CHAPTER 1

Mathematical and Statistical Preliminaries

This chapter provides a quick review of basic advanced calculus, linear algebra and probability and statistics, serving as background knowledge for the material in later chapters. Readers familiar with the relevant terminologies are encouraged to start their journey from Chapter 3 onward, and then turn back to this chapter as a reference from time to time.

With the rapid advance in computing technology and the enrichment of data storage, we often collect a huge amount of information and observations with a large number of feature variables. A dataset with p numbers of variables and n numbers of observations can be arranged in the form of an $n \times p$ data matrix:

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1^\top \\ \vdots \\ \boldsymbol{x}_n^\top \end{pmatrix},$$

where the j-th row $\boldsymbol{x}_j^\top$, for $j = 1, \ldots, n$, represents the j-th observation, the transpose of the column vector $\boldsymbol{x}_j$. Data scientists believe that there exists a data generating mechanism so that each sample of multivariate observation is randomly drawn from a common but often unknown joint probability distribution. Statistically speaking, the data matrix X is the realization of the following random matrix:

$$\mathbf{X} = \begin{pmatrix} \mathrm{x}_{11} & \cdots & \mathrm{x}_{1p} \\ \vdots & \ddots & \vdots \\ \mathrm{x}_{n1} & \cdots & \mathrm{x}_{np} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_n^\top \end{pmatrix}, \tag{1.1}$$

where the row vectors $\mathbf{x}_1^\top, \ldots, \mathbf{x}_n^\top$ are independent *random (row) vectors* following a common joint distribution. Throughout this book, we shall use roman type letters, such as x and X, to denote random variables, and use italic letters, such as x and X, to represent the corresponding observed values of sample data. In addition, we denote

vectors by bold lowercase letters, e.g. $\boldsymbol{x}$ and $\mathbf{x}$ represent deterministic and random vectors, respectively, and denote matrices by bold uppercase letters, e.g. $\boldsymbol{X}$ and $\mathbf{X}$ represent deterministic and random matrices, respectively. Due to the frequent use of vectors, matrices and their random counterparts in this book, the traditional notations in probability and statistics may cause some ambiguity. To this point our notations follow the convention in [9]; the same practice is also adopted in the companion book [4] of the present one. Also, for the notation of natural logarithm, we shall follow the notation "ln" throughout the book, while we may also use "log" and "$\log_e$" interchangeably depending on the context to avoid any ambiguity.

1.1 RANDOM VECTOR

From (1.1), we learn that a $p \times 1$ random (column) vector, denoted by $\mathbf{x} = (\mathrm{x}_1, \ldots, \mathrm{x}_p)^\top$, is just a vertical stack of p univariate random variables. The mean of $\mathbf{x}$ is the $p \times 1$ vector

$$\mathbb{E}(\mathbf{x}) := \begin{pmatrix} \mathbb{E}(\mathrm{x}_1) \\ \vdots \\ \mathbb{E}(\mathrm{x}_p) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix} =: \boldsymbol{\mu},$$

the variance-covariance matrix (or covariance matrix) of $\mathbf{x}$ is the following $p \times p$ (symmetric positive definite) matrix:

$$\mathrm{Var}(\mathbf{x}) = \mathbb{E}\left((\mathbf{x} - \mathbb{E}(\mathbf{x}))(\mathbf{x} - \mathbb{E}(\mathbf{x}))^\top\right) =: \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1p} \\ \vdots & \ddots & \vdots \\ \sigma_{p1} & \cdots & \sigma_{pp} \end{pmatrix} =: (\sigma_{ij})_{i,j=1,\ldots,p},$$

or (σ_{ij}) in shortened form if there is no cause for ambiguity; the correlation matrix of $\mathbf{x}$ is another symmetric and positive definite $p \times p$ matrix:

$$\mathrm{Corr}(\mathbf{x}) =: \boldsymbol{R} = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \cdots & \cdots & 1 \end{pmatrix} =: (\rho_{ij})_{i,j=1,\ldots,p},$$

where $\rho_{ij} := \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}}$ is the correlation coefficient between x_i and x_j. The following are some important properties that we shall use later:

1. If $\boldsymbol{A}$ is an $r \times p$ deterministic matrix, then $\mathbf{y} = \boldsymbol{A}\mathbf{x}$ is an $r \times 1$ random vector with mean $\mathbb{E}(\mathbf{y}) = \boldsymbol{A}\boldsymbol{\mu}$ and covariance matrix $\mathrm{Var}(\mathbf{y}) = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top$; moreover, for another $r \times 1$ random vector $\mathbf{z} = \boldsymbol{B}\mathbf{x}$, where $\boldsymbol{B}$ is another $r \times p$ constant matrix, the covariance of $\mathbf{y}$ and $\mathbf{z}$ is $\mathrm{Cov}(\boldsymbol{A}\mathbf{x}, \boldsymbol{B}\mathbf{x}) = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{B}^\top$.

2. The correlation matrix $\boldsymbol{R} = (\rho_{ij})$ of $\mathbf{x}$ can be computed from its covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ij})$ by the formula $\boldsymbol{R} = \boldsymbol{D}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{D}^{-1/2}$, where $\boldsymbol{D}$ is the diagonal matrix denoted by $\text{diag}(\sigma_{11}, \ldots, \sigma_{pp})$, and $\boldsymbol{D}^{-1/2} := \text{diag}\big(1/\sqrt{\sigma_{11}}, \ldots, 1/\sqrt{\sigma_{pp}}\big)$.

1.1.1 Random Sample, Sample Mean and Covariance Matrix

In many real-life applications, the underlying probability density function (or cumulative distribution function) is unknown and requires estimation. We are particularly interested in estimating the unknown mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^\top$ and covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ij})$. For a given data matrix $\boldsymbol{X}$, it is customary to estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ by using the sample mean vector $\overline{\boldsymbol{x}} := (\overline{x}_1, \ldots, \overline{x}_p)^\top$ and the sample covariance matrix $S := (s_{ij})$, respectively, where

$$\overline{x}_i = \sum_{k=1}^{n} \frac{x_{ki}}{n} \text{ and } s_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j) = \frac{1}{n-1}\left(\sum_{k=1}^{n} x_{ki}x_{kj} - n\overline{x}_i\overline{x}_j\right).$$

The next proposition demonstrates that these estimation methods enjoy the well-known desirable property known as *unbiasedness*.

Proposition 1.1 ***(Unbiasedness of the Sample Mean Vector and Covariance Matrix)*** Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be a random sample from a common joint probability distribution which has the mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Then the $p \times 1$ sample mean vector

$$\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$$

is an unbiased estimator of $\boldsymbol{\mu}$, that is, $\mathbb{E}(\overline{\mathbf{x}}) = \boldsymbol{\mu}$; besides, the $p \times p$ covariance matrix of $\overline{\mathbf{x}}$ is given by

$$\text{Var}(\overline{\mathbf{x}}) = \frac{1}{n}\boldsymbol{\Sigma}.$$

Moreover, the $p \times p$ sample covariance matrix

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

is an unbiased estimator of $\boldsymbol{\Sigma}$, that is, $\mathbb{E}(\mathbf{S}) = \boldsymbol{\Sigma}$.

Proof.

$$\mathbb{E}(\overline{\mathbf{x}}) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(\mathbf{x}_i) = \frac{1}{n}\cdot n \cdot \boldsymbol{\mu} = \boldsymbol{\mu}.$$

Note that

$$\mathrm{Var}(\overline{\mathbf{x}}) = \mathbb{E}\left((\overline{\mathbf{x}} - \boldsymbol{\mu})(\overline{\mathbf{x}} - \boldsymbol{\mu})^\top\right) = \mathbb{E}\left(\left(\frac{1}{n}\sum_{j=1}^{n}(\mathbf{x}_j - \boldsymbol{\mu})\right)\left(\frac{1}{n}\sum_{l=1}^{n}(\mathbf{x}_l - \boldsymbol{\mu})\right)^\top\right)$$
$$= \frac{1}{n^2}\sum_{j=1}^{n}\sum_{l=1}^{n}\mathbb{E}\left((\mathbf{x}_j - \boldsymbol{\mu})(\mathbf{x}_l - \boldsymbol{\mu})^\top\right).$$

For $j \neq l$, each entry in $\mathbb{E}\left((\mathbf{x}_j - \boldsymbol{\mu})(\mathbf{x}_l - \boldsymbol{\mu})^\top\right)$ vanishes because each entry is the covariance between a component of $\mathbf{x}_j$ and that of $\mathbf{x}_l$, and they are independent by the definition of random sample. Therefore,

$$\mathrm{Var}(\overline{\mathbf{x}}) = \frac{1}{n^2}\sum_{j=1}^{n}\mathbb{E}\left((\mathbf{x}_j - \boldsymbol{\mu})(\mathbf{x}_j - \boldsymbol{\mu})^\top\right).$$

Since $\boldsymbol{\Sigma} = \mathbb{E}\left((\mathbf{x}_j - \boldsymbol{\mu})(\mathbf{x}_j - \boldsymbol{\mu})^\top\right)$ is the common population covariance matrix for each $\mathbf{x}_j$, we have

$$\mathrm{Var}(\overline{\mathbf{x}}) = \frac{1}{n^2}\sum_{j=1}^{n}\mathbb{E}\left((\mathbf{x}_j - \boldsymbol{\mu})(\mathbf{x}_j - \boldsymbol{\mu})^\top\right) = \frac{1}{n^2}\cdot n\cdot\boldsymbol{\Sigma} = \frac{1}{n}\boldsymbol{\Sigma}.$$

Next, we consider the $p \times p$ matrix:

$$\mathbf{A} := \sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^\top = \sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top - n\overline{\mathbf{x}}\,\overline{\mathbf{x}}^\top, \tag{1.2}$$

which represents the matrix of sums of squares and cross products. Note that

$$\mathbb{E}\left(\mathbf{x}_i\mathbf{x}_i^\top\right) = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top \text{ and } \mathbb{E}\left(\overline{\mathbf{x}}\,\overline{\mathbf{x}}^\top\right) = \frac{1}{n}\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top.$$

Hence,

$$\mathbb{E}(\mathbf{A}) = \sum_{i=1}^{n}\mathbb{E}\left(\mathbf{x}_i\mathbf{x}_i^\top\right) - n\mathbb{E}\left(\overline{\mathbf{x}}\,\overline{\mathbf{x}}^\top\right) = n\boldsymbol{\Sigma} + n\boldsymbol{\mu}\boldsymbol{\mu}^\top - n\left(\frac{1}{n}\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top\right)$$
$$= n\boldsymbol{\Sigma} + n\boldsymbol{\mu}\boldsymbol{\mu}^\top - \boldsymbol{\Sigma} - n\boldsymbol{\mu}\boldsymbol{\mu}^\top = (n-1)\boldsymbol{\Sigma}.$$

It then follows that

$$\mathbb{E}(\mathbf{S}) = \mathbb{E}\left(\frac{1}{n-1}\mathbf{A}\right) = \boldsymbol{\Sigma}.$$

□

Example 1.1 ***(Blue Chips on the Hong Kong Stock Exchange)*** The data file `fin-ratio.csv`[1] contains financial ratios of 680 securities listed on the main board of the Hong Kong Stock Exchange in 2002. There are six financial variables in order, namely, **Earning Yield** (EV), **Cash Flow to Price** (CFTP), **logarithm of Market Value** (ln_MV), **Dividend Yield** (DY), **Book to Market Equity** (BTME), and **Debt to Equity Ratio** (DTE). The last column **HSI** is a binary variable which takes value one if the company is a Blue Chip or zero otherwise. Among these companies, there were 32 Blue Chips which were the *Hang Seng Index Constituent Stocks* at that time. For the time being we ignore this HSI variable and compute the sample mean vector, sample covariance matrix, and sample correlation matrix using Python with the codes shown in Programme 1.1; the **R** codes are shown in Programme 1.2. For those who are not yet very familiar with Python and **R**, please refer to Chapter 2 for a quick introduction.

```
import pandas as pd
import numpy as np
np.set_printoptions(precision=4) # Control display into 4 digits

HSI_2002 = pd.read_csv("fin-ratio.csv")
print(HSI_2002.columns)

X_2002 = HSI_2002.drop(columns="HSI").values # A 680x6 data matrix
mu_2002 = np.mean(X_2002, axis=0) # Mean vector
print(mu_2002)

S_2002 = np.cov(X_2002, rowvar=False) # Covariance matrix
print(S_2002)

R_2002 = np.corrcoef(X_2002, rowvar=False) # Correlation matrix
print(R_2002)
```

```
['EY' 'CFTP' 'ln_MV' 'DY' 'BTME' 'DTE' 'HSI']
[-0.6502 -0.2339  6.2668  2.4962  1.9083  0.7097]
[[18.4979  2.909   1.1602  1.9204  1.4781  0.338 ]
 [ 2.909   3.6931  0.7663  1.2371  1.8228  0.3288]
 [ 1.1602  0.7663  2.7439  0.9721 -0.7734 -0.0741]
 [ 1.9204  1.2371  0.9721 13.8716 -0.2575  0.1582]
 [ 1.4781  1.8228 -0.7734 -0.2575 68.3082  1.9618]
 [ 0.338   0.3288 -0.0741  0.1582  1.9618 12.9929]]
[[ 1.      0.352   0.1628  0.1199  0.0416  0.0218]
 [ 0.352   1.      0.2407  0.1728  0.1148  0.0475]
 [ 0.1628  0.2407  1.      0.1576 -0.0565 -0.0124]
 [ 0.1199  0.1728  0.1576  1.     -0.0084  0.0118]
 [ 0.0416  0.1148 -0.0565 -0.0084  1.      0.0659]
 [ 0.0218  0.0475 -0.0124  0.0118  0.0659  1.    ]]
```

Programme 1.1 Computations of sample mean vector, sample covariance matrix, and sample correlation matrix of `fin-ratio.csv` in Python.

[1] This dataset is available at https://github.com/kaiser1999/Financial-Data-Analytics.

```
> options(digits=4) # Control display into 4 digits
>
> HSI_2002 <- read.csv("fin-ratio.csv")
> names(HSI_2002)
[1] "EY"    "CFTP"  "ln_MV" "DY"    "BTME"  "DTE"   "HSI"
>
> X_2002 <- HSI_2002[,1:6] # A 680x6 data matrix
> (mu_2002 <- apply(X_2002, 2, mean)) # Mean vector
     EY    CFTP   ln_MV      DY    BTME     DTE
-0.6502 -0.2339  6.2668  2.4962  1.9083  0.7097
>
> (S_2002 <- var(X_2002)) # Covariance matrix
           EY   CFTP    ln_MV      DY    BTME      DTE
EY    18.498 2.9090  1.16019  1.9204  1.4781  0.33795
CFTP   2.909 3.6931  0.76630  1.2371  1.8228  0.32879
ln_MV  1.160 0.7663  2.74394  0.9721 -0.7734 -0.07413
DY     1.920 1.2371  0.97207 13.8716 -0.2575  0.15815
BTME   1.478 1.8228 -0.77342 -0.2575 68.3082  1.96177
DTE    0.338 0.3288 -0.07413  0.1582  1.9618 12.99291
>
> (R_2002 <- cor(X_2002)) # Correlation matrix
           EY    CFTP    ln_MV        DY      BTME      DTE
EY    1.00000 0.35195  0.16285  0.119884  0.041583  0.02180
CFTP  0.35195 1.00000  0.24072  0.172848  0.114767  0.04746
ln_MV 0.16285 0.24072  1.00000  0.157561 -0.056493 -0.01242
DY    0.11988 0.17285  0.15756  1.000000 -0.008366  0.01178
BTME  0.04158 0.11477 -0.05649 -0.008366  1.000000  0.06585
DTE   0.02180 0.04746 -0.01242  0.011780  0.065850  1.00000
```

Programme 1.2 Computations of sample mean vector, sample covariance matrix, and sample correlation matrix of `fin-ratio.csv` in **R**.

1.2 MATRIX THEORY

Before we embark on our journey, let us first recall some basic notations and concepts in introductory matrix theory. Readers may refer to [11, 18, 19] for more detailed introductions to matrix theory at an elementary level; more advanced topics are discussed in [8, 20, 21], including various algorithms for matrix solvers and techniques related to singular value decomposition.

(I) Determinant
Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be two $p \times p$ square matrices and let $|\boldsymbol{A}|$ denote the *determinant* of $\boldsymbol{A}$. Basic properties include the following:

1. $|\boldsymbol{A}^\top| = |\boldsymbol{A}|$.
2. $|\alpha\boldsymbol{A}| = \alpha^p|\boldsymbol{A}|$, for any scalar $\alpha \in \mathbb{R}$.
3. $|\boldsymbol{AB}| = |\boldsymbol{A}| \cdot |\boldsymbol{B}|$. [Note that $|\boldsymbol{A} + \boldsymbol{B}| \neq |\boldsymbol{A}| + |\boldsymbol{B}|$ in general.]

(II) Inverse
Let $\boldsymbol{A}$ be a $p \times p$ *nonsingular* matrix; that is, its inverse $\boldsymbol{A}^{-1}$ exists, such that $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}_p$, where $\boldsymbol{I}_p$ is the $p \times p$ identity matrix. We have the following properties:

4. $(\boldsymbol{A}^{-1})^\top = (\boldsymbol{A}^\top)^{-1}$.
5. $(\boldsymbol{A}\boldsymbol{B})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$.
6. $|\boldsymbol{A}^{-1}| = 1/|\boldsymbol{A}|$.
7. $\boldsymbol{A}$ is an *orthogonal* matrix if and only if $\boldsymbol{A}^{-1} = \boldsymbol{A}^\top$.
8. If $\boldsymbol{D} = \text{diag}(d_1, \ldots, d_p)$ with $d_i \neq 0$ for $i = 1, \ldots, p$, then $\boldsymbol{D}^{-1} = \text{diag}(d_1^{-1}, \ldots, d_p^{-1})$.

(III) Trace
Given a $p \times p$ matrix $\boldsymbol{A} = (a_{ij})$, the trace of $\boldsymbol{A}$, denoted by $\text{tr}(\boldsymbol{A})$, is the sum of the diagonal entries of $\boldsymbol{A}$:

$$\text{tr}(\boldsymbol{A}) = \sum_{i=1}^{p} a_{ii}.$$

The trace of a matrix satisfies the following properties:

9. $\text{tr}(\boldsymbol{A}) = \text{tr}(\boldsymbol{A}^\top)$.
10. $\text{tr}(\alpha\boldsymbol{A} + \beta\boldsymbol{B}) = \alpha\text{tr}(\boldsymbol{A}) + \beta\text{tr}(\boldsymbol{B})$, for $\alpha, \beta \in \mathbb{R}$.
11. Cyclic invariance: $\text{tr}(\boldsymbol{A}\boldsymbol{B}\boldsymbol{C}) = \text{tr}(\boldsymbol{C}\boldsymbol{A}\boldsymbol{B}) = \text{tr}(\boldsymbol{B}\boldsymbol{C}\boldsymbol{A})$. In general, for any k numbers of $p \times p$ matrices $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_k$, we have:

$$\text{tr}(\boldsymbol{A}_1\boldsymbol{A}_2 \cdots \boldsymbol{A}_k) = \text{tr}(\boldsymbol{A}_i\boldsymbol{A}_{i+1} \cdots \boldsymbol{A}_k\boldsymbol{A}_1 \cdots \boldsymbol{A}_{i-1}), \text{ for any } i = 2, \ldots, k.$$

The derivative of the trace of a matrix with respect to itself satisfies the following properties [5]. Using the notation $\nabla_{\boldsymbol{A}} f := \left(\frac{\partial}{\partial a_{ij}} f\right)$ for the *gradient* of f with respect to $\boldsymbol{A}$ (we sometimes also use $\frac{\partial f}{\partial \boldsymbol{A}}$ instead), we have

12. $\nabla_{\boldsymbol{A}}\text{tr}(\boldsymbol{A}\boldsymbol{B}) = \boldsymbol{B}^\top$.
13. $\nabla_{\boldsymbol{A}}\text{tr}(\boldsymbol{A}\boldsymbol{B}\boldsymbol{A}^\top\boldsymbol{C}) = \boldsymbol{C}\boldsymbol{A}\boldsymbol{B} + \boldsymbol{C}^\top\boldsymbol{A}\boldsymbol{B}^\top$.
14. $\nabla_{\boldsymbol{A}^\top}\text{tr}(\boldsymbol{A}\boldsymbol{B}\boldsymbol{A}^\top\boldsymbol{C}) = \boldsymbol{B}^\top\boldsymbol{A}^\top\boldsymbol{C} + \boldsymbol{B}\boldsymbol{A}^\top\boldsymbol{C}$.

We prove only the first statement, while the second and third statements can follow directly from the chain rule. Recall that the components of $\boldsymbol{A}$ and $\boldsymbol{B}$ are denoted as (a_{ij}) and (b_{ij}), respectively. Since

$$\text{tr}(\boldsymbol{A}\boldsymbol{B}) = \sum_{l=1}^{p}\sum_{k=1}^{p} a_{lk}b_{kl},$$

we then have, for any $i, j = 1, \ldots, p$,

$$\frac{\partial \text{tr}(\boldsymbol{AB})}{\partial a_{ij}} = \frac{\partial}{\partial a_{ij}} \sum_j \sum_k a_{lk} b_{kl} = b_{ji},$$

which confirms the assertion of Property 12.

(IV) Positive definite matrix

A $p \times p$ symmetric matrix $\boldsymbol{A}$ is called *positive definite* (usually denoted by $\boldsymbol{A} \succ \boldsymbol{0}$) if for any nonzero $p \times 1$ vector $\boldsymbol{x}$, $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} > 0$; if $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} = 0$ for some nonzero $p \times 1$ vector $\boldsymbol{x}$, while $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} \geq 0$ still holds for any $\boldsymbol{x} \in \mathbb{R}^p$, we say that $\boldsymbol{A}$ is *positive semi-definite*, denoted as $\boldsymbol{A} \succeq \boldsymbol{0}$, where $\boldsymbol{0}$ denotes a zero vector.

15. If $\boldsymbol{A} \succ \boldsymbol{0}$, then $\boldsymbol{A}^{-1} \succ \boldsymbol{0}$.
16. For any $n \times p$ matrix $\boldsymbol{B}$, $\boldsymbol{B}^\top \boldsymbol{B} \succeq \boldsymbol{0}$; indeed, for any nonzero $p \times 1$ vector $\boldsymbol{x}$, let $\boldsymbol{y} = \boldsymbol{B}\boldsymbol{x}$, which is an $n \times 1$ vector, then $\boldsymbol{x}^\top \boldsymbol{B}^\top \boldsymbol{B}\boldsymbol{x} = \boldsymbol{y}^\top \boldsymbol{y} = \sum_{i=1}^n y_i^2 \geq 0$.
17. The sample covariance matrix is positive semi-definite. It suffices to show that the SSCP (Sum of Squares and Cross Products) matrix $\boldsymbol{S} \succeq \boldsymbol{0}$. Note that

$$\boldsymbol{S} = \sum_{i=1}^n (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^\top = (\boldsymbol{X} - \boldsymbol{1}_n \overline{\boldsymbol{x}}^\top)^\top (\boldsymbol{X} - \boldsymbol{1}_n \overline{\boldsymbol{x}}^\top) \succeq \boldsymbol{0},$$

by using Property 16, where $\boldsymbol{1}_n$ denotes an n-dimensional vector whose all entries are 1.

(V) Eigenvalues and eigenvectors

For any $p \times p$ matrix $\boldsymbol{A}$, the determinant $|\boldsymbol{A} - \lambda \boldsymbol{I}_p|$ is a polynomial in λ of degree p. The p roots of the polynomial equation $|\boldsymbol{A} - \lambda \boldsymbol{I}_p| = 0$, denoted by $\lambda_1, \ldots, \lambda_p$ (counting multiplicity), are called the *eigenvalues* (or *latent roots*) of $\boldsymbol{A}$. As a result, each $\boldsymbol{A} - \lambda_i \boldsymbol{I}_p$ is a singular matrix, hence there exists a nonzero vector $\boldsymbol{h}_i$ such that $(\boldsymbol{A} - \lambda_i \boldsymbol{I}_p)\boldsymbol{h}_i = 0$ (or equivalently $\boldsymbol{A}\boldsymbol{h}_i = \lambda_i \boldsymbol{h}_i$); $\boldsymbol{h}_i$ is called an *eigenvector* of $\boldsymbol{A}$ corresponding to λ_i. If $\boldsymbol{h}_i$ is taken to have a unit length, i.e. $\boldsymbol{h}_i^\top \boldsymbol{h}_i = 1$, then it is called a *normalized (unit) eigenvector* of $\boldsymbol{A}$.

Some important properties of eigenvalues and eigenvectors are as follows:

18. If $\boldsymbol{A}$ is a real symmetric matrix, then all its eigenvalues are real.
19. If $\boldsymbol{A} \succeq \boldsymbol{0}$, then all eigenvalues of $\boldsymbol{A}$ are non-negative.
20. The eigenvalues, counting multiplicity, of $\boldsymbol{A}$ and $\boldsymbol{A}^\top$, are the same.
21. If $\boldsymbol{A}$ and $\boldsymbol{B}$ are $p \times p$ nonsingular matrices, then the eigenvalues of $\boldsymbol{AB}$ and $\boldsymbol{BA}$ are equal.
22. If $\lambda_1, \ldots, \lambda_p$ are non-zero eigenvalues of a nonsingular matrix $\boldsymbol{A}$, then $\lambda_1^{-1}, \ldots, \lambda_p^{-1}$ are the eigenvalues of $\boldsymbol{A}^{-1}$.
23. If $\boldsymbol{A}$ is a real symmetric matrix, and λ_i and λ_j are two distinct eigenvalues of $\boldsymbol{A}$, then the corresponding eigenvectors $\boldsymbol{h}_i$ and $\boldsymbol{h}_j$ are orthogonal; indeed, by definition, $\boldsymbol{A}\boldsymbol{h}_i = \lambda_i \boldsymbol{h}_i$ and $\boldsymbol{A}\boldsymbol{h}_j = \lambda_j \boldsymbol{h}_j$, and therefore $\boldsymbol{h}_j^\top \boldsymbol{A}\boldsymbol{h}_i = \lambda_i \boldsymbol{h}_j^\top \boldsymbol{h}_i$ and $\boldsymbol{h}_i^\top \boldsymbol{A}\boldsymbol{h}_j = \lambda_j \boldsymbol{h}_i^\top \boldsymbol{h}_j$. By the symmetry of $\boldsymbol{A}$, $\boldsymbol{h}_j^\top \boldsymbol{A}\boldsymbol{h}_i = \boldsymbol{h}_i^\top \boldsymbol{A}\boldsymbol{h}_j$, hence $(\lambda_i - \lambda_j)\boldsymbol{h}_j^\top \boldsymbol{h}_i = 0$. As $\lambda_i \neq \lambda_j$, we must have $\boldsymbol{h}_j^\top \boldsymbol{h}_i = 0$.

24. If $\boldsymbol{A}$ is a $p \times p$ real symmetric matrix, then it has p orthogonal eigenvectors. Let $\boldsymbol{H}$ be a $p \times p$ matrix whose columns are normalized eigenvectors $\boldsymbol{h}_1, \ldots, \boldsymbol{h}_p$ of $\boldsymbol{A}$, i.e., $\boldsymbol{H} = (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_p)$, then $\boldsymbol{H}$ is orthogonal, and

$$\boldsymbol{H}^\top \boldsymbol{A}\boldsymbol{H} = \boldsymbol{D} = \text{diag}(\lambda_1, \ldots, \lambda_p) \quad (\textit{Diagonalization} \text{ of } \boldsymbol{A}),$$

or equivalently,

$$\boldsymbol{A} = \boldsymbol{H}\boldsymbol{D}\boldsymbol{H}^\top (\textit{Spectral decomposition} \text{ of } \boldsymbol{A}).$$

To see this, one can simply write $\boldsymbol{A}\boldsymbol{H} = (\boldsymbol{A}\boldsymbol{h}_1, \ldots, \boldsymbol{A}\boldsymbol{h}_p) = (\lambda_1 \boldsymbol{h}_1, \ldots, \lambda_p \boldsymbol{h}_p) = \boldsymbol{H}\boldsymbol{D}$, therefore $\boldsymbol{H}^\top \boldsymbol{A}\boldsymbol{H} = \boldsymbol{D}$.

25. The trace of a matrix $\boldsymbol{A}$ is the sum of all its eigenvalues and the determinant of $\boldsymbol{A}$ is the product of all its eigenvalues, i.e., $\text{tr}(\boldsymbol{A}) = \sum_{i=1}^{p} \lambda_i$ and $|\boldsymbol{A}| = \prod_{i=1}^{p} \lambda_i$, respectively.

(VI) Symmetric square root of a positive semi-definite matrix

Let $\boldsymbol{A} \succeq \boldsymbol{0}$ be a $p \times p$ real symmetric matrix with non-negative eigenvalues $\lambda_1, \ldots, \lambda_p$ and their corresponding eigenvectors $\boldsymbol{h}_1, \ldots, \boldsymbol{h}_p$. As above, define the matrix $\boldsymbol{H} = (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_p), \boldsymbol{D} = \text{diag}(\lambda_1, \ldots, \lambda_p)$ and $\boldsymbol{D}^{1/2} := \text{diag}\big(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_p}\big)$. Recall that $\boldsymbol{A} = \boldsymbol{H}\boldsymbol{D}\boldsymbol{H}^\top$. The *symmetric square root* of $\boldsymbol{A}$ is given by $\boldsymbol{A}^{1/2} := \boldsymbol{H}\boldsymbol{D}^{1/2}\boldsymbol{H}^\top$; indeed, a simple calculation gives $\boldsymbol{A}^{1/2}\boldsymbol{A}^{1/2} = \boldsymbol{H}\boldsymbol{D}^{1/2}\boldsymbol{H}^\top \boldsymbol{H}\boldsymbol{D}^{1/2}\boldsymbol{H}^\top = \boldsymbol{H}\boldsymbol{D}\boldsymbol{H}^\top = \boldsymbol{A}$.

All the above matrix concepts can be implemented in both **R** and Python by using built-in matrix functions. Let us continue with the blue chips dataset in Example 1.1. The Python code `S_2002.T` returns the transpose of $\boldsymbol{S}$ from the 2002 dataset, `@` computes matrix multiplication, and so on. All the codings for other related computations can be found in Programme 1.3 for Python (resp. Programme 1.4 for **R**).

```
print(np.linalg.det(np.linalg.inv(S_2002))) # |A^{-1}| = 1/|A|
print(1/np.linalg.det(S_2002))

eig_val_2002, H_2002 = np.linalg.eig(S_2002)
print(eig_val_2002)
print(H_2002)

print(np.round(H_2002.T @ H_2002, 3)) # H^T H = I
print(H_2002[:,0].T @ H_2002[:,1]) # h_1^T h_2 = 0
print(np.round(H_2002.T @ S_2002 @ H_2002, 3)) # H^T A H = D
D_2002 = np.diag(eig_val_2002)
print(D_2002)

sqrt_S_2002 = H_2002 @ np.sqrt(D_2002) @ H_2002.T # A^{1/2} = H D^{1/2} H^T
print(sqrt_S_2002 @ sqrt_S_2002)
print(S_2002)
```

```
5.706216005035965e-07
5.706216005035965e-07
[68.4874 19.9176  3.3412  2.2574 13.2054 12.8986]
[[-0.0311 -0.9118  0.1822  0.0364 -0.3644  0.0161]
 [-0.0295 -0.1914 -0.819  -0.5398  0.0184  0.0059]
 [ 0.0109 -0.091  -0.5321  0.8403  0.0438 -0.0201]
 [ 0.003  -0.3462  0.1122 -0.0186  0.912  -0.1884]
 [-0.9984  0.0338  0.0126  0.0233  0.0076 -0.0365]
 [-0.0357 -0.0509  0.013   0.0172  0.1822  0.9811]]
[[ 1.  0.  0.  0. -0.  0.]
 [ 0.  1.  0. -0. -0.  0.]
 [ 0.  0.  1. -0.  0.  0.]
 [ 0. -0. -0.  1.  0. -0.]
 [-0. -0.  0.  0.  1. -0.]
 [ 0.  0.  0. -0. -0.  1.]]
5.637851296924623e-17
[[68.487  0.     0.     0.    -0.     0.   ]
 [ 0.    19.918  0.     0.    -0.     0.   ]
 [ 0.     0.     3.341 -0.     0.     0.   ]
 [ 0.     0.    -0.     2.257 -0.    -0.   ]
 [-0.    -0.     0.    -0.    13.205 -0.   ]
 [ 0.     0.     0.    -0.    -0.    12.899]]
[[68.4874  0.      0.      0.      0.      0.    ]
 [ 0.     19.9176  0.      0.      0.      0.    ]
 [ 0.      0.      3.3412  0.      0.      0.    ]
 [ 0.      0.      0.      2.2574  0.      0.    ]
 [ 0.      0.      0.      0.     13.2054  0.    ]
 [ 0.      0.      0.      0.      0.     12.8986]]
[[18.4979  2.909   1.1602  1.9204  1.4781  0.338 ]
 [ 2.909   3.6931  0.7663  1.2371  1.8228  0.3288]
 [ 1.1602  0.7663  2.7439  0.9721 -0.7734 -0.0741]
 [ 1.9204  1.2371  0.9721 13.8716 -0.2575  0.1582]
 [ 1.4781  1.8228 -0.7734 -0.2575 68.3082  1.9618]
 [ 0.338   0.3288 -0.0741  0.1582  1.9618 12.9929]]
[[18.4979  2.909   1.1602  1.9204  1.4781  0.338 ]
 [ 2.909   3.6931  0.7663  1.2371  1.8228  0.3288]
 [ 1.1602  0.7663  2.7439  0.9721 -0.7734 -0.0741]
 [ 1.9204  1.2371  0.9721 13.8716 -0.2575  0.1582]
 [ 1.4781  1.8228 -0.7734 -0.2575 68.3082  1.9618]
 [ 0.338   0.3288 -0.0741  0.1582  1.9618 12.9929]]
```

Programme 1.3 Illustrative examples of Matrix Theory in Python using the dataset `fin-ratio.csv`.

```
> det(solve(S_2002)) # |A^-1| = 1/|A|
[1] 5.706e-07
> 1 / det(S_2002)
[1] 5.706e-07
> eig_2002 <- eigen(S_2002)
> eig_2002$values
[1] 68.487 19.918 13.205 12.899  3.341  2.257
> (H_2002 <- eig_2002$vector)
           [,1]      [,2]       [,3]       [,4]      [,5]      [,6]
[1,]  0.031107  0.91185  0.364402  0.016136  0.18218 -0.03637
[2,]  0.029477  0.19142 -0.018447  0.005915 -0.81898  0.53980
[3,] -0.010938  0.09104 -0.043829 -0.020125 -0.53213 -0.84030
[4,] -0.003038  0.34621 -0.911976 -0.188393  0.11223  0.01856
[5,]  0.998380 -0.03383 -0.007556 -0.036516  0.01255 -0.02334
[6,]  0.035663  0.05094 -0.182190  0.981058  0.01304 -0.01720
> round(t(H_2002)%*%H_2002, 3) # H^T H = I
     [,1] [,2] [,3] [,4] [,5] [,6]
[1,]    1    0    0    0    0    0
[2,]    0    1    0    0    0    0
[3,]    0    0    1    0    0    0
[4,]    0    0    0    1    0    0
[5,]    0    0    0    0    1    0
[6,]    0    0    0    0    0    1
> t(H_2002[,1])%*%H_2002[,2] # h_1^T h_2 = 0
[,1]
[1,] -7.589e-17
> round(t(H_2002)%*%S_2002%*%H_2002, 3) # H^T A H = D
      [,1]  [,2]  [,3] [,4]  [,5]  [,6]
[1,] 68.49  0.00  0.00  0.0 0.000 0.000
[2,]  0.00 19.92  0.00  0.0 0.000 0.000
[3,]  0.00  0.00 13.21  0.0 0.000 0.000
[4,]  0.00  0.00  0.00 12.9 0.000 0.000
[5,]  0.00  0.00  0.00  0.0 3.341 0.000
[6,]  0.00  0.00  0.00  0.0 0.000 2.257
> (D_2002 <- diag(eig_2002$values))
      [,1]  [,2]  [,3] [,4]  [,5]  [,6]
[1,] 68.49  0.00  0.00  0.0 0.000 0.000
[2,]  0.00 19.92  0.00  0.0 0.000 0.000
[3,]  0.00  0.00 13.21  0.0 0.000 0.000
[4,]  0.00  0.00  0.00 12.9 0.000 0.000
[5,]  0.00  0.00  0.00  0.0 3.341 0.000
[6,]  0.00  0.00  0.00  0.0 0.000 2.257
> sqrt_S_2002 <- H_2002%*%sqrt(D_2002)%*%t(H_2002) # A^1/2 = H D^1/2 H^T
> sqrt_S_2002%*%sqrt_S_2002
       [,1]   [,2]     [,3]    [,4]    [,5]     [,6]
[1,] 18.498 2.9090  1.16019  1.9204  1.4781  0.33795
[2,]  2.909 3.6931  0.76630  1.2371  1.8228  0.32879
[3,]  1.160 0.7663  2.74394  0.9721 -0.7734 -0.07413
[4,]  1.920 1.2371  0.97207 13.8716 -0.2575  0.15815
```

```
[5,]  1.478 1.8228 -0.77342 -0.2575 68.3082  1.96177
[6,]  0.338 0.3288 -0.07413  0.1582  1.9618 12.99291
> S_2002
          EY   CFTP    ln_MV      DY    BTME      DTE
EY    18.498 2.9090  1.16019  1.9204  1.4781  0.33795
CFTP   2.909 3.6931  0.76630  1.2371  1.8228  0.32879
ln_MV  1.160 0.7663  2.74394  0.9721 -0.7734 -0.07413
DY     1.920 1.2371  0.97207 13.8716 -0.2575  0.15815
BTME   1.478 1.8228 -0.77342 -0.2575 68.3082  1.96177
DTE    0.338 0.3288 -0.07413  0.1582  1.9618 12.99291
```

Programme 1.4 Illustrative examples of Matrix Theory in **R** using the dataset `fin-ratio.csv`.

(VII) Algorithm for Cholesky Decomposition

The *Cholesky decomposition* of a $p \times p$ positive definite matrix $\boldsymbol{A}$ is to find a $p \times p$ upper triangular matrix $\mathbf{C}$ such that $\mathbf{C}^{\top}\mathbf{C} = \boldsymbol{A}$. Let $\boldsymbol{A} = (a_{ij})$ and $\mathbf{C} = (c_{ij})$ where $c_{ij} = 0$ whenever $i > j$. The following steps describe how to find $\mathbf{C}$ from a given matrix $\boldsymbol{A}$:

Step 1. Set $c_{11} = \sqrt{a_{11}}$, and $c_{1j} = a_{1j}/c_{11}$ for $j = 2, \ldots, p$.
Step 2. For $i = 2, \ldots, p$, set

$$c_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} c_{ki}^2}, \text{ and } c_{ij} = \frac{1}{c_{ii}}\left(a_{ij} - \sum_{k=1}^{i-1} c_{ki}c_{kj}\right), \text{ for } j = i+1, \ldots, p.$$

To see why this algorithm can give a valid $\mathbf{C}$, note that the (i, j) element of $\mathbf{C}^{\top}\mathbf{C} = \sum_{k=1}^{p} c_{ki}c_{kj} = a_{ij}$. When $i = 1$ and $j = 1$, $c_{11}^2 = a_{11}$, and so we simply take $c_{11} = \sqrt{a_{11}}$. When $i = 1$ and $j = 2, \ldots, p$, $a_{1j} = \sum_{k=1}^{p} c_{k1}c_{kj} = c_{11}c_{1j}$ since $c_{k1} = 0$ for $k = 2, \ldots, p$, therefore $c_{1j} = a_{1j}/c_{11}$. For $i = 2, \ldots, p$, $a_{ii} = \sum_{k=1}^{p} c_{ki}^2 = \sum_{k=1}^{i} c_{ki}^2 = \sum_{k=1}^{i-1} c_{ki}^2 + c_{ii}^2$ since $c_{ij} = 0$ for $i > j$, we then take

$$c_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} c_{ki}^2}.$$

Finally, for $j = i+1, \ldots, p$,

$$a_{ij} = \sum_{k=1}^{i} c_{ki}c_{kj} = c_{ii}c_{ij} + \sum_{k=1}^{i-1} c_{ki}c_{kj} \Rightarrow c_{ij} = \frac{1}{c_{ii}}\left(a_{ij} - \sum_{k=1}^{i-1} c_{ki}c_{kj}\right).$$

1.3 VECTORS AND MATRIX NORMS

The notion of a norm is to measure the length of different abstract objects. Here we recall some useful results and commonly used norms for vectors and matrices.

(I) Vector norm
We consider the space $\mathbb{R}^p$, such that each $\boldsymbol{x} \in \mathbb{R}^p$ is a p-dimensional vector. A function $\phi : \mathbb{R}^p \to \mathbb{R}$ is a **norm** for the Euclidean space $\mathbb{R}^p$ if it satisfies:

1. positivity: $\phi(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^p$, and $\phi(\boldsymbol{x}) = 0 \iff \boldsymbol{x} = \mathbf{0}$;
2. triangle inequality: for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$, $\phi(\boldsymbol{x} + \boldsymbol{y}) \leq \phi(\boldsymbol{x}) + \phi(\boldsymbol{y})$;
3. absolute homogeneity: for any $\boldsymbol{x} \in \mathbb{R}^p$ and scalar $a \in \mathbb{R}$, $\phi(a\boldsymbol{x}) = |a|\phi(\boldsymbol{x})$.

One of the most commonly encountered norms is the q-norm: for $\boldsymbol{x} = (x_1, \ldots, x_p)^\top \in \mathbb{R}^p$ and $q \geq 1$, the q-norm is defined by:

$$\|\boldsymbol{x}\|_q := \left(\sum_{i=1}^{p} |x_i|^q \right)^{\frac{1}{q}}.$$

The case when $q = \infty$ is defined in a different way; indeed, the ∞-norm, properly called sup-norm, is defined as

$$\|\boldsymbol{x}\|_\infty := \max_{1 \leq i \leq p} |x_i|.$$

When $q = 0$, we define the "0-norm" by the number of non-zero entries in x. The 0-norm is not a norm as it fails to satisfy the homogeneity condition.

(II) Matrix norm
The concept of a norm can be extended to the space $\mathbb{R}^{p \times k}$ of $p \times k$ matrices. A function $\phi : \mathbb{R}^{p \times k} \to \mathbb{R}$ is a **norm** for the space $\mathbb{R}^{p \times k}$ if it satisfies

1. positivity: $\phi(\boldsymbol{A}) \geq 0$ for all $\boldsymbol{A} \in \mathbb{R}^{p \times k}$, and $\phi(\boldsymbol{A}) = 0 \iff \boldsymbol{A} = \mathbf{0}$;
2. triangle inequality: for any $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{p \times k}$, $\phi(\boldsymbol{A} + \boldsymbol{B}) \leq \phi(\boldsymbol{A}) + \phi(\boldsymbol{B})$;
3. homogeneity: for any $\boldsymbol{A} \in \mathbb{R}^{p \times k}$ and scalar $a \in \mathbb{R}$, $\phi(a\boldsymbol{A}) = |a|\phi(\boldsymbol{A})$.

In addition, if $p = k$ and ϕ satisfies

4. sub-multiplicativity: $\phi(\boldsymbol{AB}) \leq \phi(\boldsymbol{A})\phi(\boldsymbol{B})$ for any $\boldsymbol{A} \in \mathbb{R}^{p \times k}$,

then ϕ is said to be a **sub-multiplicative norm**.

There are several common examples of matrix norms. In particular, the q-norm for square matrices is commonly used. A matrix norm $\|\cdot\|$ on $\mathbb{R}^{p \times p}$ is said to be *compatible* with a vector norm $\|\cdot\|_v$ on $\mathbb{R}^p$ if for any $\boldsymbol{A} \in \mathbb{R}^{p \times p}$ and $\boldsymbol{x} \in \mathbb{R}^p$, it holds that

$$\|\boldsymbol{Ax}\|_v \leq \|\boldsymbol{A}\| \cdot \|\boldsymbol{x}\|_v.$$

Let $\|\cdot\|_q$ be the vector q-norm on $\mathbb{R}^p$. Then the q-norm $\|\cdot\|$ on the space of $p \times p$ square matrices is defined as

$$\|\boldsymbol{A}\| = \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_q}{\|\boldsymbol{x}\|_q} = \max_{\|\boldsymbol{x}\|_q=1} \|\boldsymbol{A}\boldsymbol{x}\|_q, \quad \boldsymbol{A} \in \mathbb{R}^{p\times p}. \tag{1.3}$$

It is clear that this definition verifies the properties of a matrix norm. In addition, it is obvious that the q-norms for square matrices and vectors are compatible. The definition can easily be generalized to $p \times k$ matrices, by considering the norm $\|\boldsymbol{A}\boldsymbol{x}\|_q$ in (1.3) as the q-norm of $\boldsymbol{A}\boldsymbol{x}$ in $\mathbb{R}^p$ for any $\boldsymbol{x} \in \mathbb{R}^k$ which possesses the corresponding q-norm in $\mathbb{R}^k$. In the following, without the risk of causing any ambiguity, we shall denote the q-norms for both vectors and matrices by $\|\cdot\|_q$. Below are several important special cases of the q-norm for $p \times k$ matrices:

1. 1-norm:
$$\|A\|_1 = \max_{j=1,\ldots,k} \left(\sum_{i=1}^{p} |a_{ij}| \right).$$
2. 2-norm (a.k.a. the *spectral norm*):
$$\|A\|_2 = \sqrt{\lambda_{\max}(\boldsymbol{A}^\top \boldsymbol{A})},$$
where $\lambda_{\max}(\boldsymbol{A}^\top \boldsymbol{A})$ is the maximum eigenvalue of the matrix $\boldsymbol{A}^\top \boldsymbol{A}$.
3. ∞-norm/sup-norm:
$$\|A\|_\infty = \max_{i=1,\ldots,p} \left(\sum_{j=1}^{k} |a_{ij}| \right).$$

Apart from the q-norm, another commonly encountered matrix norm in machine learning and deep learning is the *Frobenius norm*, which is defined as

$$\|A\|_F := \left(\sum_{i=1}^{p} \sum_{j=1}^{k} |a_{ij}|^2 \right)^{\frac{1}{2}} = \sqrt{\mathrm{tr}(\boldsymbol{A}^\top \boldsymbol{A})}.$$

1.4 COMMON PROBABILITY DISTRIBUTIONS

In this section, we shall review some basic probability distributions learnt in elementary statistical courses. Readers may refer to [10, 13] for more a comprehensive treatment of this topic, and [1, 14, 17] for a more advanced discussion, and on tackling block matrices in particular. These distributions are important and serve as the building blocks for constructing and deriving the null distributions of many test statistics. Throughout the book, "iid" stands for "*independent and identically distributed*", referring to an underlying family of random variables.

1.4.1 Univariate Distributions

To begin with, we review the definitions and related important properties of several commonly used univariate distributions.

1. **Normal distribution:** $\mathcal{N}(\mu, \sigma^2)$. A real random variable x is said to follow a normal distribution with a mean μ and variance σ^2, denoted as $\mathrm{x} \sim \mathcal{N}(\mu, \sigma^2)$, if its *probability density function* (pdf) takes the form:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}, \quad x \in \mathbb{R}. \tag{1.4}$$

2. **Chi-square distribution:** χ^2_ν. A non-negative random variable x is said to follow a chi-square distribution with degrees of freedom $\nu \in \mathbb{N}$, denoted by $\mathrm{x} \sim \chi^2_\nu$, if its pdf takes the form:

$$f(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2}, \quad x > 0, \tag{1.5}$$

where $\Gamma(\cdot)$ is the well-known gamma function:

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\, dt, \quad \text{for} \quad z > 0.$$

If $\mathrm{z}_1, \ldots, \mathrm{z}_n$ are iid $\mathcal{N}(0,1)$ random variables, then one can easily show that $\mathrm{x} = \mathrm{z}_1^2 + \cdots + \mathrm{z}_n^2 \sim \chi^2_n$. Furthermore, if $\mathrm{y}_1, \ldots, \mathrm{y}_n$ are iid $\mathcal{N}(\mu, \sigma^2)$, we also have the following distribution of the sample variance s^2_y:

$$\frac{(n-1)\mathrm{s}^2_\mathrm{y}}{\sigma^2} = \sum_{i=1}^{n} \frac{(\mathrm{y}_i - \overline{\mathrm{y}})^2}{\sigma^2} \sim \chi^2_{n-1};$$

this result is just a special case of Proposition 1.2, which we will meet shortly.

3. **Student's *t*-distribution:** t_ν. If $\mathrm{z} \sim \mathcal{N}(0,1)$ is independent of $\mathrm{x} \sim \chi^2_\nu$, then $\mathrm{t} := \frac{\mathrm{z}}{\sqrt{\mathrm{x}/\nu}}$ is said to follow a *t*-distribution t_ν with ν degrees of freedom. As a result, if $\mathrm{y}_1, \ldots, \mathrm{y}_n$ are iid $\mathcal{N}(\mu, \sigma^2)$, then

$$\frac{\sqrt{n}(\overline{\mathrm{y}} - \mu)}{\mathrm{s}_\mathrm{y}} \sim t_{n-1};$$

indeed, since $\overline{\mathrm{y}} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$, we have $\mathrm{z} := \frac{\sqrt{n}(\overline{\mathrm{y}}-\mu)}{\sigma} \sim \mathcal{N}(0,1)$, and $\mathrm{x} := \frac{(n-1)\mathrm{s}^2_\mathrm{y}}{\sigma^2} \sim \chi^2_{n-1}$, furthermore $\overline{\mathrm{y}}$ is independent of s^2_y in accordance with Proposition 1.2. Therefore,

$$\mathrm{t} = \frac{\mathrm{z}}{\sqrt{\mathrm{x}/(n-1)}} = \frac{\sqrt{n}(\overline{\mathrm{y}} - \mu)}{\mathrm{s}_\mathrm{y}} \sim t_{n-1}.$$

4. **Fisher's *F*-distribution:** $F_{u,v}$. If $\mathrm{x} \sim \chi^2_u$ and $\mathrm{y} \sim \chi^2_v$ are independent, then $\frac{\mathrm{x}/u}{\mathrm{y}/v} \sim F_{u,v}$, an *F*-distribution with parameters u and v in order. Furthermore, if $\mathrm{x}_1, \ldots, \mathrm{x}_m$ are iid $\mathcal{N}(\mu_1, \sigma_1^2)$ and are independent of $\mathrm{y}_1, \ldots, \mathrm{y}_n$ being iid $\mathcal{N}(\mu_2, \sigma_2^2)$, then $\frac{\mathrm{s}_\mathrm{x}^2/\sigma_1^2}{\mathrm{s}_\mathrm{y}^2/\sigma_2^2} = \frac{\mathrm{s}_\mathrm{x}^2\sigma_2^2}{\sigma_1^2\mathrm{s}_\mathrm{y}^2} \sim F_{m-1,n-1}$; indeed, since $\frac{(m-1)\mathrm{s}_\mathrm{x}^2}{\sigma_1^2} \sim \chi^2_{m-1}$ and $\frac{(n-1)\mathrm{s}_\mathrm{y}^2}{\sigma_2^2} \sim \chi^2_{n-1}$ are independent,

$$\frac{\mathrm{s}_\mathrm{x}^2/\sigma_1^2}{\mathrm{s}_\mathrm{y}^2/\sigma_2^2} = \frac{\mathrm{s}_\mathrm{x}^2\sigma_2^2}{\sigma_1^2\mathrm{s}_\mathrm{y}^2} \sim F_{m-1,n-1}.$$

   In particular, if $\mathrm{t} \sim t_v$, then $\mathrm{t}^2 \sim F_{1,v}$.
5. **Sampling distributions of $\overline{\mathrm{x}}$ and s_x^2 (a special case of Proposition 1.2):** if $\mathrm{x}_1, \ldots, \mathrm{x}_n$ are iid $\mathcal{N}(\mu, \sigma^2)$, then $\overline{\mathrm{x}} \sim \mathcal{N}(\mu, \sigma^2/n)$ and is independent of $(n-1)\mathrm{s}_\mathrm{x}^2/\sigma^2 \sim \chi^2_{n-1}$.

1.4.2 The Multivariate Normal Distribution

The random vector $\mathbf{x} = (\mathrm{x}_1, \ldots, \mathrm{x}_p)^\top$ is said to follow a p-variate normal distribution with a mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$,[2] denoted by $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if its joint probability density function is given by

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}, \quad \boldsymbol{x} \in \mathbb{R}^p. \tag{1.6}$$

The following are some important properties of the multivariate normal distribution:

1. Consider $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then for a deterministic $\boldsymbol{B} \in \mathbb{R}^{q\times p}$ and $\boldsymbol{b} \in \mathbb{R}^q$, we have $\mathbf{y} = \boldsymbol{B}\mathbf{x} + \boldsymbol{b} \sim \mathcal{N}_q(\boldsymbol{B}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}^\top)$.
2. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \overset{\text{iid}}{\sim} \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then $\overline{\mathbf{x}} \sim \mathcal{N}_p(\boldsymbol{\mu}, (1/n)\boldsymbol{\Sigma})$.
3. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \overset{\text{iid}}{\sim} \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then $n(\overline{\mathbf{x}} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\overline{\mathbf{x}} - \boldsymbol{\mu}) \sim \chi^2_p$. To see this, let $\mathbf{y} = \sqrt{n}\boldsymbol{\Sigma}^{-1/2}(\overline{\mathbf{x}} - \boldsymbol{\mu})$. By Property 1, $\mathbf{y} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{I}_p)$. Therefore, $n(\overline{\mathbf{x}} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\overline{\mathbf{x}} - \boldsymbol{\mu}) = \mathbf{y}^\top\mathbf{y} = \sum_{i=1}^p \mathrm{y}_i^2 \sim \chi^2_p$ in light of Property 2 in Subsection 1.4.1.
4. (The *Central Limit Theorem*). Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be p-dimensional iid random vectors from an arbitrary distribution with a finite mean $\boldsymbol{\mu}$ and finite covariance matrix $\boldsymbol{\Sigma}$. Then $\sqrt{n}(\overline{\mathbf{x}} - \boldsymbol{\mu}) \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$ approximately for a large enough sample size n (relative to p).

   Consequently, by using Slutsky's theorem [15], we further have $n(\overline{\mathbf{x}} - \boldsymbol{\mu})^\top \mathbf{S}^{-1}(\overline{\mathbf{x}} - \boldsymbol{\mu}) \sim \chi^2_p$ approximately for a large sample size n (relative to p).

[2] Here, we only consider the nondegenerate case where $\boldsymbol{\Sigma}$ is positive definite.

5. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be iid $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ random vectors. Then the squared generalized distance (a.k.a. the *Mahalanobis distance*) $\mathrm{d}_i^2 = (\mathbf{x}_i - \overline{\mathbf{x}})^\top \mathrm{S}^{-1}(\mathbf{x}_i - \overline{\mathbf{x}}) \sim \chi_p^2$ approximately for a large value of n (relative to p). This is a direct consequence of an application of *Slutsky's theorem* by replacing $\boldsymbol{\mu}$ by $\overline{\boldsymbol{x}}$ and $\boldsymbol{\Sigma}$ by $\boldsymbol{S}$ in Property 4 above.

In simple linear regression, we often use the marginal and conditional distributions of the bivariate normal distribution; the corresponding results are readily extended to the multivariate case for their application in the context of multiple linear regression.

Let $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Partition $\mathbf{x}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ into:

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \hdashline \mathbf{x}_2 \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \hdashline \boldsymbol{\mu}_2 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \left(\begin{array}{c:c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \hdashline \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array}\right),$$

where $\mathbf{x}_1$, $\boldsymbol{\mu}_1$ are $q \times 1$ vectors; $\mathbf{x}_2$, $\boldsymbol{\mu}_2$ are $(p-q) \times 1$ vectors, and $\boldsymbol{\Sigma}_{11}, \boldsymbol{\Sigma}_{12}, \boldsymbol{\Sigma}_{21}$, and $\boldsymbol{\Sigma}_{22}$ are respectively $q \times q$, $q \times (p-q)$, $(p-q) \times q$, and $(p-q) \times (p-q)$ matrices, for some $q < p$. Note that $\boldsymbol{\Sigma}_{21} = \boldsymbol{\Sigma}_{12}^\top$. We summarize below some useful properties about the marginal and conditional distributions of the multivariate normal distribution.

1. If $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the marginal distribution of any q-component subset of $\mathbf{x}$ is q-variate normal. Without loss of generality, it suffices to show that $\mathbf{x}_1 \sim \mathcal{N}_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$. Let $\boldsymbol{B} = \left[\boldsymbol{I}_q \vdots \mathbf{0}\right]$ be a $q \times p$ matrix. By Property 1 in Subsection 1.4.2, $\mathbf{x}_1 = \boldsymbol{B}\mathbf{x} \sim \mathcal{N}_q(\boldsymbol{B}\boldsymbol{\mu}, \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}^\top) = \mathcal{N}_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$.
2. $\mathbf{x}_1$ and $\mathbf{x}_2$ are independent if and only if $\boldsymbol{\Sigma}_{12} = \mathbf{0}$.
3. $\mathbf{x}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{x}_2 \sim \mathcal{N}_q(\boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{11\cdot 2})$ and is independent of $\mathbf{x}_2$, where $\boldsymbol{\Sigma}_{11\cdot 2} := \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. To show this, we first let $C = \begin{pmatrix} \boldsymbol{I}_q & -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ \mathbf{0} & \boldsymbol{I}_{p-q} \end{pmatrix}$. Then $C\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{x}_2 \\ \mathbf{x}_2 \end{pmatrix}$ is p-variate normal with the mean $C\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\mu}_2 \\ \boldsymbol{\mu}_2 \end{pmatrix}$ and covariance matrix

$$\begin{aligned} C\boldsymbol{\Sigma}C^\top &= \begin{pmatrix} \boldsymbol{I}_q & -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ \mathbf{0} & \boldsymbol{I}_{p-q} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{I}_q & \mathbf{0} \\ -\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} & \boldsymbol{I}_{p-q} \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{\Sigma}_{11\cdot 2} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22} \end{pmatrix}. \end{aligned}$$

4. The conditional distribution of $\mathbf{x}_1$ given $\mathbf{x}_2 = \boldsymbol{x}_2$ is $\mathcal{N}_q(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11\cdot 2})$. This result follows immediately from Property 3 of this subsection by setting $\mathbf{x}_2 = \boldsymbol{x}_2$.

1.4.3 Wishart Distribution

The *Wishart distribution* is a multivariate extension of the chi-square distribution. Recall Property 2 of Subsection 1.4.1, namely if $\mathrm{x}_1, \ldots, \mathrm{x}_n$ are iid distributed as $\mathcal{N}(\mu, \sigma^2)$, then

$$\sum_{i=1}^{n} \frac{(\mathrm{x}_i - \overline{\mathrm{x}})^2}{\sigma^2} = \frac{(n-1)\mathrm{s}^2}{\sigma^2} \sim \chi^2_{n-1}.$$

We now consider iid multivariate normal random vectors $\mathbf{z}_i \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$, $i = 1, \ldots, n$, with $n > p$. Then, the $p \times p$ random matrix

$$\mathbf{A} := \sum_{i=1}^{n} \mathbf{z}_i \mathbf{z}_i^\top$$

is said to follow the Wishart distribution with degrees of freedom n and scaling matrix $\boldsymbol{\Sigma}$, denoted as $W_p(\boldsymbol{\Sigma}, n)$. The density function of $\mathbf{A}$ is given by:

$$w(\boldsymbol{a} \mid \boldsymbol{\Sigma}, n) = \frac{|\boldsymbol{a}|^{(n-p-1)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{a})\right\}}{2^{pn/2}|\boldsymbol{\Sigma}|^{n/2}\Gamma_p\left(\frac{n}{2}\right)}, \quad \boldsymbol{a} \in \mathbb{R}^{p \times p}, \tag{1.7}$$

where Γ_p is the multivariate gamma function:

$$\Gamma_p(t) = \pi^{\frac{p(p-1)}{4}} \prod_{i=1}^{p} \Gamma\left(t - \frac{1}{2}(i-1)\right). \tag{1.8}$$

We remark that if $n < p$, $\mathbf{A}$ does not admit a density; nevertheless, in that case, we still denote its distribution by $W_p(\boldsymbol{\Sigma}, n)$; see more detailed discussion in [7].

Let $\mathbf{x}_i \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $i = 1, \ldots, n$, be iid multivariate normal random vectors. The following important result gives the sampling distribution of the sample covariance matrix $\mathbf{S}$.

Proposition 1.2 The sample mean $\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$ is independent of the sample covariance matrix $\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^\top$, and $\mathbf{S} \sim W_p\left(\frac{1}{n-1}\boldsymbol{\Sigma}, n-1\right)$.

Proof. Consider the SSCP matrix $\mathbf{A} = (n-1)\mathbf{S} = \sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^\top = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top - n\overline{\mathbf{x}}\,\overline{\mathbf{x}}^\top$. Let $\boldsymbol{B} = (b_{ij})$ be an $n \times n$ orthogonal matrix with the last row being

$$b_{nj} = \frac{1}{\sqrt{n}}, \quad j = 1, \ldots, n.$$

Also let $\mathbf{z}_i := \sum_{j=1}^n b_{ij}\mathbf{x}_j$, $i = 1, \ldots, n$. The orthogonality of $\boldsymbol{B}$ gives $\sum_{i=1}^n b_{ij}b_{ik} = \delta_{jk}$, where the *Dirac delta* δ_{jk} equals 1 if $j = k$, or 0 otherwise. Then, we have

$$\begin{aligned}\sum_{i=1}^n \mathbf{z}_i\mathbf{z}_i^\top &= \sum_{i=1}^n \left(\sum_{j=1}^n b_{ij}\mathbf{x}_j\right)\left(\sum_{k=1}^n b_{ik}\mathbf{x}_k\right)^\top \\ &= \sum_{j=1}^n \sum_{k=1}^n \left(\sum_{i=1}^n b_{ij}b_{ik}\right)\mathbf{x}_j\mathbf{x}_k^\top \\ &= \sum_{j=1}^n \mathbf{x}_j\mathbf{x}_j^\top.\end{aligned}$$

By definition, as $\mathbf{z}_n = \sum_{j=1}^n b_{nj}\mathbf{x}_j = \frac{1}{\sqrt{n}}\sum_{j=1}^n \mathbf{x}_j = \sqrt{n}\overline{\mathbf{x}}$, we can rewrite

$$\mathbf{A} = \sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top - n\overline{\mathbf{x}}\,\overline{\mathbf{x}}^\top = \sum_{i=1}^n \mathbf{z}_i\mathbf{z}_i^\top - \mathbf{z}_n\mathbf{z}_n^\top = \sum_{i=1}^{n-1} \mathbf{z}_i\mathbf{z}_i^\top. \tag{1.9}$$

Note that $\mathbf{z}_i$'s, for $i = 1, \ldots, n-1$, are jointly multivariate normally distributed because they are linear combinations of the multivariate normal random vectors $\mathbf{x}_j$'s. To establish their independence, it suffices to check the vanishing of their pairwise covariance matrix. Also note that $\mathbb{E}(\mathbf{z}_n) = \mathbb{E}(\sqrt{n}\overline{\mathbf{x}}) = \sqrt{n}\boldsymbol{\mu}$. For any $i \neq n$, the orthogonality of $\boldsymbol{B}$ and the definition of $\sqrt{n} \cdot b_{nj} = 1$, for all $j = 1, 2, \ldots, n$, together give

$$\mathbb{E}(\mathbf{z}_i) = \sum_{j=1}^n b_{ij}\mathbb{E}(\mathbf{x}_j) = \sum_{j=1}^n b_{ij}\boldsymbol{\mu} = \left(\sum_{j=1}^n b_{ij}b_{nj}\right)(\sqrt{n}\boldsymbol{\mu}) = \mathbf{0}.$$

Finally, by the independence of $\mathbf{x}_i$'s and the orthogonality of $\boldsymbol{B}$, we have

$$\begin{aligned}\mathrm{Cov}(\mathbf{z}_i, \mathbf{z}_j) &= \mathbb{E}\left(\left(\sum_{k=1}^n b_{ik}(\mathbf{x}_k - \boldsymbol{\mu})\right)\left(\sum_{l=1}^n b_{jl}(\mathbf{x}_l - \boldsymbol{\mu})^\top\right)\right) \\ &= \sum_{k=1}^n \sum_{l=1}^n b_{ik}b_{jl}\mathbb{E}\left((\mathbf{x}_k - \boldsymbol{\mu})(\mathbf{x}_l - \boldsymbol{\mu})^\top\right) \\ &= \sum_{k=1}^n \sum_{l=1}^n b_{ik}b_{jl}(\delta_{kl}\boldsymbol{\Sigma}) = \sum_{k=1}^n b_{ik}b_{jk}\boldsymbol{\Sigma} = \delta_{ij}\boldsymbol{\Sigma},\end{aligned}$$

and so $\mathbf{z}_j \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$, for $j = 1, 2, \ldots, n-1$, and the $\mathbf{z}_i$'s are mutually independent. By virtue of (1.9) and the very definition of the Wishart distribution, we can then see that $\mathbf{A} \sim W_p(\boldsymbol{\Sigma}, n-1)$, and therefore $\mathbf{S} \sim W_p\left(\frac{1}{n-1}\boldsymbol{\Sigma}, n-1\right)$. The independence of $\mathbf{S}$ and $\overline{\mathbf{x}}$ is inherited from the independence of $\mathbf{A}$ and $\mathbf{z}_n$, also due to (1.9). □

1.5 INTRODUCTORY BAYESIAN STATISTICS

A Bayesian approach to the estimation of parameters is a probabilistic method in which an optimal decision rule is taken, in contrast to the usual maximum likelihood estimation (MLE) method. Compared with MLE, whose resulting $\hat{\theta}$ maximizes the probability of obtaining the given sample of training dataset, the Bayesian approach evaluates the conditional probability of θ given the sample taking various values instead; see Figure 1.1 for an illustration of the comparison. Indeed, for *Bayesian inference*, the training stage updates our prior belief on the probabilities of θ using the input data, resulting in posterior probabilities as shown in Figure 1.1b.

The theoretical foundation of Bayesian inference is based on an extended form of *Bayes' theorem*: Let $A_1, \ldots, A_n$ be mutually exclusive events and $\cup_{i=1}^{n} A_i =: A$. Then, for any event $B \subseteq A$, we have

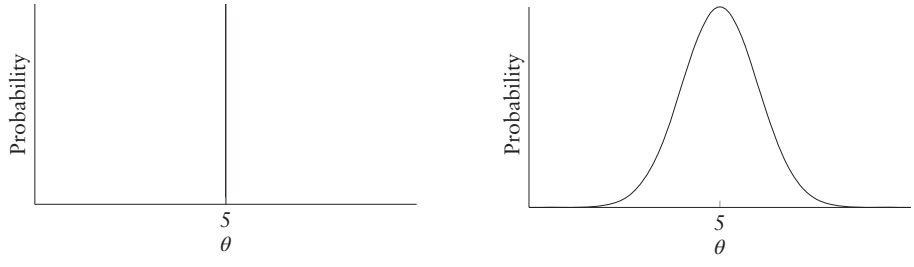
$$\mathbb{P}(A_{(j_1,\ldots,j_k)}|B) = \frac{\mathbb{P}(B|A_{(j_1,\ldots,j_k)})\mathbb{P}(A_{(j_1,\ldots,j_k)})}{\sum_{i=1}^{n}\mathbb{P}(B|A_i)\mathbb{P}(A_i)} = \sum_{\alpha=1}^{k}\frac{\mathbb{P}(B|A_{j_\alpha})\mathbb{P}(A_{j_\alpha})}{\sum_{i=1}^{n}\mathbb{P}(B|A_i)\mathbb{P}(A_i)}, \qquad (1.10)$$

where $A_{(j_1,\ldots,j_k)} = \cup_{\alpha=1}^{k} A_{j\alpha}$.

Example 1.2 John undertook a diagnostic test for heart disease. It is known that 1% of the whole population suffers from the disease. Suppose that 95% of patients (having disease) would be diagnosed by the test as positive, i.e., the accuracy of the test is 95%. On the other hand, 95% of healthy individuals would be correctly diagnosed as negative. Unfortunately, John's test result was positive.

Let A be the event of John having heart disease and B be the event of having a positive test result. Then we have:

$$\mathbb{P}(B|A) = 0.95 = \mathbb{P}(\overline{B}|\overline{A}).$$



(a) MLE ($\hat{\theta} = 5$) (b) Bayesian Inference

FIGURE 1.1 Comparison between MLE (a) and Bayesian inference (b). MLE draws the conclusion $\hat{\theta} = 5$, while Bayesian inference deduces that the conditional probability of $\theta = 5$ given the sample is the highest, yet it does not eliminate other possible values as the true parameter value of θ.

Prior to the test, we know $\mathbb{P}(A) = 0.01$. Given the positive result, the conditional probability of John having the disease becomes

$$\mathbb{P}(A|B) = \frac{0.95 \times 0.01}{0.95 \times 0.01 + 0.05 \times 0.99} \approx 0.16. \tag{1.11}$$

Therefore, even though this diagnostic test has seemingly high accuracy, the probability that John really suffered from the disease given the positive diagnosis result is fairly low. Actually, as we shall elaborate in Section 10.6, once diagnosed as positive, John can only be classified as either true positive (TP) or false positive (FP), and the quantity (1.11) is exactly the precision of the diagnosis, defined as:

$$\text{Precision} := \frac{\text{TP}}{\text{TP} + \text{FP}} = \mathbb{P}(A|B).$$

(I) Normal Distribution with Known Variance

We here present a special case of Bayesian inference for the mean parameter of a normal distribution $\mathcal{N}(\mu, \sigma^2)$; in particular, we first assume that the variance σ^2 is known and only make inference on the unknown mean μ. The case when σ is unknown is discussed at the end of this section.

Let $\mathbf{x} = \{\mathrm{x}_i\}_{i=1}^n$ be a random sample from $\mathcal{N}(\mu, \sigma^2)$, then given μ, the density of x_i is:

$$f_{\mathrm{x}_i}(x_i|\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right), \quad \text{for } x_i \in \mathbb{R}, \tag{1.12}$$

and the conditional joint density of $\mathbf{x}$ given μ is:

$$f(\mathbf{x}|\mu) = \prod_{i=1}^{n} f_{\mathrm{x}_i}(x_i|\mu). \tag{1.13}$$

In the Bayesian framework, the knowledge of μ is considered uncertain, and its uncertainty is described by a (subjective) probability distribution. By Bayes' theorem, the conditional density of μ given that the random sample $\mathbf{x} = \boldsymbol{x}$ is:

$$f_{\mu|\mathrm{x}}(\mu|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\mu)\pi(\mu)}{\int_{-\infty}^{\infty} f(\boldsymbol{x}|\nu)\pi(\nu)d\nu}, \tag{1.14}$$

where $\pi(\mu)$ is known as the *prior* distribution of μ, and $f(\mu|\boldsymbol{x})$ is known as the *posterior* distribution of μ given the sample. Without any prior knowledge, we usually prefer the prior to be uniform over $\mathbb{R}$ so that μ takes any possible real value with equal probability before training, while noting that such a distribution is ill-posed as it cannot be integrated to 1. Alternatively, we may assume the prior distribution to be normal with mean μ_0 and variance σ_0^2, i.e., $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$; indeed, it will resemble a uniform distribution over $\mathbb{R}$ as $\sigma_0^2 \to \infty$, as shown in Figure 1.2, which justifies this choice of prior.

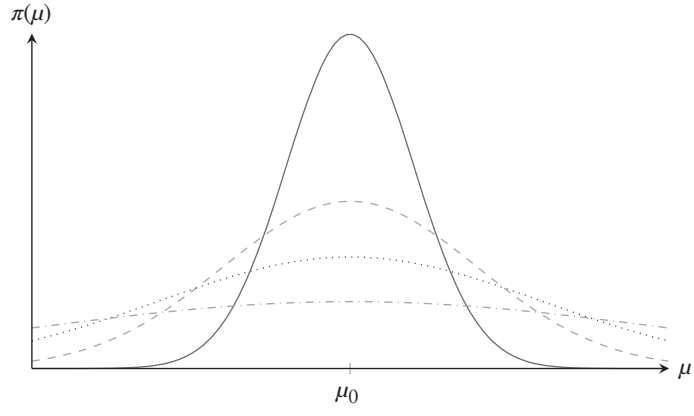


FIGURE 1.2 Density of $\mathcal{N}(\mu_0, \sigma_0^2)$ with $\sigma_0^2 = 1$ (red solid), 5 (orange dashed), 10 (blue dotted) and 20 (green dot-dashed).

We next derive the posterior distribution $f_{\mu|\mathbf{x}}(\mu|\boldsymbol{x})$. Let

$$C := \int_{-\infty}^{\infty} f(\boldsymbol{x}|\nu)\pi(\nu)d\nu$$

be the normalizing constant, then the posterior can be rewritten as:

$$f_{\mu|\mathbf{x}}(\mu|\boldsymbol{x}) = \frac{1}{C} f(\boldsymbol{x}|\mu)\pi(\mu), \tag{1.15}$$

which integrates to 1. For notational convenience, we further denote $\beta := \frac{1}{\sigma^2}$ and $\beta_0 := \frac{1}{\sigma_0^2}$; formally, such reciprocals of the variance of a normal distribution are called the *precision*. Since the prior distribution is $\mathcal{N}(\mu_0, \sigma_0^2)$, we have:

$$f_{\mu|\mathbf{x}}(\mu|\boldsymbol{x}) \propto \exp\left(-\frac{\beta}{2}\sum_{i=1}^{n}(x_i - \mu)^2 - \frac{\beta_0}{2}(\mu - \mu_0)^2\right), \tag{1.16}$$

where the proportionality constant is independent of μ. Then, we expand the exponent in (1.16) as:

$$\begin{aligned}
&-\frac{\beta}{2}\sum_{i=1}^{n}(x_i - \mu)^2 - \frac{\beta_0}{2}(\mu - \mu_0)^2 \\
&= -\frac{n\beta + \beta_0}{2}\mu^2 + \left(\beta\sum_{i=1}^{n}x_i + \beta_0\mu_0\right)\mu + k_1 \\
&= -\frac{n\beta + \beta_0}{2}\left(\mu - \frac{\beta\sum_{i=1}^{n}x_i + \beta_0\mu_0}{n\beta + \beta_0}\right)^2 + k_2,
\end{aligned} \tag{1.17}$$

where the constants k_1 and k_2 are independent of μ. By writing

$$\beta_n := n\beta + \beta_0 \quad \text{and} \quad \mu_n := \frac{\beta \sum_{i=1}^{n} x_i + \beta_0 \mu_0}{n\beta + \beta_0}, \tag{1.18}$$

and substituting (1.17) and (1.18) into (1.16), we see that

$$f_{\mu|\mathbf{x}}(\mu|\boldsymbol{x}) \propto \exp\left(-\frac{\beta_n}{2}(\mu - \mu_n)^2\right). \tag{1.19}$$

Hence, we conclude that the posterior distribution of μ is still a normal distribution but with a mean μ_n and variance β_n^{-1}, i.e., $\mu|\mathbf{x} = \boldsymbol{x} \sim \mathcal{N}(\mu_n, \beta_n^{-1})$.

As mentioned before, to mimic a uniform prior, we simply let $\sigma_0^2 \to \infty$. Then in the limiting case, $\beta_0 = 0$ and $\mu_n = \frac{1}{n}\sum_{i=1}^{n} x_i$, i.e., the sample mean $\overline{x}_n$ of the training data; this result is consistent with the maximum likelihood estimator of μ. On the other hand, as $\beta_0 \to 0$, the variance of the posterior distribution tends to

$$\beta_n^{-1} = \frac{1}{n\beta} = \frac{\sigma^2}{n},$$

which decreases in magnitude with n; graphically, the bell-shaped curve in Figure 1.3 would degenerate at μ_n.

For a small sample size n, we have a rather low confidence that the true value of μ is equal to μ_n, in the sense that the posterior variance β_n^{-1} is not too small. This confidence increases with n, and eventually becomes certain as $n \to \infty$, by then $\beta_n^{-1} \to 0$ which further implies $\mathbb{P}(|\mu - \mu_n| > \epsilon|\mathbf{x} = \boldsymbol{x}) \to 0$ for any $\epsilon > 0$ in light of *Chebyshev's inequality*. This is also consistent with the conclusion drawn by using MLE. In this way, we may consider *Bayesian inference* as supplementing MLE.

Secondly, from (1.18), the posterior mean μ_n is indeed the weighted average of the sample mean $\overline{x}_n$ and the prior mean μ_0, with their respective weights proportional

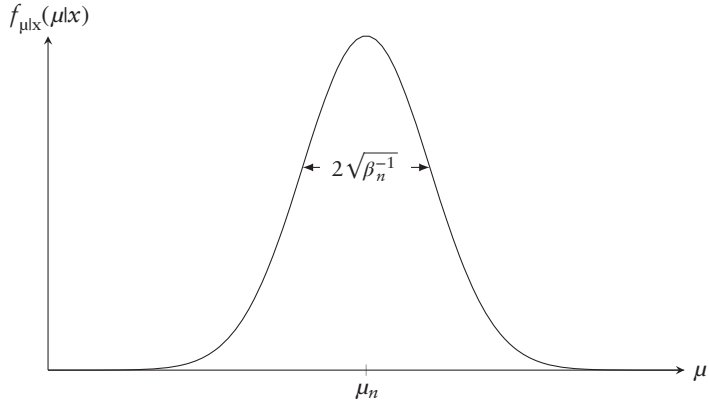


FIGURE 1.3 The posterior distribution $f_{\mu|\mathbf{x}}(\mu|\boldsymbol{x})$.

to $n\beta$ and β_0: the weight for $\overline{x}_n$ in $\mu_n = \frac{n\beta\overline{x}_n+\beta_0\mu_0}{n\beta+\beta_0}$, namely $\frac{n\beta}{n\beta+\beta_0}$, is known as the *credibility factor* in actuarial science. For any fixed value of σ_0^2 (or equivalently, β_0), μ_n always depends increasingly more on $\overline{x}_n$ than μ_0 when n is large, meaning that for a sufficiently large sample, the choice of prior does not significantly affect the posterior estimate of μ.

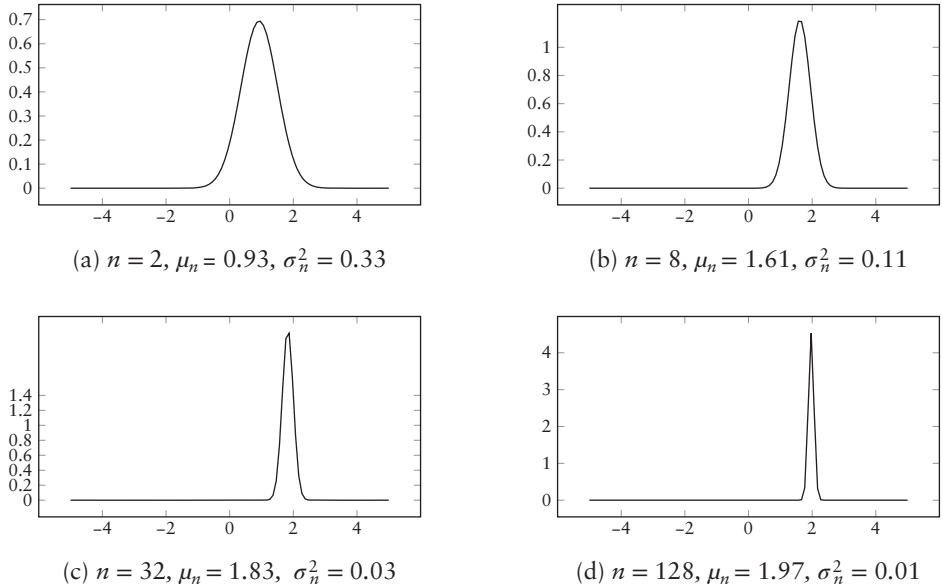
Let us illustrate this claim using a concrete example. We generate a random sample of size n from $\mathcal{N}(2, 1)$, and perform Bayesian inference on the mean $\mu = 2$ with a prior distribution of $\mathcal{N}(-1, 1)$ by pretending that we did not know the mean. Figure 1.4 shows the posterior distributions of μ, and the posterior mean and variance are denoted by μ_n and σ_n^2, respectively. We can observe that the posterior distribution of the mean detaches from the prior distribution and gets closer to the true value of 2 as n increases.

(II) Mathematical connection between MLE and Bayesian inference

We next provide another mathematical perspective to justify the connection between Bayesian inference and MLE in general. Recall the posterior distribution formula (1.15):

$$f_{\mu|x}(\mu|\boldsymbol{x}) = \frac{1}{C} f(\boldsymbol{x}|\mu)\pi(\mu).$$

For the common choice of the prior distribution $\pi(\mu)$ as a constant function independent of μ (which can be made well-posed by restricting it to compact domain), we



(a) $n = 2$, $\mu_n = 0.93$, $\sigma_n^2 = 0.33$

(b) $n = 8$, $\mu_n = 1.61$, $\sigma_n^2 = 0.11$

(c) $n = 32$, $\mu_n = 1.83$, $\sigma_n^2 = 0.03$

(d) $n = 128$, $\mu_n = 1.97$, $\sigma_n^2 = 0.01$

FIGURE 1.4 Posterior distributions of μ with different values of n.

have $f(\mu|\boldsymbol{x}) \propto f(\boldsymbol{x}|\mu)$, and the optimal point μ maximizes both the posterior distribution $f(\mu|\boldsymbol{x})$ and the likelihood function $f(\boldsymbol{x}|\mu)$ simultaneously, which implies agreement between MLE and Bayesian inference. A similar interpretation can be drawn if the likelihood function $f(\boldsymbol{x}|\mu)$ "dominates" the prior distribution $\pi(\mu)$ especially when the domain for μ is bounded. This observation serves as another justification for considering Bayesian inference to be in some sense an extension of MLE.

(III) Inference on the Sampling Distribution
Besides the parameters of the distribution of $\mathbf{x}$, sometimes we are also interested in estimating the distribution itself. Following the previous example, with the knowledge of μ and σ^2, the distribution of x is $\mathcal{N}(\mu, \sigma^2)$. However, the actual value of μ is unknown, hence we have to make use of the posterior distribution $f_{\mu|\mathbf{x}}(\mu|\boldsymbol{x})$ to describe the distribution of x, under the assumption that σ^2 is known.

By virtue of Bayes' theorem, the predictive distribution of x_{n+1} given the sample $x_1, \ldots, x_n$ is simply the conditional distribution of the feature observation x_{n+1} given the sample, so it is the weighted average of the conditional distribution $\mathcal{N}(\mu, \sigma^2)$ of $\mathbf{x}$ with weights $f_{\mu|\mathbf{x}}(\mu|\boldsymbol{x})$, i.e.,

$$\hat{f}_{\mathrm{x}}(x) := f_{\mathrm{x}_{n+1}|\mathbf{x}}(x|\boldsymbol{x}) = \int_{-\infty}^{\infty} f_{\mu|\mathbf{x}}(\mu|\boldsymbol{x})\, f_{\mathcal{N}}(x|\mu, \sigma^2) d\mu, \tag{1.20}$$

where $f_{\mathcal{N}}$ is the normal density with a mean μ and variance σ^2. Recalling that $\mu|\mathbf{x} = \boldsymbol{x} \sim \mathcal{N}(\mu_n, \beta_n^{-1})$, (1.20) becomes:

$$\hat{f}_{\mathrm{x}}(x) \propto \int_{-\infty}^{\infty} \exp\left(-\frac{\beta_n}{2}(\mu - \mu_n)^2 - \frac{\beta}{2}(x - \mu)^2\right) d\mu, \tag{1.21}$$

where the exponent of the integrand can be rearranged as:

$$\begin{aligned} & -\frac{\beta_n}{2}(\mu - \mu_n)^2 - \frac{\beta}{2}(x - \mu)^2 \\ &= -\frac{\beta_n + \beta}{2}\mu^2 + (\beta_n \mu_n + \beta x)\mu - \frac{\beta_n \mu_n^2 + \beta x^2}{2} \\ &= -\frac{\beta_n + \beta}{2}\left(\mu - \frac{\beta_n \mu_n + \beta x}{\beta_n + \beta}\right)^2 + \frac{(\beta_n \mu_n + \beta x)^2}{2(\beta_n + \beta)} - \frac{\beta_n \mu_n^2 + \beta x^2}{2}. \end{aligned} \tag{1.22}$$

Therefore, we have

$$\begin{aligned} \hat{f}_{\mathrm{x}}(x) \propto & \exp\left(\frac{(\beta_n \mu_n + \beta x)^2}{2(\beta_n + \beta)} - \frac{\beta_n \mu_n^2 + \beta x^2}{2}\right) \\ & \cdot \int_{-\infty}^{\infty} \exp\left(-\frac{\beta_n + \beta}{2}\left(\mu - \frac{\beta_n \mu_n + \beta x}{\beta_n + \beta}\right)^2\right) d\mu. \end{aligned} \tag{1.23}$$

Note that the exponent of the first term on the right-hand side can be rewritten as:

$$\frac{(\beta_n\mu_n + \beta x)^2}{2(\beta_n + \beta)} - \frac{(\beta_n\mu_n^2 + \beta x^2)}{2} = -\frac{(x - \mu_n)^2}{2(\beta^{-1} + \beta_n^{-1})} + c, \tag{1.24}$$

where the constant c is independent of x, while the integrand in the second term is just proportional to a normal density with mean $\frac{\beta_n\mu_n+\beta x}{\beta_n+\beta}$ and variance $\frac{1}{\beta_n+\beta}$, hence

$$\int_{-\infty}^{\infty} \exp\left(-\frac{\beta_n + \beta}{2}\left(\mu - \frac{\beta_n\mu_n + \beta x}{\beta_n + \beta}\right)^2\right) d\mu = \sqrt{\frac{2\pi}{\beta_n + \beta}}.$$

Altogether this yields

$$\hat{f}_{\mathrm{x}}(x) \propto \exp\left(-\frac{(x - \mu_n)^2}{2(\beta^{-1} + \beta_n^{-1})}\right),$$

which implies that the predictive distribution is $\mathcal{N}(\mu_n, \beta^{-1} + \beta_n^{-1})$.

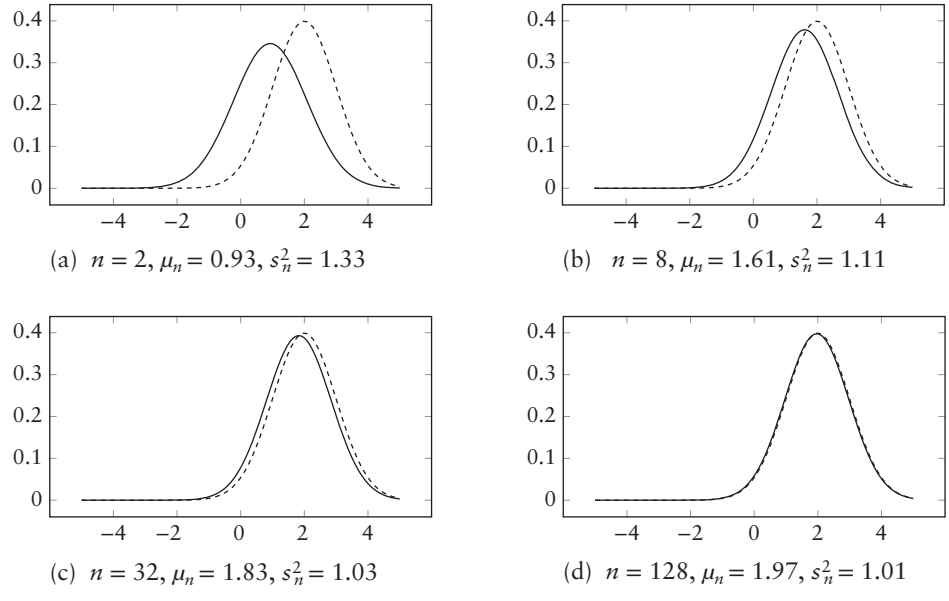
We observe that an additional factor β_n^{-1} appears in the variance of this predictive distribution compared with its original conditional distribution $\mathcal{N}(\mu, \sigma^2)$ given (μ, σ^2); this reflects the additional uncertainty regarding the convergence of μ_n to μ. Recall that β_n^{-1} is the posterior variance of μ given $\mathbf{x} = \boldsymbol{x}$, which decreases to zero with the increase of the sample size n, and so does the variance of this predictive distribution. That means as $n \to \infty$, $\beta_n^{-1} \to 0$, and the variance of x reduces to $\beta^{-1} = \sigma^2$, which echoes the convergence of μ_n towards the true value of μ. Altogether this implies that the predictive distribution of x_{n+1} converges to the true distribution as $n \to \infty$.

As an illustration, we revisit the example in Figure 1.4, where the samples are drawn from $\mathcal{N}(2, 1)$, and we perform Bayesian inference on the mean with the prior distribution $\mathcal{N}(-1, 1)$. In Figure 1.5, we plot the predictive distribution of x_{n+1} for various values of sample size n, namely $\mathcal{N}(\mu_n, s_n^2 := \beta^{-1} + \beta_n^{-1})$ (in solid lines), with the true distribution $\mathcal{N}(2, 1)$ (in dashed line), and clearly the former converges to the latter as n increases.

(IV) Normal Distribution with Unknown Variance

In Bayesian inference, we can choose virtually any arbitrary distribution as the prior $\pi(\cdot)$. However, in practice, we usually prefer to choose $\pi(\mu)$ that could simplify the expression of the posterior distribution (1.15). One common choice can be warranted by the concept of *conjugate prior* for the likelihood function, so that the prior distribution itself and its corresponding posterior one belong to the same family of distribution. For instance, the prior distribution of $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$ introduced earlier in the section is a conjugate prior for the likelihood function of a normally distributed sample, since the posterior $\mu|\mathbf{x} = \boldsymbol{x} \sim \mathcal{N}(\mu_n, \beta_n^{-1})$ is also a normal distribution.

Let us illustrate the technique by studying the extended case when both parameters μ and σ^2 of the normal distribution are unknown. For notational simplicity,



(a) $n = 2, \mu_n = 0.93, s_n^2 = 1.33$ (b) $n = 8, \mu_n = 1.61, s_n^2 = 1.11$

(c) $n = 32, \mu_n = 1.83, s_n^2 = 1.03$ (d) $n = 128, \mu_n = 1.97, s_n^2 = 1.01$

FIGURE 1.5 Predictive distributions of x_{n+1} for various values of n via Bayesian inference.

we use the precision $\lambda = \sigma^{-2}$ instead of variance, then we have $\mathrm{x}|(\mu = \mu, \lambda = \lambda) \sim \mathcal{N}(\mu, \lambda^{-1})$. With observations $\boldsymbol{x} = (x_1, \ldots, x_n)$, the likelihood function reads:

$$f(\boldsymbol{x}|\mu, \lambda) = \prod_{i=1}^{n} f_{\mathrm{x}|\mu,\lambda}(x_i|\mu, \lambda), \tag{1.25}$$

where $f_{\mathrm{x}|\mu,\lambda}(x_i|\mu, \lambda) = \sqrt{\frac{\lambda}{2\pi}} \exp\left(-\frac{\lambda(x_i-\mu)^2}{2}\right)$.

We next consider the *normal-gamma distribution*, denoted by NormalGamma $\left(\mu_0, \beta_0, \frac{\beta_0}{2} + 1, b\right)$, with the following joint probability density function:

$$\pi(\mu, \lambda) = \frac{b^{\frac{\beta_0}{2}+1}}{\Gamma\left(\frac{\beta_0}{2} + 1\right)\sqrt{2\pi}} \lambda^{\frac{\beta_0}{2}} e^{-b\lambda} \sqrt{\beta_0 \lambda} \exp\left(-\frac{\beta_0 \lambda(\mu - \mu_0)^2}{2}\right). \tag{1.26}$$

It is actually the joint distribution of (μ, λ), where $\mu|\lambda = \lambda \sim \mathcal{N}(\mu_0, \frac{1}{\beta_0 \lambda})$ and $\lambda \sim \Gamma\left(\frac{\beta_0}{2} + 1, b\right)$ with the density proportional to $\lambda^{\frac{\beta_0}{2}} e^{-b\lambda}$.[3] For the rest of this section, we aim to show that a normal-gamma distribution is a conjugate prior for the normal

[3] Equivalently, the variance σ^2 is said to follow an *inverse gamma* distribution.

likelihood by deriving the posterior distribution; also see [2, 6]. By (1.25) and (1.26), the posterior density is given by:

$$f(\mu, \lambda|\boldsymbol{x}) \propto (\beta_0\lambda)^{\frac{1}{2}} \lambda^{\frac{\beta_0+n}{2}} e^{-b\lambda} \exp\left(-\frac{\lambda}{2}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{\beta_0\lambda}{2}(\mu-\mu_0)^2\right).$$

Recall that

$$\sum_{i=1}^{n}(x_i-\mu)^2 = \sum_{i=1}^{n}(x_i-\overline{x}_n)^2 + n(\overline{x}_n-\mu)^2 =: nV + n(\overline{x}_n-\mu)^2,$$

where $V = \frac{\sum_{i=1}^{n}(x_i-\overline{x}_n)^2}{n}$. Hence, we can rewrite:

$$\begin{aligned} f(\mu, \lambda|\boldsymbol{x}) &\propto \lambda^{\frac{\beta_0+n+1}{2}} e^{-b\lambda} \exp\left(-\frac{\beta_0\lambda}{2}(\mu-\mu_0)^2 - \frac{\lambda}{2}(nV + n(\overline{x}_n-\mu)^2)\right) \\ &= \lambda^{\frac{\beta_0+n+1}{2}} \exp\left(-\lambda\left(b+\frac{nV}{2}\right)\right) \cdot \exp\left(-\frac{\beta_0\lambda}{2}(\mu-\mu_0)^2 - \frac{n\lambda}{2}(\overline{x}_n-\mu)^2\right). \end{aligned} \tag{1.27}$$

Upon rearrangement, we note that:

$$\begin{aligned} &\frac{\beta_0\lambda}{2}(\mu-\mu_0)^2 + \frac{n\lambda}{2}(\overline{x}_n-\mu)^2 \\ &= \frac{(\beta_0+n)\lambda}{2}\left(\mu^2 - 2\mu\frac{\beta_0\mu_0+n\overline{x}_n}{\beta_0+n} + \left(\frac{\beta_0\mu_0+n\overline{x}_n}{\beta_0+n}\right)^2\right) \\ &\quad - \frac{(\beta_0+n)\lambda}{2}\left(\frac{\beta_0\mu_0+n\overline{x}_n}{\beta_0+n}\right)^2 + \frac{\beta_0\lambda}{2}\mu_0^2 + \frac{n\lambda}{2}\overline{x}_n^2 \\ &= \frac{(\beta_0+n)\lambda}{2}\left(\mu - \frac{\beta_0\mu_0+n\overline{x}_n}{\beta_0+n}\right)^2 + \frac{n\beta_0\lambda}{2(\beta_0+n)}(\overline{x}_n-\mu_0)^2. \end{aligned} \tag{1.28}$$

Then substituting (1.28) into (1.27) yields:

$$\begin{aligned} f(\mu, \lambda|\boldsymbol{x}) \propto \lambda^{\frac{\beta_0+n}{2}} &\exp\left(-\lambda\left(b+\frac{nV}{2}+\frac{n\beta_0}{2(\beta_0+n)}(\overline{x}_n-\mu_0)^2\right)\right) \\ &\cdot \sqrt{(\beta_0+n)\lambda}\exp\left(-\frac{(\beta_0+n)\lambda}{2}\left(\mu - \frac{\beta_0\mu_0+n\overline{x}_n}{\beta_0+n}\right)^2\right), \end{aligned} \tag{1.29}$$

which is the joint probability density of a NormalGamma$(\mu^*, \beta^*, \frac{\beta^*}{2}+1, b^*)$ distribution, with

$$\mu^* := \frac{\beta_0\mu_0+n\overline{x}_n}{\beta_0+n}, \quad \beta^* := \beta_0+n, \quad b^* := b + \frac{nV}{2} + \frac{n\beta_0}{2(\beta_0+n)}(\overline{x}_n-\mu_0)^2;$$

in other words, $\mu|(\lambda, \mathbf{x}) = (\lambda, \boldsymbol{x}) \sim \mathcal{N}\left(\mu^*, \frac{1}{\beta^*\lambda}\right)$ and $\lambda|\mathbf{x} = \boldsymbol{x} \sim \Gamma\left(\frac{\beta^*}{2} + 1, b^*\right)$. This confirms that the NormalGamma$(\mu_0, \beta_0, \frac{\beta_0}{2} + 1, b)$ distribution is indeed a conjugate prior for the likelihood function.

In addition, we provide the predictive distribution of x_{n+1} by using:

$$\hat{f}_{\mathrm{x}}(x) = \int_0^\infty \int_{-\infty}^\infty f(\mu, \lambda|\boldsymbol{x}) f_{\mathrm{x}|\mu,\lambda}(x|\mu, \lambda) d\mu d\lambda.$$

Recall that the conditional density of x_{n+1} given μ, λ takes the following form:

$$f(x|\mu, \lambda) \propto \lambda^{\frac{1}{2}} \exp\left(-\frac{\lambda}{2}(x - \mu)^2\right), \tag{1.30}$$

while by (1.29), the posterior density is:

$$f(\mu, \lambda|\boldsymbol{x}) \propto \lambda^{\frac{\beta^*}{2}} e^{-b^*\lambda} \sqrt{\beta^*\lambda} \exp\left(-\frac{\beta^*\lambda}{2}(\mu - \mu^*)^2\right). \tag{1.31}$$

We then integrate the product of (1.30) and (1.31), first with respect to μ and then with respect to λ. By some standard yet tedious calculations, we get:

$$\hat{f}_{\mathrm{x}}(x) \propto \left(\frac{1}{b^* + \frac{\beta^*}{2(\beta^*+1)}(x - \mu^*)^2}\right)^{\frac{\beta^*+1}{2}+1},$$

which is the predictive probability density of a location-scale t-distribution; in other words, predictively $\mathrm{x}_{n+1} = \mu^* + \tau\mathrm{t}$, where $\mathrm{t} \sim t_\nu$ with $\nu = \beta^* + 2$, and $\tau = \sqrt{\frac{2b^*}{\beta^*}}$.

The model above can also be extended to high-dimensional settings as follows. Assume that given the p-dimensional mean vector $\boldsymbol{\mu}$ and the $p \times p$ covariance matrix $\boldsymbol{\Sigma}$, the random vector $\mathbf{x} = (\mathrm{x}_1, \ldots, \mathrm{x}_p)^\top$ follows a multivariate normal distribution. Via the use of the *precision matrix* $\boldsymbol{P} = \boldsymbol{\Sigma}^{-1}$, we have $\mathbf{x} \mid (\boldsymbol{\mu} = \boldsymbol{\mu}, \mathbf{P} = \boldsymbol{P}) \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{P}^{-1})$. It can be analogously shown that the conjugate prior in this scenario becomes the multivariate extension of the normal-gamma distribution, namely the *normal-Wishart distribution*, denoted by NW$(\boldsymbol{\mu}_0, \lambda, \boldsymbol{\Psi}, \nu)$, which is the joint distribution of $(\boldsymbol{\mu}, \mathbf{P})$ with $\boldsymbol{\mu} \mid \mathbf{P} = \boldsymbol{P} \sim \mathcal{N}_p(\boldsymbol{\mu}_0, (\lambda \boldsymbol{P})^{-1})$ and $\mathbf{P} \sim W_p(\boldsymbol{\Psi}, \nu)$; also see [13, 16] for more discussion.

We conclude this section with a remark on the practical use of normal-Wishart distributions in actuarial science. In credibility theory, they have been widely adopted in the formulation of hierarchical models, such as models for severity per claim, which helps to determine the pure premium by multiplying the estimated mean annual claim frequency and the estimated mean severity per claim to estimate the mean of the annual aggregate loss. In addition, these models can also facilitate an efficient estimation of the variance of each individual, leading to a justifiable pricing of the risk premium for the policyholders; see [12] for example. Meanwhile, in the rise of the trendy *evolutionary credibility theory*, which allows for time-varying mean and variance, the incorporation of temporal structures is inevitable for enhancing the practical

value of the normal-Wishart model. A possible resolution has been proposed in a more recent study [3] under a semi-parametric evolutionary setting, allowing simultaneous estimation of both mean and variance of the policyholder at each time point based on his/her past claim history; this formulation includes the class of static normal-Wishart models above as an interesting special case, and can hence be regarded as its viable extension with the presence of a temporal dependence structure. We anticipate more research developments to emerge along this direction.

REFERENCES

1. Anderson, T.W. (1962). *An Introduction to Multivariate Statistical Analysis*. Wiley.
2. Carlin, B., and Louis, T. (2008). *Bayesian Methods for Data Analysis* (3rd ed.). Chapman & Hall/CRC.
3. Chen, Y., Cheung, K.C., Choi, H.M.C., and Yam, S.C.P. (2020). Evolutionary credibility risk premium. *Insurance: Mathematics and Economics*, 93, 216–229.
4. Chen, Y., Fan, N.S., and Yam, S.C.P. (2024+). *Statistical Deep Learning with Python and R*. Preprint.
5. Deisenroth, M.P., Faisal, A.A., and Ong, C.S. (2020). *Mathematics for Machine Learning*. Cambridge University Press.
6. Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., and Rubin, D.B. (2013). *Bayesian Data Analysis* (3rd ed.). Chapman & Hall/CRC.
7. Ghosh, M., and Sinha, B.K. (2002). A simple derivation of the Wishart distribution. *The American Statistician*, 56(2), 100–101.
8. Golub, G.H., and Van Loan, C.F. (2013). *Matrix Computations*. JHU Press.
9. Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.
10. Hogg, R.V., McKean, J.W., and Craig, A.T. (2019). *Introduction to Mathematical Statistics*. Pearson.
11. Horn, R.A., and Johnson, C.R. (2012). *Matrix Analysis*. Cambridge University Press.
12. Jing, B.Y., Li, Z., Pan, G., and Zhou, W. (2016). On SURE-type double shrinkage estimation. *Journal of the American Statistical Association*, 111(516), 1696–1704.
13. Johnson, R.A., and Wichern, D.W. (2002). *Applied Multivariate Statistical Analysis*. Prentice Hall.
14. Lai, T.L., and Xing, H. (2008). *Statistical Models and Methods for Financial Markets*. Springer.
15. McLachlan, G.J. (1999). Mahalanobis distance. *Resonance*, 4(6), 20–26.
16. Morrison, D.F. (1990). *Multivariate Statistical Methods*. McGraw-Hill.
17. Rencher, A.C., and Christensen, W.F. (2012) *Methods of Multivariate Analysis* (3rd ed.). Wiley.
18. Strang, G. (2006). *Linear Algebra and its Applications*. Thomson, Brooks/Cole.
19. Strang, G. (2019). *Linear Algebra and Learning from Data*. Wellesley-Cambridge Press.
20. Watkins, D.S. (2004). *Fundamentals of Matrix Computations*. Wiley.
21. Zhang, X.D. (2017). *Matrix Analysis and Applications*. Cambridge University Press.