7 Multivariate Distributions: Definitions and Results

7.1 Probability Distributions

Let $\mathbf{X} = (X_1, \ldots, X_p)^T$ be a random vector.

The distribution of $\mathbf{X}$ can be specified via its cumulative distribution function:

Definition 7.1 (Joint c.d.f.)
The (joint) cumulative distribution function (c.d.f.) of the random vector $\mathbf{X}$ is given by

$$F_{\mathbf{X}}(\mathbf{x}) \;=\; F_{\mathbf{X}}(x_1, \ldots, x_p) \;=\; \mathbb{P}(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_p \leq x_p) \;=\; \mathbb{P}(\mathbf{X} \leq \mathbf{x}).$$

Equivalent specifications of the probabilistic behaviour of $\mathbf{X}$ can be specified in the cases where either all its components are *discrete*, so that joint probability mass functions can be defined, or all are *(absolutely) continuous* so that joint probability density functions can be defined. In these lectures we shall be considering only continuous random vectors. We shall therefore give definitions and use notation appropriate to continuous random vectors. The equivalent definitions, notation and results for discrete random vectors are immediate.

Definition 7.2 (Joint p.d.f. of a continuous random vector)
If the components of the random vector $\mathbf{X}$ are all (absolutely) continuous, i.e. $X_1, X_2, \ldots, X_p$ are (absolutely) continuous random variables, then $\mathbf{X}$ is also said to be *(absolutely) continuous* and its *(joint) probability density function* (p.d.f.) is given by

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{\partial^p F_{\mathbf{X}}(\mathbf{x})}{\partial x_1 \partial x_2 \ldots \partial x_p}.$$

Definition 7.3 (Range (or Support) for continuous random vector)
Suppose that $\mathbf{X}$ is a random vector with p.d.f. $f_{\mathbf{X}}(\cdot)$. Then the *range space* or *support* of $\mathbf{X}$ is given by

$$\Omega_{\mathbf{X}} = \{\mathbf{x} \in \mathbb{R}^p : f_{\mathbf{X}}(\mathbf{x}) > 0\}.$$

Proposition 7.4 (Characterization of a p.d.f.)
$f_{\mathbf{X}}(\cdot)$ is a p.d.f. if, and only if,
(a) $f_{\mathbf{X}}(\mathbf{x}) \geq 0$ for all $\mathbf{x}$
(b) $\int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x}) dx_1 dx_2 \ldots dx_p = 1$.

Marginal distributions
The marginal p.d.f. of any subset of $X_1, X_2, \ldots, X_p$ is found by integrating the joint p.d.f. of $\mathbf{X}$ with respect to the variables outside the subset of interest.
For example,

$$f_{X_i}(x_i) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x}) dx_1 dx_2 \ldots dx_{i-1} dx_{i+1} \ldots dx_p$$

is the marginal p.d.f of the random variable X_i.

Similarly we may consider the partitioned random vector $\mathbf{X}^T = (\mathbf{X}_1^T, \mathbf{X}_2^T)$ where $\mathbf{X}_1 = (X_1, \ldots, X_k)^T$ and $\mathbf{X}_2 = (X_{k+1}, \ldots, X_p)^T$. Then

$$f_{\mathbf{X}_1}(\mathbf{x}_1) \;=\; \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x}_1, \mathbf{x}_2) dx_{k+1} \ldots dx_p$$

is the marginal p.d.f of the random vector $\mathbf{X}_1$.

Conditional Distributions
If $\mathbf{X}$ is an absolutely continuous random vector, partitioned as before as $\mathbf{X}^T = (\mathbf{X}_1^T, \mathbf{X}_2^T)$, then for any given value of $\mathbf{X}_1, \mathbf{x}_1^0$, the conditional joint density function of $\mathbf{X}_2$, given $\mathbf{X}_1 = \mathbf{x}_1^0$, is given by

$$f_{\mathbf{X}_2|\mathbf{X}_1=\mathbf{x}_1^0}(\mathbf{x}_2|\mathbf{x}_1^0) = \frac{f_{\mathbf{X}}(\mathbf{x}_1^0, \mathbf{x}_2)}{f_{\mathbf{X}_1}(\mathbf{x}_1^0)}.$$

Similarly

$$f_{\mathbf{X}_1|\mathbf{X}_2=\mathbf{x}_2^0}(\mathbf{x}_1|\mathbf{x}_2^0) = \frac{f_{\mathbf{X}}(\mathbf{x}_1, \mathbf{x}_2^0)}{f_{\mathbf{X}_2}(\mathbf{x}_2^0)}.$$

Independence
When the conditional p.d.f $f_{\mathbf{X}_2|\mathbf{X}_1=\mathbf{x}_1^0}$ is the same for all values of $\mathbf{x}_1^0$, then $\mathbf{X_1}$ and $\mathbf{X_2}$ are said to be statistically independent. Clearly $\mathbf{X_1}$ and $\mathbf{X_2}$ statistically independent implies:

$$f_{\mathbf{X}_2|\mathbf{X}_1=\mathbf{x}_1^0}(\mathbf{x}_2|\mathbf{x}_1^0) \;=\; f_{\mathbf{X}_2}(\mathbf{x}_2) \quad \text{(the marginal density)}.$$

Proposition 7.5
$\mathbf{X_1}$ and $\mathbf{X_2}$ are statistically independent if, and only if,

$$f_{\mathbf{X}}(\mathbf{x}) \;=\; f_{\mathbf{X}_1}(\mathbf{x}_1)\, f_{\mathbf{X}_2}(\mathbf{x}_2).$$

Proposition 7.6
The component random variables $X_1, X_2, \ldots, X_p$ are mutually independent if, and only if,

$$f_{\mathbf{X}}(\mathbf{x}) \;=\; \prod_{i=1}^{p} f_{X_i}(x_i).$$

7.2 Expectation

We need to extend our notion of expectation in order to deal with matrices (including vectors) with random components.

Definition 7.7 (Expectation)
Suppose that Q is a $m \times n$ *random matrix*: that is to say, its components, the $\{Q_{ij}\}$, are random variables.

Then the *mean* or *expectation* of Q, i.e. $E[Q]$, is given by the $m \times n$ matrix whose (i,j)'th component is $E[Q_{ij}]$, $i=1,\ldots,m$, $j=1,\ldots,n$.

If g is a real-valued function of a random vector $\mathbf{X}$, then

$$E[g(\mathbf{X})] \;=\; \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} g(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) dx_1 \ldots dx_p.$$

If G is a matrix (or vector)-valued function of a random vector $\mathbf{X}$, e.g. $G(\mathbf{X}) = (g_{ij}(\mathbf{X}))$, then

$$E[G(\mathbf{X})] \;=\; (E[g_{ij}(\mathbf{X})]).$$

As usual, in all these cases expectation is linear. For example, where Q is a $m \times n$ *random matrix*, A is a $k \times m$ matrix of constants and B is a $k \times n$ matrix of constants, then

$$E[AQ + B] \;=\; AE[Q] + B.$$

7.3 Population Parameters

Definition 7.8 (Mean Vector, $\boldsymbol{\mu}$)
Suppose that $\mathbf{X}$ is a $p \times 1$ random vector. Then the *mean*, or *expectation*, of $\mathbf{X}$ is given by

$$E[\mathbf{X}] = \begin{pmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_p) \end{pmatrix} \;=\; \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} \;=\; \boldsymbol{\mu}$$

where

$$\mu_i = E[X_i] = \int_{-\infty}^{\infty} x\, f_{X_i}(x) dx \quad i=1,\ldots,p.$$

Definition 7.9 (Variance-Covariance Matrix, $\boldsymbol{\Sigma}$)
Given a random vector $\mathbf{X}$, the variously called *covariance*, *variance-covariance*, or *dispersion* matrix of $\mathbf{X}$ is the $p \times p$ matrix $\boldsymbol{\Sigma}$, where

$$\boldsymbol{\Sigma} \;=\; E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T] \;=\; (E[(X_i - \mu_i)(X_j - \mu_j)]) \;=\; (\sigma_{ij})$$

and

$$\sigma_{ij} \;=\; \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x_i \,-\, \mu_i)(x_j \,-\, \mu_j)\, f_{(X_i,X_j)}(x_i, x_j)\, dx_i\, dx_j \quad i,j=1,\ldots,p.$$

We may write this as

$$\boldsymbol{\Sigma} \;=\; \mathrm{Cov}(\mathbf{X}, \mathbf{X}) \quad \text{or} \quad \boldsymbol{\Sigma} \;=\; \mathrm{Var}(\mathbf{X}).$$

In general we find it convenient to write

$$\mathrm{Var}(X_i) \;=\; \sigma_{ii} \;=\; \sigma_i^2.$$

The following results are easy consequences of the definitions.

(i) $\boldsymbol{\Sigma} = E[\mathbf{X}\mathbf{X}^T] - \boldsymbol{\mu}\boldsymbol{\mu}^T$.

(ii) Given a $p \times 1$ vector of constants $\mathbf{a}$, consider the *linear compound* $Y = \mathbf{a}^T\mathbf{X}$, so that Y is a random variable. Then

$$E[\mathbf{a}^T\mathbf{X}] = \mathbf{a}^T\boldsymbol{\mu} \quad \text{and} \quad \text{Var}(\mathbf{a}^T\mathbf{X}) = \mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a}.$$

(iii) Given a $q \times p$ matrix of constants A, consider the linear transformation $\mathbf{Y} = A\mathbf{X}$, so that $\mathbf{Y}$ is a $q \times 1$ random vector. Then

$$E[A\mathbf{X}] = A\boldsymbol{\mu} \quad \text{and} \quad \text{Var}(A\mathbf{X}) = A\boldsymbol{\Sigma}A^T.$$

Definition 7.10 (Correlation Matrix, P)
The correlation matrix is defined to be the $p \times p$ matrix $P = (\rho_{ij})$, where

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}} = \frac{\sigma_{ij}}{\sigma_i\sigma_j}.$$

This can be written as

$$\mathbf{P} = \mathbf{D^{-1}}\boldsymbol{\Sigma}\mathbf{D^{-1}},$$

where $D = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_p)$.

Note: If $\mathbf{X}$ is a $p \times 1$ random vector, with covariance and correlation matrices given by $\boldsymbol{\Sigma}$ and $\mathbf{P}$ respectively, then:
(a) $\boldsymbol{\Sigma}$ and $\mathbf{P}$ are symmetric and positive semi-definite (psd);
(b) if $\boldsymbol{\Sigma}$, and hence $\mathbf{P}$, are of full rank, then they are both positive definite (pd);
(c) if $\boldsymbol{\Sigma}$ is of full rank, then $\text{Var}(\mathbf{a}^T\mathbf{X}) = \mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a} > 0$ for all $p \times 1$ constant vectors $\mathbf{a} \neq \mathbf{0}$.

7.4 Random Samples from a Multivariate Distribution

We shall assume that we have a random sample of size n taken from a population corresponding to a $p \times 1$ random vector $\mathbf{X}$, with population mean $\boldsymbol{\mu}$, and population covariance matrix $\boldsymbol{\Sigma}$.

7.4.1 Notation

Let n represent the number of individuals in the sample, i.e. the sample size, then, in total, we have $n \times p$ measurements, which we can represent in the *data matrix* X. Then

$$X = \begin{bmatrix} X_{11} & X_{12} & \ldots & X_{1p} \\ X_{21} & X_{22} & \ldots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \ldots & X_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1^T \\ \mathbf{X}_2^T \\ \vdots \\ \mathbf{X}_n^T \end{bmatrix}.$$

The $\mathbf{X}_i$, $i = 1, \ldots, n$, are assumed to constitute a random sample, i.e. they are mutually independent and share the same underlying multivariate distribution.

7.4.2 Estimation

We propose estimators for $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ (and indeed $\mathbf{P}$), based on the data matrix X.

Definition 7.11 (Sample Means)
The *sample mean* of the j-th variable is given by

$$\overline{X}_j = \frac{1}{n}\sum_{r=1}^{n} X_{rj}, \quad j=1,\ldots,p.$$

The *sample mean vector* is given by

$$\overline{\mathbf{X}} = (\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_p)^T = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i$$

Remarks 7.12
(i) $\overline{\mathbf{X}}$ is unbiased for $\boldsymbol{\mu}$. That is, $E[\overline{\mathbf{X}}] = \boldsymbol{\mu}$.
(ii) Observed values of $\overline{X}_j$ and $\overline{\mathbf{X}}$ are denoted by $\overline{x}_j$ and $\overline{\mathbf{x}}$ respectively.

Definition 7.13 (Sample Covariance Matrix)
The *sample covariance matrix* is given by

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})^T$$

Clearly

$$\mathbf{S} = (S_{ij}), \quad \text{where} \quad S_{ij} = \frac{1}{n-1}\sum_{r=1}^{n}(X_{ri} - \overline{X}_i)(X_{rj} - \overline{X}_j).$$

Remarks 7.14
(i) It can be shown that $\mathbf{S}$ is unbiased for $\boldsymbol{\Sigma}$. That is, $E[\mathbf{S}] = \boldsymbol{\Sigma}$.
(ii) The diagonal elements of $\mathbf{S}$, the S_{ii}, are sample variances and are denoted by $S_i^2 = S_{ii}$.

Definition 7.15 (Sample Correlation)
The *sample correlation coefficient* between X_i and X_j is given by:

$$R_{ij} = \frac{S_{ij}}{S_i S_j}$$

with corresponding *sample correlation matrix* given by $\mathbf{R} = (R_{ij})$. Setting $\widehat{\mathbf{D}} = \text{diag}(S_1, \ldots, S_p)$, we have

$$\mathbf{R} = \widehat{\mathbf{D}}^{-\mathbf{1}}\mathbf{S}\widehat{\mathbf{D}}^{-\mathbf{1}}.$$

7.5 The Multivariate Normal Distribution and Properties

Definition 7.16 (Multivariate Normal Distribution)
The Cramer-Wold Theorem tells us that the distribution of a p-dimensional random vector $\mathbf{X} = (X_1, \ldots, X_p)^T$ is completely determined by the univariate distributions of all its linear compounds $Y = \mathbf{a}^T\mathbf{X}$.

Suppose $\mathbf{X}$ is a $(p \times 1)$ random vector with population mean $\boldsymbol{\mu}$ and population covariance matrix $\boldsymbol{\Sigma}$. It is defined to have the *multivariate normal distribution* if and only if every linear compound $\mathbf{a}^T\mathbf{X}$ has a univariate normal distribution. That is if and only if

$$Y = \mathbf{a}^T\mathbf{x} \sim \mathrm{N}(\mathbf{a}^T\boldsymbol{\mu}, \mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a}), \ \forall \mathbf{a}.$$

If $\boldsymbol{\Sigma}$ is non-singular (i.e. of full rank), so that $\boldsymbol{\Sigma}^{-1}$ exists, then the joint p.d.f. of $\mathbf{X}$ is given by

$$f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{|2\pi\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \right\}$$

In either case, we can write

$$\mathbf{X} \sim \mathrm{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \text{or} \quad \mathbf{X} \sim \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

It can be shown that $E[\mathbf{X}] = \boldsymbol{\mu}$ and $E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T] = \boldsymbol{\Sigma}$.

7.5.1 Some Properties of the Multivariate Normal Distribution

Assume that $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Property 1. Let $\mathbf{Y} = A\mathbf{X}$, where A is a $q \times p$ matrix of constants.
Then

$$\mathbf{Y} \sim N_q(A\boldsymbol{\mu}, A\boldsymbol{\Sigma}A^T).$$

Property 2. It follows from Property 1, by a suitable choice of A, that the marginal distribution of *any* subset of the X_i has a multivariate normal distribution, whose mean vector and covariance matrix are obtained as the appropriate sub-vector of $\boldsymbol{\mu}$ and sub-matrix of $\boldsymbol{\Sigma}$ corresponding to the chosen subset of the X_i.

Property 3. Suppose that $\mathbf{X}$ is partitioned into two portions so that $\mathbf{X}^T = (\mathbf{X}_1^T, \mathbf{X}_2^T)$, where $\mathbf{X}_1$ has q components and $\mathbf{X}_2$ has $(p - q)$ components. Partition $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in the obviously conformable way

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

Then, for a given value of $\mathbf{X}_2$, $\mathbf{b}$, the conditional distribution of $\mathbf{X}_1$, given $\mathbf{X}_2 = \mathbf{b}$, is multivariate normal with mean

$$\boldsymbol{\mu}_1 \, + \, \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{b} \, - \, \boldsymbol{\mu}_2)$$

and covariance matrix

$$\boldsymbol{\Sigma}_{11} \, - \, \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}.$$

[Clearly we require that $\boldsymbol{\Sigma}_{22}$ is of full rank $(p-q)$.]

Property 4. If $\boldsymbol{\Sigma}$ is of full rank, contours of equi-probability of $\mathbf{X}$ in $\mathbb{R}^p$ are given by

$$(\mathbf{x} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \, = \, C,$$

for some constant C. This equation defines a hyper-ellipsoid in $\mathbb{R}^p$.

Property 5. If $\boldsymbol{\Sigma}$ is of full rank, then there exists a $p \times p$ symmetric matrix of full rank p, B, such that $B^2 = \boldsymbol{\Sigma}$. We write $B = \boldsymbol{\Sigma}^{\frac{1}{2}}$ and call it a symmetric square root of $\boldsymbol{\Sigma}$. Correspondingly B^{-1} is a symmetric square root of $\boldsymbol{\Sigma}^{-1}$ and we write $B^{-1} = \boldsymbol{\Sigma}^{-\frac{1}{2}}$. Defining

$$\mathbf{Z} = \boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{X} - \boldsymbol{\mu})$$

we have

$$\mathbf{Z} \sim N_p(\mathbf{0}, \mathbf{I}).$$

Property 6. If $\boldsymbol{\Sigma}$ is of full rank, then

$$(\mathbf{X} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu}) \sim \chi^2_p.$$

In a later chapter we shall introduce a multivariate generalization of the χ^2 distribution, and generalizations of the t and F distributions.

7.5.2 Estimation of the Parameters in the Multivariate Normal Distribution

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be a random sample from $\mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown. That is we have the $n \times p$ data matrix

$$X = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix}$$

We shall use maximum likelihood estimation. The likelihood function $L(\boldsymbol{\mu}, \boldsymbol{\Sigma}; X)$ is given by

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}; X) = |2\pi\boldsymbol{\Sigma}|^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) \right\}$$

and the log-likelihood function $\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma}; X)$ is given by

$$\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X}) = -\frac{n}{2}\ln|2\pi\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) \tag{1}$$

Now,

$$\begin{aligned}(\mathbf{x}_i - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) &= [(\mathbf{x}_i - \bar{\mathbf{x}}) + (\bar{\mathbf{x}} - \boldsymbol{\mu})]^T\boldsymbol{\Sigma}^{-1}[(\mathbf{x}_i - \bar{\mathbf{x}}) + (\bar{\mathbf{x}} - \boldsymbol{\mu})] \\ &= (\mathbf{x}_i - \bar{\mathbf{x}})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}) + 2(\mathbf{x}_i - \bar{\mathbf{x}})^T\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) + (\bar{\mathbf{x}} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu})\end{aligned} \tag{2}$$

where $\bar{\mathbf{x}} = \dfrac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$, so that summing over i gives

$$\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) = \sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}) + n(\bar{\mathbf{x}} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) \tag{3}$$

Each $(\mathbf{x}_i - \bar{\mathbf{x}})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})$ is a scalar and equals its trace. i.e.

$$\text{tr}\left((\mathbf{x}_i - \bar{\mathbf{x}})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})\right) = \text{tr}\left(\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T\right)$$

Again, summing over i gives

$$\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}) = \text{tr}\left(\Sigma^{-1}\left\{\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T\right\}\right) \tag{4}$$

Writing

$$\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T = A \qquad (= (n-1)\mathbf{S}) \tag{5}$$

then (1) becomes

$$\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X}) = -\frac{np}{2}\ln(2\pi) - \frac{n}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\text{tr}(\boldsymbol{\Sigma}^{-1}A) - \frac{n}{2}(\bar{\mathbf{x}} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) \tag{6}$$

(since $|2\pi\boldsymbol{\Sigma}| = (2\pi)^p|\boldsymbol{\Sigma}|$). The matrix A is the sums-of-squares-and-products matrix (SSP matrix) for the sample.

We will now show that the m.l.e.'s of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$ and $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}A$.

Given the data matrix X, i.e. given $\bar{\mathbf{x}}$ and A, ℓ is a function of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

Now, $\boldsymbol{\Sigma}^{-1}$ is positive-definite so that

$$\begin{aligned} &(\bar{\mathbf{x}} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}) \geq 0, \text{ and } = 0 \textit{ if and only if } \hat{\boldsymbol{\mu}} = \bar{\mathbf{x}} \\ &\Rightarrow \hat{\boldsymbol{\mu}} = \bar{\mathbf{x}} \end{aligned} \tag{7}$$

We must choose $\hat{\boldsymbol{\Sigma}}$ such that the second and third terms in (6) are maximized. i.e such that

$$-\frac{n}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}A)$$

is maximized. This expression is a real-valued function of the $p \times p$ symmetric matrix $\boldsymbol{\Sigma} = (\sigma_{ij})$, and, by definition, the derivative with respect to the matrix $\boldsymbol{\Sigma}$ is just the matrix of the derivatives with respect to the corresponding elements of $\boldsymbol{\Sigma}$. Details of such derivatives and associated results are given in Mardia, Kent and Bibby, and in the "Notes for MSc Students".

We have

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pp} \end{pmatrix}, \ \sigma_{ij} = \sigma_{ji}.$$

Then

$$\frac{\partial}{\partial \sigma_{ii}}\ln|\boldsymbol{\Sigma}| = (\boldsymbol{\Sigma}^{-1})_{ii} \tag{8}$$

$$\frac{\partial}{\partial \sigma_{ij}}\ln|\boldsymbol{\Sigma}| = 2(\boldsymbol{\Sigma}^{-1})_{ij}, \qquad i \neq j \tag{9}$$

and

$$\frac{\partial}{\partial \sigma_{ii}}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}A) = -(\boldsymbol{\Sigma}^{-1}A\boldsymbol{\Sigma}^{-1})_{ii} \tag{10}$$

$$\frac{\partial}{\partial \sigma_{ij}}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}A) = -2(\boldsymbol{\Sigma}^{-1}A\boldsymbol{\Sigma}^{-1})_{ij}, \quad i \neq j \tag{11}$$

so that,

$$(8) \text{ and } (10) \Rightarrow \frac{\partial \ell}{\partial \sigma_{ii}} = \frac{1}{2}\{\boldsymbol{\Sigma}^{-1}A\boldsymbol{\Sigma}^{-1} - n\boldsymbol{\Sigma}^{-1}\}_{ii} \tag{12}$$

and

$$(9) \text{ and } (11) \Rightarrow \frac{\partial \ell}{\partial \sigma_{ij}} = \{\boldsymbol{\Sigma}^{-1}A\boldsymbol{\Sigma}^{-1} - n\boldsymbol{\Sigma}^{-1}\}_{ij}, \quad i \neq j \tag{13}$$

Further,

$$(12) \text{ and } (13) \Rightarrow \frac{\partial \ell}{\partial \boldsymbol{\Sigma}} = \mathbf{M} - \frac{1}{2}\text{diag}(m_{ii}) \tag{14}$$

where,

$$\mathbf{M} = (m_{ij}) = \boldsymbol{\Sigma}^{-1}A\boldsymbol{\Sigma}^{-1} - n\boldsymbol{\Sigma}^{-1} \tag{15}$$

Equating (14) to the zero matrix gives

$$\begin{aligned} &\boldsymbol{\Sigma}^{-1}A\boldsymbol{\Sigma}^{-1} = n\boldsymbol{\Sigma}^{-1} \\ \Rightarrow \quad &\boldsymbol{\Sigma}^{-1}A = n\mathbf{I} \end{aligned}$$

$$\Rightarrow \hat{\boldsymbol{\Sigma}} = \frac{1}{n}A \qquad \left(= \frac{n-1}{n}\mathbf{S}\right) \tag{16}$$

(where $\hat{\boldsymbol{\Sigma}}$ uses $\frac{1}{n}$ and not $\frac{1}{n-1}$, so is a *biased* estimator).

That is, the ML estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are given by

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$$

and

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T.$$

Note 1. Need to prove that this stationary point gives a maximum - see Mardia, Kent and Bibby.

Note 2. It is sometimes useful to distinguish between the biased and unbiased estimates of $\boldsymbol{\Sigma}$ and then we set

$$\mathbf{S} = \frac{1}{n}A \quad \text{and} \quad \mathbf{S}_U = \frac{1}{n-1}A,$$

where $\mathbf{S}_U$ is now the unbiased estimator of the sample covariance matrix (with divisor $(n-1)$) and the notation $\mathbf{S}$ is used for the m.l.e of $\boldsymbol{\Sigma}$.

From (6) we have

$$\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X}) = -\frac{np}{2}\ln(2\pi) - \frac{n}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}A) - \frac{n}{2}(\bar{\mathbf{x}} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) \qquad (17)$$

and, replacing $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ by their m.l.e.'s, $\bar{\mathbf{x}}$ and $\mathbf{S} = \frac{1}{n}A$, we have that the maximized log-likelihood is

$$\ell^*(\bar{\mathbf{x}}, \mathbf{S}; \mathbf{X}) \;=\; -\frac{n}{2}\{p(1 + \ln(2\pi)) + \ln|\mathbf{S}|\}. \qquad (18)$$

7.5.3 Some Properties of the Estimators $\overline{\mathbf{X}}$, S and $\mathbf{S}_U$

Property 1. If the mean vector $\overline{\mathbf{X}}$ and the SSP matrix A are based on a random sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ from MVN$(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\overline{\mathbf{X}}$ and A are independent. Thus $\overline{\mathbf{X}}$ is independent of both $\mathbf{S}$ and $\mathbf{S}_U$.

Property 2. If $\overline{\mathbf{X}}$ is based on a random sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ from MVN$(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\overline{\mathbf{X}} \sim$ MVN$(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma})$.

This can be established by using moment generating functions in a manner similar to the univariate case.

Property 3. Even if $\mathbf{X}$ is not normally distributed, the distribution of $\overline{\mathbf{X}}$ will still be approximately normal. This result, known as the multivariate central limit theorem, is as follows:

If $\overline{\mathbf{X}}$ is the mean vector of a random sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ from a population with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ then, as $n \to \infty$, the distribution of $\sqrt{n}(\overline{\mathbf{X}} - \boldsymbol{\mu})$ approaches MVN$(\mathbf{0}, \boldsymbol{\Sigma})$. Thus, for large n, $\overline{\mathbf{X}}$ is approximately MVN$(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma})$.

Property 4. The SSP matrix A can be written as

$$A \;=\; \sum_{i=1}^{n-1} \mathbf{Z}_i\,\mathbf{Z}_i^T,$$

where the $\mathbf{Z}_i$ are iid MVN$(\mathbf{0}, \boldsymbol{\Sigma})$. [See Theorem 3.3.2 in Anderson (1984) for proofs of Properties 1 and 4.]

Note: Compare the univariate result corresponding to Property 4.

$$(n-1)s^2 \;=\; \sum_{i=1}^{n-1} Z_i^2, \quad \text{where} \quad Z_i \sim \mathrm{NID}(0, \sigma^2), \quad \Rightarrow \quad (n-1)s^2 \sim \sigma^2\chi^2_{n-1}.$$

We shall consider the distribution of A, and hence $\mathbf{S}$ and $\mathbf{S}_U$ in a later chapter.