symmetric, that is,

\[
L(X_1, \ldots, X_m)' = L(X_{\pi_1}, \ldots, X_{\pi_m})'
\]  

(2.5.11)

for each permutation $(\pi_1, \ldots, \pi_m)$ of $(1, \ldots, m)$ (Fang et al., 1990). Whereas property (2.5.11) deals with a finite sequence of random variables, exchangeability explicitly deals with an infinite sequence of random variables. From Tong (1990) we take the following definitions:
Definition 2.5.3. Let \( \{Y_i\}_{i=1}^{\infty} \) be an infinite sequence of univariate random variables. It is said to be a sequence of exchangeable random variables if, for every finite \( n \) and every permutation \( \pi = (\pi_1, \ldots, \pi_n) \) of \( (1, \ldots, n) \), the distribution of \( (Y_{\pi_1}, \ldots, Y_{\pi_n})' \) does not depend on \( \pi \).

Definition 2.5.4. \( X_1, \ldots, X_m \) are said to be exchangeable random variables if there exists a sequence of exchangeable random variables \( \{X_i^*\}_{i=1}^{\infty} \) such that \( (X_1, \ldots, X_m)' \) and \( (X_1^*, \ldots, X_m^*)' \) have the same distribution.

From definition 2.5.4, we observe that a finite sequence of random variables is exchangeable if it can be considered as a subset of an infinite sequence of exchangeable random variables. Exchangeability implies that the joint probability distribution function of any subset \( m \) of the random variables does not depend upon the subscript but only upon the number \( m \) (Loève, 1963). A random vector satisfying (2.5.11) is sometimes referred to as being finitely exchangeable, but to prevent confusion, we will avoid this term. Sphericity of the distribution of a random vector \( X \) requires that \( X \) is permutation-symmetric; exchangeability is a more restrictive requirement. We will show in the next subsection that sphericity does not necessarily imply exchangeability.

Whereas the joint density function and distribution function of exchangeable random variables must be permutation-symmetric, this argument cannot be reversed. To illustrate this, we take a very clear example from Kingman (1978), which is originally from Kendall (1967).

Example 2.5.2. If the distribution of a random vector \( X = (X_1, \ldots, X_n)' \) is permutation-symmetric, then \( \text{Var}(X_i) = \sigma^2 \) and \( \text{Cov}(X_i, X_j) = \rho \sigma^2 \) for \( 1 \leq i < j \leq n \), so that the covariance matrix of \( X \) equals \( \Sigma = (1 - \rho)\sigma^2 I_n + \rho \sigma^2 \iota_n \iota_n' \). If \( X_1, \ldots, X_n \) were exchangeable, there exists an infinite sequence \( \{X_i^*\}_{i=1}^{\infty} \) such that \( (X_1^*, \ldots, X_n^*)' \) and \( (X_1, \ldots, X_n)' \) are identically distributed. The covariance matrix of \( X^* \) is the same as that of \( X \). The eigenvalue of \( \Sigma \) corresponding to the eigenvector \( \iota_n \) equals \( \sigma^2 [\rho (n - 1) + 1] \) which is negative if \( \rho < -1/(1 - n) \). We see that \( \Sigma \) cannot be a covariance matrix if \( \rho < 0 \), since \( \Sigma \) must define a covariance matrix for all \( n \). Therefore, \( X_1, \ldots, X_n \) cannot be exchangeable if \( \rho < 0 \).

A natural question that arises is the following. If a random vector \( X = (X_1, \ldots, X_m)' \) is permutation-symmetric, when are \( X_1, \ldots, X_m \) exchangeable? The answer to this question is given by de Finetti’s theorem (see, e.g., Feller, 1971, or Loève, 1963).
**Theorem 2.5.3.** If the distribution of \( X = (X_1, \ldots, X_m)' \) is permutation-symmetric, then the random variables \( X_1, \ldots, X_m \) are exchangeable if and only if \( X_1, \ldots, X_m \) are conditionally independent and identically distributed random variables (i.e., there is a \( \sigma \)-algebra conditional on which the \( X_i \) are independently and identically distributed).

An interesting motivation for exchangeability originates from a Bayesian perspective (Tong, 1990). In the Bayesian approach, it is often assumed that the model for a population variable is one of a family of probability density functions, \( \{f(x|\theta)\} \), where \( \theta \) is the vector of parameters of interest. Prior beliefs about \( \theta \) are expressed by means of the prior density \( p(\theta) \). Thus, if an independent sample \( X_1, \ldots, X_m \) is drawn from this population, it follows that, for given \( \theta \), \( X_1, \ldots, X_m \) are independent and identically distributed and therefore exchangeable. We see that exchangeability emerges naturally in Bayesian inference.

Olshen (1974) on the other hand, remarked that de Finetti’s theorem can also be interpreted conversely. That is, an infinite, exchangeable sequence of random variables is distributed as if a probability distribution on the shape and range of the data were chosen at random, and then independent, identically distributed data subsequently were generated according to the chosen probability distribution. More practical applications of exchangeable variables can be found in Tong (1990). A nice exposition on the uses of exchangeability is Kingman (1978).

**2.5.5 A proof of Schoenberg’s theorem in the context of exchangeability**

In a compact paper by Kingman (1972), theorem 2.5.2 was proved in an alternative way. The theorem was also phrased somewhat differently:

**Theorem 2.5.2 (Schoenberg, 1938, rephrased).** Let \( X_1, X_2, \ldots \) be an infinite sequence of real random variables with the property that, for any \( n \), the distribution of the random vector \( (X_1, \ldots, X_n)' \) has spherical symmetry. Then there is a random variable \( V \), real and nonnegative, such that, conditional on \( V \), the \( X_i \) are independent and normally distributed, with zero mean and variance \( V \).

For the sake of completeness, we include this proof in somewhat more detail.
Proof. The proof given by Kingman is as follows. If \((\pi_1, \ldots, \pi_n)\) is any permutation of \((1, \ldots, n)\), then the distribution of \(X = (X_1, \ldots, X_n)'\) is the same as the distribution of \((X_{\pi_1}, \ldots, X_{\pi_n})'\) for all \(n\), because of spherical symmetry. According to definition 2.5.4, this implies that the sequence \(\{X_i\}_{i=1}^\infty\) is exchangeable, and de Finetti's theorem (theorem 2.5.3) shows there exists a \(\sigma\)-field \(\mathcal{F}\), conditional upon which the \(X_j\) are independent and have the same distribution function \(F\), say. Let

\[
\Phi(t) = \int_{-\infty}^\infty e^{itx} dF(x) = E(e^{itX_i}|\mathcal{F}),
\]

(2.5.12)

so that \(\Phi\) is a random, \(\mathcal{F}\)-measurable, continuous function and

\[
\Phi(-t) = \Phi(t), \quad |\Phi(t)| \leq 1, \quad \Phi(0) = 1.
\]

(2.5.13)

Let \(\psi(t), t \in \mathbb{R}^n\), denote the c.f. of \(X\). Because the \(X_1, \ldots, X_n\) are conditionally independent,

\[
E\{\psi(t)|\mathcal{F}\} = E\left\{\exp\left(i \sum_{j=1}^n t_jX_j\right) \mid \mathcal{F}\right\} = \prod_{j=1}^n \Phi(t_j),
\]

(2.5.14)

for all real \(t_1, \ldots, t_n\). Therefore,

\[
\psi(t) = E\left\{\exp\left(i \sum_{j=1}^n t_jX_j\right)\right\} = E\left\{\prod_{j=1}^n \Phi(t_j)\right\}.
\]

(2.5.15)

Spherical symmetry of \(X\) implies that \(\psi(t) = \phi(||t||)\), so that the left-hand side of (2.5.15), and therefore also the right-hand side, only depends on \(t_1^2 + \cdots + t_n^2\). Consider the case \(n = 2\). For any real \(u\) and \(v\), write \(t = \sqrt{u^2 + v^2}\), and use (2.5.13) and (2.5.15) to compute

\[
E|\Phi(t) - \Phi(u)\Phi(v)|^2 = E\{\Phi(t) - \Phi(u)\Phi(v)\} [\Phi(-t) - \Phi(-u)\Phi(-v)]
\]

\[
eq E\Phi(t)\Phi(-t) - E\Phi(t)\Phi(-u)\Phi(-v) +
\]

\[
- E\Phi(u)\Phi(v)\Phi(-t) + E\Phi(u)\Phi(v)\Phi(-u)\Phi(-v).
\]

The four terms in this last expression are all of the form of the right-hand side of (2.5.15), so that

\[
E\Phi(t)\Phi(-t) = \phi(2t^2)
\]

\[
E\Phi(t)\Phi(-u)\Phi(-v) = \phi(t^2 + u^2 + v^2) = \phi(2t^2)
\]

\[
E\Phi(u)\Phi(v)\Phi(-t) = \phi(u^2 + v^2 + t^2) = \phi(2t^2)
\]

\[
E\Phi(u)\Phi(v)\Phi(-u)\Phi(-v) = \phi(2u^2 + 2v^2) = \phi(2t^2).
\]
Therefore,
\[ E[\Phi(t) - \Phi(u)\Phi(v)]^2 = 0 \]
or
\[ P[\Phi(t) = \Phi(u)\Phi(v)] = 1 \]
So the function \( \Phi \) satisfies, with probability one, for each pair \((u, v)\) the functional equation
\[ \Phi(\sqrt{u^2 + v^2}) = \Phi(u)\Phi(v). \quad (2.5.16) \]
However, we cannot directly conclude that this functional equation holds with probability one for all \( u \) and \( v \), because the union of the null spaces on which this equality does not hold may have nonzero measure. To be more explicit, remember that the random variables \( X_j \) are measurable functions defined on a measure space \((\Omega, \mathcal{F}, P)\), and thus depend on \( \omega \in \Omega \). We know that
\[ P\left\{ \Phi(\sqrt{u^2 + v^2}) = \Phi(u)\Phi(v) \right\} = 1, \quad (2.5.17) \]
where we have to keep in mind that in fact \( \Phi(u) = \Phi(u, \omega) \). Define
\[ N_{u,v} = \{ \omega \mid \Phi(\sqrt{u^2 + v^2}) \neq \Phi(u)\Phi(v) \}. \]
Because of (2.5.17), \( P(N_{u,v}) = 0 \) for all pairs \((u, v)\), so in particular for all pairs \((u, v) \in \mathbb{Q}\). Define
\[ N = \bigcup_{u,v \in \mathbb{Q}} N_{u,v}, \]
then
\[ P(N) \leq \sum_{u,v \in \mathbb{Q}} P(N_{u,v}) = 0, \]
because the union of countably many null sets is also a null set. Therefore,
\[ P\left\{ \Phi(\sqrt{u^2 + v^2}) = \Phi(u)\Phi(v), \forall u, v \in \mathbb{Q} \right\} = 1. \]
Now suppose that \( \omega \notin N \). Then, with probability one,
\[ \Phi(\sqrt{u^2 + v^2}, \omega) = \Phi(u, \omega)\Phi(v, \omega) \]
for all \( u, v \in \mathbb{Q} \), and hence by continuity for all real \( u \) and \( v \). Now we can finally conclude that (2.5.16) holds for all \( u \) and \( v \), except possibly on a set with measure zero. From Feller (1971, Chapter III) we can conclude that
\[ \Phi(t) = e^{-\frac{1}{2}vt^2} \quad (2.5.18) \]
for some complex $V$, and (2.5.13) shows that $V$ is real and nonnegative. Since $V = -2 \log \Phi(1)$, $V$ is an $\mathcal{F}$-measurable random variable, so that

$$E(Z|V) = E[E(Z|\mathcal{F})|V]$$

for any random variable $Z$. Setting

$$Z = \exp \left( i \sum_{j=1}^{n} t_j X_j \right)$$

and using (2.5.14), we obtain

$$E \left\{ \exp \left( i \sum_{j=1}^{n} t_j X_j \right) \bigg| V \right\} = E \left\{ \prod_{j=1}^{n} e^{-\frac{1}{2}Vt_j^2} \bigg| V \right\} = \prod_{j=1}^{n} e^{-\frac{1}{2}Vt_j^2},$$

Hence, conditional on $V$, the $X_j$ are independent, with distribution $\mathcal{N}(0, V)$, and the proof is complete. \hfill \Box

2.5.6 Sphericity and exchangeability

Although sphericity and exchangeability can be studied individually, we now pay attention to the relationship between these two concepts. Kingman’s proof of theorem 2.5.2, explained in subsection 2.5.5, already touches upon this relationship. In this subsection, some questions will be answered concerning which conclusions can and which conclusions cannot be drawn regarding the link between exchangeability and sphericity.

We know from theorem 2.5.2 that a c.f. $\phi$ corresponding to a spherically distributed random vector $X \in \mathbb{R}^m$ belongs to the class $\Phi_\infty$, if and only if it is the c.f. corresponding to a scale mixture of normal distributions, that is, if and only if

$$\phi(t) = \int_{0}^{\infty} e^{-t^2 y^2/2} dQ(y)$$

for some probability measure $Q$ on $[0, \infty)$. In the proof of this theorem we saw that this is equivalent to saying that, given a random variable $Y$, the $X_1, \ldots, X_m$ are independently distributed, with $X_i|Y \sim \mathcal{N}(0, Y^2), i = 1, \ldots, m$ and $L(Y) = Q$. Thus, it immediately follows from theorem 2.5.3 that if $X$ has a c.f. $\phi$ which belongs to $\Phi_\infty$, then $X_1, \ldots, X_m$ are exchangeable. This statement can obviously not be reversed. A trivial example is $X \sim \mathcal{N}_m(\mu m, \Sigma)$ with $\mu \in \mathbb{R}$ and $\Sigma = (1 - \rho)\sigma^2 I_m + \rho \sigma^2 I_m' I_m$. It is easy
to show that if $\rho > 0$, then $X_1, \ldots, X_m$ are exchangeable (e.g., Tong, 1990). The distribution of $X$ for $\rho > 0$ is elliptical, however, not spherical.

Subsection 2.5.5 contains Kingman’s proof of the famous result on $\Phi_{\infty}$ by Schoenberg in the context of exchangeability. Although Schoenberg did not mention exchangeability, we just showed that if a c.f. $\phi$ of $X$ belongs to $\Phi_{\infty}$, then in fact the $X_1, \ldots, X_m$ are exchangeable. Kingman, however, started with an infinite sequence of random variables with the property that, for any $m$, the distribution of $(X_1, \ldots, X_m)'$ has spherical symmetry. This implies that $X_1, \ldots, X_m$ are exchangeable, and this feature is crucial in the proof. As a remark in his proof, Kingman stated that:

‘A curious feature of the proof is that, once exchangeability of the sequence has been established, the spherical symmetry is only used again for $n = 4$.’

We have to bear in mind, however, that spherical symmetry is needed for arbitrary $n$ to establish exchangeability. Spherical symmetry for $n = 4$ is therefore only sufficient if exchangeability holds. Moreover, Kingman referred to Wachter, who noticed that circular symmetry of the distribution of $(X_n, X_m), n \neq m$, in combination with exchangeability, is already sufficient to ensure spherical symmetry for $n = 4$. This immediately follows if we observe that the assumption of exchangeability together with the pairwise circular symmetry is characterized by equation (2.5.16).

Therefore, even though Schoenberg and Kingman basically proved the same thing, their approaches are fundamentally different. Whereas Kingman started with an infinite sequence of exchangeable random variables, Schoenberg considered a finite-dimensional random vector $X = (X_1, \ldots, X_m)'$, and established a representation theorem for arbitrary $m$ and $m = \infty$. In Schoenberg’s proof, exchangeability is a natural consequence and is not used explicitly.

Having showed that a random vector $X$, with $L(X) \in SC_m$ and $\phi \in \Phi_{\infty}$, implies exchangeability of $X_1, \ldots, X_m$, we will now consider the more general case where $L(X) \in SC_m$, and $\phi \in \Phi_m \setminus \Phi_{\infty}$. If $L(X) \in SC_m$, we know from theorem 2.5.1 that the corresponding c.f. $\phi \in \Phi_m$ can be written as

$$
\phi(t) = \int_0^{\infty} \Omega_m(ty) dQ(y)
$$

for some probability measure $Q$ on $[0, \infty)$, where $\Omega_m$ is the characteristic function of the random vector $U$ which is uniformly distributed on the $m$-dimensional unit sphere $S^{m-1}$. Equivalently, we can say that $L(X) = L(RU)$,
2.6. Conclusions

where $R$ is a random variable on $[0, \infty)$ independent of $U$. The distribution of $X$ given $R$ is now uniformly distributed on the sphere of radius $R$, and although this implies that the distribution of $X_i | R$ is the same for all $i = 1, \ldots, m$, the random variables $X_i | R$, $i = 1, \ldots, m$ need not be independent. Therefore, we cannot ensure exchangeability on the basis of $\Phi_m$, as we did for $\Phi_\infty$. A simple example which shows that sphericity of the distribution of $X$ does not in general imply exchangeability of $X_1, \ldots, X_m$ is the uniform distribution on the uniform sphere. That is, if $U \sim \mathcal{U}(S^{m-1})$, then the random variables $U_1, \ldots, U_m$ are not independent and their marginal distributions depend on $m$. Therefore, $U_1, \ldots, U_m$ cannot be exchangeable.

2.6 Conclusions

In the first part of this chapter, we discussed the normal and the Wishart distribution. We started with the well-known definition of the normal distribution and some of its properties. Our main contribution with respect to the normal distribution is the derivation of the density function of the $m$-dimensional random vector $X \sim \mathcal{N}_m(\mu, \Sigma)$, where $\text{rank}(\Sigma) = r < m$. Although the probability density function of this random vector does not exist with respect to the Lebesgue measure on $\mathbb{R}^m$, we showed that it does exist on an affine subspace. We introduced the Wishart distribution and the main theorem on partitioned Wishart distributed matrices. Whereas this theorem is known to be true for the nonsingular Wishart distribution, we proved that the theorem is also true for the singular case. As a second extension of the theorem, which only applies to the central Wishart distribution, we attempted to generalize it to the noncentral Wishart distribution. We discussed the problems we encountered and explored the boundaries of what can and what cannot be done.

In the second part of this chapter, we studied a famous theorem by Schoenberg (1938) in the theory on spherical distributions. The normal distribution plays an important part in this theorem. The proof of this theorem was originally given by Schoenberg in 1938, an alternative proof was given by Kingman in 1972. Both of these proofs are rather complicated. Other simplifications of Schoenberg’s result have been given, yet these proofs are also quite complicated and technical. We proposed a new and shorter proof using elementary techniques from probability theory. We also included Kingman’s proof which exploits the property of exchangeability of a sequence of random variables. Finally, we pointed out to the relationship between sphericity
and exchangeability.

2.A Proof of theorem 2.2.1

We prove theorem 2.2.1 by means of a couple of lemmas. We use the notation

\[ X' = (X_1, \ldots, X_n), M' = (\mu_1, \ldots, \mu_n) \]

We have that \( X_1, \ldots, X_n \) are independently distributed with \( X_i \sim \mathcal{N}_k(\mu_i, \Sigma) \). We first consider the case \( \text{rank}(\Sigma) = k \).

Lemma 2.A.1. If \( X \sim \mathcal{N}_{n \times k}(M, I_n \otimes \Sigma) \), where \( \text{rank}(\Sigma) = k \), then \( \text{rank}(X) = \min(n, k) \) with probability one.

Proof. We consider two situations, namely, \( n \leq k \) and \( n > k \).

(i) Let \( n \leq k \), then

\[
P(\text{rank}(X) < n) \leq \sum_{i=1}^{n} P(X_i \in \text{lh}\{X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n\})
\]

\[
= \sum_{i=1}^{n} EP(X_i \in \text{lh}\{X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n\} | X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)
\]

\[
= 0
\]

because \( \dim \text{lh}\{X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n\} \leq n-1 < k \). Hence \( \text{rank}(X) = n \) with probability one.

(ii) Let \( n > k \). Define the random \( k \times k \) matrix \( Y \) by \( Y' = (X_1, \ldots, X_k) \), then \( Y \sim \mathcal{N}_{k \times k}(N, I_k \otimes \Sigma) \), with \( N' = (\mu_1, \ldots, \mu_k) \). According to part (i) of this proof, \( \text{rank}(Y) = k \) with probability one. Since \( \{\text{rank}(Y) = k\} \subset \{\text{rank}(X) = k\} \), we conclude \( \text{rank}(X) = k \) with probability one.

\[ \square \]

Lemma 2.A.2. Let \( Y \sim \mathcal{N}_{n \times r}(M_1, I_n \otimes \Omega) \) with \( \text{rank}(\Omega) = r \), let \( M_2 \) be an \( n \times s \) matrix of rank \( s \), and let \( X = (Y, M_2) \). Then \( \text{rank}(X) = \min(n, r + s) \) with probability one.

Proof. Note that the case \( s = 0 \) is covered by lemma 2.A.1. So, we may assume \( s > 0 \). Obviously \( s \leq n \).

(i) Assume \( n \leq r \). According to lemma 2.A.1 we have

\[
1 = P\{\text{rank}(Y) = n\} = P\{\text{rank}(Y, M_2) = n\} = P\{\text{rank}(X) = n\}.
\]
(ii) Assume \( r < n \leq r + s \). Let \( M_2 = (m_{r+1}, \ldots, m_{r+s}) \). Define \( Z = XA \), where the \((r + s) \times (r + s)\) matrix

\[
A = \begin{pmatrix}
\Omega^{-\frac{1}{2}} & 0 \\
0 & I_s
\end{pmatrix}
\]

is of rank \( r + s \). We partition \( Z \) as follows:

\[
Z = (Y \Omega^{-\frac{1}{2}}, M_2) = (Z_1, M_2),
\]

where \( Z_1 = Y \Omega^{-\frac{1}{2}} \sim \mathcal{N}_{n,r}(M_1 \Omega^{-\frac{1}{2}}, I_n \otimes I_r) \). Let \( Z_1 = (Z_{11}, \ldots, Z_{1r}) \), then by construction \( Z_{11}, \ldots, Z_{1r} \) are independently distributed with

\[
Z_{1j} \sim \mathcal{N}_n(\lambda_j, I_n), \quad j = 1, \ldots, r
\]

and

\[
M_1 \Omega^{-\frac{1}{2}} = (\lambda_1, \ldots, \lambda_n).
\]

Subsequently, consider the \( n \times n \) submatrix \( \tilde{Z} \) of \( Z \) defined by \( \tilde{Z} = (Z_1, m_{r+1}, \ldots, m_n) \), then

\[
P\left( \text{rank}(\tilde{Z}) < n \right) \\
\leq \sum_{j=1}^{r} P \left( Z_{1j} \in \text{lh}\{Z_{11}, \ldots, Z_{1,j-1}, Z_{1,j+1}, \ldots, Z_{1r}, m_{r+1}, \ldots, m_n\} \right) \\
= 0.
\]

This implies that

\[
P(\text{rank}(X) = n) = P(\text{rank}(Z) = n) \geq P\left( \text{rank}(\tilde{Z}) = n \right) = 1.
\]

(iii) Assume that \( n > r + s \). Let \( M'_2 = (m_1, \ldots, m_n) \). Select \( m_1, \ldots, m_{r+s} \) such that the \((r + s) \times s\) matrix \( M_3 \) defined by \( M'_3 = (m_{i1}, \ldots, m_{is}) \) is of rank \( s \). Let \( Y' = (Y_1, \ldots, Y_n) \) and define \( Z' = (Y_1, \ldots, Y_{r+s}) \). Now we consider the \((r + s) \times (r + s)\) random matrix \((Z, M_3)\). According to part (ii) we have

\[
P\left\{ \text{rank}(Z, M_3) = r + s \right\} = 1.
\]

Since \( X \) is obtained from \((Z, M_3)\) by adding rows we now may conclude that \( \text{rank}(X) = r + s \) with probability one.

\[\square\]

As an immediate extension we have
Lemma 2.A.3. Let $Y \sim \mathcal{N}_{n \times r}(M_1, I_n \otimes \Omega)$ with $\text{rank}(\Omega) = r$ and let $M_2$ be an $n \times (k - r)$ matrix of rank $s$. Define $X = (Y, M_2)$, then $\text{rank}(X) = \min(n, r + s)$ with probability one.

Proof. Apply lemma 2.A.2 to $X$ after deleting $k - r - s$ columns of $M_2$ in such a way the remaining part is of the full column rank $s \leq n$. 

Proof of theorem 2.2.1. Let $\Sigma = CDC'$ be a spectral decomposition with $D = \text{diag}(\lambda_1, \ldots, \lambda_r, 0, \ldots, 0)$, such that $\lambda_1 \geq \ldots \geq \lambda_r > 0$ and let $C$ be a $k \times k$ orthogonal matrix that is partitioned as $C = (C_1, C_2)$, where $C_1$ is of order $k \times r$. Now we have $XC \sim \mathcal{N}_{n \times k}(MC, I_n \otimes D)$. Let $Y = XC_1, M_2 = MC_2, \Omega = \text{diag}(\lambda_1, \ldots, \lambda_r)$, then $XC = (Y, M_2)$ with probability one, and $Y \sim \mathcal{N}_{n \times r}(MC_1, I_n \otimes \Omega)$. If $\text{rank}(MC_2)$ is denoted by $s$, then according to lemma 2.A.3 we have

$$P\{\text{rank}(X) = \min(n, r + s)\} = P\{\text{rank}(XC) = \min(n, r + s)\} = 1.$$ 

Because

$$\text{rank}(MC_2) = \text{rank}(MC_2C_2'M') = \text{rank}(MC_2C_2'C_2'M) = \text{rank}(MC_2C_2'),$$

it follows that

$$s = \text{rank}(MC_2) = \text{rank}(MC_2C_2') = \text{rank} \left( M(I_k - C_1C_1') \right) = \text{rank} \left( M(I_k - \Sigma^+ \Sigma) \right),$$

where we used the fact that $\Sigma^+ = C_1\Omega^{-1}C_1'$. For ease of notation define

$$\Xi = (I_k - \Sigma\Sigma^+)M'M(I_k - \Sigma^+ \Sigma),$$

then $\text{rank}(\Xi) = s$. Note that $\Sigma^+ \Sigma = \Sigma\Sigma^+$. Suppose that $x \in \mathcal{R}(\Sigma) \cap \mathcal{R}(\Xi)$, then we can write for some $y, z \in \mathbb{R}^k$ that

$$x = \Sigma y = \Xi z.$$

Hence

$$||x||^2 = y'\Sigma\Xi z = 0,$$

which implies $x = 0$. Now it follows that

$$\text{rank}(\Sigma + \Xi) = \text{rank}(\Sigma) + \text{rank}(\Xi) = r + s,$$
see, for example, theorem 3.19 of Magnus and Neudecker (1988), or Marsaglia and Styan (1972). In order to show that \( \text{rank}(\Sigma + M'M) = r + s \), it suffices to show that the kernel of \( \Sigma + M'M \) and the kernel of \( \Sigma + \Xi \) are equal. Because \( \Sigma + \Xi \succeq 0 \), it can be derived that

\[
(\Sigma + \Xi)x = 0 \iff x'\Sigma x = 0 \quad \text{and} \quad M(I_k - \Sigma^+\Sigma)x = 0
\]

\[
\iff \Sigma x = 0 \quad \text{and} \quad Mx = 0
\]

\[
\iff (\Sigma + M'M)x = 0.
\]

\( \square \)