Current Natural Sciences

MULTIVARIATE ANALYSIS

Lang WU and Jin QIU

Applied Multivariate Statistical Analysis and Related Topics with R

SCIENCE PRESS

edp sciences

Current Natural Sciences



Applied Multivariate Statistical Analysis and Related Topics with R

Lang WU and Jin QIU

Multivariate analysis is a popular area in statistics and data science. This book provides a good balance between conceptual understanding, key theoretical presentation, and detailed implementation with software R for commonly used multivariate analysis models and methods in practice.

The authors

Lang Wu is Professor of Statistics at the University of British Columbia in Vancouver, Canada. His research interest includes the analysis of longitudinal data (or panel data) and survival data when there are missing values, censored values, measurement errors, and outliers.

Jin Qiu is Professor of School of Data Sciences at Zhejiang University of Finance and Economics in Hangzhou, China. Her main focus is on multivariate data, longitudinal data (or panel data), functional data, and spatial data.

ISBN : 978-2-7598-2601-8



edp sciences
www.edpsciences.org

SCIENCE PRESS

Current Natural Sciences

Lang WU and Jin QIU

Applied Multivariate Statistical Analysis and Related Topics with R

SCIENCE PRESS edp sciences

This book was originally published by Science Press, © Science Press, 2014.

Printed in France

EDP Sciences – ISBN(print): 978-2-7598-2601-8 – ISBN(ebook): 978-2-7598-2602-5

All rights relative to translation, adaptation and reproduction by any means whatsoever are reserved, worldwide. In accordance with the terms of paragraphs 2 and 3 of Article 41 of the French Act dated March 11, 1957, "copies or reproductions reserved strictly for private use and not intended for collective use" and, on the other hand, analyses and short quotations for example or illustrative purposes, are allowed. Otherwise, "any representation or reproduction – whether in full or in part – without the consent of the author or of his successors or assigns, is unlawful" (Article 40, paragraph 1). Any representation or reproduction, by any means whatsoever, will therefore be deemed an infringement of copyright punishable under Articles 425 and following of the French Penal Code.

The printed edition is not for sale in Chinese mainland. Customers in Chinese mainland please order the print book from Science Press. ISBN of the China edition: Science Press ISBN: 978-7-03-041243-0

© Science Press, EDP Sciences, 2021

Preface

Multivariate analysis models and methods are very useful in data analysis, since in practice data are often collected on more than one variables and these variables are often associated or correlated. A main consideration in multivariate analysis methods is to incorporate the associations or correlations between variables, so multivariate analysis methods are usually more efficient than univariate analysis methods which ignore the associations between variables. However, many multivariate analysis methods are mathematically intractable, so students often get lost in the complicated mathematical expressions, which prevents them to truly understand the basic ideas behind many multivariate analysis models and methods. In teaching a multivariate analysis course for undergraduate or Master level graduate students in statistics, we believe that the main goal should let students to understand the basic ideas of the models and methods and then use these models and methods in real data analysis. Theoretical proofs should be treated as secondary and should be left as exercises which may help students to better understand the methods and to prepare them for further studies or research.

A key feature of this textbook is that it focuses on detailed explanations of the basic ideas of common multivariate analysis models and methods using simple language and illustrations of the methods using software R. Tedious mathematical derivations are omitted from the main text and are left in the exercises. With this approach, students can focus on understanding of the basic ideas without being distracted by tedious mathematical arguments, as well as applications of the models and methods in data analysis using software R. Students with strong mathematical background and with strong interest in further study or research in any topics are encouraged to work on the theoretical exercises available at the end of each chapter. Moreover, many classic books on multivariate analysis contain detailed theoretical results, which are listed in the references, so interested readers can easily assess these materials.

Many books on multivariate analysis focus on inference for multivariate normal populations, such as parameter estimation and inference for multivariate normal distributions and normal regression models. This is restrictive since in practice there are

often discrete or categorical data or skewed data which do not follow normal distributions. In this textbook, we fill the gaps with chapters on multivariate discrete data, copula models, and generalized linear models, in addition to standard topics based on multivariate normal or continuous data. Categorical or discrete data are common in practice, but this topic often receives little attention in many multivariate textbooks and undergraduate and graduate curriculums. In this textbook, we provide an overview of multivariate categorical data analysis, including analysis of contingency tables and loglinear models. Copula models have received much attention in recent years, especially in finance. Copula models allow us to build multivariate distributions from any univariate distributions. Thus they are powerful tools for multivariate analysis. Generalized linear models allow non-normal response variables so they greatly extend the applicability of linear regression models.

Results of data analysis can be greatly influenced by noises in the data, such as missing data, measurement errors, and outliers. For example, in the presence of missing data, analysis results may be biased or less efficient if the missing data are simply discarded, and a few outliers in the data may completely change conclusions from data analysis so the conclusions do not represent population characteristics. These problems are especially common and bad for multivariate data. However, these problems often receive little attention in many books on multivariate analysis, so students do not know how to handle these problems in data analysis. In this textbook, we provide comprehensive discussions of these issues on separate chapters and offer practical suggestions for data analysts.

This book may be used as a textbook for a multivariate analysis course for undergraduate and Master-level graduate students in statistics, as well as students or researchers in other fields who wish to learn basics of multivariate analysis methods and apply these methods in data analysis. The English language is simpler than most English textbooks on multivariate analysis, so it is ideal for students whose first language is not English. Exercises are available at the end of each chapter. These exercises contain both theoretical problems and data analysis problems. Some of the theoretical derivations of the methods in the main text are left as exercise problems. We strongly suggest students to practice these theoretical problems, since such practice will help students to better understand the models and methods. For the applied exercise problems, students can follow the procedures explained in the examples and the R code in the examples. Some materials in certain chapters are challenging, so they are optional for undergraduate students. At the end of Chapter 1 we provide

some general advice on good data analysis practice. The datasets and R code in the book are available at www.stat.ubc.ca/∼lang/text.

We thank many colleagues and students who have provided us useful suggestions and help. We also thank Zhejiang University of Finance and Economics and School of Mathematics and Statistics for their supports and encouragements.

Lang Wu and Jin Qiu
Feb. 2014

Contents

Preface

Chapter 1 Introduction 1

1.1 Goal of Statistics 1

1.2 Univariate Analysis 3

1.3 Multivariate Analysis 7

1.4 Multivariate Normal Distribution 16

1.5 Unsupervised Learning and Supervised Learning 21

1.6 Data Analysis Strategies and Statistical Thinking 23

1.7 Outline 26

Exercises 1 27

Chapter 2 Principal Components Analysis 29

2.1 The Basic Idea 29

2.2 The Principal Components 30

2.3 Choose Number of Principal Components 34

2.4 Considerations in Data Analysis 35

2.5 Examples in R 37

Exercises 2 43

Chapter 3 Factor Analysis 45

3.1 The Basic Idea 45

3.2 The Factor Analysis Model 46

3.3 Methods for Estimation 47

3.4 Examples in R 50

Exercises 3 54

Chapter 4 Discriminant Analysis and Cluster Analysis 56

4.1 Introduction 56

4.2 Discriminant Analysis 57

4.3 Cluster Analysis 61

4.4 Examples in R 64

Exercises 4 69

Chapter 5 Inference for a Multivariate Normal Population 71
5.1 Introduction 71
5.2 Inference for Multivariate Means 72
5.3 Inference for Covariance Matrices 75
5.4 Large Sample Inferences about a Population Mean Vector 76
5.5 Examples in R 76
Exercises 5 79
Chapter 6 Discrete or Categorical Multivariate Data 80
6.1 Discrete or Categorical Data 80
6.2 The Multinomial Distribution 81
6.3 Contingency Tables 83
6.4 Associations Between Discrete or Categorical Variables 85
6.5 Logit Models for Multinomial Variables 87
6.6 Loglinear Models for Contingency Tables 89
6.7 Example in R 91
Exercises 6 95
Chapter 7 Copula Models 97
7.1 Introduction 97
7.2 Copula Models 99
7.3 Measures of Dependence 102
7.4 Applications in Actuary and Finance 103
7.5 Applications in Longitudinal and Survival Data* 106
7.6 Example in R 107
Exercises 7 110
Chapter 8 Linear and Nonlinear Regression Models 111
8.1 Introduction 111
8.2 Linear Regression Models 112
8.3 Model Selection 114
8.4 Model Diagnostics 116
8.5 Data Analysis Examples with R 117
8.6 Nonlinear Regression Models 122
8.7 More on Model Selection 125
Exercises 8 129
Chapter 9 Generalized Linear Models 131
9.1 Introduction 131

9.2 The Exponential Family 132
9.3 The General Form of a GLM 133
9.4 Inference for GLM 135
9.5 Model Selection and Model Diagnostics 137
9.6 Logistic Regression Models 140
9.7 Poisson Regression Models 146
Exercises 9 149
Chapter 10 Multivariate Regression and MANOVA Models 152
10.1 Introduction 152
10.2 Multivariate Regression Models 153
10.3 MANOVA Models 156
10.4 Examples in R 157
Exercises 10 162
Chapter 11 Longitudinal Data, Panel Data, and Repeated Measurements 164
11.1 Introduction 164
11.2 Methods for Longitudinal Data Analysis 165
11.3 Linear Mixed Effects Models 167
11.4 GEE Models 171
Exercises 11 174
Chapter 12 Methods for Missing Data 175
12.1 Missing Data Mechanisms 175
12.2 Methods for Missing Data 178
12.3 Multiple Imputation Methods 181
12.4 Multiple Imputation by Chained Equations 183
12.5 The EM Algorithm 184
12.6 Example in R 187
Exercises 12 192
Chapter 13 Robust Multivariate Analysis 193
13.1 The Need for Robust Methods 193
13.2 General Robust Methods 195
13.3 Robust Estimates of the Mean and Standard Deviation 199
13.4 Robust Estimates of the Covariance Matrix 201
13.5 Robust PCA and Regressions 203
13.6 Examples in R 205

Exercises 13 210
Chapter 14 Selected Topics 211
14.1 Likelihood Methods 211
14.2 Bootstrap Methods 214
14.3 MCMC Methods and the Gibbs Sampler 215
14.4 Survival Analysis 217
14.5 Data Science, Big Data, and Data Mining 220
References 224

Chapter 1
Introduction

1.1 Goal of Statistics

A main goal of statistics is to analyze data in order to obtain useful information and make important decisions. In other words, statisticians analyze data to extract useful information from the data in a sample and draw important conclusions about the population. In modern world, many important decisions are based on information from data. With the developments of modern computers and internet, massive data are available and can be easily obtained, but important information in the data may not be easily obtained without using modern statistical methods. Therefore, statistics is becoming one of the most important subjects in the 21 century, and statistical methods are among the most widely used tools in almost every area, including banks, insurance industries, economics, finance, medicine, and engineering. As a *New York Times* article says (August 5, 2009, **For Today's Graduate, Just One Word: Statistics**): "For many different jobs in today's world, mostly what you do is data analysis (statistics), even for jobs which seem unrelated to statistics ... Many today's decisions in industry and government are based on data analysis results. Statisticians are thus in high demand ...".

Data can be collected in many ways, such as survey, internet, company records and designed experiments. Our goal is to analyze the data to extract as much information as possible and then draw some conclusions about the whole population. For example, if a new drug is found to be effective on 20 randomly selected patients (sample) based on statistical analysis, will this drug also be effective for all patients (population)? If exam scores are found to be related to students' IQ scores as well as students' attitude on 50 randomly selected students (sample), is this also true for all students (population)? Such a generalization from sample to population is called *statistical inference*. Sometimes, however, we may just wish to obtain useful information from the sample, without necessarily making inference about the population, especially if the sample is not a random and representative sample.

In practice, data analysis often consists of two stages:

- exploratory data analysis.

- formal (or confirmatory) data analysis.

In exploratory data analyses, data are simply summarized using common statistics (e.g., means, standard deviations, correlations) and are displayed using common graphical tools (e.g., histograms, boxplots, scatterplots). In this stage, we simply present and summarize the data, without trying to generalize the conclusions obtained from summary statistics and graphs to the whole population. In this stage, we do not need to make any distributional assumptions for the data, i.e., we do not need to *assume* that data follow certain distributions such as the normal distributions. Thus, the conclusions obtained from exploratory analysis do not depend on the validity of any assumptions. Exploratory analysis can reveal important features of the data, which may lead to preliminary conclusions. **Exploratory data analysis is an important step in any data analysis and should not be skipped**. However, exploratory data analysis is usually followed by a formal or confirmatory analysis, which is used to confirm the preliminary conclusions from the exploratory analysis.

In formal (or confirmatory) data analysis, we *assume* models or distributions for the data or population, estimate unknown parameters in the models or distributions, and attempt to make statistical inference so that we may generalize the results based on the sample to the whole population. For example, we may assume that the population follows a normal distribution, and then we use data to estimate the parameters in the normal distribution (mean and variance). Note that the models or distributions are only assumptions, so they may not be true. In other words, the assumed models and distributions should be checked for their validity based on the data. **Since the assumed models or distributions rarely hold exactly, formal analysis results should only be viewed as approximate, and it is desirable to use different statistical models or methods to further validate the results.**

In practice, data are often collected on many variables, such as age, income, and education. When we analyze data on each variable separately, the data on each variable are called *univariate data*, and an analysis of univariate data is called *univariate analysis*. **In univariate analysis, the association or correlation between different variables are ignored.** In other words, when we analyze data on one variable in univariate analysis, we cannot borrow information from data of other variables. For example, suppose that we have data on income from a sample survey. In univariate analysis, we can compute the average income and the standard deviations. But if an income is missing or not reported, we cannot estimate it without information from other variables such as age and job title. Many statistical methods in introductory statistics courses, such as data summaries and hypothesis testing, are designed for univariate analysis.

Since different variables or data may be associated or correlated, analyzing data on several correlated variables *simultaneously* should be more desirable than univari-

ate analysis since additional information from the association can be used. Statistical analysis on data from more than one variables is called *multivariate analysis*. Multivariate analysis is usually more efficient than univariate analysis, although it is more complicated. In addition, multivariate analysis allows us to study the association or correlation between different variables. Thus, **multivariate analysis allows us to study the relationship between different variables and uses more information than univariate analysis.** Regression analysis can also be viewed as multivariate analysis since it studies relationship between variables.

In summary, the advantages of multivariate analysis over univariate analysis are: (i) a multivariate analysis incorporates the association between variables so it uses more information, (ii) a multivariate analysis may provide more efficient inference than a univariate analysis, (iii) a multivariate analysis may avoid biased results in a univariate analysis, and (iv) a multivariate analysis allows us to understand the association between variables. Although multivariate analysis offers many advantages over univariate analysis, multivariate analysis also has some limitations. For example, a multivariate analysis is usually more complicated than a univariate analysis, and many useful statistical tools for univariate analysis may not be available for multivariate analysis. Moreover, graphical displays of data, which are very important for statistical analysis, can be difficult for multivariate data.

In the following sections, we provide a brief overview of some simple univariate and multivariate analysis methods.

1.2 Univariate Analysis

The first step in data analysis should be exploratory analysis, which summarize the data using simple statistics and display the data using graphs. Univariate continuous data are often summarized by simple statistics such as the *mean* and *standard deviation*: the mean measures the center of the data while the standard deviation measures the variation of the data. Univariate data can be displayed by graphical tools such as *histogram* and *boxplot*: a histogram show the distribution (frequencies) of the data while the boxplot shows five number summaries of the data. It is important to display data using graphs since graphs may show unusual patterns in the data and outliers. These unusual patterns and outliers may make the mean and standard deviation misleading and unreliable. **A picture is worth a thousand words!** Graphical tools are important components of statistical analysis.

In statistical analysis, choice of statistical methods depends on the types of data in hand. Generally, there are two types of data (or variables):

- *continuous data (variables)*, which take continuous values.
- *categorical or discrete data (variables)*, which take discrete values.

Examples of continuous data include weight, age, income, blood pressure, etc. Examples of categorical data include gender, location, yes/no answers, blood type, etc. Continuous data may be transformed into categorical data by grouping the data based on some threshold values, such as subjects with ages greater than 60 and subjects with ages less than or equal to 60. Such categorizations of continuous data may simplify analysis but it may also lead to some loss of information. Note that categorical data cannot be converted to continuous data. For a categorical variable, its value may represent a category rather than a numerical value, such as gender (male/female). **Statistical methods for analyzing continuous data and categorical data are quite different.** In this textbook, we will mostly focus on continuous data, but we will also discuss categorical data.

Continuous data are probably most common in practice and thus will be our main focus. For univariate continuous data, the two most important features are

- the center of the data, measured by the sample *mean*.
- the variation of the data, measured by the sample *standard deviation*.

The mean and the standard deviation should always be reported in statistical analysis. They give us some idea about the average of the data and the variation in the data- the two most important features of continuous data. Note that the mean and the standard deviation can be very sensitive to outliers and the data distributions, i.e., a few outliers in the data or a severely skewed data distribution can greatly affect the values of the mean and standard deviation and thus may lead to misleading conclusions. Graphical tools such as histograms may reveal outliers and data distributions, so they should be used in data analysis.

Let $x_1, x_2, \cdots, x_n$ be a random sample on a continuous random variable X. Then the sample *mean* $\bar{x}$ and the sample *standard deviation* s are given respectively by

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \quad s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}.$$

Other summary statistics sometimes may also be useful in data analysis, such as the percentiles, quartiles, medians, etc. These summary statistics are useful when there may be outliers or when the data distributions may not be symmetric (so not normally distributed). For example, a 5th *percentile* is the value such that 5 % data are smaller than it and 95% data are larger than it. A *median* is the 50th percentile, i.e., half the data are smaller than the median and half the data are larger than the median. The *first quartile* is the 25th percentile, and the *third quartile* is the 75th percentile. Note that both the mean and the median measure the center of the data, but the mean can be greatly influenced by an outlier while the median is robust against outliers.

Pictures can reveal important features of the data which may not be seen from numerical summaries. For example, outliers and data distributions can be best detected by pictures. Common graphical methods for summarizing univariate data from a continuous variable include histograms and boxplots. A *histogram* shows the frequencies of the data points or the distribution of the data (see Figure 1.1). A *boxplot* shows five number summaries of the data: minimum, first quartile, median, third quartile, and maximum (see Figure 1.2).

Discrete or categorical data are summarized in different ways than continuous data. We usually use *counts* or *percentages* to summarize discrete data. For example, we can say that there are 25 (or 50%) male students and 25 (or 50%) female students in the class, or in a sample survey 20 (or 15%) participants are teachers, 40 (or 30%) participants are first year students, 20 (or 15%) participants are second year students, etc. In other words, we do not use means and standard deviations to summarize discrete or categorical data, but instead we use counts or percentages. Graphical tools for discrete data include bar-charts and pie-charts, but they are less commonly used than graphical tools for continuous data.

1.2.1 Statistical software

There are many statistical software, such as R, SAS, SPSS, Stata, etc. Any software will produce same or similar results when the same statistical model or method are used, but some software may be preferred over others. We will use software R in this book. R is free and open source. It typically includes latest statistical methods, and it produces nice pictures. R can be downloaded from the R home page: *http://cran.r-project.org/* (select a mirror and go to "Download and Install R"). Readers can go to the R webpage for information on many R packages, including some tutorials. R compiles and runs on a wide variety of UNIX platforms, Windows, and MacOS. R provides a wide variety of statistical and graphical techniques, and is highly extensible.

1.2.2 An Example in R

In the following, we show some simple R examples using a dataset which contains five quiz scores of 53 students in a semester (called *"quiz dataset"*).

```
# Import data from the data file "class.dat1"
> class.dat <- read.table("class.dat1",head=T)  # a dataframe
> attach(class.dat)  # make this data a priority in this R session.
# part of the data
> class.dat
   ID gender quiz1 quiz2 quiz3 quiz4 quiz5
1   1      M    90    79    90    90    93
2   2      F    55    60    58    70    79
3   3      F    60    72    75    80    77
```

```
4   4       F     66     48     89     70     72
..........
# save the following figures to a pdf file "histogram1.pdf"
> pdf("histogram1.pdf")
> par(mfrow=c(2,2)) # 2 by 2 figures on one page
# Show histograms of the first 4 quiz scores
> hist(quiz1, main="Scores for quiz 1",xlab="score", col="grey")
> hist(quiz2, main="Scores for quiz 2",xlab="score", col="grey")
> hist(quiz3, main="Scores for quiz 3",xlab="score", col="grey")
> hist(quiz4, main="Scores for quiz 4",xlab="score", col="grey")
> dev.off()   # save the figure
```

From the histograms in Figure 1.1, we can learn something about the *distributions* of the quiz scores. For example, in Quiz 2 the quiz scores seem roughly symmetric, while in Quiz 3 the quiz scores seem to skewed to the left and there may be outliers. Thus, a normal distribution assumption may be reasonable for Quiz 2 scores but may be inappropriate for Quiz 3 scores. Moreover, the mean and standard deviation for Quiz 3 scores may be misleading because they may be greatly affected by the outliers.

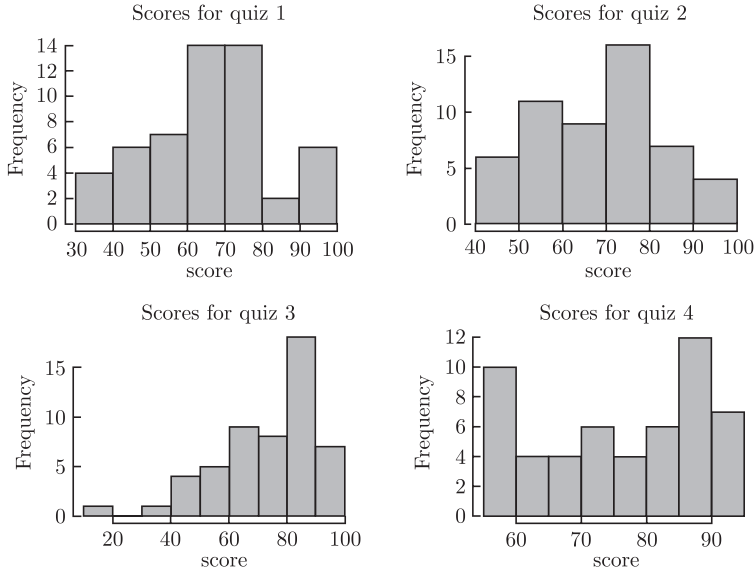


Figure 1.1 Four histograms for the first four quiz scores.

Next, let's also summarize the univariate data for each quiz using boxplots.

```
# Show boxplots of the first 4 quiz scores
> pdf("boxplot1.pdf")
> par(mfrow=c(1,1)) # one figure a page
> boxplot(quiz1,quiz2,quiz3,quiz4, main="First 4 Quiz Scores",
          names=c(1,2,3,4), col="blue")
> dev.off()   # save the figure
```

From the boxplots in Figure 1.2, we see that in Quiz 1 and Quiz 3 there are outliers (a very low score in each quiz). The median quiz scores have an increasing trend. Scores in Quiz 1 and Quiz 3 seem to have more variations than scores in the other two quizes. In Quiz 1 and Quiz 3, one student did much worse than the rest of students, but overall students' performances improve over time.

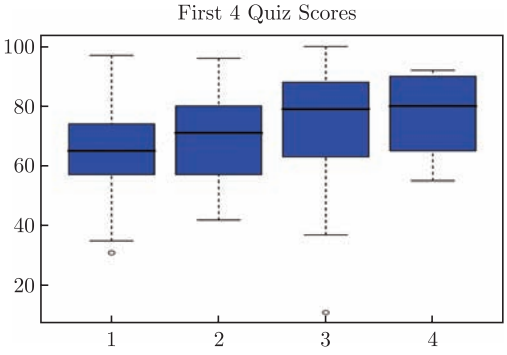


Figure 1.2 Four boxplots for the first four quiz scores.

We can also summarize the quiz data numerically. For example, the scores in Quiz 1 can be summarized as follows: the mean score is 65.7, the standard deviation is 15.65, the minimum is 31, the maximum is 97, the first quartile is 57, the median is 65, and the third quartile is 74.

```
# Summary statistics of the first quiz scores
> summary(quiz1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   31.0    57.0    65.0    65.7    74.0    97.0
> sd1 <- sqrt(var(quiz1))  # standard deviation
 15.65
```

In this example, we have seen the most common ways to summarize univariate data both graphically and numerically. These exploratory analyses give us rough ideas how students perform over time.

1.3 Multivariate Analysis

In practice, data are often collected on more than one variables, and these variables may be associated or correlated, i.e., change in one variable corresponds to changes in other variables. For example, income and education may be correlated, and gender and smoking status may be associated. When variables are associated or correlated, it is desirable to incorporate the association or correlation and analyze the data on these variables simultaneously. In other words, when several variables are associated,

it may be more desirable to use multivariate analysis than univariate analysis, since multivariate analysis uses extra information from the correlation or association. In other words, when analyzing data on one variable from a multivariate dataset, we can borrow information from data on other variables. Note that the term "correlation" is usually used for continuous data while the term "association" may be used for both continuous and discrete data. Data on two or more variables are called *multivariate data*, and statistical analysis of multivariate data is called *multivariate analysis*. **In multivariate analysis, a key consideration is to incorporate the correlation or association between the variables.**

Methods to measure the correlation among continuous data and methods to measure the association among discrete data are quite different. For continuous multivariate data, we usually use correlation matrices while for discrete multivariate data we usually use odds ratio or other measurements, which will be described in details in later chapters. We first focus on multivariate continuous data, where each variable is a continuous variable.

As an example for multivariate continuous data, suppose that a bank or a credit card company wish to classify all customers into two groups: individuals with good credit risks and individuals with bad credit risks. The classification can be based on customers' education (x_1), income (x_2), age (x_3), and past credit history (x_4). It would be easy to do the classification if we just consider one variable, say income. For example, individuals with high income may be classified as good credit risk, while those with low income may be classified as bad credit risk. However, since these variables (x_1, x_2, x_3, x_4) may be associated, we need to consider all variables simultaneously. The classification based on all variables can be challenging. For example, an individual with high income may have low education and young age. Therefore, special multivariate analysis methods are required to do the classification.

When there are many variables to be considered simultaneously, statistical analysis is challenging because: (i) it may be difficult to assume a joint distribution for all variables for statistical inference; (ii) there may be too many parameters in an assumed model for all variables; (iii) the correlation matrix may be of high dimension so it may be singular; and (iv) graphical displays of the multivariate data can be difficult. Therefore, an important question for multivariate analysis is to see if we can reduce the number of variables without much loss of information in the data. This is called *dimension reduction*, and it will greatly simplify statistical analysis. Such a dimension reduction method is often possible since some variables may be highly correlated, so they may be combined in some ways. For example, scores on algebra, geometry, and calculus may be replaced by one variable called mathematical ability. Special multivariate methods are available to do such dimension reduction.

In data analysis, the first step is to summarize the data. continuous multivariate

data are usually summarized by (i) the *mean* of each variable, (ii) the *standard deviation* or *variance* of each variable, and (iii) the *correlation* or *covariance* between the variables. These quantities reflect the center, variation, and association in multivariate data, which are the most important characteristics of multivariate data. In multivariate analysis, the covariance matrix or the correlation matrix play an important role since it contains information about the association among the variables.

Graphical displays of multivariate data can be difficult if the number of variables is more than two. A simple approach is to check the pairwise correlations for any two variables using *scatterplots*, which allow us to visually check the correlation between any two variables. Some other graphical methods such as *star-plots* can also be useful in displaying multivariate data in some cases, but their values are often limited.

1.3.1 Covariance matrix and correlation matrix

Since a main focus in multivariate analysis is to incorporate the correlation between variables, the covariance matrix or the correlation matrix play an important role in multivariate analysis for continuous data or variables. Suppose that X and Y are two continuous random variables, and suppose that $\{x_1, x_2, \cdots, x_n\}$ and $\{y_1, y_2, \cdots, y_n\}$ are data collected on X and Y respectively. The correlation between X and Y can be measured by the *covariance* $Cov(X, Y)$ or the *correlation coefficient* r:

$$Cov(X, Y) = E[(X - E(X))(Y - E(Y))],$$
$$r = \frac{Cov(X, Y)}{\sigma_X \sigma_Y},$$

where $\sigma_X = \sqrt{Var(X)}$ and $\sigma_Y = \sqrt{Var(Y)}$ are the standard deviations of X and Y respectively. The correlation coefficient r is a number between -1 and 1, and it measures the *linear correlation* between two variables. If $r > 0$, X and Y are *positively correlated*. If $r < 0$, X and Y are *negatively correlated*. The larger the value $|r|$, the stronger the correlation. When $r = 0$, X and Y are not linearly correlated.

Note that the correlation coefficient r only measures a linear relationship, not other relationships, i.e., the observed values of X and Y roughly fall on a straight line. For example, when $r = 0$, X and Y may still have a nonlinear relationship.

More generally, let $\mathbf{X} = (X_1, X_2, \cdots, X_p)^{\mathrm{T}}$ be a random vector with p continuous random variables. The mean vector of $\mathbf{X}$ is

$$E(\mathbf{X}) = \boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_p)^{\mathrm{T}}.$$

The covariance matrix of $\mathbf{X}$ is given by

$$\Sigma = Cov(\mathbf{X}) = (\sigma_{ij})_{p \times p}, \quad \text{where } \sigma_{ij} = E(X_i - E(X_i))(X_j - E(X_j)),$$

and the correlation between X_i and X_j is

$$r_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}, \quad i, j = 1, 2, \cdots, p.$$

For multivariate data, the mean vector $\boldsymbol{\mu}$ represents the *center* of the data, and the covariance matrix Σ represents the *variation* (variance) and *correlation* of the data.

The covariance matrix Σ or the correlation matrix $R = (r_{ij})_{p\times p}$ play a key role in multivariate analysis since they measure the strength of the association between continuous variables. **Much information about the distribution of multivariate continuous data is contained in the *eigenvalues* and *eigenvectors* of the covariance matrix Σ. The eigenvalues describe the *shape* of the data region, while the eigenvectors show the *directions* of the axes of the data region.** Specifically, let $\{(\lambda_i, \mathbf{a}_i),\ i = 1, \cdots, p\}$ be the eigenvalues-eigenvectors of the covariance matrix Σ. Then, the variance of data points in the $\mathbf{a}_i$ direction is $\lambda_i,\ i = 1, 2, \cdots, p$. See Figure 6 on page 23 for an example in the case of $p = 2$. This is a very important fact to keep in mind in order to understand many of the multivariate analysis methods.

Since eigenvalues represent the variation (variance) of data in certain directions, if a few eigenvalues are much larger than the rest eigenvalues of a covariance matrix, the multivariate data essentially fall onto the sub-space spanned by the eigenvectors corresponding to the few large eigenvalues, so the dimension of the multivariate data can be reduced to that subspace. This is the essential idea behind dimension-reduction methods in multivariate analysis, such as the principal components method. We will see many examples when describing the principal components method.

In practice, the parameters in the population mean vector $\boldsymbol{\mu}$, the covariance matrix Σ, and the correlation matrix $\boldsymbol{R}$ are all unknown. However, they can be estimated from data. Specifically, let $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ be an i.i.d. sample obtained from a population with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ, denoted by $\mathbf{x}_i \sim (\boldsymbol{\mu}, \Sigma)$ (note that here the population does not have to be normally distributed), where

$$\mathbf{x}_j = (x_{j1}, \cdots, x_{jp})^{\mathrm{T}}, \quad \boldsymbol{\mu} = (\mu_1, \cdots, \mu_p)^{\mathrm{T}}, \quad \Sigma = (\sigma_{kl})_{p\times p}, \qquad j = 1, 2, \cdots, n.$$

Then, $\boldsymbol{\mu}$ and Σ can be estimated by the *sample mean* $\bar{\mathbf{x}}$ and the *sample covariance matrix* S as follows

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i, \quad \boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^{\mathrm{T}} = (\hat{\sigma}_{lk})_{p\times p}.$$

In data analysis, we can use these estimates to replace the unknown population parameters. As we can imagine, the larger the sample size n, the more accurate the estimates.

Sometimes statistical analysis results may depend on the units or scales of the measurements. For example, a person's height may be measured in meters or centimetres, leading to different values, which may affect analysis results. Therefore, it is sometimes desirable to use scale-free statistics or data, which do not depend on the units or scales of measurements. Two commonly used scale-free measurements are the standardized data and correlation matrix. The *standardized data* are defined as

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{\hat{\sigma}_{jj}}}, \quad i = 1, \cdots, n;\ j = 1, \cdots, p.$$

which has a mean 0 and variance 1. The correlation between two random variables X_i and X_j can be estimated by the following *sample correlation coefficient*

$$\hat{r}_{ij} = \frac{\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)^2 \sum_{k=1}^{n}(x_{kj} - \bar{x}_j)^2}}.$$

The sample correlation matrix $\boldsymbol{R}$ is given by $\boldsymbol{R} = (\hat{r}_{ij})_{p\times p}$, which has the diagonal elements all being 1 and off-diagonal elements all being values between –1 and 1.

Outliers may be present in multivariate data. *Outliers* are unusual observations which are distinctly different from rest of the data. For example, a score of 0 in an exam may be viewed as an outlier, if most students did very well and only very few students did poorly. Outliers may lead to misleading statistical analysis results if they are not handled appropriately. For example, the sample mean and the sample covariance can be greatly influenced by a few outliers in the data. For univariate data, we often can visually detect a possible outlier from a picture. Unfortunately, for multivariate data, it is difficult to use graphical tools to detect outliers since it is difficult to graphically display multivariate data. In statistical analysis, we should always pay attention to possible outliers in the data.

Note that so far we have not assumed any distributions for the data. In other words, we do not need to assume any distributions for the random variables or the data in order to define the foregoing quantities. The only assumption we have made is that the data are continuous.

1.3.2 Examples in R

In the following, we show some simple R examples to summarize and display multivariate continuous data. These examples are for illustrations only.

Example 1. Consider the quiz score dataset described in the previous section again. We can view this dataset as a multivariate dataset if we view each quiz as a random