CHAPTER 3

Fundamental Properties and Sampling Distributions of the Multivariate Normal Distribution

In this chapter we study some fundamental properties of the multivariate normal distribution, including distribution properties and related sampling distributions.

We first observe several different definitions of the multivariate normal distribution and show their equivalence. In Section 3.3 we consider a partition of the components of a multivariate normal variable, then derive the marginal and conditional distributions and the distributions of linear transformations and linear combinations of its components. The multiple and partial correlations, the canonical correlations, and the principal components are defined and studied in Section 3.4. Finally, in Section 3.5, we derive sampling distributions of the sample mean vector, the sample covariance matrix, and the sample correlation coefficients.

3.1. Preliminaries

In order to properly define the multivariate normal distribution and to study its distribution properties more efficiently, we begin with a review of some basic facts concerning the covariance matrix and the characteristic function of an n-dimensional random variable.

For $n \geq 2$ let $\mathbf{X} = (X_1, \ldots, X_n)'$ be an n-dimensional random variable. Let μ_i and σ_{ii} denote, respectively, the mean and the variance of X_i $(i = 1, \ldots, n)$, and let σ_{ij} denote the covariance between X_i and X_j $(1 \leq i < j \leq n)$. Then

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ & & \cdots & \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{pmatrix}, \tag{3.1.1}$$

Y. L. Tong, *The Multivariate Normal Distribution*

are, respectively, the mean vector and the covariance matrix of $\mathbf{X}$. For notational convenience we shall occasionally write σ_{ii} as σ_i^2 $(i = 1, \ldots, n)$.

Fact 3.1.1. *For $k \geq 1$, let $\mathbf{C}$ be a $k \times n$ real matrix and let $\mathbf{b}$ be a $k \times 1$ real vector. Let $\mathbf{Y} = \mathbf{CX} + \mathbf{b}$. Then the mean vector and the covariance matrix of $\mathbf{Y}$ are, respectively,*

$$\boldsymbol{\mu}_{\mathbf{Y}} = \mathbf{C}\boldsymbol{\mu} + \mathbf{b}, \qquad \boldsymbol{\Sigma}_{\mathbf{Y}} = \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}'.$$

PROOF. For each fixed $i = 1, \ldots, k$, we have

$$Y_i = \sum_{s=1}^{n} c_{is} X_s + b_i, \qquad i = 1, \ldots, k.$$

Thus

$$EY_i = \sum_{s=1}^{n} c_{is}\mu_s + b_i, \qquad i = 1, \ldots, k,$$

which is the ith row of $\mathbf{C}\boldsymbol{\mu} + \mathbf{b}$. Furthermore,

$$\begin{aligned} E(Y_i - EY_i)(Y_j - EY_j) &= E\left[\left\{\sum_{s=1}^{n} c_{is}(X_s - \mu_s)\right\}\left\{\sum_{t=1}^{n} c_{jt}(X_t - \mu_t)\right\}\right] \\ &= \sum_{s=1}^{n}\sum_{t=1}^{n} c_{is}c_{jt}\sigma_{ij}, \end{aligned}$$

which is just the (i, j)th element of $\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}'$. □

Choosing $k = 1$ and $\mathbf{C} = \mathbf{c}' = (c_1, \ldots, c_n)$ in Fact 3.1.1 we have

Fact 3.1.2. *For $\mathbf{c}' = (c_1, \ldots, c_n)$ the variance of $Y = \mathbf{c}'\mathbf{X} = \sum_{i=1}^{n} c_i X_i$ is*

$$\sigma_Y^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \sigma_{ij} = \mathbf{c}'\boldsymbol{\Sigma}\mathbf{c}.$$

An $n \times n$ symmetric matrix $\boldsymbol{\Sigma}$ is said to be positive definite (p.d.) if $\mathbf{c}'\boldsymbol{\Sigma}\mathbf{c} \geq 0$ holds for all real vectors $\mathbf{c}$, and equality holds only for $\mathbf{c} = \mathbf{0}$. It is said to be positive semidefinite (p.s.d.) if $\mathbf{c}'\boldsymbol{\Sigma}\mathbf{c} \geq 0$ holds for all real vectors $\mathbf{c}$, and equality holds for some $\mathbf{c} = \mathbf{c}_0 \neq \mathbf{0}$. It is known that if $\boldsymbol{\Sigma}$ is p.d. (p.s.d.), then $|\boldsymbol{\Sigma}| > 0$ ($|\boldsymbol{\Sigma}| = 0$) or, equivalently, the rank of $\boldsymbol{\Sigma}$ is n (is less than n).

The distribution of $\mathbf{X}$ is said to be singular if there exists a vector $\mathbf{c}_0 \neq \mathbf{0}$ such that $Y = \mathbf{c}_0'\mathbf{X}$ is singular (that is, $P[Y = \mu_Y] = 1$). But the variance of Y is $\mathbf{c}'\boldsymbol{\Sigma}\mathbf{c}$ and Y is singular if and only if $\sigma_Y^2 = 0$. Thus we have

Fact 3.1.3. *A covariance matrix $\boldsymbol{\Sigma}$ is either p.d. or p.s.d. Furthermore,*

$$\begin{aligned} \boldsymbol{\Sigma} \text{ is p.s.d.} \quad &\Leftrightarrow \quad |\boldsymbol{\Sigma}| = 0, \\ &\Leftrightarrow \quad \text{the rank of } \boldsymbol{\Sigma} \text{ is less than } n, \\ &\Leftrightarrow \quad \text{the corresponding distribution is singular.} \end{aligned}$$

We shall say that the distribution of $\mathbf{X}$ is nonsingular if it is not singular. Furthermore, for notational convenience we write $\mathbf{\Sigma} > 0$ instead of $|\mathbf{\Sigma}| > 0$ when $\mathbf{\Sigma}$ is p.d.

The characteristic function (c.f.) of an n-dimensional random variable $\mathbf{X}$ is given by

$$\psi_{\mathbf{X}}(\mathbf{t}) = Ee^{i\mathbf{t}'\mathbf{X}}, \qquad \mathbf{t} \in \Re^n,$$

where $i^2 = -1$. Through an application of the following known result:

Fact 3.1.4 (Uniqueness Theorem). *The c.f. of a random variable* $\mathbf{X}$ *determines its distribution uniquely*;

c.f.'s can be used for finding the distribution of a random variable.

For linear transformations of random variables, the following fact can easily be established:

Fact 3.1.5. *Let* $\mathbf{X}$ *be an n-dimensional random variable with c.f.* $\psi_{\mathbf{X}}(\mathbf{t})$. *Let* $\mathbf{C}$ *be an* $n \times n$ *real matrix and let* $\mathbf{b}$ *be an* $n \times 1$ *vector. Then the c.f. of* $\mathbf{Y} = \mathbf{CX} + \mathbf{b}$ *is* $\psi_{\mathbf{Y}}(\mathbf{t}) = e^{i\mathbf{t}'\mathbf{b}}\psi_{\mathbf{X}}(\mathbf{C}'\mathbf{t})$.

PROOF.

$$\begin{aligned}\psi_{\mathbf{Y}}(\mathbf{t}) &= Ee^{i\mathbf{t}'\mathbf{Y}} \\ &= Ee^{i\mathbf{t}'(\mathbf{CX}+\mathbf{b})} \\ &= e^{i\mathbf{t}'\mathbf{b}}Ee^{i(\mathbf{C}'\mathbf{t})'\mathbf{X}}.\end{aligned}$$

□

Now consider the partition of the components of an n-dimensional random variable $\mathbf{Y}$ given by $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2)'$, where $\mathbf{Y}_1$ is $k \times 1$ and $\mathbf{Y}_2$ is $(n-k) \times 1$.

Fact 3.1.6. *If the c.f. of* $\mathbf{Y}$ *is* $\psi_{\mathbf{Y}}(\mathbf{t})$, $\mathbf{t} \in \Re^n$, *then the c.f. of* $\mathbf{Y}_1$ *is* $\psi_{\mathbf{Y}}(\mathbf{t}_1, \mathbf{0})$, $\mathbf{t}_1 \in \Re^k$.

PROOF.

$$\psi_{\mathbf{Y}_1}(t_1, \ldots, t_k) = E\exp\left(i\sum_1^k t_jY_j\right) = E\exp\left[i\left(\sum_1^k t_jY_j + \sum_{k+1}^n 0Y_j\right)\right] = \psi_{\mathbf{Y}}(\mathbf{t}_1, \mathbf{0}).$$

□

If $\mathbf{H}$ is a $k \times n$ real matrix $(k < n)$ and if we are interested in finding the distribution of $\mathbf{Y}_1 = \mathbf{HX}$, a standard procedure is:

(i) Find $\psi_{\mathbf{Y}}(\mathbf{t})$, the c.f. of

$$\begin{pmatrix}\mathbf{Y}_1\\ \mathbf{Y}_2\end{pmatrix} = \begin{pmatrix}\mathbf{H} & \mathbf{0}\\ \mathbf{0}' & \mathbf{I}_{n-k}\end{pmatrix}\mathbf{X}$$

by applying Fact 3.1.5, where $\mathbf{0}$ is the $k \times (n-k)$ matrix with elements 0, and $\mathbf{I}_{n-k}$ is the $(n-k) \times (n-k)$ identity matrix;

(ii) find $\psi_{\mathbf{Y}_1}(\mathbf{t}_1)$ from $\psi_{\mathbf{Y}}(\mathbf{t})$, where $\mathbf{t}_1 = (t_1, \ldots, t_k)' \in \Re^k$;

(iii) identify the density $f_1(\mathbf{y}_1)$ associated with the c.f. $\psi_{\mathbf{Y}_1}(\mathbf{t}_1)$, then apply the uniqueness theorem (Fact 3.1.4) to claim that the density function of $\mathbf{Y}_1$ is $f_1(\mathbf{y}_1)$.

This method will be used in the proof of Theorem 3.3.1 for deriving the marginal distributions of a multivariate normal distribution.

3.2. Definitions of the Multivariate Normal Distribution

We first give a definition of the nonsingular multivariate normal distribution.

Definition 3.2.1. An n-dimensional random variable $\mathbf{X}$ with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ is said to have a nonsingular multivariate normal distribution, in symbols $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} > 0$, if (i) $\boldsymbol{\Sigma}$ is positive definite, and (ii) the density function of $\mathbf{X}$ is of the form

$$f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-Q_n(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})/2}, \qquad \mathbf{x} \in \Re^n, \tag{3.2.1}$$

where

$$Q_n(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}). \tag{3.2.2}$$

Remark 3.2.1. For this definition to be consistent, we must verify that if $\mathbf{X}$ has the density function $f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the mean vector and the covariance matrix of $\mathbf{X}$ are indeed $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, respectively. This is postponed and will be given in Remark 3.3.1.

Now let $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} > 0$, and consider the transformation

$$\mathbf{Y} = \mathbf{C}\mathbf{X} + \mathbf{b}, \tag{3.2.3}$$

where $\mathbf{C} = (c_{ij})$ is an $n \times n$ real matrix and $\mathbf{b}$ is a real vector.

Theorem 3.2.1. *Let* $\mathbf{Y}$ *be defined as in* (3.2.3). *If* $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} > 0$, *and* $\mathbf{C}$ *is an* $n \times n$ *real matrix such that* $|\mathbf{C}| \neq 0$, *then* $\mathbf{Y} \sim \mathcal{N}_n(\boldsymbol{\mu}_\mathbf{Y}, \boldsymbol{\Sigma}_\mathbf{Y})$, $\boldsymbol{\Sigma}_\mathbf{Y} > 0$, *where*

$$\boldsymbol{\mu}_\mathbf{Y} = \mathbf{C}\boldsymbol{\mu} + \mathbf{b}, \qquad \boldsymbol{\Sigma}_\mathbf{Y} = \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}'. \tag{3.2.4}$$

PROOF. The mean vector and the covariance matrix of $\mathbf{Y}$ given in (3.2.4) follow immediately from Fact 3.1.1. To show normality we note that if $|\mathbf{C}| \neq 0$, then $\mathbf{C}^{-1}$ and $(\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}')^{-1} = \mathbf{C}'^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{C}^{-1}$ both exist. Thus we can write (by $\mathbf{y} = \mathbf{C}\mathbf{x} + \mathbf{b}$) $\mathbf{x} = \mathbf{C}^{-1}(\mathbf{y} - \mathbf{b})$. The density function of $\mathbf{Y}$ is then given by

$$\begin{aligned} g(\mathbf{y}; \boldsymbol{\mu}_\mathbf{Y}, \boldsymbol{\Sigma}_\mathbf{Y}) &= f(\mathbf{C}^{-1}(\mathbf{y} - \mathbf{b}), \boldsymbol{\mu}, \boldsymbol{\Sigma})|J| \\ &= \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-((\mathbf{C}^{-1}(\mathbf{y}-\mathbf{b})-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{C}^{-1}(\mathbf{y}-\mathbf{b})-\boldsymbol{\mu}))/2}|J|, \end{aligned}$$

where $|J|$ is the absolute value of $|\mathbf{C}^{-1}|$ and f is defined in (3.2.1). But $|\mathbf{C}^{-1}| = 1/|\mathbf{C}|$, so that $|\boldsymbol{\Sigma}|^{-1/2}|J| = |\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}'|^{-1/2}$. Furthermore, it is straightforward to verify that

$$\begin{aligned}(\mathbf{C}^{-1}(\mathbf{y}-\mathbf{b})-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{C}^{-1}(\mathbf{y}-\mathbf{b})-\boldsymbol{\mu}) \\ &= (\mathbf{y}-(\mathbf{C}\boldsymbol{\mu}+\mathbf{b}))'(\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}')^{-1}(\mathbf{y}-(\mathbf{C}\boldsymbol{\mu}+\mathbf{b})) \\ &= Q_n(\mathbf{y}; \boldsymbol{\mu}_{\mathbf{Y}}, \boldsymbol{\Sigma}_{\mathbf{Y}}).\end{aligned}$$

Thus we have

$$g(\mathbf{y}; \boldsymbol{\mu}_{\mathbf{Y}}, \boldsymbol{\Sigma}_{\mathbf{Y}}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}_{\mathbf{Y}}|^{1/2}} e^{-Q_n(\mathbf{y}; \boldsymbol{\mu}_{\mathbf{Y}}, \boldsymbol{\Sigma}_{\mathbf{Y}})/2}, \qquad \mathbf{y} \in \Re^n. \qquad \square$$

A special case of interest is the standard multivariate normal variable, denoted by $\mathbf{Z} = (Z_1, \ldots, Z_n)'$, with means 0, variances 1, and correlation coefficients 0. In this case we can write $\mathbf{Z} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{I}_n)$ with the density function given by

$$f(\mathbf{z}; \mathbf{0}, \mathbf{I}_n) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{j=1}^{n} z_j^2\right), \qquad \mathbf{z} \in \Re^n.$$

Thus $Z_1, \ldots, Z_n$ are independent random variables. After integrating out, we see that the marginal distribution of Z_i is univariate normal with mean 0 and variance 1.

Consider any given random variable $\mathbf{X}$ which has an $\mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution, $\boldsymbol{\Sigma} > 0$. We now show how $\mathbf{X}$ and $\mathbf{Z}$ are related. For this purpose we recall a result in linear algebra.

Proposition 3.2.1. *Let $\boldsymbol{\Sigma}$ be an $n \times n$ symmetric matrix with rank r such that $\boldsymbol{\Sigma}$ is either positive definite ($r = n$) or positive semidefinite ($r < n$).*

(i) *If $r = n$, then there exists a nonsingular $n \times n$ matrix $\mathbf{H}$ such that $\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}' = \mathbf{I}_n$.*
(ii) *If $r < n$, then there exists a nonsingular $n \times n$ matrix $\mathbf{H}$ such that*

$$\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}' = \begin{pmatrix} \mathbf{I}_r & \mathbf{0}_{12} \\ \mathbf{0}_{21} & \mathbf{0}_{22} \end{pmatrix} \equiv \mathbf{D}, \tag{3.2.5}$$

where $\mathbf{0}_{12}, \mathbf{0}_{21}, \mathbf{0}_{22}$ are $r \times (n-r)$, $(n-r) \times r$, and *$(n-r) \times (n-r)$ matrices with elements* 0.

Letting $\mathbf{B} = \mathbf{H}^{-1}$ we have:

(i)′ *if $r = n$, then there exists a nonsingular $n \times n$ matrix $\mathbf{B}$ such that $\mathbf{B}\mathbf{B}' = \boldsymbol{\Sigma}$;*
(ii)′ *if $r < n$, then there exists a nonsingular $n \times n$ matrix $\mathbf{B}$ such that $\mathbf{B}\mathbf{D}\mathbf{B}' = \boldsymbol{\Sigma}$.*

PROOF. See Anderson (1984, Theorem A.2.2). $\square$

By choosing $\mathbf{C} = \mathbf{B}$ in Proposition 3.2.1(i) we immediately have

Theorem 3.2.2. $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} > 0$, *holds if and only if there exists a nonsingular $n \times n$ matrix $\mathbf{C}$ such that*

(i) $\mathbf{CC}' = \boldsymbol{\Sigma}$; *and*
(ii) $\mathbf{X}$ *and* $\mathbf{CZ} + \boldsymbol{\mu}$ *are identically distributed, where* $\mathbf{Z} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{I}_n)$.

Next we direct our attention to the more general case in which the covariance matrix is not necessarily positive definite. To this end, we state a natural generalization of Definition 2.0.1(b).

Definition 3.2.2. An n-dimensional random variable $\mathbf{X}$ with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ is said to have a singular multivariate normal distribution (in symbols, $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $|\boldsymbol{\Sigma}| = 0$) if:

(i) $\boldsymbol{\Sigma}$ is positive semidefinite; and
(ii) for some $r < n$ there exists an $n \times r$ real matrix $\mathbf{C}$ such that $\mathbf{X}$ and $\mathbf{CZ}_r + \boldsymbol{\mu}$ are identically distributed, where $\mathbf{Z}_r \sim \mathcal{N}_r(\mathbf{0}, \mathbf{I}_r)$.

Combining the nonsingular (Definition 3.2.1) and singular (Definition 3.2.2) cases, we have

Definition 3.2.3. An n-dimensional random variable with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ is said to have a multivariate normal distribution (in symbols $\mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$) if either $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} > 0$, or $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $|\boldsymbol{\Sigma}| = 0$.

By Theorem 3.2.2 and Definition 3.2.2, Definition 3.2.3 is equivalent to:

Definition 3.2.4. An n-dimensional random variable $\mathbf{X}$ with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ is said to have a multivariate normal distribution (in symbols $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$) if there exists an $n \times r$ matrix $\mathbf{C}$ with rank $r \leq n$ such that:

(i) $\mathbf{CC}' = \boldsymbol{\Sigma}$; and
(ii) $\mathbf{X}$ and $\mathbf{CZ}_r + \boldsymbol{\mu}$ are identically distributed, where $\mathbf{Z}_r \sim \mathcal{N}_r(\mathbf{0}, \mathbf{I}_r)$.

Definition 3.2.4 was proposed by P.L. Hsu (Fang, 1988). It applies to both the nonsingular and singular cases, and is convenient for obtaining the marginal distributions and distributions of linear transformations of normal variables. Another useful application of Definition 3.2.4 is for obtaining the characteristic function (c.f.) of a multivariate normal variable. Since the c.f. of a univariate $\mathcal{N}(0, 1)$ variable is $e^{-t^2/2}$, the c.f. of $\mathbf{Z}_r$ is

$$\psi_{\mathbf{Z}_r}(\mathbf{t}) = e^{-\mathbf{t}'\mathbf{t}/2}, \qquad \mathbf{t} \in \Re^r.$$

By Definition 3.2.4 and Facts 3.1.4 and 3.1.5 we have, for all $r \leq n$:

Theorem 3.2.3. $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *holds if and only if its characteristic function is of the form*

$$\psi_{\mathbf{X}}(\mathbf{t}) = e^{i\mathbf{t}'\boldsymbol{\mu} - \mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}/2}, \qquad \mathbf{t} \in \Re^n. \tag{3.2.6}$$

The next definition involves a closure property of linear combinations of the components of $\mathbf{X}$.

Definition 3.2.5. An n-dimensional random variable $\mathbf{X}$ with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ is said to have a multivariate normal distribution if the distribution of $\mathbf{c}'\mathbf{X}$ is (univariate) $\mathcal{N}(\mathbf{c}'\boldsymbol{\mu}, \mathbf{c}'\boldsymbol{\Sigma}\mathbf{c})$ for all real vectors $\mathbf{c}$.

It should be noted that for a given n-dimensional random variable $\mathbf{X}$, $\mathbf{c}'\mathbf{X}$ may have a univariate normal distribution for some $\mathbf{c} \neq \mathbf{0}$ but not for all $\mathbf{c}$. In this case, of course, $\mathbf{X}$ is not normally distributed. To see this fact, consider the following example given in Anderson (1984, pp. 47–48).

EXAMPLE 3.2.1. Let $n = 2$, and define

$$A_1 = \{(x_1, x_2)': 0 \le x_i \le 1, i = 1, 2\},$$
$$A_2 = \{(x_1, x_2)': -1 \le x_1 \le 0, 0 \le x_2 \le 1\},$$
$$A_3 = \{(x_1, x_2)': -1 \le x_i \le 0, i = 1, 2\},$$
$$A_4 = \{(x_1, x_2)': 0 \le x_1 \le 1, -1 \le x_2 \le 0\}.$$

Let the density function of $\mathbf{X} = (X_1, X_2)'$ be

$$f(\mathbf{x}) = \begin{cases} \dfrac{1}{\pi} e^{-(x_1^2+x_2^2)/2} & \text{for} \quad \mathbf{x} \in A_1 \cup A_3, \\ 0 & \text{for} \quad \mathbf{x} \in A_2 \cup A_4, \\ \dfrac{1}{2\pi} e^{-(x_1^2+x_2^2)/2} & \text{otherwise.} \end{cases}$$

Then the marginal distributions of X_1 and X_2 are both $\mathcal{N}(0, 1)$, hence $\mathbf{c}'\mathbf{X}$ is $\mathcal{N}(0, 1)$ for $\mathbf{c} = (1, 0)'$ or $\mathbf{c} = (0, 1)'$. But clearly $\mathbf{X}$ does not have a bivariate normal distribution. □

We now prove the equivalence of all the definitions of the multivariate normal distribution stated above.

Theorem 3.2.4. *Definitions* 3.2.3, 3.2.4, *and* 3.2.5 *are equivalent.*

PROOF. The equivalence of Definitions 3.2.3 and 3.2.4 is clear. Thus it suffices to show the equivalence of Definitions 3.2.4 and 3.2.5.

It is immediate that if $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\mathbf{c}'\mathbf{X}$ is a univariate $\mathcal{N}(\mathbf{c}'\boldsymbol{\mu}, \mathbf{c}'\boldsymbol{\Sigma}\mathbf{c})$ variable for all $\mathbf{c}$. Conversely, suppose that $\mathbf{c}'\mathbf{X}$ has an $\mathcal{N}(\mathbf{c}'\boldsymbol{\mu}, \mathbf{c}'\boldsymbol{\Sigma}\mathbf{c})$ distribution for all $\mathbf{c} \in \mathfrak{R}^n$, then

$$\psi_{\mathbf{c}'\mathbf{X}}(t) = E \exp\left(it \sum_{j=1}^{n} c_j X_j\right) = e^{it\mathbf{c}'\boldsymbol{\mu} - (\mathbf{c}'\boldsymbol{\Sigma}\mathbf{c})t^2/2} \tag{3.2.7}$$

holds for all $t \in \Re$ and $\mathbf{c} \in \Re^n$. Thus

$$\psi_{\mathbf{c}'\mathbf{X}}(1) = Ee^{i\mathbf{c}'\mathbf{X}} = e^{i\mathbf{c}'\boldsymbol{\mu} - \mathbf{c}'\boldsymbol{\Sigma}\mathbf{c}/2} \equiv \psi_{\mathbf{X}}^*(\mathbf{c}), \qquad \mathbf{c} \in \Re^n.$$

But $\psi_{\mathbf{X}}^*(\mathbf{c})$ is just the characteristic function of a multivariate normal variable with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Thus by Theorem 3.2.3 and Fact 3.1.4 we have $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. □

3.3. Basic Distribution Properties

In this section we describe certain distribution properties of the multivariate normal distribution.

3.3.1. Marginal Distributions and Independence

First we show that the marginal distributions of a multivariate normal variable are normal. For fixed $k < n$, consider the partitions of $\mathbf{X}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$ given below:

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}, \qquad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}, \tag{3.3.1}$$

where

$$\mathbf{X}_1 = (X_1, \ldots, X_k)', \qquad \mathbf{X}_2 = (X_{k+1}, \ldots, X_n)', \tag{3.3.2}$$

$$\boldsymbol{\mu}_1 = (\mu_1, \ldots, \mu_k)', \qquad \boldsymbol{\mu}_2 = (\mu_{k+1}, \ldots, \mu_n)', \tag{3.3.3}$$

$\boldsymbol{\Sigma}_{ii}$ is the covariance matrix of $\mathbf{X}_i$ $(i = 1, 2)$, and $\boldsymbol{\Sigma}_{12} = (\sigma_{ij})$ is such that $\sigma_{ij} = \operatorname{cov}(X_i, X_j)$ for $1 \le i < j \le n$. Let $\mathbf{R} = (\rho_{ij})$ be such that

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}, \qquad i, j = 1, \ldots, n. \tag{3.3.4}$$

Then $\mathbf{R}$ is called the correlation matrix of $\mathbf{X}$. The diagonal elements of $\mathbf{R}$ are 1 and the off-diagonal elements are the correlation coefficients. For the partition of $\boldsymbol{\Sigma}$ defined in (3.3.1), a corresponding partition of $\mathbf{R}$ is

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{21} & \mathbf{R}_{22} \end{pmatrix}. \tag{3.3.5}$$

In the following theorem we first derive the marginal distributions of $\mathbf{X}_1$ and $\mathbf{X}_2$.

Theorem 3.3.1. *If $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then for every fixed $k < n$ the marginal distributions of $\mathbf{X}_1$ and $\mathbf{X}_2$ are $\mathcal{N}_k(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ and $\mathcal{N}_{n-k}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$, respectively.*

PROOF. By Theorem 3.2.3 and Fact 3.1.6, the characteristic function of $\mathbf{X}_1$ is

$$\begin{aligned}\psi_{\mathbf{X}_1}(t_1, \ldots, t_k) &= \psi_{\mathbf{X}}(t_1, \ldots, t_k, 0, \ldots, 0)\\ &= e^{i\mathbf{t}'\boldsymbol{\mu}_1 - \mathbf{t}'\boldsymbol{\Sigma}_{11}\mathbf{t}/2}, \qquad \mathbf{t} \in \mathfrak{R}^k.\end{aligned}$$

Thus, again by Theorem 3.2.3, $\mathbf{X}_1$ has an $\mathcal{N}_k(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ distribution. The distribution of $\mathbf{X}_2$ follows similarly. □

If $\mathbf{X}$ has a nonsingular normal distribution, then $\boldsymbol{\Sigma} > 0$. This in turn implies $\boldsymbol{\Sigma}_{11} > 0$ and $\boldsymbol{\Sigma}_{22} > 0$. Consequently, we have

Corollary 3.3.1. *If* $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} > 0$, *then* $\mathbf{X}_1 \sim \mathcal{N}_k(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$, $\boldsymbol{\Sigma}_{11} > 0$, *and* $\mathbf{X}_2 \sim \mathcal{N}_{n-k}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$, $\boldsymbol{\Sigma}_{22} > 0$.

Remark 3.3.1. Choosing $k = 2$ in Corollary 3.3.1 we observe, from Theorem 2.1.1, that if $\mathbf{X}$ has the density function given in (3.2.1), then the marginal distribution of $(X_1, X_2)'$ is bivariate normal with means μ_1, μ_2 and variances σ_{11}, σ_{22}, respectively, and covariance σ_{12}. By symmetry we conclude that if $\mathbf{X}$ has the density function given in (3.2.1), then the mean vector and the covariance matrix of $\mathbf{X}$ are, respectively, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. This observation shows that Definition 3.2.1 is indeed consistent as noted in Remark 3.2.1.

It is well known that, in general, uncorrelated random variables are not necessarily independent. But for the multivariate normal variables those two conditions are equivalent. This is shown below.

Theorem 3.3.2. *Let* $\mathbf{X}_1$, $\mathbf{X}_2$ *be the random variables defined in* (3.3.1) *where* $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. *Then they are independent if and only if* $\boldsymbol{\Sigma}_{12} = \mathbf{0}$.

PROOF. Let $\psi_{\mathbf{X}}(t_1, \ldots, t_n)$ denote the characteristic function of $\mathbf{X}$ then, by (3.2.6),

$$\begin{aligned}\boldsymbol{\Sigma}_{12} = 0 \quad &\Leftrightarrow \quad (t_1, \ldots, t_n)\boldsymbol{\Sigma}(t_1, \ldots, t_n)' = \mathbf{t}_1'\boldsymbol{\Sigma}_{11}\mathbf{t}_1 + \mathbf{t}_2'\boldsymbol{\Sigma}_{22}\mathbf{t}_2\\ &\Leftrightarrow \quad \psi_{\mathbf{X}}(\mathbf{t}) = \psi_{\mathbf{X}_1}(\mathbf{t}_1)\psi_{\mathbf{X}_2}(\mathbf{t}_2)\end{aligned}$$

for all $\mathbf{t}_1 = (t_1, \ldots, t_k)' \in \mathfrak{R}^k$ and $\mathbf{t}_2 = (t_{k+1}, \ldots, t_n)' \in \mathfrak{R}^{n-k}$. Since $\mathbf{X}_1$, $\mathbf{X}_2$ are independent if and only if their joint characteristic function is the product of the marginal characteristic functions, the proof is complete. □

Remark 3.3.2. If the distribution of $\mathbf{X}$ in Theorem 3.3.2 is nonsingular, then an alternative proof exists: Let $f_1(\mathbf{x}_1)$, $f_2(\mathbf{x}_2)$ be the marginal density functions of $\mathbf{X}_1$, $\mathbf{X}_2$ given by

$$f_1(\mathbf{x}_1) = \frac{1}{(2\pi)^{k/2}|\boldsymbol{\Sigma}_{11}|^{1/2}} e^{-Q^{(1)}(\mathbf{x}_1; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})/2},$$

$$f_2(\mathbf{x}_2) = \frac{1}{(2\pi)^{(n-k)/2}|\boldsymbol{\Sigma}_{22}|^{1/2}} e^{-Q^{(2)}(\mathbf{x}_2; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})/2},$$

where

$$Q^{(i)}(\mathbf{x}_i; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_{ii}) = (\mathbf{x}_i - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_{ii}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_i), \qquad i = 1, 2.$$

If $\boldsymbol{\Sigma} > 0$ and $\boldsymbol{\Sigma}_{12} = \mathbf{0}$, then

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{\Sigma}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22}^{-1} \end{pmatrix}. \tag{3.3.6}$$

Simple calculation yields the identity

$$Q_n(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = Q^{(1)}(\mathbf{x}_1; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) + Q^{(2)}(\mathbf{x}_2; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}). \tag{3.3.7}$$

This implies that

$$f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = f_1(\mathbf{x}_1; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) f_2(\mathbf{x}_2; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}),$$

as desired.

3.3.2. Linear Transformations and Linear Combinations

For the univariate case, the normal family of distributions is closed under linear transformations and linear combinations of random variables. In the following we show that the family of multivariate normal distributions also possesses such closure properties.

Theorem 3.3.3. *If* $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *and* $\mathbf{Y} = \mathbf{C}\mathbf{X} + \mathbf{b}$, *where* $\mathbf{C}$ *is any given* $m \times n$ *real matrix and* $\mathbf{b}$ *is any* $m \times 1$ *real vector, then* $\mathbf{Y} \sim \mathcal{N}_m(\mathbf{C}\boldsymbol{\mu} + \mathbf{b}, \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}')$.

PROOF. (i) For $m = n$, the proof follows immediately from Fact 3.1.5 and Theorem 3.2.3.

(ii) For $m < n$, consider the transformation

$$\mathbf{Y}^* = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{C} \\ \mathbf{B} \end{pmatrix}\mathbf{X} + \begin{pmatrix} \mathbf{b} \\ \mathbf{0}_{n-m} \end{pmatrix}, \tag{3.3.8}$$

where $\mathbf{B}$ is any given $(n - m) \times n$ matrix. Since

$$\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} \sim \mathcal{N}_n\left(\begin{pmatrix} \mathbf{C}\boldsymbol{\mu} + \mathbf{b} \\ \mathbf{B}\boldsymbol{\mu} \end{pmatrix}, \begin{pmatrix} \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}' & \mathbf{C}\boldsymbol{\Sigma}\mathbf{B}' \\ \mathbf{B}\boldsymbol{\Sigma}\mathbf{C}' & \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}' \end{pmatrix}\right),$$

$\mathbf{Y} = \mathbf{Y}_1 = \mathbf{C}\mathbf{X} + \mathbf{b} \sim \mathcal{N}_m(\mathbf{C}\boldsymbol{\mu}, \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}')$.

(iii) For $m > n$, by Definition 3.2.4, there exists an $n \times r$ matrix $\mathbf{C}^*$ such that $\mathbf{X}$ and $\mathbf{C}^*\mathbf{Z}_r + \boldsymbol{\mu}$ are identically distributed, where $r \leq n$ is the rank of $\boldsymbol{\Sigma}$. Thus $\mathbf{C}\mathbf{X} + \mathbf{b}$ and $\mathbf{C}\mathbf{C}^*\mathbf{Z}_r + (\mathbf{C}\boldsymbol{\mu} + \mathbf{b})$ are both distributed according to a singular $\mathcal{N}(\mathbf{C}\boldsymbol{\mu} + \mathbf{b}, \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}')$ distribution (again by Definition 3.2.4). □

If $\mathbf{X}$ is a nonsingular normal variable and, for $m < n$, if the $m \times n$ matrix $\mathbf{C}$ has rank m, then there exists an $(n - m) \times n$ matrix $\mathbf{B}$ such that the matrix

$\binom{\mathbf{C}}{\mathbf{B}}$ in (3.3.8) is nonsingular. This implies that the distribution of $\mathbf{Y}^*$, and hence the distribution of $\mathbf{Y}$, is nonsingular. Combining the result for $r = n$ already stated in Theorem 3.2.1 we have

Corollary 3.3.2. *If* $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} > 0$, $\mathbf{C}$ *is an* $m \times n$ *real matrix with rank* $m \leq n$, *and* $\mathbf{b}$ *is an* $m \times 1$ *real vector, then* $\mathbf{CX} + \mathbf{b}$ *has a nonsingular* $\mathcal{N}_m(\mathbf{C}\boldsymbol{\mu} + \mathbf{b}, \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}')$ *distribution.*

Next we consider linear combinations of the components of a multivariate normal variable. Let $\mathbf{C}_1$, $\mathbf{C}_2$ be $m \times k$ and $m \times (n-k)$ matrices. For the partition defined in (3.3.1) consider the linear combination $\mathbf{Y} = \mathbf{C}_1\mathbf{X}_1 + \mathbf{C}_2\mathbf{X}_2$. Rewriting this as $\mathbf{Y} = \mathbf{CX}$, where $\mathbf{C} = (\mathbf{C}_1 \;\; \mathbf{C}_2)$ is an $m \times n$ matrix, and applying Theorem 3.3.3 yield

Corollary 3.3.3. *Let* $\mathbf{X}$ *be partitioned as in* (3.3.1), *and let* $\mathbf{C}_1$, $\mathbf{C}_2$ *be two* $m \times k$ *and* $m \times (n-k)$ *real matrices, respectively. If* $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, *then*

$$\mathbf{Y} = \mathbf{C}_1\mathbf{X}_1 + \mathbf{C}_2\mathbf{X}_2 \sim \mathcal{N}_m(\boldsymbol{\mu}_{\mathbf{Y}}, \boldsymbol{\Sigma}_{\mathbf{Y}}),$$

where

$$\boldsymbol{\mu}_{\mathbf{Y}} = \mathbf{C}_1\boldsymbol{\mu}_1 + \mathbf{C}_2\boldsymbol{\mu}_2, \tag{3.3.9}$$

$$\boldsymbol{\Sigma}_{\mathbf{Y}} = \mathbf{C}_1\boldsymbol{\Sigma}_{11}\mathbf{C}_1' + \mathbf{C}_2\boldsymbol{\Sigma}_{22}\mathbf{C}_2' + \mathbf{C}_1\boldsymbol{\Sigma}_{12}\mathbf{C}_2' + \mathbf{C}_2\boldsymbol{\Sigma}_{21}\mathbf{C}_1'. \tag{3.3.10}$$

A special case of interest is $\mathbf{Y} = c_1\mathbf{X}_1 + c_2\mathbf{X}_2$ where c_1, c_2 are real numbers and $n = 2k$. This can be treated in Corollary 3.3.3 by taking $\mathbf{C}_i = c_i\mathbf{I}_k$ $(i = 1, 2)$. If in addition $\boldsymbol{\Sigma}_{12} = \mathbf{0}$, then clearly $\mathbf{Y}$ is distributed according to an $\mathcal{N}_k(c_1\boldsymbol{\mu}_1 + c_2\boldsymbol{\mu}_2, c_1^2\boldsymbol{\Sigma}_{11} + c_2^2\boldsymbol{\Sigma}_{22})$ distribution. Generalizing this result to several variables by induction we have

Corollary 3.3.4. *If* $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$ *are independent* $\mathcal{N}_n(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ *variables* $(j = 1, \ldots, N)$, *then* $\mathbf{Y} = \sum_{j=1}^N c_j\mathbf{X}_j$ *is distributed according to an* $\mathcal{N}_n(\sum_{j=1}^N c_j\boldsymbol{\mu}_j, \sum_{j=1}^N c_j^2\boldsymbol{\Sigma}_j)$ *distribution.*

3.3.3. Conditional Distributions

For $1 \leq k < n$ consider the partition of $\mathbf{X}$ defined in (3.3.1) and the linear transformation

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{I}_k & -\mathbf{B} \\ \mathbf{0} & \mathbf{I}_{n-k} \end{pmatrix} \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \equiv \mathbf{CX}, \tag{3.3.11}$$

where $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are $k \times 1$ and $(n-k) \times 1$ random variables, respectively, and $\mathbf{B}$ is a $k \times (n-k)$ real matrix. If $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then by Theorem 3.3.3 the joint distribution of $\mathbf{Y}_1, \mathbf{Y}_2$ is $\mathcal{N}_n(\boldsymbol{\mu}_\mathbf{Y}, \boldsymbol{\Sigma}_\mathbf{Y})$, where

$$\boldsymbol{\mu}_\mathbf{Y} = \begin{pmatrix} \boldsymbol{\mu}_1 - \mathbf{B}\boldsymbol{\mu}_2 \\ \boldsymbol{\mu}_2 \end{pmatrix},$$

$$\boldsymbol{\Sigma}_\mathbf{Y} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} + \mathbf{B}\boldsymbol{\Sigma}_{22}\mathbf{B}' - \mathbf{B}\boldsymbol{\Sigma}_{21} - \boldsymbol{\Sigma}_{12}\mathbf{B}' & \boldsymbol{\Sigma}_{12} - \mathbf{B}\boldsymbol{\Sigma}_{22} \\ (\boldsymbol{\Sigma}_{12} - \mathbf{B}\boldsymbol{\Sigma}_{22})' & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

If $\boldsymbol{\Sigma}$ is nonsingular, then $\boldsymbol{\Sigma}_{11}^{-1}$ and $\boldsymbol{\Sigma}_{22}^{-1}$ both exist. Thus if we choose $\mathbf{B}$ to satisfy $\boldsymbol{\Sigma}_{12} - \mathbf{B}\boldsymbol{\Sigma}_{22} = \mathbf{0}$, that is, if $\mathbf{B}$ is chosen to be

$$\mathbf{B} = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}, \tag{3.3.12}$$

then $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are uncorrelated (and thus independent). Consequently, we have

$$\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2 \end{pmatrix} \sim \mathcal{N}_n\left(\begin{pmatrix} \mathbf{v}_{1\cdot 2} \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11\cdot 2} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22} \end{pmatrix}\right), \tag{3.3.13}$$

where

$$\mathbf{v}_{1\cdot 2} = \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\mu}_2, \tag{3.3.14}$$

$$\boldsymbol{\Sigma}_{11\cdot 2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}. \tag{3.3.15}$$

Since $\mathbf{X}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{X}_2$ and $\mathbf{X}_2$ are independent normal variables with marginal densities

$$g(\mathbf{x}_1; \boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{x}_2) = \frac{1}{(2\pi)^{k/2}|\boldsymbol{\Sigma}_{11\cdot 2}|^{1/2}} e^{-Q_k(\mathbf{x}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{x}_2; \mathbf{v}_{1\cdot 2}, \boldsymbol{\Sigma}_{11\cdot 2})/2},$$

$$f_2(\mathbf{x}_2; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{(n-k)/2}|\boldsymbol{\Sigma}_{22}|^{1/2}} e^{-Q_{n-k}(\mathbf{x}_2; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})/2},$$

respectively, their joint density is given by $g(\mathbf{x}_1; \boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{x}_2)f_2(\mathbf{x}_2; \boldsymbol{\mu}, \boldsymbol{\Sigma})$. From this joint density function we can rewrite the joint density of $(\mathbf{X}_1, \mathbf{X}_2)'$ by a linear transformation, which yields

$$f(\mathbf{x}_1, \mathbf{x}_2; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = g(\mathbf{x}_1; \boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{x}_2)f_2(\mathbf{x}_2; \boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{3.3.16}$$

But

$$f(\mathbf{x}_1, \mathbf{x}_2; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = f_{1|2}(\mathbf{x}_1; \boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{x}_2)f_2(\mathbf{x}_2; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{3.3.17}$$

also holds where $f_{1|2}$ is the conditional density function of $\mathbf{X}_1$, given $\mathbf{X}_2 = \mathbf{x}_2$. Thus the conditional density function of $\mathbf{X}_1$, given $\mathbf{X}_2 = \mathbf{x}_2$, must be $g(\mathbf{x}_1; \boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{x}_2)$. Since

$$Q_k(\mathbf{x}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{x}_2; \nu_{1\cdot 2}, \boldsymbol{\Sigma}_{11\cdot 2}) = (\mathbf{x}_1 - \boldsymbol{\mu}_{1\cdot 2})'\boldsymbol{\Sigma}_{11\cdot 2}^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_{1\cdot 2}), \tag{3.3.18}$$

where

$$\boldsymbol{\mu}_{1\cdot 2} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \tag{3.3.19}$$

we then obtain

Theorem 3.3.4. *Let* $\mathbf{X}$ *be partitioned as in* (3.3.1). *If* $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} > 0$, *then for any fixed* $k < n$ *the conditional distribution of* $\mathbf{X}_1$, *given* $\mathbf{X}_2 = \mathbf{x}_2$, *is* $\mathcal{N}_k(\boldsymbol{\mu}_{1\cdot 2}, \boldsymbol{\Sigma}_{11\cdot 2})$ *where* $\boldsymbol{\mu}_{1\cdot 2}$ *and* $\boldsymbol{\Sigma}_{11\cdot 2}$ *are defined in* (3.3.19) *and* (3.3.15), *respectively.*

We note in passing that $\boldsymbol{\mu}_{1\cdot 2}$ *is the conditional mean vector and* $\boldsymbol{\Sigma}_{11\cdot 2}$ is the conditional covariance matrix of $\mathbf{X}_1$, given $\mathbf{X}_2 = \mathbf{x}_2$. Furthermore, $\boldsymbol{\mu}_{1\cdot 2}$ is a linear function of $\mathbf{x}_2$ and $\boldsymbol{\Sigma}_{11\cdot 2}$ does not depend on $\mathbf{x}_2$. The matrix $\mathbf{B} = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}$ is called the regression matrix of $\mathbf{X}_2$ on $\mathbf{X}_1$, and will be discussed more extensively in the next section.

3.4. Regression and Correlation

Consider the partition of the components of $\mathbf{X}$ into $\mathbf{X}_1$ and $\mathbf{X}_2$ defined in (3.3.1). In this section we study:

(a) the best predictor of a component of $\mathbf{X}_1$ based on $\mathbf{X}_2 = \mathbf{x}_2$;
(b) the multiple correlation coefficient between a component of $\mathbf{X}_1$ and the components of $\mathbf{X}_2$;
(c) the partial correlation coefficient between two components of $\mathbf{X}_1$, given $\mathbf{X}_2 = \mathbf{x}_2$;
(d) the canonical correlation coefficients between $\mathbf{X}_1$ and $\mathbf{X}_2$; and
(e) the principal components of $\mathbf{X}$.

3.4.1. Best (Linear) Predictors

For fixed $1 \le i \le k$ suppose that we are interested in predicting the value of X_i, given $\mathbf{X}_2 = \mathbf{x}_2$. Let $\hat{x}_i = \lambda(\mathbf{x}_2)$ denote a predictor which is a function of $\mathbf{x}_2$. The problem of interest is to find the optimal choice of such a function. For this purpose we define

Definition 3.4.1. $\hat{x}_i^* = \lambda^*(\mathbf{x}_2)$ is said to be the best predictor of X_i based on $\mathbf{X}_2 = \mathbf{x}_2$, using the loss function

$$L(X_i, \lambda(\mathbf{x}_2)) = (X_i - \lambda(\mathbf{x}_2))^2,$$

if

$$\inf_{\lambda} E[(X_i - \lambda(\mathbf{x}_2))^2 | \mathbf{X}_2 = \mathbf{x}_2] = E[(X_i - \lambda^*(\mathbf{x}_2))^2 | \mathbf{X}_2 = \mathbf{x}_2]$$

holds for all $\mathbf{x}_2$.

For certain multivariate distributions, the best predictor is difficult to find. Since linear functions of $\mathbf{x}_2$ are simpler, we often restrict attention to the subset of all linear functions of $\mathbf{x}_2$ and then obtain the best linear predictor. In the following we show that, for the multivariate normal distribution, the "overall" best predictor is in fact a linear predictor.

To see this, first note that

$$\begin{aligned} E[(X_i - \lambda(\mathbf{x}_2))^2 | \mathbf{X}_2 = \mathbf{x}_2] \\ &= E[\{(X_i - \mu_{i\cdot 2}(\mathbf{x}_2)) + (\mu_{i\cdot 2}(\mathbf{x}_2) - \lambda(\mathbf{x}_2))\}^2 | \mathbf{X}_2 = \mathbf{x}_2] \\ &= \mathrm{Var}(X_i | \mathbf{X}_2 = \mathbf{x}_2) + (\mu_{i\cdot 2}(\mathbf{x}_2) - \lambda(\mathbf{x}_2))^2, \end{aligned}$$

where

$$\mu_{i\cdot 2}(\mathbf{x}_2) = E(X_i | \mathbf{X}_2 = \mathbf{x}_2)$$

is the conditional mean. If $\mathrm{Var}(X_i | \mathbf{X}_2 = \mathbf{x}_2)$ does not depend on $\mathbf{x}_2$, then clearly $E[(X_i - \lambda(\mathbf{x}_2))^2 | \mathbf{X}_2 = \mathbf{x}_2]$ is minimized when the second term is zero.

If $X \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} > 0$, then by Theorems 3.3.4 and 3.3.1 the conditional distribution of X_i, given $\mathbf{X}_2 = \mathbf{x}_2$, is normal with mean

$$\mu_{i\cdot 2}(\mathbf{x}_2) = \mu_i + \boldsymbol{\sigma}_i \boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \tag{3.4.1}$$

(a linear function of $\mathbf{x}_2$) and variance

$$\sigma_{ii\cdot k+1,\ldots,n} = \sigma_{ii} - \boldsymbol{\sigma}_i \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\sigma}_i', \tag{3.4.2}$$

where

$$\boldsymbol{\sigma}_i = (\sigma_{i,k+1}, \ldots, \sigma_{i,n}) \tag{3.4.3}$$

is the ith row of the submatrix $\boldsymbol{\Sigma}_{12}$. Since $\sigma_{ii\cdot k+1,\ldots,n}$ does not depend on $\mathbf{x}_2$, we have

Theorem 3.4.1. *If* $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} > 0$, *then for all* $i \le k$ *the best predictor of* X_i, *based on* $\mathbf{X}_2 = \mathbf{x}_2$, *is* $\mu_{i\cdot 2}(\mathbf{x}_2)$ *given in* (3.4.1).

We note that for given $\mathbf{X}_2 = \mathbf{x}_2$ the smallest value of $E(X_i - \lambda(\mathbf{x}_2))^2$ is $\sigma_{ii\cdot k+1,\ldots,n}$. The infimum occurs, of course, at $\lambda^*(\mathbf{x}_2) = \mu_{i\cdot 2}(\mathbf{x}_2)$. Also note that $\mu_{i\cdot 2}(\mathbf{x}_2)$ is just the ith row of the vector $\boldsymbol{\mu}_1 + \mathbf{B}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$ where $\mathbf{B}$ is the regression matrix defined in (3.3.12).

3.4.2. Multiple Correlation Coefficient

The theory of partial and multiple correlation coefficients treated in this section was originally developed by Pearson (1896) and Yule (1897a, b). The reader is referred to Pearson (1920) for the historical developments.

Assume that $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} > 0$. Then, for fixed $1 \le i \le k$ and given real vector $\mathbf{c}$, the joint distribution of $(X_i, \mathbf{c}'\mathbf{X}_2)'$ can be obtained by the transformation

$$\begin{pmatrix} X_i \\ \mathbf{c}'\mathbf{X}_2 \end{pmatrix} = \begin{pmatrix} \delta_{i1}\ldots\delta_{ik} & 0\ldots 0 \\ 0\ldots 0 & c_1\ldots c_{n-k} \end{pmatrix} \mathbf{X},$$

where

$$\delta_{ij} = \begin{cases} 1 & \text{for} \quad i = j, \\ 0 & \text{otherwise.} \end{cases}$$

By Theorem 3.3.3, this distribution is bivariate normal with means μ_i and $\mathbf{c}'\boldsymbol{\mu}_2$, and variances σ_{ii} and $\mathbf{c}'\boldsymbol{\Sigma}_{22}\mathbf{c}$, respectively, and covariance $\mathbf{c}'\boldsymbol{\sigma}_i$ where $\boldsymbol{\sigma}_i$ is given in (3.4.3). Thus for $\mathbf{c} \neq \mathbf{0}$ the correlation coefficient between X_i and $\mathbf{c}'\mathbf{X}_2$ is simply $\mathbf{c}'\boldsymbol{\sigma}_i'/(\sigma_{ii}\mathbf{c}'\boldsymbol{\Sigma}_{22}\mathbf{c})^{1/2}$.

In certain applications, we are interested in the best linear combination of the components of $\mathbf{X}_2$ such that the correlation coefficient between X_i and $\mathbf{c}'\mathbf{X}_2$ is maximized.

Definition 3.4.2. Let $\mathbf{X}$ be partitioned as in (3.3.1). For $1 \leq i \leq k$ the multiple correlation coefficient between X_i and $\mathbf{X}_2$ is defined by

$$R_{i\cdot k+1,\ldots,n} = \sup_{\mathbf{c}} \operatorname{Corr}(X_i, \mathbf{c}'\mathbf{X}_2). \tag{3.4.4}$$

In the following theorem we show that the $\mathbf{c}'$ vector which maximizes the right-hand side of (3.4.4) is $\boldsymbol{\sigma}_i\boldsymbol{\Sigma}_{22}^{-1}$, the *same* vector that yields the best predictor for X_i, when $\mathbf{X}_2 = \mathbf{x}_2$ is given.

Theorem 3.4.2. *If* $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \boldsymbol{\Sigma} > 0$, *and the components of* $\mathbf{X}$ *are partitioned as in* (3.3.1), *then for every fixed* $i = 1, \ldots, k$ *the supremum of the right-hand side of* (3.4.4) *is attained at* $\mathbf{c}' = \boldsymbol{\sigma}_i\boldsymbol{\Sigma}_{22}^{-1}$, *and*

$$R_{i\cdot k+1,\ldots,n} = \left(\frac{\boldsymbol{\sigma}_i\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_i'}{\sigma_{ii}}\right)^{1/2}. \tag{3.4.5}$$

Proof. We shall follow the core of the argument given in Anderson (1984, p. 40). Since the correlation coefficient does not depend on the means, without loss of generality it may be assumed that $\boldsymbol{\mu} = \mathbf{0}$. By Theorem 3.4.1, the inequality

$$E[(X_i - \boldsymbol{\sigma}_i\boldsymbol{\Sigma}_{22}^{-1}\mathbf{x}_2)^2|\mathbf{X}_2 = \mathbf{x}_2] \leq E[(X_i - \alpha\mathbf{c}'\mathbf{x}_2)^2|\mathbf{X}_2 = \mathbf{x}_2]$$

holds for all real numbers α, real vectors $\mathbf{c}$, and all $\mathbf{x}_2 \in \Re^{n-k}$. After unconditioning we have

$$E(X_i - \boldsymbol{\sigma}_i\boldsymbol{\Sigma}_{22}^{-1}\mathbf{X}_2)^2 \leq E(X_i - \alpha\mathbf{c}'\mathbf{X}_2)^2 \tag{3.4.6}$$

for all α and $\mathbf{c}$. Expanding both sides of (3.4.6) we have

$$\operatorname{Var}(\boldsymbol{\sigma}_i\boldsymbol{\Sigma}_{22}^{-1}\mathbf{X}_2) - 2\operatorname{Cov}(X_i, \boldsymbol{\sigma}_i\boldsymbol{\Sigma}_{22}^{-1}\mathbf{X}_2) \leq \alpha^2 \operatorname{Var}(\mathbf{c}'\mathbf{X}_2) - 2\alpha\operatorname{Cov}(X_i, \mathbf{c}'\mathbf{X}_2).$$

After rearranging the terms and dividing $(\sigma_{ii}\operatorname{Var}(\boldsymbol{\sigma}_i\boldsymbol{\Sigma}_{22}^{-1}\mathbf{X}_2))^{1/2}$ throughout, we then obtain

$$\frac{\operatorname{Cov}(X_i, \boldsymbol{\sigma}_i\boldsymbol{\Sigma}_{22}^{-1}\mathbf{X}_2) - \alpha\operatorname{Cov}(X_i, \mathbf{c}'\mathbf{X}_2)}{(\sigma_{ii}\operatorname{Var}(\boldsymbol{\sigma}_i\boldsymbol{\Sigma}_{22}^{-1}\mathbf{X}_2))^{1/2}}$$
$$\geq \frac{1}{2}\left[\left(\frac{\operatorname{Var}(\boldsymbol{\sigma}_i\boldsymbol{\Sigma}_{22}^{-1}\mathbf{X}_2)}{\sigma_{ii}}\right)^{1/2} - \frac{\alpha^2\operatorname{Var}(\mathbf{c}'\mathbf{X}_2)}{(\sigma_{ii}\operatorname{Var}(\boldsymbol{\sigma}_i\boldsymbol{\Sigma}_{22}^{-1}\mathbf{X}_2))^{1/2}}\right].$$

The inequality

$$\operatorname{Corr}(X_i, \boldsymbol{\sigma}_i\boldsymbol{\Sigma}_{22}^{-1}\mathbf{X}_2) - \operatorname{Corr}(X_i, \mathbf{c}'\mathbf{X}_2) \geq 0$$

now follows by choosing

$$\alpha = \left(\frac{\operatorname{Var}(\boldsymbol{\sigma}_i\boldsymbol{\Sigma}_{22}^{-1}\mathbf{X}_2)}{\operatorname{Var}(\mathbf{c}'\mathbf{X}_2)}\right)^{1/2}.$$

Consequently, we have

$$\begin{aligned} R_{i\cdot k+1,\ldots,n} &= \frac{\operatorname{Cov}(X_i, \boldsymbol{\sigma}_i\boldsymbol{\Sigma}_{22}^{-1}\mathbf{X}_2)}{(\sigma_{ii}\operatorname{Var}(\boldsymbol{\sigma}_i\boldsymbol{\Sigma}_{22}^{-1}\mathbf{X}_2))^{1/2}} \\ &= \frac{\boldsymbol{\sigma}_i\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_i'}{(\sigma_{ii}(\boldsymbol{\sigma}_i\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_i'))^{1/2}} \\ &= \left(\frac{\boldsymbol{\sigma}_i\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_i'}{\sigma_{ii}}\right)^{1/2}. \end{aligned}$$ □

For the nonsingular multivariate normal distribution, the multiple correlation coefficient given in (3.4.5) is always larger than or equal to zero and less than or equal to one. Furthermore, since $\boldsymbol{\Sigma}_{22}^{-1}$ is positive definite (because $\boldsymbol{\Sigma}_{22}$ is positive definite), it is equal to zero if and only if $\boldsymbol{\sigma}_i = \mathbf{0}$; that is, if and only if X_i and $\mathbf{X}_2$ are independent.

Since $(X_i, \boldsymbol{\sigma}_i\boldsymbol{\Sigma}_{22}^{-1}\mathbf{X}_2)'$ has a bivariate normal distribution with means μ_i, $\boldsymbol{\sigma}_i\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\mu}_2$, variances σ_{ii}, $\boldsymbol{\sigma}_i\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_i'$, and correlation coefficient $R_{i\cdot k+1,\ldots,n}$, the conditional distribution of X_i, given $\boldsymbol{\sigma}_i\boldsymbol{\Sigma}_{22}^{-1}\mathbf{x}_2$, is normal with variance

$$\sigma_{ii}(1 - R_{i\cdot k+1,\ldots,n}^2) = \sigma_{ii} - \boldsymbol{\sigma}_i\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_i'. \tag{3.4.7}$$

This is the smallest possible variance of the conditional distribution of X_i, given $\mathbf{c}'\mathbf{X}_2 = \mathbf{c}'\mathbf{x}_2$, where $\mathbf{c}$ is a nonzero real vector, and is obtained when $\mathbf{c}'$ is chosen to be $\boldsymbol{\sigma}_i\boldsymbol{\Sigma}_{22}^{-1}$.

3.4.3. Partial Correlation Coefficients

The partial correlation coefficient between two random variables is their correlation coefficient after allowing for the effects of a set of other variables. For $i, j = 1, \ldots, k$, $i \neq j$, if we consider the correlation between X_i and X_j, when $\mathbf{X}_2 = (X_{k+1}, \ldots, X_n)'$ is fixed, then this correlation coefficient can be obtained from the conditional distribution of $(X_i, X_j)'$, given $\mathbf{X}_2 = \mathbf{x}_2$.

Definition 3.4.3. Let $\mathbf{X}$ be partitioned as in (3.3.1). Then for given $\mathbf{X}_2 = \mathbf{x}_2$ the partial correlation coefficient between X_i and X_j is

$$\rho_{ij\cdot k+1,\ldots,n} = \frac{\operatorname{Cov}((X_i, X_j)|\mathbf{X}_2 = \mathbf{x}_2)}{(\operatorname{Var}(X_i|\mathbf{X}_2 = \mathbf{x}_2)\operatorname{Var}(X_j|\mathbf{X}_2 = \mathbf{x}_2))^{1/2}}$$

for $i, j = 1, \ldots, k$, $i \neq j$.

For the general case, the partial correlation coefficient might depend on $\mathbf{x}_2$. But for the multivariate normal distribution the result is quite simple, depending only on the elements of the covariance matrix.

Theorem 3.4.3. *If* $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \boldsymbol{\Sigma} > 0$, *and the components of* $\mathbf{X}$ *are partitioned as in* (3.3.1), *then*

$$\rho_{ij\cdot k+1,\ldots,n} = \frac{\sigma_{ij} - \boldsymbol{\sigma}_i \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\sigma}_j'}{((\sigma_{ii} - \boldsymbol{\sigma}_i \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\sigma}_i')(\sigma_{jj} - \boldsymbol{\sigma}_j \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\sigma}_j'))^{1/2}} \tag{3.4.8}$$

for $i, j = 1, \ldots, k, i \neq j$.

PROOF. By Theorem 3.3.4, the conditional distribution of $\mathbf{X}_1$, given $\mathbf{X}_2 = \mathbf{x}_2$, is normal with the conditional covariance matrix $\boldsymbol{\Sigma}_{11\cdot 2} = \boldsymbol{\Sigma}_{12} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. Thus by Theorem 3.3.1 the conditional distribution of $(X_i, X_j)'$, given $\mathbf{X}_2 = \mathbf{x}_2$, is bivariate normal with the conditional covariance matrix

$$\begin{pmatrix} \sigma_{ii} & \sigma_{ij} \\ \sigma_{ij} & \sigma_{jj} \end{pmatrix} - \begin{pmatrix} \boldsymbol{\sigma}_i \\ \boldsymbol{\sigma}_j \end{pmatrix} \boldsymbol{\Sigma}_{22}^{-1} (\boldsymbol{\sigma}_i' \ \ \boldsymbol{\sigma}_j'). \qquad \square$$

Note that the partial correlation coefficient in (3.4.8) is nonnegative if and only if $\sigma_{ij} \geq \boldsymbol{\sigma}_i \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\sigma}_j'$. Thus it is possible to have a covariance matrix $\boldsymbol{\Sigma}$ such that the correlation coefficient between X_i and X_j (which is $\rho_{ij} = \sigma_{ij}/\sqrt{\sigma_{ii}\sigma_{jj}}$) is positive while the partial correlation coefficient is negative. As an illustration consider the following example:

EXAMPLE 3.4.1. Let $\mathbf{X} = (X_1, X_2, X_3)'$ be distributed according to an $\mathcal{N}_3(\mathbf{0}, \boldsymbol{\Sigma})$ distribution, where $\sigma_{ii} = 1$ $(i = 1, 2, 3)$ and $\sigma_{12} = 1 - 2\varepsilon$, $\sigma_{13} = \sigma_{23} = 1 - \varepsilon$, $0 < \varepsilon < \frac{1}{2}$. For every fixed $\mathbf{c} = (c_1, c_2, c_3)' \in \Re^3$ we have

(a) $$\begin{aligned} \mathbf{c}'\boldsymbol{\Sigma}\mathbf{c} &= (c_1^2 + c_2^2 + c_3^2) + 2(1 - \varepsilon)(c_1c_2 + c_1c_3 + c_2c_3) - 2\varepsilon c_1 c_2 \\ &= (1 - \varepsilon)(c_1 + c_2 + c_3)^2 + \varepsilon(c_1 - c_2)^2 + \varepsilon c_3^2. \end{aligned}$$

Since $\mathbf{c}'\boldsymbol{\Sigma}\mathbf{c} \geq 0$ holds, and the equality holds if and only if $c_1 = c_2 = c_3 = 0$, $\boldsymbol{\Sigma}$ is a positive definite matrix.

(b) The conditional distribution of $(X_1, X_2)'$, given $X_3 = x_3$, is bivariate normal with the covariance matrix

$$\boldsymbol{\Sigma}_{11\cdot 2} = \begin{pmatrix} 1 & 1 - 2\varepsilon \\ 1 - 2\varepsilon & 1 \end{pmatrix} - (1 - \varepsilon)^2 \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} = \varepsilon \begin{pmatrix} 2 - \varepsilon & -\varepsilon \\ -\varepsilon & 2 - \varepsilon \end{pmatrix}.$$

Thus $\rho_{12} = 1 - 2\varepsilon > 0$ and $\rho_{12\cdot 3} = -\varepsilon/(2 - \varepsilon) < 0$. $\square$

3.4.4. Canonical Correlation Coefficients

The theory of canonical correlation was developed by Hotelling (1936), and may be regarded as a generalization of the notion of the multiple correlation.

Let us again consider the partition of $\mathbf{X}$ into $\mathbf{X}_1$ and $\mathbf{X}_2$, defined in (3.3.1), where $\mathbf{X}_1 = (X_1, \ldots, X_k)'$ and $\mathbf{X}_2 = (X_{k+1}, \ldots, X_n)'$. Recall that the multiple correlation coefficient between X_i and $\mathbf{X}_2$ is the largest possible correlation coefficient between X_i and $\mathbf{c}'\mathbf{X}_2$ over all possible choices of real vectors $\mathbf{c}$, where $1 \le i \le k$ is fixed. In canonical correlation analysis we are interested in finding two real vectors $\mathbf{c}_1, \mathbf{c}_2$ such that the correlation coefficient between $\mathbf{c}_1'\mathbf{X}_1$ and $\mathbf{c}_2'\mathbf{X}_2$ is maximized.

This maximization process can be carried out in the following fashion: First, we choose $\mathbf{c}_2$ to maximize the correlation coefficient between $\mathbf{c}_1'\mathbf{X}_1$ and $\mathbf{c}_2'\mathbf{X}_2$ for fixed $\mathbf{c}_1$, then we find its maximum over all possible choices of $\mathbf{c}_1$. Second, we choose $\mathbf{c}_1$ to maximize the correlation coefficient between $\mathbf{c}_2'\mathbf{X}_2$ and $\mathbf{c}_1'\mathbf{X}_1$ for fixed $\mathbf{c}_2$, then we find the optimal solution for $\mathbf{c}_2$. After completing these two steps we then choose the larger of the two resulting correlation coefficients.

Now, for every fixed $\mathbf{c}_1$ and $X_0 = \mathbf{c}_1'\mathbf{X}_1$ (say), the best choice of $\mathbf{c}_2'$ is simply the regression vector of $\mathbf{X}_2$ on X_0. Thus the largest possible correlation coefficient is just $(\mathbf{c}_1'\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\mathbf{c}_1/\mathbf{c}_1'\boldsymbol{\Sigma}_{11}\mathbf{c}_1)^{1/2}$, the multiple correlation coefficient between X_0 and $\mathbf{X}_2$. Furthermore, since correlation coefficients are scale invariant, without loss of generality it may be assumed that

$$\mathbf{c}_1'\boldsymbol{\Sigma}_{11}\mathbf{c}_1 = 1. \tag{3.4.9}$$

Using Lagrange's method of multipliers, this amounts to the maximization of

$$g(\mathbf{c}_1, \lambda) = \mathbf{c}_1'\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\mathbf{c}_1 - \lambda(\mathbf{c}_1'\boldsymbol{\Sigma}_{11}\mathbf{c}_1 - 1), \tag{3.4.10}$$

subject to the constraint in (3.4.9). After taking partial derivatives with respect to the components of $\mathbf{c}_1$ and letting them equal zero, we have

$$(-\lambda\boldsymbol{\Sigma}_{11} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})\mathbf{c}_1 = \mathbf{0}. \tag{3.4.11}$$

Multiplying the left-hand side of (3.4.11) by $\mathbf{c}_1'$ and using the identity in (3.4.9), we then obtain

$$\lambda = \mathbf{c}_1'\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\mathbf{c}_1. \tag{3.4.12}$$

For $\mathbf{c}_1$ to have a nontrivial solution in (3.4.11) we must have

$$h_1(\lambda) = |-\lambda\boldsymbol{\Sigma}_{11} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}| = 0. \tag{3.4.13}$$

But $h_1(\lambda)$ is a polynomial of degree k. It can be verified that (see Anderson (1984, p. 483)) if $\boldsymbol{\Sigma}$ is positive definite, then $h_1(\lambda)$ has k nonnegative real roots.

Similarly, the maximization of the multiple correlation coefficient $(\mathbf{c}_2'\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\mathbf{c}_2/\mathbf{c}_2'\boldsymbol{\Sigma}_{22}\mathbf{c}_2)^{1/2}$, subject to the constraint

$$\mathbf{c}_2'\boldsymbol{\Sigma}_{22}\mathbf{c}_2 = 1, \tag{3.4.14}$$

leads to the equation

$$(-\lambda\boldsymbol{\Sigma}_{22} + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})\mathbf{c}_2 = \mathbf{0}. \tag{3.4.15}$$

In order to have a nonsingular solution we must have

$$h_2(\lambda) = |-\lambda\mathbf{\Sigma}_{22} + \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}| = 0, \tag{3.4.16}$$

which has $n - k$ nonnegative real roots.

Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the collection of roots of the two equations $h_1(\lambda) = 0$ and $h_2(\lambda) = 0$ and, without loss of generality, assume that

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0. \tag{3.4.17}$$

Then by (3.4.12) the largest canonical correlation coefficient is simply $\sqrt{\lambda_1}$. The vectors $\mathbf{c}_1, \mathbf{c}_2$, which yield this largest canonical correlation coefficient, can be obtained from either (3.4.11) or (3.4.15) with $\lambda = \lambda_1$, depending on which equation has the root λ_1. Let $(\mathbf{c}_{1,1}, \mathbf{c}_{2,1})$ denote such a solution. Then the random variables $\mathbf{c}'_{1,1}\mathbf{X}_1, \mathbf{c}'_{2,1}\mathbf{X}_2$ are called the first pair of canonical variables.

This process can be continued to find all the λ_j values and the corresponding canonical variables. Without loss of generality, let

$$\lambda_1 > \lambda_2 > \cdots > \lambda_r \geq 0$$

denote the r $(r \leq n)$ distinct roots of $h_1(\lambda) = 0$ and $h_2(\lambda) = 0$. Then $\sqrt{\lambda_j}$ is called the jth canonical correlation coefficient, and the corresponding vector $(\mathbf{c}'_{1,j}\mathbf{X}_1, \mathbf{c}'_{2,j}\mathbf{X}_2)'$ is called the jth pair of canonical variables. Using Lagrange's method of multipliers it can be shown that (see Anderson (1984, p. 484)) the vectors $\{\mathbf{c}_{1,j}\}_{j=1}^r, \{\mathbf{c}_{2,j}\}_{j=1}^r$ also satisfy the conditions, for all $s \neq t$,

(i) $\mathbf{c}'_{1,s}\mathbf{X}_1$ and $\mathbf{c}'_{1,t}\mathbf{X}_1$ are independent, (3.4.18)

(ii) $\mathbf{c}'_{2,s}\mathbf{X}_2$ and $\mathbf{c}'_{2,t}\mathbf{X}_2$ are independent, (3.4.19)

(iii) $\mathbf{c}'_{1,s}\mathbf{X}_1$ and $\mathbf{c}'_{2,t}\mathbf{X}_2$ are independent. (3.4.20)

Summarizing the above result, we say

Definition 3.4.4. Let $\mathbf{X}$ be partitioned as in (3.3.1), and let

$$\tau_j = \sup_{\mathbf{c}_{1,j},\mathbf{c}_{2,j}} \operatorname{Corr}(\mathbf{c}'_{1,j}\mathbf{X}_1, \mathbf{c}'_{2,j}\mathbf{X}_2)$$

subject to (3.4.18)–(3.4.20) and the condition

$$\mathbf{c}'_{1,j}\mathbf{\Sigma}_{11}\mathbf{c}_{1,j} = \mathbf{c}'_{2,j}\mathbf{\Sigma}_{22}\mathbf{c}_{2,j} = 1.$$

The distinct values $\tau_1 > \tau_2 > \cdots > \tau_r$ $(r \leq n)$ are called the canonical correlation coefficients between $\mathbf{X}_1$ and $\mathbf{X}_2$.

If $\mathbf{X}$ is a nonsingular $\mathcal{N}_n(\boldsymbol{\mu}, \mathbf{\Sigma})$ variable, then $\mathbf{\Sigma}_{11}$ and $\mathbf{\Sigma}_{22}$ are both nonsingular. Thus we obtain

Theorem 3.4.4. *If* $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \mathbf{\Sigma})$, $\mathbf{\Sigma} > 0$, *and* $\mathbf{X}$ *is partitioned as in* (3.3.1), *then the jth canonical correlation coefficient between* $\mathbf{X}_1$ *and* $\mathbf{X}_2$ *is* $\sqrt{\lambda_j}, j = 1, \ldots, r$, *where* $\lambda_1 > \lambda_2 > \cdots > \lambda_r \geq 0$ *are the r distinct roots of* $h_1(\lambda) = 0$ *and* $h_2(\lambda) = 0$, *defined in* (3.4.13) *and* (3.4.16), *respectively.*

3.4.5. Principal Components

Principal component analysis, originally proposed and studied by Hotelling (1933), concerns a method for obtaining a set of linear combinations of components of an n-dimensional random variable with certain desirable properties. Suppose that $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \boldsymbol{\Sigma} > 0$. When the X_i's are independent, then a measure of dispersion of the distribution of $\mathbf{X}$ is the sum of the variances (which are the diagonal elements of $\boldsymbol{\Sigma}$). Furthermore, the larger the variance of X_i, the more it contributes to this dispersion. Thus the problem of interest is to define and obtain the principal (or the most influential) components with large variances when the X_i's are correlated. In this case it is not adequate just to consider each of the components separately because they tend to hang together.

In principal component analysis, we look for linear combinations of the X_i's such that the variances are maximized under certain constraints. Let $\mathbf{c}_1' = (c_{11}, \ldots, c_{1n})$ be a real vector such that $\mathbf{c}_1'\mathbf{c}_1 = 1$ and

$$\sup_{\{\boldsymbol{\alpha}:\boldsymbol{\alpha}'\boldsymbol{\alpha}=1\}} \mathrm{Var}(\boldsymbol{\alpha}'\mathbf{X}) = \mathrm{Var}(\mathbf{c}_1'\mathbf{X}).$$

Then $\mathbf{c}_1$ is the vector with norm 1 such that the variance of $\mathbf{c}_1'\mathbf{X}$ is maximized over all linear combinations of the components under this constraint. To find $\mathbf{c}_1$ note that $\mathrm{Var}(\boldsymbol{\alpha}'\mathbf{X}) = \boldsymbol{\alpha}'\boldsymbol{\Sigma}\boldsymbol{\alpha}$ for all $\boldsymbol{\alpha}$. Thus, by Lagrange's method of multipliers, this amounts to maximizing the function

$$g_1(\boldsymbol{\alpha}, \lambda) = \boldsymbol{\alpha}'\boldsymbol{\Sigma}\boldsymbol{\alpha} - \lambda(\boldsymbol{\alpha}'\boldsymbol{\alpha} - 1),$$

subject to $\boldsymbol{\alpha}'\boldsymbol{\alpha} = 1$. By calculus it follows that

$$\frac{\partial}{\partial \boldsymbol{\alpha}} g_1(\boldsymbol{\alpha}, \lambda) = 2(\boldsymbol{\Sigma}\boldsymbol{\alpha} - \lambda\boldsymbol{\alpha}) = 2(\boldsymbol{\Sigma} - \lambda\mathbf{I}_n)\boldsymbol{\alpha},$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix. The system of linear equations $(\partial/\partial\boldsymbol{\alpha})g_1(\boldsymbol{\alpha}, \lambda) = 0$ has a nontrivial solution if and only if

$$h(\lambda) \equiv |\boldsymbol{\Sigma} - \lambda\mathbf{I}_n| = 0 \tag{3.4.21}$$

holds. Thus λ must be an eigenvalue of $\boldsymbol{\Sigma}$. Furthermore, if $\mathbf{c}_1$ satisfies $2(\boldsymbol{\Sigma} - \lambda\mathbf{I}_n)\mathbf{c}_1 = 0$ then, by $\boldsymbol{\alpha}'\boldsymbol{\alpha} = 1$, we must have

$$\boldsymbol{\Sigma}\mathbf{c}_1 = \lambda\mathbf{c}_1, \tag{3.4.22}$$

and

$$\mathbf{c}_1'\boldsymbol{\Sigma}\mathbf{c}_1 = \lambda\mathbf{c}_1'\mathbf{I}_n\mathbf{c}_1 = \lambda. \tag{3.4.23}$$

Thus λ is actually the variance of $\mathbf{c}_1'\mathbf{X} \equiv Y_1$. Let λ_1 denote the value of this λ.

After Y_1 and λ_1 are obtained, we then look for another randm variable $Y_2 = \mathbf{c}_2'\mathbf{X}$ such that:

(i) $\mathbf{c}_2'\mathbf{c}_2 = 1$;
(ii) Y_2 is independent of Y_1; and

(iii) Y_2 has the largest variance among all linear combinations of components of $\mathbf{X}$ that satisfy (i) and (ii).

If $\boldsymbol{\alpha}$ is any vector that satisfies (i) and (ii), then $\boldsymbol{\alpha}'\boldsymbol{\alpha} = 1$ and, by (3.4.22),

$$\operatorname{Cov}(Y_1, Y_2) = \mathbf{c}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha} = \boldsymbol{\alpha}'\boldsymbol{\Sigma}\mathbf{c}_1 = \lambda_1\boldsymbol{\alpha}'\mathbf{c}_1 = 0; \tag{3.4.24}$$

that is, $\boldsymbol{\alpha}$ and $\mathbf{c}_1$ must be orthogonal. Applying Lagrange's method of multipliers one more time leads to the maximization of the function

$$g_2(\boldsymbol{\alpha}, \lambda, \eta) = \boldsymbol{\alpha}'\boldsymbol{\Sigma}\boldsymbol{\alpha} - \lambda(\boldsymbol{\alpha}'\boldsymbol{\alpha} - 1) - \eta\boldsymbol{\alpha}'\boldsymbol{\Sigma}\mathbf{c}_1,$$

subject to $\boldsymbol{\alpha}'\boldsymbol{\alpha} = 1$ and (3.4.24). By

$$\frac{\partial}{\partial\boldsymbol{\alpha}} g_2(\boldsymbol{\alpha}, \lambda, \eta) = 2(\boldsymbol{\Sigma}\boldsymbol{\alpha} - \lambda\boldsymbol{\alpha} - \eta\boldsymbol{\Sigma}\mathbf{c}_1)$$

and (3.4.24) it follows that if $\mathbf{c}_2$ is a solution of $(\partial/\partial\boldsymbol{\alpha})g_2(\boldsymbol{\alpha}, \lambda, \eta) = 0$, then

$$\mathbf{c}_1'\boldsymbol{\Sigma}\mathbf{c}_2 - \lambda\mathbf{c}_1'\mathbf{c}_2 - \eta\mathbf{c}_1'\boldsymbol{\Sigma}\mathbf{c}_1 = -\eta\mathbf{c}_1'\boldsymbol{\Sigma}\mathbf{c}_1 = 0,$$

$$\mathbf{c}_2'\boldsymbol{\Sigma}\mathbf{c}_2 - \lambda\mathbf{c}_2'\mathbf{c}_2 - \eta\mathbf{c}_2'\boldsymbol{\Sigma}\mathbf{c}_1 = 0.$$

Thus we have $\eta = 0$. This implies that $\mathbf{c}_2$ and λ also satisfy the equations

$$\lambda = \mathbf{c}_2'\boldsymbol{\Sigma}\mathbf{c}_2, \qquad (\boldsymbol{\Sigma} - \lambda\mathbf{I}_n)\mathbf{c}_2 = 0.$$

Consequently, if $\mathbf{c}_2$ has a nontrivial solution, then λ also satisfies (3.4.21). Let the value of this λ be denoted by λ_2.

It is known that if $\boldsymbol{\Sigma}$ is an $n \times n$ positive definite matrix, then it has n positive real eigenvalues. Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues and, without loss of generality, assume that

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0. \tag{3.4.25}$$

Then using a similar argument we can continue this process to find n real vectors $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_n$ such that:

(i) $\mathbf{c}_i'\mathbf{c}_i = 1$ $(i = 1, \ldots, n)$;
(ii) $\mathbf{c}_i'\mathbf{c}_j = 0$ for all $i \neq j$; and
(iii) the variance of $Y_i = \mathbf{c}_i'\mathbf{X}$ is λ_i $(i = 1, \ldots, n)$.

Expressing the linear transformation in a matrix form we have

Theorem 3.4.5. *Let* $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} > 0$, *and let* $\lambda_1, \ldots, \lambda_n$ *be the eigenvalues of* $\boldsymbol{\Sigma}$ *satisfying* (3.4.25). *Then there exists an orthogonal matrix* $\mathbf{C} = (\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_n)$ *satisfying* $\mathbf{Y} = \mathbf{C}'\mathbf{X} \sim \mathcal{N}_n(\mathbf{C}'\boldsymbol{\mu}, \mathbf{D})$, *where* $\mathbf{D} = (d_{ij})$ *is a diagonal matrix such that* $d_{ii} = \lambda_i$ $(i = 1, \ldots, n)$.

We now provide a formal definition of the principal components of $\mathbf{X}$ when it has a multivariate normal distribution.

Definition 3.4.5. Let $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} > 0$, and let $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ be the random variable defined in Theorem 3.4.5. Then Y_i is said to be the ith principal component of $\mathbf{X}$ $(i = 1, \ldots, n)$.

Remark 3.4.1. If the components are independent, then $\mathbf{\Sigma}$ is already a diagonal matrix. In this case, Y_1 is the component of $\mathbf{X}$ with the largest variance, Y_2 is the component of $\mathbf{X}$ with the second largest variance, and so on; and $\mathbf{c}_i' = (0, \ldots, 0, 1, 0, \ldots, 0)$ which has a "1" in one of the n positions.

Remark 3.4.2. As a measure of the contributions to the sum of the variances of the Y_i's, the ratios $\lambda_i / \sum_{i=1}^n \lambda_j$ $(i = 1, \ldots, n)$ are of interest. In particular, $\lambda_1 / \sum_{i=1}^n \lambda_j$ represents the contribution of the first principal component of $\mathbf{X}$.

We note in passing that applications of the results of principal component analysis are not limited to the multivariate normal distribution because Theorem 3.4.5 does not require the assumption of normality.

3.4.6. An Example

We complete this section with an example.

Example 3.4.2. Let $n = 5$, $k = 2$, and $\mathbf{X} \sim \mathcal{N}_5(\boldsymbol{\mu}, \mathbf{\Sigma})$. Consider the partition $\mathbf{X}_1 = (X_1, X_2)'$, $\mathbf{X}_2 = (X_3, X_4, X_5)'$,

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \qquad \mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix},$$

where

$$\mathbf{\Sigma}_{11} = \begin{pmatrix} 1 & \rho_2 \\ \rho_2 & 1 \end{pmatrix}, \qquad \mathbf{\Sigma}_{22} = \begin{pmatrix} 1 & \rho_2 & \rho_2 \\ \rho_2 & 1 & \rho_2 \\ \rho_2 & \rho_2 & 1 \end{pmatrix},$$

$$\mathbf{\Sigma}_{12} = \mathbf{\Sigma}_{21}' = \begin{pmatrix} \rho_1 & \rho_1 & \rho_1 \\ \rho_1 & \rho_1 & \rho_1 \end{pmatrix},$$

and $0 \leq \rho_1 \leq \rho_2 < 1$. That is, the random variables are partitioned into two groups; the correlation coefficients within each group are ρ_2, and the correlation coefficients between groups are ρ_1.

(a) $\mathbf{X}$ is a nonsingular normal variable, i.e., $\mathbf{\Sigma} > 0$. To see this, for all nonzero vectors $\mathbf{c}' = (c_1, c_2, c_3, c_4, c_5) \neq \mathbf{0}$ we have

$$\begin{aligned} &\mathbf{c}'\mathbf{\Sigma}\mathbf{c} \\ &= \sum_{i=1}^{5} c_i^2 + 2\rho_2(c_1c_2 + c_3(c_4 + c_5) + c_4c_5) + 2\rho_1(c_1 + c_2)(c_3 + c_4 + c_5) \\ &= \left(\sqrt{\rho_1} \sum_{i=1}^{5} c_i\right)^2 + (\rho_2 - \rho_1)((c_1 + c_2)^2 + (c_3 + c_4 + c_5)^2) + (1 - \rho_2) \sum_{i=1}^{5} c_i^2 \\ &> 0. \end{aligned}$$

(b) Simple calculation shows

$$\Sigma_{22}^{-1} = \frac{1}{(1+2\rho_2)(1-\rho_2)} \begin{pmatrix} 1+\rho_2 & -\rho_2 & -\rho_2 \\ -\rho_2 & 1+\rho_2 & -\rho_2 \\ -\rho_2 & -\rho_2 & 1+\rho_2 \end{pmatrix}.$$

Thus

$$\Sigma_{12}\Sigma_{22}^{-1} = \frac{\rho_1}{1+2\rho_2}\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix},$$

$$\Sigma_{11\cdot 2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

$$= \frac{1}{1+2\rho_2}\begin{pmatrix} 1+2\rho_2-3\rho_1^2 & \rho_2+2\rho_2^2-3\rho_1^2 \\ \rho_2+2\rho_2^2-3\rho_1^2 & 1+2\rho_2-3\rho_1^2 \end{pmatrix}; \qquad (3.4.26)$$

and the conditional distribution of $\mathbf{X}_1$, given $\mathbf{X}_2 = \mathbf{x}_2$, is normal with mean vector

$$\begin{pmatrix} \mu_1 + \dfrac{\rho_1}{1+2\rho_2}\sum_{i=3}^{5}(x_j-\mu_j) \\ \mu_2 + \dfrac{\rho_1}{1+2\rho_2}\sum_{j=3}^{5}(x_j-\mu_j) \end{pmatrix}$$

and covariance matrix $\Sigma_{11\cdot 2}$ given in (3.4.26).

(c) The best predictor for X_i $(i = 1, 2)$, given $(X_3, X_4, X_5)' = (x_3, x_4, x_5)'$, is

$$\hat{x}_i = \mu_i + \frac{\rho_1}{1+2\rho_2}\sum_{j=3}^{5}(x_j-\mu_j).$$

(d) The multiple correlation coefficient between X_i and $\mathbf{X}_2$ is

$$R_{i\cdot 345} = \frac{\sqrt{3}\rho_1}{\sqrt{1+2\rho_2}}, \qquad i = 1, 2,$$

and $R_{i\cdot 345} = 0$ if and only if $\rho_1 = 0$. When $\rho_1 = \rho_2 = \rho$, it becomes $\sqrt{3}\rho/\sqrt{1+2\rho}$.

(e) The partial correlation coefficient between X_1 and X_2 is

$$\rho_{12\cdot 345} = \frac{\rho_2+2\rho_2^2-3\rho_1^2}{1+2\rho_2-3\rho_1^2},$$

and is equal to ρ_2 when $(X_1, X_2)'$ and $(X_3, X_4, X_5)'$ are independent. When $\rho_1 = \rho_2 = \rho$, it reduces to $\rho/(1+3\rho)$.

(f) The determinants of the matrices $-\lambda\Sigma_{11} + \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ and $-\lambda\Sigma_{22} + \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$ are, respectively,

$$h_1(\lambda) = (1-\rho_2)\lambda\left[(1+\rho_2)\lambda - \frac{6\rho_1^2}{1+2\rho_2}\right], \qquad (3.4.27)$$

$$h_2(\lambda) = -(1-\rho_2)^2\lambda^2\left[(1+2\rho_2)\lambda - \frac{6\rho_1^2}{1+\rho_2}\right]. \qquad (3.4.28)$$

Thus $h_1(\lambda) = 0$ and $h_2(\lambda) = 0$ have a common unique positive root $\lambda_1 = 6\rho_1^2/((1+\rho_2)(1+2\rho_2))$ and all the other roots are zero. Consequently, it follows that:

(i) the largest canonical correlation coefficient between $\mathbf{X}_1$ and $\mathbf{X}_2$ is $\sqrt{6}\rho_1/\sqrt{(1+\rho_2)(1+2\rho_2)}$, which is $\sqrt{6}\rho/\sqrt{(1+\rho)(1+2\rho)}$ when $\rho_1 = \rho_2 = \rho$;

(ii) the canonical variables that yield this canonical correlation coefficient can be obtained by either finding a solution for $\mathbf{c}_1$ in (3.4.11) or finding a solution for $\mathbf{c}_2$ in (3.4.15), with $\lambda = \lambda_1$;

(iii) all other pairs of canonical variables that are uncorrelated with (hence independent of) the first pair must also be independent.

This is so because all other canonical correlation coefficients are zero.

(g) It is straightforward to verify that

$$|\mathbf{\Sigma} - \lambda \mathbf{I}_5| = (1 - \lambda - \rho_2)^3[(1 - \lambda + \rho_2)(1 - \lambda + 2\rho_2) - 6\rho_1^2]. \quad (3.4.29)$$

Thus the eigenvalues of $\mathbf{\Sigma}$ are

$$\lambda_1 = 1 + \tfrac{3}{2}\rho_2 + \tfrac{1}{2}(\rho_2^2 + 24\rho_1^2)^{1/2},$$
$$\lambda_2 = 1 + \tfrac{3}{2}\rho_2 - \tfrac{1}{2}(\rho_2^2 + 24\rho_1^2)^{1/2},$$
$$\lambda_3 = \lambda_4 = \lambda_5 = 1 - \rho_2.$$

In the special case when $\rho_1 = \rho_2 = \rho$, we have

$$\lambda_1 = 1 + 4\rho, \qquad \lambda_2 = \lambda_3 = \lambda_4 = \lambda_5 = 1 - \rho;$$

and a set of solutions for $\mathbf{c}_i$ in $(\mathbf{\Sigma} - \lambda_i \mathbf{I}_5)\mathbf{c}_i = \mathbf{0}$ is

$$\mathbf{c}_1' = \frac{1}{\sqrt{5}}(1\ \ 1\ \ 1\ \ 1\ \ 1), \qquad \mathbf{c}_2' = \frac{1}{\sqrt{20}}(-4\ \ 1\ \ 1\ \ 1\ \ 1),$$

$$\mathbf{c}_3' = \frac{1}{\sqrt{12}}(0\ \ -3\ \ 1\ \ 1\ \ 1), \qquad \mathbf{c}_4' = \frac{1}{\sqrt{6}}(0\ \ 0\ \ -2\ \ 1\ \ 1),$$

and

$$\mathbf{c}_5' = \frac{1}{\sqrt{2}}(0\ \ 0\ \ 0\ \ -1\ \ 1).$$

Thus the orthogonal matrix $\mathbf{C} = (\mathbf{c}_1\ \mathbf{c}_2\ \mathbf{c}_3\ \mathbf{c}_4\ \mathbf{c}_5)'$ satisfies the condition that $\mathbf{Y} = \mathbf{C}'\mathbf{X} \sim N_5(\mathbf{C}'\boldsymbol{\mu}, \mathbf{D})$, where $\mathbf{D}$ is the diagonal matrix with diagonal elements $\lambda_1, \ldots, \lambda_5$. The components of $\mathbf{Y}$, $Y_i = \mathbf{c}_i'\mathbf{X}$ $(i = 1, \ldots, 5)$, are the principal components of $\mathbf{X}$. When $\rho_1 = \rho_2 = \rho$, the variance of the first principal component $Y_1 = \mathbf{c}_1'\mathbf{X}$ is $\lambda_1 = \mathbf{c}_1'\mathbf{\Sigma}\mathbf{c}_1 = 1 + 4\rho$, and its contribution to the sum of the variances of the principal components is $20(1 + 4\rho)\%$. □

3.5. Sampling Distributions

For fixed positive integer N let $\mathbf{X}_1, \ldots, \mathbf{X}_N$ be a random sample of size N from an $\mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution, that is, $\mathbf{X}_1, \ldots, \mathbf{X}_N$ are i.i.d. random variables with a common $\mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution. Let

$$\bar{\mathbf{X}}_N = \frac{1}{N}\sum_{t=1}^{N} \mathbf{X}_t = (\bar{X}_1, \ldots, \bar{X}_n)', \tag{3.5.1}$$

$$\mathbf{S} = \begin{pmatrix} S_{11} & S_{12} & \cdots & S_{1n} \\ S_{21} & S_{22} & \cdots & S_{2n} \\ & & \cdots & \\ S_{n1} & S_{n2} & \cdots & S_{nn} \end{pmatrix} \tag{3.5.2}$$

denote the sample mean vector and the sample covariance matrix, respectively, where

$$\bar{X}_i = \frac{1}{N}\sum_{t=1}^{N} X_{it}, \tag{3.5.3}$$

$$S_{ij} = \frac{1}{N-1}\sum_{t=1}^{N} (X_{it} - \bar{X}_i)(X_{jt} - \bar{X}_j), \tag{3.5.4}$$

for $i, j = 1, \ldots, n$ (X_{it} is the ith component of $\mathbf{X}_t$). After arranging $\mathbf{X}_1, \ldots, \mathbf{X}_N$ in a matrix form by defining the $n \times N$ data matrix

$$\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_N), \tag{3.5.5}$$

the sample covariance matrix can be expressed as

$$\begin{aligned} \mathbf{S} &= \frac{1}{N-1}\left(\sum_{t=1}^{N} \mathbf{X}_t\mathbf{X}_t' - N\bar{\mathbf{X}}_N\bar{\mathbf{X}}_N'\right) \\ &= \frac{1}{N-1}(\mathbf{X}\mathbf{X}' - N\bar{\mathbf{X}}_N\bar{\mathbf{X}}_N'). \end{aligned} \tag{3.5.6}$$

By the identity

$$N\bar{\mathbf{X}}_N\bar{\mathbf{X}}_N' = \mathbf{X}\left(\frac{1}{N}\mathbf{1}_N\right)\mathbf{X}', \tag{3.5.7}$$

where $\mathbf{I}_N$ is the $N \times N$ identity matrix and $\mathbf{1}_N$ is the $N \times N$ matrix with elements one, we can write

$$\mathbf{S} = \frac{1}{N-1}\mathbf{X}\left(\mathbf{I}_N - \frac{1}{N}\mathbf{1}_N\right)\mathbf{X}'. \tag{3.5.8}$$

Note that $\mathbf{S}$ is symmetric, thus it involves only $n(n+1)/2$ random variables.

It is known that for $N > n$ $(\bar{\mathbf{X}}_N, (N-1)\mathbf{S}/N)$ is the maximum likelihood estimator of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (see Anderson (1984, Sec. 3.2)). Furthermore, almost all of the useful inference procedures in multivariate analysis depend on the data matrix $\mathbf{X}$ only through $(\bar{\mathbf{X}}_N, \mathbf{S})$. Thus the (marginal and joint) distributions of $\bar{\mathbf{X}}_N$ and $\mathbf{S}$ are of great interest.

3.5.1. Independence of $\bar{\mathbf{X}}_N$ and S

Before deriving their distributions we first observe a basic fact. For the univariate normal distribution, it is well known that the sample mean and the sample variance are independent. We show below that a similar statement holds for the multivariate normal distribution.

Theorem 3.5.1. *For* $N > n$, *let* $\mathbf{X}_1, \ldots, \mathbf{X}_N$ *be i.i.d.* $\mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *variables,* $\boldsymbol{\Sigma} > 0$. *Let* $\bar{\mathbf{X}}_N$ *and* $\mathbf{S}$ *be defined as in* (3.5.1) *and* (3.5.2), *respectively. Then* $\bar{\mathbf{X}}_N$ *and* $\mathbf{S}$ *are independent.*

There exist two independent proofs for this result.

FIRST PROOF. The proof depends on the following known result: Let $\mathbf{X}$ be defined as in (3.5.5), and let $\mathbf{C}_1$, $\mathbf{C}_2$ be two given $N \times N$ symmetric real matrices. If $\mathbf{C}_1\mathbf{C}_2 = \mathbf{0}$, then the quadratic forms $\mathbf{X}\mathbf{C}_1\mathbf{X}'$ and $\mathbf{X}\mathbf{C}_2\mathbf{X}'$ are independent. (See, e.g., Anderson and Styan (1982); a less general result was given earlier by Craig (1943).) Thus, by (3.5.7), (3.5.8) and

$$\left(\frac{1}{N}\mathbf{1}_N\right)\left(\mathbf{I}_N - \frac{1}{N}\mathbf{1}_N\right) = \mathbf{0},$$

$\bar{\mathbf{X}}_N\bar{\mathbf{X}}_N'$ and $\mathbf{S}$ are independent. Consequently, $\bar{\mathbf{X}}_N$ and $\mathbf{S}$ are independent. □

SECOND PROOF. The second proof depends on an orthogonal transformation of the elements of $\mathbf{X}$. For every fixed $N \geq 2$ there exists an $N \times N$ orthogonal matrix $\mathbf{C} = (c_{rt})$ satisfying

$$c_{N1} = \cdots = c_{NN} = \frac{1}{\sqrt{N}}. \tag{3.5.9}$$

Since $\mathbf{C}\mathbf{C}' = \mathbf{C}'\mathbf{C} = \mathbf{I}_N$, we must have

$$\sum_{t=1}^{N} c_{rt}^2 = 1 \qquad \text{for all } r, \tag{3.5.10}$$

and

$$\sum_{t=1}^{N} c_{rt}c_{st} = 0 \qquad \text{for all} \quad r \neq s. \tag{3.5.11}$$

This implies

$$\sum_{t=1}^{N} c_{rt} = \sqrt{N}\sum_{t=1}^{N} c_{rt}c_{Nt} = 0 \qquad \text{for all} \quad r < N. \tag{3.5.12}$$

Let us define an $n \times N$ random matrix $\mathbf{Y}$ by

$$\mathbf{Y} = (\mathbf{Y}_1 \dots \mathbf{Y}_N) = \mathbf{X}\mathbf{C}', \qquad \text{or equivalently,} \qquad \mathbf{Y}' = \mathbf{C}\mathbf{X}'.$$

Obviously, the joint distribution of the nN elements of $\mathbf{Y}$ is multivariate normal. Their means, variances, and covariances can be obtained from (3.5.9)–(3.5.12):

(i) For $1 \le i \le n$ and $1 \le r \le N-1$,

$$EY_{ir} = \sum_{t=1}^{N} c_{rt} EX_{it} = \mu_i \sum_{t=1}^{N} c_{rt} = 0.$$

(ii) For $1 \le i, j \le n$ and $1 \le r \le N$,

$$\begin{aligned} \operatorname{Cov}(Y_{ir}, Y_{jr}) &= \operatorname{Cov}\left(\sum_{t=1}^{N} c_{rt} X_{it}, \sum_{t=1}^{N} c_{rt} X_{jt}\right) \\ &= \sum_{t=1}^{N} c_{rt}^2 \operatorname{Cov}(X_{it}, X_{jt}) \\ &= \sigma_{ij}, \end{aligned}$$

which is the (i, j)th element of $\mathbf{\Sigma}$.

(iii) For $1 \le r < s \le N$,

$$\begin{aligned} \operatorname{Cov}(Y_{ir}, Y_{js}) &= \operatorname{Cov}\left(\sum_{t=1}^{N} c_{rt} X_{it}, \sum_{t=1}^{N} c_{st} X_{jt}\right) \\ &= \sum_{t=1}^{N} c_{rt} c_{st} \operatorname{Cov}(X_{it}, X_{jt}) \\ &= 0. \end{aligned}$$

It is easy to verify that the last row of $\mathbf{Y}'$, and hence the transpose of the last column of $\mathbf{Y}$, is

$$\sqrt{N}\bar{\mathbf{X}}'_N = (\sqrt{N}\bar{X}_1, \dots, \sqrt{N}\bar{X}_n).$$

Combining (i)–(iii) with this fact we conclude that: The column vectors $\mathbf{Y}_1, \dots, \mathbf{Y}_{N-1}$ of $\mathbf{Y}$ are i.i.d. $\mathcal{N}_n(\mathbf{0}, \mathbf{\Sigma})$ variables and are independent of its last column (which is $\sqrt{N}\bar{\mathbf{X}}$ and thus has an $\mathcal{N}_n(\sqrt{N}\boldsymbol{\mu}, \mathbf{\Sigma})$ distribution). Now by $\mathbf{C}'\mathbf{C} = \mathbf{I}_N$ we have $\mathbf{X}\mathbf{X}' = \mathbf{Y}\mathbf{Y}'$. But we also have

$$\mathbf{Y}\mathbf{Y}' = \sum_{t=1}^{N-1} \mathbf{Y}_t \mathbf{Y}'_t + N\bar{\mathbf{X}}_N \bar{\mathbf{X}}'_N,$$

and (by (3.5.6))

$$(N-1)\mathbf{S} = \mathbf{X}\mathbf{X}' - N\bar{\mathbf{X}}_N \bar{\mathbf{X}}'_N.$$

Thus $(N-1)\mathbf{S}$ and $\sum_{t=1}^{N-1} \mathbf{Y}_t \mathbf{Y}'_t$ are identically distributed. Consequently, $\mathbf{S}$ and $\bar{\mathbf{X}}_N$ are independent. □

Remark 3.5.1. It should be pointed out that, although the statement of Craig's (1943) result is correct, his proof contains an error that cannot be patched up easily. Correct proofs seem to be first obtained independently by Ogawa (1949) and P.L. Hsu (Fang, 1988). For details, see Anderson ad Styan (1982) and Fang and Zhang (1988, Sec. 2.8).

This useful by-product, obtained in the second proof of Theorem 3.5.1 and stated below, will be applied to derive the Wishart distribution and the Hotelling T^2 distribution.

Proposition 3.5.1. *Let* $\mathbf{X}_1, \ldots, \mathbf{X}_N$ *be i.i.d.* $\mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *variables,* $\boldsymbol{\Sigma} > 0$, *and let*

$$\mathbf{S} = \frac{1}{N-1}\left(\sum_{t=1}^{N} \mathbf{X}_t\mathbf{X}_t' - N\overline{\mathbf{X}}\overline{\mathbf{X}}'\right)$$

be the sample covariance matrix. Then $\mathbf{S}$ *and* $(N-1)^{-1}\sum_{t=1}^{N-1}\mathbf{Y}_t\mathbf{Y}_t'$ *are identically distributed where* $\mathbf{Y}_1, \ldots, \mathbf{Y}_{N-1}$ *are i.i.d.* $\mathcal{N}_n(\mathbf{0}, \boldsymbol{\Sigma})$ *variables.*

3.5.2. Sampling Distributions Concerning $\overline{\mathbf{X}}_N$

In view of the fact that $\overline{\mathbf{X}}_N$ and $\mathbf{S}$ are independent, their joint distribution is uniquely determined from the marginal distributions.

For the univariate normal case, it is well known that $\overline{X}_N$ and $N(\overline{X}_N - \mu)^2/\sigma^2$ are, respectively, $\mathcal{N}(\mu, \sigma^2/N)$ and $\chi^2(1)$ variables. We show that similar results hold for the sample mean vector $\overline{\mathbf{X}}_N$ of a multivariate normal distribution.

Theorem 3.5.2. *Let* $\mathbf{X}_1, \ldots, \mathbf{X}_N$ *be i.i.d.* $\mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *variables,* $\boldsymbol{\Sigma} > 0$, *and* $\overline{\mathbf{X}}_N$ *be the sample mean vector defined in* (3.5.1). *Then* $\overline{\mathbf{X}}_N$ *has an* $\mathcal{N}_n(\boldsymbol{\mu}, (1/N)\boldsymbol{\Sigma})$ *distribution.*

PROOF. Immediate by Corollary 3.3.4. □

Theorem 3.5.3. *Under the conditions stated in Theorem* 3.5.2, $N(\overline{\mathbf{X}}_N - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\overline{\mathbf{X}}_N - \boldsymbol{\mu})$ *has a* $\chi^2(n)$ *distribution.*

PROOF. Let $\mathbf{C}$ be a nonsingular $n \times n$ matrix such that $\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}' = \mathbf{I}_n$ (the existence of $\mathbf{C}$ follows from Proposition 3.2.1). Let $\mathbf{Z} = \sqrt{N}\mathbf{C}(\overline{\mathbf{X}}_N - \boldsymbol{\mu})$. Then, by Theorems 3.5.2 and 3.3.3, $\mathbf{Z}$ has an $\mathcal{N}_n(\mathbf{0}, \mathbf{I}_n)$ distribution; thus $\mathbf{Z}'\mathbf{Z}$ has a $\chi^2(n)$ distribution. But by $\sqrt{N}(\overline{\mathbf{X}}_N - \boldsymbol{\mu}) = \mathbf{C}^{-1}\mathbf{Z}$ we have

$$\begin{aligned} N(\overline{\mathbf{X}}_N - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\overline{\mathbf{X}}_N - \boldsymbol{\mu}) &= \mathbf{Z}'\mathbf{C}^{-1\prime}\boldsymbol{\Sigma}^{-1}\mathbf{C}^{-1}\mathbf{Z} \\ &= \mathbf{Z}'(\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}')^{-1}\mathbf{Z} \\ &= \mathbf{Z}'\mathbf{Z}. \end{aligned}$$

Consequently, $N(\overline{\mathbf{X}}_N - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\overline{\mathbf{X}}_N - \boldsymbol{\mu})$ also has a $\chi^2(n)$ distribution. □

3.5.3. The Wishart and Related Distributions

The Wishart distribution is the joint distribution of the $n(n+1)/2$ variables $(N-1)S_{ij}$, $1 \le i \le j \le n$, which are elements of the random matrix $(N-1)\mathbf{S}$. The density function of this distribution is given in the following theorem.

Theorem 3.5.4. *Let* $\mathbf{X}_1, \ldots, \mathbf{X}_N$ *be i.i.d.* $\mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *variables,* $\boldsymbol{\Sigma} > 0$. *Let* $\mathbf{S}$ *be the sample covariance matrix defined in* (3.5.2). *Then for* $N > n$, *the density function of* $\mathbf{W} = (N-1)\mathbf{S}$ *is*

$$f_{\boldsymbol{\Sigma}, N-1}(\mathbf{w}) = \frac{c_{N-1}|\mathbf{w}|^{(N-n-2)/2}}{|\boldsymbol{\Sigma}|^{(N-1)/2}} e^{-tr\boldsymbol{\Sigma}^{-1}\mathbf{w}/2} \tag{3.5.13}$$

for $\mathbf{w}$ *in the set of all* $n \times n$ *positive definite matrices and* 0 *otherwise, where*

$$c_{N-1} = \left[2^{n(N-1)/2}\pi^{n(n-1)/4} \prod_{j=1}^{n} \Gamma\left(\frac{N-j}{2}\right)\right]^{-1}. \tag{3.5.14}$$

There exist many different methods and approaches for deriving this density function. Wishart's (1928) original proof has a strong geometric flavor. Other proofs were given by Mahalanobis, Bose, and Roy (1937), Hsu (1939), Olkin and Roy (1954), and others. In view of Proposition 3.5.1 we may consider the distribution of the random matrix $\mathbf{W} = \mathbf{Y}\mathbf{Y}'$, where $\mathbf{Y}_1, \ldots, \mathbf{Y}_{N-1}$ are i.i.d. $\mathcal{N}_n(\mathbf{0}, \boldsymbol{\Sigma})$ variables and $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_{N-1})$. The proof adopted here depends on the following lemma give in Anderson (1984, p. 533):

Lemma 3.5.1. *If the density function of the* $n \times (N-1)$ *random matrix* $\mathbf{Y}$ *is* $g(\mathbf{y}\mathbf{y}')$, *then the density function of* $\mathbf{W} = \mathbf{Y}\mathbf{Y}'$ *is*

$$f(\mathbf{w}) = \frac{\pi^{(1/2)n[(N-1)-(n-1)/2]}|\mathbf{w}|^{(N-n-2)/2}g(\mathbf{w})}{\prod_{j=1}^{n}\Gamma((N-j)/2)}.$$

The proof of this lemma involves the joint distribution of the characteristic roots of $\mathbf{W}$, as shown in Anderson (1984, p. 533).

PROOF OF THEOREM 3.5.4. The joint density function of $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_{N-1})'$ is

$$\begin{aligned} g(\mathbf{y}) &= \prod_{t=1}^{N-1} \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\mathbf{y}_t'\boldsymbol{\Sigma}^{-1}\mathbf{y}_t/2} \\ &= [(2\pi)^{n(N-1)/2}|\boldsymbol{\Sigma}|^{(N-1)/2}]^{-1} \exp\left(-\frac{1}{2}\sum_{t=1}^{N-1} \mathbf{y}_t'\boldsymbol{\Sigma}^{-1}\mathbf{y}_t\right) \\ &= [(2\pi)^{n(N-1)/2}|\boldsymbol{\Sigma}|^{(N-1)/2}]^{-1} e^{-tr\boldsymbol{\Sigma}^{-1}\mathbf{y}\mathbf{y}'/2}. \end{aligned}$$

The statement now follows immediately from Lemma 3.5.1. □

A special case of interest for $\boldsymbol{\Sigma} = \mathbf{I}_n$ in Theorem 3.5.4 is:

Corollary 3.5.1. *Let* $\mathbf{X}_1, \ldots, \mathbf{X}_N$ *be i.i.d.* $\mathcal{N}_n(\boldsymbol{\mu}, \mathbf{I}_n)$ *variables and let* $\mathbf{S} = (S_{ij})$ *be the sample covariance matrix. Then for* $N > n$, *the density function of* $(N-1)\mathbf{S}$ *is*

$$f_{\mathbf{I}_n, N-1}(\mathbf{w}) = c_{N-1} |\mathbf{w}|^{(N-n-2)/2} e^{-\mathrm{tr}\mathbf{w}/2} \tag{3.5.15}$$

for $\mathbf{w}$ *in the set of all* $n \times n$ *positive definite matrices, and* 0 *otherwise, where* c_{N-1} *is the constant defined in* (3.5.14).

A problem of great importance concerns the distribution of a transformation of the submatrices of a Wishart matrix. Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_{N-1}$ be i.i.d. $\mathcal{N}_n(\mathbf{0}, \boldsymbol{\Sigma})$ variables such that, for $t = 1, \ldots, N-1$, $\mathbf{Y}_t$ is partitioned as

$$\mathbf{Y}_t = \begin{pmatrix} \mathbf{Y}_{1,t} \\ \mathbf{Y}_{2,t} \end{pmatrix} \sim \mathcal{N}_n\left(\mathbf{0}, \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}\right), \tag{3.5.16}$$

where $\mathbf{Y}_{1,t}$ is $k \times 1$ and $\mathbf{Y}_{2,t}$ is $(n-k) \times 1$. Consider the corresponding partition of $\mathbf{W} = \mathbf{Y}\mathbf{Y}'$ given by

$$\mathbf{W} = \begin{pmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} \\ \mathbf{W}_{21} & \mathbf{W}_{22} \end{pmatrix}, \tag{3.5.17}$$

where $\mathbf{W}_{11}$, $\mathbf{W}_{12} = \mathbf{W}'_{21}$, and $\mathbf{W}_{22}$ are, respectively, $k \times k$, $k \times (n-k)$, and $(n-k) \times (n-k)$. Clearly, we have

$$\mathbf{W}_{ii} = (\mathbf{Y}_{i,1} \cdots \mathbf{Y}_{i,N-1})(\mathbf{Y}_{i,1} \cdots \mathbf{Y}_{i,N-1})', \qquad i = 1, 2$$

and

$$\mathbf{W}_{12} = (\mathbf{Y}_{1,1} \cdots \mathbf{Y}_{1,N-1})(\mathbf{Y}_{2,1} \cdots \mathbf{Y}_{2,N-1})'.$$

The following lemma concerns the distribution of the matrix $\mathbf{W}_{11}$ $\mathbf{W}_{12}\mathbf{W}_{22}^{-1}\mathbf{W}_{21}$.

Lemma 3.5.2. *Let* $\mathbf{Y}_1, \ldots, \mathbf{Y}_{N-1}$ *be defined as in* (3.5.16) *and let* $\mathbf{W}$ *be partitioned as in* (3.5.17). *If* $\boldsymbol{\Sigma} > 0$, *then* $\mathbf{V} = \mathbf{W}_{11} - \mathbf{W}_{12}\mathbf{W}_{22}^{-1}\mathbf{W}_{21}$ *and* $\sum_{t=1}^{(N-1)-(n-k)} \mathbf{U}_t\mathbf{U}'_t$ *are identically distributed where* $\mathbf{U}_1, \ldots, \mathbf{U}_{(N-1)-(n-k)}$ *are i.i.d.* $\mathcal{N}_k(\mathbf{0}, \boldsymbol{\Sigma}_{11\cdot 2})$ *variables and* $\boldsymbol{\Sigma}_{11\cdot 2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$.

PROOF. For given $\mathbf{Y}_{2t} = \mathbf{y}_{2t}$ the conditional distribution of $\mathbf{Y}_{1t}$ is $\mathcal{N}_k(\mathbf{B}\mathbf{x}_{2t}, \boldsymbol{\Sigma}_{11\cdot 2})$ for $t = 1, \ldots, N-1$, where $\mathbf{B} = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}$ is the regression matrix (Theorem 3.3.4). We show that for given $\mathbf{y}_{21}, \ldots, \mathbf{y}_{2,N-1}$ the conditional distribution of

$$\sum_{t=1}^{N-1} \mathbf{Y}_{1t}\mathbf{Y}'_{1t} - \mathbf{T}\mathbf{w}_{22}\mathbf{T}' = \sum_{t=1}^{N-1} \mathbf{Y}_{1t}\mathbf{Y}'_{1t} - \left(\sum_{t=1}^{N-1} \mathbf{Y}_{1t}\mathbf{y}'_{2t}\right)\mathbf{w}_{22}^{-1}\left(\sum_{t=1}^{N-1} \mathbf{Y}_{1t}\mathbf{y}'_{2t}\right)'$$

and the distribution of $\sum_{t=1}^{(N-1)-(n-k)} \mathbf{U}_t\mathbf{U}'_t$ are identical, where

$$\mathbf{w}_{22} = \sum_{t=1}^{N-1} \mathbf{y}_{2t}\mathbf{y}'_{2t}$$

and $\mathbf{T}$ is a $k \times (n-k)$ random matrix given by

$$\mathbf{T} = \left(\sum_{t=1}^{N-1} \mathbf{Y}_{1t}\mathbf{y}'_{2t}\right)\mathbf{w}_{22}^{-1}.$$

The lemma then follows from the fact that the underlying conditional distribution does not depend on the $\mathbf{y}_{2t}$'s. The proof given below, which follows the steps of Anderson's (1984, pp. 130–131) proof, depends on an orthogonal transformation of the matrix

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_{11}\mathbf{Y}_{12}\cdots\mathbf{Y}_{1,N-1} \\ \mathbf{y}_{21}\mathbf{y}_{22}\cdots\mathbf{y}_{2,N-1} \end{pmatrix} \equiv \begin{pmatrix} \mathbf{Y}^{(1)} \\ \mathbf{y}^{(2)} \end{pmatrix}$$

when the $\mathbf{y}_{2t}$'s are given. The basic idea is similar to that in the second proof of Theorem 3.5.1, except that it is more general.

Let $\mathbf{C}$ be a nonsingular matrix such that $\mathbf{C}\mathbf{w}_{22}\mathbf{C}' = \mathbf{I}_{n-k}$ and, for given $\mathbf{y}^{(2)} = (\mathbf{y}_{21}, \ldots, \mathbf{y}_{2,N-1})$, let $\mathbf{G}_2 = \mathbf{C}\mathbf{y}^{(2)}$ or, equivalently, $\mathbf{y}^{(2)} = \mathbf{C}^{-1}\mathbf{G}_2$. Then

$$\begin{aligned} \mathbf{G}_2\mathbf{G}_2' &= \mathbf{C}\mathbf{y}^{(2)}(\mathbf{y}^{(2)})'\mathbf{C}' \\ &= \mathbf{C}\mathbf{w}_{22}\mathbf{C}' = \mathbf{I}_{n-k}. \end{aligned}$$

By Anderson (1984, p. 598), there exists an $((N-1)-(n-k)) \times (N-1)$ matrix $\mathbf{G}_1$ such that

$$\mathbf{G} = \begin{pmatrix} \mathbf{G}_1 \\ \mathbf{G}_2 \end{pmatrix}$$

is an orthogonal matrix. Now consider the orthogonal transformation of the matrix $\mathbf{Y}$ given by

$$\mathbf{U} = (\mathbf{U}_1, \ldots, \mathbf{U}_{N-1}) = \mathbf{Y}^{(1)}\mathbf{G}', \tag{3.5.18}$$

or, equivalently, $\mathbf{Y}^{(1)} = \mathbf{G}\mathbf{U}$. Clearly, we have

$$\begin{aligned} \sum_{t=1}^{N-1} \mathbf{U}_t\mathbf{U}_t' &= \mathbf{U}\mathbf{U}' \\ &= \mathbf{Y}^{(1)}\mathbf{G}'\mathbf{G}(\mathbf{Y}^{(1)})' = \mathbf{Y}^{(1)}(\mathbf{Y}^{(1)})'. \end{aligned}$$

On the other hand, by $\mathbf{T} = \mathbf{Y}^{(1)}(\mathbf{y}^{(2)})'\mathbf{w}_{22}^{-1}$ and $\mathbf{T}\mathbf{w}_{22}\mathbf{T}' = (\mathbf{T}\mathbf{w}_{22}\mathbf{T}')'$, we have

$$\begin{aligned} \mathbf{T}\mathbf{w}_{22}\mathbf{T}' &= (\mathbf{G}\mathbf{U}(\mathbf{C}^{-1}\mathbf{G}_2)'\mathbf{w}_{22}^{-1}(\mathbf{C}^{-1}\mathbf{G}_2)(\mathbf{G}\mathbf{U})')' \\ &= \mathbf{U}(\mathbf{G}\mathbf{G}_2')(\mathbf{C}\mathbf{w}_{22}\mathbf{C}')^{-1}(\mathbf{G}_2\mathbf{G}')\mathbf{U}' \\ &= \mathbf{U}\begin{pmatrix} \mathbf{0} \\ \mathbf{I}_{n-k} \end{pmatrix}(\mathbf{0}\ \ \mathbf{I}_{n-k})\mathbf{U}' \\ &= \sum_{t=N-(n-k)}^{N-1} \mathbf{U}_t\mathbf{U}_t'. \end{aligned}$$

This implies that $\sum_{t=1}^{N-1} \mathbf{Y}_{1t}\mathbf{Y}_{1t}' - \mathbf{T}\mathbf{w}_{22}\mathbf{T}'$ and $\sum_{t=1}^{(N-1)-(n-k)} \mathbf{U}_t\mathbf{U}_t'$ are identically distributed. By (3.5.18) and the fact that $\mathbf{G}$ is an orthogonal matrix, it is easy to verify that $\mathbf{U}_1, \ldots, \mathbf{U}_{(N-1)-(n-k)}$ are independent $\mathcal{N}_k(\mathbf{0}, \boldsymbol{\Sigma}_{11\cdot 2})$ variables. Hence Lemma 3.5.2 follows. □

Combining Proposition 3.5.1 and Lemma 3.5.2 we immediately have

Theorem 3.5.5. *Let $\mathbf{X}_1, \ldots, \mathbf{X}_N$ be i.i.d. $\mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ variables and let $\mathbf{S}$ be the sample covariance matrix defined in* (3.5.2). *For fixed $1 \le k < n$, let $\mathbf{S}$ be partitioned as*

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{pmatrix},$$

where $\mathbf{S}_{11}, \mathbf{S}_{12} = \mathbf{S}'_{21}$, and $\mathbf{S}_{22}$ are, respectively, $k \times k, k \times (n-k)$, and $(n-k) \times (n-k)$. If $N > n$, then

$$\mathbf{V} = (N-1)(\mathbf{S}_{11} - \mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}) \tag{3.5.19}$$

and $\sum_{t=1}^{(N-1)-(n-k)} \mathbf{U}_t\mathbf{U}'_t$ are identically distributed where $\mathbf{U}_1, \ldots, \mathbf{U}_{(N-1)-(n-k)}$ are i.i.d. $\mathcal{N}_k(\mathbf{0}, \boldsymbol{\Sigma}_{11\cdot 2})$ variables ad $\boldsymbol{\Sigma}_{11\cdot 2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. Consequently, the density function of $\mathbf{V}$ can be obtained by substituting (i) $\boldsymbol{\Sigma}_{11\cdot 2}$ *for $\boldsymbol{\Sigma}$,* (ii) *$N-(n-k)$ for N, and* (iii) *k for n in the density function given in* (3.5.13).

Of special interest is the case $\boldsymbol{\Sigma}_{12} = \mathbf{0}$. This result is stated below and will be used to derive the distribution of the sample multiple correlation coefficient.

Corollary 3.5.2. *Let $\mathbf{X}_1, \ldots, \mathbf{X}_N$ be i.i.d. $\mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ variables, $\boldsymbol{\Sigma} > 0$, and $\boldsymbol{\Sigma}_{12} = \mathbf{0}$. Then*:

(a) $\mathbf{V}$ *in* (3.5.19) *is distributed as $\sum_{t=1}^{(N-1)-(n-k)} \mathbf{U}_t\mathbf{U}'_t$, and $(N-1)\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}$ is distributed as $\sum_{t=N-(n-k)}^{N-1} \mathbf{U}_t\mathbf{U}'_t$, where $\mathbf{U}_1, \ldots, \mathbf{U}_{N-1}$ are i.i.d. $\mathcal{N}_n(\mathbf{0}, \boldsymbol{\Sigma}_{11})$ variables; and*
(b) $\mathbf{V}$ *and $\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}$ are independent.*

Note that in (3.5.19) the random matrix $\mathbf{V}$ is properly defined only if $\mathbf{S}_{22}$ is invertible with probability one. This is possible when $\mathbf{S}$ itself is invertible with probability one. A more important question is whether $\mathbf{S}$ is positive definite (in symbols $\mathbf{S} > 0$) which, of course, implies that $\mathbf{S}$ is invertible. The answer to this question is given below.

Theorem 3.5.6. *Let $\mathbf{X}_1, \ldots, \mathbf{X}_N$ be i.i.d. $\mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ variables, $\boldsymbol{\Sigma} > 0$, and let $\mathbf{S}$ be the sample covariance matrix defined in* (3.5.2). *Then $P[\mathbf{S} > 0] = 1$ holds if and only if $N > n$.*

Proof. The proof given here is due to Dykstra (1970). Note that the assumption of normality is not needed, so that the statement also holds for other multivariate distributions.

By Proposition 3.5.1, $(N-1)\mathbf{S}$ and $\sum_{t=1}^{N-1} \mathbf{U}_t\mathbf{U}'_t = \mathbf{U}\mathbf{U}'$ are identically distributed where $\mathbf{U}_1, \ldots, \mathbf{U}_{N-1}$ are i.i.d. $\mathcal{N}_n(\mathbf{0}, \boldsymbol{\Sigma})$ variables and $\mathbf{U} =$

$(\mathbf{U}_1, \ldots, \mathbf{U}_{N-1})$. Thus, it is equivalent to showing that $P[\mathbf{UU}' > 0] = 1$ holds if and only if $N > n$. Scheffé (1959, p. 399) states that:

(i) $\mathbf{U}$ and $\mathbf{UU}'$ have the same rank; and
(ii) $\mathbf{UU}' > 0$ ($\mathbf{UU}'$ is positive semidefinite) if and only if the rank of $\mathbf{U}$ is n (is $< n$).

If $N < n$, then clearly the rank of $\mathbf{U}$ is $< n$. On the other hand, since

$$\operatorname{rank}(\mathbf{U}_1, \ldots, \mathbf{U}_m) \leq \operatorname{rank}(\mathbf{U}_1, \ldots, \mathbf{U}_{m+1})$$

for all $m \geq n$, it suffices to show that

$$P[\operatorname{rank}(\mathbf{U}_1, \ldots, \mathbf{U}_n) < n] = 0.$$

For every fixed $i = 1, \ldots, n$, and for given

$$\mathbf{U}^{(i)} \equiv (U_1, \ldots, U_{i-1}, U_{i+1}, \ldots, U_n) = (u_1, \ldots, u_{i-1}, u_{i+1}, \ldots, u_n) \equiv \mathbf{u}^{(i)},$$

let $B_i(\mathbf{u}^{(i)})$ be the subspace spanned by $\mathbf{u}^{(i)}$. Then by $\boldsymbol{\Sigma} > 0$,

$$P[\mathbf{U}_i \in B_i(\mathbf{u}^{(i)})] = 0$$

holds for all $\mathbf{u}^{(i)}$ except perhaps on a set of probability zero. Consequently,

$$\begin{aligned} P[\operatorname{rank}(\mathbf{U}_1, \ldots, \mathbf{U}_n) < n] &= P[\mathbf{U}_1, \ldots, \mathbf{U}_n \text{ are linearly dependent}] \\ &\leq \sum_{i=1}^{n} EP[\mathbf{U}_i \in B_i(\mathbf{u}^{(i)}) | \mathbf{U}^{(i)} = \mathbf{u}^{(i)}] \\ &= 0. \end{aligned}$$

□

3.5.4. Hotelling's T^2 Distribution

When applying Theorem 3.5.2 or 3.5.3 to make statistical inference on $\boldsymbol{\mu}$ based on $\bar{\mathbf{X}}_N$, the covariance matrix $\boldsymbol{\Sigma}$ must be known. If $\boldsymbol{\Sigma}$ is unknown, then a new statistic (a generalization of Student's t statistic) is needed. This was proposed and studied by Hotelling (1931): The statistic

$$T^2 = N(\bar{\mathbf{X}}_N - \boldsymbol{\mu})'\mathbf{S}^{-1}(\bar{\mathbf{X}}_N - \boldsymbol{\mu}) \tag{3.5.20}$$

is called Hotelling's T^2 statistic. Note that, by Theorem 3.5.6, $\mathbf{S}$ is positive definite with probability one, so that $\mathbf{S}^{-1}$ is positive definite and T^2 is a properly defined quadratic form with probability one.

Theorem 3.5.7. *Let $\mathbf{X}_1, \ldots, \mathbf{X}_N$ be i.i.d. $\mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ variables, $\boldsymbol{\Sigma} > 0$. Let $\bar{\mathbf{X}}$ and $\mathbf{S}$ be the sample mean vector and sample covariance matrix, respectively, and let T^2 be given in* (3.5.20). *Then for $N > n$, $((N - n)/((N - 1)n))T^2$ has an $F(n, N - n)$ distribution.*

Remark 3.5.2. We first note that the distribution of T^2 is invariant under the transformation $\mathbf{Z}_t = \mathbf{HX}_t + \mathbf{b}$ $(t = 1, \ldots, N)$ where $\mathbf{H}$ is any nonsingular $n \times n$

matrix and $\mathbf{b}$ is any real vector. The fact that it does not depend on $\mathbf{b}$ is easy to verify. To show that it also does not depend on $\mathbf{H}$, suppose that $\boldsymbol{\mu} = \mathbf{b} = \mathbf{0}$ and $\mathbf{Z}_t = \mathbf{H}\mathbf{X}_t$. Then $\bar{\mathbf{Z}}_N = \mathbf{H}\bar{\mathbf{X}}_N$ and $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_N) = \mathbf{H}(\mathbf{X}_1, \ldots, \mathbf{X}_N) = \mathbf{H}\mathbf{X}$. Consequently,

$$(N-1)\mathbf{S}_{\mathbf{Z}} = \mathbf{Z}\mathbf{Z}' - N\bar{\mathbf{Z}}_N\bar{\mathbf{Z}}_N'$$
$$= \mathbf{H}(\mathbf{X}\mathbf{X}' - N\bar{\mathbf{X}}_N\bar{\mathbf{X}}_N')\mathbf{H}',$$

and this implies

$$\bar{\mathbf{Z}}_N'\mathbf{S}_{\mathbf{Z}}^{-1}\bar{\mathbf{Z}}_N = (N-1)\bar{\mathbf{X}}_N'\mathbf{H}'\mathbf{H}'^{-1}(\mathbf{X}\mathbf{X}' - N\bar{\mathbf{X}}_N\bar{\mathbf{X}}_N')^{-1}\mathbf{H}^{-1}\mathbf{H}\bar{\mathbf{X}}_N$$
$$= \bar{\mathbf{X}}_N'\mathbf{S}_{\mathbf{X}}^{-1}\bar{\mathbf{X}}_N.$$

PROOF OF THEOREM 3.5.7. There are different methods for deriving the distribution of T^2. The proof given here, adopted from Anderson (1984, p. 161–162), depends on an orthogonal transformation for given $\bar{\mathbf{X}}_N = \bar{\mathbf{x}}_N$. After it is shown that the conditional distribution of T^2 does not depend on $\bar{\mathbf{x}}_N$, the statement follows by unconditioning.

Without loss of generality we may assume that $\mathbf{X}_1, \ldots, \mathbf{X}_N$ are i.i.d. $\mathcal{N}_n(\mathbf{0}, \mathbf{I}_n)$ variables. For given $\bar{\mathbf{X}}_N = \bar{\mathbf{x}}_N$ let $\mathbf{C}$ be an $n \times n$ orthogonal matrix such that the first row of $\mathbf{C}$ is $\bar{\mathbf{x}}_N/\sqrt{\bar{\mathbf{x}}_N'\bar{\mathbf{x}}_N}$, and let the c_{ij}'s $(i = 2, \ldots, n; j = 1, \ldots, n)$ depend on $\bar{\mathbf{x}}_N$ through the c_{1j}'s. Now consider the linear transformation $\mathbf{U} = (U_1, \ldots, U_n)' = \mathbf{C}\bar{\mathbf{x}}_N$, and define $\mathbf{B} = \mathbf{C}\mathbf{S}\mathbf{C}'$. Since

$$\sum_{j=1}^{n} c_{ij}c_{1j} = \frac{(c_{i1} \ldots c_{in})\bar{\mathbf{x}}_N}{\sqrt{\bar{\mathbf{x}}_N'\bar{\mathbf{x}}_N}} = 0,$$

clearly we have

$$U_1 = \sqrt{\bar{\mathbf{x}}_N'\bar{\mathbf{x}}_N} \qquad \text{and} \qquad U_i = 0 \qquad \text{for} \quad i = 2, \ldots, n.$$

Thus we have, for given $\bar{\mathbf{X}}_N = \bar{\mathbf{x}}_N$,

$$\frac{T^2}{(N-1)} = \bar{\mathbf{x}}_N'\mathbf{S}^{-1}\bar{\mathbf{x}}_N = \mathbf{U}'\mathbf{B}^{-1}\mathbf{U} = U_1^2 b^{11},$$

where $\mathbf{B}^{-1} = (b^{ij})$. But $1/b^{11} = b_{11} - \mathbf{b}_1'\mathbf{B}_{22}^{-1}\mathbf{b}_1$. Since $\bar{\mathbf{X}}_N$ and $\mathbf{S}$ are independent (Theorem 3.5.1) and $\mathbf{C}$ depends only on $\bar{\mathbf{x}}_N$, the conditional distribution of $\mathbf{B}$ is identical to that of $\sum_{t=1}^{N-1} \mathbf{Z}_t\mathbf{Z}_t'$ where the $\mathbf{Z}_t$'s are i.i.d. $\mathcal{N}_n(\mathbf{0}, \mathbf{I}_n)$ variables. Thus by Theorem 3.5.5 (with $k = 1$), $(N-1)/b^{11}$ has a chi-square distribution with $(N - n)$ degrees of freedom for every fixed $U_1^2 = \bar{\mathbf{x}}_N'\bar{\mathbf{x}}_N$. But $NU_1^2 = N\bar{\mathbf{X}}_N'\bar{\mathbf{X}}_N$ itself has a chi-square distribution with n degrees of freedom (Theorem 3.5.3) and is independent of b^{11}. Consequently,

$$\frac{(N-n)}{(N-1)n}T^2 = \frac{NU_1^2/n}{((N-1)/b^{11})/(N-n)}$$

has an $F(n, N-n)$ distribution. □

Other results for the distribution of T^2 can be obtained by studying the random variable

$$T_0^2 = N(\overline{\mathbf{X}}_N - \boldsymbol{\mu}_0)'\mathbf{S}^{-1}(\overline{\mathbf{X}}_N - \boldsymbol{\mu}_0),$$

where μ_0 is not necessarily the mean vector of $\overline{\mathbf{X}}_N$. The distribution of T_0^2 involves a noncentral F distribution, and was given by Bose and Roy (1938), Hsu (1938), Bowker (1960), and others. The distribution of T^2 can be obtained from the distribution of T_0^2 by letting $\boldsymbol{\mu}_0 = \boldsymbol{\mu}$ (the mean vector of $\overline{\mathbf{X}}_N$) as a special case.

3.5.5. Sample Correlation Coefficients

To investigate the distributions of the sample multiple correlation coefficient and the sample partial correlation coefficient, we once again consider the partition of the components of an n-dimensional normal variable defined in (3.3.1). For fixed $N > n$, let $\{\mathbf{X}_t\}_{t=1}^N$ be a sequence of i.i.d. $\mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ variables, $\boldsymbol{\Sigma} > 0$, and let $\mathbf{S}$ be the sample covariance matrix defined in (3.5.2). We consider a corresponding partition of this sample covariance matrix given by

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{pmatrix} = \frac{1}{N-1}\begin{pmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} \\ \mathbf{W}_{21} & \mathbf{W}_{22} \end{pmatrix} = \frac{1}{N-1}\mathbf{W}, \qquad (3.5.21)$$

where $\mathbf{S}_{11}$ is $k \times k$, $\mathbf{S}_{12} = \mathbf{S}_{21}'$ is $k \times (n-k)$, and $\mathbf{S}_{22}$ is $(n-k) \times (n-k)$.

The population multiple correlation coefficient between X_i and $(X_{k+1}, \ldots, X_n)'$, defined in Definition 3.4.2, can be estimated by substituting S_{ij}'s for σ_{ij}'s in (3.4.5). Without loss of generality it may be assumed that $i = k = 1$. (Because otherwise we need to consider only the marginal distribution of $(X_i, X_{k+1}, \ldots, X_n)'$ instead of the joint distribution of $\mathbf{X}$.) Then the sample multiple correlation coefficient is given by

$$\hat{R}_{1\cdot 2\ldots n} = \left(\frac{\mathbf{S}_1\mathbf{S}_{22}^{-1}\mathbf{S}_1'}{S_{11}}\right)^{1/2}, \qquad (3.5.22)$$

where $\mathbf{S}_1 = (S_{12}, \ldots, S_{1n})$. It is known that $\hat{R}_{1\cdot 2\ldots n}$ is the maximum likelihood estimator of the population correlation coefficient $R_{1\cdot 2\ldots n}$ and has certain desirable properties.

In the following theorem we state a result for the distribution of $\hat{R}_{1\cdot 2\ldots n}$. When $R_{1\cdot 2\ldots n} > 0$, its density function has several different expressions, and the one given here is due to Fisher (1928). Since the proof is quite involved algebraically, it is outlined without details.

Theorem 3.5.8. *Let $\mathbf{X}_1, \ldots, \mathbf{X}_N$ be i.i.d. $\mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ variables, $\boldsymbol{\Sigma} > 0$. Let $\hat{R} = \hat{R}_{1\cdot 2\ldots n}$ be the sample multiple correlation coefficient defined in* (3.5.22).

(a) *If the population multiple correlation coefficient $R_{1\cdot 2\ldots n}$ is zero, then $(N-n)\hat{R}^2/((n-1)(1-\hat{R}^2))$ has an $F(n-1, N-n)$ distribution.*

(b) *If $R_{1\cdot 2\ldots n}$ is not zero, the density function of $\hat{R}^2$ is*

$$g_{\hat{R}^2}(r^2) = \frac{(1-r^2)^{(N-n-2)/2}(1-R^2)^{(N-1)/2}}{\Gamma(\frac{1}{2}(N-n))\Gamma(\frac{1}{2}(N-1))} \sum_{j=1}^{\infty} \frac{(R^2)^j (r^2)^{(n-1)/2+j-1}\Gamma^2(\frac{1}{2}(N-1)+j)}{j!\Gamma(\frac{1}{2}(n-1)+j)}, \tag{3.5.23}$$

where $R^2 = R^2_{1\cdot 2\ldots n}$. Thus the density function of $\hat{R}$ is $2rg_{\hat{R}^2}(r^2)$ for $r \geq 0$.

PROOF. (a) First note the identity

$$\frac{\hat{R}^2}{1-\hat{R}^2} = \frac{\mathbf{S}_1\mathbf{S}_{22}^{-1}\mathbf{S}_1'}{S_{11} - \mathbf{S}_1\mathbf{S}_{22}^{-1}\mathbf{S}_1'}.$$

Since $R_{1\cdot 2\ldots n} = 0$ if and only if $\boldsymbol{\Sigma}_{12} = \mathbf{0}$, by Corollary 3.5.2 we immediately have:

(i) $(N-1)\mathbf{S}_1\mathbf{S}_{22}^{-1}\mathbf{S}_1'/\sigma_{11}$ has a $\chi^2(n-1)$ distribution;
(ii) $(N-1)(S_{11} - \mathbf{S}_1\mathbf{S}_{22}^{-1}\mathbf{S}_1')/\sigma_{11}$ has a $\chi^2(N-n)$ distribution; and
(iii) the two random variables in (i) and (ii) are independent.

Thus $(N-n)\hat{R}^2/((n-1)(1-\hat{R}^2))$ has an $F(n-1, N-n)$ distribution.

(b) Without loss of generality assume that $\boldsymbol{\mu} = \mathbf{0}$.

(i) First, let us consider the conditional distribution of the random variable $(N-n)\hat{R}^2/((n-1)(1-\hat{R}^2))$ for given

$$\mathbf{X}_{2t} \equiv (X_{2t}, \ldots, X_{nt})' = (x_{2t}, \ldots, x_{nt})' \equiv \mathbf{x}_{2t}, \qquad t = 1, \ldots, N.$$

Since the conditional distribution of X_{1t} is

$$\mathcal{N}(\boldsymbol{\sigma}_1\boldsymbol{\Sigma}_{22}^{-1}\mathbf{x}_{2t}, \sigma_{11} - \boldsymbol{\sigma}_1\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_1' \equiv \sigma_{11\cdot 2}),$$

by applying the transformation in the proof of Lemma 3.5.2 we can show that $(N-1)\mathbf{S}_1\mathbf{S}_{22}^{-1}\mathbf{S}_1')/\sigma_{11\cdot 2}$ has a noncentral chi-square distribution with $n-1$ degrees of freedom and noncentrality parameter $(N-1)\boldsymbol{\beta}\mathbf{s}_{22}\boldsymbol{\beta}'/\sigma_{11\cdot 2}$, where $\boldsymbol{\beta} = \boldsymbol{\sigma}_1\boldsymbol{\Sigma}_{22}^{-1}$. Thus, for given $\mathbf{X}_{2t} = \mathbf{x}_{2t}$ $(t = 1, \ldots, N)$, the conditional distribution of $(N-n)\hat{R}^2/((n-1)(1-\hat{R}^2))$ is a noncentral F distribution with degrees of freedom $(n-1, N-n)$ and noncentrality parameter $(N-1)\boldsymbol{\beta}\mathbf{s}_{22}\boldsymbol{\beta}'/\sigma_{11\cdot 2}$.

(ii) By the result in (i) we can write out the joint density function of $(\hat{R}^2, \mathbf{X}_{21}, \ldots, \mathbf{X}_{2N})'$ and then integrate out $\mathbf{x}_{21}, \ldots, \mathbf{x}_{2N}$ over the $(n-1)N$-dimensional space to obtain $g_{\hat{R}^2}(r^2)$. □

The population partial correlation coefficient $\rho_{ij\cdot k+1,\ldots,n}$ defined in Definition 3.4.3 is the correlation between X_i and X_j $(1 \leq i < j \leq k)$ in the conditional distribution of $(X_1, \ldots, X_k)'$, given $(X_{k+1}, \ldots, X_n)' = (x_{k+1}, \ldots, x_n)'$. Now let the sample covariance matrix $\mathbf{S}$ be partitioned as in (3.5.21). Then for fixed $(X_{2t}, \ldots, X_{nt})' = (x_{2t}, \ldots, x_{nt})'$ $(t = 1, \ldots, N)$ the sample partial correlation coefficient is

$$r_{ij\cdot k+1,\ldots,n} = \frac{s_{ij\cdot k+1,\ldots,n}}{(s_{ii\cdot k+1\ldots n}\cdot s_{jj\cdot k+1\ldots n})^{1/2}},$$

where $s_{ij\cdot k+1,\ldots,n}$ is the (i, j)th element of the matrix $\mathbf{S}_{11} - \mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}$. The following theorem concerns the distribution of $r_{ij\cdot k+1,\ldots,n}$.

Theorem 3.5.9. *Let* $\mathbf{X}_1, \ldots, \mathbf{X}_N$ *be i.i.d.* $\mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *variables,* $\boldsymbol{\Sigma} > 0$. *Let*

$$\boldsymbol{\Sigma}_{11\cdot 2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} = (\sigma_{ij\cdot k+1,\ldots,n}).$$

Let r_{12} *be the sample correlation coefficient between* Y_1 *and* Y_2 *based on a random sample of size* $N - (n - k)$ *from a bivariate normal distribution with means* 0, *variances* 1, *and correlation coefficient*

$$\rho = \frac{\sigma_{ij\cdot k+1,\ldots,n}}{(\sigma_{ii\cdot k+1,\ldots,n}\sigma_{jj\cdot k+1,\ldots,n})^{1/2}}.$$

Then $r_{ij\cdot k+1,\ldots,n}$ *and* r_{12} *are identically distributed.*

This result, due to Fisher (1924), can be obtained by applying Theorem 3.5.5. The details are left to the reader.

Problems

3.1. Let $\mathbf{X}$ be any $r \times m$ real matrix for $r \leq m$. Show that $\mathbf{X}\mathbf{X}'$ is either positive definite (p.d.) or positive semidefinite (p.s.d.). Furthermore, show that if the rank of $\mathbf{X}$ is r, then $\mathbf{X}\mathbf{X}'$ is p.d.

3.2. Show that $\boldsymbol{\Sigma}$ is a p.d. matrix if and only if $\boldsymbol{\Sigma}^{-1}$ is a p.d. matrix.

3.3. Show that if $\boldsymbol{\Sigma}$ is a p.d. matrix, then its determinant is positive.

3.4. Show that if $\boldsymbol{\Sigma}$ is a p.d. matrix, then $c\boldsymbol{\Sigma}$ is a p.d. matrix for all $c > 0$.

3.5. Show that if $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$ are two $n \times n$ p.d. matrices, then $\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2$ is a p.d. matrix.

3.6. Show that if $\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_m$ are $n \times n$ p.d. matrices, then $\sum_{j=1}^m c_j\boldsymbol{\Sigma}_j$ is a p.d. matrix for all $c_j > 0$ $(j = 1, \ldots, m)$.

3.7. Let $\boldsymbol{\Sigma} = (\sigma_{ij})$ be a 3×3 symmetric matrix such that

$$\sigma_{11} = \sigma_{22} = \sigma_{33} = 1, \qquad \sigma_{12} = 0.$$

Show that, at least for $(\sigma_{13} + \sigma_{23}) > \frac{3}{2}$, $\boldsymbol{\Sigma}$ is *not* a p.d. matrix.

In Problems 3.8–3.10, $\boldsymbol{\Sigma}$ denotes an $n \times n$ symmetric matrix; $\boldsymbol{\Sigma}_{11}$ and $\boldsymbol{\Sigma}_{22}$ are the corresponding submatrices defined in (3.3.1).

3.8. Show that if $\boldsymbol{\Sigma}$ is p.d., then both $\boldsymbol{\Sigma}_{11}$ and $\boldsymbol{\Sigma}_{22}$ are p.d.

3.9. Show that if $\boldsymbol{\Sigma}_{11}$ is not p.d., then $\boldsymbol{\Sigma}$ is not p.d.

3.10. Verify the statement in (3.3.6) cooncerning the inverse of $\boldsymbol{\Sigma}$.

3.11. Show that if $\mathbf{\Sigma}$ is p.s.d., then there exists a sequence of p.d. matrices $\{\mathbf{\Sigma}_t\}_{t=1}^{\infty}$ such that $\lim_{t\to\infty} \mathbf{\Sigma}_t = \mathbf{\Sigma}$.

3.12. Verify the equivalence statements in the proof of Theorem 3.3.2.

3.13. Verify the identity in (3.3.18).

3.14. Let $\mathbf{Z} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{I}_n)$. Let $\mathbf{Y}_1 = \mathbf{C}_1\mathbf{Z}$ and $\mathbf{Y}_2 = \mathbf{C}_2\mathbf{Z}$ where $\mathbf{C}_i$ is $k_i \times n$, $k_i \le n$ $(i = 1, 2)$. Find a necessary and sufficient condition for the independence of $\mathbf{Y}_1, \mathbf{Y}_2$.

3.15. Let $\mathbf{X}$ be partitioned as in (3.3.1), for fixed $k + 1 \le m \le n$ let $\mathbf{X}_2^{(m)} = (X_{k+1}, \ldots, X_m)'$. Let $\lambda^*(\mathbf{x}_2^{(m)})$ be the best predictor of X_i $(1 \le i \le k)$, given $\mathbf{X}_2^{(m)} = \mathbf{x}_2^{(m)}$. Show that $E((X_i - \lambda^*(\mathbf{x}_2^{(m)}))^2 | \mathbf{X}_2^{(m)} = \mathbf{x}_2^{(m)})$ is a nonincreasing function of m.

3.16. Let $\mathbf{X}$ and $\mathbf{X}_2^{(m)}$ be defined as in Problem 3.15, and let $R_{i\cdot k+1,\ldots,n}$ be the multiple correlation coefficient between X_i and $\mathbf{X}_2^{(m)}$. Show that $R_{i\cdot k+1,\ldots,m}$ is a nondecreasing function of m.

3.17. Show that the multiple correlation coefficient $R_{i\cdot k+1,\ldots,n}$ is nonnegative and is bounded above by one.

3.18. Show that the canonical correlation coefficients are nonnegative and are bounded above by one.

3.19. Verify the statement in (3.4.10).

3.20. Verify (3.4.27) and (3.4.28).

3.21. Verify (3.4.29).

In Problems 3.22–3.28, $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)' = ((X_1, X_2), (X_3, X_4))'$ is assumed to have a multivariate normal distribution with means μ, variances σ^2, and correlation coefficients

$$\rho_{12} = \rho_{34} = \rho_2, \qquad \rho_{ij} = \rho_1 \qquad \text{for} \quad i \le 2 \text{ and } j \ge 3,$$

where $0 \le \rho_1 \le \rho_2$.

3.22. Find the marginal distributions of $\mathbf{X}_1$ and $\mathbf{X}_2$.

3.23. Find the conditional distribution of $\mathbf{X}_1$ given $\mathbf{X}_2 = \mathbf{x}_2$.

3.24. Find the best predictor $\lambda^*(\mathbf{x}_2)$ of X_1 and find $E((X_1 - \lambda^*(\mathbf{x}_2))^2 | \mathbf{X}_2 = \mathbf{x}_2)$.

3.25. Find the multiple correlation coefficient $R_{1\cdot 34}$.

3.26. Find the partial correlation coefficient $\rho_{12\cdot 34}$.

3.27. Find the canonical correlation coefficients between $\mathbf{X}_1$ and $\mathbf{X}_2$.

3.28. Find the principal components of $\mathbf{X}$ and their variances.

3.29. Verify the identities in (3.5.6), (3.5.7), and (3.5.8).

It is known that a sequence of m-dimensional random vectors $\{(U_{1N}, \ldots, U_{mN})'\}_{N=1}^{\infty}$ converges to a constant vector $\mathbf{c} = (c_1, \ldots, c_m)'$ in probability if and only if $\{U_{iN}\}_{N=1}^{\infty}$ converges to c_i in probability for each $i = 1, \ldots, m$. Use this result to establish the facts in Problems 3.30 and 3.31.

3.30. Let $\bar{\mathbf{X}}_N$ denote the sample mean vector from an $\mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ population. Show that $\bar{\mathbf{X}}_N$ converges to $\boldsymbol{\mu}$ in probability as $N \to \infty$.

3.31. Let $\mathbf{S}_N$ be the sample covariance matrix of a random sample of size N from an $\mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ population, $\boldsymbol{\Sigma} > 0$. Show that $\mathbf{S}_N$ converges to $\boldsymbol{\Sigma}$ in probability as $N \to \infty$.

3.32. Let $\mathbf{W}$ denote a Wishart matrix and $\mathbf{W}_{11}$ the submatrix defined in (3.5.17). Show that $\mathbf{W}_{11}$ has a Wishart distribution.

3.33. Verify that in the proof of Lemma 3.5.2, $\mathbf{U}_1, \ldots, \mathbf{U}_{(N-1)-(n-k)}$ are i.i.d. $\mathcal{N}_k(\mathbf{0}, \boldsymbol{\Sigma}_{11\cdot 2})$ variables.

3.34. Let T^2 be Hotelling's T^2 statistic defined in (3.5.20). Show that as $N \to \infty$, the limiting distribution of T^2 is $\chi^2(n)$.

3.35. Let $\hat{R} = \hat{R}_{1\cdot 2\ldots n}$ be the sample multiple correlation coefficient of a random sample of size N from an $\mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ population, $\boldsymbol{\Sigma} > 0$. Show that if $R_{1\cdot 2\ldots n} = 0$, then the limiting distribution of $N\hat{R}^2/(1 - \hat{R}^2)$ is $\chi^2(n-1)$.

3.36. Show that when $n = 2$, the distribution of the sample multiple correlation coefficient given in Theorem 3.5.8 reduces to that given in Theorems 2.2.1 and 2.2.2.

3.37. Complete the proof of Theorem 3.5.9.