On high–dimensional Mahalanobis distances

Deliang Dai

Dissertation
Department of Economics and Statistics
Linnaeus University
Box 451
351 06 Växjö

ISBN: 978-91-88357-71-7

Abstract

This thesis investigates on the properties of several forms of MDs under different circumstances. For high–dimensional data sets, the classic MD doesn't work satisfyingly because the complexity of estimating the inverse covariance matrix increases drastically. Thus, we propose a few solutions based on two directions: first, find a proper estimation of the covariance matrix. Second, find explicit distributions of MDs with sample mean and sample covariance matrix of normally distributed random variables and the asymptotic distributions of MDs without assumption of normal distribution. Some of the methods are implemented with empirical datasets.

We also combine the factor model with MDs since the factor model simplifies the estimation of both the covariance matrix and its inverse for structured data sets. The results offer a new way of detecting outliers from this type of structured variables. An empirical application presents the differences between the classic method and the one we derived.

Besides the estimations, we also investigated the qualitative measures of MDs. The distributional properties, first moments and asymptotic distributions for different types of MDs are derived.

The MDs are also derived for complex random variables. The MDs first moments are derived under the assumption of normally distribution. Then we relax the distribution assumption on the complex random vector. The asymptotic distribution is derived with regard to the estimated MD and the leave-one-out MD. Simulations are also supplied to verify the results.

Sammanfattning

Denna avhandling studeras egenskaperna hos olika former av Mahalanobis avstånd, på Engelska Mahalanobis distance (MD), under olika förhållanden. För högdimensionella data fungerar klassiska skattningar av MD inte tillfredställande eftersom komplexiteten med att skatta den inversa kovariansmatrisen ökar drastiskt. Därför föreslår några lösningar baserat på två ansatser: För det första, finn en lämplig skattning av kovariansmatrisen. För det andra, finn en explicit fördelning av MD med medelvärde och kovariansmatris skattade från stickprov av normalfördelade variabler, och asymptotisk fördelning av MD utan normaltantagande. Några av metoderna tilämpas med empiriska data.

Vi kombinerar också faktormodell med MD då faktormodellen förenklar skattning av både kovariansmatrisen och dess invers för strukturerade datamängder. Resultaten ger en ny metod för att upptäcka extremvärden från denna typ av strukturerade variabler. En empirisk tillämpning visar skillnaderna mellan den klassiska metoden och den som härlett.

Förutom skattningar har också de kvalitativa egenskaperna på MD undersökts. Fördelningsegenskaper, första moment och asymptotisk fördelning för olika typer av MD härleds.

MD härleds även för komplexa slumpvariabler. Vi definierar MD för den reala delen och den imaginära delen av en komplex slumpmässig vektor. Deras första moment härleds under antagande om normalfördelning. Sedan lättar vi på antagandet om fördelningen på den komplexa slumpmässiga vektorn. Den asymptotiska fördelningen har härletts under mer generella antaganden. Simuleringar presenteras också för att bekräfta resultaten.

Acknowledgements

I would like to express my deepest appreciation to the people who helped me finish this thesis.
First and foremost, my greatest gratitude goes to my supervisor Prof. Thomas Holgersson for his comments, discussions and patience. He led me in the right direction of my research. His unlimited knowledge and generous guidance have been invaluable to me throughout this amazing research journey. I am deeply grateful to him for introducing such an interesting topic to me.
I am also very grateful to my secondary supervisor Prof. Ghazi Shukur for all his supports. He makes my life easier all the time. Many thanks to Dr. Peter Karlsson who helped to improve my knowledge of both academia and vehicles. Thank him for showing me the real meanings of humility and kindness. Thanks also to Dr. Hyunjoo Karlsson, for interesting conversations and meals.
Many thanks to Prof. Rolf Larsson for numerous valuable and important comments on my licentiate thesis. Thanks to Assoc. Prof. Taras Bodnar for all helpful comments that have improved my thesis. Thanks to Prof. Fan Yang Wallentin and Prof. Adam Taube for introducing me to the world of statistics. Thanks to Prof. Dietrich von Rosen and Assoc. Prof. Tatjana von Rosen for their kind help and valuable suggestions. Thanks to Prof. Jianxin Pan for his inspirational discussions during my visit at the University of Manchester, UK.
Many thanks also go to my office colleagues. Thanks to Aziz who is always very interesting to chat with and who has given me much useful knowledge ranging from research to practical tips about living in Sweden. Thanks to Chizheng for the Chinese food and all the chatting. Thanks to Abdulaziz for all the interesting chats on football and casual life. All these amazing people make our office a fantastic place.
I would also like to thank all my colleagues at the Department of Economics and Statistics as well as all friends in Stockholm, Uppsala, Tianjin and all over the world.
Last but not least, I would like to thank my family who encourage me all the time. Mum, I made it as you wished. Thanks to my wife Yuli for her support and patience during difficult times.

Deliang Dai
24 March 2017 on the train from Växjö to Stockholm

List of papers

This thesis includes four papers as follows:

- Dai D. Holgersson T. Karlsson P. Expected and unexpected values of Individual Mahalanobis Distances. Forthcoming in *Comminications in Statistics – Theory and Methods.*
- Dai D. Holgersson T. High-dimensional CLTs for individual Mahalanobis distances. Forthcoming in *Trends and Perspectives in Linear Statistical Inference - LinStat, Istanbul, August 2016, Springer.*
- Dai D. Mahalanobis distances of factor structured data. Manuscript.
- Dai D. Mahalanobis distances of complex normal random vectors. Manuscript.

Contents

1 Introduction **1**
1.1 Outline 4

2 Mahalanobis distance **5**
2.1 Definitions of Mahalanobis distances 6

3 Random matrices **11**
3.1 Wishart distribution 11
3.2 The Wigner matrix and semi circle law 12

4 Complex random variables **17**
4.1 Definition of general complex random variables 17
4.2 Circularly-symmetric complex normal random variables 19
4.3 Mahalanobis distance on complex random vectors 19

5 MDs under model assumptions **21**
5.1 Autocorrelated data 21
5.2 The factor model 23

6 Future work and unsolved problems **25**

7 Summary of papers **27**
7.1 Paper I: Expected and unexpected values of Mahalanobis distances in high–dimensional data 28
7.2 Paper II: High-dimensional CLTs for individual Mahalanobis distances 28
7.3 Paper III: Mahalanobis distances of factor structure data 28
7.4 Paper IV: Mahalanobis distances of complex random variables . . . 29

8 Conclusions **31**

Chapter 1

Introduction

Multivariate analysis is an important direction of statistics that analyses the relationships between more than one variable. Due to the practice, multiple variables data sets appear more commonly than the univariate cases since we usually concern ourselves with several features of the observations in an analysis. Thus, the measurements and analysis of the dependence between variables and between groups of variables are important for most of the multivariate analysis methods.

One of the multivariate methods is called Mahalanobis distance (herein after MD) (Mahalanobis, 1930). It is used as a measure of the distance between two individuals with several features (variables). In daily life, the most common measure of distance is the Euclidean distance. Then what is the difference between the MD and the Euclidean distance? Why do we need the MD instead of the Euclidean distance in some specific situations? We introduce the advantages of MD here.

Assume we have a data set with the scatter plot as in Figure 1.1. We would like to find out the distances between any two individuals of this data set. The shape of this plot is close to an ellipse, whose two axes are labelled in the figure as well. The origin of the ellipse is the centroid of the points, which is the intersection of the two axes. Assume we draw a unit circle on the scatter plot together with the axes. The distances between the points on the circle and the origin are all equal and labelled as X_i, $i = 1, ...n$. But it does not seem so obvious if we rotate the whole space as $X \mapsto AX + B$ as in Figure 1.2, where A and B are some constant matrices. The distances are turned into a different space as in Figure 1.3.

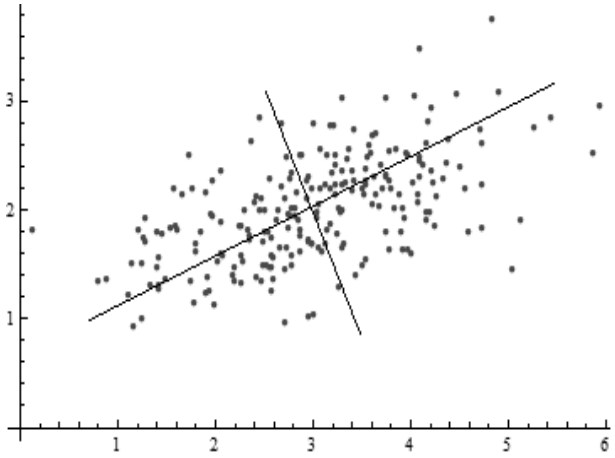

Figure 1.1: Scatter plot.

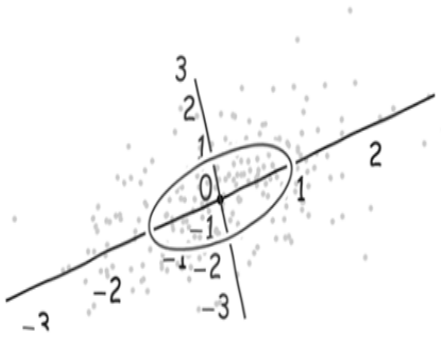

Figure 1.2: MD to a rotated space.

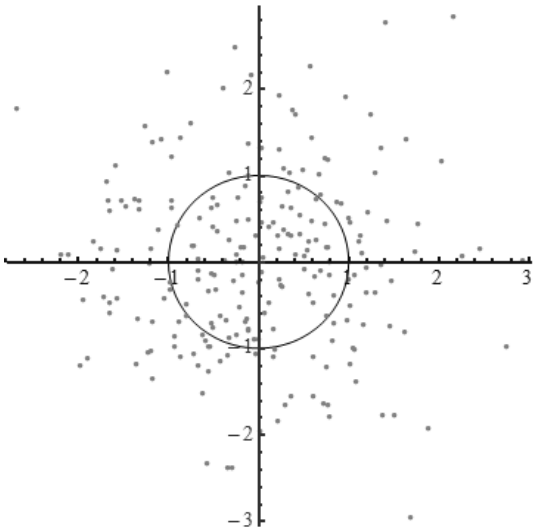

Figure 1.3: Euclidean distance.

Figure 1.3 shows a straightforward circle, which is more conform to common sense. As a result, if we still use the Euclidean distance to measure the distance between the points on the ellipse and the origin, they will be different. However, we know that they are still the same while the measure does not perform equivalently well for this situation. Therefore, some other measures that remain unaffected by the rotations than the Euclidean distance should be implemented, in this example a linear transformation.

The MD is implemented in this thesis in order to avoid the problems above. It was proposed by Mahalanobis (1930) in order to measure the similarity of pairwise individuals. Later on, the idea was extended to several applications related to the measure of difference between observations. We will introduce some details later.

For multivariate analysis problems, most of the classic statistical methods are investigated under a typical data set that satisfies two conditions: first, the data set has a dimension of n observations and p variables where n is much larger than p; second, there is a sample, like a randomly selected normally distributed population. Under these conditions, various statistical methods have been well–developed in the last hundred years. However, with the development of information technology, collecting data is becoming more and more easy. The size of a data set, both horizontally (p) and vertically (n), is increasing drastically. Thus, needs of new statistical methods arise with regard to the new types of data sets. Furthermore, the case of a high–dimensional data set violates the first assumption more frequently as well. As a result, high–dimensional data analysis arises as a new research direction in statistics. Some asymptotic results have been well–developed for several years. However, they mainly focus on the situation where the number

of observations n is increasing while the number of variables p is fixed. Thus, with an increasing p, especially when p is comparably close to n, the classic methods of multivariate analysis would not be the most proper way for analysis.
In addition, many data sets are not normally distributed. There are many potential reasons for this problem, such as outliers in the data set. Thus, the second assumption is also rarely satisfied. Therefore, some statistics are developed in this thesis in order to investigate the problems above and some new estimators are proposed for high–dimensional data sets.

1.1 Outline

Section 2 is a general introduction to MD. We give a short review of random matrices in Section 3. Complex random variables are defined in Section 4. Section 5 discusses some model-based MDs. Section 6 introduces some future research topics. In Section 7, we summarise the contributions of the papers in this thesis. We draw the conclusions of this thesis in Section 8.

Chapter 2

Mahalanobis distance

In multivariate analysis, MD has been a fundamental statistic, proposed by Mahalanobis (1930). It has been applied by researchers in several different areas. The MD is used for measuring the distance between vectors with regard to different practical uses, such as the difference between pairwise individuals, comparing the similarity of observations, etc. Based on this idea, MD is developed into different forms of definitions. Distinct forms of MDs are referred to in the literature (Gower, 1966; Khatri, 1968; Diciccio and Romano, 1988). Before we introduce the definitions of MDs, we should first define the Mahalanobis space. The definition is given as follows:

Definition 1. *Let $\boldsymbol{x}_i, i = 1, ..., n$ be random vectors with p components, its mean be $E(\boldsymbol{x}_i) = \boldsymbol{\mu}$ and covariance matrix be $Cov(\boldsymbol{x}_i) = \boldsymbol{\Sigma}$, the Mahalanobis space $\boldsymbol{y}_i$ is generated by*

$$\boldsymbol{y}_i = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x}_i - \boldsymbol{\mu}), \qquad i = 1, ...n.$$

The definition of Mahalanobis space shows several of its advantages. First, it takes the correlation between random vectors into account. By standardising the random vectors with their covariance matrix, the measures on the individuals are more reasonable and comparable. Second, the definition shows that the MDs are invariant to linear transformations. This could be understood by some steps of simple derivations which will be illustrated later in this thesis. Third, it gives the MDs some convenient properties. We list them below.

Proposition 1. *Let $D(P, Q)$ be the distance between two points P and Q in Mahalanobis space, then we have*

1. *symmetry. $D(P, Q) = D(Q, P)$,*
2. *non-negativity. $D(P, Q) \geq 0$. We have $D(P, Q) = 0$ if and only if $P = Q$,*

3. triangle inequality. $D(P,Q) \geq D(P,R) + D(R,Q)$.

The formal definitions of MDs are given below.

2.1 Definitions of Mahalanobis distances

We present the definitions of MDs as follows:

Definition 2. *Let* $\boldsymbol{X}_i : p \times 1$ *be a random vector such that* $E[\boldsymbol{X}_i] = \boldsymbol{\mu}$ *and* $E\left[(\mathbf{X}_i - \boldsymbol{\mu})(\mathbf{X}_i - \boldsymbol{\mu})'\right] = \boldsymbol{\Sigma}_{p\times p}$, *then the MD (Mahalanobis, 1936) between the the random vector and its mean vector is defined as*

$$D(\boldsymbol{\Sigma}, \boldsymbol{X}_i, \boldsymbol{\mu}) = (\mathbf{X}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}_i - \boldsymbol{\mu}). \tag{2.1}$$

where $'$ stands for the transpose. The form above is the well–known form of MD frequently seen in the literature. Furthermore, for different considerations, there are several types of MDs. In this thesis, we consider several types of MDs according to different aims. Their definitions are presented below.

Definition 3. *Let* $\boldsymbol{X}_i : p \times 1$ *be a random vector such that* $E[\boldsymbol{X}_i] = \boldsymbol{\mu}$ *and* $E\left[(\mathbf{X}_i - \boldsymbol{\mu})(\mathbf{X}_i - \boldsymbol{\mu})'\right] = \boldsymbol{\Sigma}_{p\times p}$, $\boldsymbol{X}_i$, $\boldsymbol{X}_j$ *independent. Then we make the following definitions:*

$$\dot{D}(\boldsymbol{\Sigma}, \boldsymbol{X}_i, \boldsymbol{X}_j) = (\mathbf{X}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}_j - \boldsymbol{\mu}), \tag{2.2}$$

$$D(\boldsymbol{\Sigma}, \boldsymbol{X}_i, \boldsymbol{X}_j) = (\mathbf{X}_i - \mathbf{X}_j)'\boldsymbol{\Sigma}^{-1}(\mathbf{X}_i - \mathbf{X}_j). \tag{2.3}$$

The statistic (2.1) measures the scaled distance between an individual variable $\mathbf{X}_i$ and its expected value $\boldsymbol{\mu}$ and is frequently used to display data, assess distributional properties and detect influential values, etc. The MD (2.2) measures the distance between two scaled and centred observations. This measure is used in cluster analysis and also to calculate the Mahalanobis angle between $\mathbf{X}_i$ and $\mathbf{X}_j$ subtended at $\boldsymbol{\mu}$, defined by $cos\theta(\mathbf{X}_i, \mathbf{X}_j) = \dot{D}(\boldsymbol{\Sigma}, \boldsymbol{X}_i, \boldsymbol{X}_j) \Big/ \sqrt{D(\boldsymbol{\Sigma}, \boldsymbol{X}_i, \boldsymbol{\mu}) D(\boldsymbol{\Sigma}, \boldsymbol{X}_j, \boldsymbol{\mu})}$. The third statistic, (2.3), is related to (2.2) but centres the observation $\mathbf{X}_i$ about another independent observation $\mathbf{X}_j$ and is thereby independent of an estimate of $\boldsymbol{\mu}$.

On the applications, the mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ are usually unknown. Thus, the sample mean and sample covariance are used for the estimators above instead. Estimators of (2.1) – (2.3) may be obtained by simply replacing the unknown parameters with appropriate estimators. If both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown and replaced by the standard estimators, we get the well-known estimators defined below.

Definition 4. *Let* $\{\mathbf{X}_i\}_{i=1}^n$ *be* n *independent realizations of the random vector* $\mathbf{X}$, $\bar{\mathbf{X}} = n^{-1}\sum_{i=1}^n \mathbf{X}_i$ *and* $\mathbf{S} = n^{-1}\sum_{i=1}^n \left(\mathbf{X}_i - \bar{\mathbf{X}}\right)\left(\mathbf{X}_i - \bar{\mathbf{X}}\right)'$. *Following the ideas above, we make the following definition:*

$$D\left(\mathbf{S}, \boldsymbol{X}_i, \bar{\mathbf{X}}\right) = \left(\mathbf{X}_i - \bar{\mathbf{X}}\right)'\mathbf{S}^{-1}\left(\mathbf{X}_i - \bar{\mathbf{X}}\right).$$

This is the MD with sample mean $\bar{\mathbf{X}}$ and sample covariance matrix $\mathbf{S}$. It is used for many applications, based on two different forms of random vectors and its hypothesis mean vector (Rao, 1945; Hotelling, 1933).

Definition 5. *Let* $\mathbf{S}_{(i)} = (n-1)^{-1}\sum_{k=1,k\neq i}^n \left(\mathbf{X}_k - \bar{\mathbf{X}}_{(i)}\right)\left(\mathbf{X}_k - \bar{\mathbf{X}}_{(i)}\right)'$, $\bar{\mathbf{X}}_{(i)} = (n-1)^{-1}\sum_{k=1,k\neq i}^n \mathbf{X}_k$, $\mathbf{S}_{(ij)} = (n-2)^{-1}\sum_{k=1,k\neq i,k\neq j}^n \left(\mathbf{X}_k - \bar{\mathbf{X}}_{(ij)}\right)\left(\mathbf{X}_k - \bar{\mathbf{X}}_{(ij)}\right)'$, $\bar{\mathbf{X}}_{(ij)} = (n-2)^{-1}\sum_{k=1,k\neq i,k\neq j}^n \mathbf{X}_k$,

$$D\left(\mathbf{S}_{(i)}, \boldsymbol{X}_i, \bar{\mathbf{X}}_{(i)}\right) = \left(\mathbf{X}_i - \bar{\mathbf{X}}_{(i)}\right)'\mathbf{S}_{(i)}^{-1}\left(\mathbf{X}_i - \bar{\mathbf{X}}_{(i)}\right).$$

This MD is built with the so-called "leave-one-out" and "leave-two-out" random vectors (De Maesschalck et al., 2000; Mardia, 1977). By leaving the i^{th} observation out, we get independence between the sample covariance matrix and the centred vector. Further, it will not contaminate the sample mean and covariance matrix if there is an outlier in the data set. Therefore, it is an alternative to the classic MD as in Definition 2 when the data set is not badly contaminated. The independence between the sample covariance matrix and the mean vector makes the investigations on the MDs neat and simple.

The MDs are widely implemented in many statistical applications due to their advantageous properties. First, Mahalanobis's idea was proposed to solve the problem of identifying the similarities in biological topics based on measurements in 1927. MD is used as the measure between two random vectors as discriminant analysis on the linear and quadratic discriminations (Fisher, 1936; Srivastava and Khatri, 1979; Fisher, 1940; Hastie et al., 1995; Fujikoshi, 2002; Pavlenko, 2003; McLachlan, 2004) and classification with covariates (Anderson, 1951; Friedman et al., 2001; Berger, 1980; Blackwell, 1979; Leung and Srivastava, 1983a,b). It is closely related to Hotelling's T-square distribution which is used for multivariate statistical testing and Fisher's linear discriminant analysis. The later method is used for supervised classification. In order to use the MD to classify a target individual into one of N classes, one first estimates the covariance matrix of each class, usually based on samples known to belong to each class. Then, given a test sample, one computes the MD to each class, and classifies the test point as belonging to that class based on the value of MDs. The observations with the minimal distances are chosen as the classified observations. Second, MD is also used for detection of multivariate outliers (Mardia et al., 1980; Wilks, 1963). MD

and leverage are often used to detect outliers, especially in applications related to linear regression models. The observation with a larger value of MD than the rest of the sample population of points is said to have leverage since it has a considerable influence on the slope or coefficients of the regression equation. Outliers can affect the results of any multivariate statistical methods from several aspects. First, outliers may lead to abnormal values of correlation coefficients (Osborne and Overbay, 2004; Marascuilo and Serlin, 1988). A correlation with outliers will produce biassed sample estimations, since the linearity among a pair of variables can not be trusted (Osborne and Overbay, 2004). Another common estimator is the sample mean, which is used in ANOVA and many other analyses (Osborne and Overbay, 2004). An outlier would make the sample mean drastically biassed, and the result of ANOVA would be flawed. Further, methods based on the correlation coefficient such as factor analysis and structural equation modelling are also affected by outliers. Their estimations depend on the estimation accuracy of the correlation structure among the variables while outliers will cause the collinearity problem (Brown, 2015; Pedhazur, 1997).

Regression techniques can be used to determine if a specific case within a sample population is an outlier via the combination of two or more variables. Even for normal distributions, a point could be a multivariate outlier even if it is not a univariate outlier for any variable, making MD a more sensitive measure than checking dimensions individually. Third, as a connection to Hotelling's T^2, MD is also applied in hypothesis testing (Fujikoshi et al., 2011; Mardia et al., 1980). Fourth, Mardia (1974); Mardia et al. (1980); Mitchell and Krzanowski (1985); Holgersson and Shukur (2001) use MD as part of some statistics such as skewness and kurtosis as a criteria statistic for assessing the assumption of multivariate normality. Mardia (1974) has defined two statistics in order to test multi-normality – skewness and kurtosis. They are given by

$$b_{1,p} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[D\left(\boldsymbol{S}, \boldsymbol{X}_i, \boldsymbol{X}_j\right) \right]^3,$$

and

$$b_{2,p} = \frac{1}{n^2} \sum_{i=1}^{n} \left[D\left(\boldsymbol{S}, \boldsymbol{X}_i, \bar{\boldsymbol{X}}\right) \right]^2.$$

To the population case, they could be expressed as follows:

$$\beta_{1,p} = E\left[\left(\boldsymbol{X} - \boldsymbol{\mu}\right)' \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{Y} - \boldsymbol{\mu}\right) \right]^3,$$

and

$$\beta_{2,p} = E\left[\left(\boldsymbol{X} - \boldsymbol{\mu}\right)' \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{X} - \boldsymbol{\mu}\right) \right]^2,$$

where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are distributed identically and independently. Note also that, for the sample covariance matrix and leave-one-out covariance matrix, working with dimension n instead of $n-1$ is harmless to our results since the majority of them are derived under asymptotic conditions.
We investigate some of the properties of MD in this thesis under several different considerations.

Chapter 3

Random matrices

In the 1950s, a huge number of experiments related to nuclei were made in order to measure the behaviours of heavy atoms. The experiments produced high–dimensional data due to the fact that the energy level of heavy atoms changes very quickly. Thus, to track and label the energy levels was a difficult but necessary task for researchers. Wigner and Dyson (Dyson, 1962) proposed an idea that, by finding the distribution of energy levels, one can get an approximate solution for the nuclear system. The idea of random matrices was thus employed to describe the properties of heavy nucleus. Wigner assumed the elements of a random matrix to be the heavy nucleus which is independently chosen from a distribution. One simple scenario of the random matrices is the Wishart matrix. We describe it in the coming section.

3.1 Wishart distribution

The Wishart distribution can be considered as a generalised multivariate distribution of the chi–square distribution. It is used to describe the distribution of symmetric, non-negative definite matrix–valued random variables. One notable example is the sample covariance matrix $\boldsymbol{S} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})'$ where $\boldsymbol{x}_i, i = 1, ..., n$, is a p dimensional random sample from a normal distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The Wishart distribution is defined as follows:
Let $\boldsymbol{X}$ be an $n \times p$ matrix, each row of which is following a p-variate normal distribution with zero mean:

$$\boldsymbol{x}_i \sim N_p(\boldsymbol{0}, \boldsymbol{\Sigma}).$$

Then the Wishart distribution is the probability distribution of the $p \times p$ random matrix $\boldsymbol{S} = \boldsymbol{X}'\boldsymbol{X}'$ with the presentation

$$\boldsymbol{S} \sim W_p(\boldsymbol{\Sigma}, n),$$

where n is the number of degrees of freedom. The joint distribution of several independent Wishart distributions is also important. One of them is the multivariate beta distribution. We show its definition as follows:

Definition 6. *Let* $\boldsymbol{W}_1 \sim \boldsymbol{W}_p(\boldsymbol{I}, n)$, $p \leq n$, *and* $\boldsymbol{W}_2 \sim \boldsymbol{W}_p(\boldsymbol{I}, m)$, $p \leq m$ *be independently distributed. Then,*

$$\boldsymbol{F} = (\boldsymbol{W}_1 + \boldsymbol{W}_2)^{1/2} \boldsymbol{W}_2 (\boldsymbol{W}_1 + \boldsymbol{W}_2)^{1/2}$$

has a multivariate beta distribution with density function given by

$$f_{\boldsymbol{F}}(\boldsymbol{F}) = \begin{cases} \frac{c(p,n)c(p,m)}{c(p,n+m)} |\boldsymbol{F}|^{\frac{1}{2}(m-p-1)} |\boldsymbol{I} - \boldsymbol{F}|^{\frac{1}{2}(n-p-1)}, & |\boldsymbol{I} - \boldsymbol{F}| > 0, |\boldsymbol{F}| > 0, \\ 0, & otherwise, \end{cases}$$

where $c(p,n) = \left(2^{\frac{pn}{2}} \Gamma_p\left(\frac{n}{2}\right)\right)^{-1}$ *and* $(\boldsymbol{W}_1 + \boldsymbol{W}_2)^{1/2}$ *is a symmetric square root.*

By the definition of Wishart distribution, we can investigate the properties of the sample covariance matrix and its related statistics such as MDs. But for the high–dimensional data, there are some difficulties with regard to investigations of the MDs. Thus, some other results can be used in order to derive the sample covariance matrix and related statistics. A more general case of the Wishart matrix is the Wigner matrix, which was actually proposed even before the Wishart matrix. We introduce it in the next section.

3.2 The Wigner matrix and semi circle law

First let us specify some notations. Recall that a matrix $\boldsymbol{H} = (\boldsymbol{H}_{ij})_{i,j=1}^{n}$ is Hermitian if and only if

$$\boldsymbol{H} = \boldsymbol{H}'.$$

In terms of the matrix elements, the Hermitian properties read

$$\boldsymbol{H}_{ij} = \boldsymbol{H}_{ji}^{*};$$

where $*$ stands for the complex conjugate. If we need to split the real and complex components of the elements, we write

$$\boldsymbol{H}_{ij} = \boldsymbol{H}_{ij}^{R} + i\boldsymbol{H}_{ij}^{I};$$

where $\boldsymbol{H}_{ij}^{R}$ is the real part and $i\boldsymbol{H}_{ij}^{I}$ is the complex part. A particularly important case is that of real symmetric matrices. A matrix $\boldsymbol{H}$ is real symmetric if and only if all its entries are real and

$$\boldsymbol{H} = \boldsymbol{H}'.$$

By using these notations, we introduce the definition of the Wigner matrix as follows:

Definition 7. *A Wigner matrix ensemble is a random matrix ensemble of Hermitian matrices* $\boldsymbol{H} = (\boldsymbol{H}_{ij})_{i,j=1}^{n}$ *such that*
– the upper-triangular entries $H_{ij}, i > j$ *are i.i.d. complex random variables with mean zero and unit variance,*
– the diagonal entries H_{ii} *are i.i.d. real variables, independent of the upper triangular entries, with bounded mean and variance.*

Then we can specify Wigner's semicircle law:

Theorem 1. *Let* $\boldsymbol{H}_n$ *be a sequence of Wigner matrices and I an interval. Then we introduce the distribution of the random variables below*

$$E_n(I) = \frac{\#\left\{\lambda_j\left(H/\sqrt{n}\right) \in I\right\}}{n}.$$

Then $E_n(I) \to \mu_{sc}(I)$ *in probability as* $n \to \infty$.

It is possible to study the behaviour of the $E_n(I)$ without computing the eigenvalues directly. This is accomplished in terms of a random measure, the empirical law of eigenvalues.

Definition 8. *The empirical law of eigenvalues* μ_n *is the random discrete probability measure*

$$\mu_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{\lambda_j}\left(\boldsymbol{H}/\sqrt{n}\right).$$

Clearly this implies that for any continuous function $f \in \mathbb{C}(R)$ we obtain

$$\int f d\mu_n = \frac{1}{n}\sum_{i=1}^{n} f\left(\lambda_j\right).$$

As a result, the summation of the eigenvalues of a matrix which is equivalent to the trace of a matrix can be connected with the random matrix theory. Several such results are used in this thesis.
One concern of this thesis is that, under some extreme situations, the classic MD as in (2.1) can not be applied directly for analysis since the dimension of the variables is too large. An example is given below in order to illustrate the problem in details.

Let $\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_n$ be a sample from a p-dimensional Gaussian distribution $N(\mathbf{0}, \boldsymbol{I}_p)$ with mean zero and identity covariance matrix. Let the sample covariance matrix be $\mathbf{S}_n = 1/n \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i'$. An important statistic in multivariate analysis is $W_n = log(|\boldsymbol{S}_n|) = \sum_{j=1}^{p} log(\gamma_{n,j})$, where $\gamma_{n,j}, 1 \leq j \leq p$ are the eigenvalues of $\boldsymbol{S}_n$, $||$ is the determinant. It is used in several statistical analysis methods such as coding, communications (Cai et al., 2015), signal processing (Goodman, 1963) and statistical inference (Girko, 2012). When p is fixed, $\gamma_{n,j} \to 1$ almost surely as $n \to \infty$, and thus $W_n \to 0$. Furthermore, by taking a Taylor expansion of $log(1+x)$, when $p/n = c \in (0,1)$ as $n \to \infty$, it is shown that,

$$\sqrt{n/p} W_n = d\,(c)\,\sqrt{np} \xrightarrow{a.s.} -\infty,$$

where $d(c) = \dfrac{1}{p} W_n \to \displaystyle\int_{a(c)}^{b(c)} \frac{log\pi}{2\pi cx} [\{b(c) - x\}\{x - a(c)\}]^{1/2} dx = \frac{c-1}{c} log(1-c) - 1$, $a\,(c) = (1 - \sqrt{c})^2$ and $b\,(c) = (1 + \sqrt{c})^2$. Thus, any test which assumes asymptotic normality of W_n will result in a serious error as shown in Figure 3.1 below.

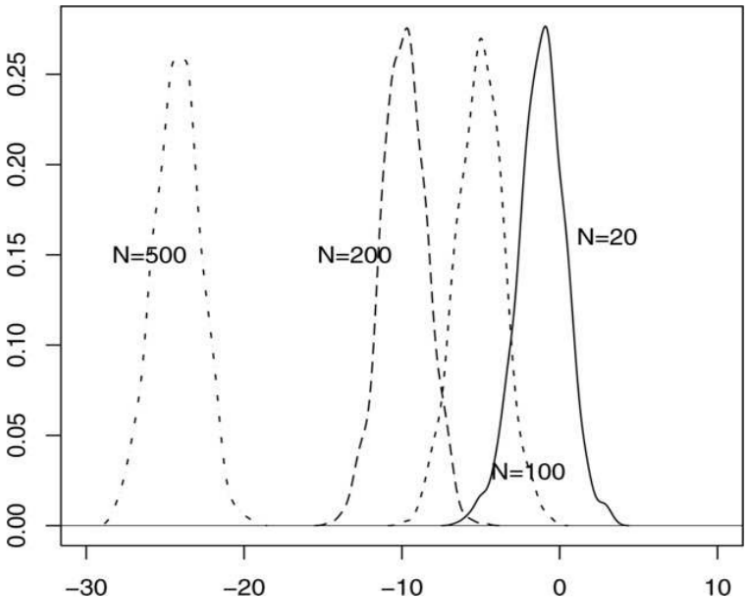

Figure 3.1: Density of W_n under different sample sizes w.r.t. c = 0.2.

As a consequence, methods involving $W_n = log(|\boldsymbol{S}_n|)$ would be suffering from serious weaknesses. One common example is the log likelihood function of a normally distributed sample with sample covariance matrix $\boldsymbol{S}_n$. To high–dimensional data, the common log likelihood function will be varying drastically with the changing of sample sizes and dimensions of variables. Thus, some alternative methods should

be developed in order to investigate the behaviours of the sample covariance matrix under some extreme situations.
So far, many studies of the inverse covariance matrix are developed in the non-classic dataset. Here, classic data stands for the case when the sample size (n) is much larger than the dimensions of variables (p). But for high–dimensional data with both large and close values of (n) and (p), the classic methods perform poorly in most situations. We are concerned with developing some new methods that could be implied in some of these situations. This is implemented by deriving the asymptotic distributions of the MDs. Some useful results of the connection between different types of MDs are also investigated. The other method focuses on reduction of dimensions. Factor analysis and principal component analysis are two methods for dimension reduction. They both maintain the necessary information while reducing the dimension of the variables into a few combinations. Factor models have another advantage in that they could be used to estimate the covariance matrix efficiently. This property is also used to build a new type of MD. This thesis utilises both ideas.

Chapter 4

Complex random variables

As mentioned before, MDs are used for many different aims related to methods of multivariate analysis. One of them is finding meaningful information from multiple inputs, such as signals, which are measured in the form of complex random variables. The complex random variable is an important concept in many fields, such as signal processing (Wooding, 1956), magnetotelluric method (Chave and Thomson, 2004), communication technologies (Bai and Silverstein, 2010) and time series analysis (Brillinger, 2012). Compared with their wide applications, MDs on complex random vectors are rarely mentioned. Hence, investigations on some inferentially related properties and MD on complex random vectors are worthwhile. In the last part of this thesis, we will investigate some properties of MDs on complex random vectors under both normal and non-normal distributions.

4.1 Definition of general complex random variables

We introduce some basic definitions of complex random variables here. Due to its differences from the random variables in real space, we define the covariance matrix of a general complex random vector first as follows:

Definition 9. *Let* $\boldsymbol{z}_j = (z_1, \ldots, z_p)' \in \mathbb{C}^p$, $j = 1, \ldots, n$ *be a complex random vector with known mean* $E[\boldsymbol{z}_j] = \boldsymbol{\mu}_{z,j}$ *where* $z_j = x_j + iy_j$, $i = \sqrt{-1}$. *Let* $\boldsymbol{\Gamma}_{p\times p}$ *be the covariance matrix and* $\boldsymbol{C}_{p\times p}$ *be the relation matrix. The covariance matrix of the complex random vector* $\boldsymbol{z}_j$ *is defined as follows:*

$$\boldsymbol{\Gamma} = E\left[(\boldsymbol{z}_j - \boldsymbol{\mu}_{z,j})(\boldsymbol{z}_j - \boldsymbol{\mu}_{z,j})^*\right].$$

Switching between a complex random vectors $\boldsymbol{z}$ and its expanded form $\boldsymbol{z} = \boldsymbol{x} + i\boldsymbol{y}$ is straightforward. Let $\boldsymbol{z}_j$ be a complex random sample and we get

$$\boldsymbol{z}_j = \begin{pmatrix} 1 & i \end{pmatrix} \begin{pmatrix} \boldsymbol{x}_j \\ \boldsymbol{y}_j \end{pmatrix}.$$

This connection makes the derivation simpler. For different considerations of research, the expanded form is clearer and easily used to explain the results (Chave and Thomson, 2004). The connection between a complex random vector and its extended real components is illustrated as follows.
The covariance matrix of a p - dimensional complex random vector can also be represented in the form of $\boldsymbol{x}$ and $\boldsymbol{y}$, as follows:

$$\boldsymbol{\Gamma}_{z,2p\times 2p} = \begin{pmatrix} \boldsymbol{\Gamma}_{xx} & \boldsymbol{\Gamma}_{xy} \\ \boldsymbol{\Gamma}_{yx} & \boldsymbol{\Gamma}_{yy} \end{pmatrix},$$

where $\boldsymbol{\Gamma}_{xx,p\times 1} = \frac{1}{2}\operatorname{Re}(\boldsymbol{\Gamma}+\boldsymbol{C}) = E\left[(\boldsymbol{x}-\operatorname{Re}\boldsymbol{\mu})(\boldsymbol{x}-\operatorname{Re}\boldsymbol{\mu})'\right]$; $\boldsymbol{\Gamma}_{yy,p\times 1} = \frac{1}{2}\operatorname{Re}(\boldsymbol{\Gamma}-\boldsymbol{C}) = E\left[(\boldsymbol{y}-\operatorname{Im}\boldsymbol{\mu})(\boldsymbol{y}-\operatorname{Im}\boldsymbol{\mu})'\right]$; $\boldsymbol{\Gamma}_{xy,p\times 1} = \frac{1}{2}\operatorname{Im}(\boldsymbol{C}-\boldsymbol{\Gamma}) = E\left[(\boldsymbol{x}-\operatorname{Re}\boldsymbol{\mu})(\boldsymbol{y}-\operatorname{Im}\boldsymbol{\mu})'\right]$; $\boldsymbol{\Gamma}_{yx,p\times 1} = \frac{1}{2}\operatorname{Im}(\boldsymbol{\Gamma}+\boldsymbol{C}) = E\left[(\boldsymbol{y}-\operatorname{Im}\boldsymbol{\mu})(\boldsymbol{x}-\operatorname{Re}\boldsymbol{\mu})'\right]$.

Theorem 2. *The quadratic form of the real random vectors and the quadratic form of the complex random vectors can be connected as:*

$$q(\boldsymbol{x},\boldsymbol{y}) = q'(\boldsymbol{z},\boldsymbol{z}^*) = \boldsymbol{\nu}^*\boldsymbol{\Gamma}_{\boldsymbol{\nu}}^{-1}\boldsymbol{\nu}$$

where $\boldsymbol{\Gamma}_{\boldsymbol{\nu}}^{-1} = \boldsymbol{M}^*\boldsymbol{\Gamma}_{2p\times 2p}^{-1}\boldsymbol{M}$.

Proof. Following Picinbono (1996) we have that $(Re\boldsymbol{\Gamma})^{-1} = (\boldsymbol{\Gamma}_{xx}+\boldsymbol{\Gamma}_{yy})^{-1} = 2\boldsymbol{\Gamma}^{-1}$, $(Im\boldsymbol{\Gamma})^{-1} = [i(\boldsymbol{\Gamma}_{xx}+\boldsymbol{\Gamma}_{yy})]^{-1} = i^{-1}(\boldsymbol{\Gamma}_{xx}+\boldsymbol{\Gamma}_{yy})^{-1} = \boldsymbol{0}$; the inverse matrix of $\boldsymbol{\Gamma}$ is

$$\boldsymbol{\Gamma}^{-1} = (2\boldsymbol{\Gamma}_{xx}+\boldsymbol{0})^{-1} = 2^{-1}\boldsymbol{\Gamma}_{xx}^{-1}.$$

By the results above, the quadratic form of the complex random vector can be expressed as follows:

$$q(\boldsymbol{z},\boldsymbol{z}^*) = 2[\boldsymbol{z}^*\boldsymbol{P}^{-1*}\boldsymbol{z} - \boldsymbol{R}(\boldsymbol{z}^T\boldsymbol{R}^T\boldsymbol{P}^{-1*}\boldsymbol{z})],$$

where $\boldsymbol{P}^{-1*} = \boldsymbol{\Gamma}^{-1} + \boldsymbol{\Gamma}^{-1}\boldsymbol{C}\boldsymbol{P}^{-1}\boldsymbol{C}^*\boldsymbol{\Gamma}^{-1}$; $\boldsymbol{R} = \boldsymbol{C}^*\boldsymbol{\Gamma}^{-1}$; $\boldsymbol{\Gamma} = \boldsymbol{\Gamma}_x + \boldsymbol{\Gamma}_y + i(\boldsymbol{\Gamma}_{yx} - \boldsymbol{\Gamma}_{xy})$ and $\boldsymbol{C} = \boldsymbol{\Gamma}_x - \boldsymbol{\Gamma}_y + i(\boldsymbol{\Gamma}_{yx} + \boldsymbol{\Gamma}_{xy})$. ■

4.2 Circularly-symmetric complex normal random variables

A circularly–symmetric complex random variable is an assumption used for many situations as a standardised form of complex Gaussian distributed random variables. We introduce them as follows:

Definition 10. *A p-dimension complex random variable* $\boldsymbol{z}_{p\times 1} = \boldsymbol{x}_{p\times 1} + i\boldsymbol{y}_{p\times 1}$ *is circularly–symmetric complex normal if the vector* $vec[\boldsymbol{x}\ \boldsymbol{y}]$ *is bivariate normally distributed as follows:*

$$\begin{pmatrix}\boldsymbol{x}_{p\times 1}\\ \boldsymbol{y}_{p\times 1}\end{pmatrix} \sim N\left(\begin{bmatrix}\mathrm{Re}\ \boldsymbol{\mu}_{z,p\times 1}\\ \mathrm{Im}\ \boldsymbol{\mu}_{z,p\times 1}\end{bmatrix}, \frac{1}{2}\begin{bmatrix}\mathrm{Re}\ \boldsymbol{\Gamma}_{z,p\times 1} & -\mathrm{Im}\ \boldsymbol{\Gamma}_{z,p\times 1}\\ \mathrm{Im}\ \boldsymbol{\Gamma}_{z,p\times 1} & \mathrm{Re}\ \boldsymbol{\Gamma}_{z,p\times 1}\end{bmatrix}\right),$$

where $\boldsymbol{\mu}_{z,p\times 1} = E[\boldsymbol{z}]$ *and* $\boldsymbol{\Gamma}_{z,p\times p} = E\left[(\boldsymbol{z}-\boldsymbol{\mu}_z)(\boldsymbol{z}-\boldsymbol{\mu}_z)^*\right]$.

The circularly–symmetric normally distributed complex random variable is one way to simplify the analysis of complex random variables. By this condition, we get a simplified form of probability density function on a complex normal random vector as follows:

Definition 11. *The circularly–symmetric complex random vector* $\boldsymbol{z} = (z_1, \ldots, z_p)' \in \mathbb{C}^p$ *assumes that the mean vector* $\boldsymbol{\mu}_z = \boldsymbol{0}$ *and the relation matrix of the complex vector* $\boldsymbol{C} = \boldsymbol{0}$. *Its probability density function is*

$$f(\boldsymbol{z}) = \frac{1}{\pi^p|\boldsymbol{z}|} exp(-\boldsymbol{z}^*\boldsymbol{\Gamma}_z^{-1}\boldsymbol{z}).$$

The circularly–symmetric complex normal shares many properties with the standard normal random variables in the real plane. Some of the results here will be used to define the MDs.

4.3 Mahalanobis distance on complex random vectors

We now turn to the definitions of MDs with complex random variables.

Definition 12. *The original Mahalanobis distance of the complex random vector* $\boldsymbol{z}_i : p\times 1, i = 1, \ldots, n$ *with known mean* $\boldsymbol{\mu}_{p\times 1}$ *and known covariance matrix* $\boldsymbol{\Gamma}_z$: $p \times p$ *can be formulated as follows:*

$$D\left(\boldsymbol{\Gamma}_z, \boldsymbol{z}_i, \mathrm{Re}\,\boldsymbol{\mu}\right) = \left(\boldsymbol{z}_i - \mathrm{Re}\,\boldsymbol{\mu}\right)^*\boldsymbol{\Gamma}_z^{-1}\left(\boldsymbol{z}_i - \mathrm{Re}\,\boldsymbol{\mu}\right). \tag{4.1}$$

As we know, there are two parts of a complex random vector. In each separate component of a complex random vector, we can also find the corresponding MDs. The MDs on separate parts of a complex random vectors are defined as follows.

Definition 13. *The Mahalanobis distance on the real part* $\boldsymbol{x}_i : p \times 1$ *and imaginary part* $\boldsymbol{y}_i : p \times 1$ *of a complex random vector* $\boldsymbol{z}_i : p \times 1, i = 1, \ldots, n$ *with known mean* $\boldsymbol{\mu}$ *and known covariance matrix* $\boldsymbol{\Gamma}_{\cdot\cdot}$ *is defined as follows:*

$$D\left(\boldsymbol{\Gamma}_{xx}, \boldsymbol{x}_i, \operatorname{Re}\boldsymbol{\mu}\right) = \left(\boldsymbol{x}_i - \operatorname{Re}\boldsymbol{\mu}\right)' \boldsymbol{\Gamma}_{xx}^{-1} \left(\boldsymbol{x}_i - \operatorname{Re}\boldsymbol{\mu}\right), \tag{4.2}$$

$$D\left(\boldsymbol{\Gamma}_{yy}, \boldsymbol{y}_i, \operatorname{Im}\boldsymbol{\mu}\right) = \left(\boldsymbol{y}_i - \operatorname{Im}\boldsymbol{\mu}\right)' \boldsymbol{\Gamma}_{yy}^{-1} \left(\boldsymbol{y}_i - \operatorname{Im}\boldsymbol{\mu}\right). \tag{4.3}$$

Definition 13 specifies the MDs on each part of a complex random vector separately. Next, we turn to another definition of MD that compares the real random vectors $\boldsymbol{x}$ and $\boldsymbol{y}$.

Chapter 5

MDs under model assumptions

5.1 Autocorrelated data

Autocorrelation is a characteristic of data frequently occurring in economic and other data. The violation of the assumption of independence makes most of the statistical models infeasible since most of them assume independence. Practically, the presence of autocorrelation is more frequent than one may expect. For example, when analysing time series data, the correlation between a variable's current value and its past value is usually non-zero. In a sense, they are dependent all the time. It is only a matter of stronger or weaker autocorrelation. Many statistical methods fail to work properly when the assumption of independence is violated. Thus, some methods that can handle this type of situation are needed.
One example is the VAR (vector autoregression) model (Lütkepohl, 2007). A VAR model is a generalisation of the univariate autoregressive model for forecasting a collection of variables, that is, a vector of time series. It comprises one equation per variable considered in the system. The right hand side of each equation includes a constant and lags of all the variables in the system. For example, we write a two–dimensional VAR(1) as follows:

$$y_{1,t} = c_1 + \phi_{11,1} y_{1,t1} + \phi_{12,1} y_{2,t1} + e_{1,t}, \tag{5.1}$$
$$y_{2,t} = c_2 + \phi_{21,1} y_{1,t1} + \phi_{22,1} y_{2,t1} + e_{2,t}. \tag{5.2}$$

where $e_{1,t}$ and $e_{2,t}$ are white noise processes that may be contemporaneously correlated. The coefficient $\phi_{ii,k}$ captures the influence of the k^{th} lag of variable y_i on itself, while coefficient $\phi_{ij,k}$ captures the influence of the k^{th} lag of variable y_j on y_i etc. By extending the lag order, we can generalize the VAR(1) to a p^{th} order VAR, denoted VAR(p):

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + e_t, \tag{5.3}$$

where the m – period's back observation y_{tm} is called the m^{th} lag of $\boldsymbol{y}$, c is a $k \times 1$ vector of constants (intercepts), $\boldsymbol{\phi}_i$ is a time-invariant $k \times k$ matrix and $\boldsymbol{e}_t$ is a $k \times 1$ vector of error terms satisfying $\mathrm{E}(\boldsymbol{e}_t) = 0$ – every error term has mean zero; $\mathrm{E}(\boldsymbol{e}_t \boldsymbol{e}_t') = \boldsymbol{\Omega}$ – the corresponding covariance matrix of error terms is $\boldsymbol{\Sigma}_{k \times k}$; $\mathrm{E}(\boldsymbol{e}_t \boldsymbol{e}_{t-k}') = 0$ for any non-zero k – the error terms are independent across time; in particular, there is no serial correlation in individual error terms.
The connection between the VAR model and MD is given:
Let the data be in a matrix form as follows:

$$\boldsymbol{\Gamma} = \begin{bmatrix} Y_0 & Y_{-1} & \cdots & Y_{-P+1} \\ Y_1 & Y_0 & \cdots & Y_{-P+2} \\ \vdots & \vdots & \cdots & \vdots \\ Y_{n-1} & Y_{n-2} & \cdots & Y_{n-p} \end{bmatrix}.$$

The MD can be estimated with the help of the matrix $\boldsymbol{\Gamma}$ as

$$D\left(\frac{\boldsymbol{\Gamma}\boldsymbol{\Gamma}}{n}, \boldsymbol{Y}_i, \overline{\boldsymbol{Y}}\right) = \left(\boldsymbol{Y}_i - \overline{\boldsymbol{Y}}\right)' \left(\frac{\boldsymbol{\Gamma}\boldsymbol{\Gamma}}{n}\right)^{-1} \left(\boldsymbol{Y}_i - \overline{\boldsymbol{Y}}\right),$$

which is the measure of the systematic part of the model. It does not take the error term into account. On the other hand, if one is interested in the error term part, the MD of the error terms could be computed as follows. Let the hat matrix $\mathbf{H}$ be

$$\boldsymbol{H} = \boldsymbol{\Gamma}\left(\boldsymbol{\Gamma}^T\boldsymbol{\Gamma}\right)^{-1}\boldsymbol{\Gamma}^T.$$

Then the estimation of the error term ε is

$$\boldsymbol{R} = (\boldsymbol{I} - \boldsymbol{H})\,\boldsymbol{Y} = (\boldsymbol{I} - \boldsymbol{H})\,(\boldsymbol{\Gamma}\boldsymbol{\phi} + \boldsymbol{\varepsilon}) = (\boldsymbol{I} - \boldsymbol{H})\,\boldsymbol{\varepsilon}.$$

The covariance of the error term is

$$var\,(\boldsymbol{R}) = (\boldsymbol{I} - \boldsymbol{H})\,cov\,(\boldsymbol{\varepsilon}) = (\boldsymbol{I} - \boldsymbol{H})\,\boldsymbol{\sigma}^2.$$

Thus, the MD $D\left((\boldsymbol{I} - \boldsymbol{H})\,\boldsymbol{\sigma}^2, \boldsymbol{R}, \mathbf{0}\right)$ could be obtained with the inverse of the covariance matrix.

5.2 The factor model

Factor analysis is a multivariate statistical method that summarises the observable correlated variables into fewer unobservable latent variables. These unobserved latent variables are also called common factors of the factor model. The factor model can simplify and present the observed variables with much fewer latent variables while still containing most information of a data set. It represents another way of dealing with correlated variables. Further, the factor model offers a method for estimating the covariance matrix and its inverse with the simplified latent variables. We introduce them as follows:

Definition 14. *Let $\boldsymbol{x}_{p\times 1} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be a random vector with known mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The factor model of $\boldsymbol{x}_{p\times 1}$ is*

$$\boldsymbol{x}_{p\times 1} - \boldsymbol{\mu}_{p\times 1} = \boldsymbol{L}_{p\times m}\boldsymbol{F}_{m\times 1} + \boldsymbol{\varepsilon}_{p\times 1},$$

where m is the number of factors in this model, $\boldsymbol{x}$ are the observations (p>m), $\boldsymbol{L}$ is the factor loading matrix, $\boldsymbol{F}$ is an $m \times 1$ vector of common factors and $\boldsymbol{\varepsilon}$ an error term.

The definition above shows the factor model, which represents the random vector $\boldsymbol{x}$ with fewer latent variables. The factor model simplifies the estimation of many statistics such as the covariance matrix. We will introduce the idea as follows: By using Definition 1, we can transform the definition of covariance matrix in the form of $\mathbf{x}_{p\times 1}$ into the covariance matrix in the form of factor model:

Proposition 2. *Let $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \boldsymbol{\Psi})$ where $\boldsymbol{\Psi}$ is a diagonal matrix and $\boldsymbol{F} \sim N(\boldsymbol{0}, \boldsymbol{I})$ are distributed independently so that $Cov(\boldsymbol{\varepsilon}, \boldsymbol{F}) = \boldsymbol{0}$, the covariance structure for $\boldsymbol{x}$ is given as follows:*

$$Cov(\boldsymbol{x}) = \boldsymbol{\Sigma}_f = E(\boldsymbol{LF} + \boldsymbol{\varepsilon})(\boldsymbol{LF} + \boldsymbol{\varepsilon})' = \boldsymbol{LL}' + \boldsymbol{\Psi},$$

where $\boldsymbol{\Sigma}_f$ is the covariance matrix for $\boldsymbol{x}$ under the assumption of a factor model, which generally differs from the classic covariance matrix. The joint distribution of the components of the factor model is

$$\begin{pmatrix} \boldsymbol{LF} \\ \boldsymbol{\varepsilon} \end{pmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{LL}' & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Psi} \end{bmatrix} \right).$$

It must be pointed out that Definition 14 implies the rank of $\boldsymbol{LL}'$, $r(\boldsymbol{LL}') = m \leq p$. Thus, the inverse of a singular matrix $\boldsymbol{LL}'$ is not unique. More details will be discussed later. By using the covariance matrix above, we define the MD on a factor model as follows:

Definition 15. *Under the assumptions in Definition 14, the MD for a factor model with known mean* $\boldsymbol{\mu}$ *is*

$$D\left(\boldsymbol{\Sigma}_f, \boldsymbol{x}_i, \boldsymbol{\mu}\right) = \left(\boldsymbol{x}_i - \boldsymbol{\mu}\right)' \boldsymbol{\Sigma}_f^{-1} \left(\boldsymbol{x}_i - \boldsymbol{\mu}\right),$$

where $\boldsymbol{\Sigma}_f$ *is defined in Proposition 1.*

The way of estimating the covariance matrix from a factor model is different from the classic way. This alternative way makes the estimation of the covariance matrix not only much simpler but also quite informative due to the factor model's properties (Lawley and Maxwell, 1971; McDonald, 2014). Definition 14 shows that a factor model consists of two parts, the systematic part and the residual part. Hence there is an option to build the covariance matrix with the two independent parts separately. By splitting a factor model we can detect the source of the outlier. This is also another part of the thesis that we investigate.

Chapter 6

Future work and unsolved problems

There are several potential research projects related to the MDs in this thesis. First, as we have shown in this thesis, the sample covariance matrix and its inverse do not perform very well under high–dimensional data. Thus, some improved estimators of the inverse sample covariance matrix should be developed in order to find a well approximated estimator. Some work has been done by the author; the results are quite promising. Second, the higher moments of the MDs are still unknown. In this thesis, we focus on their first two moments and the asymptotic distributions. Their higher moments and exact distributions could be undertaken in future studies. Third, this thesis concerns the case of $c = p/n \in (0, 1)$. The $c > 1$ situation can be a topic of further study. Fourth, in this thesis we only derive the point–wise limits on the MDs. Further, the uniform weak limits could be investigated.

Chapter 7

Summary of papers

This thesis investigates the properties of a number of forms of MDs under different circumstances. For high–dimensional data sets, the classic MD does not work satisfyingly because the complexity of estimating the inverse covariance matrix increases drastically. Thus, we propose a few solutions based on two directions: First, find a proper estimation of the covariance matrix. Second, find explicit distributions of MDs with sample mean and sample covariance matrix of normally distributed random variables and the asymptotic distributions of MDs without assumption of normally distributed. Some of the methods are implemented with empirical datasets.

We also combine the factor model with MDs since the factor model simplifies the estimation of both covariance matrix and its inverse for factor–structured data sets. The results offer a new way of detecting outliers from this type of structured variables. An empirical application presents the differences between the classic method and the one we derived.

Besides the estimations, we also investigated the qualitative measures of MDs. The distributional properties, first moments and asymptotic distributions for different types of MDs are derived.

The MDs are also derived for complex random variables. We define the MDs for the real part and the imaginary part of a complex random vector. Their first moments are derived under the assumption of normal distribution. Then we relax the distribution assumption on the complex random vector. The asymptotic distribution is derived with regard to the estimated MD and the leave-one-out MD. Simulations are also supplied to verify the results.

7.1 Paper I: Expected and unexpected values of Mahalanobis distances in high–dimensional data

In Paper I, several different types of MDs are defined. They are built in different forms corresponding to different definitions of means and covariance matrices. The first two moments of MDs are derived. The limits of the first moments reveal some unexpected results such that, in order to find the unbiassed estimator under high–dimensional data sets, there is no unique solution of a constant to make all these MDs asymptotically unbiassed. The reason is that the sample covariance matrix is not an appropriate estimator for the high–dimensional data set. Some asymptotic results of the MDs are also investigated under the high–dimensional set.

The results we get in this paper reveals the need for further investigation of the properties of the MDs under high–dimensional data.

7.2 Paper II: High-dimensional CLTs for individual Mahalanobis distances

In Paper II, we investigate some asymptotic properties of MDs by assuming the sample size n and dimension of variables p go to infinity as $n, p \to \infty$ simultaneously. Their ratio converges into a constant $p/n \to c \in (0, 1)$. Some simulations have been carried out in order to confirm the results.

A duality connection between the estimated MD and the leave-one-out MD is derived. The connection between these two MDs shows a straightforward transformation. The asymptotic distributions for different types of MDs are investigated.

7.3 Paper III: Mahalanobis distances of factor structure data

In Paper III, we use a factor model to reduce the dimensions of the data set and build a factor–structure–based inverse covariance matrix. The inverse covariance matrix estimated from a factor model is then used to construct new types of MDs. The distributional properties of the new MDs are derived. The split–form of MDs based on the factor model is also derived. MDs are used to detect the source of outliers from a factor–structured data set. Detections of the source of outliers are also studied on additive types of outliers. In the last section, the methods are implemented with an empirical study. The results show a difference between the new method and the results from classic MDs.

7.4 Paper IV: Mahalanobis distances of complex random variables

This paper defines some different types of MDs on complex random vectors with considerations of known and unknown mean and covariance matrix. Their first moments and the distributions of MD with known mean and covariance matrix are derived. Further, some asymptotic distributions of the sample MD and leave-one-out MDs under non-normal distribution are investigated. Simulations show a promising result that confirms our derivations.
In conclusion, the MDs on complex random vectors are useful tools when dealing with complex random vectors in many situations, such as outlier detection. The asymptotic properties of MDs we derived could be used in some inferential studies. The connection between the estimated MD and the leave-one-out MD is a contribution due to the special property of the leave-one-out MD. Some statistics that involve the estimated MD could be simplified by substituting the leave–one–out MD. Further study could be developed by the MDs on the real and imaginary parts of a complex random sample with sample mean and sample covariance matrix.

Chapter 8

Conclusions

This thesis has defined eighteen types of MDs. They could be used to measure several types of distances and similarity between the observations in a data set. The explicit first moments in real space for the fixed dimension (n, p) are derived. Then the asymptotic moments are also investigated. By using the asymptotic assumption that as $n, p \to \infty$, the results can be used over some inferential methods when the value of ratio $p/n = c \in (0, 1)$. The results confirm an important conclusion that the sample covariance matrix performs poorly for high dimensional data sets. Their second moments are also derived under the fixed dimension circumstances, which fills a gap in the literature.

Further, our contributions also include the explicit distributions for the MDs under normal distributions in both real and complex spaces. The asymptotic distributions of MDs are also derived for both sample MD and the leave–one–out MD under non–normal distribution. One relationship between the leave-one-out MD and the estimated MD is investigated. The transformation is a substantial tool for some other derivations since the independence of the leave-one-out MD can further simplify the derivations. It shows its preponderance especially under asymptotic circumstances.

We also utilise the factor model to construct the covariance matrix. This factor based covariance matrix is used to build a new type of MD in this thesis. This method makes the estimation simple via classifying the observations or the variables into several fewer numbers of groups. The idea offers a better way when dealing with the structured data. Another new contribution is also made to the detection of outliers with regard to the structured data. The exact outlying distance is also derived with regard to two types of contaminated data sets. This type of MD shed light on the source of an outlier, which has never been considered in literature.

Bibliography

Anderson, T. W. (1951). Classification by multivariate analysis, *Psychometrika* **16**(1): 31–50.

Bai, Z. and Silverstein, J. W. (2010). *Spectral Analysis of Large Dimensional Random Matrices*, Vol. 20, Springer.

Berger, J. (1980). *Statistical decision theory, foundations, concepts, and methods*, Springer series in statistics: Probability and its applications, Springer-Verlag.

Blackwell, D. (1979). *Theory of games and statistical decisions*, Courier Dover Publications.

Brillinger, D. R. (2012). *Asymptotic properties of spectral estimates of second order*, Springer.

Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research*, Guilford Publications.

Cai, T. T., Liang, T. and Zhou, H. H. (2015). Law of log determinant of sample covariance matrix and optimal estimation of differential entropy for high-dimensional Gaussian distributions, *Journal of Multivariate Analysis* **137**: 161–172.

Chave, A. D. and Thomson, D. J. (2004). Bounded influence magnetotelluric response function estimation, *Geophysical Journal International* **157**(3): 988–1006.

De Maesschalck, R., Jouan-Rimbaud, D. and Massart, D. L. (2000). The Mahalanobis distance, *Chemometrics and Intelligent Laboratory Systems* **50**(1): 1–18.

Diciccio, T. and Romano, J. (1988). A review of bootstrap confidence intervals, *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 338–354.

Dyson, F. J. (1962). Statistical theory of the energy levels of complex systems. I, *Journal of Mathematical Physics* **3**(1): 140–156.

Fisher, R. (1936). The use of multiple measurements in taxonomic problems, *Annals of Human Genetics* **7**(2): 179–188.

Fisher, R. A. (1940). The precision of discriminant functions, *Annals of Human Genetics* **10**(1): 422–429.

Friedman, J., Hastie, T. and Tibshirani, R. (2001). *The Elements of Statistical Learning*, Springer Series in Statistics.

Fujikoshi, Y. (2002). Selection of variables for discriminant analysis in a high-dimensional case, *Sankhyā: The Indian Journal of Statistics, Series A* **64**(2): 256–267.

Fujikoshi, Y., Ulyanov, V. and Shimizu, R. (2011). *Multivariate Statistics: High-Dimensional and Large-Sample Approximations*, Vol. 760, Wiley.

Girko, V. L. (2012). *Theory of Random Dterminants*, Vol. 45, Springer Science & Business Media.

Goodman, N. (1963). The distribution of the determinant of a complex Wishart distributed matrix, *The Annals of mathematical statistics* **34**(1): 178–180.

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis, *Biometrika* **53**(3-4): 325–338.

Hastie, T., Buja, A. and Tibshirani, R. (1995). Penalized discriminant analysis, *The Annals of Statistics* **23**(1): 73–102.

Holgersson, H. and Shukur, G. (2001). Some aspects of non-normality tests in systems of regression equations, *Communications in Statistics-Simulation and Computation* **30**(2): 291–310.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components., *Journal of educational psychology* **24**(6): 417.

Khatri, C. (1968). Some results for the singular normal multivariate regression models, *Sankhyā: The Indian Journal of Statistics, Series A* **30**(3): 267–280.

Lawley, D. N. and Maxwell, A. E. (1971). *Factor Analysis as a Statistical Method*, Butterworths.

Leung, C. and Srivastava, M. (1983a). Asymptotic comparison of two discriminants used in normal covariate classification, *Communications in Statistics-Theory and Methods* **12**(14): 1637–1646.

Leung, C. and Srivastava, M. (1983b). Covariate classification for two correlated populations, *Communications in Statistics-Theory and Methods* **12**(2): 223–241.

Lütkepohl, H. (2007). *New Introduction to Multiple Time Series Analysis*, Springer Berlin Heidelberg.

Mahalanobis, P. (1930). On tests and measures of group divergence, **26**: 541–588.

Mahalanobis, P. (1936). On the generalized distance in statistics, **2**(1): 49–55.

Marascuilo, L. A. and Serlin, R. C. (1988). Statistical methods for the social and behavioral sciences.

Mardia, K. (1974). Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies, *Sankhyā: The Indian Journal of Statistics, Series B* **36**(2): 115–128.

Mardia, K. (1977). Mahalanobis distances and angles, *Multivariate analysis IV* **4**(1): 495–511.

Mardia, K., Kent, J. and Bibby, J. (1980). *Multivariate Analysis*, Academic press.

McDonald, R. P. (2014). *Factor Analysis and Related Methods*, Psychology Press.

McLachlan, G. (2004). *Discriminant analysis and statistical pattern recognition*, Vol. 544, John Wiley & Sons.

Mitchell, A. and Krzanowski, W. (1985). The Mahalanobis distance and elliptic distributions, *Biometrika* **72**(2): 464–467.

Osborne, J. W. and Overbay, A. (2004). The power of outliers (and why researchers should always check for them), *Practical assessment, research & evaluation* **9**(6): 1–12.

Pavlenko, T. (2003). On feature selection, curse-of-dimensionality and error probability in discriminant analysis, *Journal of statistical planning and inference* **115**(2): 565–584.

Pedhazur, E. (1997). Multiple regression in behavioral research: Explanation and prediction., *Inc: New York, NY* .

Picinbono, B. (1996). Second-order complex random vectors and normal distributions, *IEEE Transactions on Signal Processing* **44**(10): 2637–2640.

Rao, C. R. (1945). Familial correlations or the multivariate generalisations of the intraclass correlations, *Current Science* **14**(3): P66–67.

Srivastava, S. and Khatri, C. (1979). *An Introduction to Multivariate Statistics*, North-Holland/New York.

Wilks, S. (1963). Multivariate statistical outliers, *Sankhyā: The Indian Journal of Statistics, Series A* **25**(4): 407–426.

Wooding, R. A. (1956). The multivariate distribution of complex normal variables, *Biometrika* **43**(1/2): 212–215.