Essays on Empirical likelihood for Heaviness Estimation, Outlier Detection and Clustering

by

Zhuojing Zhang

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Applied Economics

Waterloo, Ontario, Canada, 2024

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner: Zhonghui Zhang
Assistant Professor
School of Finance
Nanjing Audit University

Supervisor(s): Tao Chen
Associate Professor
Department of Economics
University of Waterloo

Internal Member: Pierre Chaussé
Associate Professor
Department of Economics
University of Waterloo

Thomas Parker
Associate Professor
Department of Economics
University of Waterloo

Internal-External Member: Fan Yang
Assistant Professor
Department of Statistics and Actuarial Science
University of Waterloo

Author's Declaration

This thesis consists of material all of which I co-authored: see Statement of Contributions included in the thesis.

This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

While the first and second chapters are co-authored with my supervisor, Dr. Tao Chen, from the Department of Economics at the University of Waterloo, I have made a major contribution to the work, including the collection (or generation) and analysis of data, mathematical derivation and the writing of manuscripts.

Abstract

Empirical likelihood (EL) is a non-parametric likelihood method of inference, first proposed by Owen (1988). There are a large number of studies about the extensions and applications of EL. Most studies discuss the EL ratio for constructing confidence regions and testing hypotheses, while this thesis focuses on the EL weight assigned to each observation in the dataset by the EL ratio function. This thesis contains three chapters on studying the behaviour and application of EL weights.

Specifically, chapter 1 provides a novel approach based on the EL weights to estimate a threshold that separates the bulk part and tail part of datasets of datasets with a heavy-tailed histogram. Because the transition between the bulk and tail parts can not be fully disjointed in many cases, we allow the threshold to be a random variable instead of a fixed number. In addition, the threshold is relative to a benchmark since heaviness is a relative concept.

In Chapter 2, we focus on outlier detection. We develop an unsupervised method based on EL to identify outliers. In particular, we calculate the EL weights through the EL ratio function with the bootstrap mean constraint and show that the EL weights have different behaviours for datasets with and without outliers. Additionally, the EL weights provide a measure of outlierness for all observations, which might reduce the cost of time.

In Chapter 3, I consider a clustering algorithm based on the EL weights. Clustering is an unsupervised method that aims to group unlabeled datasets based on their similarities. Numerous clustering methods have been proposed. The performance of these methods is typically related to the characteristics of the dataset in the specific applications. The proposed EL weights based clustering algorithm is available to work on datasets with outliers. Moreover, it might suggest the number of clusters for well-separated clusters.

Acknowledgements

I would like to express my deepest gratitude to Dr. Tao Chen, Dr. Pierre Chaussé, Dr. Thomas Parker and Dr. Fan Yang for their excellent guidance and valuable knowledge. Without their support, this thesis would not have been possible. Special thanks to my supervisor, Dr. Tao Chen, for profoundly shaping my PhD journey. His rigorous guidance, profound knowledge, and endless patience have inspired me to become a better version of myself. I've gained invaluable insights from him, encompassing not only academic research but also life-changing experiences. It has been an honour to be his student.

I would also like to extend my thanks to all the members of the faculty, staff, and colleagues from the Economics Department at the University of Waterloo for their assistance. Together, they have contributed to making my experience at Waterloo memorable.

Finally, I express my deepest gratitude to my parents for their unwavering love and encouragement. They have always believed in me and supported me in all my endeavours.

Dedication

To my parents, I dedicate this thesis.

Table of Contents

Examining Committee Membership ii

Author's Declaration iii

Statement of Contributions iv

Abstract v

Acknowledgements vi

Dedication vii

List of Figures xi

List of Tables xiii

Preliminary 1

1 Where Does the Heaviness Start? 6

1.1 Introduction 6

1.2 EL-based algorithm 8

1.3 Simulation studies 10

1.4 An empirical example 14

1.5 Conclusion 17

2 An Algorithm Based on Empirical Likelihood to Identify Outliers 18

2.1 Introduction . . . 18

2.2 Methodology . . . 20

2.2.1 Algorithm . . . 20

2.2.2 The implementation problems . . . 25

2.3 Simulation . . . 26

2.4 Conclusion . . . 29

3 A Clustering Algorithm Based on Empirical Likelihood 31

3.1 Introduction . . . 31

3.2 Methodology . . . 33

3.2.1 Algorithm . . . 33

3.2.2 The implementation problems . . . 35

3.3 Simulation studies . . . 36

3.3.1 Well-separated clusters . . . 36

3.3.2 Non-disjoint clusters . . . 40

3.4 An empirical example . . . 44

3.5 Conclusion . . . 45

References 47

APPENDICES 59

A Appendices of Chapter 1 60

A.1 Additional Hill plots . . . 60

A.2 The plots for simulation results . . . 62

A.3 The EVT method . . . 64

B Appendices of Chapter 2 **70**

B.1 Proof of Lemma 2.2.1 70

B.2 Proof of Theorem 2.2.1 70

B.3 Proof of Lemma 2.2.2 71

B.4 Proof of Theorem 2.2.2 72

B.5 Proof of Theorem 2.2.3 72

B.6 Proof of Theorem 2.2.4 72

B.7 Proof of Theorem 2.2.5 72

B.8 Proof of Remark 2.2.1 73

B.9 Skewness of EL weigths 73

B.10 Plot of EL weights 74

B.11 Additional simulations 75

C Appendices of Chapter 3 **86**

C.1 The difference plots for well-separated clusters 86

C.2 Results for mixture Gaussian with different parameters 88

C.3 The hierarchical clustering dendrogram 88

C.4 Additional results for mixture Von Mises-Fisher case using DBSCAN 89

C.5 K-based methods for the iris dataset 92

C.6 Hierarchical clustering for the iris dataset 94

C.7 DBSCAN for the iris dataset 97

List of Figures

1.1 An example of the Hill plot 11

1.2 Histograms of household-level total asset values 15

1.3 The log-log plot of household-level total asset values 15

1.4 The Hill plot of the SIPP dataset 16

3.1 The plots of mixture of Gaussian distributions 37

3.2 The plots of the 3D atom dataset 38

3.3 The plots of the mixture of Von Mises-Fisher distributions 39

3.4 Results for the 2D spherical dataset using k-means and k-medoids 40

3.5 Results for the 2D spherical dataset using hierarchical clustering with average and complete linkages 41

3.6 Results for the 2D spherical dataset using DBSCAN 42

3.7 Results for the 2D spherical dataset using hierarchical clustering with single linkage 42

3.8 The plots of non-disjoint clusters with the same sample size and densities . . 43

3.9 The plots of non-disjoint clusters with different sample sizes and densities . . 43

3.10 The scatterplot of the iris dataset 45

3.11 The results of the iris dataset 46

A.1 The Hill plots for Mixed distribution 61

A.2 The Hill plots for Pareto(6) 62

A.3 The Hill plots with the selected region 63

A.4 Plots for simulation data with exponential benchmark 64
A.5 Plots for simulation data with other benchmarks 65

B.1 The plot of EL weights of the 1D dataset with outliers 75
B.2 The plot of EL weights of the 1D dataset without outliers 76
B.3 The plot of EL weights of the 2D dataset with outliers 76
B.4 The plot of EL weights of the 2D dataset without outliers 77
B.5 Plots of EL weights for additional 1D cases with outliers 78
B.6 Plots of EL weights for additional 1D cases without outliers 78
B.7 Plots of EL weights for additional 2D cases with outliers 82
B.8 Plots of EL weights for additional 2D cases without outliers 82

C.1 The difference plot for the mixture Gaussian case 87
C.2 The difference plot for the 3D atom dataset 87
C.3 The difference plot for the mixture Von Mises-Fisher case 88
C.4 The plots for well-separated Mixture Gaussian cases 89
C.5 The dendrogram with complete linkage 90
C.6 The dendrogram with average linkage 90
C.7 The dendrogram with single linkage 91
C.8 Additional results for mixture Von Mises-Fisher case using DBSCAN (1) . . 92
C.9 Additional results for mixture Von Mises-Fisher case using DBSCAN (2) . . 93
C.10 Additional results for mixture Von Mises-Fisher case using DBSCAN (3) . . 94
C.11 Additional results for mixture Von Mises-Fisher case using DBSCAN (4) . . 95
C.12 Results for the iris dataset using K-based methods 96
C.13 The dendrogram for the iris dataset with complete linkage 97
C.14 The dendrogram for the iris dataset with average linkage 98
C.15 The dendrogram for the iris dataset with single linkage 98
C.16 Results for the iris dataset using hierarchal clustering methods 99
C.17 Results for the iris dataset using DBSCAN with $MinPts = 5$ 100
C.18 Results for the iris dataset using DBSCAN with $MinPts = 10$ 101

List of Tables

1.1 Simulation results for the variance restriction 12
1.2 Simulation results for the multiple restrictions 13
1.3 Threshold selection by approaches from EVT 14
1.4 Summary statistics of SIPP total asset values 14
1.5 Summary statistics of K 16

2.1 P-value of the outliers identification test for 1D cases 27
2.2 P-value of the outliers identification test for 2D cases 28
2.3 Kurtosis of EL weights for 1D cases 29
2.4 Kurtosis of EL weights for 2D cases 30

3.1 The parameters of three bi-variate Gaussian clusters 36

B.1 Skewness of EL weights for 1D cases 74
B.2 Skewness of EL weights for 2D cases 74
B.3 P-value of the outliers identification test for additional 1D cases 79
B.4 Skewness of EL weights for additional 1D cases 80
B.5 Kurtosis of EL weights for additional 1D cases 81
B.6 P-value of the outliers identification test for additional 2D cases 83
B.7 Skewness of EL weights for additional 2D cases 84
B.8 Kurtosis of EL weights for additional 2D cases 85

Preliminary

1. Introduction

Likelihood method is an effective and flexible method which is widely used in Economics. For example, maximum likelihood estimation is a general method for measuring the parameters of economic models, as well as for inference and testing hypotheses. However, one problem of parametric likelihood methods is the requirement for assumptions of the distribution of the models.

Then, non-parametric likelihood methods are developed to avoid the problem of model misspecification. Empirical likelihood (EL), first introduced by Owen (1988), is a non-parametric likelihood method that captures some advantages of likelihood methods without assuming the distribution of the dataset. Then, some creative works, such as Owen (1990), Qin and Lawless (1994), DiCiccio and Romano (1989), DiCiccio et al. (1991), Newey and Smith (2004) and Hjort et al. (2009), built good properties and greatly improved the implementation of EL. In this section, we briefly introduce the EL, while comprehensive reviews of EL are referred to Owen (2001) and Lazar (2021).

Let $X_1, \dots, X_n \in \mathbb{R}$ be the independent and identical distributed random variable from an unknown distribution F with mean μ. The empirical likelihood function is

$$L(F) = \prod_{i=1}^{n} w_i,$$

where w_i is the weight that the distribution F places on x_i with $w_i \geq 0, \sum_{i=1}^{n} w_i = 1$. Then, the empirical likelihood ratio function is

$$R(F) = \frac{L(F)}{L(F_n)} = \prod_{i=1}^{n} n w_i,$$

where F_n is the empirical distribution function such that put $1/n$ to each observation.

Owen (1988) considered the mean μ as the parameter of interest. The profile empirical likelihood ratio function for the μ is defined as

$$R(\mu) = \max\left\{\prod_{i=1}^{n} nw_i | \sum_{i=1}^{n} w_i(X_i - \mu) = 0, w_i \geq 0, \sum_{i=1}^{n} w_i = 1\right\}. \tag{1}$$

The empirical likelihood ratio statistic $-2\log R(\mu)$ asymptotically follows χ_1^2 under some mild conditions, serving as a non-parametric version of the Wilks' theorem (Owen, 1988, 1990).

Moreover, Qin and Lawless (1994) combined the empirical likelihood with estimating equations. Consider random vectors, $X_1, ..., X_n \in \mathbb{R}^d$, and the parameter vector $\theta \in \mathbb{R}^p$ such that $\mathbb{E}[m(X,\theta)] = 0$, where $m(X,\theta) \in \mathbb{R}^s$ is the estimating equation. The profile empirical likelihood ratio for θ is written as

$$R(\theta) = \max\left\{\prod_{i=1}^{n} nw_i | \sum_{i=1}^{n} w_i m(X,\theta) = 0, w_i \geq 0, \sum_{i=1}^{n} w_i = 1\right\}. \tag{2}$$

Similar to the result on $-2\log R(\mu)$, the empirical likelihood ratio statistic $-2\log R(\theta)$ asymptotically follows χ_p^2 under some mild conditions.

We should note that the selection of $m(X,\theta)$ depends on the research problems and the prior information. For example, if we have the information about the mean and variance of a variable, that is, $\mathbb{E}(X) = \mu$ and $Var(X) = \sigma^2$, we could take

$$m(X,\theta) = (X - \mu)^2 - \sigma^2,$$

where $\theta = (\mu, \sigma)$. When considering the regression model for $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}$, we might take

$$m(X,\theta) = X(Y - X'\beta),$$

where $\theta = \beta$ is the regression parameter.

Now, to solve (2), which is a maximization problem, the Lagrange multiplier method is employed and then gives the weight function

$$w_i = \frac{1}{n}\frac{1}{1 + \lambda^T m(X,\theta)}, \tag{3}$$

where the λ is the Lagrange multiplier satisfying the following non-linear equation,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{m(X,\theta)}{1 + \lambda^T m(X,\theta)} = 0.$$

However, the solution for (2) might not exist. Hence, the adjustment of EL has been studied (Chen et al., 2008; Grendár and Judge, 2009; Liu and Chen, 2010). Besides the adjusted empirical likelihood, there are many variants of empirical likelihood, including the bayesian empirical likelihood (Lazar, 2003; Schennach, 2007; Zhong and Ghosh, 2016), the weighted empirical likelihood (Wu, 2004; Glenn and Zhao (2007)), and the jackknife empirical likelihood (Jing et al., 2009; Gong et al., 2010; Cheng et al., 2018). A recent review for variants of EL is referred to Liu and Zhao (2022).

2. Literature review

Some good properties of EL have been established, such as the ability to handle complex data structures and provide robust inference without specific assumptions about the distribution of datasets. Thus, EL has been widely used in many problems.

Owen (1991) considered the EL for regression which is an important tool that might help to analyze the relationships between variables or predict the behaviour of variables. Chen (1993, 1994) studied the higher-order properties of EL inference for the regression coefficients and the linear combinations of regression coefficients. La Rocca (1998) investigated the coverage error of empirical likelihood for linear regression models using a bootstrap method. The coverage error of EL for the non-parametric regression is studied by Chen and Qin (2000) with a local linear kernel estimator and by Chen and Qin (2003) with a Nadaraya-Watson estimator. The EL inference for the non-parametric linear regression with missing data is investigated in Wang and Rao (2002). Whang (2006) considered the EL for quantile regression and discussed the Bartlett correctable of smoothed EL.

Moreover, Kolaczyk (1994) extended EL to generalized linear models, introduced by Nelder and Wedderburn (1972), with a fixed amount of overdispersion. Song and Cui (2003) proposed an extended empirical likelihood with incorporating extra constraints to improve the efficiency of estimation for generalized linear models. The missing data problem in the generalized linear models is studied by Xue et al. (2011), which proposed the bias-correction approaches using the inverse selection probability weighted imputation technique. Zang et al. (2017) applied EL to test both the entire set of regression coefficients and a part of the regression coefficients in high dimensional generalized linear models.

The regression with censored data has also been discussed. Qin and Jing (2001) studied the EL for censored linear regression with random censoring, while Wang et al. (2011) and Cai and Chen (2018) discussed the EL inference for the censored data with fixed censoring points. Additionally, many researchers considered the right censored regression data which is often found in survival data. Qin and Tsao (2003) investigated EL inference for the median regression model of the right censored data based on the normal-approximation-based method. Li and Wang (2003) proposed an adjusted empirical likelihood for regression coefficients under the right censored data, as well as the linear combinations of the regression coefficients.

Using EL for the problems associated with the use of instrumental variables was also discussed. Guggenberger and Smith (2005) considered the problem of instrumental variable models in regression using EL. Anatolyev (2005), Smith (2011) and Chang et al. (2015) investigated the higher order properties of EL in the case of weakly dependent processes. Zhao and Xue (2013) and Sun et al. (2019) constructed the confidence intervals for the varying-coefficient functions by using a proposed instrumental variable-based empirical.

Although EL was first designed to deal with independent observations, it has been extended to dependent observations. Monti (1997) developed EL to time series models based on Whittle's estimation method. Then, Nordman and Lahiri (2006) devised a frequency domain empirical likelihood based on the periodogram and spectral estimating equations to both short and long range dependence. Baragona et al. (2013) focused on structural breaks in regression and auto-regression models based on EL. Furthermore, a goodness-of-fit statistic based on the EL ratio for parametric time series regression models is proposed by Chen et al. (2003). A review of EL for time series is referred to Nordman and Lahiri (2014).

Other than time series data, there are also many works on EL for longitudinal data. Tang and Leng (2011) and Wang and Zhu (2011) were interested in EL for quantile regression with longitudinal data. Bai et al. (2010) studied a weighted EL for generalized linear models with longitudinal data, and showed that the efficiency of the weighted EL inference is affected by the selection of weights. Tian and Xue (2014) constructed the EL inference for the generalized linear model with longitudinal data by combining the EL method with the quadratic inference function. Recently, Yin et al. (2023) considered the EL inference for fixed design generalized linear model.

3. Outline of the thesis

In this thesis, we study the behaviour of EL weights assigned to each observation by (2), and investigate the applications of EL weights.

Specifically, chapter 1 provides a novel approach based on the EL weights to estimate a threshold that separates the bulk part and tail part of datasets of datasets with a heavy-tailed histogram. Because the transition between the bulk and tail parts can not be fully disjointed in many cases, we allow the threshold to be a random variable instead of a fixed number. In addition, the threshold is relative to a benchmark since heaviness is a relative concept.

In Chapter 2, we focus on outlier detection. We develop an unsupervised method based on EL to identify outliers. In particular, we calculate the EL weights through the EL ratio function with the bootstrap mean constraint and show that the EL weights have different behaviours for datasets with and without outliers. Additionally, the EL weights provide a measure of outlierness for all observations, which might reduce the cost of time.

In Chapter 3, I consider a clustering algorithm based on the EL weights. Clustering is an unsupervised method that aims to group unlabeled datasets based on their similarities. Numerous clustering methods have been proposed. The performance of these methods is typically related to the characteristics of the dataset in the specific applications. The proposed EL weights based clustering algorithm is available to work on datasets with outliers. Moreover, it might suggest the number of clusters for well-separated clusters.

Chapter 1

Where Does the Heaviness Start?[1]

1.1 Introduction

Datasets with a heavy-tailed histogram tend to have a large number of outliers that provide important information. In many fields, datasets are characterized by this feature, such as Psychology (Barabási, 2005; Malmgren et al., 2008), Economics (Rossi-Hansberg and Wright, 2007; Giesen and Sudekum, 2010), Finance (Mandelbrot, 1963; Gabaix et al., 2003; Gabaix, 2009), Statistics (Richardson, 1948; Clauset and Woodard, 2013; Cirillo and N., 2016) and Hydrology (Anderson and Meerschaert, 1998; Katz et al., 2002; Bernardara et al., 2007), to name a few. One important challenge for analyzing datasets with this feature is that the behaviours of the bulk part and tail part are different. For example, it is generally accepted that Pareto distributions are useful when describing the distributions of high incomes, which are represented in the tail parts of income datasets. However, Pareto distributions perform poorly over the whole range of incomes (Reed, 2003). Pareto distributions have infinite variance when the shape parameter is smaller than two, and the usual least square method cannot be used directly when datasets have infinite variance (Kanter and Steiger, 1974). Thus, the properties of one part of the dataset can affect the choice of method and mislead the analysis of the other part or the overall pattern, leading to questions about which part of the dataset should be dropped to study non-extreme events or be used to predict rare events. We loosely use the word “threshold” to denote the solution. Because the transition between the bulk and tail parts can not be fully disjointed in many cases, we allow the threshold to be a random variable instead of a fixed number.

[1] This chapter is co-authored with Tao Chen.

We might borrow approaches from Extreme Value Theory (EVT) to estimate the threshold. In EVT, a tail parameter k, which is the number of upper observations used to estimate the tail part of the dataset, is a possible threshold. It is possible to choose k by detecting the change of slope in the mean excess plot (Embrechts et al., 1997) or the first "stable" region of the Hill plot which is based on the hill estimator (Hill, 1975 and Drees et al., 2000). However, the most critical issue of these graphical diagnostics is that the results are subjective. A glance ahead to Figure 1.1 will indicate why this is so. To avoid these problems, completely programmed estimators that can automatically choose the k are widely studied (Caeiro and Gomes, 2016). Hall and Welsh (1985) derived a formula by minimizing the asymptotic mean squared error (AMSE) of the Hill estimator to find the optimal k. However, it requires extra knowledge of an unknown second-order parameter. The estimation of the second-order parameters has been studied in several papers, like Gomes and Pestana (2007) and Caeiro and Gomes (2014). Bootstrap methods based on the minimization of the AMSE criterion are also developed due to the need to know the prior knowledge of the second order parameter (Hall, 1990; Draisma et al., 1999; Danielsson et al., 2001; Gomes and Oliveira, 2001; Gomes et al., 2012). However, the goal of these AMSE minimization approaches is to find an optimal k that balances the bias and variance of the tail index estimator which focuses on the tail part. They are not designed to estimate the transition region between the bulk and tail parts.

Besides these, some papers compared the empirical distribution of data above a threshold k with the fitted generalized Pareto distribution (GPD) by using some goodness-of-fit tests (Northrop and Coleman, 2014; Bader et al., 2018; Schneider et al., 2021), others minimized the distance between the L-moments of the datasets and the fitted GPD (Silva Lomba and Fraga Alves, 2020) or a standard Exponential distribution (Kiran and V., 2021). These approaches are also not designed to study the transition region. They aim to find a k such that the empirical distribution of data above the k certainly fits the GDP. In EVT, k is reasonable as long as the tail part gives enough information to estimate the tail index, so the bulk part is ignored and a fixed k is acceptable. But, the threshold is a fixed number only if the bulk part and tail part can be fully separated, which is much less likely to happen. Otherwise, we cannot find a fixed number threshold, because the transition between the two parts is gradual. Although extreme value mixture models, which combine a model of the bulk distribution with an extreme value tail model, treat k as a parameter and use the information from the bulk part, the major purpose of these models is also not to study the threshold. Moreover, these models might lack of robustness of the bulk and tail fits to each other and are the computational complexity and implementation difficult (Scarrott and MacDonald, 2012).

To study the transition, we propose an approach based on the Empirical Likelihood

(EL) method. In the transition region, the likelihood of a variable falling in the tail part increases with the value of the variable, whereas the likelihood of a variable falling in the bulk part increases as the value of the variable decreases. Naturally, the threshold can be viewed as a random variable K with a density function representing the likelihood of possible values of K that separate the bulk and tail parts.

EL, as a non-parametric method (Owen, 1988; Owen, 1990; Imbens (2002)) is a good candidate for analyzing datasets with a heavy-tailed histogram, since these datasets are usually not amenable to a known distribution. EL assigns weights to each observation of the sample dataset without parametric assumptions through an empirical likelihood ratio function under certain constraints. By imposing the right restrictions, we expect these weights to shed light on the transition region and indicate where the heaviness of a distribution starts.

In addition, we view the heaviness as a relative concept. Many concepts only make sense when relative to a benchmark. For example, what height is considered tall? The answer to this question is very subjective without a benchmark. Also, the choice of benchmark should align with the specific characteristics and goals of the research. Therefore, our approach allows researchers to choose benchmarks based on their specific research questions. For example, in environmental science, the choice of benchmark may depend on the type of pollutant or environmental variable being studied. In finance, different benchmarks might be used to assess or compare tail risk in various financial products or asset classes.

The rest of the paper is structured as follows. Section 1.2 describes the methodology proposed. Section 1.3 reports and analyzes the simulation results. Section 1.4 presents an application of our method to an empirical dataset. Section 1.5 concludes the study.

1.2 EL-based algorithm

We model the transition between the bulk and tail parts through the empirical likelihood ratio function (2), i.e.,

$$\max_{w_1,\ldots,w_n} \left\{ \prod_{i=1}^{n} w_i \,\middle|\, \sum_{i=1}^{n} w_i m(X_i,\theta) = 0, w_i \geq 0, \sum_{i=1}^{n} w_i = 1 \right\},$$

where $m(X,\theta)$ is a set of appropriately chosen restrictions that link the dataset of interest to the benchmark. We should provide the intuition behind our procedure: by aligning the “center” and “spread” of the benchmark and data of interest, the weights implied by the pre-specified constraints disclose the relative rareness of realizations with similar values.

Accordingly, the variance constraint is used in this chapter. Other constraints related to the “spread” of the datasets also work. As an exposition, we pick the benchmark to be an Exponential random variable and we only focus on the right tail. Our method can be summarized as follows:

(a) From a random sample of F, we obtain its estimated mean ($\hat{\mu}$) and variance ($\hat{\sigma}^2$).

(b) Under the “variance restriction”, the empirical likelihood ratio function becomes

$$\max_{w_1,\dots,w_n} \left\{ \prod_{i=1}^{n} w_i \middle| \sum_{i=1}^{n} w_i \left(X_i - \hat{\mu}\right)^2 = \hat{\sigma}^2, w_i \geq 0, \sum_{i=1}^{n} w_i = 1 \right\},$$

and the weight function for X_i's are

$$w_i = n^{-1} \left(1 + \lambda \left[\left(X_i - \hat{\mu}\right)^2 - \hat{\sigma}^2 \right] \right)^{-1}, \tag{1.1}$$

where the Lagrange multiplier λ can be found by numerical search. Denote the weights of X_i's by $\{w^X\}_{i=1}^n$.

(c) Simulate a random sample $Y_1, \dots, Y_n$ from an Exponential distribution with mean being $\hat{Q}/\ln 2$, where $\hat{Q}$ is the sample median of X_i's.

(d) Calculate weights of Y_i's by substituting Y_i's, its true mean, $\hat{Q}/\ln 2$ and true variance, $(\hat{Q}/\ln 2)^2$, into the left side of equation (1.1), denoted as $\{w^Y\}_{i=1}^n$.

(e) Sort $\{w^X\}_{i=1}^n$ and $\{w^Y\}_{i=1}^n$ in ascending order, respectively. Denote the sorted weights as $\left\{w_{(i)}^X\right\}_{i=1}^n$ and $\left\{w_{(i)}^Y\right\}_{i=1}^n$.

(f) Find the set of crossing points I^c between $\left\{w_{(i)}^X\right\}_{i=1}^n$ and $\left\{w_{(i)}^Y\right\}_{i=1}^n$ by collecting the index numbers j such that the sign of $w_{(j)}^X - w_{(j)}^Y$ differs from the sign of $w_{(j+1)}^X - w_{(j+1)}^Y$, for $j = 1, \dots, n-1$.

(g) Select the minimum value k^c of I^c that is bigger than $n/2$.

(h) Repeat (c)-(g) m times.

This procedure gives $\{k_1^c, \dots, k_m^c\}$ that is the set of all possible values of the threshold K which models the transition region, denoted as $R(K)$. Now we are ready to address all the assumptions we have imposed through the following remarks.

Remark 1.2.1. *This algorithm does not require F itself to have μ and/or σ because the sample mean and variance, which are well-defined, can serve as the same device. Also, we can replace the "variance restriction" with restrictions in terms of a combination of low and high quantiles, which always exist.*

In the chapter, the mean and variance are estimated by the bootstrap method.

Remark 1.2.2. *The median of the benchmark sample Y_i's is equal to the median of the random sample X_i's.*

The choice of the benchmark should be based on the researchers' interests and sample datasets. The next two remarks are from the perspective of numerical stability, and we drop some simulated samples if the conditions within are not met.

Remark 1.2.3. *As the weight function is monotone in X (Y) when exceeding its mean, the maximum crossing point is well-defined. We let k^c be the smallest value that is bigger than $n/2$.*

Remark 1.2.4. *The $\left\{w^Y_{(i)}\right\}_{i=1}^n$ is comparable to $\left\{w^X_{(i)}\right\}_{i=1}^n$ only if the sign of λ in the weight function of X_i's equals to the sign of λ in the weight function of Y_i's.*

In many studies, researchers would prefer a simplified version of K, e.g., a representative number, denoted by τ, from K. There are many possible choices for τ and in the current paper,

Remark 1.2.5. *τ is defined to be the average of K.*

Further inference is now feasible because τ has variance attached to it.

1.3 Simulation studies

We focus on the heaviness of two types of distributions: Pareto distribution and mixed distribution, which is a linear combination of Pareto and Normal distributions. A Pareto distribution is defined by its scale and shape parameters. Here, we fix the scale parameter to be 1 and will only change the value of the shape parameter.

We consider four cases, each with 10000 observations. Of the first two, one involves Pareto distribution where the mean and variance exist (shape = 6) and the other focuses on truncated Pareto distribution. The truncated Pareto distribution is a truncated version of the Pareto distribution, which is truncated at 0.99 quantile of the Pareto distribution with shape = 1.5. We denote them as Pareto(6) and TPareto(1.5), respectively. For a Pareto distribution, the tail index is the inverse of the value of the shape parameter. Thus, the tail index γ is around 0.1667 for Pareto(6) and around 0.667 for TPareto(1.5). The other two cases are mixed distributions: 90% of a Pareto(6) or TPareto(1.5) and 10% of a Normal random variable. Using Pareto(6) as an example, since we are interested in the right tail, we let the 40^{th} and 45^{th} percentile values of the Normal distribution equal the same percentiles of the Pareto(6) so that the shape of the right part of the mixed distribution is not distorted. We denote this mixed distribution as Mixed(6). Similarly, Mixed(1.5) for the other mixed distribution.

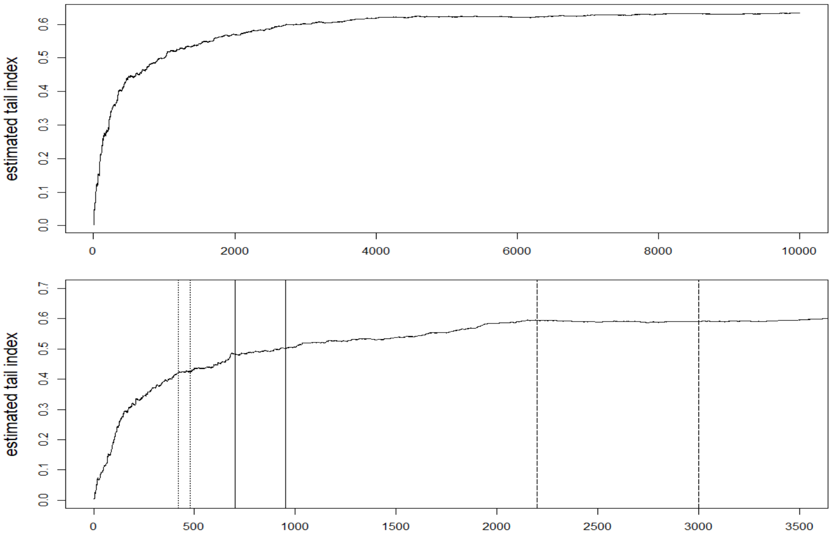


Figure 1.1: An example of the Hill plot

Figure 1.1 provides an example of a TPareto(1.5) Hill plot to explain why graphical methods can be subjective. Three intervals, $[420, 550]$, $[700, 950]$ and $[2200, 3000]$, are indicated in the bottom panel of Figure 1.1 with different types of lines. All of these three intervals can be selected as the first "stable" region under different standards. Thus, the result computed from the first "stable" region is not objective. With substantial expertise, it is possible to choose an acceptable unique solution. For example, the minimum value of the largest intervals $[2200, 3000]$ can be selected. However, it requires expert experience and is time-consuming when there are many datasets. More Hill plots for Pareto distribution and Mixed distribution are provided in Appendix A.1.

The results for the four cases with 1000 replicates can be found in the first four rows of Table 1.1, and are illustrated in A.2. In this chapter, $\hat{\tau}$ is selected as the mean of K and $m = 100$. To make the value of $\hat{\tau}$ easier to interpret, we present it in a percentage format. For example, $\hat{\tau}$ is 86.1% for Pareto(6), meaning that the heaviness starts at the top 13.9% of Pareto(6), that is the 0.861 quantile of the Pareto(6) distribution. The difference between the values of $\hat{\tau}$ for Pareto and mixed distributions is very small (less than 0.31%), since the heaviness is only driven by the Pareto component. As a side result, the table also presents the estimated tail-index, $\hat{\gamma}$, by using $\hat{\tau}$ as the tail parameter in Hill's tail-index estimator. Given the true value of the tail index, we calculate the mean squared error (MSE) of $\hat{\gamma}$. The last column of Table 1.1 shows that the MSE for all cases is fairly small. Though it is not our goal in this paper to propose another tail parameter estimator, which is one of the focal points in EVT, our approach can be viewed as an addition to that literature. For the sake of completeness, we leave the preliminary information on EVT in Appendix A.3.

Table 1.1: Simulation results for the variance restriction

Target distribution	Benchmark	$\hat{\tau}$	$\hat{\gamma}$	MSE
Pareto(6)	Exponential	86.10%	0.1694	0.00002
Mixed(6)	Exponential	86.24%	0.1562	0.00018
TPareto(1.5)	Exponential	84.97%	0.5477	0.01428
Mixed(1.5)	Exponential	85.28%	0.5185	0.02205
TPareto(1.5)	Pareto(6)	86.67%	0.4865	0.03258
Beta(2,5)	Exponential	81.20%	0.0603	0.00436

Next, we set Pareto(6) instead of an Exponential distribution as the benchmark for TPareto(1.5). In Table 1.1, the value of $\hat{\tau}$ is 84.97% when the target is TPareto(1.5) and the benchmark is an Exponential distribution, whereas the value of $\hat{\tau}$ is 86.67% when the benchmark is Pareto(6). The tail of Pareto(6) is heavier than the tail of Exponential distribution, therefore, the heaviness starts at a higher quantile. Then, the $\hat{\gamma}$ for TPareto(1.5) is calculated by setting the tail parameter in Hill's tail-index estimator to be 84.97% and 86.67%, respectively. More interestingly, when the benchmark is an Exponential distribution, the estimated shape parameter is 1.826, which implies that the distribution has an infinite variance; whereas, the estimated shape parameter is 2.055, meaning that the distribution has a finite variance when the benchmark is Pareto(6).

Note that if we force TPareto(1.5) to be the benchmark and Pareto(6) to be the distribution of interest, we get 86.67% back. The "role reversal" property of our algorithm ensures relative heaviness. In addition, the tail index of the target distribution and benchmark are both allowed to be non-positive. For example, we can force a Beta distribution with two shape parameters $(2, 5)$ which has a negative tail index, denoted by Beta(2,5), to be the target distribution and an Exponential distribution which has a zero tail index to be the benchmark. Then, it gives the threshold K of the Exponential distribution with benchmark Beta(2,5) by the "role reversal" property, since the tail of the Exponential distribution is heavy relative to a Beta distribution. The results are shown in the last row of Table 1.1. The value of $\hat{\tau}$ is 81.20% and $\hat{\gamma}$ is 0.0603 which is close to 0. Note that the $\hat{\gamma}$ in this case is the estimated tail-index of the Exponential distribution and is calculated by using the Moment tail-index estimator since the Hill estimator only works for $\gamma > 0$.

Table 1.2: Simulation results for the multiple restrictions

Target distribution	Benchmark	$\hat{\tau}$	$\hat{\gamma}$	MSE
Pareto(6)	Exponential	66.68%	0.1678	0.00001
Mixed(6)	Exponential	67.61%	0.1529	0.00018
TPareto(1.5)	Exponential	72.04%	0.5899	0.26473
Mixed(1.5)	Exponential	71.18%	0.5496	0.26474
TPareto(1.5)	Pareto(6)	79.39 %	0.5711	0.00926
Beta(2,5)	Exponential	71.05%	0.0822	0.00724

The choice of restriction of the EL ratio function is also based on the specific research questions and will affect the result. We add a median restriction to the empirical likelihood ratio function. The median restriction can be written as $m(X, p, q) = 1_{X_i \le q} - p$, where $-\infty < q < \infty$ and $0 < p < 1$. Then, the results for all cases by using the median and variance restriction can be found in Table 1.2. Again, the difference between the values of $\hat{\tau}$ for Pareto and mixed distributions is very small. The values of $\hat{\tau}$ by using the median and variance restriction are all smaller than the values of $\hat{\tau}$ by using only variance restriction, since a median restriction is not sensitive to the presence of outliers.

Although $\hat{\tau}$ and k are two different concepts, we still present the value of k in Table 1.3. Two approaches are considered: 1. method by Caeiro and Gomes (2016), denoted as DAMSE; 2. method by Gomes et al. (2012) and Caeiro and Gomes (2016), denoted GC. DAMSE and GC perform best for Pareto(6) which is a standard Pareto distribution with finite variance, because DAMSE and GC depend on the AMSE minimization of the Hill estimator.

Table 1.3: Threshold selection by approaches from EVT

Target distribution	Method	k	$\hat{\gamma}$
Pareto(6)	DAMSE	81.32%	0.1660
Mixed(6)	DAMSE	50.84%	0.1525
TPareto(1.5)	DAMSE	98.38%	0.2756
Mixed(1.5)	DAMSE	96.92%	0.3634
Pareto(6)	GC	90.26%	0.1659
Mixed(6)	GC	25.30%	0.1528
TPareto(1.5)	GC	94.85%	0.4343
Mixed(1.5)	GC	78.07%	0.5557

1.4 An empirical example

We apply our approach to a U.S. household asset value dataset, which draws from the 2018 Survey of Income and Program Participation (SIPP) and contains 743,753 observations. The SIPP, sponsored by the U.S. Census Bureau, collects information on households' economic status, such as assets and liabilities. Table 1.4 presents descriptive statistics of the household-level total asset values. The top panel of Figure 1.2 shows a histogram of the complete dataset and the bottom panel presents the values up to 6 million to make the histogram's trends easier to see.

Table 1.4: Summary statistics of SIPP total asset values

Minimum	Mean	Maximum	Skewness	Kurtosis
1	690,824	142,631,700	21	897

From the table and histograms, we see that total asset values tend to have a large number of observations with large values and its kurtosis is fairly large; that is, the dataset is heavy-tailed. Figure 1.3 shows the log-log plot of household-level total asset values. The horizontal axis is the log value of household-level total asset values and the vertical axis is the log value of its survival probability. The tail part of the log-log plot is almost a straight line, which suggests that this portion of the dataset is a member of the Pareto distribution family.

Again, we set Exponential distribution to be the benchmark and align the median. The summary statistics of K are reported in Table 1.5 with $m = 1000$ and Figure 1.4 is the

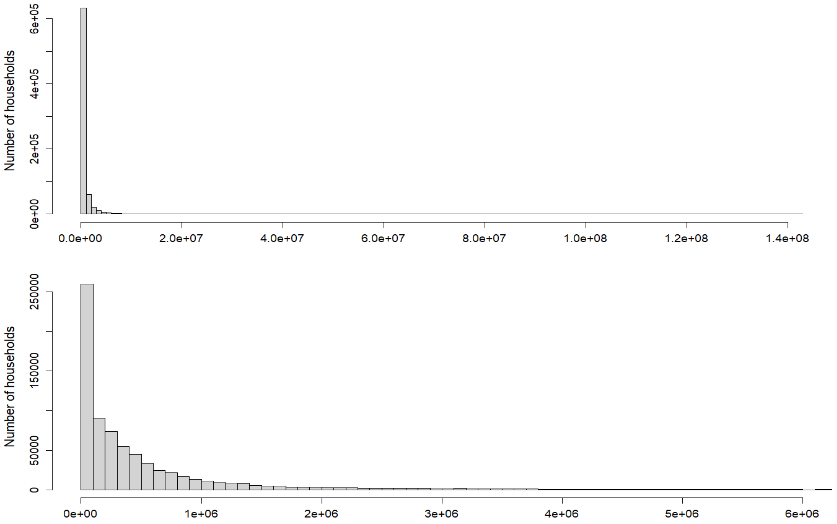


Figure 1.2: Histograms of household-level total asset values

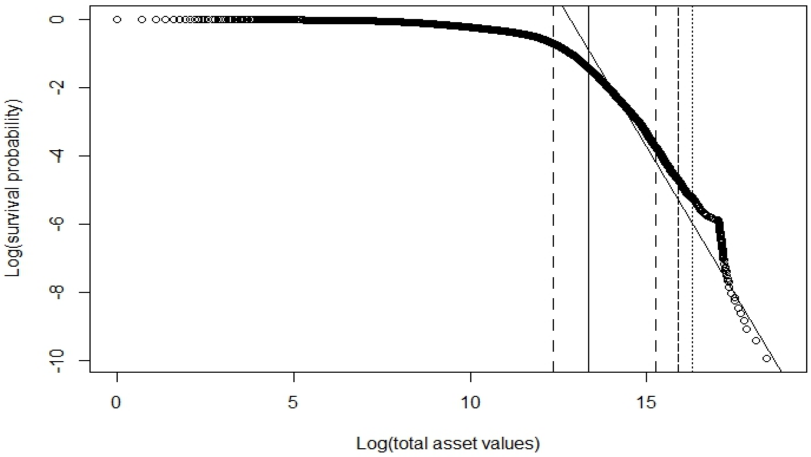


Figure 1.3: The log-log plot of household-level total asset values

corresponding Hill plot. Two intervals, $[2700, 4200]$ and $[5500, 7000]$, are chosen as the "first" possible stable region, where one is between dotted lines, and the other is between two dashed lines. The two intervals are $[99.44\%, 99.64\%]$ and $[99.06\%, 99.26\%]$ in terms of

Table 1.5: Summary statistics of K

Min.	1st Qu.	$\hat{\tau}$	Median	3rd Qu.	Max.
50.04%	75.65%	75.66%	76.28%	76.57%	97.65%

percentiles. Any numbers in those two intervals are admissible choices of the threshold for the EVT method. If we denote the middle points by $k_{\text{dotted}} = 99.54\%$ and $k_{\text{dashed}} = 99.16\%$ and together with $\hat{\tau} = 75.56\%$ and $R(K) = [50.04\%, 97.65\%]$, five vertical lines (dotted, dashed, solid and two long dashes) are added to the aforementioned log-log plot, as shown in the Figure 1.3. In addition, the k is 99.50% by using DAMSE and is 97.52% by using GC.

An encouraging finding is that the solid line and longdash lines, which represent $\hat{\tau}$ and K are close to the turning point, beyond which Pareto distribution becomes a good approximation.

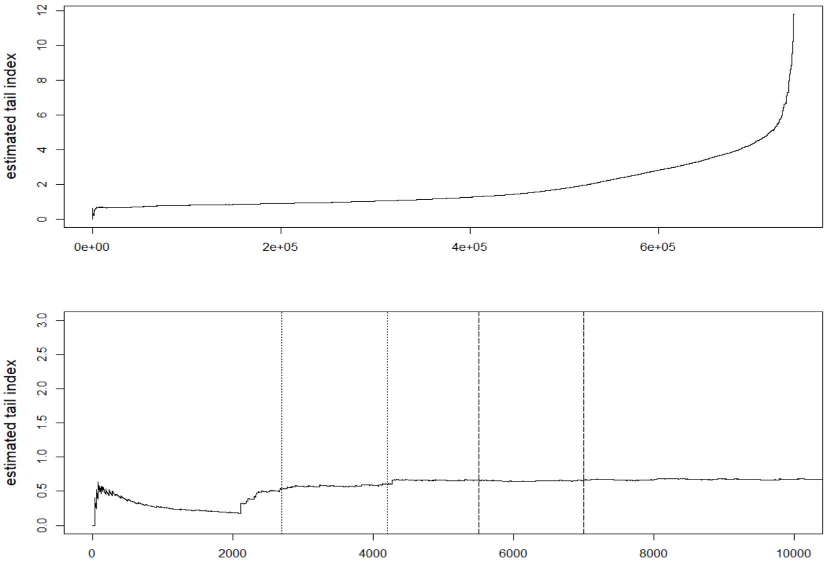


Figure 1.4: The Hill plot of the SIPP dataset

1.5 Conclusion

We propose a novel approach based on EL to characterize the transition between the bulk and tail parts of a dataset. There are many desirable features attached to this method. It remains to show the convergence rate and limiting distribution of $\hat{\tau}$ (or some other important statistics of K). We defer those topics to future research projects.

Chapter 2

An Algorithm Based on Empirical Likelihood to Identify Outliers[1]

2.1 Introduction

Empirical likelihood (EL) first proposed by Owen (1988) has been used to find efficient estimators and construct confidence regions. As a nonparametric method, EL is versatile, as it can be used to deal with incomplete data, sampling bias, and information from multiple sources (Owen, 2001).

Applications of EL have been suggested for many statistical problems: diagnostic measures (Zhu et al., 2008), generalized linear models (Xi Chen and Cui, 2003; Li and Pan, 2013), heaviness identification (Chen and Zhang, 2022), and time series analysis (Fan et al., 2012; Chen and Huang, 2021), just to name a few. Of particular interest to us is applying EL to outlier detection which is an important problem in data analysis. Methods for outlier detection can be classified into three types based on the availability of labels: supervised outlier detection, semi-supervised outlier detection, and unsupervised outlier detection. Supervised outlier detection methods require the training set to include both labelled normal[2] observations and labelled outliers. However, labelled data is very costly and can be scarce in the real world.

The semi-supervised outlier detection methods utilize either only normal observations or outliers. Nevertheless, it is difficult to collect only normal observations or cover all

[1] This chapter is co-authored with Tao Chen.

[2] To avoid confusion, we use the word ‘normal’ only for data points that are usual values in the dataset in this Chapter. For distributions, we use the word ‘Gaussian’.

possible types of unusual behaviours. Although some semi-supervised methods allow the dataset to contain an extremely small amount of outliers (Schölkopf et al., 1999; Tax and Duin, 2004; Binbusayyis and Vaiyapuri, 2021), the performance of these methods might be affected by the outliers. Furthermore, semi-supervised methods require a train phasing before identifying outliers so they are mainly used in the detection of additional outliers. In general, both supervised and semi-supervised outlier detection methods are hard to implement in the real world due to the high cost of labelled data. Therefore, unsupervised outlier detection methods are often employed to avoid this issue.

A variety of unsupervised outlier detection methods have been developed, such as k-nearest neighbour (KNN) based methods (Hautamaki et al., 2004; Chen et al., 2010) and Isolation Forest (Liu et al., 2008). KNN-based methods determine outliers according to the k nearest neighbourhood of each observation. Isolation Forest detects outliers based on the idea that outliers are more susceptible to isolation than normal observations. However, KNN-based methods can be computationally expensive when the dataset is very large, while Isolation Forest might need a sufficient amount of data to provide reliable identifications.

This paper develops an unsupervised method based on EL to identify outliers without a training phase. Unlike KNN and Isolation Forest, which need to carefully select parameters to fit datasets, this method only requires to calculate a bootstrap mean as the parameter. In addition, we utilize a chi-squared test instead of selecting a threshold to distinguish outliers from normal observations.

Some earlier works have used EL to detect outliers in the autoregressive time series (Baragona et al., 2016) and regression models (Baragona et al., 2018). They examine the influence of an observation based on the asymptotic behaviour of EL statistics, which is computed by using the EL ratio function with constraints for the autoregressive parameters and regression parameters, respectively. In contrast, this paper uses EL weights that are assigned to each observation of the dataset under the bootstrap mean constraint. This constraint allows us to take advantage of the fact that outliers can distort the mean, while normal observations generally do not. We show that EL weights have different behaviours for normal observations and outliers under the bootstrap mean constraint. Since EL weights are assigned to each observation, the distribution of EL weights is more sensitive to the influence of each outlier than the EL statistic. We measure the goodness of fit between the distribution of EL weights and a uniform distribution by the chi-squared test. In contrast to the EL statistic, the influence of outliers is amplified in the chi-squared statistic. Additionally, EL weights also provide a measure of outlierness for all observations, which might reduce the cost of time of our method, especially for large and high-dimensional datasets.

The article is organized as follows. Section 2.2 describes the methodology proposed. Section 2.3 reports and analyzes the simulation results. Section 2.4 concludes.

2.2 Methodology

2.2.1 Algorithm

Let $X_1, X_2, \dots, X_n \in \mathbb{R}^d$ be independent random vectors with $\mathbb{E}(X_i) = \mu_x$, denoted as $\{X\}_{i=1}^n$, and let $\{w\}_{i=1}^n$ be the corresponding EL weights which are assigned to each observation in $\{X\}_{i=1}^n$ by the EL ratio function (2), i.e.,

$$\max_{w_1,\dots,w_n} \left\{ \prod_{i=1}^n w_i \,\middle|\, \sum_{i=1}^n w_i m(X_i,\theta) = 0, w_i \geq 0, \sum_{i=1}^n w_i = 1 \right\}, \tag{2.1}$$

where $m(X,\theta)$ is a set of appropriately chosen constraints. We propose a method to identify outliers in $\{X\}_{i=1}^n$ based on the $\{w\}_{i=1}^n$. The intuition behind our method is that the EL weights of datasets with or without outliers under proper conditions are significantly different.

Now, we don't have prior information regarding the presence or quantity of outliers. In this chapter, the outlier is defined as an observation that is far removed from the rest of the observations, as provided by Maddala (1992). To identify all outliers in the dataset, we first assume there are m outliers in $\{X\}_{i=1}^n$, denoted as $\{X^o\}_{s=1}^m$, where $m \in \mathbb{N}$ and $m << n$. The value of m is much less than the value of n by the definition of outliers. Note that m is usually unknown in practice, we will discuss the choice of m in Section 2.2.2. Let $\{X\}_{j=1}^{n-m}$ be a subset of $\{X\}_{i=1}^n$ obtained by removing all elements in $\{X^o\}_{s=1}^m$ from $\{X\}_{i=1}^n$. Then $\{X\}_{j=1}^{n-m}$ only contains normal observations, i.e., the null hypothesis is that

$$H_0 : \{X\}_{j=1}^{n-m} \text{ does not contains outliers.} \tag{2.2}$$

The alternative hypothesis is that

$$H_1 : \{X\}_{j=1}^{n-m} \text{ contains one or more outliers,} \tag{2.3}$$

In this paper, the corresponding EL weights of $\{X\}_{j=1}^{n-m}$, which is denoted as $\{w\}_{j=1}^{n-m}$, is utilized in the test for (2.2). $\{w\}_{j=1}^{n-m}$ are calculated by the EL ratio function under the mean constraint (1), that is,

$$\max_{w_1,\ldots,w_{n-m}} \left\{ \prod_{j=1}^{n-m} w_j \,\middle|\, \sum_{j=1}^{n-m} w_j \left(X_j - \mu_x\right) = 0, w_j \geq 0, \sum_{j=1}^{n-m} w_j = 1 \right\},$$

Then, from Owen (1990), We readily get:

Lemma 2.2.1. *Let $X_1, X_2, \ldots, X_n \in \mathbb{R}^d$ be independent and identically distributed random vectors with $\mathbb{E}(X_i) = \mu_x$ and $Var(X_i) < \infty$. Let $w_1, w_2, ..., w_n \in \mathbb{R}$ be the corresponding EL weights of $\{X\}_{i=1}^n$ where $w_i \geq 0$, $\sum_{i=1}^n w_i = 1$ and $\sum_{i=1}^n w_i \left(X_i - \mu_x\right) = 0$. Then, $|w_i - \frac{1}{n}| \overset{P}{\rightarrow} 0$ as $n \rightarrow \infty$ for $i = 1, 2, ..., n$.*

Under the mean constraint in Lemma 2.2.1, $|w_i - \frac{1}{n}| \overset{P}{\rightarrow} 0$ as $n \rightarrow \infty$ for all i whenever $\{X\}_{i=1}^n$ contains outliers or not. If there are outliers in $\{X\}_{i=1}^n$, the outliers are actually treated as normal observations. This is because the value of μ_x is distorted by the outliers, consequently affecting the mean constraint.

To avoid this problem, assuming the mean of normal observations is known, we employ it in the mean constraint. In this case, Lemma 2.2.1 still holds for $\{X\}_{i=1}^n$ without outliers since the mean of normal observations is μ_x. Note that a constant can be thought of as following a Discrete Uniform distribution that assigns equal mass $\frac{1}{n}$ to each observation in the dataset with sample size $=n$. We denote the Discrete Uniform distribution by $DU(k)$, where k is the number of distinct observations, as from Balakrishnan and Nevzorov (2004).

Theorem 2.2.1. *Let $X_1, X_2, \ldots, X_{n-m} \in \mathbb{R}^d$ be independent and identically distributed random vectors with $\mathbb{E}(X_j) = \mu_x$ and $Var(X_j) < \infty$. Let $w_1, w_2, ..., w_{n-m} \in \mathbb{R}$ be the corresponding EL weights of $\{X\}_{j=1}^{n-m}$ where $w_j \geq 0$, $\sum_{j=1}^{n-m} w_j = 1$ and $\sum_{j=1}^{n-m} w_j \left(X_j - \mu_x\right) = 0$. If (2.2) be true, $w_j \overset{D}{\rightarrow} DU(1)$ as $n - m \rightarrow \infty$ for $j = 1, 2, ..., n - m$.*

Now, if $\{X\}_{i=1}^n$ contains outliers, the μ_x does not equal to the mean of normal observations. Then, Lemma 2.2.1 does not hold, and we have

Lemma 2.2.2. *Let $X_1, X_2, \ldots, X_n \in \mathbb{R}^d$ be independent and identically distributed random vectors with $\mathbb{E}(X_i) = \mu_x$, which consists of m outliers and $n - m$ normal observations. Assume the mean of normal observations in $\{X\}_{i=1}^n$ is finite, that is, $\mu_{normal} < \infty$, and $\mu_{normal} \neq \mu_x$. Let $w_1, w_2, ..., w_n \in \mathbb{R}$ be the corresponding EL weights of $\{X\}_{i=1}^n$ where $w_i \geq 0$, $\sum_{i=1}^n w_i = 1$ and $\sum_{i=1}^n w_i \left(X_i - \mu_{normal}\right) = 0$. Assume $\frac{m}{n} \nrightarrow 0$ as $n \rightarrow \infty$, then $|w_i - \frac{1}{n}| \overset{P}{\nrightarrow} 0$ as $n \rightarrow \infty$ for $i = 1, 2, ..., n$.*

Theorem 2.2.2. *Let $X_1, X_2, \ldots, X_{n-m} \in \mathbb{R}^d$ be independent and identically distributed random vectors. Assume the mean of normal observations in $\{X\}_{j=1}^{n-m}$ is finite, that is,*

$\mu_{normal} < \infty$, and $\mu_{normal} \neq \mathbb{E}(X_j)$. Let $w_1, w_2, ..., w_{n-m} \in \mathbb{R}$ be the corresponding EL weights of $\{X\}_{j=1}^{n-m}$ where $w_j \geq 0$, $\sum_{j=1}^{n-m} w_j = 1$ and $\sum_{j=1}^{n-m} w_j(X_j - \mu_{normal}) = 0$. Assume (2.3) is true and there exists k outliers in $\{X\}_{j=1}^{n-m}$ where $k << n-m$ by the definite of outliers and $\frac{k}{n-m} \nrightarrow 0$ as $n-m \to \infty$. Then, $w_j \overset{D}{\nrightarrow} DU(1)$ as $n-m \to \infty$ for $j = 1, 2, ..., n-m$.

From Theorem 2.2.1 and Theorem 2.2.2, the $\{w\}_{j=1}^{n-m}$ demonstrates a Discrete Uniform distribution with $[\frac{1}{n-m}, \frac{1}{n-m}]$ under the null hypothesis, whereas this uniformity is disrupted when the alternative hypothesis is true. If $\{X\}_{j=1}^{n-m}$ contains outliers, the distribution of $\{w\}_{j=1}^{n-m}$ depends on the distribution of $\{X\}_{j=1}^{n-m}$.

These suggest that a test for (2.2) can be obtained by comparing the distribution of EL weights with a Uniform distribution. One possible statistic for testing (2.2) is Pearson's chi-squared statistic which tests whether an observed frequency distribution differs from a theoretical distribution. The Pearson's chi-squared statistic is

$$\chi^2 = \sum_{i=1}^{b} \frac{(O_i - E_i)^2}{E_i}, \tag{2.4}$$

where observations are divided among b bins. O_i and E_i are the observed frequency and expected frequency in each bin, respectively.

Theorem 2.2.3. *Assume the observations $w_1, w_2, \dots w_n$ are divided among b bins independently of each other with probabilities $p_1, p_2, \dots p_b$, where $b \in \mathbb{N}$, $b < n$ and $\sum_{i=1}^{b} p_i = 1$. Let E_i be the expected frequency, np_i, in each bins for $i = 1, ..., b$. Let O_i be the observed frequency in each bin for $i = 1, ..., b$. We have $\sum_{i=1}^{b} \frac{(O_i - E_i)^2}{E_i} \xrightarrow{D} \chi^2_{b-1}$.*

In this paper, we measure how closely the observations resemble a Discrete Uniform distribution. Hence, our outliers identification test for (2.2) can be defined as

$$\chi^2_o = \frac{b}{n-m} \sum_{i=1}^{b} \left(O_i - \frac{n-m}{b}\right)^2. \tag{2.5}$$

Of course, in practice, the mean of normal observations is unknown and has to be estimated. We employ the estimated mean of normal observations, $\hat{\mu}_{normal}$, instead. This estimation problem will be discussed later in Section 2.2. Additionally, the m potential outliers are unobserved as they are to be identified, we use $\hat{m}$ potential outliers that are suggested by $\{w\}_{i=1}^{n}$ instead. Then, our test statistic becomes

$$\hat{\chi}^2_o = \frac{b}{n-\hat{m}} \sum_{i=1}^{b} \left(\hat{O}_i - \frac{n-\hat{m}}{b}\right)^2, \tag{2.6}$$

where $\hat{O}_i$ is the observed frequency for EL weights, which is calculated based on the EL ratio function under mean constraints with $\hat{\mu}_{normal}$, in each bin. Now, if m observations in $\{X\}_{i=1}^n$ have been selected as outliers already. Then, we have

Theorem 2.2.4. *Let the conditions in Theorem 2.2.1 hold and $w_1, w_2, \ldots, w_{n-m}$ are divided among b bins independently of each other with probabilities $p_1, p_2, \ldots p_b$, where $b \in \mathbb{N}$ is a fixed number and $b < n - m$ and $\sum_{i=1}^b p_i = 1$. If (2.2) be true,*

(i) $\chi_o^2 \overset{D}{\to} \chi_{b-1}^2$.

(ii) Assume $\text{plim}(\hat{\mu}_{normal}) = \mu_{normal}$, $\hat{\chi}_o^2 \overset{D}{\to} \chi_{b-1}^2$.

Theorem 2.2.5. *Let the conditions in Theorem 2.2.2 hold and $w_1, w_2, \ldots, w_{n-m}$ are divided among b bins independently of each other with probabilities $p_1, p_2, \ldots p_b$, where $b \in \mathbb{N}$ is a fixed number and $b < n - m$ and $\sum_{i=1}^b p_i = 1$. If (2.3) be true,*

(i) $\lim_{n\to\infty} \chi_o^2 = \infty$ *w.p.a.1.*

(ii) Assume $\text{plim}(\hat{\mu}_{normal}) = \mu_{normal}$, $\lim_{n\to\infty} \hat{\chi}_o^2 = \infty$ *w.p.a.1.*

Theorem 2.2.4 and 2.2.5 state that, as the sample size increase, under (2.2) the distribution of $\hat{\chi}_o^2$ is approximated by χ_{b-1}^2 distribution and under (2.3) $\hat{\chi}_o^2$ goes to ∞. Therefore, the null hypothesis (2.2) is rejected if the observed value of $\hat{\chi}_o^2$ is large enough. In this chapter, we use the 5% critical value. However, multiple testing corrections can be used, especially for datasets that have a possibility of a higher number of outliers. To improve readability, all proofs are placed in Appendix B.

Now, let us discuss the selection of potential outliers. If $\{X\}_{i=1}^n$ contains outliers. The value of $X_i - \mu_{normal}$ for outliers is either much larger or smaller than those for normal observations, so is the value of $w_i = n^{-1}\left(1 + \lambda\left(X_i - \mu_{normal}\right)\right)^{-1}$ with $\lambda \neq 0$. Hence, there is a disparity in the values of $|w_i - \frac{1}{n}|$ between outliers and normal observations. The suggested potential outlier is $X_s^o \in \{X\}_{i=1}^n$ such that $|w_s^o - \frac{1}{n}|$ is among the m largest values, where $s \in \mathbb{N}$ and $s <= n$.

Note that the value of m is dependent on n, which only affects the computation time and is also discussed in Section 2.2.2. If $\{X^o\}_{s=1}^m$ do not contains all outliers in $\{X\}_{i=1}^n$, implying $\{X\}_{j=1}^{n-m}$ still contains some outliers, H_0 will be rejected. Otherwise, H_0 will be accepted. When H_0 is accepted, $\{X^o\}_{s=1}^m$ is allowed to contain some normal observations. Thus, test for $\{X\}_{j=1}^{n-(m-k)}$ are required, where $k \in \mathbb{N}$, $k < m$ and $\{X\}_{j=1}^{n-(m-k)}$ is a subset of $\{X\}_{i=1}^n$ obtained by removing all elements in $\{X^o\}_{s=1}^{m-k}$ from $\{X\}_{i=1}^n$. To identify $\{X_m\}_{s=1}^m$

which exclusively consists of all outliers from $\{X\}_{i=1}^{n}$, we need H_0 of absence of $\{X_m\}_{s=1}^{m}$ is accepted, while H_0 of absence of $\{X^o\}_{s=1}^{m-k}$ is rejected for any $k < m$. Our method can be summarized as follows:

(a) Obtain an estimated mean vector $\hat{\mu}$ of $\{X\}_{i=1}^{n}$ by using resample method.

   (i) Randomly select l observations from $\{X\}_{i=1}^{n}$. Then, calculate the average of $\{X\}_{i=1}^{l}$, denoted as $\hat{\mu}^1$.

   (ii) Repeat (i) B times. Let $\hat{\mu}$ be the average of $\hat{\mu}^1, ..., \hat{\mu}^B$.

(b) Assign the weight to X_i by using the EL ratio function with the bootstrap mean constraint. Under the bootstrap mean constraint, i.e., $m(X, \hat{\mu}) = X - \hat{\mu}$, the empirical likelihood ratio function becomes

$$\max_{w_1,...,w_n} \left\{ \prod_{i=1}^{n} w_i \middle| \sum_{i=1}^{n} w_i \left(X_i - \hat{\mu}\right) = 0, w_i \geq 0, \sum_{i=1}^{n} w_i = 1 \right\},$$

and the weight function for X_i's are

$$w_i = n^{-1} \left(1 + \lambda \left(X_i - \hat{\mu}\right)\right)^{-1}, \tag{2.7}$$

where the Lagrange multiplier λ can be found by numerical search. Denote the weights of X_i's by $\{w\}_{i=1}^{n}$.

(c) Test if the distribution of $\{w\}_{i=1}^{n}$ is close to a uniform distribution by using the outlier identification test. If it is close to a uniform distribution, then all observations in $\{X\}_{i=1}^{n}$ are normal. Otherwise, follow the steps below to detect outliers.

(d) Calculate $d_i = |w_i - \frac{1}{n}|$ and sort $\{w\}_{i=1}^{n}$ according to the ascending sort of d_i. Denote the sorted weights as $\left\{w_{(i)}\right\}_{i=1}^{n}$ and $w_{(i)}$ is the weight assigned to $X_{(i)}$ for $i = 1, ..., n$.

(e) Remove $X_{(n-m+1)}, ..., X_{(n)}$ from $\left\{X_{(i)}\right\}_{i=1}^{n}$, then test if the distribution of $\left\{w_{(i)}\right\}_{i=1}^{n-m}$ is close to a uniform distribution by using the outlier identification test. If it is close to a uniform distribution, test $\left\{w_{(i)}\right\}_{i=1}^{n-m+1}$. Otherwise, we proceed with further testing.

(f) Repeat (e) until the test result shows that $\left\{w_{(i)}\right\}_{i=1}^{n-q}$ is not close to a uniform distribution and $\left\{w_{(i)}\right\}_{i=1}^{n-q-1}$ is close to a uniform distribution, where $q = 0, 1, ..., n-1$. The outliers are $X_{(n-q)}, ..., X_{(n)}$.

2.2.2 The implementation problems

We now turn to discuss some implementation problems. First, we should note that this chapter uses the mean constraint since it allows us to take advantage of the fact that outliers can distort the mean, while normal observations generally do not. However, the mean constraint can be replaced by other constraints $m(X,\theta)$ on the condition that θ is affected by outliers.

Next, we need to estimate the mean of normal observations in $\{X\}_{i=1}^{n}$. The sample mean of $\{X\}_{i=1}^{n}$ is not suitable in this case because it is highly distorted by the outliers.

Remark 2.2.1. *If $w_1, w_2, ..., w_n \in \mathbb{R}$, $w_i \geq 0$, $\sum_{i=1}^{n} w_i = 1$, and $\sum_{i=1}^{n} w_i(X_i - \frac{1}{n}\sum_{i=1}^{n} X_i) = 0$, then the function $\prod_{i=1}^{n} w_i$ is maximized by taking $w_i = \frac{1}{n}$ for all i.*

In this paper, we employ a bootstrap mean which is less distorted by outliers compared to the sample mean such that the outliers are not treated as normal observations. The estimation error from using the estimated mean of normal observations instead of the true mean of normal observations affects the value of w_i, so that we might have $w_i \neq \frac{1}{n}$ for the dataset without outliers for some i. However, the difference in EL weights between datasets with and without outliers is still significant as long as the estimated mean of normal observations is less distorted than the sample mean. The reason behind this is that the function $\prod_{i=1}^{n} w_i$ is maximized when $\left(w_i - \frac{1}{n}\right)^2$ is minimized given the conditional of Remark 2.2.1. When employing the estimated mean of normal observations for the dataset without outliers, it is unnecessary for the values of w_i for normal observations to offset the effects of outliers as in the dataset with outliers. This suggests that the shift caused by estimation error is considerably smaller.

Due to the estimation error, the distribution of the EL weights for the dataset without outliers also might not exactly follow a Uniform distribution. Hence, the next implementation problem is about measuring the goodness of fit between the distribution of the EL weights and a Uniform distribution. We compare them by partitioning the EL weights into bins and testing if the EL weights are equally distributed among the bins using our outlier identification test. Employing a strict criterion for the test might lead to detecting normal observations as outliers when the distribution of the EL weights is only close to but does not exactly follow a Uniform distribution, particularly for datasets with highly skewed or heavy-tailed. The strictness of the test can be modified by the selection of b. To impose a loose restriction for the test, we select a high value for b. In this paper, we impose a loose restriction for the test with $b = n - m$. Additionally, the value of b does not require change as removing potential outliers each time, since the number of potential outliers is much smaller than n by the definition of outliers.

Finally, we present three remarks which are for reducing the computational burden. In this paper, m is the number of outliers we initially assumed. Since m is an estimated number of outliers, it is possible that $\{X\}_{j=1}^{n-m}$ contains outliers or $\{X^o\}_{s=1}^{m}$ contains normal observations. However, the initially selection of m does not affect the result of identification since we repeat the test until all outliers are detected. So, m only impacts the time cost. In this chapter, we determine the value of m based on the Interquartile Range (IQR), that is, the difference between the upper quartiles ($Q3$) and lower quartiles ($Q1$).

Remark 2.2.2. *Calculate the IQR of $\{w_i\}_{i=1}^{n}$, and find the number of weights that less than $Q1 - 1.5 \times IQR$ and more than $Q3 + 1.5 \times IQR$. This number is denoted as $N_{outlier}$. If $N_{outlier} <= 50$, set $m = 1$. Otherwise, m equals the round of the $N_{outlier}/100$.*

Remark 2.2.3. *For large datasets, the $\{w_i\}_{i=1}^{n-m+1}$ and $\left\{w_{(i)}\right\}_{i=1}^{n-m-1}$ in step (e) can be replaced by $\left\{w_{(i)}\right\}_{i=1}^{n-m+k}$ and $\left\{w_{(i)}\right\}_{i=1}^{n-m-k}$ for $k = 1, 2, ..., m-1$ to reduce the time cost. Note that the smaller the value of k, the more accurate the result becomes.*

Although this algorithm is designed to detect outliers that might already exist in the input dataset, it can effectively examine the new observation.

Remark 2.2.4. *Suppose a large dataset without outliers is given, denoted as X. To reduce the cost of time for examining the influence of a new observation, a subset X_{subset} is randomly selected from X. Then, add the new observation into X_{subset} instead of X and then implement the examination.*

2.3 Simulation

In this section, we present results for 1D and 2D simulation studies, both including cases with and without outliers. Of the 1D datasets, one case consists of 9990 normal observations drawn from a Gaussian distribution with mean 2 and variance 1 and 10 outliers drawn from another Gaussian distribution with mean 20 and variance 5, denoted as X. The other one is outliers free, which only contains 10000 observations drawn from a Gaussian distribution with mean 2 and variance 1, denoted as X'.

Using the method introduced in the previous section, we found that the EL weights of normal observations are around $1/n$, that is, $1/10000$ in both X and X_{normal}. However, the EL weights of outliers are relatively larger than $1/10000$. The plot of the EL weigths for X is shown in Appendix B.1. The first column of Table 2.1 is the number of removed elements from the datasets, which is mentioned in the algorithm step (e). The second column presents the p-value of our outliers identification test for $\left\{X_{(i)}\right\}_{i=1}^{n-1}, ..., \left\{X_{(i)}\right\}_{i=1}^{n-12}$, respectively. The p-values of EL weights of $\left\{X_{(i)}\right\}_{i=1}^{n-1}, ..., \left\{X_{(i)}\right\}_{i=1}^{n-9}$ are statistically significant so that their distributions are not close to uniform distributions. Starting from the $\left\{X_{(i)}\right\}_{i=1}^{n-10}$, the p-values are not significant. This is because all outliers are removed so that their EL weights are around $1/n$. Thus, our method gives that the outliers in X are $X_{(n)}, X_{(n-1)}, ..., X_{(n-9)}$. The third column of Table 2.1 shows the p-value from the chi-square test of EL weights for $\left\{X'_{(i)}\right\}_{i=1}^{n-1}, ..., \left\{X'_{(i)}\right\}_{i=1}^{n-12}$, which are all not significant since X' does not contain outliers.

Table 2.1: P-value of the outliers identification test for 1D cases

Num of Removed elements	P-value for X	P-value for X'
1	1.2265e-280	0.9999
2	3.5797e-268	0.9999
3	4.9536e-261	0.9999
4	6.2142e-210	0.9999
5	4.6449e-209	0.9999
6	9.1252e-199	0.9999
7	6.9710e-146	1
8	1.0750e-96	1
9	7.1828e-31	1
10	0.9999	1
11	0.9999	1
12	0.9999	1

For the 2D datasets, one case consists of 9990 normal observations drawn from a bivariate Gaussian distribution with mean $[2, 5]$ and variance-covariance matrix $\left[\begin{smallmatrix}1 & 0\\ 0 & 1\end{smallmatrix}\right]$ and 10 outliers drawn from another bivariate Gaussian distribution with mean $[20, 30]$ and variance-covariance matrix $\left[\begin{smallmatrix}5 & 0\\ 0 & 5\end{smallmatrix}\right]$, denoted by Y. The other one only contains 10000 observations drawn from the bivariate Gaussian distributions with mean $[2, 5]$ and variance-covariance matrix $\left[\begin{smallmatrix}1 & 0\\ 0 & 1\end{smallmatrix}\right]$, denoted by Y'.

Again, the EL weights for normal observations are around 1/10000 in both Y and Y', which are much smaller than the EL weights for outliers. The plot of the EL weights for Y is shown in Appendix B.3. Table 2.2 shows that the p-value of our outliers identification test for $\left\{Y_{(i)}\right\}_{i=1}^{n-1}, ..., \left\{Y_{(i)}\right\}_{i=1}^{n-9}$ are statistically significant so that their distributions are not

Table 2.2: P-value of the outliers identification test for 2D cases

Num of Removed elements	P-value for Y	P-value for Y'
1	1.1612e-196	0.9999
2	0	0.9999
3	5.683e-197	0.9999
4	0	0.9999
5	0	0.9999
6	0	0.9999
7	0	0.9999
8	6.3346e-153	0.9999
9	0	0.9999
10	0.9999	0.9999
11	0.9999	0.9999
12	0.9999	1

close to uniform distributions. The p-values of the outliers identification test for Y are not monotonically increasing as the number of removed elements increased in Table 2.2, because EL weights in Y are contributed by two variables from different dimensions. However, the p-values are still not significant starting from $\left\{Y_{(i)}\right\}_{i=1}^{n-10}$. Thus, the outliers in Y are $Y_{(n)}, Y_{(n-1)}, ..., Y_{(n-9)}$. The third column of Table 2.2 shows the p-value of the outliers identification test for $\left\{Y'_{(i)}\right\}_{i=1}^{n-1}, ..., \left\{Y'_{(i)}\right\}_{i=1}^{n-12}$, which are all not significant since Y' is outlier free.

Furthermore, the difference between the behaviours of EL weights for normal observations and outliers is also found by other indicators of EL weights for these four cases, such as skewness or kurtosis. Here we just present the kurtosis of the EL weights for the simulated datasets, the skewness of the EL weights for the simulated datasets is presented in Appendix B.1 and B.2. Table 2.3 and 2.4 present the kurtosis of EL weights for all four cases. For both X' and Y', the kurtosis of EL weights does not have significant change. However, the kurtosis of EL weights for X decreases rapidly as the number of removed elements increases in the first 10 rows of Table 2.3, since the tailedness is reduced by removing outliers. In addition, the decreasing of kurtosis becomes slow from $\left\{X_{(i)}\right\}_{i=1}^{n-10}$ since the normal observations $X_{(n-11)}$ and $X_{(n-12)}$ do not distort the kurtosis. Therefore, we also get that $X_{(n)}, X_{(n-1)}, ..., X_{(n-9)}$ are outliers by using the change of kurtosis. The second column of Table 2.4 shows that the kurtosis of EL weights for Y is not monotonically decreasing as the number of removed elements increased because EL weights in the 2D case are contributed by two variables from different dimensions. However, the change of kurtosis still becomes small starting from $\left\{Y_{(i)}\right\}_{i=1}^{n-10}$. Thus, it also gives that the outliers are $Y_{(n)}, Y_{(n-1)}, ..., Y_{(n-9)}$ in Y. Again, the reason is that EL weights for outliers heavily affect the value of kurtosis whereas EL weights for normal observations generally do not distort the kurtosis. A test based on the skewness or kurtosis of the EL weights for comparing datasets with and without outliers is an opportunity for further exploration.

Table 2.3: Kurtosis of EL weights for 1D cases

Num of Removed elements	Kurtosis of X	Kurtosis of X'
1	66.1735	2.9957
2	53.9540	2.9821
3	42.1980	2.9720
4	31.2103	2.9612
5	25.7994	2.9509
6	17.2385	2.9410
7	9.9809	2.9322
8	5.2887	2.9239
9	3.3157	2.9157
10	2.9041	2.9074
11	2.8921	2.8995
12	2.8835	2.8923

2.4 Conclusion

This chapter explores the use of EL weights for identifying outliers. The EL weights are computed by the EL ratio function with bootstrap mean constraint. In this paper, the mean constraint is used due to the fact that outliers distort the mean whereas normal observations generally do not. This leads the EL weights for outliers and normal observations to behave differently. The distribution of EL weights for datasets without outliers is close to a Discrete Uniform distribution, while EL weights for datasets with outliers do not. Different constraints are also effective as long as they are influenced by outliers. We test

Table 2.4: Kurtosis of EL weights for 2D cases

Num of Removed elements	Kurtosis of Y	Kurtosis of Y'
1	247.0733	2.9763
2	118.1280	2.9978
3	87.3886	3.0620
4	153.5379	2.9452
5	6.6021	2.9911
6	186.0335	2.9882
7	124.2261	2.9920
8	47.3130	3.0367
9	34.6328	2.9129
10	2.9485	3.0086
11	2.9535	2.9494
12	3.015	2.9597

the distribution of EL weights by Pearson's chi-squared test. The selection of tests might be investigated in future work.

Chapter 3

A Clustering Algorithm Based on Empirical Likelihood

3.1 Introduction

Clustering is an important problem in data mining, aiming to group datasets into different clusters based on their similarities. Compared to classification, clustering is more challenging since it is an unsupervised learning problem, that is, the datasets are unlabelled. In the real world, labelling datasets usually are very expensive or hard to collect. Thus, clustering has wide applications in many fields, such as biomedical science (Wiwie et al., 2015), education (Abdullah et al., 2011; Dutt et al., 2015), social networks (Handcock et al., 2007; Chang et al., 2017), and climatology (Benmouiza and Cheknane, 2019).

Various clustering methods have been proposed (Ezugwu et al., 2022). Two primary types of clustering are partitional clustering and hierarchical clustering. Partitional clustering includes methods that iterative partition groups based on the similarity of datasets. Two of the most well-known methods in this type are K-means (Steinhaus, 1956; Forgy, 1965; MacQueen, 1967) and K-medoids (Kaufman and Rousseuw, 1990; Park and Jun, 2009). K-means assigns each observation into K clusters according to the centroid of the clusters, while K-medoids relocates observations according to the medoid of the clusters. Both require specifying the number of clusters (K), the centroid/medoid of each cluster, and the distance metric. The value of K and the initial setting of centroids/medoids highly affect the results. Some papers have discussed the selection of optimal K (Makles, 2012) However, they require several runs for different K with different selections of initial centroids/medoids, which might be computationally expensive when the dataset is very large (Ikotun et al., 2023). Additionally, the standard K-means or K-medoids use the Euclidean distance metric to measure the similarity. Then, they have a challenge in detecting arbitrary-shaped clusters.

Regarding hierarchical clustering, it groups clusters in a hierarchical format, which is a dendrogram that reveals the dataset's structure. Hierarchical clustering can be divided into two types based on the direction of clustering, that is, agglomerative methods, which are bottom-up approaches, and divisive methods, which are top-down approaches. None of them need to specify the number of clusters, however, they require a distance threshold to customize the number of clusters. In addition, the performance of hierarchical clustering depends on the selection of the distance function and linkage criteria. The most widely used distance functions are the Euclidean distance function, and the popular linkage criteria are average (Sokal and Michener, 1958), complete (Sokal and Michener, 1948), single (Sneath, 1957), and Ward's (Ward, 1963). Moreover, it is well-known that they are sensitive to outliers and computationally expensive.

Besides these two primary types of clustering, other types of clustering have also been developed. For example, density-based clustering is designed to deal with arbitrary-shaped clusters. A representative method of this type is DBSCAN (Ester et al., 1996; Schubert et al., 2017). It does not require a parameter to specify the number of clusters. However, it is sensitive to the choice of two parameters, the distance that specifies the neighbourhoods (Eps) and the minimum number of data points to define a cluster (MinPts). Furthermore, it is not suitable for the case of varied-density clusters. Next, parametric model-based clustering algorithms, such as the Gaussian Mixture Model, are also well-known (Melnykov and Maitra, 2010). They group the clusters based on the mixture of pre-defined models where each model represents a cluster. It works better for datasets with complex structures. However, it requires prior information about the models, and the performance is poor if the models are misspecified.

In this paper, we present a clustering method based on the Empirical likelihood (EL) introduced by Owen (1988). EL is a non-parametric method, implying it is a good candidate for analyzing datasets with little or even no prior information. EL has been employed in clustering problems by Melnykov and Shen (2013) who propose an EL-based iterative algorithm. However, their method is designed to identify which cluster a new observation belongs to. Our method uses the EL weights assigned to each observation to group the clusters, since the behaviour of EL weights for different clusters differs under appropriate constraints. Although a threshold that customizes the number of clusters is required, the difference between the EL weights for each cluster makes selecting a threshold relatively easy. Additionally, our method is available to work on datasets with outliers.

The article is organized as follows. Section 3.2 describes the proposed EL-based algorithm. Section 3.3 reports and discusses the performance of simulation datasets using the proposed clustering algorithm. Section 3.4 applies the method to the celebrated Iris dataset. Section 3.5 provides a conclusion.

3.2 Methodology

3.2.1 Algorithm

We begin by recalling some necessary preliminaries on EL, i.e. the EL ratio function (2) and EL weight function (3). Let $\{X\}_{i=1}^{n}$ be a collection of n independent and identically distributed random vectors with unknown distribution F. The EL ratio function is

$$\max_{w_1,\dots,w_n}\left\{\prod_{i=1}^{n} w_i \,\middle|\, \sum_{i=1}^{n} w_i m(X_i,\theta) = 0, w_i \geq 0, \sum_{i=1}^{n} w_i = 1\right\},$$

where $\{w\}_{i=1}^{n}$ is the corresponding EL weights that are assigned to each observation in $\{X\}_{i=1}^{n}$, and $m(X,\theta)$ is a set of appropriately chosen constraints that depend on the parameter θ and the true parameter value of θ should satisfy $\mathbb{E}[m(X,\theta)] = 0$.

Using a Lagrange multiplier method, the EL weight function for X_i's are

$$w_i = n^{-1}\left(1 + \lambda m(X,\theta)\right)^{-1},$$

where the Lagrange multiplier λ can be found by numerical search.

We propose a clustering method based on the EL weights $\{w\}_{i=1}^{n}$. The intuition behind our method is that the EL weights for different clusters are behaviour different under appreciated constraints. Without constraints, all EL weights are equal to $\frac{1}{n}$ by the inequality of arithmetic and geometric means. When constraints are added, the constraint function $m(X,\theta)$ changes the EL weights based on certain criteria related to the behaviour of observations. If the constraint function $m(X,\theta)$ correctly captures similarities between observations, the EL weights will deviate from $\frac{1}{n}$ in a distinct manner for each cluster and close to EL weights within the same cluster. Thus, the clusters of the datasets can be detected based on their corresponding EL weights.

The utilization of constraint in this EL-based clustering method might closely resemble the utilization of distance metric or linkage metric in other clustering methods. The purpose of utilizing them is to capture the inherent structure of the datasets. The selection of them influences clustering results, however, there is no best constraint or metric for clustering. Different type of datasets or the researchers' questions requires different constraints or metrics. In this chapter, we discuss mean constraint and variance constraint as exposition.

Let $X_{k\,i}$ is the i^{th} observation from the k^{th} cluster for $k = 1, ..., K$ and $i = 1, ..., n$. Let μ_k be the mean of each cluster and μ be the overall mean of X. The EL ratio function for the mean is (1) and the EL weights can be written as

$$w_i = n^{-1} \left(1 + \lambda(X_i - \mu)\right)^{-1},$$

where μ is the expected value of X, that is, $\mathbb{E}(X_i) = \mu$. When the mean constraint is appropriate for the dataset, the observations within a cluster are closer to the center of that cluster than to the centers of other clusters, that is, $|X_{k,i} - \mu_k|$ is smaller than $|X_{k,i} - \mu_j|$. Thus, $X_{k,i} - \mu$ is closer to $\mu_k - \mu$ than $\mu_j - \mu$ for $j \neq k$. When $\mu_k - \mu$ is not equal to $\mu_j - \mu$, the difference of EL weights from different clusters is smaller than those in the same clusters. In other words, the EL weights from different clusters contribute differently to satisfy the constraints. Additionally, to solve (1), all EL weights are as close to $1/n$ as possible, as previously stated. Therefore, the difference in EL weights from the same clusters is compressed to meet the constraints. Similar to the variance constraints.

In general, our method to group the clusters in the dataset X can be summarized as follows:

(a) Assign the EL weight to X_i using the EL ratio function under constraint $m(X, \hat{\theta})$. Denote the EL weights of X_i's by $\{w\}_{i=1}^{n}$.

(b) Sort $\{w\}_{i=1}^{n}$ in ascending order, denote the sorted weights as $\left\{w_{(i)}\right\}_{i=1}^{n}$ and $w_{(i)}$ is the weight assigned to $X_{(i)}$ for $i = 1, ..., n$.

(c) Calculate the differences between the consecutive values in the sorted EL weights $\left\{w_{(i)}\right\}_{i=1}^{n}$, denoted as $\{d_{(i)}\}_{i=1}^{n-1}$. Then, $d_{(i)}$ is the difference between $w_{(i)}$ and $w_{(i+1)}$.

(d) Find all $d_{(i)}$ which is bigger than a threshold D, and their index values in $\{d_{(i)}\}_{i=1}^{n-1}$. Denote the index values as $\{r_j\}_{j=1}^{m}$.

(e) Split the $\left\{w_{(i)}\right\}_{i=1}^{n}$ into $m+1$ clusters based on $\{r_j\}_{j=1}^{m}$. Then, the $m+1$ clusters contains values in X whose indices range from 1 to r_1, $r_1 + 1$ to r_2,..., $r_{m-1} + 1$ to r_m and $r_m + 1$ to n, respectively.

(f) Check the length of each cluster. If the length of the cluster j is smaller than a selected number, $MinNum$, the observations in the cluster will be treated as outliers or merged into the cluster $j-1$ or $j+1$, if they exist.

(g) If cluster $j-1$ and $j+1$ both exist, merge the cluster j to the cluster $j-1$ if $d_{(r_{j-1})} < d_{(r_j)}$. Otherwise, merge it to the cluster $j+1$.

3.2.2 The implementation problems

We now turn to discuss some implementation problems. First, to compute the EL weights, we need to select the θ for constraint $m(X, \theta)$, denoted by $\hat{\theta}$. The selection of θ influences the clustering results. For example, the sample mean is not suitable in this case of the mean constraint because it makes the EL weights view all observations as in one cluster. Numerically, it is $\sum_{i=1}^{n} w_i X_i = \frac{1}{n}\sum_{i=1}^{n} X_i$ which makes the $\lambda = 0$ and $w_i = \frac{1}{n}$ for all i. For some well-separated clusters, it is accepted to use the bootstrap method for determining the value of θ. Other selections of θ are also available depending on the datasets. Note that we do not need to accurately estimate θ, since the purpose for selecting θ is to give constraint for clustering. In this chapter, we use the bootstrap method to select θ for simulated datasets, while the 3rd quantile of the dataset is used for the empirical dataset which contains both well-separated clusters and non-disjoint clusters.

Next, let us consider the selection of threshold D, which affects the proposed number of clusters. For well-separated clusters, there are usually some huge gaps between the sorted EL weights of different clusters, which suggests the number of clusters. Therefore, the threshold D can be easily selected. However, it is relatively complex to select the threshold D for non-disjoint clusters which is also a challenging problem for many other clustering methods. For non-disjoint clusters, the observations within a cluster might share some characteristics with the observations from other clusters. Unlike well-separated clusters, the change in the sorted EL weights occurs relatively gradually. Therefore, there may not be huge individual differences in the sorted EL weights that directly give the splitting point for non-disjoint clusters. Then, the threshold D is strategically set to capture distinct variations in the sorted EL weights for different clusters. We present examples of non-disjoint clusters in detail in the following section.

Finally, the value of $MinNum$ must also be chosen for the EL-based clustering method. The selection of $MinNum$ might depend on the sample size of the dataset. When the sample size of a cluster is smaller than the value of $MinNum$, the observations in the cluster are either outliers or located at the outer boundaries of another cluster. Thus, this EL-based clustering method is able to deal with datasets with outliers, and it is possible to identify outliers when grouping clusters.

3.3 Simulation studies

In this section, we identify similar groups of synthetic datasets using the EL weights. As an exposition, we present three datasets with well-separated clusters that can be grouped by EL weights under the mean constraint or variance constraint. Then, we discuss two datasets with non-disjoint clusters.

3.3.1 Well-separated clusters

Mixture of three Gaussian distributions

We consider three bi-variate Gaussian clusters with different sizes, variances, and co-variances. The parameters of the employed mixture model are presented in Table 1. The n_k, μ_k and Σ_k are the size, mean vector, and covariance matrix of the k^{th} cluster of the synthetic dataset, respectively.

Table 3.1: The parameters of three bi-variate Gaussian clusters

Parameters	Black	Green	Red
n_k	500	300	300
μ_k	$\begin{pmatrix} 4 \\ 4 \end{pmatrix}$	$\begin{pmatrix} 8 \\ 10 \end{pmatrix}$	$\begin{pmatrix} 12 \\ 4 \end{pmatrix}$
Σ_k	$\begin{pmatrix} 1 & -0.1 \\ -0.1 & 1 \end{pmatrix}$	$\begin{pmatrix} 0.1 & 0 \\ 0 & 0.1 \end{pmatrix}$	$\begin{pmatrix} 0.3 & -0.3 \\ -0.3 & 0.31 \end{pmatrix}$

We standardize the dataset and then compute the EL weights under the mean constraint with the bootstrap mean, while a result by using the mean constraint with 3^{rd}-quantile of the dataset is presented in Appendix 3.2. The left panel of Figure 3.1 shows the plot of sorted EL weights. There are 2 gaps in the sorted EL weights that split the weights into 3 parts, so our method indicates the dataset includes 3 clusters. The gaps are detected by setting the threshold D as the 99.9%-quantile of the difference of sorted EL weights. The plot for the differences of the sorted EL weights is provided in Appendix C.1. The clustering result obtained by the EL weights is shown in the right panel of Figure 3.1.

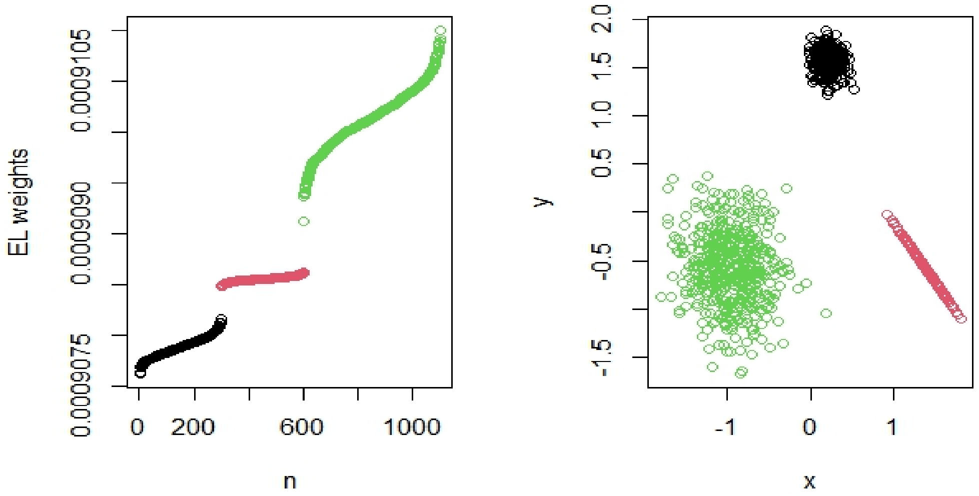


Figure 3.1: The plots of mixture of Gaussian distributions

3D atom dataset

We present another simulation study for the 3D atom dataset defined in Ultsch (2004). As shown in the right panel of Figure 3.2, the 3D atom dataset contains two clusters where each cluster contains 400 observations and is indicated by different symbols. The shape of this atom dataset is a core enclosed by a hull, which is a completely overlapping convex hull (Steinhaus, 2020). Thus, the mean constraint cannot deal with this problem, and we use the variance constraint with the bootstrap variance.

Figure 3.2 presents the sorted EL weights and the clustering result for the dataset after standardization. By setting the threshold D to be the 99.9%-quantile of differences of sorted EL weights, the sorted EL weights are split at the 400^{th} EL weight. The plot for the differences of the sorted EL weights is presented in Appendix C.2. Thus, the 3D atom dataset is split into 2 clusters (shown as red and black) based on EL weights.

Mixture of Von Mises-Fisher distributions

We simulate a spherical dataset with two clusters, each containing 100 observations. The first cluster follows a Von Mises-Fisher distribution with the mean direction being $(-1, 1)$

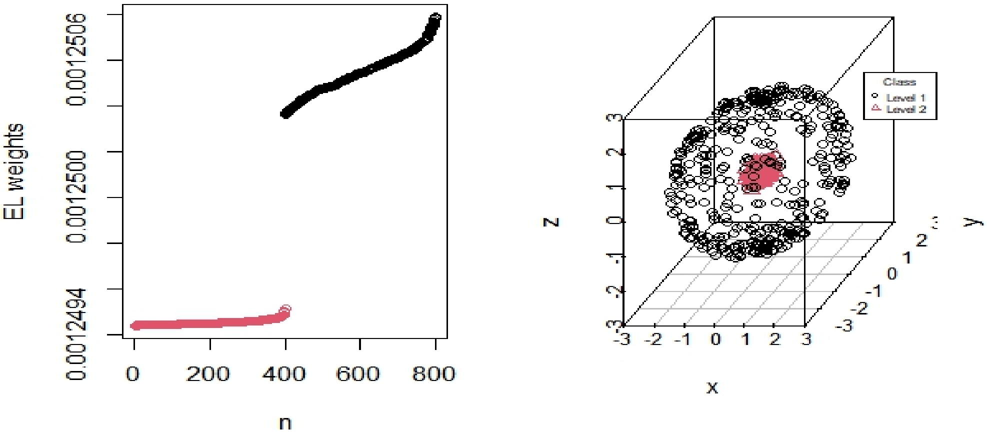


Figure 3.2: The plots of the 3D atom dataset

and the concentration parameter being 1, and the second cluster follows a Von Mises-Fisher distribution with the mean direction being $(1, -1)$ and the concentration parameter being 1. Then, we scale the first cluster 15 times, so that the two clusters have the same center and the radius of the outer cluster is 15 times the radius of the inner cluster. In addition, four observations are added in four corners, that is, {(-15, -15), (-15, 15), (15, -15), (15, 15)}. We might treat them as outliers or observations from a small cluster.

As we can see from the left panel of Figure 3.3, there are three relatively large gaps for the sorted EL weights of the dataset after standardization. The gaps are at the 1^{st}, 3^{rd} and 104^{th} EL weights. The plot for the differences of the sorted EL weights is presented in Appendix C.3. Then, we split the EL weights into four groups. The first group only contains one element and the second contains two. The number of EL weights in these two groups is too small to be a cluster. Additionally, these three EL weights are far away from the remaining EL weights. Then, the first three EL weights might be treated as outliers or observations from a small cluster when $MinNum <= 3$. If $MinNum > 3$, they might be treated as a part of the outer cluster. In this chapter, we set $MinNum = 2$ and treat the four additional observations as outliers. Then, our method gives that the synthetic dataset consists of 2 clusters and three outliers. The observations are grouped by the corresponding EL weights and are shown in the right panel of Figure 3.3, which shows that our method only detects three outliers and treats one outlier as the observation of the outer cluster. The undetected outlier might be because of the significant density inequality, implying the varied density might affect the result.

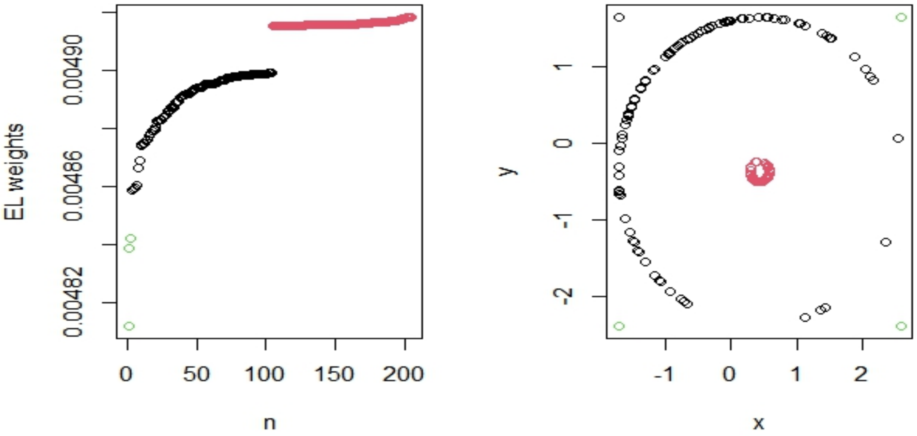


Figure 3.3: The plots of the mixture of Von Mises-Fisher distributions

We also compare the performance with those obtained using other popular algorithms, including k-means, k-medoids, DBSCAN, and hierarchical clustering with average, single and complete linkages for this dataset. Figure 3.4 and 3.5 provides the results of using k-means, k-medoids, and hierarchical clustering with average and complete linkages with $K = 2$ and $K = 3$. These four methods fail to group the clusters correctly. The dendrogram of hierarchical clustering with average and complete linkages is provided in Appendix C.5 and C.6.

Figure 3.6 shows the results of using DBSCAN with different parameters (Eps and MinPts). Additional results with a broader range of parameters are provided in Appendix C.4. As we know, DBSCAN is not suitable for the dataset with varied density clusters and is very sensitive to the value of Eps and MinPts (Chauhan, 2014). If this dataset does not include the outliers, DBSCAN can group the clusters correctly with appropriate parameters. However, with the four outliers, DBSCAN might be unable to correctly group the clusters.

Figure 3.7 illustrates the results using hierarchical clustering with single linkage with different threshold settings. Its dendrogram is provided in Appendix C.7. Again, if this dataset does not include the outliers, it can group the clusters correctly with an appropriate threshold. However, hierarchical clustering approaches are usually sensitive to outliers. In

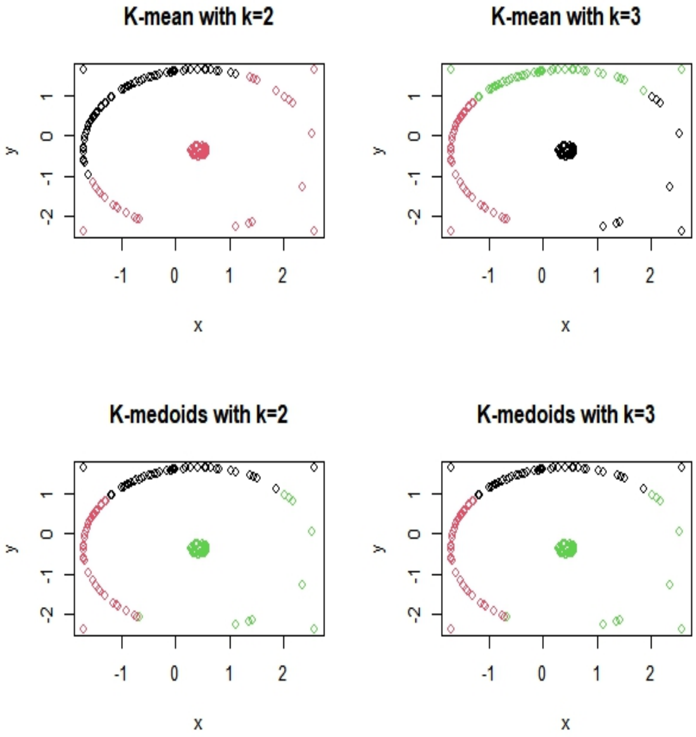


Figure 3.4: Results for the 2D spherical dataset using k-means and k-medoids

this case, it detects either all outliers as observations of the outer cluster or an observation of the outer cluster as an outlier.

3.3.2 Non-disjoint clusters

Non-disjoint cluster is a challenging issue in clustering since it allows an observation to belong to more than one cluster. We now turn to discuss two examples of non-disjoint clusters.

First, we simulate two bi-variate Gaussian clusters, one containing 500 observations with the mean vector $\begin{pmatrix}1.5\\1.5\end{pmatrix}$ and the covariance matrix $\begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}$, and the other containing 500 observations with the mean vector $\begin{pmatrix}5\\5\end{pmatrix}$ and the covariance matrix $\begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}$. The right

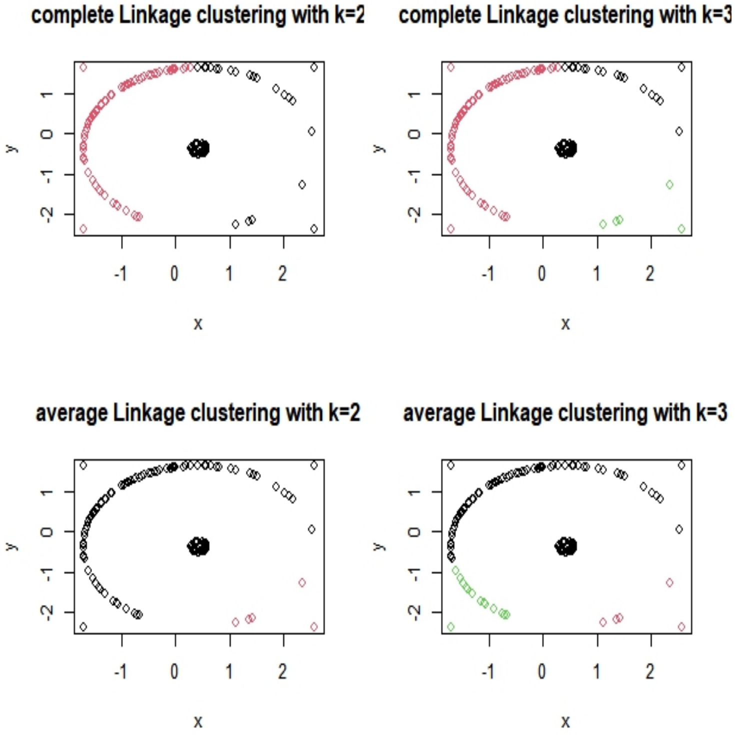


Figure 3.5: Results for the 2D spherical dataset using hierarchical clustering with average and complete linkages

panel of Figure 3.8 shows that these two clusters are slightly overlapped, where circle and triangle symbols indicate the two clusters, respectively.

Then, we calculate the EL weights of the dataset and present the sorted EL weights in the left panel of Figure 3.8. There is no huge gap in the sorted EL weights since the two clusters are non-disjoint. We visualize the differences between all consecutive values of the sorted EL weights. In the middle panel of Figure 3.8, the red line indicates the threshold D which is the 99%-quantile of differences of sorted EL weights. After combining the clusters with lengths less than 5, we partition the dataset into two clusters (red and black), as shown in the right panel of Figure 3.8. The precision and recall are both 0.978 for each cluster.

Next, we consider non-disjoint clusters with different sample sizes and densities. We simulate two bi-variate Gaussian clusters, one containing 1000 observations with the mean

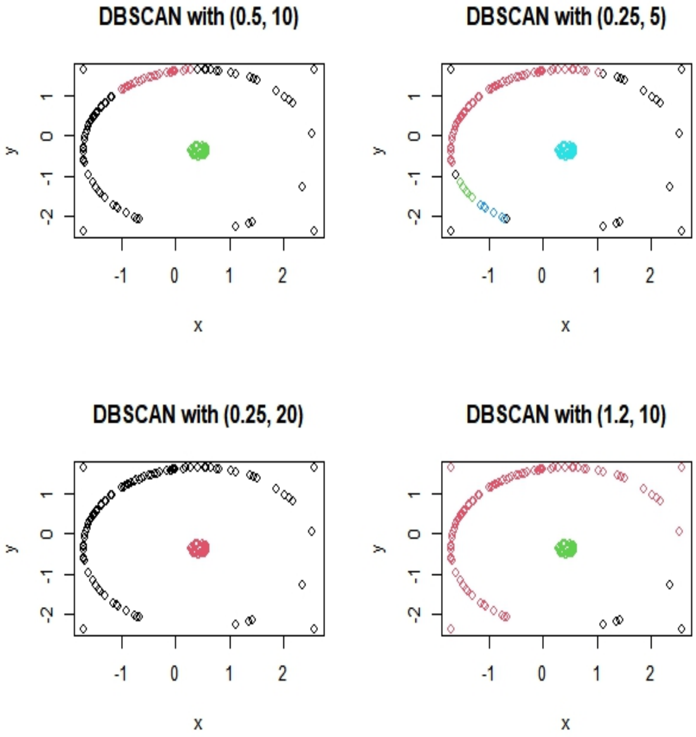


Figure 3.6: Results for the 2D spherical dataset using DBSCAN

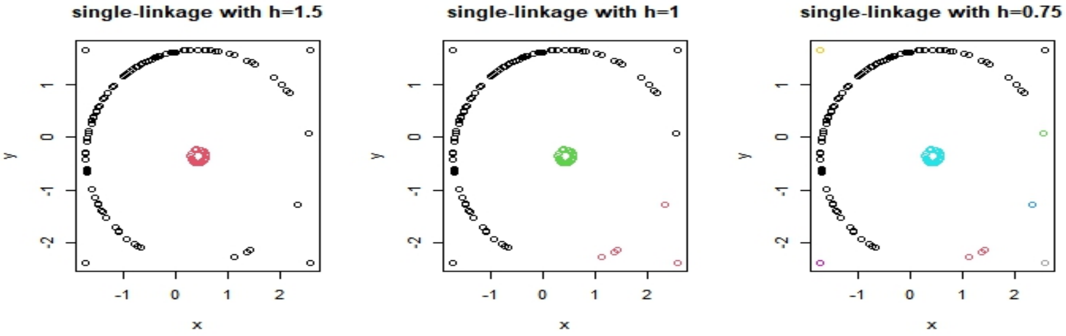


Figure 3.7: Results for the 2D spherical dataset using hierarchical clustering with single linkage

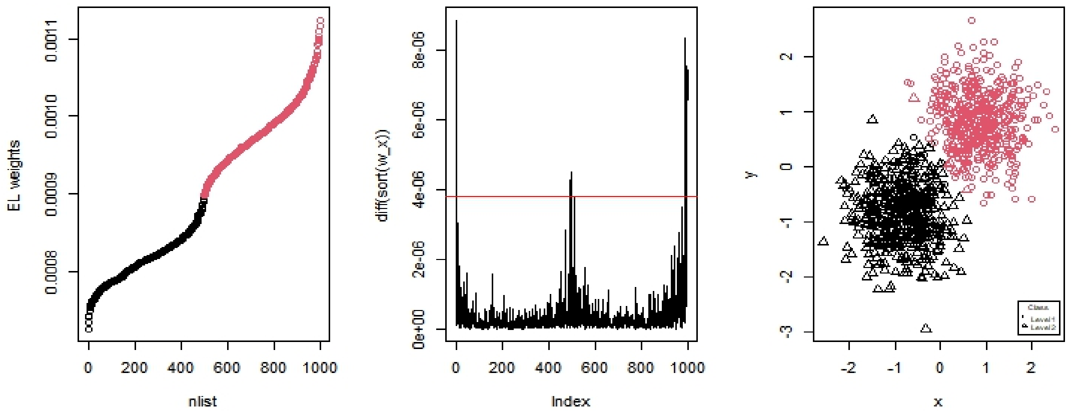


Figure 3.8: The plots of non-disjoint clusters with the same sample size and densities

vector $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ and the covariance matrix $\begin{pmatrix} 3 & -0.9 \\ -0.9 & 3 \end{pmatrix}$ and the other containing 50 observations with the mean vector $\begin{pmatrix} 5 \\ 5 \end{pmatrix}$ and the covariance matrix $\begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}$. From Figure 3.9, we see that these two clusters are slightly overlapped, where the two clusters are indicated by the circle and triangle symbols, respectively.

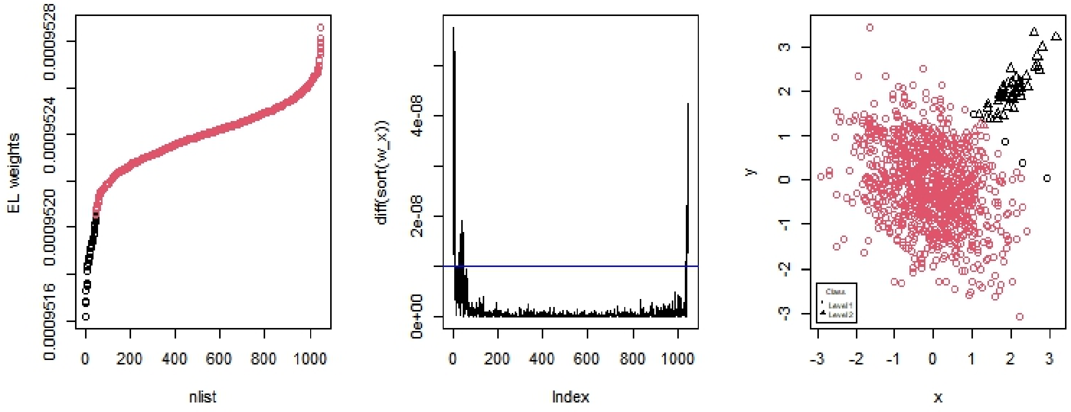


Figure 3.9: The plots of non-disjoint clusters with different sample sizes and densities

Again, we calculate the EL weights of the dataset and present the sorted EL weights in the left panel of Figure 3.9. From the middle panel of Figure 3.9, we find that the variations in the initial segment of the differences between all consecutive values of the sorted EL weights are relatively larger than those in the remaining part. In this case, the threshold is not selected to find the value of differences that can directly split the clusters, but to capture the change of variations in the sorted EL weights for different clusters. We set the threshold D to be the 98%-quantile of differences of sorted EL weights, indicated by the blue line in Figure 3.9. Since most difference values in the initial segment are larger than D, the dataset is split into two clusters (red and black) after combining the clusters with lengths less than 5. The result is provided in the right panel of Figure 3.9. The precision and recall for the red cluster are 0.995 and 0.996, and for the black cluster are 0.918 and 0.9, respectively.

3.4 An empirical example

We apply our method to the famous Iris dataset (Anderson, 1935; Fisher1936), which consists of 3 species (Setosa, Virginica, and Versicolor) with 50 samples and 4 features (the length and the width of the sepals and petals) from each species. As can be seen from Figure 3.10, Setosa (black) is well-separated from the others, while Virginia (blue) and Versicolor (red) overlap considerably. Thus, clustering for the Virginia and Versicolor groups is a well-known challenge problem.

The Iris dataset contains 4 features, however, some features might contain redundant or little information. Thus, feature selection is suggested in clustering (Alelyani et al., 2014; Witten and Tibshirani, 2010). For the Iris dataset, the length and the width of petals are usually selected as the relevant features to help group clusters (Dash et al., 2002). Then, we compute the EL weights for these two features under mean constraints using a vector value of $(0.7602, 0.7880)$ which is the 3rd quantile of the dataset after standardization.

There is a huge gap and some relatively large gaps in the sorted EL weights, as shown in Figure 3.11. We set the threshold D to be the 10^{th} largest value of the difference of sorted EL weights. After merging clusters with observations less than 10, the iris dataset is split into three clusters in the dataset. The Setosa group is partitioned correctly, while the Virginia and Versicolor groups are split with 4 misclassified observations.

Then, we show the performance of using other clustering methods to compare with the above results, and these results are plotted in Appendix C. DBSCAN might suggest there are two clusters in the Iris dataset, which identifies the Virginia and Versicolor groups

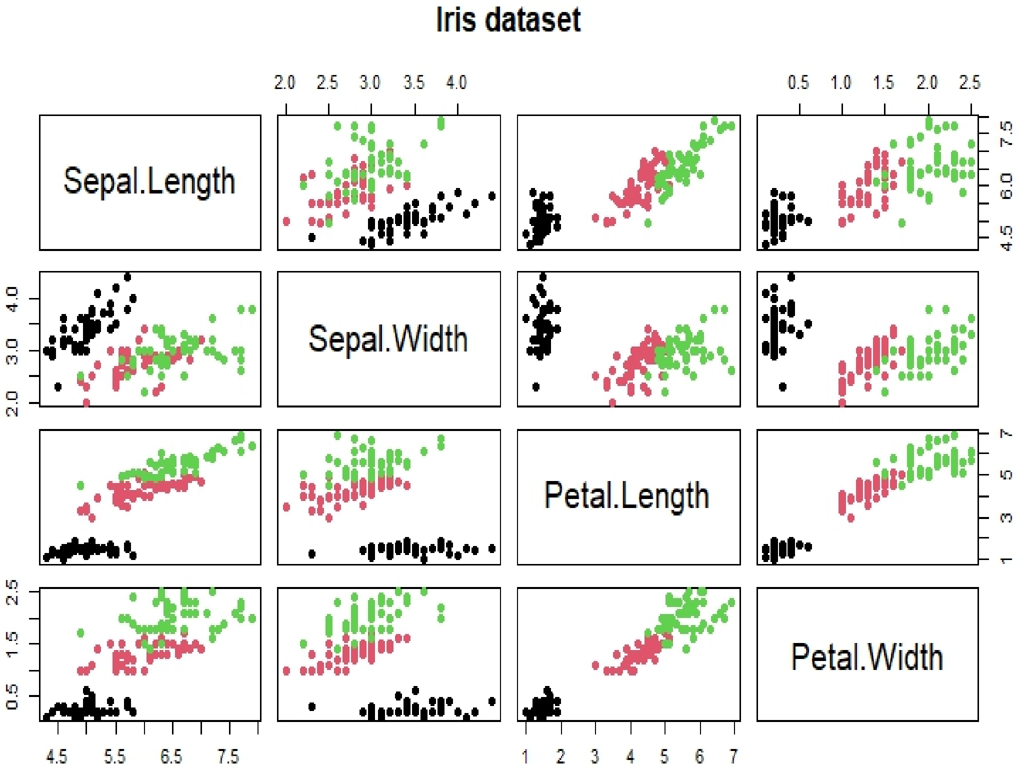


Figure 3.10: The scatterplot of the iris dataset

as one cluster. Moreover, the dendrogram of hierarchical clustering may yield different suggestions about the number of clusters. Given the number of clusters is three, K-means and K-medoids might give 6 misclassified observations from Virginia and Versicolor groups. Therefore, our EL-based clustering method is a feasible competitor to these methods in this case.

3.5 Conclusion

This paper introduces a clustering algorithm based on the EL weights. The proposed algorithm does not require specifying the number of clusters. Although a threshold of the difference of sorted EL weights is required, it is easy to select for the well-separated

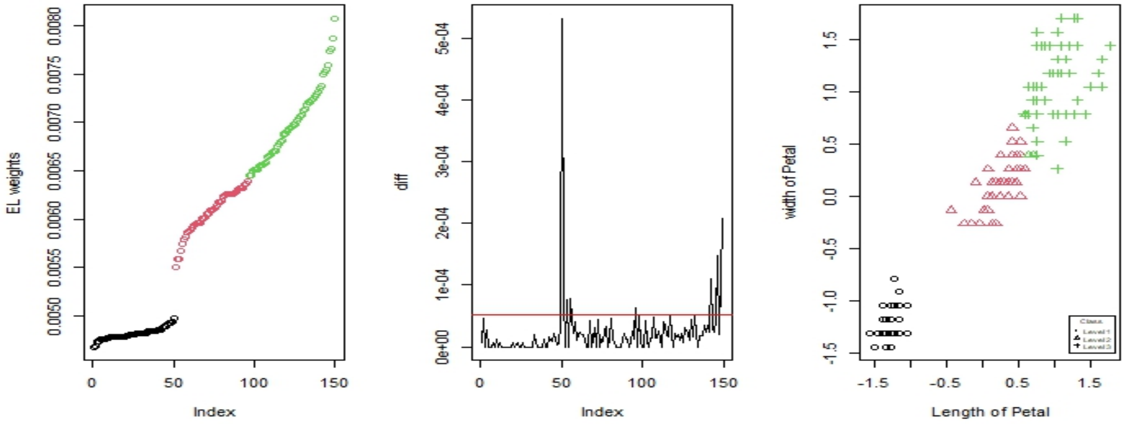


Figure 3.11: The results of the iris dataset

clusters. Moreover, this algorithm is less sensitive to the outliers. It might detect the outliers while grouping the dataset into clusters. In addition, this paper focuses on the EL weights under mean or variance constraint as exposition. However, the selection of constraints is a potential future work. Future research might also explore the feature selection for clustering based on the EL method.

References

Abdullah, Z., T. Herawan, N. Ahmad, and M. Deris (2011): "Extracting highly positive association rules from students' enrollment data," *Procedia-Social and behavioural Sciences*, 28, 107–111.

Adam, M. B. and J. A. Tawn (2016): "Modelling record times in sport with extreme value methods," *Malaysian Journal of Mathematical Sciences*, 10, 1–21.

Alelyani, S., J. Tang, and H. Liu (2014): *Feature selection for clustering: a review*, United Kingdom: CRC Press.

Anatolyev, S. (2005): "GMM, GEL, serial correlation, and asymptotic bias," *Econometrica*, 73, 983–1002.

Anderson, E. (1935): "The irises of the gaspé peninsula," *Bulletin of the American Iris Society*, 59, 2–5.

Anderson, P. L. and M. M. Meerschaert (1998): "Modeling river flows with heavy tails," *Water Resources Research*, 34, 2271–2280.

Bader, B., J. Yan, and X. Zhang (2018): "Automated threshold selection for extreme value analysis via ordered goodness-of-fit tests with adjustment for false discovery rate," *The Annals of Applied Statistics*, 12, 310–329.

Bai, Y., W. Fung, and Z. Zhu (2010): "Weighted empirical likelihood for generalized linear models with longitudinal data," *Journal of Statistical Planning and Inference*, 140, 3446–3456.

Balakrishnan, N. and V. Nevzorov (2004): *A primer on statistical distributions*, New Jersey: Wiley.

Barabási, A. (2005): "The origin of bursts and heavy tails in human dynamics," *Nature*, 435, 207–211.

Baragona, R., F. Battaglia, and D. Cucina (2013): "Empirical likelihood for break detection in time series," *Electronic Journal of Statistics*, 7, 3089–3123.

——— (2016): "Empirical likelihood for outlier detection and estimation in autoregressive time series," *Journal of Time Series Analysis*, 37, 315–336.

——— (2018): "Empirical likelihood for outlier detection in regression models," *Journal of Statistical Theory and Practice*, 12, 255–281.

Benmouiza, K. and A. Cheknane (2019): "Clustered ANFIS network using fuzzy c-means, subtractive clustering, and grid partitioning for hourly solar radiation forecasting," *Theoretical and Applied Climatology*, 137, 31–43.

Bernardara, P., D. Schertzer, E. Sauquet, I. Tchiguirinskaia, and M. Lang (2007): "The flood probability distribution tail: how heavy is it?" *Stochastic Environmental Research and Risk Assessment*, 22, 107–122.

Binbusayyis, A. and T. Vaiyapuri (2021): "Unsupervised deep learning approach for network intrusion detection combining convolutional autoencoder and one-class SVM," *Applied Intelligence*, 51, 7094–7108.

Caeiro, F. and M. I. Gomes (2014): "A semi-parametric estimator of a shape second-order parameter," in *New Advances in Statistical Modeling and Applications*, Springer, 137–144.

——— (2016): "Threshold selection in extreme value analysis," *Extreme Value Modeling and Risk Analysis: Methods and Applications*, 1, 69–86.

Cai, S. and J. Chen (2018): "Empirical likelihood inference for multiple censored samples," *The Canadian Journal of Statistics*, 46, 212–232.

Chang, J., S. X. Chen, and X. Chen (2015): "High dimensional generalized empirical likelihood for moment restrictions with dependent data," *Journal of Econometrics*, 185, 283–304.

Chang, Y. H., K. K. Lai, C. Y. Lin, F. P. Su, and M. C. Yang (2017): "A hybrid clustering approach to identify network positions and roles through social network and multivariate analysis," *Scientometrics*, 113, 1733–1755.

Chauhan, R. (2014): "A comprehensive study of various clustering techniques," *International Journal of Advanced Research in Computer Science*, 5.

Chen, J., A. Variyath, and B. Abraham (2008): "Adjusted empirical likelihood and its properties," *Journal of Computational and Graphical Statistics*, 17, 426–443.

Chen, K. and R. Huang (2021): "Robust empirical likelihood for time series," *Journal of Time Series Analysis*, 42, 4–18.

Chen, S., W. Härdle, and L. Ming (2003): "An empirical likelihood goodness-of-fit test for time series," *Journal of the Royal Statistical Society. Series B*, 65, 663–678.

Chen, S. and Y. Qin (2000): "Empirical likelihood confidence intervals for local linear smoothers," *Biometrika*, 87, 946–953.

——— (2003): "Coverage accuracy of confidence intervals in nonparametric regression," *Acta Mathematicae Applicatae Sinica*, 19, 387–396.

Chen, S. X. (1993): "On the accuracy of empirical likelihood confidence regions for linear regression model," *Annals of the Institute of Statistical Mathematics*, 45, 621–637.

——— (1994): "Empirical likelihood confidence intervals for linear regression coefficients," *Journal of Multivariate Analysis*, 49, 24–40.

Chen, T. and Z. Zhang (2022): "Where does the heaviness start?" *Working Paper.*

Chen, Y., D. Miao, and H. Zhang (2010): "Neighborhood outlier detection," *Expert Systems with Applications*, 37, 8745–8749.

Cheng, C., Y. Liu, Z. Liu, and W. Zhou (2018): "Balanced augmented jackknife empirical likelihood for two sample U-statistics," *Science China Mathematics*, 61, 1129–1138.

Chryssolouris, G., V. Subramanian, and M. Lee (1994): "Use of extreme value theory in engineering decision making," *Journal of Manufacturing Systems*, 13, 302–312.

Cirillo, P. and T. N. (2016): "On the statistical properties and tail risk of violent conflicts," *Physica A*, 452, 29–45.

Clauset, A. and R. Woodard (2013): "Estimating the historical and future probabilities of large terrorist events," *The Annals of Applied Statistics*, 7, 1838–1865.

Danielsson, J., L. de Haan, L. Peng, and C. G. de Vries (2001): "Using a bootstrap method to choose the sample fraction in tail index estimation," *Journal of Multivariate Analysis*, 76, 226–248.

Dash, M., K. Choi, P. Scheuermann, and L. H. (2002): "Feature selection for clustering - a filter solution," *2002 IEEE International Conference on Data Mining, Proceedings*, 115–122.

Dekkers, A., J. Einmahl, and L. de Haan (1989): "A moment estimator for the index of an extreme-value distribution," *The Annals of Statistics*, 17, 1833–1855.

DiCiccio, T., P. Hall, and J. Romano (1991): "Empirical likelihood is Bartlett-correctable," *The Annals of Statistics*, 19, 1053–1061.

DiCiccio, T. and J. Romano (1989): "On adjustments based on the signed root of the empirical likelihood ratio statistic," *Biometrika*, 76, 447–456.

Diebold, F., T. Schuermann, and J. Stroughair (2000): "Pitfalls and opportunities in the use of extreme value theory in risk management," *The Journal of Risk Finance*, 1, 30–35.

Draisma, G., L. de Haan, L. Peng, and T. T. Pereira (1999): "A bootstrap-based method to achieve optimality in estimating the extreme-value index," *Extremes*, 2, 367–404.

Drees, H., L. de Haan, and S. Resnick (2000): "How to make a hill plot," *The Annals of Statistics*, 28, 254–274.

Dutt, A., S. Aghabozrgi, M. A. B. Ismail, and H. Mahroeian (2015): "Clustering algorithms applied in educational Sata mining," *International Journal of Information and Electronics Engineering*, 5, 112.

Einmahl, J. H. J. and J. R. Magnus (2008): "Records in athletics through extreme-value theory," *Journal of the American Statistical Association*, 103, 1382–1391.

Embrechts, P., C. Klüppelberg, and T. Mikosch (1997): *Modelling eextremal events: for insurance and finance*, Berlin: Springer.

Ester, M., H. P. Kriegel, J. Sander, and X. Xu (1996): "A density-based algorithm for discovering clusters in large spatial databases with noise," *In kdd*, 96, 226–231.

Ezugwu, A. E., A. M. Ikotun, O. O. Oyelade, L. Abualigah, J. O. Agushaka, C. I. Eke, and A. A. Akinyelu (2022): "A comprehensive survey of clustering algorithms: state-of-the-art machine learning applications, taxonomy, challenges, and future research prospects," *Engineering Applications of Artificial Intelligence*, 110, 104743.

Fan, G., H. Xu, and H. Liang (2012): "Empirical likelihood inference for partially time-varying coefficient errors-in-variables models," *Electronic Journal of Statistics*, 6.

Forgy, E. (1965): "Cluster analysis of multivariate data: Efficiency vs. Interpretability of classifications," *Biometrics*, 21, 768–780.

Fuller, W. (1914): "Flood flows," *Transactions of the American Society of Civil Engineers*, 77, 564–617.

Gabaix, X. (2009): "Power laws in economics and finance," *Annual Review of Economics*, 1, 255–294.

Gabaix, X., P. Gopikrishnan, V. Plerou, and H. E. Stanley (2003): "A theory of power-law distributions in financial market fluctuations," *Nature*, 423, 267–270.

Gencay, R. and F. Selçuk (2004): "Extreme value theory and Value-at-Risk: Relative performance in emerging markets," *International Journal of Forecasting*, 20, 287–303.

Giesen, K. and J. Sudekum (2010): "Zipf's law for cities in the regions and the country," *Journal of Economic Geography*, 11, 667–686.

Glenn, N. and Y. Zhao (2007): "Weighted empirical likelihood estimates and their robustness properties," *Computational statistics & data analysis*, 51, 5130–5141.

Gnedenko, B. (1943): "Sur la distribution limite du terme maximum d'une série aléatoire," *Annals of Mathematics*, 44, 423–453.

Gomes, M. I., F. Figueiredo, and M. M. Neves (2012): "Adaptive estimation of heavy right tails: resampling-based methods in action," *Extremes*, 15, 463–489.

Gomes, M. I. and A. Guillou (2015): "Extreme value theory and statistics of univariate extremes: A review," *International Statistical Review*, 83, 263–292.

Gomes, M. I. and O. Oliveira (2001): "The bootstrap methodology in statistics of extremes – choice of the optimal sample fraction," *Extremes*, 4, 331–358.

Gomes, M. I. and D. Pestana (2007): "A sturdy reduced-bias extreme quantile (VaR) estimator," *Journal of the American Statistical Association*, 102, 280–292.

Gong, Y., L. Peng, and Y. Qi (2010): "Smoothed jackknife empirical likelihood method for ROC curve," *Journal of Multivariate Analysis*, 101, 1520–1531.

Grendár, M. and G. Judge (2009): "Empty set problem of maximum empirical likelihood methods," *Electronic Journal of Statistics*, 3, 1542–1555.

Guggenberger, P. and R. Smith (2005): "Genralized empirical likelihood estimators and tests under partial, weak, and strong identification," *Econometric Theory*, 21, 667–709.

Gumbel, E. (1958): *Statistics of extremes.*, London: Oxford University Press.

Hall, P. (1990): "Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems," *Journal of Multivariate Analysis*, 32, 177–203.

Hall, P. and A. H. Welsh (1985): "Adaptive estimates of parameters of regular variation," *The Annals of Statistics*, 13, 331–341.

Handcock, M. S., A. S. Raftery, and J. S. Tantrum (2007): "Model-based clustering for social networks," *Journal of the Royal Statistical Society: Series A*, 170, 301–354.

Hautamaki, V., I. Karkkainen, and P. Franti (2004): "Outlier detection using k-nearest neighbour graph," *Proceedings of the 17th International Conference on Pattern Recognition*, 3, 430–433.

Hill, B. (1975): "A simple general approach to inference about the tail of a distribution," *The Annals of Statistics*, 3, 1163–1174.

Hjort, N. L., I. McKeague, and I. Van Keilegom (2009): "Extending the scope of empirical likelihood," *The Annals of Statistics*, 37, 1079–1111.

Ikotun, A. M., A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming (2023): "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Information Sciences*, 622, 178–210.

Imbens, G. W. (2002): "Generalized method of moments and empirical likelihood," *Journal of Business & Economic Statistics*, 20, 493–506.

Jing, B., J. Yuan, and W. Zhou (2009): "Jackknife empirical likelihood," *Journal of the American Statistical Association*, 104, 1224–1232.

Juncosa, M. (1949): "The asymptotic behavior of the minimum in a sequence of random variables," *Duke Mathematical Journal*, 16, 609–618.

Kanter, M. and W. L. Steiger (1974): "Regression and autoregression with infinite variance," *Advances in applied probability*, 6, 768–783.

Katz, R. (1999): "Extreme value theory for precipitation: sensitivity analysis for climate change," *Advances in water resources*, 23, 133–139.

Katz, R. W., P. M. B., and P. Naveau (2002): "Statistics of extremes in hydrology," *Advances in Water Resources*, 25, 1287–1304.

Kaufman, L. and P. Rousseeuw (1990): *Finding groups in data: An introduction to cluster analysis*, New York: John Wiley & Sons.

Kiran, K. G. and S. V. V. (2021): "A Mahalanobis distance-based automatic threshold selection method for peaks over threshold model," *Water Resources Research*, 57.

Kolaczyk, E. (1994): "Empirical likelihood for generalized linear models," *Statistica Sinica*, 199–218.

Kotz, S. and S. Nadarajah (2000): *Extreme value distributions: Theory and applications*, London: Imperial College Press.

La Rocca, M. (1998): "Bootstrapping empirical likelihood for linear regression models," *NTTS-Conferences on New Techniques and Technologies for Statistics*, 277–282.

Lazar, N. (2003): "Bayesian empirical likelihood," *Biometrika*, 90, 319–326.

Lazar, N. A. (2021): "A review of empirical likelihood," *Annual Review of Statistics and its Application*, 8, 329–344.

Li, D. and J. Pan (2013): "Empirical likelihood for generalized linear models with longitudinal data," *Journal of Multivariate Analysis*, 114, 63–73.

Li, G. and Q. H. Wang (2003): "Empirical likelihood regression analysis for right censored data," *Statistica Sinica*, 51–68.

Liu, F. T., K. M. Ting, and Z.-H. Zhou (2008): "Isolation forest," *Proceedings of the 8th IEEE International Conference on Data Mining*, 413–422.

Liu, P. and Y. Zhao (2022): "A review of recent advances in empirical likelihood," *Computational Statistics*, 15, e1599.

Liu, Y. and J. Chen (2010): "Adjusted empirical likelihood with high-order precision," *The Annals of Statistics*, 38, 1341–1362.

MacQueen, J. (1967): "Some methods for classification and analysis of multivariate observations," *In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 281–297.

Maddala, G. (1992): *Introduction to Econometrics*, Macmillan Pub. Co.

Makles, A. (2012): "Stata tip 110: How to get the optimal k-means cluster solution," *The Stata Journal*, 12, 347–351.

Malmgren, R. D., D. B. Stouffer, A. E. Motter, and L. A. N. Amaral (2008): "A poissonian explanation for heavy tails in e-mail communication," *Proceedings of the National Academy of Sciences - PNAS*, 105, 18153–18158.

Mandelbrot, B. (1963): "The variation of certain speculative prices," *The Journal of Business*, 36, 394–419.

McNeil, A. (1999): "Extreme value theory for risk managers," *Departement Mathematik ETH Zentrum*, 12, 217–237.

Melnykov, V. and R. Maitra (2010): "Finite mixture models and model-based clustering," *Statistics Surveys*, 4, 80–116.

Melnykov, V. and G. Shen (2013): "Clustering through empirical likelihood ratio," *Computational Statistics & Data Analysis*, 62, 1–10.

Monti, A. (1997): "Empirical likelihood confidence regions in time series models," *Biometrika*, 84, 395–405.

Nelder, J. and R. Wedderburn (1972): "Generalized linear models," *Journal of the Royal Statistical Society, Series A*, 135, 370–384.

Newey, W. K. and R. J. Smith (2004): "Higher order properties of gmm and generalized empirical likelihood estimators," *Econometrica*, 72, 219–255.

Nordman, D. and S. Lahiri (2006): "A frequency domain empirical likelihood for short-and long-range dependence," *Annals of Statistics*, 34, 3019–3050.

——— (2014): "A review of empirical likelihood methods for time series," *Journal of Statistical Planning and Inference*, 155, 1–8.

Northrop, P. J. and C. L. Coleman (2014): "Improved threshold diagnostic plots for extreme value analyses," *Extremes*, 17, 289–303.

Owen, A. B. (1988): "Empirical likelihood ratio confidence intervals for a single functional," *Biometrika*, 75, 237–249.

——— (1990): "Empirical likelihood ratio confidence regions," *The Annals of Statistics*, 18, 90–120.

——— (1991): "Empirical likelihood for linear models," *The Annals of Statistics*, 19, 1725–1747.

——— (2001): *Empirical likelihood*, Chapman & Hall/CRC.

Park, H. S. and C. H. Jun (2009): "A simple and fast algorithm for k-medoids clustering," *Expert systems with applications*, 36, 3336–3341.

Qin, G. and B. Y. Jing (2001): "Empirical likelihood for censored linear regression," *Scandinavian Journal of Statistics*, 28, 661–673.

Qin, G. and M. Tsao (2003): "Empirical likelihood inference for median regression models for censored survival data." *Journal of Multivariate Analysis*, 85, 416–430.

Qin, J. and J. Lawless (1994): "Empirical likelihood and general estimating equations," *The Annals of Statistics*, 22, 300–325.

Raggiotto, F., D. Scarpi, and M. C. Mason (2019): "Faster! More! Better! Drivers of upgrading among participants in extreme sports events," *Journal of Business Research*, 102, 1–11.

Reed, W. J. (2003): "The pareto law of incomes – an explanation and an extension," *Physica A*, 319, 469–486.

Richardson, L. F. (1948): "Variation of the frequency of fatal quarrels with magnitude," *Journal of the American Statistical Association*, 43, 523–546.

Rossi-Hansberg, E. and M. L. J. Wright (2007): "Urban structure and growth," *The Review of Economic Studies*, 74, 597–624.

Scarf, P. and P. Laycock (1994): "Applications of extreme value theory in corrosion engineering," *Journal of Research of the National Institute of Standards and Technology*, 99, 313.

Scarrott, C. and A. MacDonald (2012): "A review of extreme value threshold estimation and uncertainty quantification," *Revstat-Statistical Journal*, 10, 33–60.

Schennach, S. (2007): "Point estimation with exponentially tilted empirical likelihood," *The Annals of Statistics*, 35, 634–672.

Schneider, L., A. Krajina, and T. Krivobokova (2021): "Threshold selection in uni-variate extreme value analysis," *Extremes*, 24, 881–913.

Schölkopf, B., R. C. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt (1999): "Support vector method for novelty detection," *Advances in Neural Information Processing Systems*, 12.

Schubert, E., J. Sander, M. Ester, H. P. Kriegel, and X. Xu (2017): "DBSCAN Revisited, Revisited: Why and how you should (still) use dbscan," *ACM Transactions on Database Systems (TODS)*, 42, 1–21.

Silva Lomba, J. and M. Fraga Alves (2020): "L-moments for automatic threshold selection in extreme value analysis," *Stochastic Environmental Research and Risk Assessment*, 34, 465–491.

Smith, R. J. (2011): "GEL criteria for moment condition models," *Econometric Theory*, 27, 1192–1235.

Sneath, P. (1957): "The application of computers to taxonomy," *Journal of General Microbiology*, 17, 201–226.

Sokal, R. and C. Michener (1948): "A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons," *Biologiske Skrifter*, 5, 1–34.

——— (1958): "A statistical method for evaluating systematic relationships," *University of Kansas Science Bulletin*, 38, 1409–1438.

Song, X. and H. Cui (2003): "An extended empirical likelihood for generalized linear models," *Statistica Sinica*, 69–81.

Steinhaus, H. (1956): "Sur la division des corps matériels en parties," *Bulletin de l'Académie Polonaise des Sciences*, 3, 801–804.

——— (2020): "Clustering benchmark datasets exploiting the fundamental clustering problems," *Data in Brief*, 30, 105501.

Sun, Z., Y. Jiang, and X. Ye (2019): "Improved statistical inference on semiparametric varying-coefficient partially linear measurement error model," *Journal of Nonparametric Statistics*, 31, 549–566.

Tang, C. and C. Leng (2011): "Empirical likelihood and quantile regression in longitudinal data analysis," *Biometrika*, 98, 1001–1006.

Tax, D. and R. Duin (2004): "Support vector data description," *Machine Learning*, 54, 45–66.

Tian, R. and L. Xue (2014): "Generalized empirical likelihood inference in generalized linear models for longitudinal data," *Communications in Statistics - Theory and Methods*, 43, 3893–3904.

Toulemonde, G., P. Ribereau, and P. Naveau (2015): "Applications of extreme value theory to environmental data analysis," *Extreme Events: Observations, Modeling, and Economics*, 7–1.

Ultsch, A. (2004): "Strategies for an artificial life system to cluster high dimensional data," *Abstracting and Synthesizing the Principles of Living Systems*, 6, 128–137.

Wang, C., Z. Tan, and T. Louis (2011): "Exponential tilt models for two-group comparison with censored data," *Journal of Statistical Planning and Inference*, 141, 1102–1117.

Wang, H. and Z. Zhu (2011): "Empirical likelihood for quantile regression models with longitudinal data," *Journal of Statistical Planning and Inference*, 141, 1603–1615.

Wang, Q. and J. Rao (2002): "Empirical likelihood-based inference in linear models with missing data," *Scandinavian Journal of Statistics*, 29, 563–576.

Ward, J. (1963): "Hierarchical grouping to optimize an objective function," *Journal of the American Statistical Association*, 58, 236–244.

Whang, Y. J. (2006): "Smoothed empirical likelihood methods for quantile regression models," *Econometric Theory*, 22, 173–205.

Witten, D. and R. Tibshirani (2010): "A framework for feature selection in clustering," *Journal of the American Statistical Association*, 105, 713–726.

Wiwie, C. B., J. Baumbach, and R. Röttger (2015): "Comparing the performance of biomedical clustering methods," *Nature Methods*, 12, 1033–1038.

Wu, C. (2004): "Weighted empirical likelihood inference," *Statistics & Probability Letters*, 66, 67–79.

Xi Chen, S. and H. Cui (2003): "An extended empirical likelihood for generalized linear models," *Statistica Sinica*, 69–81.

Xue, D., L. Xue, and W. Cheng (2011): "Empirical likelihood for generalized linear models with missing responses," *Journal of Statistical Planning and Inference*, 141, 2007–2020.

Yin, C., M. Ai, X. Chen, and X. Kong (2023): "Empirical likelihood for generalized linear models with longitudinal data," *Journal of Systems Science and Complexity*, 36, 2100–2124.

Zang, Y., Q. Zhang, S. Zhang, Q. Li, and S. Ma (2017): "Empirical likelihood test for high dimensional generalized linear models," *Big and Complex Data Analysis: Methodologies and Applications*, 29–50.

Zhao, P. and L. Xue (2013): "Empirical likelihood inferences for semiparametric instrumental variable models," *Journal of Applied Mathematics and Computing*, 43, 75–90.

Zhong, X. and M. Ghosh (2016): "Higher-order properties of Bayesian empirical likelihood," *Electronic Journal of Statistics*, 10, 3011–3044.

Zhu, H., J. G. Ibrahim, N. Tang, and H. Zhang (2008): "Diagnostic measures for empirical likelihood of general estimating equations," *Biometrika*, 95, 489–507.

APPENDICES

Appendix A

Appendices of Chapter 1

A.1 Additional Hill plots

This appendix provides additional examples of Hill plots for Pareto and Mixed distributions. As we know, k can be chosen from the "first stable" region in the Hill plot, such as randomly selecting a fixed number from the region or using the average of the number value in the region. However, the selection of the "first stable" region in the Hill plot is very subjective. To choose an optimal k based on the Hill plot, we sometimes need to change the range a few times.

The top panels of Figure A.1 and A.2 display the Hill plots for Pareto and Mixed distributions with the entire dataset range. However, it's not easy to detect the pattern and find the "first stable" region in the top panels of Figure A.1 and A.2. We might change the range for better clarity, as shown in the bottom panels of Figure A.1 and A.2. Even if a "first stable" region is found in the Hill plots of datasets with the entire dataset range, the selection of this region might change after zooming in on the Hill plots. This is because additional fluctuations that were not visible at a broader scale might be uncovered.

Moreover, researchers might study a bundle of datasets at one time. Then, it is an exceedingly heavy workload and extremely time-consuming to check each Hill plot and find the "first stable" region. Thus, a region might be chosen for the Hill plot for the entire bundle of datasets. Typically, k is bigger than $10\%n$ and smaller than $20\%n$. For example, if $n = 500$, we might select k from $[10, 11, ..., 100]$; if $n = 1000$, we might select k from $[20, 21, ..., 200]$; if $n = 2000$, we might select k from $[50, 52, ..., 250]$. In Figure A.3, we present the Hill plots for a bundle of Pareto(6) with 10000 observations, and the region

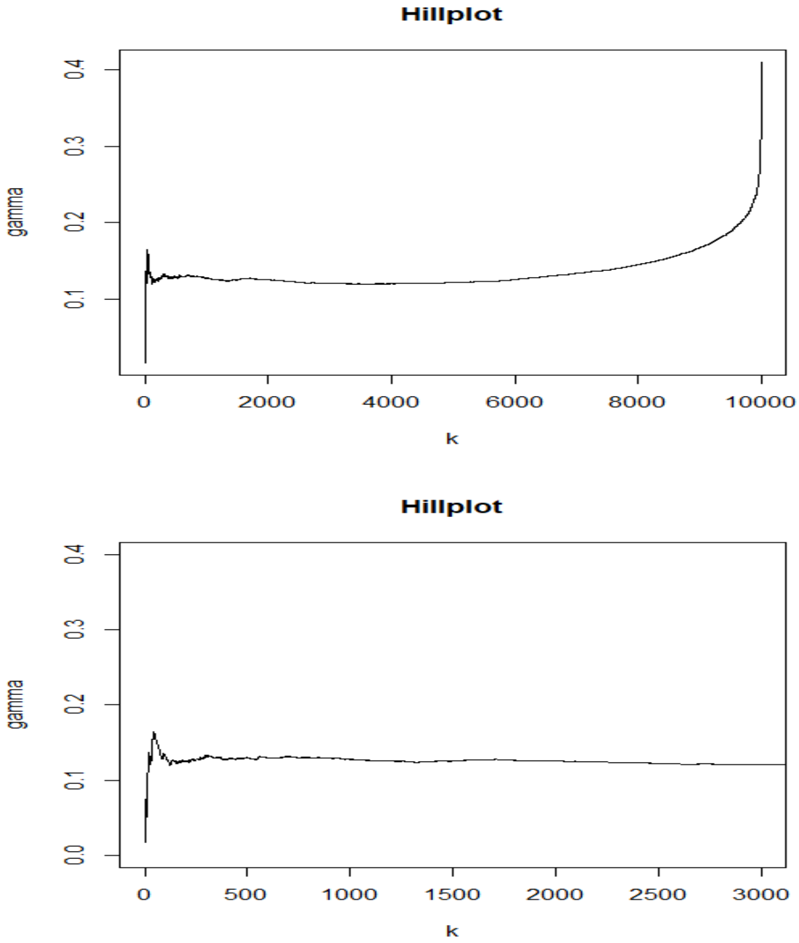


Figure A.1: The Hill plots for Mixed distribution

[500, 1000] is indicated by two solid lines. From Figure A.3, we can see that the region [500, 1000] is not the optimal selection for all the datasets, even though the datasets are generated from the same distribution with the same sample size. For example, the region [500, 1000] could be the “first stable” region for the first plot in the top panel of Figure A.3, however, the “first stable” region of the next plot seems should be [250, 400].

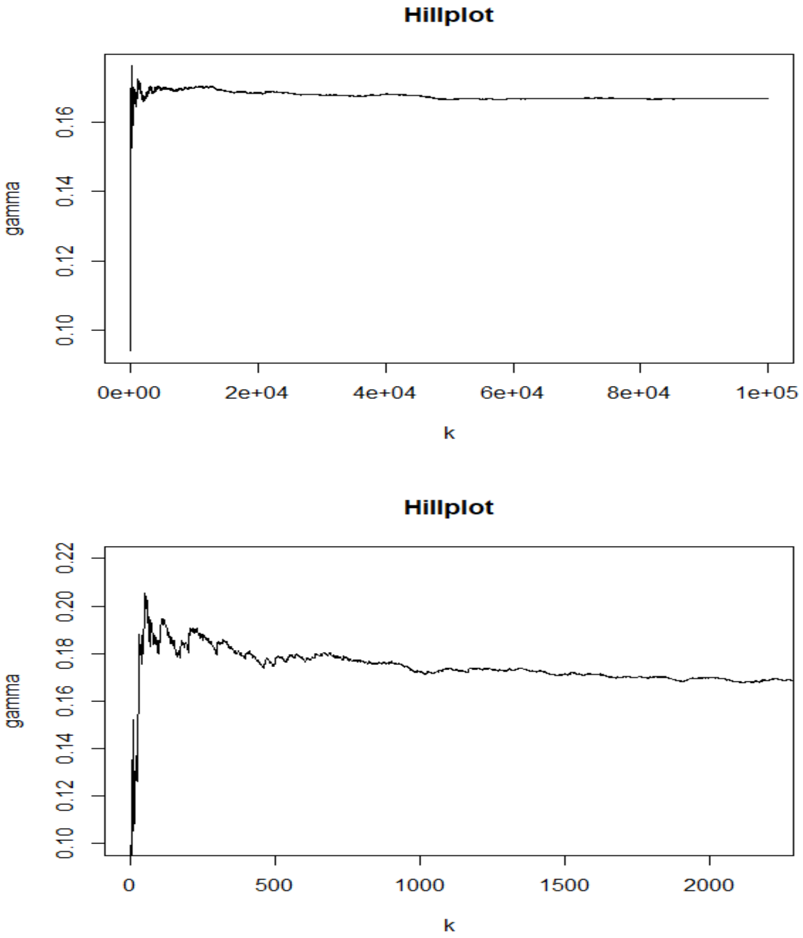


Figure A.2: The Hill plots for Pareto(6)

A.2 The plots for simulation results

This section plots the probability for simulation datasets and adds lines indicating threshold K to each plot to enhance the visibility of the results, as illustrated in Figure A.4 and A.5. The two dashed lines indicate the minimum and maximum value of threshold K, while the solid line indicates the representative number $\hat{\tau}$, that is, the mean of K.

Figure A.4 includes four plots, which are the probability plots for Pareto(6), TPareto(1.5), Mixed(6) and Mixed(1.5). In these four plots, the threshold K is estimated using our EL weights-based method, with the benchmark set as Exponential distribution. We see that

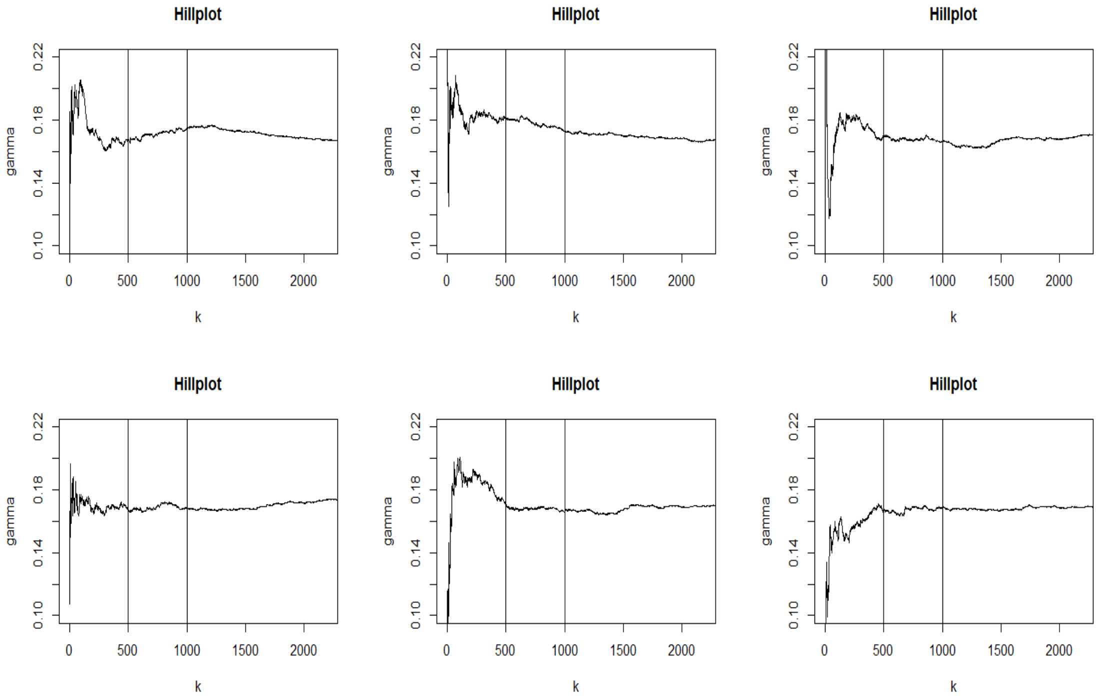


Figure A.3: The Hill plots with the selected region

the value of K is close to the turning point for all cases. Note that the mixed distribution is a mixture of 90% Pareto distribution and 10% Normal distribution such that the heaviness of the mixed distribution is only driven by the Pareto component. Thus, the difference between the results for Pareto and mixed distributions is very small.

Next, the Exponential distribution benchmark is replaced by Pareto(6) for the targeted dataset TPareto(1.5). The tail of Pareto(6) is heavier than the tail of Exponential distribution, therefore, the heaviness starts at a higher quantile, as shown in the first plot of Figure A.5. Then, if we force TPareto(1.5) to be the benchmark and Pareto(6) to be the distribution of interest, we get the same K due to the “role reversal” property of our algorithm.

Moreover, our method allows the tail index of the target distribution and benchmark both to be non-positive numbers. We force a Beta(2,5) which has a negative tail index to be the target distribution and an Exponential distribution which has a zero tail index to be the benchmark. Since the tail of the Exponential distribution is heavy relative to a Beta distribution, it gives the threshold K of the Exponential distribution with benchmark

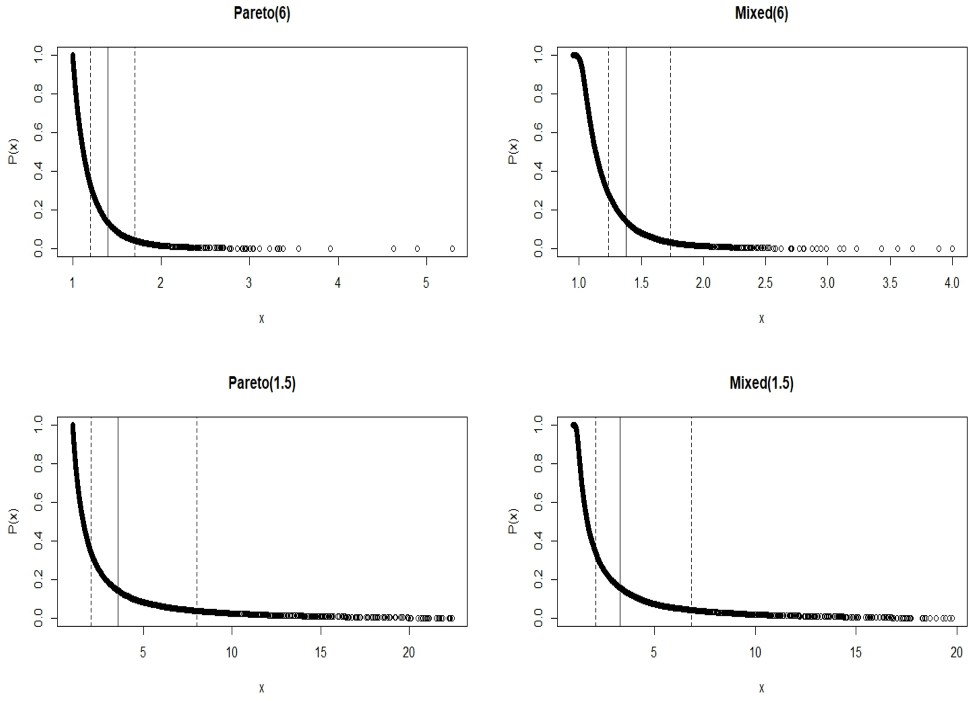


Figure A.4: Plots for simulation data with exponential benchmark

Beta(2,5) by the “role reversal” property. The results for Beta(2,5) with Exponential as the benchmark and Exponential with Beta(2,5) as the benchmark are shown in the two plots at the bottom of Figure A.5.

A.3 The EVT method

In this section, we give preliminary information on EVT. EVT studies the behaviour of extreme events, that is, the tail distribution. As early as 1709, the idea of EVT was mentioned by Niclas Bernouilli (Kotz and Nadarajah, 2000). EVT was first applied to study massive floods, such as 100 years flood, by Fuller, 1914. Then, the theoretical properties of EVT were developed by numerous works (Gnedenko, 1943; Juncosa, 1949 and Gumbel, 1958). As a consequence, EVT has been widely used in many fields, such as engineering (Chryssolouris et al., 1994; Scarf and Laycock, 1994), environmental science(Katz, 1999; Toulemonde et al., 2015), finance (McNeil, 1999; Diebold et al., 2000; Gencay and Selçuk

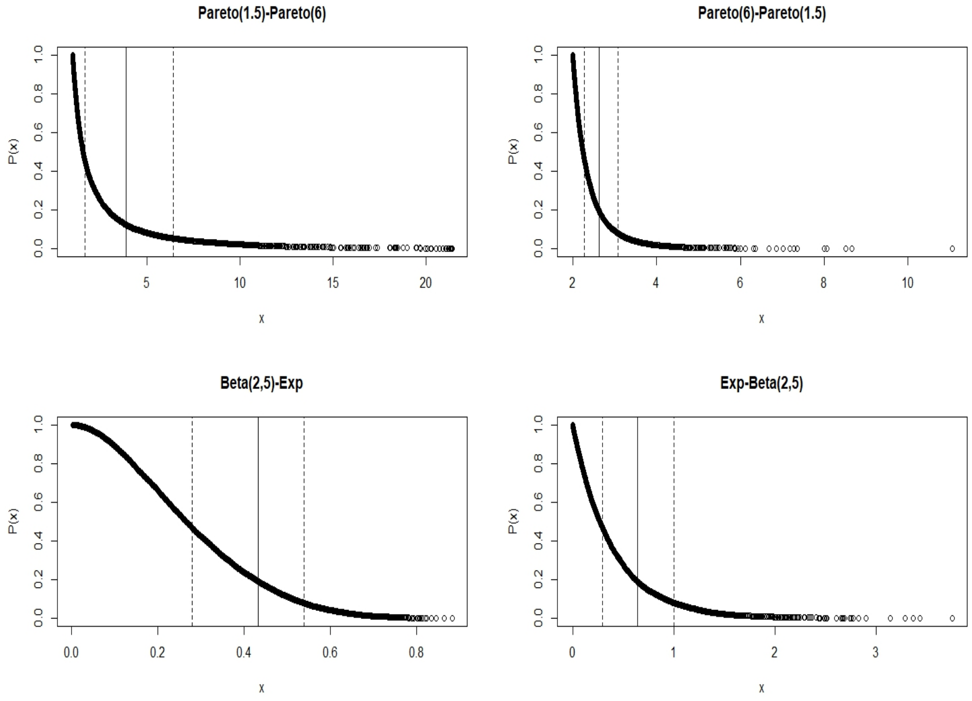


Figure A.5: Plots for simulation data with other benchmarks

(2004)), and sports (Einmahl and Magnus, 2008; Raggiotto et al., 2019; Adam and Tawn, 2016).

Let $X_1, X_2, ..., X_n$ be independent and identically distributed random variables with the same cumulative distribution function F. Let $X_{1,n}, X_{2,n}, ..., X_{n,n}$ be the associated order statistics, where $X_{n,n}$ is the maximum. Then, we have

$$P(X_{n,n} \leq x) = P(X_{1,n} \leq x, X_{1,n} \leq x, ..., X_{n,n} \leq x) = F^n(x).$$

Based on EVT, we have that if the maximum $X_{n,n}$, with appropriate centring and scaling, converges to a nondegenerate random variable, then

$$P\left(\frac{X_{n,n} - b_n}{a_n} \leq x\right) = F^n(a_n x + b_n) \rightarrow G_\gamma(x) \tag{A.1}$$

as $n \rightarrow \infty$, where

$$G_\gamma(x) := \exp(-(1 + \gamma x)^{-1/\gamma})$$

from some $\gamma \in \mathbb{R}$ with x such that $1 + \gamma x > 0$, and $\{a_n\}$ and $\{b_n\}$ are sequences of constant with $a_n > 0$. We say that F is in the domain of attraction of the extreme value distribution function $G_\gamma(x)$, and γ is the extreme value index or tail index which quantifies the heaviness of the tail part of the distributions. In general, the larger the tail index, the heavier the tail, implying that extreme events are more likely to occur.

Then, we can have

$$\lim_{t\to\infty} t(1 - F(a_t x + b_t)) = -\log G_{\gamma(x)} = (1 + \gamma x)^{-1/\gamma}, \quad G_{\gamma(x)} > 0,$$

where $t \in \mathbb{R}^+$, a_t and b_t are defined by interpolation.

Let $\bar{F}(1 - 1/t) := \inf\{x : F(x) \geq 1 - 1/t\}$ with $t > 1$. Then, the equation A.1 can be written as

$$\lim_{t\to\infty} \frac{U(tx) - U(t)}{a(t)} = \psi_\gamma(x) =: \begin{cases} \frac{x^\gamma - 1}{\gamma} & \gamma \neq 0 \\ \ln x & \gamma = 0, \end{cases}$$

where

$$U(t) := \bar{F}(1 - 1/t)$$

for all $x > 0$.

This is known as the first-order condition. However, a first-order condition is in general not sufficient to study the properties of estimators for tail parameters. Therefore, the second-order condition and third-order condition are required.

The second-order condition is about the rate of convergence, and is defined as follows,

$$\lim_{t\to\infty} \frac{\frac{U(tx)-U(t)}{a(t)} - \psi_\gamma(x)}{A(t)} = \begin{cases} \frac{1}{\rho}\left(\frac{x^\rho - 1}{\rho} - \ln x\right) & \gamma = 0, \rho \neq 0 \\ \frac{1}{\gamma}\left(x^\gamma \ln x - \frac{x^\gamma - 1}{\gamma}\right) & \gamma \neq 0, \rho = 0, \\ \ln^2 x/2 & \gamma = \rho = 0, \\ \frac{1}{\rho}\left(\frac{x^{\gamma+\rho} - 1}{\gamma+\rho} - \frac{x^\gamma - 1}{\gamma}\right) & \text{otherwise}, \end{cases}$$

for all $x > 0$, where ρ is a second-order parameter and A is a function possibly not changing in sign and $A \to 0$ as $t \to 1$.

The rate of convergence is also specified by the third-order condition,

$$U(t) = Ct^\gamma \left(1 + A(t)/\rho + \eta A^2(t) + o(A^2(t))\right),$$

where $A(t) := \gamma\beta t^{\rho}$, with $\gamma > 0, \rho < 0$ and $\beta, \eta \neq 0$. Note that, β and $\eta 0$ can more generally be slowly varying functions.

Now, let us discuss the estimation of γ, a_t and b_t. Since EVT focuses on the tail part whose behaviour might differ from the bulk part, we might not use the whole datasets to estimate γ, a_t and b_t. Let k be the tail parameter, which is the number of upper observations used to estimate the tail part of the dataset. We assume $k \to \infty$ and $k/n \to 0$ as $n \to \infty$.

There are different methods to estimate γ for different types of the extreme value distribution function $G_\gamma(x)$. The types of $G_\gamma(x)$ are presented as follows:

$$\text{Fréchet with } \gamma > 0: \quad \Phi_\gamma = \begin{cases} 0 & x < 0 \\ \exp\left(-x^{-1/\gamma}\right) & x \geq 0, \end{cases}$$

$$\text{Weibull with } \gamma < 0: \quad \Psi_\gamma = \begin{cases} \exp\left(-(-x)^{-1/\gamma}\right) & x \leq 0 \\ 1 & x > 0, \end{cases}$$

$$\text{Gumbel with } \gamma = 0: \quad \lambda_\gamma = \exp\left(-\exp(x)^{\gamma}\right) \quad x \in \mathbb{R}.$$

In this appendix, we only present two estimators mentioned in Chapter 1: Hill estimator and Moment estimator. We refer to Gomes and Guillou (2015) for a review of EVT estimation.

First, the Hill estimator introduced by Hill (1975) is the most famous estimator for $\gamma > 0$. It is defined as

$$\hat{\gamma}^{H}_{k,n} := \frac{1}{k}\sum_{i=0}^{k-1} \log X_{n-i,n} - \log X_{n-k,n}.$$

Second, the Moment estimator is an alternative estimator proposed by Dekkers et al. (1989). It does not restrict to the case $\gamma > 0$, and is defined as

$$\hat{\gamma}^{M}_{k,n} = M^{(1)}_{k,n} + 1 - \frac{1}{2}\left(1 - \frac{\left(M^{(1)}_{k,n}\right)^2}{M^{(2)}_{k,n}}\right)^{-1}, \tag{A.2}$$

where

$$M^{(j)}_{k,n} := \frac{1}{k}\sum_{i=0}^{k-1} \left(\log X_{n-i,n} - \log X_{n-k,n}\right)^{j}, \quad j = 1, 2.$$

One might find that $M_{k,n}^{(1)} \equiv \hat{\gamma}_{k,n}^{H}$.

Then, estimators for $a_{n,k}$, and $b_{n,k}$ are defined as follow

$$\hat{a}_{k,n} = \begin{cases} M_{k,n}^{(1)} X_{n-k,n}(1-\hat{\gamma}), & \hat{\gamma} \leq 1 \\ M_{k,n}^{(1)} X_{n-k,n}, & \hat{\gamma} > 1, \end{cases}$$

and

$$\hat{b}_{k,n} = X_{n-k,n}.$$

We should note that the estimated $\hat{\gamma}$, $\hat{a}_{k,n}$ and $\hat{b}_{k,n}$ are dependent on the selection of the tail parameter k. Ideally, the results for estimating γ should not be highly sensitive to small changes in the value of k. However, this is not always the case, especially for Hill estimator. If k is too large, $\hat{\gamma}$ is a biased estimator. If k is too small, $\hat{\gamma}$ is a volatile estimator. This might lead to a very peaked mean square error. Hence, some works, such as (Hall and Welsh, 1985), derived a formula by minimizing the asymptotic mean squared error for selecting k to balance the variance and bias.

For $k \to \infty$ and $k/n \to 0$, $\sqrt{k}A(\frac{n}{k}) \to \lambda$ as $n \to \infty$,

$$\sqrt{k}(\hat{\gamma} - \gamma) \xrightarrow{d} \gamma N + \frac{\lambda}{\rho},$$

where N is the standard normal distribution. Then, the asymptotic variance is

$$\frac{\gamma^2}{k},$$

and the asymptotic bias is

$$\frac{A^2(t)}{(1-\rho)^2}.$$

Nowm, we have the asymptotic mean squared error of γ is defined as

$$\frac{A^2(t)}{(1-\rho)^2} + \frac{\gamma^2}{k}.$$

Then, the k minimized the asymptotic mean squared error of γ is defined as

$$k_{\hat{\gamma}}(n) = \arg\min_{k} \left(\frac{A^2(t)}{(1-\rho)^2} + \frac{\gamma^2}{k} \right)$$

.

To minimize $k_{\hat{\gamma}}(n)$, the knowledge of the second-order parameter ρ is required. There are several methods to estimate ρ and find an optimal k based on minimizing the asymptotic mean squared error. In this thesis, we only present the results using the method discussed in Gomes et al. (2012) and Caeiro and Gomes (2016).

Appendix B

Appendices of Chapter 2

B.1 Proof of Lemma 2.2.1

Proof. By using the Lagrange multiplier method, we have $w_i = n^{-1}\left(1+\lambda\left(X_i-\mu_x\right)\right)^{-1}$, where the Lagrange multiplier λ is the unique solution of

$$f(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\frac{X_i-\mu_x}{1-\lambda(X_i-\mu_x)}.$$

From Owen (1990), we have that

$$\|\lambda\| = O_P(n^{-1/2}),$$

Therefore,

$$(1+\lambda\left(X_i-\mu_x\right)) \to 1 \text{ as } n\to\infty.$$

Recall that $w_i = n^{-1}\left(1+\lambda\left(X_i-\mu_x\right)\right)^{-1}$. Thus, $|w_i-\frac{1}{n}| \overset{P}{\to} 0$ as $n\to\infty$ for $i=1,2,...,n$. □

B.2 Proof of Theorem 2.2.1

Proof. Generally, we say the distribution of a constant follows Degenerate or Point Mass distribution. But, a constant can also be thought of as following a Discrete Uniform distribution. By the definition of Discrete Uniform distribution, the probability mass function (PMF) and cumulative distribution function (CDF) can be expressed as

$$P_X(X = x) = \begin{cases} \frac{1}{k} & \text{if } x = a, a+1, ..., b-1, b, \\ 0 & \text{otherwise,} \end{cases} \tag{B.1}$$

and

$$F_X(x) = \begin{cases} 0 & \text{if } x < a, \\ \frac{\lfloor x \rfloor - a + 1}{b - a + 1} & \text{if } x = a, a+1, ..., b-1, b, \\ 1 & \text{if } x > b, \end{cases} \tag{B.2}$$

where $a, b \in N$ with $a \geq b$ and $k = b - a + 1$.

Let $a = b = \frac{1}{n-m}$, then $k = 1$. It gives $P_X(X = x) = 1$ if $x = \frac{1}{n-m}$, and 0 otherwise. Similarly, $F_X(x) = 1$ if $x \geq \frac{1}{n-m}$, and 0 otherwise. The constant $\frac{1}{n-m}$ follows a Discrete Uniform distribution with $[\frac{1}{n-m}, \frac{1}{n-m}]$.

Based on Lemma 2.2.1, we have $|w_j - \frac{1}{n-m}| \overset{P}{\to} 0$ as $n - m \to \infty$ for $j = 1, 2, ..., n - m$. Therefore, $w_j \overset{D}{\to} DU(1)$ for $j = 1, 2, ..., n - m$. □

B.3 Proof of Lemma 2.2.2

Proof. Similar to the proof of Lemma 2.2.1, $w_i = n^{-1}(1 + \lambda(X_i - \mu_{normal}))^{-1}$, where the Lagrange multiplier λ is the unique solution of

$$f(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \frac{X_i - \mu_{normal}}{1 - \lambda(X_i - \mu_{normal})}.$$

It remains to show that $\lambda \nrightarrow 0$ as $n \to \infty$. Again, from Owen (1990), we have that

$$\lambda = S^{-1}(\bar{X} - \mu_{normal}) + o_P(n^{-1/2}),$$

where $S = \frac{1}{n} \sum (X_i - \mu_{normal})(X_i - \mu_{normal})'$.

If $\frac{m}{n} \nrightarrow 0$ as $n \to \infty$, $\mu_x \nrightarrow \mu_{normal}$ as $n \to \infty$ since μ_x is highly distorted by the outliers, implying that $\lambda \overset{P}{\nrightarrow} 0$.

□

B.4 Proof of Theorem 2.2.2

Proof. From Lemma 2.2.2, we know that

$$w_j = \frac{1}{n-m}\frac{1}{1+\lambda\left(X_j-\mu_{normal}\right)},$$

where $\lambda \overset{P}{\nrightarrow} 0$, so that $P_X(X = \frac{1}{n-m}) \neq 1$. □

B.5 Proof of Theorem 2.2.3

Proof. Pearson's limit theorem gives that the random variable $\sum_{i=1}^{b} \frac{(O_i-E_i)^2}{E_i}$ converges in distribution to a chi-square random variable with $b-1$ degrees of freedom as the sample size n increases to infinity. □

B.6 Proof of Theorem 2.2.4

Proof. Theorem 2.2.3 leads immediately to this result. □

B.7 Proof of Theorem 2.2.5

Proof. When (2.3) is true, $p_i \neq \frac{1}{b}$ for some i. Then,

$$\begin{aligned}\frac{O_i-(n-m)/b}{\sqrt{n-m/b}} &= \frac{O_i-(n-m)p_i+(n-m)(p_i-1/b)}{\sqrt{n-m/b}} \\ &= \sqrt{bp_i}\frac{O_i-(n-m)p_i}{\sqrt{(n-m)p_i}} + \sqrt{n-m}\frac{p_i-1/b}{\sqrt{1/b}}.\end{aligned}$$

The first term converges to $\mathcal{N}\left(0, \frac{b(1-p_i}{p_i}\right)$ and the second term diverges to ∞, which implies $\frac{(O_i-(n-m)/b)^2}{n-m/b} \to \infty$ as $n-m \to \infty$. Similarly, this divergence occurs in the $\frac{(\hat{O}_i-(n-\hat{m})/b)^2}{n-\hat{m}/b}$. □

B.8 Proof of Remark 2.2.1

Proof. We first recall the inequality of arithmetic and geometric mean. For a set of non-negative real numbers $X_1+,...,+X_n$, we have

$$\frac{X_1+,...,+X_n}{n} \geq \sqrt[n]{X_1\cdot,...,\cdot X_n}.$$

And, the equality holds if and only if $X_1 = X_2,...,= X_n$.

From the inequality of arithmetic and geometric means, we have

$$\prod_{i=1}^{n} w_i \leq \left(\frac{\sum_{i=1}^{n} w_i}{n}\right)^n,$$

and the equality holds if and only if $w_1 = w_2 = \cdots = w_n$. Thus, $\prod_{i=1}^{n} w_i$ is maximized when $w_1 = w_2 = \cdots = w_n$. Since $\sum_{i=1}^{n} w_i = 1$, we have $w_i = \frac{1}{n}$.

If a sample mean constraint is added, the optimal result for maximizing the function $\prod_{i=1}^{n} w_i$ is still $w_i = \frac{1}{n}$ for all i since the sample mean constraint is actually $\sum_{i=1}^{n} w_i X_i = \frac{1}{n}\sum_{i=1}^{n} X_i$. □

B.9 Skewness of EL weigths

This section provides the skewness of EL weights for all four cases in the simulation section of Chapter 2.

Table B.1 presents the skewness of EL weights for the 1D datsets, while Table B.2 presents the skewness of EL weights for the 2D datsets. The sign of the skewness of EL weights is not consistent, but we only focus on the absolute value of the skewness of EL weights. The absolute values of the skewness of EL weights for both X' and Y' do not have significant change, since these two datasets only contain normal observations. However, the absolute values of the skewness of EL weights for X decrease rapidly as the number of removed elements increases in the first 10 rows of Table B.1, similar to the absolute values of the skewness of EL weights for Y in the first 10 rows of Table B.2. In addition, the decreasing of the absolute values of the skewness becomes slow from $\left\{X_{(i)}\right\}_{i=1}^{n-10}$ and $\left\{Y_{(i)}\right\}_{i=1}^{n-10}$ since the normal observations do not distort the skewness. Therefore, we also get that $X_{(n)}, X_{(n-1)},...,X_{(n-9)}$ and $Y_{(n)}, Y_{(n-1)},...,Y_{(n-9)}$ are outliers for X and Y using the change of the skewness, respectively.

Table B.1: Skewness of EL weights for 1D cases

Num of Removed elements	Skewness of X	Skewness of X'
1	3.3377	0.0260
2	-2.8334	-0.0246
3	-2.3622	-0.0299
4	1.8709	0.0339
5	1.4601	-0.0299
6	1.0414	-0.0264
7	-0.6215	0.0345
8	-0.2670	-0.0279
9	-0.1231	0.0260
10	-0.0239	0.0271
11	-0.0280	-0.0212
12	-0.0251	0.0248

Table B.2: Skewness of EL weights for 2D cases

Num of Removed elements	Skewness of Y	Skewness of Y'
1	-10.6779	0.0124
2	-5.9027	0.0460
3	4.5306	-0.015
4	6.7020	-0.0396
5	-0.3912	0.0222
6	7.0295	-0.0175
7	-4.8056	0.0221
8	-2.0485	0.0241
9	-1.3487	0.0254
10	-0.0062	0.0281
11	0.0168	0.0037
12	0.0003	0.0071

B.10 Plot of EL weights

The plot of the EL weights for the four simulated datasets is provided in this section. The EL weights for normal observations are indicated by gray points, while the outliers are indicated by blue points. The red lines indicated $1/n$, that is, 1/10000 for each case. Figure B.1 and B.3 show the EL weights for datasets with outliers. Figure B.2 and B.4 show the EL weights for datasets without outliers. From the figures, we find that the EL weights for normal observations are located around 1/10000, while the EL weights for outliers are relatively far from 1/10000.

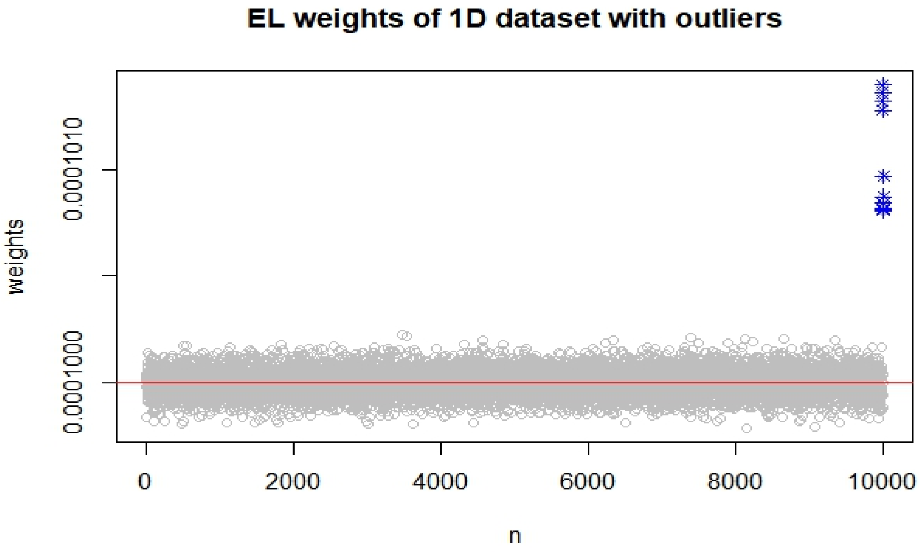


Figure B.1: The plot of EL weights of the 1D dataset with outliers

B.11 Additional simulations

In this section, we provide additional 1D and 2D simulation studies. Both 1D and 2D cases include datasets with and without outliers. For the 1D datasets, the case without outliers consists of 9990 and 10 outliers. The normal observations are drawn from a Gaussian distribution with mean 15 and variance 3. The 10 outliers are generated from two different distributions, where 5 outliers are drawn from a Gaussian distribution with mean 70 and variance 5 and the other 5 outliers are drawn from a Gaussian distribution with mean -10 and variance 2. Let us denote this 1D dataset as A. The case without outliers only contains 10000 observations drawn from a Gaussian distribution with mean 25 and variance 3, denoted as A'.

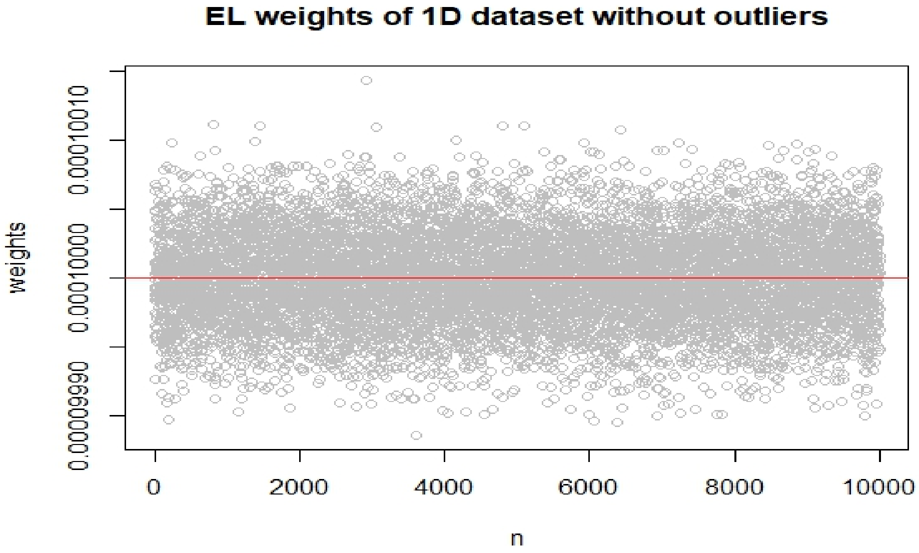


Figure B.2: The plot of EL weights of the 1D dataset without outliers

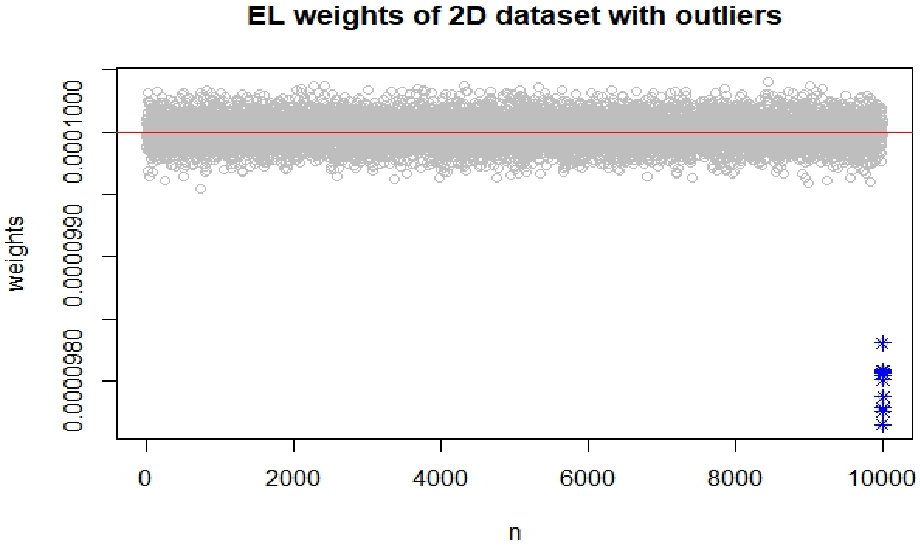


Figure B.3: The plot of EL weights of the 2D dataset with outliers

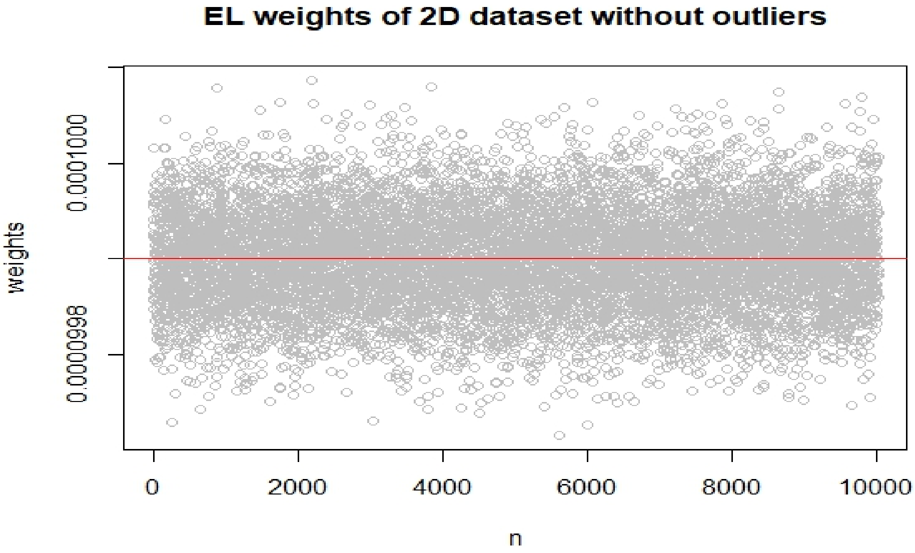


Figure B.4: The plot of EL weights of the 2D dataset without outliers

The EL weights for A and A' are illustrated in Figure B.5 and B.6, respectively. Figure B.5 shows the EL weights for A, where the outliers and normal observations are indicated by blue points and gray points, respectively. Figure B.6 shows the EL weights for A'. Since A' does not contains outliers, all observations in A' are around 1/10000 indicated by red lines. For both A and A', normal observations are around 1/10000, while outliers are relatively far from the 1/10000. Moreover, the outliers which are generated from two different distributions deviate 1/10000 in different directions.

Table B.3 presents the p-value of our outliers identification test for A and A'. The first column is the number of removed elements from the datasets. The second column presents the p-value of our outliers identification test for $\left\{A_{(i)}\right\}_{i=1}^{n-1}, ..., \left\{A_{(i)}\right\}_{i=1}^{n-12}$, respectively. The p-values of EL weights of $\left\{A_{(i)}\right\}_{i=1}^{n-1}, ..., \left\{A_{(i)}\right\}_{i=1}^{n-9}$ are statistically significant, implying the distribution of EL weights for them are not close to uniform distributions. After removing $A_{(n)}, A_{(n-1)}, ..., A_{(n-1)}$, the distribution of the EL weights starts to close to the uniform distribution. Thus, the p-values of EL weights of $\left\{A_{(i)}\right\}_{i=1}^{n-10}, ..., \left\{A_{(i)}\right\}_{i=1}^{n-15}$ in Table B.3 are not significant. Therefore, our method gives that the outliers in A are $A_{(n)}, A_{(n-1)}, ..., A_{(n-9)}$. The third column of Table B.3 shows the p-value from our outliers identification test of EL weights for $\left\{A'_{(i)}\right\}_{i=1}^{n-1}, ..., \left\{A'_{(i)}\right\}_{i=1}^{n-15}$, which are all not significant

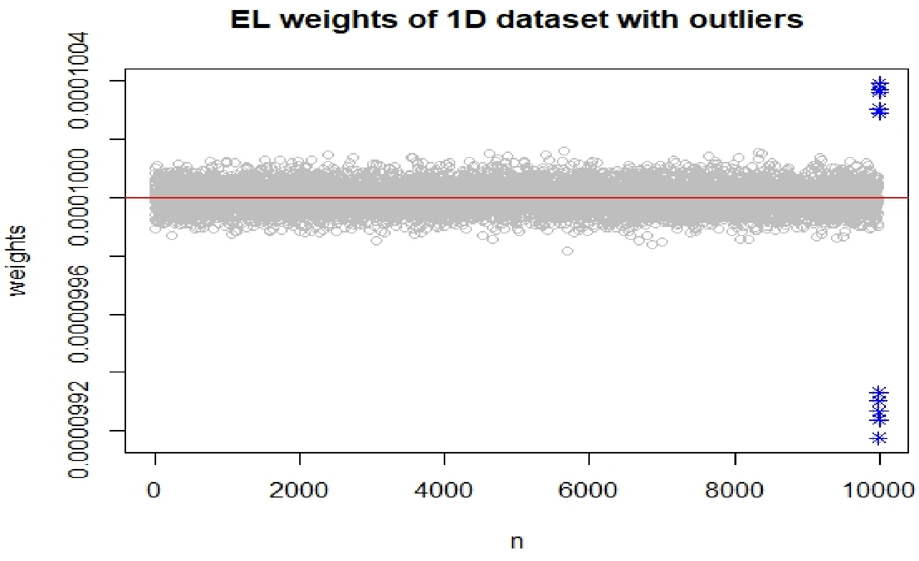


Figure B.5: Plots of EL weights for additional 1D cases with outliers

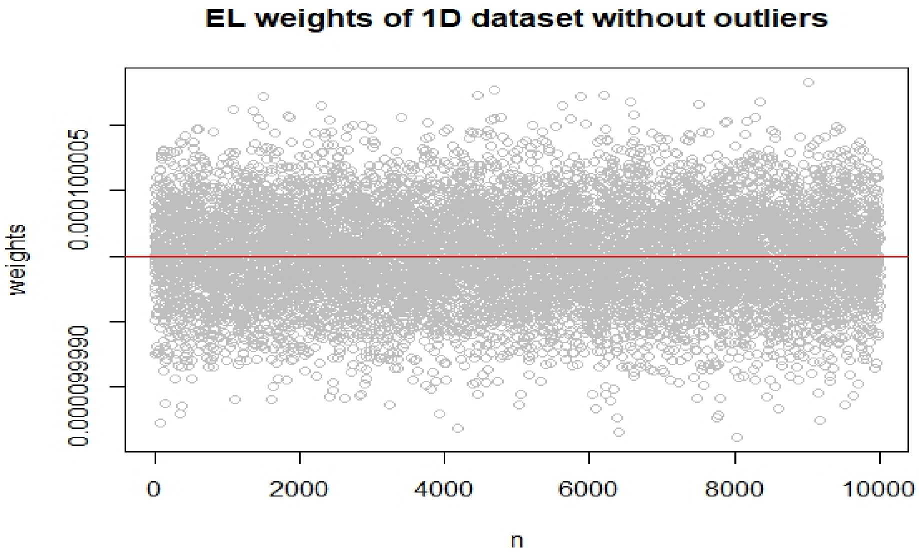


Figure B.6: Plots of EL weights for additional 1D cases without outliers

since A' does not contain outliers.

Table B.3: P-value of the outliers identification test for additional 1D cases

Num of Removed elements	P-value for A	P-value for A'
1	0	1
2	0	0.9999
3	8.2659e-309	1
4	2.1776e-284	1
5	3.4658e-59	1
6	3.2365e-51	1
7	2.3321e-54	1
8	5.3877e-27	0.9999
9	6.7099e-19	1
10	0.9999	0.9999
11	0.9999	0.9999
12	1	1
13	1	1
14	1	1
15	1	1

The skewness and kurtosis of the EL weights for both A and $A\prime$ are presented in Table B.4 and B.5, respectively. For the dataset $A\prime$, the skewness and kurtosis of EL weights do not significantly change as removing observations from the datasets. The kurtosis of EL weights for the dataset A is shown in the second column of Table B.5. We find that the kurtosis of EL weights in the first 6 row is relatively large. Although the kurtosis value in the 7^{th} row to 9^{th} row is close to those in the 10^{th} row to 15^{th} row, the difference of kurtosis in 9^{th} and the 10^{th} is still relatively large. Starting from $\left\{A_{(i)}\right\}_{i=1}^{n-10}$, the difference of kurtosis is about 0.01, while the difference between kurtosis of EL weights for $\left\{A_{(i)}\right\}_{i=1}^{n-9}$ and for $\left\{A_{(i)}\right\}_{i=1}^{n-10}$ is 0.169. It is one decimal point greater than 0.01. Similar results hold for the skewness.

Next, we generate two 2D datasets, one case consists of 9990 normal observations drawn from a bivariate Gaussian distribution with mean $[30, 25]$ and variance-covariance matrix $\left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$ and 10 outliers are from two different distributions. Of the 10 outliers, 7 outliers drawn from a bivariate Gaussian distribution with mean $[60, 70]$ and variance-covariance matrix $\left[\begin{smallmatrix} 5 & 0 \\ 0 & 5 \end{smallmatrix}\right]$ and 3 outliers drawn from a bivariate Gaussian distribution with mean $[1, 1]$

Table B.4: Skewness of EL weights for additional 1D cases

Num of Removed elements	Skewness of A	Skewness of A'
1	1.2850	-0.0023
2	0.8622	0.0007
3	-0.4514	-0.0019
4	0.0846	-0.0034
5	-0.2660	0.0057
6	0.2053	0.0043
7	0.1506	0.0004
8	0.0999	0.0075
9	-0.0638	0.0124
10	0.0395	0.0105
11	0.0470	-0.0035
12	0.0436	0.0024
13	0.0496	0.0062
14	0.0438	0.0039
15	-0.0353	-0.0033

and variance-covariance matrix $\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$. We denote this 2D dataset by B. The other case only contains 10000 observations drawn from the bivariate Gaussian distributions with mean $[30, 25]$ and variance-covariance matrix $\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$, denoted by B'.

The EL weights for B and B' are illustrated in Figure B.7 and B.8, respectively. Again, the outliers are indicated by blue points, while normal observations are indicated by gray points. For both B and B', we find that the EL weights for normal observations are around 1/10000 indicated by red lines. The EL weights of outliers deviate 1/10000 in different directions since they come from two different distributions. The distance between EL weights for normal observations and 1/10000 is much smaller than the distance between EL weights for outliers and 1/10000.

Table B.6 shows that the p-value of our outliers identification test for $\left\{B_{(i)}\right\}_{i=1}^{n-1}$, ..., $\left\{B_{(i)}\right\}_{i=1}^{n-9}$ are statistically significant so that their distributions are not close to uniform distributions. The p-values of the outliers identification test for B are not monotonically increasing as the number of removed elements increased in Table B.6, because EL weights in B are contributed by two variables from different dimensions. However, the p-values are still not significant starting from $\left\{B_{(i)}\right\}_{i=1}^{n-10}$. Thus, the outliers in B are

Table B.5: Kurtosis of EL weights for additional 1D cases

Num of Removed elements	Kurtosis of A	Kurtosis of A'
1	27.9327	3.0069
2	21.4092	2.9967
3	15.2061	2.9869
4	9.8338	2.9775
5	4.7951	2.9696
6	4.2682	.9619
7	3.8262	2.9534
8	3.4112	2.9494
9	3.2074	2.9416
10	3.0384	2.9350
11	3.0175	2.9288
12	3.0072	2.9223
13	2.9966	2.9160
14	2.9868	2.9097
15	2.9780	2.9038

$B_{(n)}, B_{(n-1)}, ..., B_{(n-9)}$. The third column of Table B.6 shows the p-value of the outliers identification test for $\left\{B'_{(i)}\right\}_{i=1}^{n-1}, ..., \left\{B'_{(i)}\right\}_{i=1}^{n-12}$, which are all not significant since B' is outlier free.

Furthermore, the skewness and kurtosis of the EL weights for the both B and B' are presented in Table B.7 and B.8. We see that both skewness and kurtosis of EL weights do not have significant change for the dataset B'. Again, the second column of Table B.7 and Table B.8 shows that the skewness and kurtosis of EL weights for the dataset B are not monotonically decreasing as the number of removed elements increased, because EL weights in the 2D case are contributed by two variables from different dimensions. However, the change of skewness and kurtosis still becomes relatively small starting from $\left\{B_{(i)}\right\}_{i=1}^{n-10}$. Thus, it also gives that the outliers are $B_{(n)}, B_{(n-1)}, ..., B_{(n-9)}$ in B.

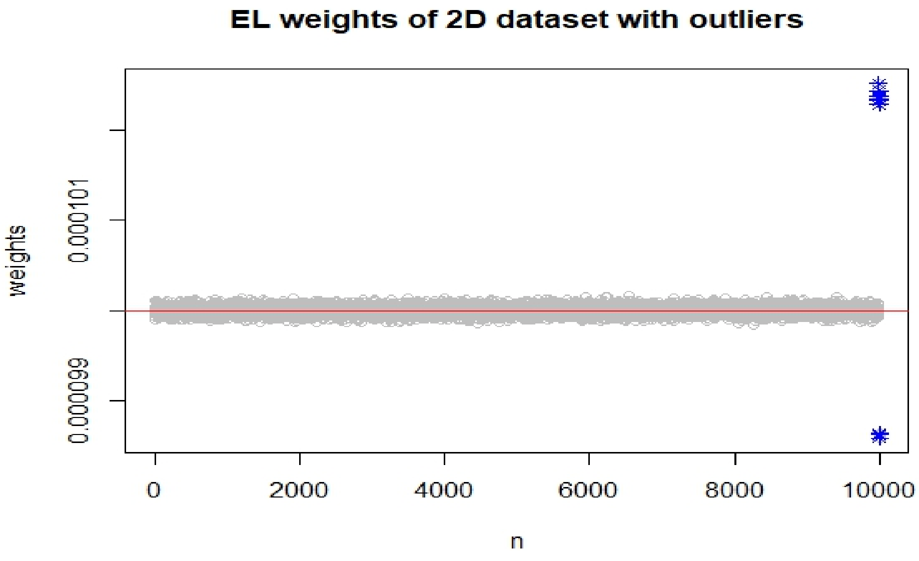


Figure B.7: Plots of EL weights for additional 2D cases with outliers

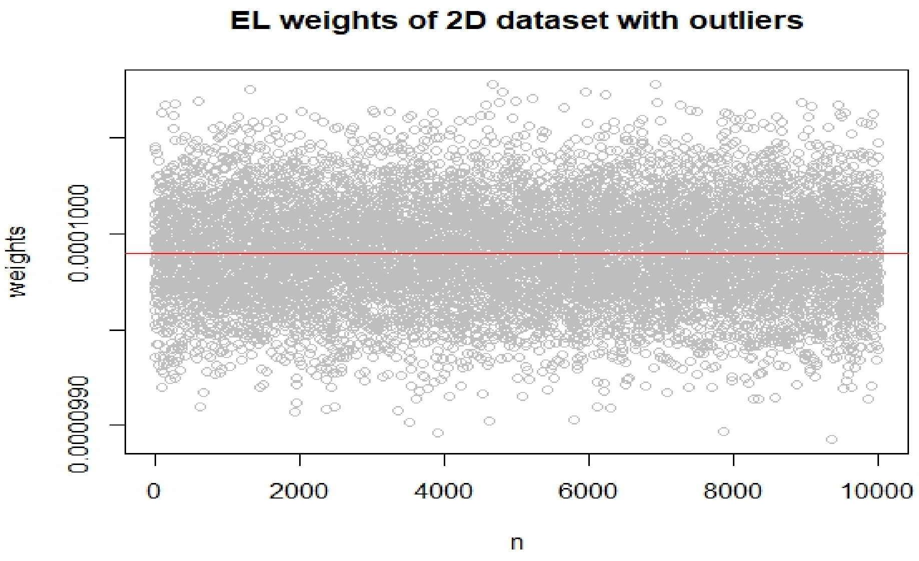


Figure B.8: Plots of EL weights for additional 2D cases without outliers

Table B.6: P-value of the outliers identification test for additional 2D cases

Num of Removed elements	P-value for B	P-value for B'
1	0	0.9999
2	0	0.9999
3	0	0.9729
4	8.3614e-89	0.9999
5	0	0.9908
6	3.5455e-310	1
7	0	1
8	1.1956e-60	0.9999
9	3.9790e-109	0.9999
10	0.9961	0.9999
11	0.9810	0.9999
12	1	0.9999
13	0.9635	0.9999
14	0.9999	0.9985
15	0.9999	0.9999

Table B.7: Skewness of EL weights for additional 2D cases

Num of Removed elements	Skewness of B	Skewness of B'
1	-1.4004	0.0054
2	15.6738	-0.0041
3	10.8835	-0.024
4	-0.2982	-0.0003
5	-0.6475	-0.0197
6	1.1077	-0.0109
7	-8.7595	0.0207
8	-0.1627	-0.0162
9	0.2041	-0.0063
10	0.0258	0.0138
11	0.0197	-0.0279
12	0.0253	0.0102
13	0.0102	0.0261
14	0.0089	-0.0311
15	-0.0004	0.0383

Table B.8: Kurtosis of EL weights for additional 2D cases

Num of Removed elements	Kurtosis of B	Kurtosis of B'
1	570.1453	3.0069
2	632.2354	2.9797
3	569.8329	3.0892
4	5.8513	2.9857
5	324.4267	3.0646
6	21.7094	2.9485
7	273.0982	2.9304
8	4.4007	3.0478
9	5.2599	2.9863
10	2.9859	2.9824
11	3.01988	3.0906
12	3.0094	2.9678
13	2.9859	2.9820
14	2.9736	3.0400
15	2.9711	3.178

Appendix C

Appendices of Chapter 3

C.1 The difference plots for well-separated clusters

In this section, Figure C.1, C.2 and C.2 illustrate the differences of sorted EL weights for the three well-separated clusters discussed in the simulation section of Chapter 3, respectively. The red lines in these three figures indicate the threshold D. From these figures, we find that selecting threshold D is relatively easy for well-separated clusters.

Figure C.1 is for the three bi-variate Gaussian cluster case. The 2 large different values at around the 250^{th} and 600^{th} index are obvious. Thus, we can the threshold can be easily selected such that these two difference values are bigger than the threshold. Thus, it gives the dataset containing 3 clusters.

Next, the differences of sorted EL weights for the 3D atom dataset is shown in Figure C.2. There is a huge difference value located at the 400^{th} index. It suggests that the dataset contains 2 clusters and the sample size of the two clusters is the same.

Then, Figure C.3 presents the differences of sorted EL weights for the mixture Von Mises-Fisher case. The selection of the threshold D is also easy since there are 3 large difference values. It implies the dataset has 4 clusters. However, the first huge difference value is positioned next to the second one. Then, the first and second clusters are too small to be the individual clusters. Thus, the elements in these two clusters might be treated as outliers or observations from one small cluster.

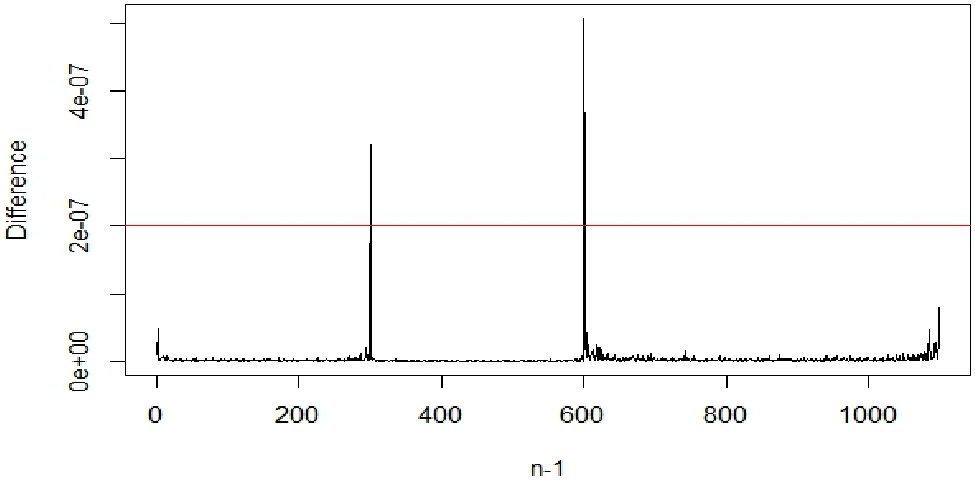


Figure C.1: The difference plot for the mixture Gaussian case

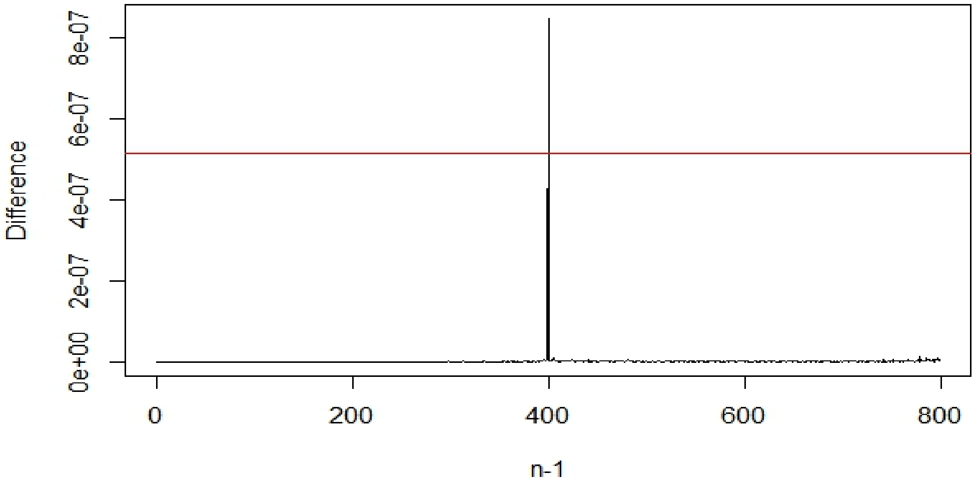


Figure C.2: The difference plot for the 3D atom dataset

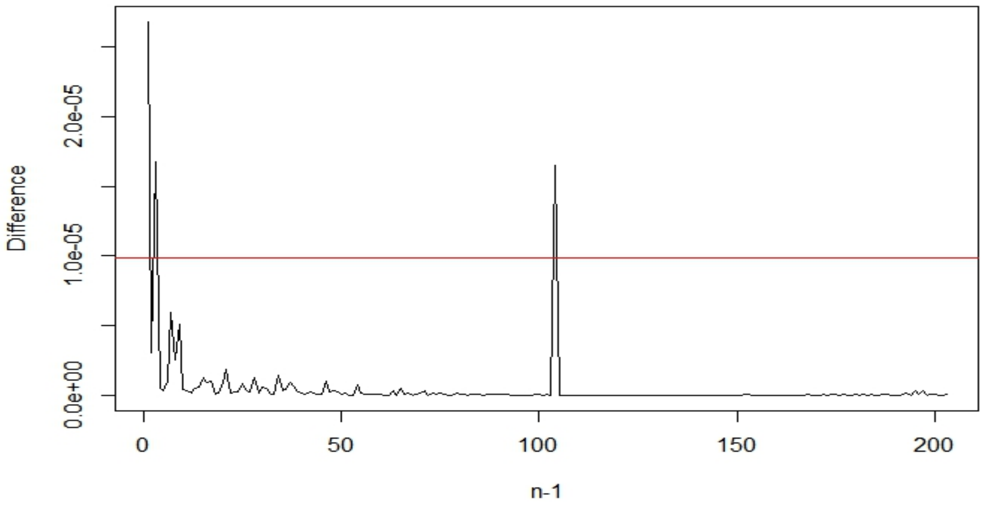


Figure C.3: The difference plot for the mixture Von Mises-Fisher case

C.2 Results for mixture Gaussian with different parameters

This appendix presents the result for the three bi-variate Gaussian cluster cases with mean constraint by setting mean be the 3^{rd}-quantile of the dataset. In Figure C.4, there are 3 huge difference values in the middle plot. Then, the dataset first be split into 4 clusters. However, the sample size of the last cluster is small, so it is merged into another cluster. Although the EL weights in Figure C.4 differ from those obtained using bootstrap mean, the clustering results do not change.

C.3 The hierarchical clustering dendrogram

In this section, we show the dendrogram of hierarchical clustering with complete, average and single linkage for the case that generated from two Von Mises-Fisher distributions with four additional observations, respectively. The horizontal axis of the dendrogram shows the clusters, while the vertical axis represents distance between the clusters. From

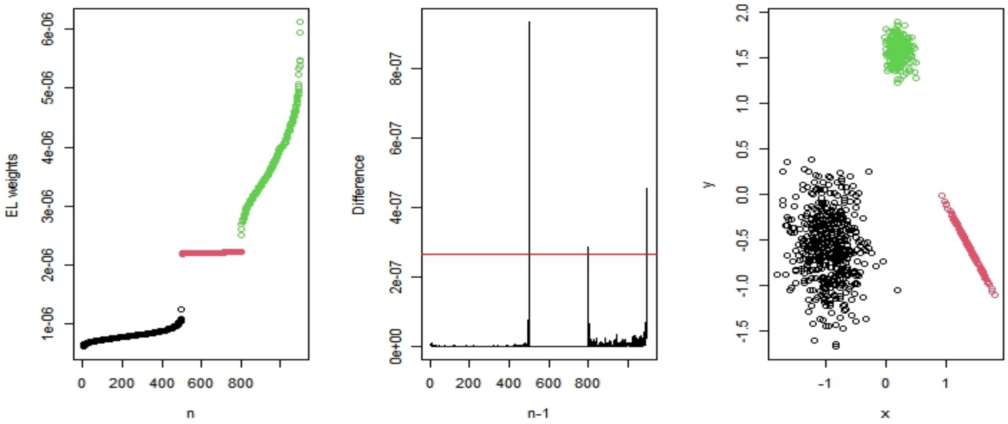


Figure C.4: The plots for well-separated Mixture Gaussian cases

Figure C.5, C.6 and C.7, we see that it is relatively subjective to select the threshold to customize the number of clusters in hierarchical clustering dendrogram.

In Figure C.5, it is reasonable to choose a threshold that splits the dataset into 2 or 4 clusters. In Figure C.6, the selection of threshold is more difficult since the tree structure is tighter. The reasonable threshold could be the value that splits the dataset into 2 to 5 clusters. Then, Figure C.7 seems to suggest that the datasets should be grouped into 2 clusters.

C.4 Additional results for mixture Von Mises-Fisher case using DBSCAN

Since DBSCAN is sensitive to the choice of two parameters: Eps and $MinPts$, this section provides additional results for DBSCAN with a broader range of parameters to improve reliability for claiming that DBSCAN does not perform well in the mixture Von Mises-Fisher case. Now, the value of Eps takes from the set $\{0.01, 0.05, 0.1, 0.2, 0.3, ..., 0.8, 1, 1.5\}$ and $MinPts$ takes values from the set $\{2, 5, 10\}$.

The Eps indicates the distance that specifies the neighbourhoods, so that the observations within the distance Eps to each other are viewed as neighbours. Generally, a smaller value of Eps gives a higher number of clusters and each cluster is tighter, while a larger

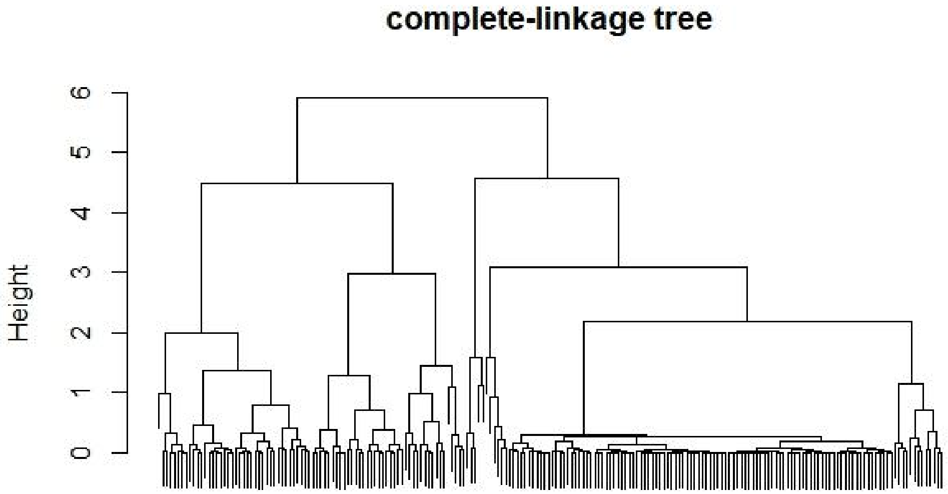


Figure C.5: The dendrogram with complete linkage

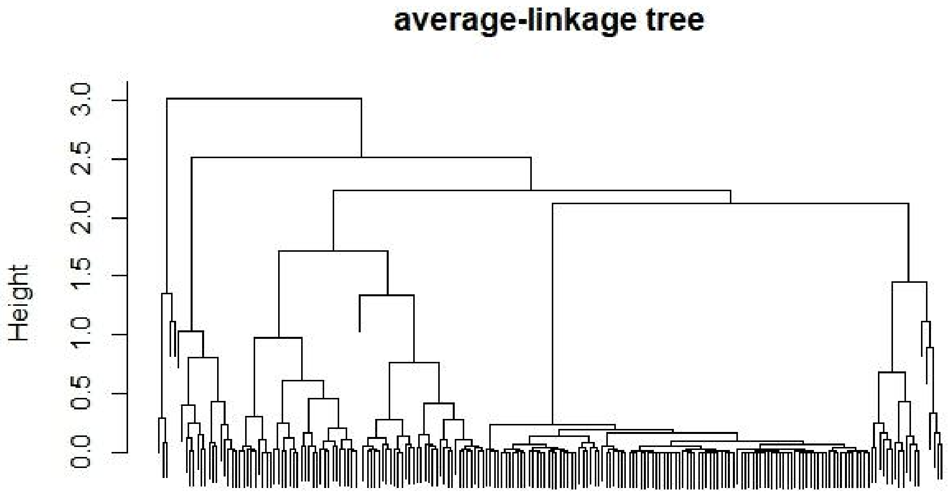


Figure C.6: The dendrogram with average linkage

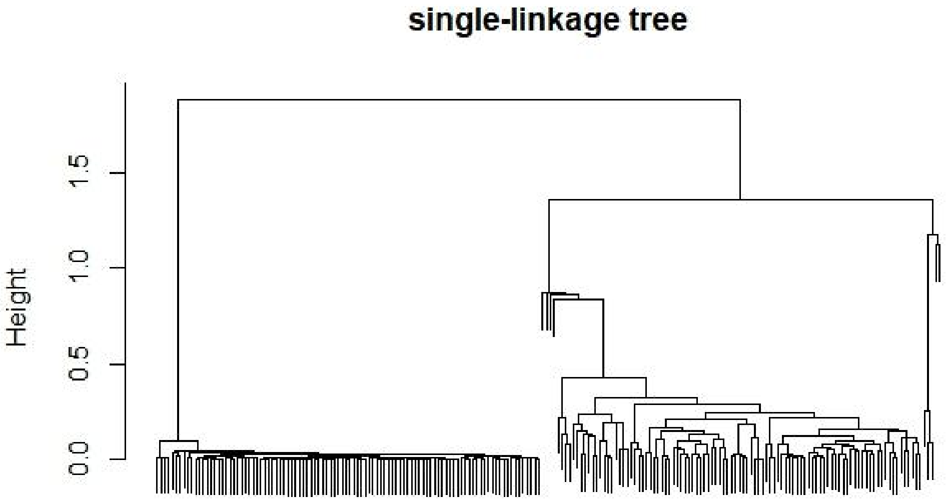
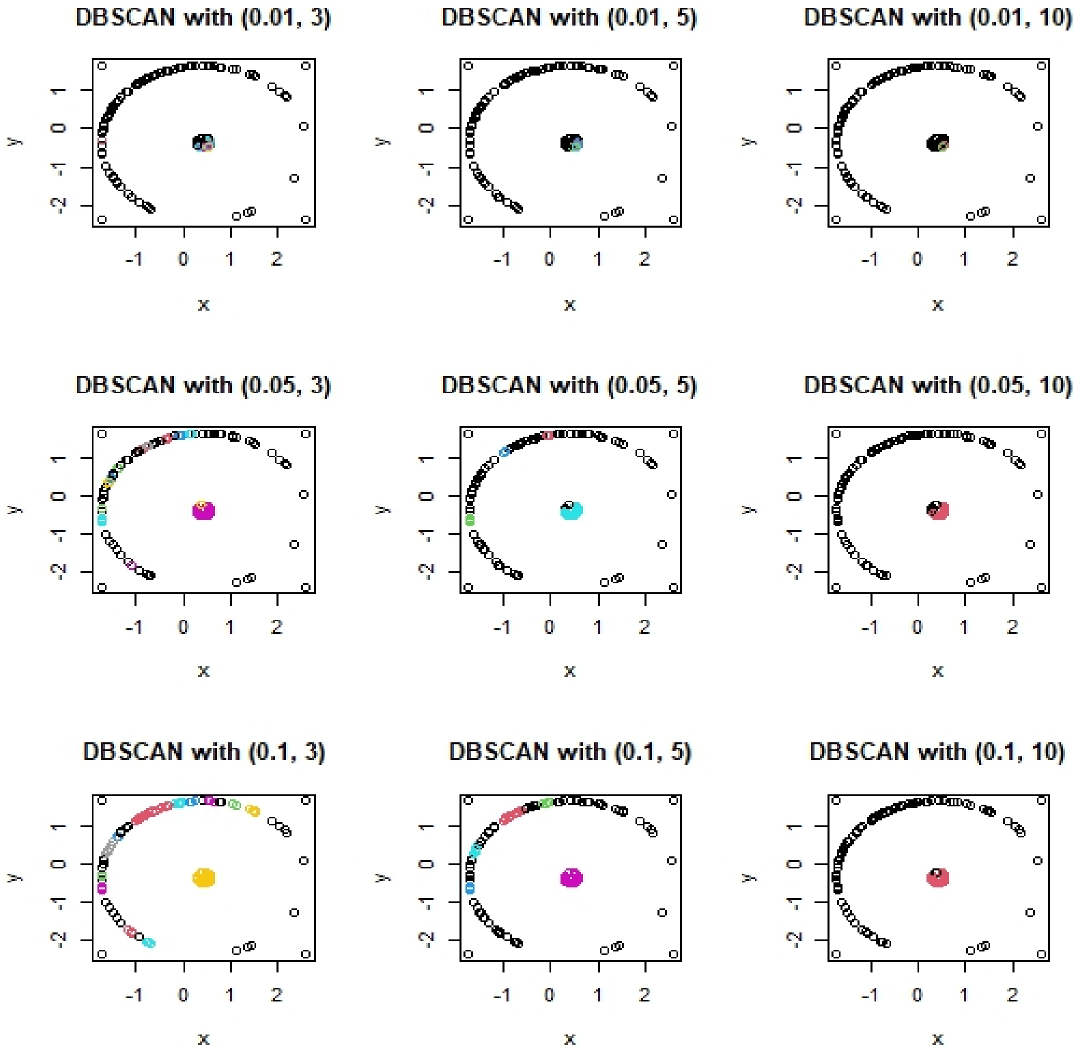


Figure C.7: The dendrogram with single linkage

value of *Eps* potentially leads to a lower number of clusters. The results in Figure C.9 to Figure C.11 align with this pattern pattern.

However, if the value of *Eps* is too small, we might have a smaller number of clusters. This is because most observations do not have any neighbours when the value of *Eps* is too small. Then, the setting of $MinPts$ makes the isolated observations merged. Thus, the number of clusters in the top panel of Figure C.8, which with *Eps* = 0.01, is less than the number of clusters in the other figures. Further, if the value of *Eps* is too large, some details might be lost, as shown in the bottom panel of Figure C.11. Additionally, the value of $MinPts$ should be bigger than the number of features and smaller than half of the sample size.

From the figures in this section, we find that DBSCAN with these parameters all fail to correctly group the clusters. The relatively better one from the additional results is produced by DBSCAN with $(1.5, 2)$ or $(1.5, 5)$, as shown in the bottom panel of Figure C.11. They detect there are 2 clusters, while the 4 additional observations are detected as elements from the outer cluster.



DBSCAN with (0.01, 3)
DBSCAN with (0.01, 5)
DBSCAN with (0.01, 10)
DBSCAN with (0.05, 3)
DBSCAN with (0.05, 5)
DBSCAN with (0.05, 10)
DBSCAN with (0.1, 3)
DBSCAN with (0.1, 5)
DBSCAN with (0.1, 10)

Figure C.8: Additional results for mixture Von Mises-Fisher case using DBSCAN (1)

C.5 K-based methods for the iris dataset

This appendix presents the results using K-means and K-medoids with different K. From Figure C.12, we can see that the results using K-means and K-medoids are highly dependent on the selection of the number of clusters. When we have the correct number of clusters, they can detect the Setosa groups. However, they give 6 misclassified observations for the

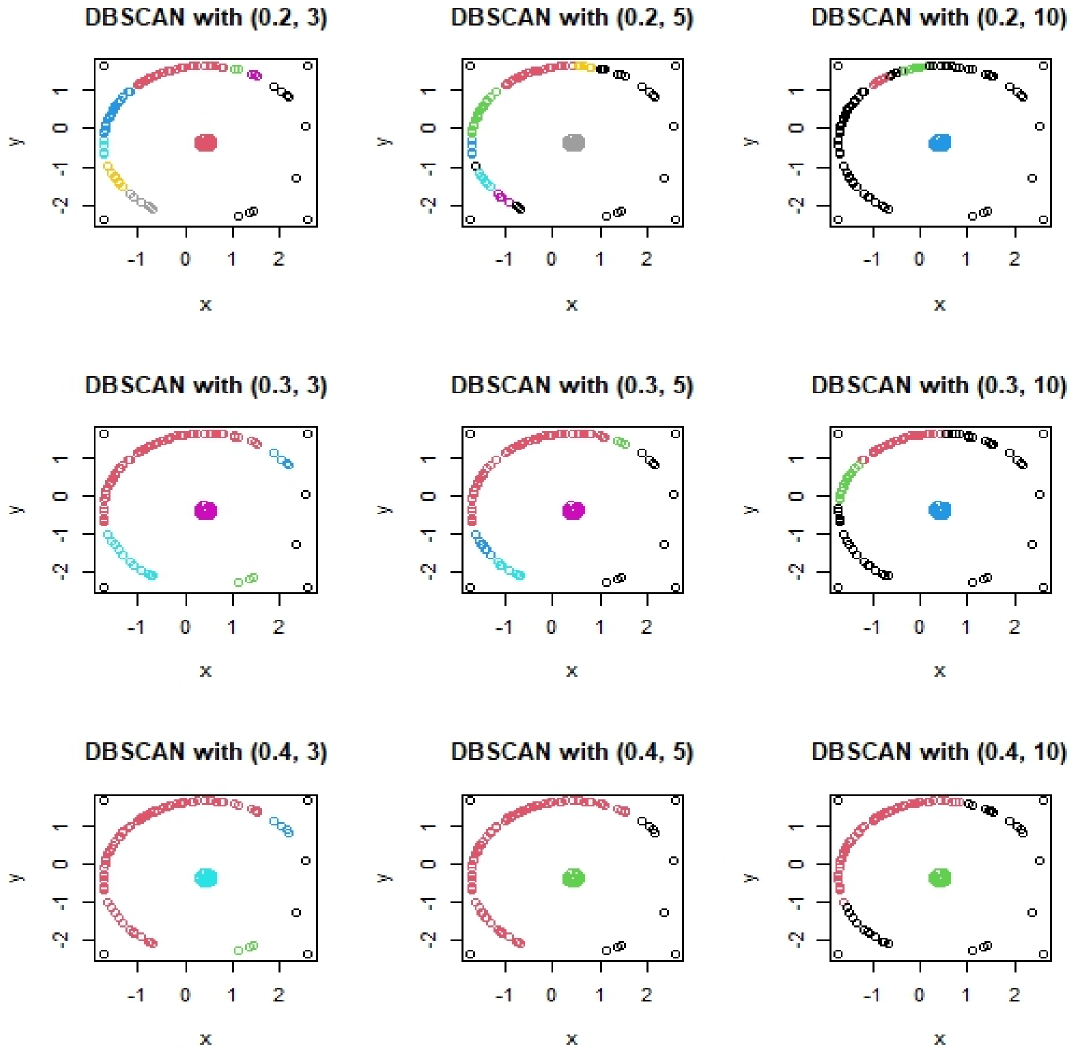
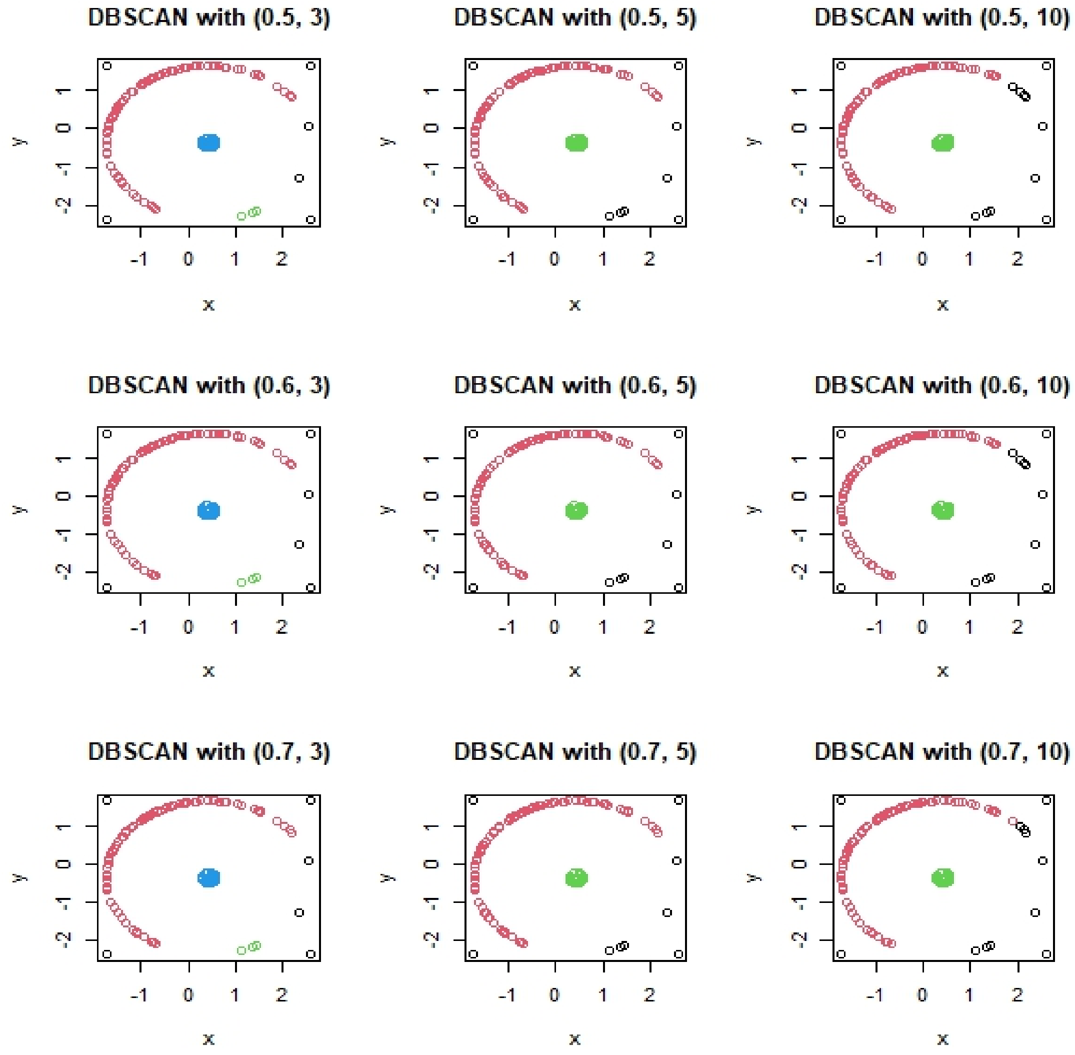


Figure C.9: Additional results for mixture Von Mises-Fisher case using DBSCAN (2)

Virginia and Versicolor groups.



DBSCAN with (0.5, 3)

DBSCAN with (0.5, 5)

DBSCAN with (0.5, 10)

DBSCAN with (0.6, 3)

DBSCAN with (0.6, 5)

DBSCAN with (0.6, 10)

DBSCAN with (0.7, 3)

DBSCAN with (0.7, 5)

DBSCAN with (0.7, 10)

Figure C.10: Additional results for mixture Von Mises-Fisher case using DBSCAN (3)

C.6 Hierarchical clustering for the iris dataset

Although hierarchical clustering methods do not require the number of clusters, they need to set a threshold to customize the number of clusters. This section presents the results using hierarchical clustering methods with different thresholds.

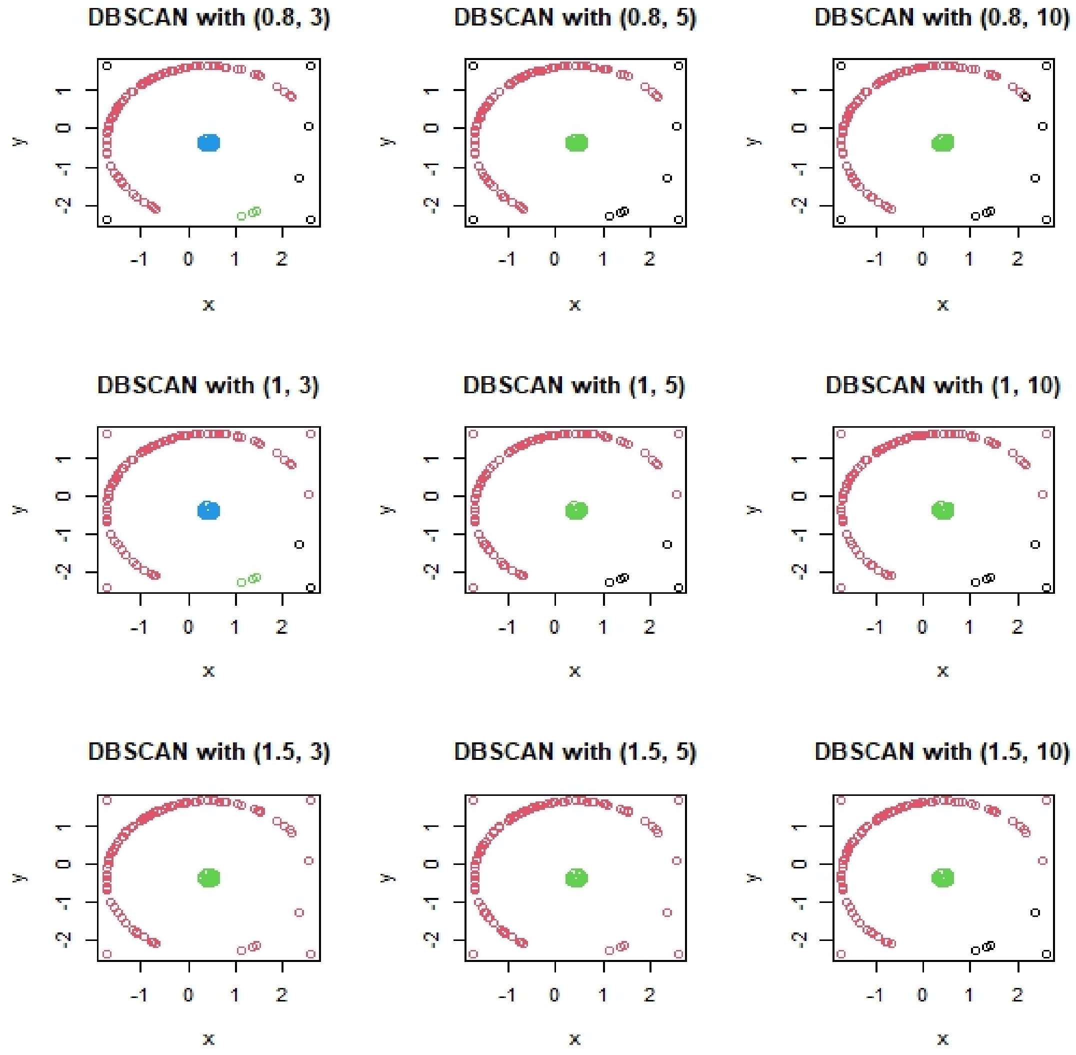


Figure C.11: Additional results for mixture Von Mises-Fisher case using DBSCAN (4)

The selection of the threshold is based on the dendrogram. For hierarchical clustering with complete linkage, any threshold that makes the number of clusters be 2 and 4 is highly possible based on the structure of the dendrogram, as shown in Figure C.13. For hierarchical clustering with average linkage, a threshold which causes the number of clusters to equal 2 or 3 is more likely to be chosen, as shown in Figure C.14. Then, Figure C.15 provides the dendrogram of hierarchical clustering with single linkage, it seems highly

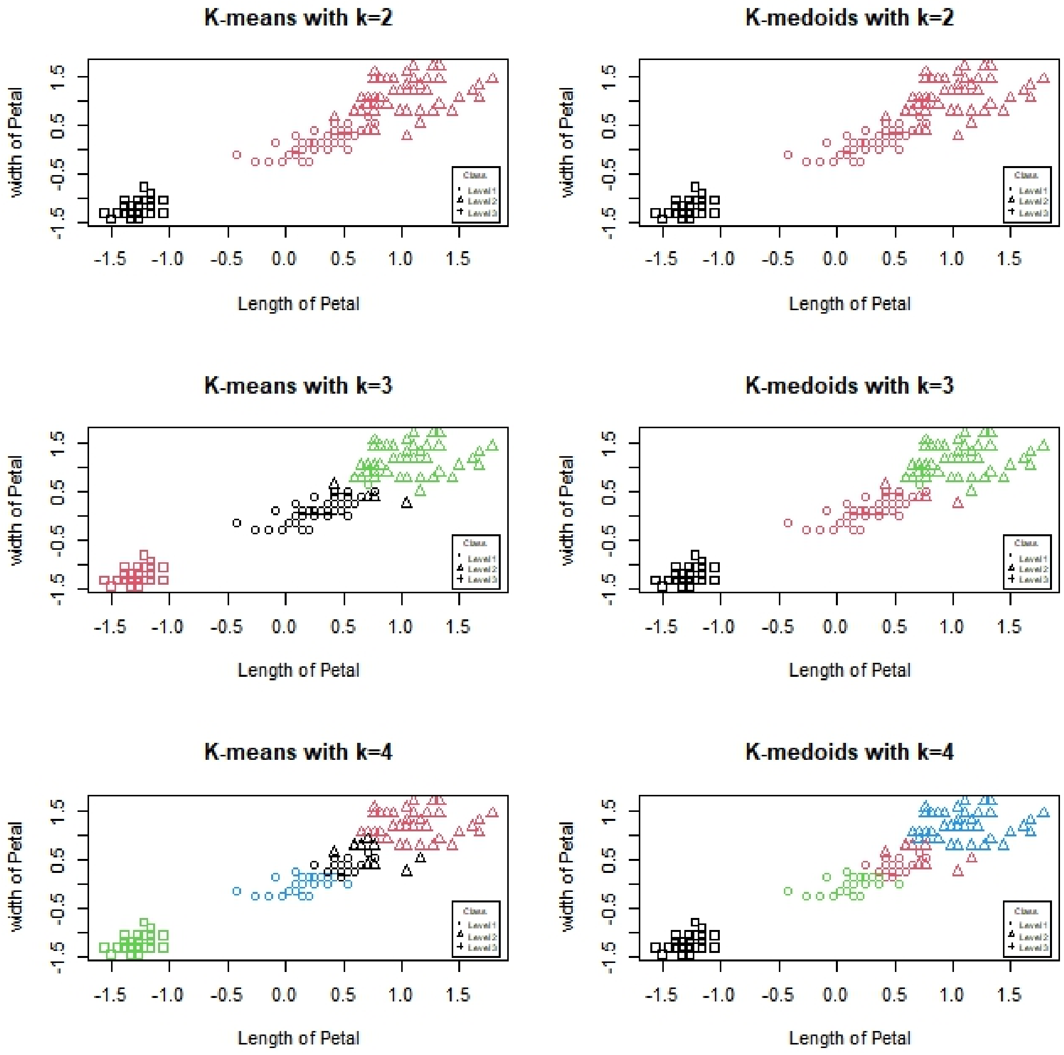


Figure C.12: Results for the iris dataset using K-based methods

suggest the number of clusters should be 2 is most suitable. However, these selections are subjective.

Here, we present the results using hierarchical clustering with complete, average and single linkages when the number of clusters is $2, 3, 4, 5$, respectively. From Figure C.16, we can see that hierarchical clustering with complete and single linkages is not suitable for this dataset.

From Figure C.16, we can see that hierarchical clustering with single linkages is not suitable for this dataset. It can not correctly detect the Setosa group when the number of clusters is 2. Also, when the threshold is selected to make the number of clusters be 3, it falls to split the Virginia and Versicolor groups. Then, hierarchical clustering with complete linkages is better than with single linkages. But, it gives around 20 misclassified observations from the Virginia and Versicolor groups. Hierarchical clustering with average linkage performs best, which gives 3 misclassified observations from the Virginia and Versicolor groups if we correctly select the threshold. However, it is most possible to select a threshold such that the number of clusters is 2 which mixes Versicolor and Setosa groups.

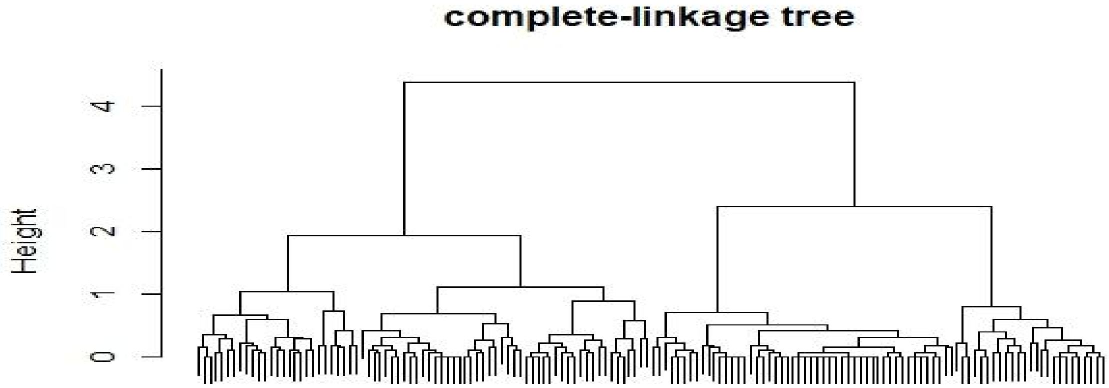


Figure C.13: The dendrogram for the iris dataset with complete linkage

C.7 DBSCAN for the iris dataset

Again, this appendix provides the results using DBSCAN with different parameters to improve reliability to calim DBSCAN method does not perform well in this case, since the results using DBSCAN is very sensitive to the value of parameters. Now, the value of *Eps* takes from the set $\{0.02, 0.05, 0.1, 0.2, 0, 4, 0.6, 0.8, 1, 1.2, 1.5, 2\}$ and $MinPts$ takes values from the set $\{5, 20\}$. Unlike K-based methods or hierarchical clustering, DBSCAN automatically gives the number of clusters, but the performance of results in Figure C.17 and C.18 are all poor.

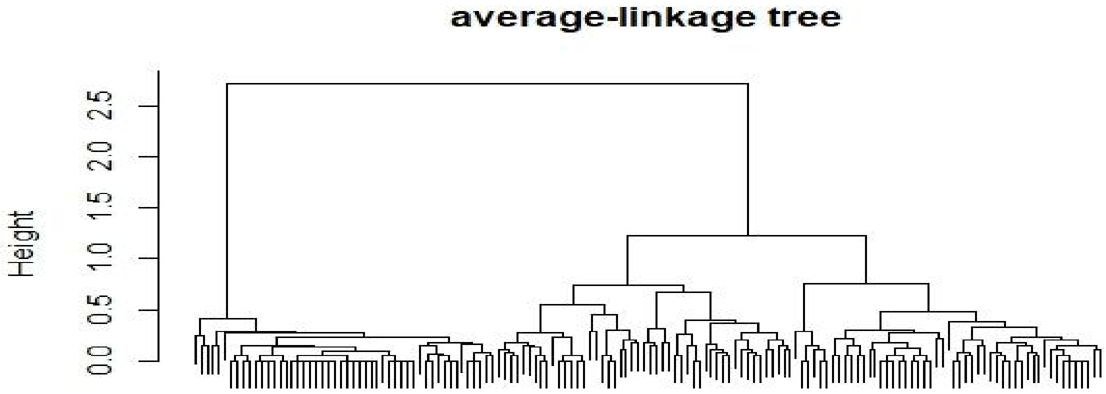


Figure C.14: The dendrogram for the iris dataset with average linkage

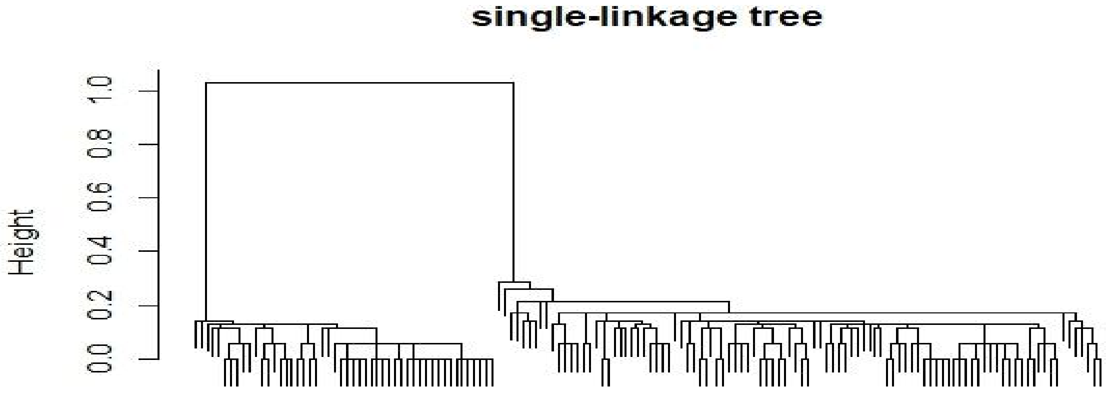


Figure C.15: The dendrogram for the iris dataset with single linkage

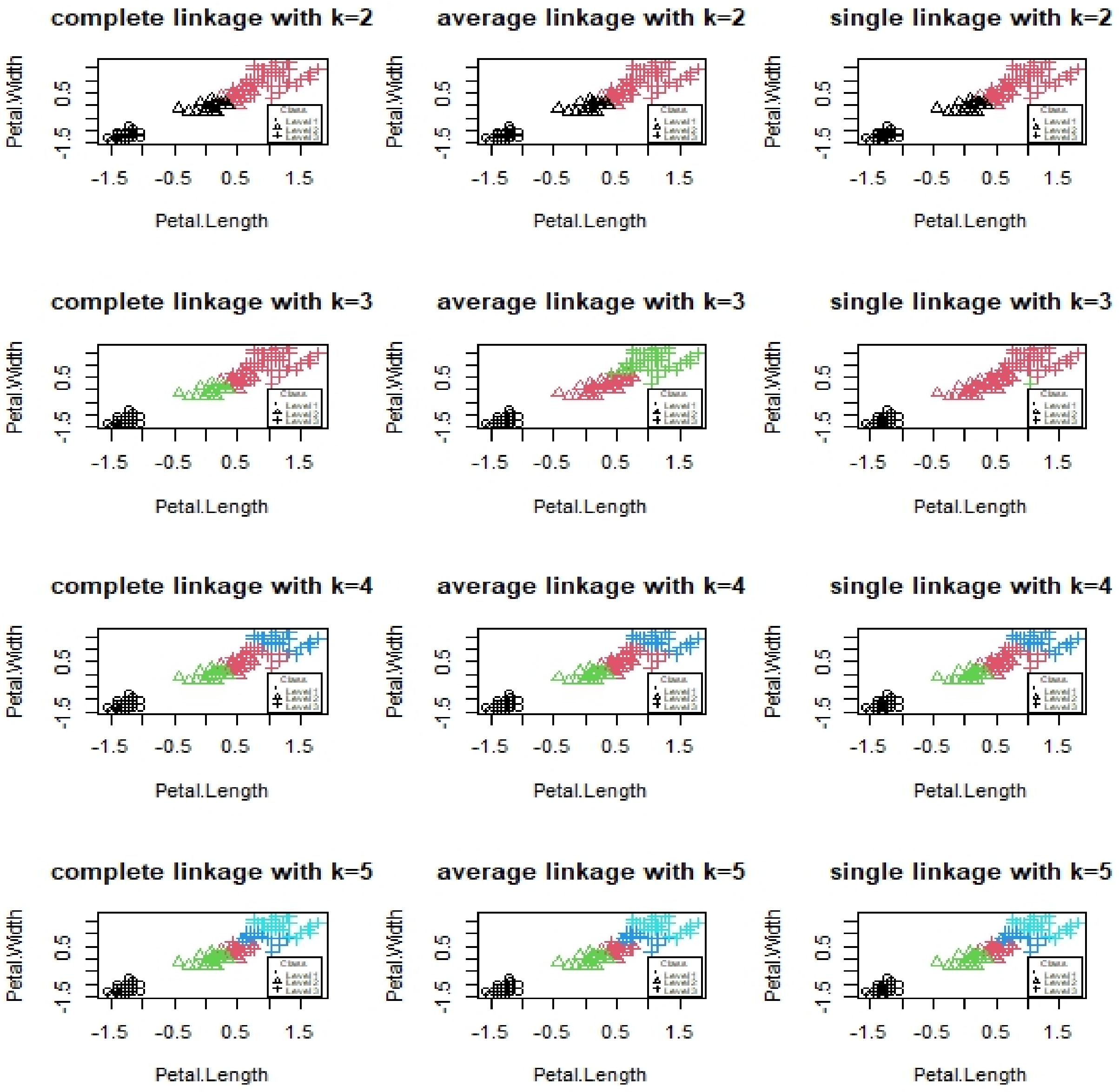


Figure C.16: Results for the iris dataset using hierarchal clustering methods

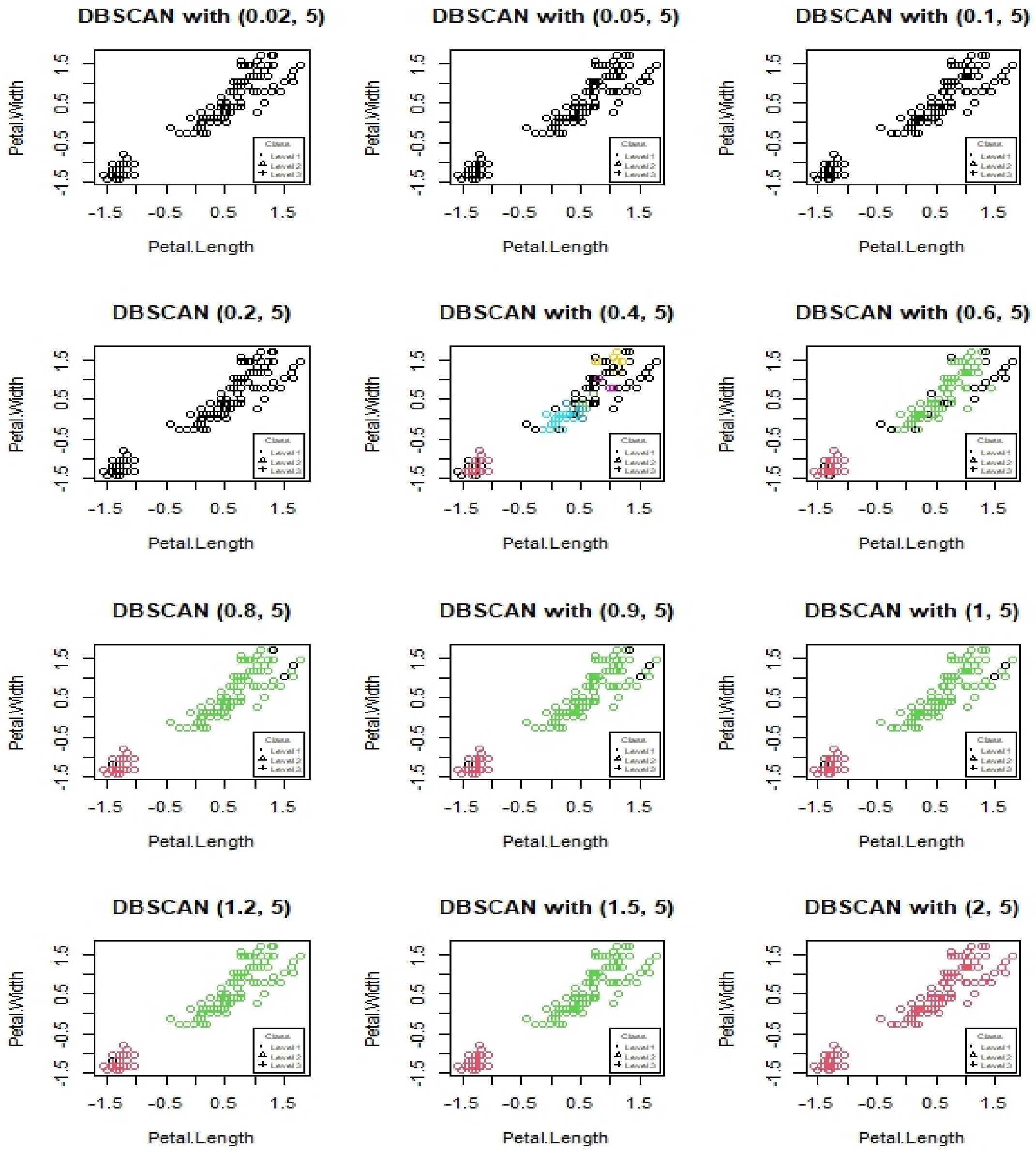


Figure C.17: Results for the iris dataset using DBSCAN with $MinPts = 5$

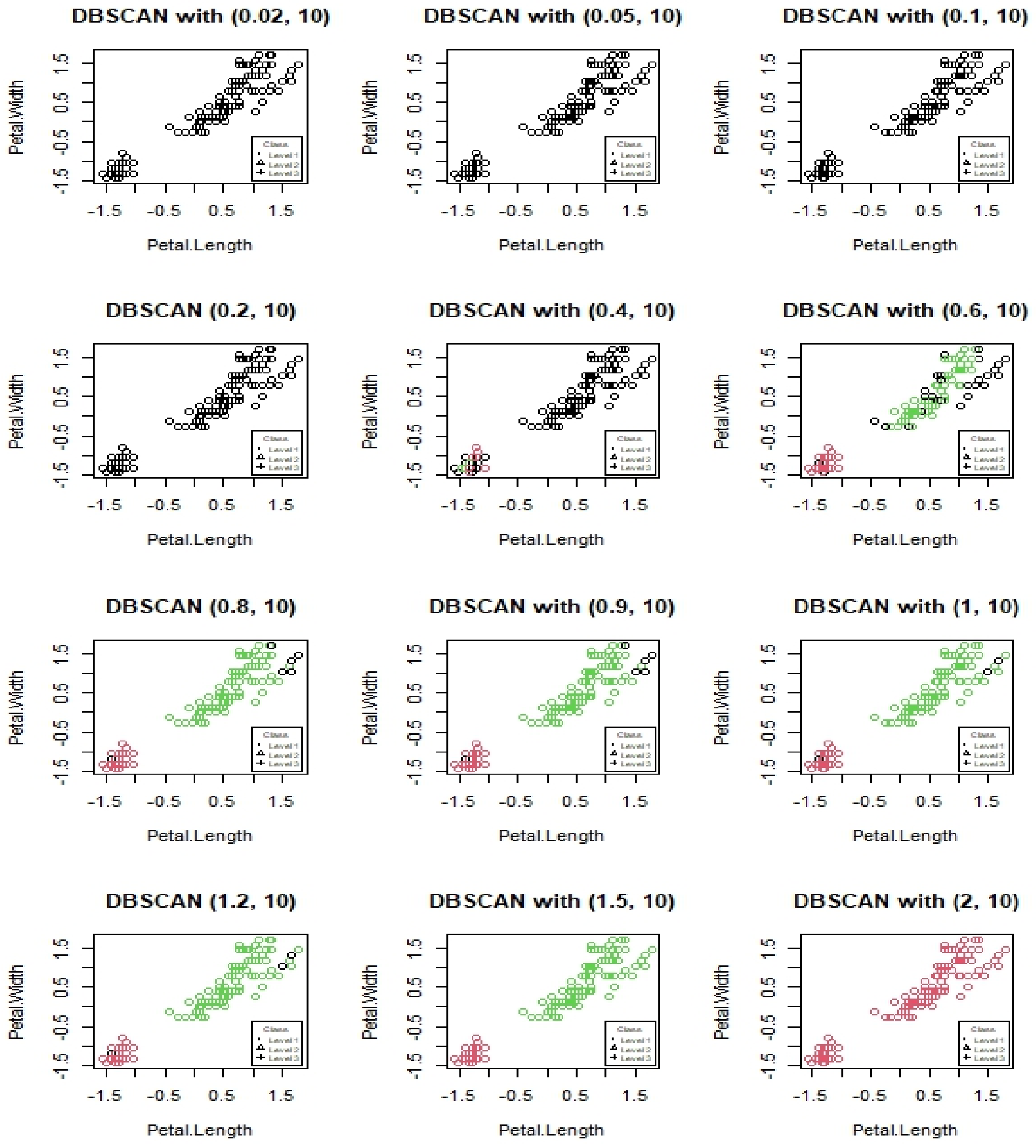


Figure C.18: Results for the iris dataset using DBSCAN with $MinPts = 10$