Indian Statistical Institute
Doctoral Thesis

On Robust Estimation of Multivariate Location and Scale with Applications

Author:
Soumya Chakraborty

Supervisors:
Prof. Ayanendranath Basu
& Dr. Abhik Ghosh

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy in
Statistics
Interdisciplinary Statistical Research Unit
Indian Statistical Institute, Kolkata
January, 2026

Acknowledgement

I would like to convey my deepest gratitude to my alma mater Indian Statistical Institute (ISI), Kolkata for providing me the opportunity to carry out my research with modern facilities including an enriched library, computing servers, stipend, contingency support and most importantly, a healthy academic environment. The colleagues of the Interdisciplinary Statistical Research Unit (ISRU) at ISI were enormously cooperative and supportive during my entire research tenure.

I consider myself extremely fortunate to have Prof. Ayanendranath Basu and Dr. Abhik Ghosh as my doctoral advisors. I feel deeply indebted to their overall academic guidance which have shaped the present thesis. During my tenure of research, my advisors have carefully guided me about the effective ways of doing literature review, framing the research problems, designing the statistical methods, applying the essential mathematics for establishing their theoretical properties and their computational improvisations. I am really grateful to my advisors for their enormous support and numerous useful suggestions which shaped the present form of this thesis. In addition, the valuable suggestions from the Research Fellow Advisory Committee (RFAC) and the Ph.D.-D.Sc. committee during various evaluation sessions have greatly helped me in my research assignments.

It is my great fortune to study my undergraduate (Bachelor of Statistics) and postgraduate (Master of Statistics) courses under the guidance of the eminent faculty of ISI. The courses offered by Prof. Parthanil Roy, Prof. Rajat Subhra Hazra, Prof. Gopal Krishna Basak, Prof. Ayanendranath Basu, Prof. (Lt.) Saurabh Ghosh, Prof. Subir Kumar Bhandari, Prof. Probal Chaudhuri, Prof. Tapas Samanta, Prof. Arijit Chakrabarti, Prof. Anil K. Ghosh, Prof. Sourabh Bhattacharya and Prof. Kiranmoy Das have made an everlasting impression on my young mind which has driven me onto the path of research and teaching.

From May, 2022, I have started working as an Assistant Professor in the Department of Statistics at Bethune College, Kolkata and have simultaneously continued my research work as an external fellow at ISI. I must convey my gratitude to the entire fraternity of the faculty members of Bethune College for their constant support and cooperation.

I must admit the enormous support, cooperation and valuable advice received from my friends at ISI and other Universities. Especially, my classmates from ISI during the bachelors and masters courses and the research scholars from ISI do need to be

mentioned in this regard.

Finally, I would like to convey my gratitude to my parents without whose affection and mental support it would not be possible for me to live a balanced life by managing all the things simultaneously. In the conclusion, I humbly express my heartfelt gratitude to the Almighty, whose presence has been the guiding force behind my academic endeavours.

Soumya Chakraborty

Contents

List of Figures

List of Tables

List of Acronyms

ARE	Asymptotic Relative Efficiency
BP	Breakdown Point
CDF	Cumulative Distribution Function
CLT	Central Limit Theorem
DPD	Density Power Divergence
FS	Fisher's Scoring
GD	Gradient Descent
GK	Gnanadesikan-Kettenring
HD	Hellinger Distance
IF	Influence Function
IRLS	Iteratively Reweighted Least Squares
KLD	Kullback-Leibler Divergence
LD	Likelihood Disparity
MCD	Minimum Covariance Determinant
MDPDE	Minimum Density Power Divergence Estimator
MLE	Maximum Likelihood Estimator
MPLE	Maximum Pseudo β-Likelihood Estimator
MM	Modified M (Estimator)
MVE	Minimum Volume Ellipsoid
NR	Newton-Raphson
NCS	Neyman's Chi-square
PCS	Pearson's Chi-square
PDF	Probability Density Function
SMDPDE	Sequential Minimum Density Power Divergence Estimator
SPD	Symmetric and Positive Definite

Chapter 1

Introduction

1.1 Preamble

The objective of all scientific endeavor is to get a proper understanding of different natural phenomena taking place around us. Technology helps us build tools, based on such scientific information, to provide us the conveniences of science. Even though there is great variation in the scope and the approach of the different branches of science, there is one aspect that is common to all of them; they all generate loads of data (of different types). This is true not only for disciplines which are explicitly mathematical (such as physics and engineering), but also for such disciplines which have large observational or clinical components (such as biology or medical science). In one sense statistics is the single discipline which binds together all these different branches by providing quantitative methods that analyze the available data and can lead to the proposal of and/or validation of suitable models for the unknown data generating process operating in the background, as well as for testing the veracity of postulated hypotheses that naturally arise as part of the process. Since practically all data generating processes are probabilistic and follow appropriate distributional rules, statistical analysis is now an integral part of any scientific decision making process which is not wholly deterministic. The majority of statistical models used in practice are parametric in nature. Such models help us to describe the physical data generation process in terms of a finite number of interpretable (but unknown) parameters. In such cases, the first step of statistical inference is to optimally estimate the values of the unknown parameters. These parameter estimates are then utilized for testing different hypotheses as well as for making future predictions. Thus, the estimation problem is of primary importance in the context of statistical inference which directly relies on both the available data and the fitted statistical models. One of the pioneering ideas in the field of classical statistical inference is the likelihood principle (Fisher (1912) [48], (1922) [49], (1925) [50]) which has been the basis of a large part of statistical study and analysis done throughout the last century. The estimator obtained by maximizing the likelihood function is known as the maximum likelihood estimator (MLE) which is utilized in

statistical applications in various fields. Maximum likelihood estimators (see Section 1.5) are known to have superior statistical properties, such as consistency, asymptotic normality, minimax rate of convergence and most importantly, first order efficiency (i.e., lowest asymptotic variance among all the estimators) under standard regularity assumptions. However, these estimators as well as some other related classical estimators are severely affected by contamination or presence of anomalous observations in the sample in many real life datasets (which is not entirely unexpected). In fact, consistency and minimum asymptotic variance do not guarantee such resistance of the estimators against anomalous observations (referred to as "outliers") when model conditions fail. In a broad sense, the property of statistical procedures which ensures the stability of the same towards data contamination and model misspecification is known as "robustness".

Robustness of statistical procedures have been formalized and studied at least since the 1960s and 1970s when the fundamental works of Box (1953) [17], Box and Andersen (1955) [18], Tukey (1960 [146], 1962 [147]), Huber (1964 [77], (1965) [78]), Hampel (1971 [66], 1974 [67]) and some others were published and popularized. These works have mainly described the fundamental theoretical issues associated with the philosophy of robustness, such as quantification of the same via influence analysis and breakdown behaviour, the trade-off between robustness and asymptotic efficiency and the study and development of M-estimators (see Section 1.6) which were regarded, at least initially, as the primary robust alternatives to the maximum likelihood estimator. Later, the methodological development and subsequent real life applications of robust methods have been done massively through the works of Maronna (1976) [109], Stahel (1981) [141], Donoho (1982) [38], Ronchetti (1982 [125], 1982 [126], 1985 [127]), Tyler (1983, 1987, 2014) [148, 149, 150], Kent and Tyler (1991, 1996) [86, 87], Rousseeuw (1985) [129], Rousseeuw and Leroy (1987) [131], Rousseeuw and Driessen (1999) [130] and many others, particularly in the fields of multivariate statistics and regression analysis. Sometimes, however, these methods have been found to be computationally challenging especially with growing data dimension and increasing complexity of statistical models. Thus, computationally efficient robust methods are essential for applications in different domains including the field of multivariate and high dimensional set-ups. The main goal of this thesis is to develop theoretically efficient and computationally efficient methods for robust estimation of parameters, including multivariate location and scale parameters, and their subsequent application in some machine learn-

ing problems (parametric clustering, classification and anomaly detection) through the minimization of a particular statistical distance (see Section 1.7). The background and illustrations of useful prerequisites are discussed in the next few sections.

1.2 General Notation

(i) Unless otherwise mentioned, the notations "log" and "exp" are used to denote the natural logarithm and the usual exponential function, respectively.

(ii) Bold symbols (e.g., $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$) are used to denote vectors (column vectors unless otherwise specified) and matrices and for a matrix $\boldsymbol{A}$, $\boldsymbol{A}^{'}$ is its usual transpose matrix.

(iii) $I(A)$ is used to denote the indicator function of the set A.

(iv) In general, Θ is used to denote the parameter space where θ is the unknown parameter of interest.

(v) Upper case letters like F, F_θ, G, G_θ are used to denote cumulative distribution functions (CDF) while the corresponding lower case letters f, f_θ, g, g_θ are used to denote probability density functions (PDF). The PDFs are considered with respect to the standard Lebesgue measure or the counting measure in general. In multivariate cases, for simplicity, we use the same notation to represent probability measures and their corresponding CDFs. We will also specify it explicitly, wherever it will be required.

(vi) In particular, g will denote the true unknown probability density function and f_θ will denote the model density function.

(vii) For a cumulative distribution function G, the corresponding empirical cumulative distribution function based on an identically and independently (i.i.d.) distributed random sample $\{X_1, \ldots, X_n\}$ from G is denoted by

$$\begin{array}{l} G_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x), \text{ in case of univariate samples,} \\ G_n(A) = \frac{1}{n}\sum_{i=1}^{n} I_A(X_i), \text{ in case of multivariate samples.} \end{array} \tag{1.1}$$

Additionally, all the random samples considered in this thesis are assumed to be i.i.d. samples (unless otherwise mentioned).

(viii) Let $\boldsymbol{\theta}$ be an m-dimensional parameter vector. The gradient vector and the Hessian matrix of the model density $f_{\boldsymbol{\theta}}$ with respect to $\boldsymbol{\theta}$ are denoted by $\nabla f_{\boldsymbol{\theta}}$ and $\nabla^2 f_{\boldsymbol{\theta}}$, respectively. The first order partial derivative of the function $f_{\boldsymbol{\theta}}$ with respect to θ_j, the j-th component of $\boldsymbol{\theta}$ is denoted by $\frac{\partial f_{\boldsymbol{\theta}}}{\partial \theta_j}$ which is the j-th component of the gradient vector $\nabla f_{\boldsymbol{\theta}}$; the second order partial derivative of $f_{\boldsymbol{\theta}}$ with respect to the i-th and j-th components of $\boldsymbol{\theta}$ is denoted by $\frac{\partial^2 f_{\boldsymbol{\theta}}}{\partial \theta_i \partial \theta_j}$.

(ix) The notations $\overset{a.s.}{\to}$, $\overset{p}{\to}$ and $\overset{d}{\to}$ are used to denote convergence in the almost sure, in probability and in distribution (weak mode) senses, respectively.

(x) In multivariate cases, we characterize the probability distribution with its probability measure instead of its CDF.

1.3 Asymptotic Efficiency

For an estimator $\hat{\theta}_n$ of the unknown parameter θ which is $\sqrt{n}$ consistent, "efficiency" will be measured by the inverse of the variance of $\sqrt{n}\hat{\theta}_n$. That is, the precision of the estimator increases with its efficiency and decreases with its variance. Exact small sample probability distributions of these estimators may not be possible to determine all the time. Asymptotic distributions of the properly standardized versions of these estimators are utilized in this situation for drawing inferences. In particular, various versions of the central limit theorem (CLT) provide the basis of asymptotic normality of many of the well-known estimators. The concept of first order efficiency deals with the comparison of the asymptotic variance of a particular estimator with the minimum possible value of the asymptotic variance among all the estimators. To understand the concept more precisely, let us first delineate the notion of Fisher information.

Let us consider the random variable X with true unknown PDF g that is modelled by the parametric family of densities $\mathcal{F}_\Theta = \{f_\theta : \theta \in \Theta\}$ which exists with respect to a σ-finite measure ν on $\mathbb{R}$. Let us also assume appropriate regularity conditions so that for any measurable $B \subset \mathbb{R}$, the following differentiation can be taken under the integral sign (Loève (1977) [96]):

$$\frac{\partial}{\partial \theta} \int_B f_\theta(x) d\nu(x) = \int_B \frac{\partial}{\partial \theta} f_\theta(x) d\nu(x). \tag{1.2}$$

Regular exponential families of probability distributions are known to satisfy (1.2).

The score function $u_\theta(x)$ of the model is defined as

$$u_\theta(x) = \frac{\partial}{\partial\theta}\log(f_\theta(x)) = \frac{\frac{\partial}{\partial\theta}f_\theta(x)}{f_\theta(x)}.$$

Using (1.2), it can be shown that,

$$E_{f_\theta}(u_\theta(X)) = 0,$$

where the expectation is taken with respect to the density f_θ. The Fisher information function $I(\theta)$ is defined as the variance of the score function, that is,

$$I(\theta) = \mathrm{var}_{f_\theta}(u_\theta(X)) = E_{f_\theta}(u_\theta^2(X)). \tag{1.3}$$

Remark 1.1. *We have assumed θ to be a scalar parameter in the aforesaid discussion. However, if $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^{'}$ is a m-dimensional parameter vector in $\boldsymbol{\Theta} \subset \mathbb{R}^m$, then*

$$\boldsymbol{u_\theta}(x) = (u_{\boldsymbol{\theta},1}(X), \ldots, u_{\boldsymbol{\theta},m}(X))^{'} = \left(\frac{\partial}{\partial\theta_1}\log f_{\boldsymbol{\theta}}(x), \ldots \frac{\partial}{\partial\theta_m}\log f_{\boldsymbol{\theta}}(x)\right)^{'},$$

and consequently, the Fisher information matrix is defined as

$$\boldsymbol{I}(\boldsymbol{\theta}) = ((I_{j,k}(\boldsymbol{\theta})))_{1\le j,k\le m}, \text{ where, } I_{j,k}(\boldsymbol{\theta}) = E_{f_{\boldsymbol{\theta}}}(u_{\boldsymbol{\theta},j}(X)u_{\boldsymbol{\theta},k}(X)).$$

Now, we are in a position to introduce the concept of first order efficiency. Let θ be a scalar parameter and suppose that we restrict our attention within the class of consistent and asymptotic normal (CAN) estimators T_n of the parametric function $\gamma(\theta)$ which satisfy

$$\sqrt{n}(T_n - \gamma(\theta)) \xrightarrow{d} N(0, v(\theta)),$$

where $v(\theta)$ is the asymptotic variance of $\sqrt{n}T_n$. It was believed earlier that the lower bound for $v(\theta)$ was provided by the Cramér-Rao lower bound (CRLB)

$$\mathrm{CRLB}(\gamma(\theta)) = \frac{\{\gamma^{'}(\theta)\}^2}{I(\theta)}.$$

However, there exists a class of estimators, known as "superefficient estimators" (Le

Cam (1953) [91], Cox and Hinkley (1974) [28]), which are asymptotically normal with asymptotic variances never exceeding and sometimes strictly smaller than the CRLB. The latter work, however, suggests that the idea of superefficiency is not statistically important and thus we do not consider such estimators in this thesis. We, therefore, restrict our attention only to the class of consistent and uniformly asymptotic normal (CUAN) estimators which are also CAN estimators but the distributional convergence of $\sqrt{n}(T_n - \gamma(\theta))$ towards normality is uniform in compact intervals of θ. Such an estimator T_n of $\gamma(\theta) = \theta$ is called first order efficient if $\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0, I^{-1}(\theta))$. This will be our working requirement for first order efficiency. More generally, if $\boldsymbol{\theta}$ is a vector parameter, $\boldsymbol{T}_n$ will be referred to as first order efficient for $\boldsymbol{\theta}$ if

$$\sqrt{n}(\boldsymbol{T}_n - \boldsymbol{\theta}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{I}^{-1}(\boldsymbol{\theta})). \tag{1.4}$$

1.4 Robustness

Statistical models aim to fit complex datasets with probability distributions which can approximately capture the shape of the data generating distribution as represented by the observed data. Some discrepancies between these models and the actual structures of the datasets are not unexpected due to the random fluctuations that occur naturally. However, in some cases such discrepancies, even though apparently small, may lead to unsatisfactory model fits and incorrect insight generation. The term "robustness" refers to a very important property which loosely means the capacity of the statistical procedure to guard against model misspecification and outliers. To quantify robustness, certain measures have been developed in the literature. We mainly focus on two such measures, namely, the influence function and the breakdown point.

1.4.1 Influence Function

The influence function measures the effect of an infinitesimal contamination on the resulting procedure. To define it formally, let us first introduce the concept of statistical functionals. Let $\{X_1, \ldots, X_n\}$ be a random sample drawn from a probability distribution with distribution function G and let $T_n(X_1, \ldots, X_n)$ be a statistic based on the aforesaid random sample. Let

$$G_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x)$$

be the empirical cumulative distribution function. If one can formally express T_n as a functional of the empirical CDF G_n, i.e., $T_n = T(G_n)$, where $T(\cdot)$ is independent of n, then $T(\cdot)$ is called a statistical functional. It is defined on a space of distribution functions including G, all the model distribution functions $\{F_\theta\}$, all empirical CDFs G_n and this space is assumed to be closed under convex combinations. In practice, $T(G)$ will be our unknown parameter of interest and we want to study the properties of $T(G_n)$ as an estimator of $T(G)$. The notion of statistical functionals were introduced and studied by von Mises $(1936, 1937, 1947)$ [152, 153, 154].

The concept of influence functions initiates from the idea of the famous von Mises derivatives. Let, $\mathcal{D}^*$ be a space of distribution functions and for $F, G \in \mathcal{D}^*$, consider the following functional

$$T(\alpha F + (1-\alpha)G) = T(G + \alpha(F - G)).$$

The von Mises derivative $T_G^{'}$ of T at the distribution function G is defined as

$$T_G^{'}(F - G) = \frac{d}{d\alpha} T(G + \alpha(F - G))\big|_{\alpha=0}, \tag{1.5}$$

if there exists a real-valued function $\phi_G(x)$, independent of F, such that,

$$T_G^{'}(F - G) = \int \phi_G(x) d(F - G)(x). \tag{1.6}$$

The function $\phi_G(x)$ is uniquely defined only upto an additive constant as

$$\int \phi_G(x) d(F - G)(x) = \int (\phi_G(x) + c) d(F - G)(x), \forall\ c \in \mathbb{R}.$$

However, this non-uniqueness can be removed by imposing the constraint

$$\int \phi_G(x) dG(x) = 0. \tag{1.7}$$

The function $\phi_G(\cdot)$ is the influence function (IF) of the functional T at the distribution function G and is denoted by $\phi_G(y) = IF(y, T, G)$. Equation (1.7) shows that the influence function has mean zero at the true distribution. Existence of the von Mises derivative is, however, not needed for the influence function to exist, and the latter

may be explicitly defined as

$$IF(y, T, G) = \frac{d}{d\alpha} T((1-\alpha)G + \alpha \wedge_y)\Big|_{\alpha=0}, \tag{1.8}$$

under appropriate conditions, where $\wedge_y$ is the cumulative distribution function corresponding to the Dirac mass at the point y. In fact, in the subsequent sections, we will follow the definition of influence function as given in Equation (1.8). The intuition behind the concept of influence function can be explicitly understood following Equation (1.8). It essentially demonstrates the effect of adding an infinitesimal contamination at the point mass y to the existing random sample on the resulting estimator $T(G_n)$. When the von Mises derivative exists, the function $\phi_G(\cdot)$ defined in Equation (1.6) is the same as the $IF(\cdot, T, G)$ as defined in Equation (1.8).

In practice, $IF(y, T, G)$ is treated as a function of y and is plotted against y. Boundedness of this function is viewed as being desirable in terms of robustness as it indicates an infinitesimal contamination cannot exert unbounded influence on the estimator even when the contamination is at a highly unlikely value. In the following, we give examples of both bounded and unbounded influence functions.

(E1) The mean functional is defined as

$$T_{\text{Mean}}(G) = \int x dG,$$

whose influence function is given by

$$IF(y, T_{\text{mean}}, G) = y - T_{\text{Mean}}(G)$$

which is unbounded in the contamination point y. Thus, the mean is not robust from the viewpoint of influence analysis.

(E2) The median functional is defined as

$$T_{\text{Median}}(G) = G^{-1}\left(\frac{1}{2}\right),$$

whose influence function is given by

$$IF(y, T_{\text{median}}, G) = \frac{\frac{1}{2} - I(y \leq T_{\text{median}}(G))}{g(T_{\text{median}}(G))},$$

which is bounded as a function of y. Thus, the median is robust unlike the mean functional as per the influence function analysis.

The aforesaid description indicates the robustness aspects of the influence function in respect of the mean and the median. But, it also plays an important role in the asymptotic analysis of the estimators (Basu et al. (2011) [12]). To understand this role, let us consider,

$$A(\alpha) = T(G + \alpha(F - G)), \text{ for } \alpha \in [0, 1]$$

and its Taylor series expansion around $\alpha = 0$, which is

$$A(\alpha) = A(0) + \alpha A^{'}(0) + \text{Higher Order Terms.} \tag{1.9}$$

Now, replacing F with G_n, denoting the higher order terms by R_n and evaluating (1.9) at $\alpha = 1$, we have

$$\begin{aligned} T(G_n) &= T(G) + T^{'}_{G}(G_n - G) + R_n \\ &= T(G) + \int \phi_G(x) dG_n(x) + R_n \end{aligned}$$

which implies

$$\sqrt{n}(T(G_n) - T(G)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_G(X_i) + \sqrt{n} R_n, \tag{1.10}$$

assuming that the von Mises derivative exists. The expansion in (1.10) provides the linearization of the estimator and is crucial in establishing the asymptotic normality of the properly standardized estimator $\sqrt{n}(T(G_n) - T(G))$, particularly if the higher order terms (properly scaled) can be ignored, i.e.,

$$\sqrt{n} R_n = o_p(1). \tag{1.11}$$

In fact, this idea will be the key step to establish the asymptotic normality of our one-step estimators, which represent a novel class of estimators studied in Chapter 5 of this thesis. Assuming (1.11), it can be trivially concluded that

$$\sqrt{n}(T(G_n) - T(G)) \xrightarrow{d} N(0, \text{var}(\phi_G(X))), \text{ as } n \to \infty$$

following CLT and Slutsky's Theorem. Thus, the asymptotic variance of an estimator is actually the variance of its influence function under appropriate conditions.

1.4.2 Breakdown Point

The influence function measures the effect of a single contaminating observation on the estimators or statistical functionals of interest. It is a local measure of robustness. However, a larger proportion of the sample could be outlying in many real life problems. To quantify robustness in these scenarios, a different measure of robustness, formally called the breakdown point of an estimator or functional, has been proposed in the literature. Intuitively, the breakdown point (BP) of an estimator $\hat{\theta}$ (for the parameter θ) accounts for the largest proportion of sample observations that can be contaminated such that $\hat{\theta}$ still carries some "reasonable" information about θ. Suppose Θ be the parameter space. Assuming Θ to be non-empty, $\hat{\theta}$ is able to provide some information about the unknown parameter θ only if it is bounded and it lies in the interior of Θ, i.e., $\hat{\theta}$ stays away from the boundary of the parameter space Θ. As for example, suppose θ is a scale parameter with the parameter space $\Theta = (0, \infty)$. Then, $\hat{\theta}$ should preferably remain bounded and also remain bounded away from 0 in order to extract helpful information about θ. The formal definition of asymptotic breakdown point (Hampel (1968, 1971) [65, 66]) of an estimator $\hat{\theta}$ is as follows.

Definition 1.1. *The asymptotic breakdown point $\epsilon^*(\hat{\theta}, F)$ of an estimator $\hat{\theta}$ at the distribution function F is the largest $\epsilon^* \in (0, 1)$ such that for all $\epsilon < \epsilon^*$, $\hat{\theta}_\infty((1-\epsilon)F + \epsilon G)$ as a functional of G remains bounded and bounded away from the boundary of Θ.*

Here,

$$\hat{\theta}_\infty((1-\epsilon)F + \epsilon G) = \lim_{n\to\infty} \hat{\theta}_n((1-\epsilon)F + \epsilon G), \tag{1.12}$$

where $\hat{\theta}_n((1-\epsilon)F + \epsilon G)$ denotes the estimator of θ based on a random sample of size n from the contaminated distribution (with distribution function $((1-\epsilon)F + \epsilon G)$). One may analogously define the finite sample breakdown point (Donoho (1982) [38], Donoho and Huber (1983) [39]) of $\hat{\theta}_n$ based on a random sample $\boldsymbol{X}$ of size n as

$$\epsilon^*(\hat{\theta}_n, \boldsymbol{X}) = \max_{1\le m\le n} \left[\frac{m}{n} : \sup_{\boldsymbol{Y}_m} ||\hat{\theta}_n(\boldsymbol{X}) - \hat{\theta}_n(\boldsymbol{Y}_m)|| < \infty \right], \tag{1.13}$$

where the sample $\boldsymbol{Y}_m$ is obtained from the original sample $\boldsymbol{X}$ by replacing its m observations with arbitrary values. Here $\hat{\theta}_n(\boldsymbol{X})$ and $\hat{\theta}_n(\boldsymbol{Y}_m)$ are the values of the

estimators calculated from the samples $\boldsymbol{X}$ and $\boldsymbol{Y}_m$, respectively.

1.5 Maximum Likelihood Estimation

The maximum likelihood principle is perhaps the most prolific cornerstone in the domain of classical inference. Although this methodology was evidently used by the great mathematician Carl F. Gauss in the nineteenth century, the statistical theory and methodology of the maximum likelihood procedure was systematically developed and studied by the famous statistician Sir Ronald A. Fisher in the twentieth century. Some of his notable works in this field include Fisher (1912, 1922, 1925) [48, 49, 50]. To delineate the concept, let $\{X_1, \ldots, X_n\}$ be an i.i.d. random sample from some unknown probability distribution with PDF g (CDF G) and we model this true unknown PDF by the parametric model family $\mathcal{F}_\Theta = \{f_\theta : \theta \in \Theta\}$ of densities. The maximum likelihood philosophy estimates the unknown parameter θ by maximizing the likelihood function with respect to θ over Θ which is defined as

$$L(\theta) = L(\theta | X_1, \ldots, X_n) = \prod_{i=1}^{n} f_\theta(X_i). \tag{1.14}$$

Note that $L(\theta)$ is actually the joint density of the random sample but the same is observed as a function of θ. To maximize the likelihood function with respect to θ, it is convenient to use the logarithm of the likelihood function instead, i.e.,

$$\log L(\theta) = \sum_{i=1}^{n} \log f_\theta(X_i).$$

Note that, maximizing $L(\theta)$ with respect to θ is equivalent to maximizing $\log L(\theta)$ with respect to θ. Generally, it is more convenient to maximize $\log L(\theta)$ as most of our standard parametric models belong to the exponential family. Assuming the model density f_θ to be differentiable with respect to θ, the maximizer of the log likelihood function can be found by equating its derivative with respect to θ to 0 and then solving for θ. Let us observe that,

$$\frac{\partial}{\partial \theta} \log L(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f_\theta(X_i) = \sum_{i=1}^{n} u_\theta(X_i).$$

Consequently, the maximum likelihood estimator of θ is defined as the solution of the estimating equation (the likelihood score equation)

$$\sum_{i=1}^{n} u_\theta(X_i) = 0. \tag{1.15}$$

Standard regularity conditions are required to prove asymptotic results for the maximum likelihood estimators, such as consistency and asymptotic normality (see Lehmann (1983) [92]). Under these assumptions, the maximum likelihood estimating equations have a consistent root $\hat{\theta}_n$ of Equation (1.15), such that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, I^{-1}(\theta_0)),$$

where θ_0 is the true value of the parameter. However, the maximum likelihood estimator faces several robustness issues under anomalous observations and model misspecification. To understand the robustness of the maximum likelihood estimator, let us first define the maximum likelihood functional in terms of the true distribution function G. The maximum likelihood functional $T_{\mathrm{ML}}(G)$ can be implicitly defined as the solution of

$$\int u_\theta(x) dG(x) = 0.$$

It can be easily shown that (Basu et al. (2011) [12]) the influence function of the maximum likelihood functional is

$$IF(y, T_{\mathrm{ML}}, G) = I^{-1}(T_{\mathrm{ML}}(G)) u_{T_{\mathrm{ML}}(G)}(y). \tag{1.16}$$

Thus, the robustness of the maximum likelihood functional solely depends on the nature of the score function u_θ as a function of the contaminating point mass y. In most parametric models including the exponential models, $u_\theta(y)$ is an unbounded function of y indicating the robustness issues with the maximum likelihood estimators.

A robust generalization of the maximum likelihood approach was later proposed by means of the weighted likelihood philosophy. The weighted likelihood framework generalizes the notion of maximum likelihood estimation by attaching data-driven weights to the likelihood contributions (i.e., individual scores). Often these weights are chosen with the aim of downweighting outlying observations which ensures robustness of the

estimators obtained by solving the weighted likelihood estimating equations. Many of the minimum distance methods (discussed in Section 1.8) ultimately boil down to solving weighted likelihood based estimating equations with weight functions which potentially downweight the outlying observations present in the data and thus these minimum distance inference procedures become robust. Weighted likelihood based estimating equations have deep connection with the M-estimation approach. Some of the phenomenal works in the area of weighted likelihood approach include the pioneering work of Green (1984) [62], which discussed the application of iteratively reweighted least squares technique in the maximum likelihood estimation and proposed an alternative to the maximum likelihood estimator by considering a weighted version of the score equation, Lenth and Green (1987) [93], the works by Hu (1997) [74], Hu and Zidek (2001) [75] and (2002) [76]. The works of Markatou et al. (1997) [106], (1998) [107] and Markatou (2000) [105] provide further advancements in the literature of weighted likelihood based inference. In particular, Markatou (2000) [105] assessed the performance of weighted likelihood based robust inference in the context of mixture models. Agostinelli and Markatou (2001) [3] proposed the application of weighted likelihood approach in the development of robust hypothesis testing procedures. One very appealing feature of the weighted likelihood methodology as proposed by Markatou and her collaborators is that many of these methods achieve strong robustness properties simultaneously with full asymptotic model efficiency, and the downweighting effect eventually vanishes at the correct model.

1.6 M-Estimation

As we have seen, the maximum likelihood estimators may not be robust against data contamination and model misspecification. In trying to develop robust alternatives to the maximum likelihood estimator, let us note that the left hand side of Equation (1.15) is a sum of identically and independently distributed random variables $u_\theta(X_i)$. This idea can further be generalized by replacing u_θ with some general function $\psi : \mathcal{X} \times \Theta \to \mathbb{R}$ and define the estimator $\hat{\theta}_n$ as the solution of

$$\sum_{i=1}^{n} \psi(X_i, \theta) = 0. \tag{1.17}$$

One can then choose the function ψ, so that the corresponding estimator has suitable bounded influence function. This type of estimators are known as M-estimators and has

held a prominent place in the literature of robust statistics for several decades. In fact, many of the well-known robust statistical tools ultimately boil down to M-estimation. In the aforesaid discussion, the M-estimator $\hat{\theta}_n$ is defined as a solution of the estimating equation (1.17). This type of M-estimators are called ψ-type M-estimators. However, an M-estimator can also be defined as the minimizer of some "reasonable" objective function which is very common in practice. That is, the M-estimator $\hat{\theta}_n$ can also be thought of as a minimizer of

$$\sum_{i=1}^{n} \rho(X_i, \theta),$$

where $\rho : \mathcal{X} \times \Theta \to \mathbb{R}$ and the estimator is based on the random sample $\{X_1, \ldots, X_n\}$ of size n. This type of M-estimators are known as ρ-type M-estimators.

We will focus on ψ-type M-estimators for mathematical convenience. The functional $T_\psi(G)$ corresponding to the ψ-type M-estimators can be defined implicitly as the solution of

$$\int \psi(x, T_\psi) dG(x) = 0,$$

where G is the true distribution function. Under the differentiability of the ψ function and certain other regularity assumptions, it can be shown that the influence function of T_ψ is given by

$$IF(y, T_\psi, G) = \frac{\psi(y, T_\psi(G))}{\int \psi'(x, T_\psi(G)) dG(x)},$$

and

$$\sqrt{n}(T_\psi(G_n) - T_\psi(G)) \xrightarrow{d} N(0, \sigma_g^2),$$

where, $\sigma_g^2 = \frac{\int \psi^2(x, T_\psi(G)) dG(x)}{\left(\int \psi'(x, T_\psi(G)) dG(x)\right)^2}$. Let us consider some of the popular examples of M-estimators.

(i) The maximum likelihood estimator is itself an M-estimator with $\rho(x, \theta) = -\log\ f_\theta(x)$ and, consequently, $\psi(x, \theta) = u_\theta(x)$.

(ii) The Huber's M-estimator is defined as either the minimizer of $\sum_{i=1}^{n} \rho_k(X_i, \theta)$, where $\rho_k(x, \theta) = \rho_k(x - \theta)$ is defined through the relation

$$\rho_k(y) = \begin{cases} y^2, & \text{if } |y| \leq k, \\ 2k|y| - k^2, & \text{if } |y| > k, \end{cases}$$

or, equivalently, it is the solution of $\sum_{i=1}^{n} \psi_k(X_i, \theta) = 0$, where $\psi_k(x, \theta) = \psi_k(x - \theta)$ is defined through the relation

$$\psi_k(y) = \begin{cases} y, & \text{if } |y| \leq k, \\ \text{sign}(y)k, & \text{if } |y| > k. \end{cases}$$

Here, k is a positive tuning constant. This M-estimator is useful in the context of the location model.

It is to be noted that in the first example (i.e., the maximum likelihood estimator), the choice of the ψ function is not bounded in general, whereas, the choice of the same in the second example (i.e., Huber's M-estimator) is bounded. Consequently, the maximum likelihood estimators are non-robust in general (depending on the boundedness of the corresponding choice of ψ), but the Huber's M-estimator is always robust. In particular, the Huber's M-estimator has much improved robustness properties in the location model. There are other types of M-estimators which have been studied and utilized for different purposes. We refer to Tyler (1987, 2014) [149, 150], Maronna (1976) [109], Maronna et al. (2019) [110], Hampel et al. (1986) [68] and Huber (2004) [79] for detailed discussions on M-estimators.

1.7 Statistical Distances

The principal objective of parametric statistical modelling is to understand the overall structure of the data by means of probability distributions. Statistical distances (also referred to as divergences) may be conveniently utilized for this purpose. The notion of a mathematical distance is formalized by the definition of metrics in the literature of mathematical analysis. However, statistical distances need not be proper metrics in that sense. To understand this, let us recall that if (M, d) denotes a metric space with the underlying metric d, then $d : M \times M \to \mathbb{R}$ has to satisfy

(i) $d(x, x) = 0 \ \forall x$,

(ii) $d(x, y) > 0$, if $x \neq y$,

(iii) $d(x, y) = d(y, x)\ \forall x,\ y$, and,

(iv) $d(x, y) + d(y, z) \geq d(x, z)\ \forall x,\ y$ and z.

Here, the first two conditions describe the non-negativity, the third one expresses the symmetry and the last one states the triangular property of the metric. However, the last two conditions are not necessary for d to be a divergence. In particular, many divergences are asymmetric in their arguments and this plays a major role in controlling the theoretical properties of the estimators based on these measures. It is necessary for statistical distance measures to be non-negative and to be equal to zero if and only if the two arguments are identically equal. Let us now formally define statistical distances or divergences.

Let $\mathcal{F}$ be a collection of non-negative real-valued functions. Then $d : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ is a valid statistical distance if,

(i) $d(G, F) \geq 0\ \forall G, F \in \mathcal{F}$ and

(ii) $d(G, F) = 0$ if and only if $G = F$.

Here the arguments of the statistical distances are distribution functions. Some of the well-known examples of distribution function based divergences include the Kolmogorov-Smirnov distance, the Cramér-von Mises distance, Anderson-Darling distance and many more (see Basu et al. (2011) [12] for an illustrative discussion). On the other hand, when the arguments of the measure $d(\cdot, \cdot)$ in the above formulation are PDFs, it results in a density based divergence. As divergences belonging to this class represent the main medium for the approach in this thesis, we provide an expanded description in the following.

1.7.1 Density based Divergences

Suppose, we have a random sample $\{X_1, \ldots, X_n\}$ from the true unknown probability distribution with CDF G and suppose its PDF g is modelled by a parametric family of densities $\{f_\theta : \theta \in \Theta\}$. A density based divergence $d(g, f)$ between two probability density functions g and f with respect to the same measure quantifies the separation between these two densities (or the distributions that they represent). In our context, we are interested in measuring the separation between the true unknown density function g and the model PDF f_θ. Since g is unknown, either we have to plug-in a nonparametric estimate of g (based on the aforesaid random sample) in the expression

of the divergence (when explicitly required) or make use of the empirical distribution function G_n to get an empirical estimate of the true divergence. Let us consider some examples in the following:

1.7.2 Discrete Models

To describe the treatment of discrete models, we assume (without loss of generality) that the true unknown density function g (with respect to the counting measure) is supported on $\mathcal{X} = \{0, 1, 2, \ldots, \infty\}$. Let $\boldsymbol{d}_n$ be the vector of relative frequencies. We use $\boldsymbol{d}_n$ to estimate the true unknown probability density function g.

Let C be a convex, thrice differentiable function on $[-1, \infty)$ with $C(0) = 0$ and $\delta(x) = \frac{d_n(x)}{f_\theta(x)} - 1$ (known as the Pearson residual). Then a particular but rich class of divergences between the estimated true density d_n and the model density f_θ generated by the function C can be defined as,

$$\rho_C(d_n, f_\theta) = \sum_{x=0}^{\infty} C(\delta(x)) f_\theta(x). \tag{1.18}$$

The expression $\rho_C(d_n, f_\theta)$ can be shown to represent a valid divergence (as per the definition) using the conditions imposed on C and Jensen's inequality (see Basu et al. (2011) [12]). Different choices of C generate various divergences between d_n and f_θ. Some of the well-known divergences of this type are listed in the following table:

Divergence	$C(\delta)$	Algebraic form
Likelihood Disparity (LD)	$(\delta+1)\log(\delta+1) - \delta$	$\sum d_n \log \frac{d_n}{f_\theta}$
Kullback-Leibler Divergence (KLD)	$\delta - \log(\delta+1)$	$\sum f_\theta \log \frac{f_\theta}{d_n}$
Hellinger Distance (HD)	$2(\sqrt{\delta+1}-1)^2$	$2\sum(\sqrt{d_n} - \sqrt{f_\theta})^2$
Pearson's Chi-square (PCS)	$\frac{\delta^2}{2}$	$\sum \frac{(d_n - f_\theta)^2}{2 f_\theta}$
Neyman's Chi-square (NCS)	$\frac{\delta^2}{2(\delta+1)}$	$\sum \frac{(d_n - f_\theta)^2}{2 d_n}$

Table 1.1: Examples in Discrete cases.

The divergences of the type described in Equation (1.18) are referred to in the literature as ϕ-divergences, f-divergences or disparities. See Csiszár (1963) [30] and Lindsay (1994) [94] for further details in this topic.

1.7.3 Continuous Models

In case of continuous distributions, let us consider the following examples:

(i) The twice, squared Hellinger distance between the densities g and f_θ is defined as

$$\mathrm{HD}(g, f_\theta) = 2 \int (\sqrt{g(x)} - \sqrt{f_\theta(x)})^2 \, dx.$$

Here g is unknown and a natural substitute is a kernel density estimate of g based on the random sample $\{X_1, X_2, \ldots, X_n\}$ which is defined as

$$g_n^*(x) = \frac{1}{n} \sum_{i=1}^{n} K(x, X_i, h_n),$$

where $K(\cdot, X_i, h_n)$ is a smooth kernel function, usually symmetric about X_i, with bandwidth h_n. One of the disadvantages of this distance is the additional theoretical complexity and computational burden due to the kernel density estimation especially with growing data dimensions along with increasing estimation bias. The choice of the bandwidth h_n is another critical issue in kernel density estimation. However, the second example below can circumvent this difficulty by virtue of a special algebraic form.

(ii) The Bregman divergence, indexed by a strictly convex and differentiable function $B(\cdot)$, between the densities g and f_θ is defined as

$$\begin{aligned} D_B(g, f_\theta) &= \int [B(g(x)) - B(f_\theta(x)) - (g(x) - f_\theta(x))B^{'}(f_\theta(x))]dx \\ &= \int B(g(x))dx - \int B(f_\theta(x))dx - \int g(x)B^{'}(f_\theta(x))dx \\ &+ \int f_\theta(x)B^{'}(f_\theta(x))dx. \end{aligned}$$

Since we wish to minimize $D_B(g, f_\theta)$ as a function of θ, we may ignore the first term in the right hand side of the aforesaid equation as it is free of θ. Only the third term depends on the unknown g in the remainder of the expansion. But

this term can be rewritten as

$$\int g(x)B^{'}(f_\theta(x))dx = E_G(B^{'}(f_\theta(X)))$$

which facilitates to approximate the same by replacing the true distribution function G with the empirical distribution function G_n as defined in Equation (1.1). Here, we do not need to make use of kernel density estimate of g which may be challenging in case of higher data dimensions.

1.8 Minimum Distance Estimation

The notion of statistical distances has been introduced in the previous section. As we have seen, a generic density based divergence D ideally proposes to quantify the discrepancy between the data and the model by measuring the distance between the true density g and the model density f_θ. Hence, the optimal choice $\hat{\theta}$ of $\theta \in \Theta$ should minimize the aforesaid distance between g and f_θ, i.e.,

$$\hat{\theta} = \underset{\theta\in\Theta}{\operatorname{argmin}}\, D(g, f_\theta). \tag{1.19}$$

In practice, of course, the true density g will be unknown and $D(g, f_\theta)$ will have to be empirically estimated, either through nonparametric density estimation, or through the substitution of G by G_n.

The principal aim of this thesis is to provide robust and efficient estimators of the unknown parameter θ. Minimization of certain statistical distances are utilized in this purpose. However, the asymptotic and robustness properties of the minimum distance estimators heavily depend on the form of the divergence which is being minimized. In particular, the asymptotic normality (with the corresponding convergence rate), boundedness of the influence functions and the breakdown behaviour explicitly depend on the statistical distance which is being minimized. In fact, some of the distances (e.g., the likelihood disparity) may produce completely non-robust estimators with very high (often, maximum possible) efficiency. On the other hand, there are many distances which produce highly robust estimators with a minor loss in asymptotic efficiency. The trade-off between robustness and efficiency significantly depends on the structure of the underlying distance which is being minimized.

Many of the minimum distance estimators have been found to be M-estimators and

the standard treatment for M-estimators can be utilized in order to derive the asymptotic and robustness properties of the same under general assumptions. However, the asymptotics of some of these minimum distance estimators have later been established using less stringent model assumptions. See, e.g., Basu et al. (1998) [10] for a development of the minimum divergence estimators in case of the density power divergence (DPD) in a similar spirit.

1.9 The Density Power Divergence

One of the main objectives of this thesis is to provide robust and efficient estimators of multivariate location and scale parameters with less computer intensive algorithms and apply them in various machine learning problems. Estimation through minimization of suitable statistical distances is one possible way to do that. As we have seen in Section 1.7.3, the true unknown density has to be estimated using some nonparametric method like the kernel density estimation procedure in case of continuous models for many of the distances. Consequently, the overall distance minimization procedure inherits all the complications involved with the aforesaid nonparamtric smoothing, such as selecting the optimal bandwidth or choosing the appropriate kernel function.

But we have also observed in the second example of Section 1.7.3 that such complications can be bypassed if the distance assumes a particular type of algebraic form. DPD is one such distance which was originally proposed by Basu et al. (1998) [10]. This distance is indexed by a single non-negative tuning parameter β. The role of β is crucial in controlling the trade-off between the robustness and asymptotic efficiency of the estimators derived by minimizing the DPD. To formally define the distance, let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from the true unknown distribution with PDF g (CDF G) and assume this unknown density g to be modelled by the family $\mathcal{F}_\Theta = \{f_\theta : \theta \in \Theta\}$. Then the DPD between the true density g and the model density f_θ is defined as

$$D_\beta(g, f_\theta) = \int \left[f_\theta^{1+\beta}(x) - \left(1 + \frac{1}{\beta}\right) g(x) f_\theta^\beta(x) + \frac{1}{\beta} g^{1+\beta}(x) \right] dx, \tag{1.20}$$

and consequently, the minimum DPD estimator (MDPDE) is defined as

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}}\, D_\beta(g, f_\theta). \tag{1.21}$$

Here, the tuning parameter β crucially controls the nature of the DPD itself as well as that of the MDPDE. As $\beta \to 0$, the DPD reduces to $\int g(x) \log \frac{g(x)}{f_\theta(x)} dx$ which is precisely the likelihood disparity and for $\beta = 1$, the divergence becomes the usual squared error loss function $\int (g(x) - f_\theta(x))^2 dx$. As a consequence, the MDPDE coincides with the usual MLE for the limiting case $\beta \to 0$ which is highly non-robust but most model efficient and as $\beta = 1$ the minimum DPD estimator becomes very stable and robust but with a fair loss in efficiency (see Basu et al. (2011) [12]). Thus, the tuning parameter β actually controls the trade-off between the asymptotic efficiency and the robustness of the resulting MDPDE with lower β values leading to greater efficiency and higher β values correspond to greater stability. From efficiency considerations, we avoid very large values of β and in practice, we restrict β to vary within the interval $[0, 1]$. The DPD can also be recognised a member of the Bregman divergence family as $B(f) = f^{1+\beta}$ provides β times the right hand side of Equation (1.20).

This distance (i.e., the DPD) is the main theme of this thesis. Our contributions are based on this distance as well as the estimators obtained by minimizing the same in various set-ups. Thus, it is essential to delineate the properties of the MDPDE at first to set up the background of this thesis.

(i) **Minimization:** We need to minimize the DPD measure in order to derive the estimator $\hat{\theta}$. To do that, let us first observe that

$$\begin{aligned} D_\beta(g, f_\theta) &= \int \left[f_\theta^{1+\beta}(x) - \left(1 + \frac{1}{\beta}\right) g(x) f_\theta^\beta(x) + \frac{1}{\beta} g^{1+\beta}(x) \right] dx \\ &= \int f_\theta^{1+\beta}(x) dx - \left(1 + \frac{1}{\beta}\right) \int g(x) f_\theta^\beta(x) dx + \frac{1}{\beta} \int g^{1+\beta}(x)\, dx, \end{aligned}$$

where the last term in the right hand side is free of θ and thus can be ignored in the minimization process. The second term involves the true unknown density g, but fortunately, this term can be written as a constant multiple of $E_G(f_\theta^\beta(X))$ (as in case of the Bregman divergence). So, we may follow a similar approach (as in case of the Bregman divergence) by approximating $E_G(f_\theta^\beta(X))$ with $E_{G_n}(f_\theta^\beta(X))$. Thus, it is enough to minimize

$$H_n(\theta) = \int f_\theta^{1+\beta}(x)\, dx - \left(1 + \frac{1}{\beta}\right) \frac{1}{n} \sum_{i=1}^{n} f_\theta^\beta(X_i) \tag{1.22}$$

with respect to θ in order to get $\hat{\theta}$. However, the actual minimization of $H_n(\theta)$

is not straightforward as a closed form solution may not exist in many cases. Iterative algorithms have to be invoked in order to achieve the minimization of DPD. Although some research has already been done in this area, one of the principal objectives of this thesis is to provide computationally efficient algorithms to minimize the DPD in various complex multivariate set-ups either explicitly or implicitly and derive the corresponding theoretical properties with relevant applications in other scientific disciplines.

(ii) **Minimum DPD Functional:** The estimator $\hat{\theta}$ is shown to be the minimizer of $H_n(\theta)$ which is a functional of the empirical distribution function G_n based on the random sample $\{X_1, X_2, \ldots, X_n\}$. Analogously, we need to express the unknown parameter θ as a functional of the true distribution function G in order to derive the asymptotic and robustness properties of the MDPDE $\hat{\theta}$. We call this the minimum DPD functional and define it as

$$\theta^g = \underset{\theta \in \Theta}{\operatorname{argmin}}\, H(\theta), \tag{1.23}$$

where

$$H(\theta) = \int f_\theta^{1+\beta}(x)\, dx - \left(1 + \frac{1}{\beta}\right) \int g(x) f_\theta^{\beta}(x) dx,$$

which can be thought of as the population version of $H_n(\theta)$.

(iii) **Some Prerequisite Expressions:** We need to introduce some notation and expressions for stating the asymptotic properties of MDPDEs (Basu et al. (1998, 2011) [10, 12]). In the following, we allow $\boldsymbol{\theta}$ to be a vector valued parameter. Let us recall that the score function is

$$u_{\boldsymbol{\theta}}(x) = \nabla \log f_{\boldsymbol{\theta}}(x),$$

and the information function is defined as

$$i_{\boldsymbol{\theta}}(x) = -\nabla u_{\boldsymbol{\theta}}(x).$$

Let us also define $\boldsymbol{J} = \boldsymbol{J}(\boldsymbol{\theta}^g)$ and $\boldsymbol{K} = \boldsymbol{K}(\boldsymbol{\theta}^g)$ as

$$\boldsymbol{J} = \int u_{\boldsymbol{\theta}^g}(x) u_{\boldsymbol{\theta}^g}^t(x) f_{\boldsymbol{\theta}^g}^{1+\beta}(x) dx + \int \left[(i_{\boldsymbol{\theta}^g}(x) - \beta u_{\boldsymbol{\theta}^g}(x) u_{\boldsymbol{\theta}^g}^t(x))(g(x) - f_{\boldsymbol{\theta}^g}(x)) f_{\boldsymbol{\theta}^g}^{\beta}(x) \right] dx, \tag{1.24}$$

and

$$\boldsymbol{K} = \int u_{\boldsymbol{\theta}^g}(x) u_{\boldsymbol{\theta}^g}^t(x) f_{\boldsymbol{\theta}^g}^{2\beta}(x) g(x) dx - \boldsymbol{\xi}\boldsymbol{\xi}^t, \tag{1.25}$$

where

$$\boldsymbol{\xi} = \int u_{\boldsymbol{\theta}^g}(x) f_{\boldsymbol{\theta}^g}^{\beta}(x) g(x) dx. \tag{1.26}$$

(iv) **Technical Assumptions:** We need the following technical assumptions in order to establish the asymptotic properties of the minimum DPD estimators. These assumptions have been explicitly made by Basu et al. (1998, 2011) [10, 12]. These assumptions (or similar ones) will be assumed over and over again in this thesis in the following chapters to prove various theoretical results in our contributions.

(C1) Both of the true and the model distributions share the same support $\mathcal{X} = \{x | f_{\boldsymbol{\theta}}(x) > 0\}$ which is independent of the parameter $\boldsymbol{\theta}$.

(C2) There exists an open subset $\boldsymbol{\omega}$ of the parameter space $\boldsymbol{\Theta}$ containing the best fitting parameter $\boldsymbol{\theta}^g$ such that $\forall \boldsymbol{\theta} \in \boldsymbol{\omega}$ and for almost all $x \in \mathcal{X}$, the density $f_{\boldsymbol{\theta}}(x)$ is three times differentiable with respect to $\boldsymbol{\theta}$ with continuous third derivatives.

(C3) The integrals $\int f_{\boldsymbol{\theta}}^{1+\beta}(x) dx$ and $\int f_{\boldsymbol{\theta}}^{\beta}(x) g(x) dx$ can be differentiated three times with respect to $\boldsymbol{\theta}$ and the derivatives can be taken under the integral sign.

(C4) The matrix $\boldsymbol{J}(\boldsymbol{\theta})$ defined in Equation (1.24) is positive definite.

(C5) There exists a function $M_{jkl}(x)$ such that

$$|\nabla_{jkl} V_{\boldsymbol{\theta}}(x)| \leq M_{jkl}(x), \text{ for all } \boldsymbol{\theta} \in \boldsymbol{\omega},$$

where $E_g(M_{jkl}(X)) = m_{jkl} < \infty$, for all j, k and l. Here $V_{\boldsymbol{\theta}}(x) = \int f_{\boldsymbol{\theta}}^{1+\beta}(x) dx - \left(1 + \frac{1}{\beta}\right) f_{\boldsymbol{\theta}}^{\beta}(X)$.

(v) **Asymptotic Properties:**

Theorem 1.1 (Basu et al. (1998) [10])**.** *Under the aforesaid regularity assumptions, we have the following properties:*
(Consistency) The minimum DPD estimating equation

$$\nabla H_n(\boldsymbol{\theta}) = \mathbf{0} \tag{1.27}$$

has a consistent sequence of roots.
(Asymptotic Normality) $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^g)$ *has an asymptotic multivariate normal distribution with zero mean and covariance matrix* $\boldsymbol{J}^{-1}\boldsymbol{K}\boldsymbol{J}^{-1}$*.*

(vi) **Influence Function:** The minimum DPD functional $\boldsymbol{T}_\beta(G)$ can implicitly be defined as the solution of

$$\nabla H(\boldsymbol{\theta}) = \mathbf{0}. \tag{1.28}$$

The influence function of the minimum DPD functional $\boldsymbol{T}_\beta(G)$ (Basu et al. (2011) [12]) is given by

$$\text{IF}(y, \boldsymbol{T}_\beta, G) = \boldsymbol{J}^{-1}(u_{\boldsymbol{T}_\beta(G)}(y) f^\beta_{\boldsymbol{T}_\beta(G)}(y) - \boldsymbol{\xi}).$$

1.10 Outline of the Thesis

The principal objective of this thesis is, in a nutshell, to provide robust estimators of multivariate location and scale which have reasonable to high model efficiency but avoid high computational complexity so as to be practically useful in real problems. We utilize the minimum DPD and the related philosophy to invoke robustness. There are some computational issues while minimizing the DPD in different multivariate set-ups. We will work on this problem rigorously and come up with three types of estimation procedures which are explicitly or implicitly related to the minimum DPD methodology, keeping the computational issue in mind each time. We derive the theoretical properties (asymptotic and robustness features) of these methods, empirically validate them with extensive simulation studies in various set-ups and apply them in different problems in the domains of pattern recognition and machine learning.

In Chapter 2 (Chakraborty et al. (2023) [22]), we develop an iteratively reweighted least squares (IRLS) algorithm to compute the MDPDEs in case of multivariate normal

models and subsequently apply them to derive suitable minimum divergence estimators of the component parameters of the multivariate normal mixture model in the spirit of the MDPDE. Fujisawa and Eguchi (2006) [54] took a direct approach for minimizing the DPD objective function in case of univariate normal mixture models. However, we observe that direct minimization of the DPD in case of multivariate normal mixture models is computationally challenging especially under growing data dimensions. To handle this problem, we utilize a particular discriminant rule to construct the initial cluster configurations which allows us to estimate the parameters of each multivariate normal component separately rather than under the composite mixture set-up. This approach leads to the development of the maximum pseudo β-likelihood (MPLE$_\beta$) algorithm which combines robust estimation of the parameters of the different multivariate normal components followed by a robust clustering and anomaly detection exercise. In our numerical studies, we have found this new algorithm MPLE$_\beta$ to be competitive or better than many of the well-known robust clustering algorithms including the TCLUST, trimmed K-means, MCLUST and K-medoids, apart from leading to very substantial computational improvement over the direct minimization of the DPD objective function. Note that, because of the modification of the objective function to bypass the direct minimization of the DPD, the resultant estimator is not strictly the same as the MDPDE (although similar in spirit). We will refer to this estimator as the maximum pseudo β-likelihood estimator (MPLE$_\beta$). The MPLE$_\beta$ clustering algorithm is also applied to a satellite image for its reconstruction. To understand the robustness of this algorithm, we perform the influence function analysis and the method is found to be robust at multivariate normal mixture models. Existence and weak consistency of the estimators are established under certain regularity assumptions in Chapter 3 (Chakraborty et al. (2022) [21]) with further simulations and real data examples. However, the IRLS algorithm developed for the minimum DPD estimation in case of multivariate normal model is a naive algorithm and it is sometimes found to be computationally problematic, particularly, in case of higher data dimensions and higher values of the tuning parameter β (typically for $\beta > 0.5$).

Motivated by the need to bypass this difficulty, we propose a new componentwise robust estimation procedure of multivariate location and scale parameters in Chapter 4 (Chakraborty et al. (2024) [23]). It is a method which is sequential in spirit where the MDPDEs of component means, variances and correlations are obtained, separately. This method, with high probability, is computationally far more tractable and the

convergence of this algorithm is guaranteed unlike the aforesaid IRLS procedure. Theoretical properties, such as consistency and asymptotic normality of the estimators, the asymptotic positive definiteness of the scale estimators and the influence function analyses of the estimators (suitably defined functionals) are studied. Apart from the MDPDE which is obtained as the simultaneous minimizer of the overall DPD objective function, this new method is also compared with many of the popular robust estimation tools through simulation experiments in various set-ups. A real data (credit card fraudulent transactions) application is also presented to understand the robust nature of our method in case of higher data dimensions. But, it is found that this procedure may have very high time complexity specifically with growing data dimensions.

Two types of robust estimation procedures using the idea of minimum DPD estimation have been described in the preceding two paragraphs. The first one may be computationally problematic in case of higher data dimensions and higher values of β while the second one has been found to have higher time complexity with growing data dimensions although failure to convergence is no longer an issue. It is essential to develop a robust estimation method which is efficient (from theoretical angle) as well as computationally fast to facilitate practitioners. We develop one such robust estimation methodology primarily with reduced computational cost without compromising the efficiency to the extent possible in Chapter 5. We utilize the idea of one-step M-estimators to achieve this goal. M-estimators (ψ-type) are defined as solutions of Equation (1.17) which need iterative procedures most of the time in practice. One-step estimators are derived from initial highly robust estimators by performing just the first iteration step of any iterative method. Existing literature shows some of its remarkable statistical properties, such as preservation of breakdown (i.e., the breakdown point of the one-step update is same as that of the initial highly robust estimate) and increased asymptotic efficiency under theoretical assumptions. Motivated by these, we study the one-step version of the original minimum DPD estimation method to build up a computationally faster and theoretically efficient robust procedure for multivariate location-scale estimation. Here we consider the Newton-Raphson, gradient descent, iteratively reweighted least squares and Fisher's scoring iterations with different robust initializations to solve the minimum DPD estimating equation (1.27) and study the corresponding one-step estimators. Theoretical properties like consistency and asymptotic normality and influence function analyses of these one-step estimators

are studied under certain regularity assumptions. These one-step estimators are also assessed through extensive simulation studies and two real data examples. Later, a slightly modified version of the one-step IRLS estimators are studied especially in multivariate location-scale set-up where the location estimator is a weighted sample mean and the scale estimator is proportional to the weighted sample covariance matrix with the same set of weights. We consider density power weights in this context. Theoretical properties of these estimators (consistency, asymptotic normality and preservation of breakdown point) are studied following Lopuhaä and Rousseeuw (1991) [98] and Lopuhaä (1999) [100] along with extensive simulation experiments. Finally, this robust method is applied to a classification problem in case of prediction involving survival of heart failure patients.

We conclude our thesis with concluding remarks and possible future directions of our research in Chapter 6.

Chapter 2

Robust Clustering via Maximum Pseudo β-Likelihood Estimation

2.1 Introduction

Mixture distributions arise in many common practical situations. In particular, when the population is not homogeneous due to the presence of different categorical attributes, the variable of interest has different behaviour over distinct homogeneous subgroups which come together to generate an overall heterogeneous mixture system. To draw statistical inference based on this kind of heterogeneous datasets, a single probability distribution may not be adequate to model the data; mixture distributions provide the appropriate structure in these situations. Mixtures of many different distributions with varying shapes have been used in the literature to model datasets coming from different disciplines ranging over astronomy, clinical psychology, economics, finance, DNA sequencing, image processing, voice recognition, criminology, species counting and many others. In parametric mixture modelling, one uses a model probability distribution constructed as a mixture (convex combination) of several probability distributions from a particular parametric class with different parameter values. Mathematically, a mixture probability distribution can be described in terms of its probability density function $f_{\boldsymbol{\theta}} = \sum_{j=1}^{k} \pi_j f_{\boldsymbol{\theta}_j}$, where π_j is the weight given to the j-th component of the mixture having PDF $f_{\boldsymbol{\theta}_j}$ ($\boldsymbol{\theta}_j$ being the j-th component parameter) for $j = 1, \ldots, k$ with $\sum_{j=1}^{k} \pi_j = 1$, and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_k)$ is the parameter vector of interest. Here the component densities $f_{\boldsymbol{\theta}_j}$, $j = 1, \ldots, k$, are coming from a particular parametric family (e.g., normal or Cauchy). In practice, the weights π_js are unknown and hence they also have to be estimated along with the parameters ($\boldsymbol{\theta}_j$s) of the component distributions. Normal, Cauchy and Laplace are perhaps the most common symmetric mixture examples. The normal mixture models are flexible, can fit a large variety of shapes, and are among the most popular and most used statistical tools in practice. Moreover, gamma scale mixtures of normal components cover a

very large class of symmetric mixture distributions (e.g., Andrews and Mallows (1974) [4]). For non-symmetric or skewed mixture distributions, uniform mixtures of normal distributions are used (e.g., Qin et al. (2003) [121]). For general references covering different areas of mixture models (including non-normal mixtures), see Titterington et al. (1985) [144], Lindsay (1995) [95] and McLachlan and Peel (2004) [114].

Motivated by their huge applicability, here we develop a robust estimation procedure for normal mixture models. However, likelihood based inference, which is asymptotically the most efficient under the model, is strongly affected by outliers and model misspecification. To address the robustness issue, here we take a minimum distance approach in the spirit of the DPD. We are going to view the problem of minimizing the DPD as a maximization of a generalized likelihood function. One of our primary objectives in performing robust inference in normal mixtures is its subsequent use in robust clustering. Clustering is an active area of research and has many real life applications. Some of the existing clustering methods available in the literature are K-means, K-medoids and standard likelihood discrimination. The problem with the K-means method is that it tends to find only spherical or elliptical clusters with "roughly" equal cluster sizes and equal component covariance matrices. Also the method is based on traditional cluster mean estimates which makes the algorithm non-robust in the presence of "anomalous" observations. The presence of anomaly is not a rare thing in practice. Subjective deletion of outliers, inadvisable as it is, cannot be done in high dimensions where the data cannot be visualized. Hence, some of the small components of the cluster may be misspecified and the observations coming from these irregular clusters become disturbing.

Various modifications of the K-means clustering algorithm have been proposed in the literature with the aim of robustifying the algorithm. The trimmed K-means method (Cuesta-Albertos et al. (1997) [31]) is one such example where the same objective function as the K-means is used but only using a subsample after trimming the extreme observations from the whole sample. To further generalize the K-means and trimmed K-means algorithms beyond spherical or elliptical clusters, the concept of heterogeneous clustering has been considered in the literature. Gallegos and Ritter (2005) [57] proposed a normal mixture set-up in this context under the "spurious-outliers model" and developed an algorithm for estimation which is a naive extension of the Minimum Covariance Determinant (MCD) algorithm (Rousseeuw (1984) [128], (1985) [129]). But the estimation procedure under this model is too difficult because

the algorithm often ends up finding clusters made up of observations lying on a low dimensional subspace. Moreover, the likelihood function can be unbounded; when one of the observations is equal to one of the component means, the likelihood tends to infinity as the determinant of the dispersion matrix of that component goes to zero. To bypass this difficulty, Gallegos (2002) [56] and Gallegos and Ritter (2005) [57] proposed two additional methods based on the structure of the component covariance matrices. The first one proposed an algorithm assuming similar scales of the components while the second one assumes that the dispersion matrices have an unknown but same covariance structure. Later, García-Escudero et al. (2008) [58] proposed the TCLUST approach which performs a likelihood based discrimination with trimming a certain proportion of outlying observations; it is still based on the maximum likelihood estimates of the parameters but obtained only from non-trimmed observations. This method is very popular and heavily used for robust clustering under normal mixture data.

Some of the other existing robust clustering algorithms based on finite Gaussian mixture models include Punzo and McNicholas (2016) [119], Scrucca et al. (2016) [137], Banfield and Raftery (1993) [8] which mainly model the contaminated datasets with finite mixture Gaussian models and perform parameter estimation with subsequent data clustering and outlier detection.

Divergence based clustering procedures are also used to derive robust and efficient parametric clustering literature. Here, the clustering methods are based on robust estimation of the model parameters by minimizing suitable divergence measures (between probability density functions). The γ-divergence is one of those divergences which are utilized to build robust clustering tools. The γ-divergence was originally proposed by Jones et al. (2001) [82] under a different name and was motivated later by Fujisawa and Eguchi (2006) [55] through the γ-cross entropy. Two interesting robust clustering algorithms were proposed by utilizing the γ-divergence, viz., those proposed by Chen et al. (2014) [24] and Notsu et al. (2014) [116]. The former developed a robust clustering algorithm by combining the q-Gaussian mixture model and minimum γ-divergence estimation procedure and applied the algorithm on cryo-EM data whereas the latter developed a hierarchical robust clustering algorithm by minimizing the γ-divergence.

In the present chapter, we develop a fully parametric robust approach to estimate the parameters under the normal mixture model which leads to a subsequent robust clustering strategy. An iteratively reweighted least squares approach is developed for estimating the component means and covariances. It is useful in the same estimation

and clustering set-up for which TCLUST is one of the existing robust procedures, but the source of robustness of our estimators and our algorithm lies in suitable density power downweighting of the observations rather than through invoking likelihood based trimming. A section of our assumptions are similar to those necessary for TCLUST, although there are also some assumptions specific to the form of the DPD. The contributions of the present chapter are highlighted as follows.

Firstly, we present a new robust method for model based clustering under the Gaussian mixture set-up developed in the spirit of the DPD. Thus, the proposed set of techniques represent a general class of estimation methods which contains likelihood based methods as a particular case. A single scalar tuning parameter controls the trade-off between efficiency and robustness and we demonstrate how positive but small values of the tuning parameter provide more stable performance under noisy data with little loss in efficiency compared to the likelihood based results.

Secondly, we develop an approximate EM-like algorithm to solve the problem efficiently even in higher dimensions. This is important since a straightforward optimization of the proposed objective function is difficult particularly in high dimensions. The proposed algorithm performs parameter estimation as a precursor to the detection of the clusters and helps to detect anomalous observations (if present) in the dataset. Subsequently, the algorithm leads to robust detection of clusters on the basis of the robust parameter estimates obtained earlier.

Thirdly, under noisy data, the proposed method provides improved results in terms of the estimated misclassification rates and outlier detection compared to the TCLUST (which possibly represents the most popular robust and clustering algorithm currently in use in this area which has an easily implementable software) and the trimmed K-means methods. Along with the robust methods like the TCLUST, trimmed K-means and K-medoids algorithms, we have also included the MCLUST method (Scrucca et al. (2016) [137]) in our analysis which optimally selects the number of clusters unlike the others. However, we use the MCLUST method with fixed number of clusters (with uniform noise component) in this work.

Finally, the usefulness of the proposed clustering procedure in image processing is illustrated through identifying differently colored (anomalous) regions from a colour image. With appropriately proposed additional refinements, our methodology is seen to outperform the aforesaid procedures also for this special application as illustrated through analysis of a real satellite image.

The rest of the chapter is organized as follows. In Section 2.2, we propose our algorithm along with the underlying theoretical formulations. In Section 2.3, we present some of the theoretical properties of our estimators and the behaviour of the influence functions is explored to justify the claimed robustness. A large scale comparative simulation study is presented in Section 2.4. Analysis of two real life datasets are considered in Section 2.5 while an application of our method to image processing is provided in Section 2.6. Technical proofs and derivations are presented in Section 2.7.

2.2 Proposed Parameter Estimation and Clustering Procedure

Let $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n\}$ be a random sample drawn from a p-dimensional multivariate normal mixture distribution with k components (k is fixed in our set-up) having an unknown PDF (since the true values of the component parameters are unknown) which is modelled by the model density $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{j=1}^{k} \pi_j \phi_p(\boldsymbol{x}, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ for $\boldsymbol{x} \in \mathbb{R}^p$, where $\phi_p(\cdot, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the PDF of a p-dimensional normal distribution with mean $\boldsymbol{\mu}$ and dispersion matrix $\boldsymbol{\Sigma}$. The parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is given by $\boldsymbol{\theta} = (\pi_1, \pi_2, \ldots, \pi_k, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \ldots, \boldsymbol{\Sigma}_k)$ where $\boldsymbol{\mu}_j \in \mathbb{R}^p$, $\boldsymbol{\Sigma}_j$ is a real symmetric, positive definite $p \times p$ matrix and $0 \leq \pi_j \leq 1$ is the weight of the j-th component, $j = 1, 2, \ldots, k$, with $\sum_{j=1}^{k} \pi_j = 1$. Our objective is to estimate the parameter $\boldsymbol{\theta}$ robustly and to detect the true clusters. Instead of the ordinary likelihood method, we propose a generalized likelihood approach which is motivated by the minimum DPD methodology, and subsumes the ordinary maximum likelihood approach.

2.2.1 Theoretical Formulation

As discussed in Chapter 1 (Section 1.9), the MDPDE of the unknown parameter $\boldsymbol{\theta}$ can be obtained by minimizing

$$\int f_{\boldsymbol{\theta}}^{1+\beta}(\boldsymbol{x})\, d\boldsymbol{x} - \left(1 + \frac{1}{\beta}\right) \frac{1}{n} \sum_{i=1}^{n} f_{\boldsymbol{\theta}}^{\beta}(\boldsymbol{X}_i)$$

with respect the $\boldsymbol{\theta}$. Let us observe that the aforesaid minimization problem can also be viewed as the maximization of

$$l_\beta(\boldsymbol{\theta}) = \left(1 + \frac{1}{\beta}\right) \frac{1}{n} \sum_{i=1}^{n} f_{\boldsymbol{\theta}}^{\beta}(\boldsymbol{X}_i) - \int f_{\boldsymbol{\theta}}^{1+\beta}(\boldsymbol{x})\, d\boldsymbol{x} \tag{2.1}$$

with respect the $\boldsymbol{\theta}$. For a general $\beta > 0$, the quantity in the right hand side of Equation (2.1) has been referred to as the β-likelihood in Fujisawa and Eguchi (2006) [54] and its maximizer is termed as the maximum β-likelihood estimator. Other robust generalizations of the ordinary likelihood function (and inference based on these generalized likelihoods) can be found in the works of Cichocki and Amari (2010) [27] (α, β and γ divergences) and Ferrari and Vecchia (2012) [45] (q-entropy and its relationship with the DPD). The same quantity (in the right hand side of Equation (2.1)) is indeed also equal to the L_q-likelihood function with $q = 1-\beta$; its use in robust statistical inferences has been explored by [44, 46, 47, 64, 120, 157].

Returning to the original problem of normal mixture models, the joint likelihood is given by

$$L_M(\boldsymbol{\theta}, F_n) = \prod_{i=1}^{n} f_{\boldsymbol{\theta}}(\boldsymbol{X}_i), \qquad \text{with} \quad f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{j=1}^{k} \pi_j \phi_p(\boldsymbol{x}, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j), \tag{2.2}$$

where F_n is the empirical probability measure of the data. The classical MLE is defined as the maximizer of $L_M(\boldsymbol{\theta}, F_n)$ with respect to $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. The corresponding β-likelihood can again be defined by Equation (2.1), but now $f_{\boldsymbol{\theta}}$ is as given in (2.2). This may be maximized with respect to $\boldsymbol{\theta}$ to obtain the maximum β-likelihood estimator (or the MDPDE) of $\boldsymbol{\theta}$ for a given value of β. However, due to the presence of a summation term in $f_{\boldsymbol{\theta}}$ and the integral of its power in the objective function, the associated optimization problem becomes extremely difficult. Fujisawa and Eguchi (2006) [54] have proposed an algorithm for this particular optimization problem to obtain the MDPDEs for univariate ($p = 1$) normal mixture models. But the algorithm is very difficult to implement in higher dimensions and hence the computation of the maximum β-likelihood estimator in a normal mixture model with $p > 1$ still remains a challenging problem.

We hereby propose an alternative EM like algorithm for robust parameter estimation in case of normal mixture models without directly maximizing the β-likelihood (equivalently minimizing the DPD) as done by Fujisawa and Eguchi (2006) [54]. In particular, we consider an alternative version of the likelihood using the β-likelihood of the individual component densities, as described below, rather than considering the β-likelihood for the overall mixture density $f_{\boldsymbol{\theta}}$ as in Fujisawa and Eguchi (2006) [54]. Our approach leads to a valid objective function which has a much simpler form that is fairly straightforward to maximize through EM type iterative algorithms even for

higher dimensions ($p > 1$). As a result, our algorithm also leads to clustering and outlier detection and is structurally similar to the TCLUST algorithm (García-Escudero et al. (2008) [58]) but the source of robustness is different. Instead of performing a likelihood based trimming, we invoke the β-likelihood from the minimum DPD approach. The motivation comes from the fact that outliers (if present) may also provide useful information about the system; so they should be further scrutinized rather than be eliminated by trimming.

In order to describe our proposed algorithm, let us note that even the likelihood function of the normal mixture model given in (2.2) is difficult to maximize directly with respect to $\boldsymbol{\theta}$ and a different expression for the likelihood function is used for the computation of MLE via EM algorithms. Consider the missing assignment functions

$$Z_j(\boldsymbol{X}_i, \boldsymbol{\theta}) = \begin{cases} 1, & \text{if } \boldsymbol{X}_i \in C_j \\ 0, & \text{otherwise} \end{cases}$$

with C_j as the j-th cluster, $j = 1, 2, \ldots, k$. If these assignment functions are known, the likelihood function can also be presented as,

$$L(\boldsymbol{\theta}, F_n) = \prod_{j=1}^{k} \prod_{i \in C_j} \pi_j \phi_p(\boldsymbol{X}_i, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) = \prod_{i=1}^{n} \prod_{j=1}^{k} \pi_j^{Z_j(\boldsymbol{X}_i, \boldsymbol{\theta})} \phi_p(\boldsymbol{X}_i, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)^{Z_j(\boldsymbol{X}_i, \boldsymbol{\theta})}.$$

It is mathematically equivalent and convenient to maximize,

$$\begin{aligned} l(\boldsymbol{\theta}, F_n) = & \log L(\boldsymbol{\theta}, F_n) \\ = & \sum_{i=1}^{n} \sum_{j=1}^{k} Z_j(\boldsymbol{X}_i, \boldsymbol{\theta}) [\log \ \pi_j + \log \ \phi_p(\boldsymbol{X}_i, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)] \\ = & \sum_{j=1}^{k} [n_j \log \ \pi_j + \sum_{i \in C_j} \log \ \phi_p(\boldsymbol{X}_i, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)], \end{aligned}$$

where $n_j = \sum_{i=1}^{n} Z_j(\boldsymbol{X}_i, \boldsymbol{\theta})$ represents the number of observation in the j-th cluster for $j = 1, \ldots, k$.

Now our goal is to use the β-likelihood instead of the ordinary log-likelihood for the estimation of the parameter $\boldsymbol{\theta}$. Hence, for $j = 1, \ldots, k$, we replace separately the individual term $\sum_{i \in C_j} \log \ \phi_p(\boldsymbol{X}_i, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ by $\frac{n_j}{1+\beta} l_\beta^{(j)}(\boldsymbol{\theta})$, an appropriate constant

multiple of the β-likelihood function $l_\beta^{(j)}(\boldsymbol{\theta})$ of the j-th component density given by,

$$l_\beta^{(j)}(\boldsymbol{\theta}) = \left(1+\frac{1}{\beta}\right)\frac{1}{n_j}\sum_{i\in C_j}\phi_p^\beta(\boldsymbol{X}_i,\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j) - \int \phi_p^{1+\beta}(\boldsymbol{x},\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)\,d\boldsymbol{x}. \tag{2.3}$$

Our β modified objective function thus becomes, $\sum_{j=1}^k \left[n_j \log\ \pi_j + \frac{n_j}{1+\beta} l_\beta^{(j)}(\boldsymbol{\theta})\right]$, which after some algebra simplifies to

$$\begin{aligned} nE_{F_n}\Bigg[\sum_{j=1}^k Z_j(\boldsymbol{X},\boldsymbol{\theta})\log\ \pi_j + \frac{1}{\beta}\sum_{j=1}^k Z_j(\boldsymbol{X},\boldsymbol{\theta})\phi_p^\beta(\boldsymbol{X},\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j) \\ -\tfrac{1}{1+\beta}\sum_{j=1}^k Z_j(\boldsymbol{X},\boldsymbol{\theta})\int \phi_p^{1+\beta}(\boldsymbol{x},\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)\,d\boldsymbol{x}\Bigg]. \end{aligned} \tag{2.4}$$

Hence, it is enough to maximize,

$$L_\beta(\boldsymbol{\theta},F_n) = E_{F_n}\left[\sum_{j=1}^k Z_j(\boldsymbol{X},\boldsymbol{\theta})\left[\log\ \pi_j + \frac{1}{\beta}\phi_p^\beta(\boldsymbol{X},\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j) - \frac{1}{1+\beta}\int \phi_p^{1+\beta}(\boldsymbol{x},\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)\,d\boldsymbol{x}\right]\right]. \tag{2.5}$$

Equation (2.5) represents the empirical objective function. Assuming F to be the true probability measure of $\boldsymbol{X}_1$, the corresponding theoretical objective function is given by,

$$L_\beta(\boldsymbol{\theta},F) = E_{F}\left[\sum_{j=1}^k Z_j(\boldsymbol{X},\boldsymbol{\theta})\left[\log\ \pi_j + \frac{1}{\beta}\phi_p^\beta(\boldsymbol{X},\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j) - \frac{1}{1+\beta}\int \phi_p^{1+\beta}(\boldsymbol{x},\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)\,d\boldsymbol{x}\right]\right]. \tag{2.6}$$

To solve the aforesaid estimation problem, we need a specific algebraic form of $Z_j(\boldsymbol{X}_i,\boldsymbol{\theta})$. That is, we need a discrimination rule which can assign a particular observation $\boldsymbol{X}_i$ to a cluster C_j systematically. The most well-known discrimination rule is based on the likelihood method, originally proposed by R.A. Fisher, and was used in the TCLUST method by García-Escudero et al. (2008) [58]. We are also going to use the likelihood based discrimination rule which is defined below:

Discriminant Function: Given $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, we define the discriminant functions

$$D_j(\boldsymbol{X},\boldsymbol{\theta}) = \pi_j\phi_p(\boldsymbol{X},\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j) \text{ and } D(\boldsymbol{X},\boldsymbol{\theta}) = \max_{1\le j\le k} D_j(\boldsymbol{X},\boldsymbol{\theta})$$

and we include a particular observation $\boldsymbol{X}_i$ to the j-th cluster C_j if $D(\boldsymbol{X}_i,\boldsymbol{\theta}) =$

$D_j(\boldsymbol{X}_i, \boldsymbol{\theta})$.

Note that although the discrimination is based on the likelihood, we compute its empirical values by substituting the robust parameter estimates obtained through the maximum pseudo β-likelihood method, which guarantees proper stability. In terms of these discriminant functions, the assignment functions can be written as $Z_j(\boldsymbol{X}_i, \boldsymbol{\theta}) = I\left[D(\boldsymbol{X}_i, \boldsymbol{\theta}) = D_j(\boldsymbol{X}_i, \boldsymbol{\theta})\right]$.

These discrimination functions will also be used for outlier detection at the end of our algorithm. A small value of the discriminant function is a good indicator of possible anomaly in respect of a particular observation in relation to the presumed cluster (as the discriminant function is proportional to the posterior likelihood of that particular observation).

We refer to the right hand side of Equation (2.5) as the empirical pseudo β-likelihood function and the right hand side of Equation (2.6) as the theoretical pseudo β-likelihood function. We define the maximizers of these empirical and theoretical pseudo β-likelihood functions, with respect to $\boldsymbol{\theta}$, as the maximum pseudo β-likelihood estimator (MPLE$_\beta$) and the maximum pseudo β-likelihood functional (MPLF$_\beta$), respectively.

It may be noted that after using the alternative version of the β-likelihood for the individual component densities, our objective function in Equation (2.5) is no longer the objective function of the actual MDPDE of the normal mixture model. We differ in this respect from the Fujisawa and Eguchi (2006) [54] approach, although the two approaches coincide for $\beta = 0$ (the case of the ordinary likelihood). The source of robustness of our procedure as well as our motivation and philosophy are, however, strictly in line with those of the MDPDEs. Accordingly we feel that the "pseudo β-likelihood" and the "maximum pseudo β-likelihood estimator" represent logical nomenclature for our method and our estimator.

However, as we have already noted in the previous section, the mixture normal likelihood is unbounded as a function of the parameters, so that, its direct maximization is not a well defined problem. The same difficulty also arises in case of the pseudo β-likelihood of the mixture normal model leading to singularities in the estimates of covariance matrices. To circumvent this problem in one dimension, Hathaway (1985) [69] proposed a constraint on the ratios of component standard deviations. Later, García-Escudero et al. (2008) [58] generalized this constraint in the multivariate set-up in terms of eigenvalues to avoid singularity of the dispersion matrix estimators. We will impose the same eigenvalue ratio constraint in our case. Let us denote λ_{jl} to be

the l-th eigenvalue of the covariance matrix $\mathbf{\Sigma}_j$ for $1 \leq j \leq k$ and $1 \leq l \leq p$, and put $M = \max\limits_{1\leq j\leq k} \max\limits_{1\leq l\leq p} \lambda_{jl}$ and $m = \min\limits_{1\leq j\leq k} \min\limits_{1\leq l\leq p} \lambda_{jl}$, the largest and smallest eigenvalues, respectively.

Eigenvalue Ratio (ER) Constraint: For a prespecified constant $c \geq 1$, the system satisfies the condition

$$\frac{M}{m} \leq c. \tag{2.7}$$

Along with the eigenvalue ratio constraint, we will make the following additional assumption to avoid singularity and establish the existence and consistency of our proposed estimators in the next chapter.

Non-singularity (NS) Constraint: We assume that the smallest eigenvalue m satisfies $m \geq c_1$ for some small positive constant c_1 which is prespecified.

Under the above two constraints, characterized by constants $C = (c, c_1)$, our search for the estimator can be confined with the restricted parameter space defined as

$$\mathbf{\Theta}_C = \left\{ \boldsymbol{\theta} : \boldsymbol{\theta} = (\pi_1, \pi_2, \ldots, \pi_k, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k, \mathbf{\Sigma}_1, \mathbf{\Sigma}_2, \ldots, \mathbf{\Sigma}_k) \text{ with } \frac{M}{m} \leq c \text{ and } m \geq c_1 \right\}. \tag{2.8}$$

For the sake of completeness, note that $c = 1$ provides the strongest possible restriction in case of the eigenvalue ratio constraint. But a large value of c is more pragmatic in the sense that the estimation problem becomes less restrictive in this case. Along with that, a small but positive value of c_1 is preferred both in terms of theoretical aspects (such as existence and consistency) as well as practical aspects (in the sense that the constraint does not become too stringent).

The non-singularity constraint is not really a stringent assumption in the presence of the eigenvalue ratio constraint. We need the non-singularity constraint (in the presence of the eigenvalue ratio constraint) only when the sequence of the smallest eigenvalue tends to 0 and the sequence of the largest eigenvalue is of same order as that of the sequence of smallest eigenvalues. This scenario is quite rare in practice especially under the positive definiteness of the dispersion matrices. This assumption is crucial to establish existence and consistency of the proposed estimators even in case of the above mentioned pathological case. We will study these theoretical properties

(i.e., existence and consistency of our estimators) in the next chapter.

2.2.2 Computational Algorithm for the MPLE$_\beta$

To estimate the unknown parameters, to form the clusters and to detect the outliers present in the dataset, we need to optimize the empirical objective function on the right hand side of Equation (2.5). We hereby propose an approximate EM like algorithm which solves this empirical problem and provides reasonable estimates of the unknown parameters. We refer to this algorithm as the MPLE$_\beta$ algorithm. Before describing this algorithm, we need to derive another iterative procedure to find the minimum DPD estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ based on a random sample $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n\}$ (note that this sample is not the sample mentioned at the beginning of Section 2.2 which was modelled by a multivariate normal mixture distribution) which is modelled with a $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ model. To find the aforesaid estimators, we need to minimize the objective function

$$\frac{1}{1+\beta}\int \phi_p^{1+\beta}(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\boldsymbol{x} - \frac{1}{n\beta}\sum_{i=1}^{n}\phi_p^{\beta}(\boldsymbol{X}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

(which is a constant multiple $\left(\frac{1}{1+\beta}\right)$ of the original DPD objective function) with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ (which is same as the maximization of β-likelihood with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$). This minimization can be done by simultaneously solving the equations:

$$\frac{1}{n}\sum_{i=1}^{n} e^{-\frac{\beta}{2}(\boldsymbol{X}_i-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_i-\boldsymbol{\mu})}(\boldsymbol{X}_i - \boldsymbol{\mu}) = \boldsymbol{0}, \tag{2.9}$$

$$\frac{1}{n}\sum_{i=1}^{n} e^{-\frac{\beta}{2}(\boldsymbol{X}_i-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_i-\boldsymbol{\mu})}\left(\boldsymbol{\Sigma} - (\boldsymbol{X}_i - \boldsymbol{\mu})(\boldsymbol{X}_i - \boldsymbol{\mu})'\right) = \frac{\beta}{(1+\beta)^{\frac{p}{2}+1}}\boldsymbol{\Sigma}. \tag{2.10}$$

This is formally stated and described in Theorem 2.1 and the mathematical derivations of the aforesaid system of equations are presented in Section 2.7.2. Here we propose an iteratively reweighted least squares algorithm to solve the aforesaid system of equations as follows.

Algorithm 2.1 (IRLS)

1. **Starting Value:** A non-robust starting value may affect a robust algorithm

severely. Thus, we will use robust starting values for both the mean and the dispersion matrix. For the mean vector, we use the componentwise sample medians, that is, $\hat{\boldsymbol{\mu}}^0 = (u_1, u_2, \ldots, u_p)$ where $u_j = \text{median}\{X_{1j}, X_{2j}, \ldots, X_{nj}\}$ for $1 \leq j \leq p$. For the dispersion matrix we have chosen the starting value as $\hat{\boldsymbol{\Sigma}}^0$ such that,

$$\hat{\boldsymbol{\Sigma}}^0_{ij} = \begin{cases} 1.4826^2 \text{ median}\{(X_{li} - u_i)^2, 1 \leq l \leq n\}, & \text{if } i = j, \\ 1.4826^2 \text{ median}\{(X_{li} - u_i)(X_{lj} - u_j), 1 \leq l \leq n\}, & \text{otherwise.} \end{cases}$$

$\hat{\boldsymbol{\Sigma}}^0$ can be treated as the multivariate generalization of the median absolute deviation (MAD) estimator of dispersion in one dimension.

2. **Update:** Let, $\hat{\boldsymbol{\mu}}^l$ and $\hat{\boldsymbol{\Sigma}}^l$ be the estimates at the l-th step of iteration. Calculate the current weights,

$$w_i^l = e^{-\frac{\beta}{2}(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}^l)'(\hat{\boldsymbol{\Sigma}}^l)^{-1}(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}^l)}.$$

Now update,

$$\hat{\boldsymbol{\mu}}^{l+1} = \frac{\sum_{i=1}^n w_i^l \boldsymbol{X}_i}{\sum_{i=1}^n w_i^l}$$

and

$$\hat{\boldsymbol{\Sigma}}^{l+1} = \frac{\sum_{i=1}^n w_i^l (\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}^{l+1})(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}^{l+1})'}{\sum_{i=1}^n w_i^l - \frac{n\beta}{(1+\beta)^{\frac{p}{2}+1}}}.$$

3. **Stopping Rule:** Repeat step 2 for a large number of times until, $||\hat{\boldsymbol{\mu}}^l - \hat{\boldsymbol{\mu}}^{l+1}|| \leq \epsilon$ and $||\hat{\boldsymbol{\Sigma}}^l - \hat{\boldsymbol{\Sigma}}^{l+1}|| \leq \epsilon$ for some small (prespecified) $\epsilon > 0$.

Now, let us introduce the MPLE$_\beta$ algorithm for clustering, the main focus of the present chapter.

Algorithm 2.2 (MPLE$_\beta$)

1. **Initialization:** Initially, k many random observations from the dataset are chosen as initial cluster centers, identity matrices of proper dimensions as initial dispersion matrices and the vector $(\frac{1}{k}, \frac{1}{k}, \ldots, \frac{1}{k})$ as initial weights. Then the initial clusters $C_1^0, C_2^0, \ldots, C_k^0$ are constructed by the maximum likelihood principle which assigns a particular data point to the cluster which maximizes its likelihood. (See subsequent Remark 2.1 for the effects of different initialization schemes on our algorithm).

2. **Update:** Let, $C_1^l, C_2^l, \ldots, C_k^l$ be the clusters at the l-th step of the algorithm $(l = 0, 1, \ldots)$.

   (a) For each $1 \leq j \leq k$, obtain $n_j^l = |C_j^l|$, $\hat{\pi}_j^l = \frac{n_j^l}{n}$ (see Theorem 2.2 for the justification).

   (b) For each $1 \leq j \leq k$, given n_j^l, obtain $\hat{\boldsymbol{\mu}}_j^l$ and $\hat{\boldsymbol{\Sigma}}_j^l$ by maximizing the β-likelihood of the observations which are currently assigned to the j-th cluster C_j^l. Specifically,

   $$(\hat{\boldsymbol{\mu}}_j^l, \hat{\boldsymbol{\Sigma}}_j^l) = \underset{\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j}{\operatorname{argmax}} \left[\frac{1}{n_j^l \beta} \sum_{i \in C_j^l} \phi_p^\beta(\boldsymbol{X}_i, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) - \frac{1}{1+\beta} \int \phi_p^{1+\beta}(\boldsymbol{x}, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\, d\boldsymbol{x} \right]. \tag{2.11}$$

   (by the above mentioned IRLS Algorithm 2.1.)

   (c) If the full set of eigenvalues $\boldsymbol{\Lambda} = (\boldsymbol{\Lambda}_1, \ldots, \boldsymbol{\Lambda}_k)$ of the component dispersion matrix estimates does not satisfy either of the ER or NS constraints, we replace $\boldsymbol{\Lambda}$ by another vector $\tilde{\boldsymbol{\Lambda}}$ which minimizes $||\tilde{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}||^2$ subject to the aforesaid constraints. That is, we need to obtain another vector "closest" to the existing set of eigenvalues which satisfies both the constraints. The ER constraint can be mathematically rephrased as $\lambda_{jl} - c\lambda_{su} \leq 0$ for all $(j, l) \neq (s, u)$. The NS constraint can mathematically be rephrased as $\lambda_{jl} \geq c_1$ for all (j, l). Both of these constraints are linear constraints. Dykstra's algorithm (Dykstra (1983) [42]) can be used to solve the aforesaid constrained minimization problem.

(d) The estimates of the current step are then given by

$$\hat{\boldsymbol{\theta}}^{l} = (\hat{\pi}_1^l, \hat{\pi}_2^l, \ldots, \hat{\pi}_k^l, \hat{\boldsymbol{\mu}}_1^l, \hat{\boldsymbol{\mu}}_2^l, \ldots, \hat{\boldsymbol{\mu}}_k^l, \hat{\boldsymbol{\Sigma}}_1^l, \hat{\boldsymbol{\Sigma}}_2^l, \ldots, \hat{\boldsymbol{\Sigma}}_k^l). \tag{2.12}$$

(e) Construct the updated clusters $C_1^{l+1}, C_2^{l+1}, \ldots, C_k^{l+1}$ as follows. For each $1 \leq i \leq n$, assign $\boldsymbol{X}_i$ to C_j^{l+1} if $D_j^l(\boldsymbol{X}_i, \hat{\boldsymbol{\theta}}^l) = D^l(\boldsymbol{X}_i, \hat{\boldsymbol{\theta}}^l)$, where

$$D_j^l(\boldsymbol{X}_i, \hat{\boldsymbol{\theta}}^l) = \hat{\pi}_j^l \phi_p(\boldsymbol{X}_i, \hat{\boldsymbol{\mu}}_j^l, \hat{\boldsymbol{\Sigma}}_j^l) \quad \text{and} \quad D^l(\boldsymbol{X}_i, \hat{\boldsymbol{\theta}}^l) = \max_{1 \leq j \leq k} D_j^l(\boldsymbol{X}_i, \hat{\boldsymbol{\theta}}^l).$$

3. **Stopping Rule:** Repeat step 2 for a large (preassigned) number of times or until the cluster configurations become stable.

4. **Outlier Detection:** After the process terminates and the final clusters $C_1, C_2, \ldots, C_k$ and their configurations are available, let the final parameter estimate be $\hat{\boldsymbol{\theta}} = (\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_k, \hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \ldots, \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_1, \hat{\boldsymbol{\Sigma}}_2, \ldots, \hat{\boldsymbol{\Sigma}}_k)$. Now, for each $1 \leq i \leq n$, if $\boldsymbol{X}_i$ is assigned to C_j for some $1 \leq j \leq k$, calculate $D_j(\boldsymbol{X}_i, \hat{\boldsymbol{\theta}}) = \hat{\pi}_j \phi_p(\boldsymbol{X}_i, \hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j)$. If $D_j(\boldsymbol{X}_i, \hat{\boldsymbol{\theta}}) \leq T$ for some small positive prespecified constant T, classify $\boldsymbol{X}_i$ as an outlier.

Remark 2.1. *To initialize the aforesaid algorithm, we have applied a random initialization scheme as stated in the initialization step of the algorithm. But randomly selected data points can produce very good as well as very bad estimators after completing the iterations. Hence, the algorithm should be repeated several times using different choices of initialization and then provide the solution which leads to the maximum value of the objective function. This initialization scheme has also been proposed in García-Escudero et al.* (2008) *[58], although an improved version has been proposed in Fritz et al.* (2013) *[53]. Non-robust initial choices produce spurious maxima and the misclassification rates along with the bias and mean squared errors of the parameter estimates increase drastically. This problem also arises in case of other robust clustering algorithms like TCLUST and trimmed K-means.*

Remark 2.2. *For positive β, our method smoothly discounts the ill effects of anomalous observations without physically deleting them. On the other hand, TCLUST forcefully trims those anomalous observations along with others that may not be anomalous.*

2.2.3 Selection of Tuning Parameters

To implement the algorithm, we have to choose the tuning parameters β, c, c_1 and T with appropriate justification. These choices are not straightforward in general; some comments are provided in the following discussions.

As we have seen in Chapter 1, the tuning parameter β in the DPD balances robustness and asymptotic efficiency of the resulting MDPDE; a small value (close to 0) of β is appropriate to achieve higher asymptotic efficiency whereas a large value is appropriate for higher stability. Thus, as β approaches 0, our algorithm becomes non-robust. To achieve robustness as well as high asymptotic efficiency, we use a small but positive value of β (in the interval $(0, 0.5]$). Sophisticated theoretical techniques for choosing an optimal value of β have been described by Warwick and Jones (2005) [155] and Basak et al. (2021) [9]. What these methods essentially do is that they create an empirical estimate of the true mean square error as a function of the tuning parameter (and a suitable pilot estimator) which can then be appropriately minimized over the tuning parameter to obtain an "optimal" estimate of the unknown tuning parameter. But, in our algorithm, component parameters need to be estimated separately for each cluster, in each iteration. Each of these cases would, ideally, require different optimal choices of the parameter β and thus a single optimal choice of β is not reasonable to derive for the full clustering problem. Further mathematical details of these procedures can be found in Section 2.7.8.

On the other hand, a large value of the tuning parameter c makes the optimization problem almost unrestricted. For some reasonable choices of c, we refer again to García-Escudero et al. (2008) [58] and Farcomeni and Punzo (2020) [43]. It is natural to choose the value of c_1 to be close to 0 as the value of this constant is needed to be positive to serve certain theoretical purposes but a small value of the same is less restrictive. To find an approximate optimal choice of the tuning parameter T, a "maximal-gap" approach may be useful. This approach declares those observations as outliers whose estimated discriminant values (the realized value of the discriminant function: $\hat{\pi}_j \phi_p(\cdot, \hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j)$ if the observation belongs to the j-th cluster) drop below the threshold value of T. So, we first rearrange these discriminant values of all the n observations in an increasing order. This rearrangement should bring the discriminant values corresponding to the outliers (if any) at the first few positions of the sequence. Since the distant outliers are expected to have very low discriminant values in comparison with those of the regular observations, the sorted vector of discriminant values should

contain a (possibly big) jump between the region containing the outliers and the region containing the regular observations. One can then choose T to correspond to this "jump" region; a pictorial illustration can be found in Section 2.7.9. This maximal-gap idea can also be implemented with the TCLUST methodology (and possibly to some other robust clustering methodologies) to improve outlier detection. It could be the subject of a future research to investigate how this type of refinement might improve the performances of these methods. However, although the maximal-gap strategy has worked satisfactorily in practically all of our numerical studies, we do not expect that choosing T by this philosophy will work perfectly in every possible situation. For example, when we have a not-so-remote background contamination together with an additional extremely remote outlying observation, only the remotest data point would be discarded by the maximal-gap strategy, without discarding the not-so-remote background noise. There might be a way to generalize the idea of maximal (largest) gap by choosing the value of T around the position of the t-th largest gap ($t = 2$ will serve the purpose for this example) depending on the situation. But choosing the optimal value of t in this refinement may be a very difficult problem in itself. On the whole, the choice of the tuning parameter T is, at the least, a complicated proposition and substantial future research will be needed for a completely satisfactory solution.

Finally note that, although we have performed robust estimation in the multivariate normal mixture model, the primary focus of our proposed algorithm has been on robust clustering. We need to choose the number of clusters k appropriately. In many of the well-known clustering techniques, performance improves with increasing the value of k, but increasing the value of k indefinitely may be inappropriate. Hence an optimal choice of k is needed. Rate distortion theory gives nice insights into the problem of detecting optimal number of clusters. It applies the "jump" method which detects k by maximizing efficiency and minimizing error using information based measurements. We refer to Sugar and James (2003) [142] for details. A novel penalized likelihood based method was also proposed by Cerioli et al. (2018) [20] to optimally select the pair (k, c) (c of ER constraint).

2.3 Properties of the Proposed Algorithm

2.3.1 Theoretical Results

Theorem 2.1 (MDPDEs). *Suppose* $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n\}$ *be a random sample drawn from an unknown PDF* g *which is modelled by a family of* p*-dimensional normal distributions with mean vector* $\boldsymbol{\mu}$ *and dispersion matrix* $\boldsymbol{\Sigma}$*. Then the MDPDEs of* $\boldsymbol{\mu}$ *and* $\boldsymbol{\Sigma}$ *can be obtained by solving Equations (2.9) and (2.10).*

The detailed derivation of the aforesaid system of equations can be found in Section 2.7.2.

Our next theorem provides the mathematical justification behind the update of the estimate $\hat{\pi}_j$ in Step 2(a) of our proposed clustering algorithm.

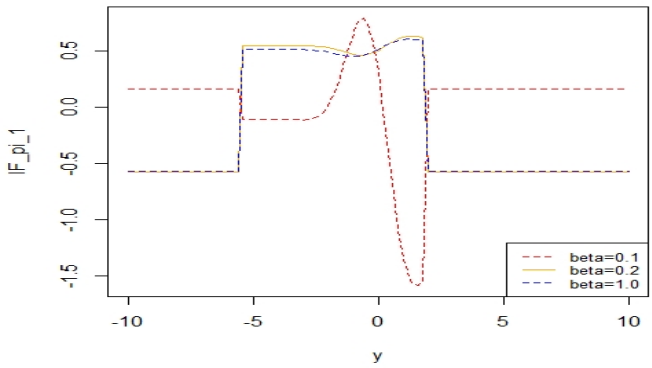
Theorem 2.2. *Given a particular cluster assignment* $C_1,\ C_2, \ldots,\ C_k$ *and the estimates* $\hat{\boldsymbol{\mu}}_j,\ \hat{\boldsymbol{\Sigma}}_j$*, the optimal value of* π_j *that maximizes (2.5) is given by,* $\hat{\pi}_j = \frac{n_j}{n}$ *where* $n_j = |C_j|$.

The proof is presented in Section 2.7.3. We have also studied further theoretical and asymptotic properties, such as (i) existence of a solution to the optimization problems in Equations (2.5) (sample version) and (2.6) (population version) for the proposed procedures and (ii) consistency of the resulting parameter estimates, as defined in Equation (2.5), which also yields the consistency of the estimated cluster centers, dispersions and proportions. A detailed discussion (along with the mathematical proofs) can be found in Chapter 3 of this dissertation (also in Chakraborty et al. (2022) [21]).

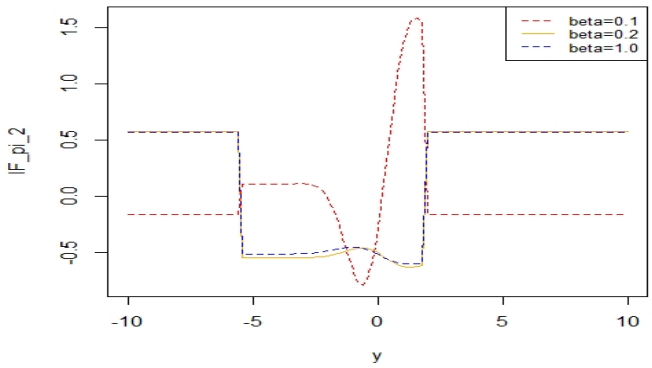
2.3.2 Robustness: Influence Function

To justify the robustness of our proposed estimators of cluster proportions, means and dispersion matrices, we will study the behaviour of their influence functions. Deriving these influence functions in higher dimensions is substantially difficult with respect to computational aspects. So, we will focus on the one dimensional case with two components ($p = 1,\ k = 2$); the implications will be in the same direction for higher dimensions. Ruwet et al. (2012) [136], the only existing literature (as per our knowledge) for studying the robustness of TCLUST, also studied this special case only. We make the following assumption in order to circumvent cumbersome calculations.

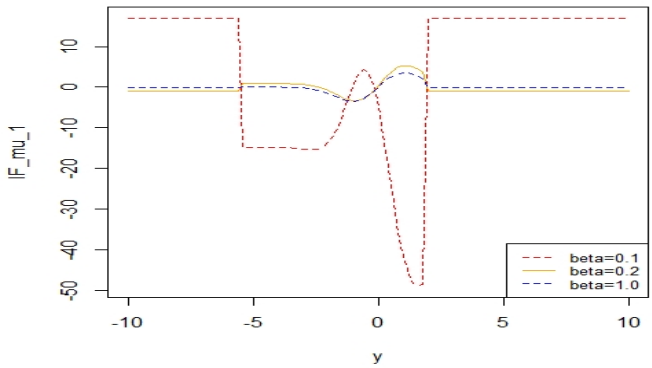
(**IF**) We assume that $\boldsymbol{\theta} \in \text{interior}(\boldsymbol{\Theta}_C)$, that is, $\frac{M}{m} < c$ and $m > c_1$.



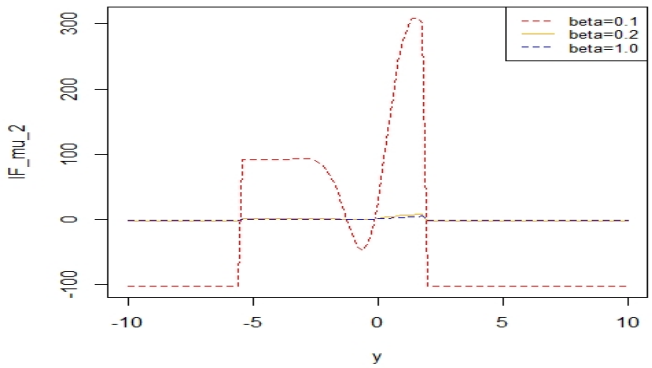
(a) $IF(\pi_1, P, y)$



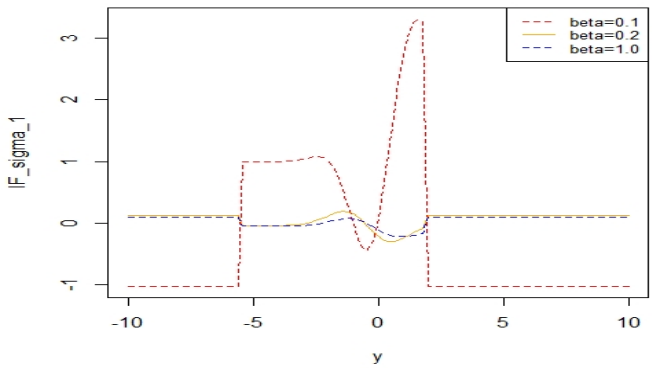
(b) $IF(\pi_2, P, y)$



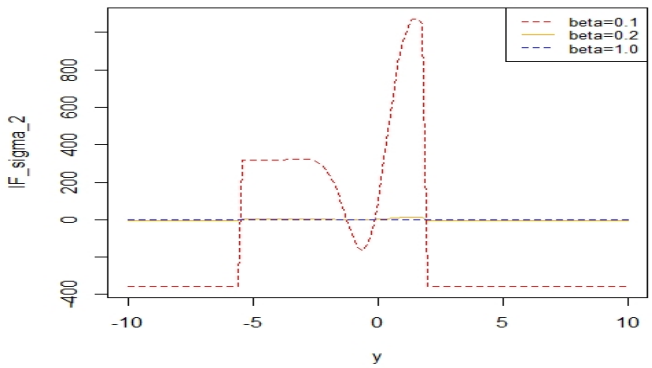
(c) $IF(\mu_1, P, y)$



(d) $IF(\mu_2, P, y)$



(e) $IF(\sigma_1^2, P, y)$



(f) $IF(\sigma_2^2, P, y)$

Figure 2.1: Influence functions of different functionals.

To derive the influence functions under our simple case ($p = 1,\ k = 2$), we first have to present our MPLF$_\beta$ (maximum pseudo β-likelihood functional) of the parameter $\boldsymbol{\theta} = (\pi_1, \pi_2, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$ as a functional of the true CDF P (corresponding PDF $p(\cdot)$), namely,

$$\boldsymbol{\theta}_\beta(P) = (\pi_1(P), \pi_2(P), \mu_1(P), \mu_2(P), \sigma_1^2(P), \sigma_2^2(P)).$$

In our simple case (i.e., $p = 1$ and $k = 2$), the true distribution is itself a 2-component, univariate normal distribution. This functional $\boldsymbol{\theta}_\beta(P)$ can be implicitly described through the following system of equations (see Section 2.7.4, for their mathematical details):

$$\begin{aligned}
&\pi_1(P) = \textstyle\int_a^b p(x)\, dx,\\
&\pi_1(P) + \pi_2(P) = 1,\\
&D_1(c, \boldsymbol{\theta}(P)) = D_2(c, \boldsymbol{\theta}(P)) \text{ for } c = a \text{ and } b,\\
&\textstyle\int_a^b f^\beta(x, \mu_1, \sigma_1^2)(x - \mu_1)p(x)\, dx = 0,\\
&\textstyle\int_{x\notin(a,b)} f^\beta(x, \mu_2, \sigma_2^2)(x - \mu_2)p(x)\, dx = 0,\\
&\textstyle\int_a^b f^\beta(x, \mu_1, \sigma_1^2)\left(\frac{(x-\mu_1)^2}{2\sigma_1^2} - 1\right)p(x)\, dx + \frac{\beta(P(b)-P(a))}{2(2\pi)^{\frac{\beta}{2}}(\sigma_1^2)^{1+\frac{\beta}{2}}(1+\beta)^{\frac{3}{2}}} = 0,\\
&\textstyle\int_{x\notin(a,b)} f^\beta(x, \mu_2, \sigma_2^2)\left(\frac{(x-\mu_2)^2}{2\sigma_2^2} - 1\right)p(x)\, dx + \frac{\beta(1-P(b)+P(a))}{2(2\pi)^{\frac{\beta}{2}}(\sigma_2^2)^{1+\frac{\beta}{2}}(1+\beta)^{\frac{3}{2}}} = 0,
\end{aligned} \tag{2.13}$$

where $f(\cdot, \mu, \sigma^2)$ is the PDF of univariate normal distribution with mean μ and variance σ^2.

The aforesaid system of equations will lead us to the influence functions of the necessary functionals through simple differentiation.

Let $\boldsymbol{IF}(\boldsymbol{\theta}_\beta, P, y) = (IF(\pi_1, P, y), IF(\pi_2, P, y), IF(a, P, y), IF(b, P, y), IF(\mu_1, P, y), IF(\mu_2, P, y), IF(\sigma_1^2, P, y), IF(\sigma_2^2, P, y))'$ be the vector of influence functions of the aforesaid functionals in $\boldsymbol{\theta}_\beta$. Now considering the contaminated version of the System (2.13) (replacing P with $P_\epsilon = (1 - \epsilon)P + \epsilon\wedge_y$) and differentiating that with respect to ϵ at $\epsilon = 0$, we obtain the following system of linear equations:

$$\boldsymbol{A}_\beta(\boldsymbol{\theta}_0, a_0, b_0)\boldsymbol{IF}(\boldsymbol{\theta}_\beta, P, y) = \boldsymbol{B}_\beta(y, \boldsymbol{\theta}_0, a_0, b_0), \tag{2.14}$$

where $\boldsymbol{\theta}_0, a_0, b_0$ are the values of $\boldsymbol{\theta}$, a and b, respectively, at the true model, $\boldsymbol{A}_\beta(\boldsymbol{\theta}_0, a_0, b_0)$ is a 8×8 coefficient matrix whose entries are independent of the contamination point y and $\boldsymbol{B}_\beta(y, \boldsymbol{\theta}_0, a_0, b_0)$ is an element in $\mathbb{R}^8$ which depends on y only through

$I(y \in (a_0, b_0))$, $(y-\mu_{j0})^l exp(-\frac{\beta(y-\mu_{j0})^2}{2\sigma_{j0}^2})$ for $j = 1, 2$, $l = 1, 2$ and $p(y)$, the true density function corresponding to P. Detailed expressions for A_β and B_β are given in Section 2.7.5 along with the derivation of Equation (2.14).

The functions $I(y \in (a_0, b_0))$, $(y-\mu_{j0})^l e^{-\frac{\beta(y-\mu_{j0})^2}{2\sigma_{j0}^2}}$ for $j = 1, 2$ and $l = 1, 2$ are bounded while the PDF p by itself is not in general. This observation leads to the boundedness of the influence functions so that our estimators (and hence the clustering) are robust against outliers.

Theorem 2.3. *The influence function vector* $\boldsymbol{IF}(P, y)$ *exists if the coefficient matrix* $\boldsymbol{A}_\beta(\boldsymbol{\theta}_0, a_0, b_0)$ *is invertible. Moreover, in case* $\beta > 0$*, it is componentwise bounded as a function of* y*, if the true PDF* p *is bounded.*

The proof is trivial from the above discussion and the form of B_β given in Section 2.7.5. Note that, the density p is always bounded for a non-singular normal mixture model, which is our case.

Now, let us study the behaviour of the influence functions graphically in a special case. Let us take the true distribution P as a mixture of $N(0, 1)$ and $N(5, 4)$ with mixing proportions $\pi_1 = \pi_2 = 0.5$. We have taken $\beta = 0.1,\ 0.2$ and 1 and the true values of the boundaries a and b are found to be -5.5 and 1.95 respectively. We have taken c and c_1 to be 5 and 0.1 so that the restrictions $\frac{M}{m} = 4 < 5 = c$ and $m = 1 > 0.1 = c_1$ are satisfied. The influence functions of the functionals are plotted in Figure 2.1.

The boundedness of the curves in Figure 2.1 indicates the stability and the robustness of our estimators. Additionally, the respective ranges of each of the influence functions shrink drastically as β increases. It may also be noted that the influence functions are practically identical for $\beta = 0.2$ and $\beta = 1$. This observation indicates that very strong levels of stability have been already attained, at least in this scenario, for very small values of β. In case of $\beta = 0$, the influence functions are unbounded; in fact at $\beta = 0$ the curve increases so fast, that a proper depiction of this case together with the positive β cases in the same frame is not informative at all.

2.4 Simulation Studies

2.4.1 Simulation Set-up

We now present some simulation experiments for investigating the finite sample performance of our algorithm in terms of the obtained estimators and the subsequent clustering, and compare it with the TCLUST, trimmed K-means (TKMEANS), K-medoids (with data matrix and the Manhattan distance, abbreviated as KMEDOIDS) (Kaufman and Rousseeuw (1987) [84]) and the MCLUST (Scrucca et al. (2016) [137]) algorithms, some of the well-known and/or state-of-the-art robust clustering methods. To carry out the simulation study, we have generated samples of size $n = 1000$ from 3-component (i.e., $k = 3$) and p-dimensional normal mixtures with component means $\boldsymbol{\mu}_1 = (0, 0, \ldots, 0)^t$, $\boldsymbol{\mu}_2 = (5, 5, \ldots, 5)^t$ and $\boldsymbol{\mu}_3 = (-5, -5, \ldots, -5)^t$ and identical covariance matrices $\boldsymbol{\Sigma}$. Different choices of p and $\boldsymbol{\Sigma}$ are taken to cover a reasonable range of data shapes. To study the robustness and efficiency of our algorithm, both pure (contamination free) and contaminated datasets are used. Three types of data contamination are used for this purpose: (i) uniform noise contamination from the p-dimensional cuboid $[-10, 10]^p$, where only those data points whose squared Mahalanobis distances from any of the cluster centers are more than the 97.5-th percentile of the $\chi^2(p)$ distribution are chosen; this type of contamination will be referred to as "uniform (chi-squared method) contamination"; (ii) uniform noise contamination from the p-dimensional annulus (centered at the origin with the inner and the outer radii 15 and 20, respectively), and (iii) outlying cluster contamination with the outlying cluster center at $(20, 20, \ldots, 20)^t$ and identity dispersion matrices with 10% (approximately) contamination in each case. A similar motivational example is provided in Section 2.7.1. For the pure datasets, the cluster assignment probabilities are taken as 0.33, 0.33 and 0.34 and in case of contaminated datasets, the cluster assignment probabilities are 0.3 for each of them and the rest of the observations (approximately 10% of the sample) are outlying observations. As accuracy measures, we have focused on the estimated misclassification rates and the proportions of regular observations misclassified as outliers in case of the pure datasets; smaller values of these measures indicate greater accuracy. In each type of contaminated datasets the estimated misclassification rates of the regular observations (which are not outliers) and the estimated proportion of undetected outliers are the accuracy measures considered; again smaller values of these measures indicate greater accuracy. Estimated bias and mean squared errors of the estimated cluster means are also presented in Section 2.7.7. Datasets of

five different dimensions, namely, $p = 2,\ 4,\ 6,\ 8$ and 10, have been generated with three choices of the common dispersion matrix $\boldsymbol{\Sigma}$, namely, $\boldsymbol{I}_p,\ 3\boldsymbol{I}_p$ and $5\boldsymbol{I}_p$ ($\boldsymbol{I}_p$ is the $p \times p$ identity matrix). Together with the above, another simulation set-up with differentially dispersed clusters has also been considered where the component means are again $\boldsymbol{\mu}_1 = (0, 0, \ldots, 0)^t$, $\boldsymbol{\mu}_2 = (5, 5, \ldots, 5)^t$ and $\boldsymbol{\mu}_3 = (-5, -5, \ldots, -5)^t$ with data dimensions $p = 2,\ 6$. But the component dispersion matrices are no longer identical and are taken to be $\boldsymbol{I}_p,\ 3\boldsymbol{I}_p$ and equicorrelation matrix of order p with common correlation $\rho = 0.5$, respectively, and the cluster assignment probabilities are taken as 0.30, 0.35, 0.35 for pure datasets. For the contaminated datasets, the cluster assignment probabilities are 0.25, 0.30, 0.35; the remaining 10% observations are outliers generated by either of the three respective contamination schemes as described in (i), (ii) and (iii). Tables 2.1, 2.2, 2.3, 2.4 and 2.5 exhibit the estimated (mean) misclassification rates in case of pure datasets, uniformly (chi-squared method) contaminated datasets, uniformly (from annulus) contaminated datasets, outlying cluster contaminated datasets, and datasets with differentially dispersed clusters, respectively, based on 100 replications.

Further, different tuning parameters are chosen as follows in all our simulation experiments. In particular, we have taken $c = 5$, $c_1 = 0.1$ and $T = 10^{-3},\ 10^{-5},\ 10^{-8},\ 10^{-18}$ and 10^{-24} for $p = 2,\ 4,\ 6,\ 8$ and 10, respectively, in our proposed algorithm (as discussed in Section 2.2.3). Also for the TCLUST method, we have used $c = 5$.

The trimming proportion α in TCLUST and trimmed K-means methods are taken as 0.0 and 0.05 for pure datasets and 0.10 and 0.15 for contaminated cases. The MCLUST method is used with a fixed number of clusters ($G = 3$) and uniform noise component (only in case of contaminated datasets) to make it resistant against outliers according as its R implementation. Several values of β are taken in the range $[0, 0.5]$ for our method; values of β larger than 0.5 have been avoided in order to limit the loss in model efficiency. The R packages tclust ([52]), trimcluster ([71]), cluster ([103]) and mclust ([137]) are used to carry out the simulations for the TCLUST, trimmed K-means, K-medoids and MCLUST algorithms, respectively.

2.4.2 Discussion of Simulation Results

The simulations that have been performed here are quite extensive, and it is necessary to clearly pinpoint what the salient features of these numbers are. These features are described in the following.

		MPLE$_\beta$				TCLUST		TKMEANS		KMEDOIDS	MCLUST
p	$\boldsymbol{\Sigma}$	$\beta = 0$	$\beta = 0.1$	$\beta = 0.3$	$\beta = 0.5$	$\alpha = 0.0$	$\alpha = 0.05$	$\alpha = 0.0$	$\alpha = 0.05$		
2	$\boldsymbol{I}_2$	0.0003	0.0003	0.0004	0.0004	0.0003	0.050	0.0012	0.051	0.0003	0.0004
		(0.0001)	(0.0001)	(0.0002)	(0.0002)	(0.000)	(0.050)	(0.001)	(0.050)	(0.000)	(0.000)
	$3\boldsymbol{I}_2$	0.029	0.029	0.029	0.029	0.030	0.076	0.028	0.074	0.029	0.029
		(0.0002)	(0.0002)	(0.0003)	(0.0004)	(0.000)	(0.050)	(0.001)	(0.050)	(0.000)	(0.000)
	$5\boldsymbol{I}_2$	0.082	0.082	0.082	0.082	0.093	0.129	0.077	0.122	0.079	0.078
		(0.001)	(0.001)	(0.001)	(0.001)	(0.000)	(0.050)	(0.001)	(0.050)	(0.000)	(0.000)
4	$\boldsymbol{I}_4$	0.000	0.000	0.0001	0.0002	0.000	0.050	0.001	0.050	0.000	0.0004
		(0.001)	(0.001)	(0.0001)	(0.0002)	(0.000)	(0.050)	(0.0001)	(0.050)	(0.000)	(0.000)
	$3\boldsymbol{I}_4$	0.003	0.004	0.004	0.004	0.003	0.051	0.0034	0.051	0.0034	0.003
		(0.001)	(0.001)	(0.001)	(0.002)	(0.000)	(0.050)	(0.001)	(0.050)	(0.000)	(0.000)
	$5\boldsymbol{I}_4$	0.020	0.019	0.021	0.022	0.019	0.066	0.017	0.064	0.029	0.019
		(0.002)	(0.002)	(0.003)	(0.003)	(0.000)	(0.050)	(0.001)	(0.050)	(0.000)	(0.000)
6	$\boldsymbol{I}_6$	0.000	0.000	0.000	0.000	0.000	0.050	0.001	0.050	0.000	0.000
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.050)	(0.001)	(0.050)	(0.000)	(0.000)
	$3\boldsymbol{I}_6$	0.000	0.000	0.000	0.000	0.0004	0.051	0.001	0.050	0.000	0.0002
		(0.000)	(0.0001)	(0.0003)	(0.0004)	(0.000)	(0.050)	(0.001)	(0.050)	(0.000)	(0.000)
	$5\boldsymbol{I}_6$	0.005	0.005	0.006	0.006	0.004	0.053	0.005	0.053	0.007	0.004
		(0.000)	(0.001)	(0.001)	(0.001)	(0.000)	(0.050)	(0.001)	(0.050)	(0.000)	(0.000)
8	$\boldsymbol{I}_8$	0.000	0.000	0.000	0.000	0.000	0.050	0.001	0.050	0.000	0.000
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.050)	(0.001)	(0.050)	(0.000)	(0.000)
	$3\boldsymbol{I}_8$	0.000	0.000	0.000	0.000	0.000	0.052	0.001	0.05	0.000	0.000
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.050)	(0.001)	(0.050)	(0.000)	(0.000)
	$5\boldsymbol{I}_8$	0.001	0.001	0.001	0.001	0.001	0.051	0.002	0.051	0.002	0.002
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.050)	(0.001)	(0.050)	(0.000)	(0.000)
10	$\boldsymbol{I}_{10}$	0.000	0.000	0.000	0.000	0.000	0.050	0.001	0.050	0.000	0.000
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.050)	(0.001)	(0.050)	(0.000)	(0.000)
	$3\boldsymbol{I}_{10}$	0.000	0.000	0.000	0.000	0.000	0.050	0.001	0.050	0.000	0.000
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.050)	(0.001)	(0.050)	(0.000)	(0.000)
	$5\boldsymbol{I}_{10}$	0.0002	0.0002	0.0002	0.0003	0.0003	0.051	0.001	0.051	0.001	0.000
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.050)	(0.001)	(0.050)	(0.000)	(0.000)

Table 2.1: Estimated misclassification rates of regular observations (and proportions of regular observations misclassified as outliers within parentheses) for pure datasets.

		MPLE$_\beta$				TCLUST		TKMEANS		KMEDOIDS	MCLUST
p	$\boldsymbol{\Sigma}$	$\beta=0$	$\beta=0.1$	$\beta=0.3$	$\beta=0.5$	$\alpha=0.1$	$\alpha=0.15$	$\alpha=0.1$	$\alpha=0.15$		
2	$\boldsymbol{I}_2$	0.018	0.013	0.029	0.031	0.009	0.056	0.013	0.055	0.002	0.011
		(0.123)	(0.081)	(0.007)	(0.005)	(0.085)	(0.000)	(0.080)	(0.000)	(1.000)	(0.068)
	$3\boldsymbol{I}_2$	0.061	0.053	0.056	0.061	0.040	0.078	0.036	0.077	0.028	0.053
		(0.184)	(0.158)	(0.057)	(0.011)	(0.096)	(0.0002)	(0.096)	(0.000)	(1.000)	(0.002)
	$5\boldsymbol{I}_2$	0.135	0.127	0.129	0.134	0.108	0.133	0.085	0.127	0.084	0.117
		(0.213)	(0.198)	(0.146)	(0.060)	(0.107)	(0.001)	(0.085)	(0.000)	(1.000)	(0.000)
4	$\boldsymbol{I}_4$	0.008	0.007	0.009	0.010	0.007	0.057	0.008	0.056	0.000	0.001
		(0.060)	(0.008)	(0.006)	(0.006)	(0.041)	(0.000)	(0.036)	(0.000)	(1.000)	(0.022)
	$3\boldsymbol{I}_4$	0.009	0.007	0.009	0.010	0.010	0.058	0.009	0.056	0.003	0.008
		(0.271)	(0.150)	(0.064)	(0.060)	(0.072)	(0.000)	(0.067)	(0.000)	(1.000)	(0.063)
	$5\boldsymbol{I}_4$	0.029	0.028	0.033	0.036	0.028	0.073	0.028	0.071	0.022	0.028
		(0.286)	(0.190)	(0.072)	(0.053)	(0.091)	(0.0002)	(0.079)	(0.000)	(1.000)	(0.061)
6	$\boldsymbol{I}_6$	0.003	0.0005	0.0006	0.0008	0.006	0.055	0.005	0.056	0.000	0.0004
		(0.061)	(0.004)	(0.004)	(0.004)	(0.034)	(0.000)	(0.032)	(0.000)	(1.000)	(0.004)
	$3\boldsymbol{I}_6$	0.008	0.006	0.010	0.011	0.005	0.053	0.004	0.053	0.001	0.002
		(0.079)	(0.026)	(0.016)	(0.014)	(0.051)	(0.000)	(0.053)	(0.000)	(1.000)	(0.003)
	$5\boldsymbol{I}_6$	0.010	0.007	0.009	0.011	0.011	0.060	0.009	0.056	0.007	0.008
		(0.093)	(0.017)	(0.083)	(0.077)	(0.073)	(0.001)	(0.067)	(0.000)	(1.000)	(0.074)
8	$\boldsymbol{I}_8$	0.002	0.000	0.000	0.000	0.005	0.056	0.004	0.055	0.000	0.000
		(0.672)	(0.007)	(0.008)	(0.008)	(0.022)	(0.000)	(0.035)	(0.000)	(1.000)	(0.000)
	$3\boldsymbol{I}_8$	0.003	0.000	0.000	0.000	0.005	0.058	0.005	0.057	0.001	0.001
		(0.711)	(0.000)	(0.000)	(0.058)	(0.041)	(0.000)	(0.030)	(0.000)	(1.000)	(0.013)
	$5\boldsymbol{I}_8$	0.009	0.007	0.011	0.014	0.007	0.055	0.007	0.057	0.003	0.003
		(0.839)	(0.048)	(0.025)	(0.022)	(0.052)	(0.000)	(0.049)	(0.000)	(1.000)	(0.048)
10	$\boldsymbol{I}_{10}$	0.005	0.000	0.000	0.000	0.004	0.055	0.003	0.054	0.000	0.000
		(0.479)	(0.004)	(0.003)	(0.003)	(0.048)	(0.000)	(0.037)	(0.000)	(1.000)	(0.001)
	$3\boldsymbol{I}_{10}$	0.004	0.000	0.000	0.000	0.004	0.055	0.004	0.055	0.000	0.000
		(0.612)	(0.022)	(0.019)	(0.019)	(0.044)	(0.000)	(0.037)	(0.000)	(1.000)	(0.005)
	$5\boldsymbol{I}_{10}$	0.007	0.002	0.004	0.006	0.005	0.054	0.004	0.055	0.001	0.002
		(0.997)	(0.032)	(0.018)	(0.016)	(0.043)	(0.000)	(0.039)	(0.000)	(1.000)	(0.025)

Table 2.2: Estimated misclassification rates of regular observations (and proportion of undetected outliers within parentheses) for uniformly (chi-squared method) contaminated datasets.

In case of pure datasets, the estimated misclassification rates (averaged over the 100 simulated samples) along with the proportion of regular observations misclassified as outliers are presented in Table 2.1. All the methods are very similar in terms of the estimated misclassification rates. As expected, the proposed method performed the best in case of $\beta = 0$, the case corresponding to maximum likelihood along with the ER and NS constraints. Further, in the present proposal, the proportions of regular observations misclassified as outliers (presented in parenthesis below the misclassification rates) are very small indicating the method is doing the correct thing when the data are from pure models.

In case of uniformly (chi-squared method) contaminated datasets, the estimated

		MPLE$_\beta$				TCLUST		TKMEANS		KMEDOIDS	MCLUST
p	$\boldsymbol{\Sigma}$	$\beta=0$	$\beta=0.1$	$\beta=0.3$	$\beta=0.5$	$\alpha=0.1$	$\alpha=0.15$	$\alpha=0.1$	$\alpha=0.15$		
2	$\boldsymbol{I}_2$	0.004	0.001	0.001	0.001	0.004	0.057	0.006	0.058	0.001	0.002
		(0.497)	(0.000)	(0.000)	(0.000)	(0.029)	(0.000)	(0.025)	(0.000)	(1.000)	(0.000)
	$3\boldsymbol{I}_2$	0.038	0.040	0.029	0.029	0.053	0.079	0.032	0.079	0.029	0.032
		(0.632)	(0.280)	(0.000)	(0.000)	(0.034)	(0.000)	(0.039)	(0.000)	(1.000)	(0.000)
	$5\boldsymbol{I}_2$	0.092	0.095	0.096	0.093	0.146	0.132	0.081	0.126	0.080	0.085
		(0.698)	(0.396)	(0.038)	(0.032)	(0.039)	(0.000)	(0.041)	(0.000)	(1.000)	(0.000)
4	$\boldsymbol{I}_4$	0.001	0.0002	0.0002	0.0002	0.004	0.053	0.003	0.053	0.0003	0.0002
		(0.071)	(0.000)	(0.000)	(0.000)	(0.036)	(0.000)	(0.043)	(0.000)	(1.000)	(0.000)
	$3\boldsymbol{I}_4$	0.010	0.003	0.003	0.004	0.007	0.055	0.007	0.058	0.003	0.004
		(0.105)	(0.014)	(0.009)	(0.008)	(0.044)	(0.000)	(0.028)	(0.000)	(1.000)	(0.007)
	$5\boldsymbol{I}_4$	0.026	0.021	0.021	0.021	0.023	0.072	0.022	0.070	0.024	0.021
		(0.147)	(0.065)	(0.038)	(0.037)	(0.046)	(0.004)	(0.045)	(0.002)	(1.000)	(0.032)
6	$\boldsymbol{I}_6$	0.0001	0.000	0.000	0.000	0.006	0.054	0.005	0.060	0.002	0.000
		(0.147)	(0.000)	(0.000)	(0.000)	(0.031)	(0.000)	(0.033)	(0.000)	(1.000)	(0.001)
	$3\boldsymbol{I}_6$	0.002	0.0005	0.0008	0.0008	0.004	0.055	0.005	0.055	0.0006	0.001
		(0.239)	(0.012)	(0.009)	(0.009)	(0.038)	(0.001)	(0.038)	(0.001)	(1.000)	(0.010)
	$5\boldsymbol{I}_6$	0.007	0.004	0.005	0.005	0.010	0.059	0.010	0.060	0.008	0.006
		(0.329)	(0.046)	(0.040)	(0.039)	(0.045)	(0.004)	(0.044)	(0.004)	(1.000)	(0.032)
8	$\boldsymbol{I}_8$	0.0001	0.000	0.000	0.000	0.005	0.056	0.003	0.053	0.003	0.000
		(0.893)	(0.003)	(0.003)	(0.002)	(0.021)	(0.000)	(0.043)	(0.000)	(1.000)	(0.000)
	$3\boldsymbol{I}_8$	0.0003	0.000	0.000	0.000	0.006	0.057	0.005	0.056	0.0002	0.001
		(0.925)	(0.014)	(0.012)	(0.012)	(0.031)	(0.000)	(0.035)	(0.000)	(1.000)	(0.005)
	$5\boldsymbol{I}_8$	0.004	0.002	0.003	0.005	0.008	0.057	0.006	0.057	0.003	0.003
		(0.997)	(0.020)	(0.017)	(0.015)	(0.032)	(0.002)	(0.038)	(0.002)	(1.000)	(0.023)
10	$\boldsymbol{I}_{10}$	0.000	0.000	0.000	0.000	0.004	0.056	0.004	0.056	0.000	0.000
		(1.000)	(0.0002)	(0.0002)	(0.0002)	(0.031)	(0.000)	(0.026)	(0.000)	(1.000)	(0.000)
	$3\boldsymbol{I}_{10}$	0.000	0.000	0.000	0.000	0.005	0.057	0.005	0.057	0.000	0.001
		(1.000)	(0.025)	(0.022)	(0.024)	(0.029)	(0.000)	(0.031)	(0.000)	(1.000)	(0.002)
	$5\boldsymbol{I}_{10}$	0.0005	0.0003	0.0004	0.0008	0.005	0.055	0.004	0.054	0.001	0.001
		(1.000)	(0.060)	(0.026)	(0.025)	(0.038)	(0.002)	(0.044)	(0.001)	(1.000)	(0.015)

Table 2.3: Estimated misclassification rates of regular observations (and proportion of undetected outliers within parentheses) for uniformly (from annulus) contaminated datasets.

misclassification rates of the regular observations (which are not outliers) along with the estimated proportion of undetected outliers are presented in Table 2.2. In many cases, the proposed method with $\beta \approx 0.1$ or 0.3 have lower estimated regular misclassification rates in comparison with the trimming based methods, while being competitive in other cases. The K-medoids method generates slightly lower missclassification rates than the proposed method for several of the cases. However, the K-medoid method is not adaptable for detection of outliers and fails in this respect. The MCLUST method produces marginally better results in comparison to almost all the methods in this case.

The relative performance (presented in Table 2.3) of the MPLE$_\beta$ method in compar-

		MPLE$_\beta$				TCLUST		TKMEANS		KMEDOIDS	MCLUST
p	$\boldsymbol{\Sigma}$	$\beta=0$	$\beta=0.1$	$\beta=0.3$	$\beta=0.5$	$\alpha=0.1$	$\alpha=0.15$	$\alpha=0.1$	$\alpha=0.15$		
2	$\boldsymbol{I}_2$	0.042	0.019	0.019	0.019	0.041	0.056	0.030	0.056	0.413	0.563
		(0.058)	(0.000)	(0.000)	(0.000)	(0.099)	(0.000)	(0.080)	(0.000)	(1.000)	(0.990)
	$3\boldsymbol{I}_2$	0.108	0.107	0.068	0.067	0.213	0.088	0.106	0.080	0.486	0.471
		(0.238)	(0.209)	(0.000)	(0.000)	(0.470)	(0.009)	(0.151)	(0.000)	(1.000)	(0.999)
	$5\boldsymbol{I}_2$	0.251	0.260	0.169	0.156	0.414	0.248	0.268	0.126	0.505	0.510
		(0.354)	(0.401)	(0.068)	(0.000)	(0.992)	(0.492)	(0.405)	(0.000)	(1.000)	(0.995)
4	$\boldsymbol{I}_4$	0.218	0.008	0.009	0.010	0.033	0.055	0.062	0.055	0.376	0.541
		(0.849)	(0.000)	(0.000)	(0.000)	(0.095)	(0.000)	(0.124)	(0.000)	(1.000)	(0.999)
	$3\boldsymbol{I}_4$	0.113	0.111	0.053	0.057	0.219	0.057	0.094	0.054	0.414	0.459
		(0.930)	(0.869)	(0.000)	(0.000)	(0.560)	(0.000)	(0.204)	(0.000)	(1.000)	(0.999)
	$5\boldsymbol{I}_4$	0.218	0.222	0.134	0.135	0.368	0.320	0.153	0.071	0.464	0.490
		(0.857)	(0.900)	(0.000)	(0.000)	(0.920)	(0.480)	(0.226)	(0.000)	(1.000)	(0.999)
6	$\boldsymbol{I}_6$	0.339	0.001	0.001	0.001	0.050	0.057	0.021	0.058	0.369	0.516
		(0.990)	(0.000)	(0.000)	(0.000)	(0.154)	(0.000)	(0.037)	(0.000)	(1.000)	(0.992)
	$3\boldsymbol{I}_6$	0.309	0.027	0.011	0.013	0.165	0.056	0.034	0.054	0.441	0.518
		(1.000)	(0.720)	(0.000)	(0.000)	(0.421)	(0.000)	(0.058)	(0.000)	(1.000)	(1.000)
	$5\boldsymbol{I}_6$	0.290	0.076	0.035	0.036	0.311	0.192	0.109	0.058	0.450	0.533
		(1.000)	(1.000)	(0.020)	(0.000)	(0.800)	(0.360)	(0.173)	(0.000)	(1.000)	(1.000)
8	$\boldsymbol{I}_8$	0.337	0.000	0.000	0.000	0.070	0.053	0.047	0.053	0.382	0.479
		(1.000)	(0.000)	(0.000)	(0.000)	(0.242)	(0.000)	(0.123)	(0.000)	(1.000)	(0.999)
	$3\boldsymbol{I}_8$	0.344	0.004	0.001	0.000	0.116	0.054	0.107	0.053	0.408	0.499
		(1.000)	(0.400)	(0.000)	(0.000)	(0.301)	(0.000)	(0.175)	(0.000)	(1.000)	(1.000)
	$5\boldsymbol{I}_8$	0.357	0.012	0.002	0.002	0.293	0.100	0.096	0.057	0.425	0.471
		(1.000)	(0.980)	(0.000)	(0.000)	(0.760)	(0.120)	(0.151)	(0.000)	(1.000)	(1.000)
10	$\boldsymbol{I}_{10}$	0.335	0.000	0.000	0.000	0.035	0.056	0.064	0.055	0.356	0.495
		(1.000)	(0.000)	(0.000)	(0.000)	(0.119)	(0.000)	(0.117)	(0.000)	(1.000)	(0.999)
	$3\boldsymbol{I}_{10}$	0.344	0.004	0.000	0.000	0.145	0.053	0.048	0.056	0.395	0.535
		(1.000)	(0.520)	(0.000)	(0.000)	(0.382)	(0.000)	(0.115)	(0.000)	(1.000)	(1.000)
	$5\boldsymbol{I}_{10}$	0.352	0.029	0.005	0.002	0.200	0.101	0.180	0.054	0.443	0.467
		(1.000)	(0.980)	(0.000)	(0.000)	(0.520)	(0.120)	(0.293)	(0.000)	(1.000)	(1.000)

Table 2.4: Estimated misclassification rates of regular observations (and proportion of undetected outliers within parentheses) for outlying cluster contaminated datasets.

ison with the other methods is better in case of uniformly (from annulus) contaminated datasets (as compared to that for uniformly (chi-squared) contaminated datasets). The proposed method clearly beats trimmed K-means and has a very similar performance to that of the MCLUST method. The K-medoids have slightly lower misclassification rates but have no outlier detection capability.

The simulation outputs for the outlying cluster contaminated datasets, presented in Table 2.4, indicate that the proposed method clearly outperforms the other methods in this case. It should be noted that the outlying cluster simulation scheme, being stochastic in nature, sometimes generates a contaminating proportion slightly larger than 10% (although the assignment probability is exactly 0.1 for the contaminating part), and consequently, some of these contaminated observations (very distant) are

		MPLE$_\beta$				TCLUST		TKMEANS		KMEDOIDS	MCLUST
Type	p	$\beta = 0$	$\beta = 0.1$	$\beta = 0.3$	$\beta = 0.5$	$\alpha = 0$	$\alpha = 0.05$	$\alpha = 0$	$\alpha = 0.05$		
Pure	2	0.004 (0.000)	0.004 (0.0002)	0.004 (0.0002)	0.004 (0.0002)	0.003 (0.000)	0.051 (0.050)	0.010 (0.001)	0.055 (0.050)	0.009 (0.000)	0.004 (0.000)
Pure	6	0.0002 (0.0001)	0.0002 (0.0002)	0.0003 (0.0002)	0.0003 (0.0003)	0.000 (0.000)	0.052 (0.050)	0.001 (0.001)	0.001 (0.050)	0.0004 (0.000)	0.0005 (0.000)
						α =0.10	α =0.15	α =0.10	α = 0.15		
Uniform (chi-squared) Contaminated	2	0.029 (0.180)	0.028 (0.147)	0.022 (0.000)	0.024 (0.000)	0.008 (0.030)	0.058 (0.000)	0.012 (0.038)	0.057 (0.000)	0.009 (1.000)	0.021 (0.000)
Uniform (chi-squared) Contaminated	6	0.003 (0.054)	0.002 (0.000)	0.003 (0.000)	0.003 (0.000)	0.005 (0.032)	0.054 (0.000)	0.005 (0.036)	0.056 (0.000)	0.0003 (1.000)	0.001 (0.000)
Uniform (Annulus) Contaminated	2	0.014 (0.000)	0.013 (0.022)	0.006 (0.000)	0.005 (0.000)	0.008 (0.038)	0.053 (0.000)	0.013 (0.042)	0.057 (0.000)	0.010 (1.000)	0.009 (0.000)
Uniform (Annulus) Contaminated	6	0.003 (0.002)	0.0002 (0.004)	0.0004 (0.003)	0.0004 (0.004)	0.004 (0.044)	0.057 (0.0003)	0.005 (0.038)	0.056 (0.000)	0.0002 (1.000)	0.002 (0.004)
Outlying Cluster	2	0.067 (0.271)	0.068 (0.240)	0.026 (0.000)	0.025 (0.000)	0.104 (0.274)	0.057 (0.000)	0.090 (0.255)	0.058 (0.000)	0.358 (1.000)	0.340 (0.999)
Outlying Cluster	6	0.036 (0.996)	0.018 (0.700)	0.003 (0.000)	0.003 (0.000)	0.094 (0.266)	0.053 (0.000)	0.084 (0.188)	0.055 (0.000)	0.336 (1.000)	0.332 (0.999)

Table 2.5: Estimated misclassification rates with proportions of regular observations misclassified as outliers (in case of pure datasts) and proportion of undetected outliers (in case of contaminated datasets) (within parentheses) for datasets with differentially dispersed clusters.

not trimmed by the $\alpha = 0.1$ trimming level in TCLUST. This can have a potentially negative impact on the performance of TCLUST in such cases. The K-medoids performs poorly in terms of the misclassification rate of regular observations. MCLUST performs the worst in terms of classification of regular observations and also fails poorly in detecting outlying observations.

For the differentially dispersed simulation set-up, the proposed method has more or less performed the best as compared to the other methods in terms of misclassification rates as well as bias and mean squared errors of the cluster means. The superiority is more prominent in case of outlying cluster contaminated datasets.

Although the K-medoids method performs better than some of its competitors in terms of estimated regular misclassification rates in some cases, it completely fails to

detect outliers and also has higher bias and mean squared errors of the estimated cluster centers compared to those obtained by the present proposal with moderate values of β (see Section 2.7.7).

The estimated bias and mean squared errors of the estimated cluster means are presented in Section 2.7.7. The proposed method performs quite well on the average. The use of TCLUST with an $\alpha = 0.15$ trimming level (i.e., a trimming level larger than the actual contamination rate) also performs quite well with respect to these accuracy measures.

2.4.3 Empirical Running Times

To compute the empirical running times of the MPLE$_\beta$ algorithm and its other alternatives (considered in the simulation experiments), we have simulated 100 uniformly (chi-squared method) contaminated datasets of sample size $n = 1000$ with number of clusters $k = 3$, component means at $\boldsymbol{\mu}_1 = (0, 0, \ldots, 0)^t$, $\boldsymbol{\mu}_2 = (5, 5, \ldots, 5)^t$ and $\boldsymbol{\mu}_3 = (-5, -5, \ldots, -5)^t$, cluster covariance matrices $\boldsymbol{I}_p$, cluster proportions 0.25, 0.30 and 0.35 (rest 10% of the observations are the contaminating observations) and data dimensions $p = 2,\ 4,\ 6,\ 8$ and 10. The average running times per sample (in seconds) of our method along with its robust and non-robust competitors based on these simulated datasets are provided in Table 2.6.

p	MPLE$_\beta$				TCLUST	TKMEANS	KMEDOID	MCLUST
	β =0.0	β =0.1	β =0.3	β =0.5				
2	6.11	14.02	15.34	15.95	0.26	4.13	0.10	0.001
4	6.63	16.22	16.83	17.06	0.26	4.84	0.14	0.001
6	7.09	16.92	18.26	20.34	0.26	5.10	0.16	0.001
8	19.95	20.05	21.51	26.05	0.87	8.92	0.18	0.02
10	20.31	23.34	24.79	31.17	1.03	11.01	0.20	0.05

Table 2.6: Empirical running times (in seconds) per sample of different clustering algorithms under the uniformly contaminated (chi-squared method) set-up.

As observed, our method requires comparatively higher running times to converge possibly because of the minimum DPD estimation of each of the component means and covariance matrices at each of the iterations. One of our future plans is to improve this algorithm with the one-step estimators (discussed in Chapter 5) of component means and covariance matrices which will surely reduce the time complexity of the present clustering algorithm.

2.4.4 Cluster Stability

From a statistical perspective, cluster stability essentially means that the derived cluster orientations do not change drastically if the datasets become slightly modified. In order to observe this empirically, we have simulated 100 uniformly (chi-squared method) contaminated datasets of sample size $n = 1000$ with number of clusters $k = 3$, component means at $\boldsymbol{\mu}_1 = (0, 0, \ldots, 0)^t$, $\boldsymbol{\mu}_2 = (5, 5, \ldots, 5)^t$ and $\boldsymbol{\mu}_3 = (-5, -5, \ldots, -5)^t$, cluster covariance matrices $\boldsymbol{I}_p$, cluster proportions 0.25, 0.30 and 0.35 (rest 10% of the observations are the contaminating observations) and data dimensions $p = 2,\ 4,\ 6,\ 8$ and 10. The average misclassification rates of the regular

p	Sample	MPLE_β		
		β =0.1	β =0.3	β =0.5
2	Original	0.013	0.029	0.031
		(0.081)	(0.007)	(0.005)
	Reduced	0.012	0.027	0.032
		(0.078)	(0.009)	(0.006)
4	Original	0.007	0.009	0.010
		(0.008)	(0.006)	(0.006)
	Reduced	0.007	0.008	0.009
		(0.007)	(0.005)	(0.006)
6	Original	0.0005	0.0006	0.0008
		(0.004)	(0.004)	(0.004)
	Reduced	0.0006	0.001	0.001
		(0.003)	(0.004)	(0.005)
8	Original	0.0004	0.001	0.0001
		(0.007)	(0.008)	(0.007)
	Reduced	0.0005	0.0009	0.0002
		(0.006)	(0.008)	(0.009)
10	Original	0.0003	0.0003	0.0003
		(0.004)	(0.003)	(0.003)
	Reduced	0.0004	0.0003	0.0003
		(0.002)	(0.002)	(0.004)

Table 2.7: Average misclassification rates of the regular observations and the average proportions of undetected outliers (within parentheses) of our method based on the original and reduced datasets under the uniformly (chi-squared method) contaminated set-up.

observations and the average proportions of undetected outliers have been determined based on the aforesaid simulated datasets. Along with these, we have obtained 100 additional datasets of sample size $n = 950$ from the aforesaid 100 simulated datasets (sample size $n = 1000$) by randomly discarding 50 observations from each of the samples of size $n = 1000$. The average misclassification rates of the regular observations and the average proportions of undetected outliers have also been determined based on

the aforesaid reduced datasets (of size $n = 950$). The average misclassification rates of the regular observations (based on both original and reduced datasets) and the average proportions of undetected outliers (based on both original and reduced datasets) are provided in Table 2.7.

It can be easily observed that the estimated misclassification rates or the proportions of undetected outliers for the original and the reduced samples do not differ much which indicates the stability of the clusters derived by the MPLE$_\beta$ algorithm.

2.4.5 Other Accuracy Measures

We have only considered the estimated misclassification rate and the proportion of undetected outliers as the accuracy measures to assess our proposed clustering algorithm along with its competitors. To understand them in terms of macro-precision, macro-recall and F-score, we have again considered the same simulation set-up (100 uniformly (chi-squared method) contaminated samples of size $n = 1000$) as considered in the last couple of subsections. The outputs are presented in Table 2.8.

p	Measure	MPLE$_\beta$				TCLUST	TKMEANS	KMEDOID	MCLUST
		β =0.0	β =0.1	β =0.3	β =0.5				
2	Precision	0.979	0.985	0.998	0.997	1.000	1.000	0.906	0.994
	Recall	0.995	0.994	0.991	0.991	0.938	0.938	0.998	0.994
	F-score	0.987	0.990	0.995	0.994	0.968	0.969	0.950	0.994
4	Precision	0.993	1.000	1.000	1.000	1.000	1.000	0.907	1.000
	Recall	0.995	0.996	0.995	0.995	0.939	0.938	0.999	1.000
	F-score	0.994	0.998	0.997	0.997	0.968	0.967	0.951	1.000
6	Precision	0.992	1.000	1.000	1.000	1.000	1.000	0.897	1.000
	Recall	1.000	1.000	1.000	1.000	0.949	0.950	1.000	1.000
	F-score	0.996	1.000	1.000	1.000	0.974	0.977	0.946	1.000
8	Precision	0.926	0.999	0.999	0.999	1.000	1.000	0.909	1.000
	Recall	1.000	1.000	1.000	1.000	0.939	0.939	1.000	1.000
	F-score	0.961	0.999	0.999	0.999	0.969	0.969	0.952	1.000
10	Precision	0.924	0.999	0.999	0.999	1.000	1.000	0.910	1.000
	Recall	1.000	1.000	1.000	1.000	0.938	0.939	1.000	1.000
	F-score	0.960	0.999	0.999	0.999	0.968	0.967	0.953	1.000

Table 2.8: Estimated macro-precision, macro-recall and F-scores of different clustering algorithms under the uniformly (chi-squared method) contaminated set-up.

This is to be noted from the aforesaid values of the macro-precision, macro-recall and F-scores of different clustering algorithms that almost all the methods are performing reasonably in terms of these accuracy measures. However, the MPLE$_\beta$ and the MCLUST algorithms remain superior compared to the other methods in terms of these measures.

2.5 Real Data Examples

2.5.1 Swiss Bank Notes Data

These data, originally considered in Flury and Riedwyl (1988) [51], have been accessed from the R-cloud of datasets. The data consist of 200 old Swiss 1000-franc bank notes. It is known that the first 100 notes are genuine and the remaining are counterfeit. Our interest is in determining whether our algorithm can detect the counterfeit notes based on these data. Six measurements are made on each bank note: (i) length of the bank note, (ii) height of the bank note (measured along the left side), (iii) height of the bank note (measured along the right side), (iv) distance of inner frame to the lower border, (v) distance of inner frame to the upper border and (vi) length of the diagonal.

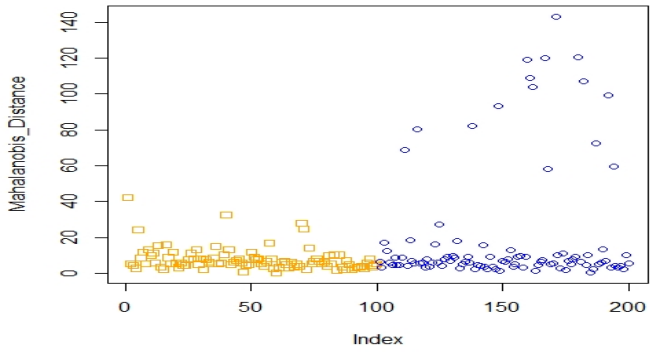
To study these data, we begin with an exploratory data analysis. We split the data into two groups according to the true nature of the notes (i.e., genuine or counterfeit) and then estimate the location and dispersion of each of these groups using the MCD method. We then estimate the squared Mahalanobis distances of the observations from their respective group (or cluster) centers. For the i-th observation $\boldsymbol{X}_i$, therefore, we compute,

$$d_i = \begin{cases} (\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_1)' \hat{\boldsymbol{\Sigma}}_1^{-1} (\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_1), & \text{for } 1 \leq i \leq 100, \\ (\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_2)' \hat{\boldsymbol{\Sigma}}_2^{-1} (\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_2), & \text{for } 101 \leq i \leq 200, \end{cases} \tag{2.15}$$

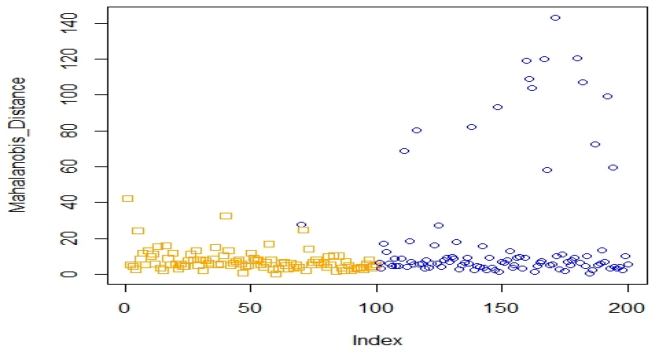
where $\hat{\boldsymbol{\mu}}_j$ and $\hat{\boldsymbol{\Sigma}}_j$ are the MCD based location and dispersion estimates of the j-th group, $j = 1, 2$.

Figure 2.2a presents the index plot of these (robust) squared Mahalanobis distances; the first 100 indices represent the squared Mahalanobis distances of the genuine notes while the last 100 represent the same for the counterfeit notes. In this plot, some points are far above the baseline with ordinates that are much larger compared to the ordinates of the general cloud of points concentrated near the horizontal axis. These observations are "far away" from their true cluster centers in terms of their estimated squared Mahalanobis distances and are therefore anomalous. It should also be noted that these anomalous observations are primarily from the group of counterfeit notes. In the figure, the genuine notes are represented as orange squares, and the counterfeit notes as blue circles.

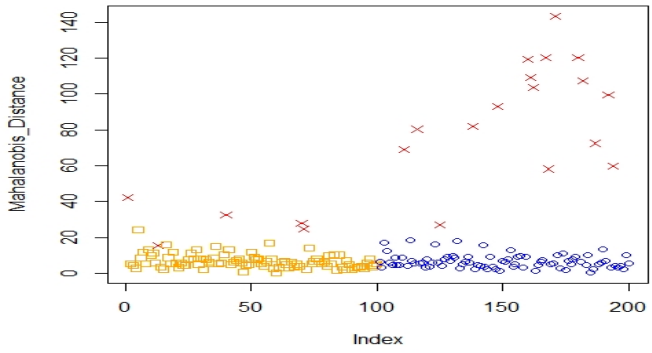
Now, we apply our robust method on these data along with the K-medoids, trimmed K-means, TCLUST and MCLUST methods. For our proposed method, we have taken $\beta = 0.5$. For the TCLUST we have taken $\alpha = 0.10$, while for the trimmed K-means



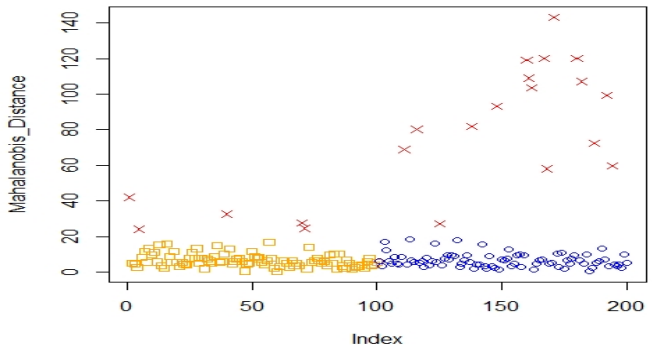
(a) Original



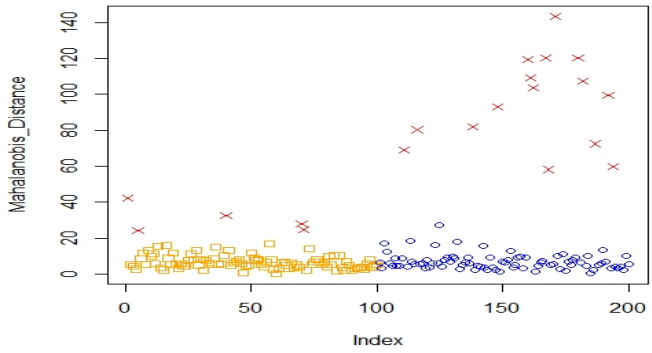
(b) K-medoids



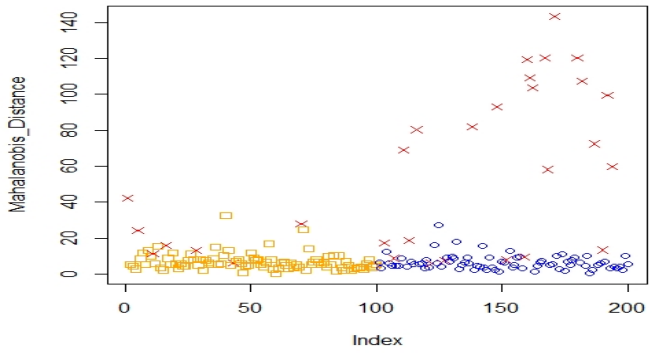
(c) MPLE$_\beta$, $\beta = 0.5$



(d) MCLUST



(e) TCLUST, $\alpha = 0.10$



(f) Trimmed K-means, $\alpha = 0.15$

Figure 2.2: Clusters derived from different methods for the Swiss Bank Notes data. The vertical axis presents estimated squared Mahalanobis distances of the observations from their respective estimated (MCD) cluster centers.

$\alpha = 0.15$ has been used. The MCLUST method has been applied with 2 clusters (and uniform noise component). The other panels of Figure 2.2 contain the clusters derived from each of these methods. Since the data are 6 dimensional, it is not possible to present the clusters along with the scatterplot of the data. So, in the remaining panels of Figure 2.2, we present the squared Mahalanobis distance values (as depicted in Figure 2.2a) for each index (through the observed magnitudes), the classification results according to the specified algorithm (orange squares representing observations classified as genuine and blue circles representing observations classified as counterfeit) and the outliers detected by the algorithm (depicted by red crosses).

In general it appears that the robust methods are all successful in doing the classifications and labelling the outliers correctly. There are rare misclassifications in the K-medoids case, but, more importantly, the latter makes no contribution to the issue of anomaly detection.

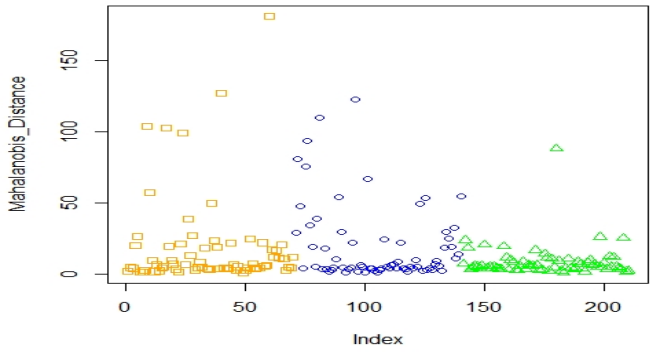
2.5.2 Seed Data

These data[1] contain measurements of geometrical properties of three different varieties of wheat, namely, Kama, Rosa and Canadian. This dataset consists of 210 (70 of each type) observations on seven attributes that are all continuous and real-valued. The attributes are (i) area A, (ii) perimeter P, (iii) compactness $C = 4\pi A/P^2$, (iv) length of kernel, (v) width of kernel, (vi) asymmetry coefficient and (vii) length of kernel groove, respectively.

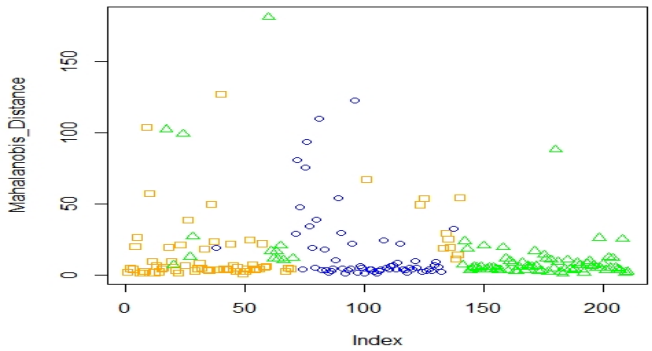
As we have done in case of the Swiss Bank Notes data, we first perform an exploratory data analysis by calculating the MCD estimates of cluster centers and dispersion matrices and calculate the estimated squared Mahalanobis distances of the observations from their respective cluster center estimates (MCD). These distances are presented in Figure 2.3a which confirms the presence of some outlying observations. These observations are "far" away from the baseline as the squared Mahalanobis distances of these points from their respective estimated (MCD) cluster centers are much larger compared to majority of the points.

The presence of such outlying observations suggests the need for robust clustering tools to analyze these data. We will apply our method with $\beta = 0.3$ and compare it with the K-medoids, TCLUST (with $\alpha = 0.2$), trimmed K-means (with $\alpha = 0.2$) and MCLUST algorithm with 3 clusters (with uniform noise component). Once again

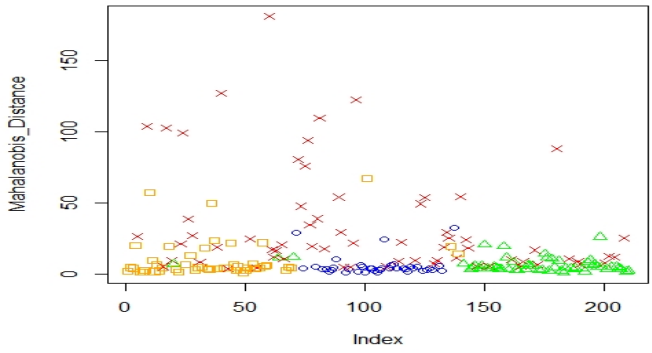
[1]Source: https://archive.ics.uci.edu/ml/datasets/seeds



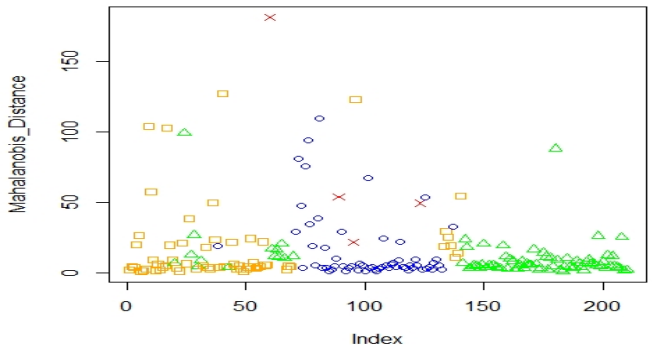
(a) Original



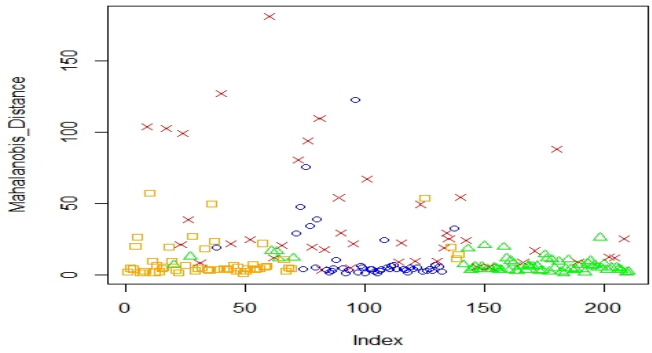
(b) K-medoids



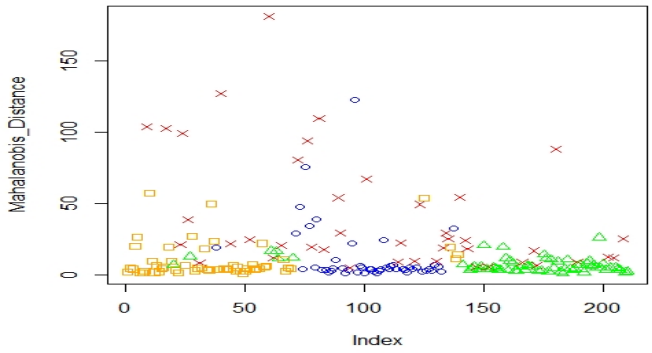
(c) MPLE$_\beta$, $\beta = 0.3$



(d) MCLUST



(e) TCLUST, $\alpha = 0.20$



(f) Trimmed K-means, $\alpha = 0.20$

Figure 2.3: Clusters derived from different methods for the Seed data. The vertical axis presents estimated squared Mahalanobis distances of the observations from their respective estimated (MCD) cluster centers.

we represent the derived clusters and outliers, in Figure 2.3, by expressing the squared Mahalanobis distance index plot using different colours and shapes, with orange square (cluster 1), blue circle (cluster 2) and green triangle (cluster 3) representing the three clusters, and red crosses indicating the outliers.

The classifications observed in Figure 2.3 indicate that the K-medoids algorithm, apart from failing in terms of outlier detection, lead to too many misclassified observations. Our proposed algorithm, the TCLUST and the trimmed K-means algorithms provide improvements through better classification and outlier detection, although neither classification nor outlier detection is done as perfectly as in case of the Swiss Bank Notes data. Our proposal does marginally better than the TCLUST and the trimmed K-means methods in terms of controlling the misclassification. The MCLUST method failed to detect the outlying observations properly.

2.6 Extension to Image Processing

Unsupervised methods for image analysis with anomaly detection is an important class of techniques in computer vision with applications in astronomy, biology, geology and many other fields. We analyze a satellite image[1] presented in Figure 2.4 with our method along with the aforesaid algorithms. For this purpose, we divide the original high resolution picture into 500×500 pixels (Figure 2.5a). Each pixel consists of a combination of three different colours (red, blue and green) with different intensities ranging from 0 to 1. We observe the intensities of the colours in a pixel as a multivariate three dimensional observation.

The two main components of the picture are the ocean body (blue region along the bottom and the left border of the image) and the coastal area (rest of the image along the top and right border) adjacent to it. But the flying body (the white body with black wings) and the shadow like regions (comparatively black regions in the coastal part, just adjacent to the ocean) should be treated as anomalies as they do not really belong to any of the two primary components. Thus we describe the image as a sample of size $500 \times 500 = 2.5 \times 10^5$ from a three dimensional, two component normal mixture population where the dimensions are the intensities of red, blue and green colours and the mixing components are the ocean body and the coastal area along with the aforesaid anomalies. We apply our method (along with the TCLUST, trimmed K-means,

[1]Source: https://medium.com/swlh/using-deep-learning-semantic-segmentation-method-for-ship-detection-on-satellite-optical-imagery-ffeaae8c1ab, see [140]

Figure 2.4: An Example of Satellite Image.

MCLUST and the ordinary K-means algorithms; K-medoids is avoided due to very large sample size of this data (250000) which is not permissible in the "PAM" function of the R software that is used to implement the K-medoids algorithm) with two clusters and check whether these methods can correctly identify those components and detect the anomalous structures (the flying object and the shadow like regions). In analyzing the dataset and reconstructing the original image, we make the following modifications to our algorithm.

Initialization: The initialization and estimation methods are same as before.

Modification 1: The first modification is in the assignment step. We use the minimum distance principle rather than the maximum likelihood principle in assigning the observations to different clusters. That is, we now assign $\boldsymbol{X}$ (the intensity vector of red, green and blue colours for a particular pixel) to C_j if the intensity vector of the cluster centre corresponding to C_j is closest to $\boldsymbol{X}$ (in the Euclidean norm) compared to all other cluster centres (instead of assigning $\boldsymbol{X}$ to the cluster which maximizes its likelihood).

Modification 2: The second modification is in the outlier detection step. Although in our data analysis examples in Section 2.5, declaration of outliers has not been cluster

specific, outliers may have different sources of possible anomalies, so a further classification among them may be useful. Note that, prior to the identification of outliers, our algorithm assigns all data points among the regular clusters. While we have not distinguished between the outliers so far, we can use these class memberships to consider a classification of the outliers; thus with k clusters, there can be k different types of outliers according to their final cluster assignments. It is likely that these points will end up representing different things in the reconstructed image.

It may be noted that the above modifications are not specific to the image under considerations; they can and should be appropriately incorporated while applying our proposed clustering techniques for any image under study. The distance based assignment (modification 1) is common in most image processing techniques. To see the requirement of the second modification, let us consider the example image given in Figure 2.4; in this image the shadow like regions and the flying object both can be regarded as outliers compared to the ocean body and coastal area but they are actually different objects. The MPLE$_\beta$ as well as the TCLUST, trimmed K-means and ordinary K-means methods will recognize these regions as outliers ignoring the structural difference between them. But this will not be helpful if the clustering algorithms aim to separately identify the anomalous flying object. Such problems are of great practical importance, e.g., in aeronautics and marine science. Although in our example we have illustrated the classification of outliers in two groups, the proposed method can be extended to similarly classify more than two types of outliers, as required, depending on the image under consideration.

Now, we implement our proposed method to analyze the image under study in Figure 2.5a (divided into 500×500 pixels); here we take $\beta = 0.2$, $T = 0.02$, $c = 20$ and $c_1 = 0.1$ in our proposal, whereas for the TCLUST method $c = 20$ and $\alpha = 0.1$ are taken and for the trimmed K-means method $\alpha = 0.1$ is taken; MCLUST is applied with 2 clusters (with uniform noise component). The reconstructed images with different methods are presented in Figure 2.5. The water body and the coastal region can be clearly identified from the reconstructed images using all the algorithms. The brown shades indicate the possible anomalous regions in case of the TCLUST method and the trimmed K-means method. The trimmed K-means method is somewhat inefficient in detecting the outliers as some of the shadow like regions in the coastal area are misclassified as water body (blue regions within the brown outlying parts). But in our method the brown shades correspond to the shadow like regions whereas the white

(a) Original image with 500×500 pixels

(b) Output of MPLE$_\beta$ method

(c) Output of trimmed K-means method

(d) Output of TCLUST method

(e) Output of K-means method

(f) Output of MCLUST method

Figure 2.5: Original and reconstructed images after applying different methods of clustering.

shades correspond to the flying object. Some regions in the coastal area which are close to white in the original image, are also detected as white outliers in our method.

So, we may conclude that the MPLE$_\beta$ and the TCLUST algorithms successfully point out these areas except for some areas on the wings of the flying object but the trimmed K-means method cannot classify those areas perfectly. On the other hand, the MCLUST algorithm could not detect either the flying object or the shadow like anomalous region in the coastal area properly.

Additionally, our method specifically points out and distinguishes the flying object separately from the shadow like region through further classification of identified out-

liers. The existing implementations of the TCLUST or the trimmed K-means methods (in CRAN) do not distinguish between the outlier types. This additional refinement of our proposed methodology specifically helps to identify different small parts in the image making it a very useful robust clustering technique for image processing.

For all the three clustering methods considered here, however, a little amount of misclassification occurs possibly due to the reduced resolution of the original image.

2.7 Appendices

2.7.1 Robust clustering tools: A motivation

While Appendices are mainly for the technical proofs, here we provide a small description by way of motivation. Let us consider the following scatter plots in Figure 2.6. In the left panel of the plots, along with three main clusters, two outlying clusters (in red) are present. In the right panel, the three main clusters are contaminated with uniform noise (red points) from an annulus. If a clustering procedure with three clusters based on the estimation of the cluster mean using classical methods is run on the contaminated data, the cluster means are strongly affected and as a result the detection of the actual clusters could be highly inaccurate. The outliers affect both the estimators and the misclassification rates. Clearly, robust clustering methods may be useful in these situations to properly control the noise in the data.

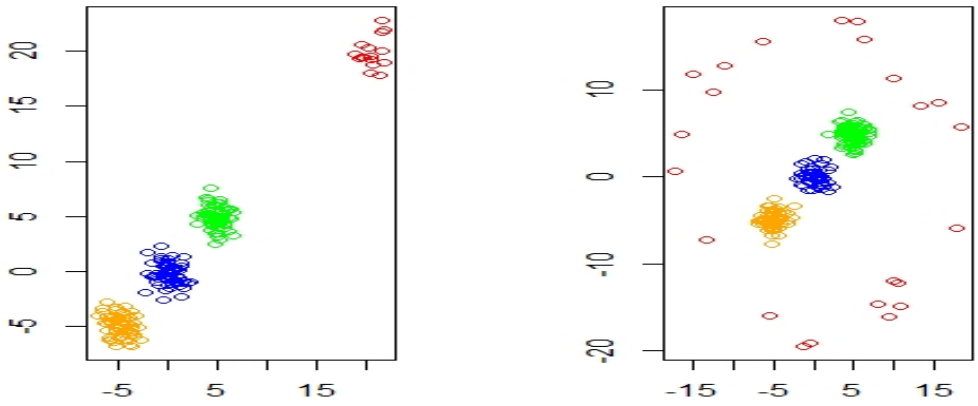


Figure 2.6: Scatter plots with outlying cluster contamination (left panel) and uniform (from annulus) contamination (right panel), for a three cluster bivariate dataset.

2.7.2 Proof of Theorem 2.1

To find the MDPDEs of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, it is enough to minimize,

$$H(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{1+\beta} \int \phi_p^{1+\beta}(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\boldsymbol{x} - \frac{1}{n\beta} \sum_{i=1}^{n} \phi_p^{\beta}(\boldsymbol{X}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ where,

$$\phi_p(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{p}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})} \text{ for } \boldsymbol{x} \in \mathbb{R}^p.$$

Since $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ model is a location-scale model, the first integral term in $H(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ $\left(\text{given by} \frac{1}{(2\pi)^{\frac{p\beta}{2}} |\boldsymbol{\Sigma}|^{\frac{\beta}{2}} (1+\beta)^{\frac{p}{2}+1}}\right)$ is independent of $\boldsymbol{\mu}$. So differentiating $H(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with respect to $\boldsymbol{\mu}$ and equating the derivative to zero, the first estimating equation becomes,

$$\frac{1}{n} \sum_{i=1}^{n} \phi_p^{\beta}(\boldsymbol{X}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma})\, \frac{\partial \log\ \phi_p(\boldsymbol{X}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\mu}} = \mathbf{0}. \tag{2.16}$$

But, we have

$$\log\ \phi_p(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{p}{2} \log\ (2\pi) - \frac{1}{2} \log\ |\boldsymbol{\Sigma}| - \frac{1}{2} (\boldsymbol{x}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})$$

and hence,

$$\frac{\partial \log\ \phi_p(\boldsymbol{X}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\mu}} = \boldsymbol{\Sigma}^{-1} (\boldsymbol{X}_i - \boldsymbol{\mu}).$$

So, the first estimating equation in Equation (2.16) simplifies to

$$\frac{1}{n} \sum_{i=1}^{n} \phi_p^{\beta}(\boldsymbol{X}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma})(\boldsymbol{X}_i - \boldsymbol{\mu}) = \mathbf{0}. \tag{2.17}$$

Next, to obtain the second estimating equation, our task is to differentiate $H(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with respect to $\boldsymbol{\Sigma}$. But it will be cumbersome and thus we will carry on the calculation with respect to $\boldsymbol{\Sigma}^{-1}$ as was done to find the maximum likelihood estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in Mardia et al. (1979) [104]. To do that, the following lemma (Mardia et al. (1979) [104]) is required.

Lemma 2.4 (Mardia et al. (1979) [104]). *Suppose $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ be a differentiable function. Then,*

1. *The derivative of $f(\boldsymbol{X})$ with respect to $\boldsymbol{X} = [[x_{ij}]]$ is given by,*

$$\frac{\partial f(\boldsymbol{X})}{\partial \boldsymbol{X}} = \left[\left[\frac{\partial f(\boldsymbol{X})}{\partial x_{ij}}\right]\right].$$

2. *If $\boldsymbol{X}$ is symmetric,*

$$\frac{\partial |\boldsymbol{X}|}{\partial x_{ij}} = \begin{cases} \boldsymbol{X}_{ii}, & \text{if } i = j, \\ 2\boldsymbol{X}_{ij}, & \text{otherwise,} \end{cases}$$

   where $\boldsymbol{X}_{ij}$ being the (i, j)-th cofactor of $\boldsymbol{X}$.

3. *If $\boldsymbol{X}$ is symmetric,*

$$\frac{\partial tr(\boldsymbol{X}\boldsymbol{A})}{\partial \boldsymbol{X}} = \boldsymbol{A} + \boldsymbol{A}^{'} - Diag(\boldsymbol{A}),$$

   where $tr(\boldsymbol{A})$ denotes the trace of the matrix $\boldsymbol{A}$ and $Diag(\boldsymbol{A})$ denotes the diagonal matrix whose diagonal elements are that of the matrix $\boldsymbol{A}$.

Coming back to our problem,

$$\begin{aligned} \frac{\partial \log\ \phi_p(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}^{-1}} &= \frac{\partial}{\partial \boldsymbol{\Sigma}^{-1}} \left[\frac{1}{2} \log\ |\boldsymbol{\Sigma}^{-1}| - \frac{1}{2} tr(\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^{'})\right] \\ &= \frac{\partial}{\partial \boldsymbol{V}} \left[\frac{1}{2} \log\ |\boldsymbol{V}| - \frac{1}{2} tr(\boldsymbol{V}(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^{'})\right],\ \boldsymbol{V} = \boldsymbol{\Sigma}^{-1}. \end{aligned}$$

Now, using Lemma 2.4,

$$\frac{\partial}{\partial \boldsymbol{V}} \log\ |\boldsymbol{V}| = \frac{1}{|\boldsymbol{V}|} \frac{\partial |\boldsymbol{V}|}{\partial \boldsymbol{V}} \text{ whose } (i,j)\text{-th element is } \begin{cases} \frac{2\boldsymbol{V}_{ij}}{|\boldsymbol{V}|}, & \text{if } i \neq j, \\ \frac{\boldsymbol{V}_{ii}}{|\boldsymbol{V}|}, & \text{if } i = j, \end{cases}$$

where $\boldsymbol{V}_{ij}$ is the (i, j)-th cofactor of $\boldsymbol{V}$ and

$$\frac{\partial}{\partial \boldsymbol{V}} tr\left(\boldsymbol{V}(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^{'}\right) = 2(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^{'} - Diag(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^{'}.$$

Since, $\boldsymbol{V} = \boldsymbol{\Sigma}^{-1}$ is symmetric, the matrix with elements $\frac{\boldsymbol{V}_{ij}}{|\boldsymbol{V}|}$ equals $\boldsymbol{V}^{-1} = \boldsymbol{\Sigma}$. Hence,

$$\begin{aligned}\frac{\partial \log\ \phi_p(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}^{-1}} &= \frac{\partial}{\partial \boldsymbol{V}} \left[\frac{1}{2} \log\ |\boldsymbol{V}| - \frac{1}{2} tr(\boldsymbol{V}(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^{'})\right]\ , \boldsymbol{V} = \boldsymbol{\Sigma}^{-1} \\ &= \frac{1}{2}\boldsymbol{M} - \frac{1}{2}(2\boldsymbol{S} - Diag(\boldsymbol{S})),\end{aligned}$$

where $\boldsymbol{S} = (\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^{'}$ and $\boldsymbol{M} = 2\boldsymbol{\Sigma} - Diag(\boldsymbol{\Sigma})$. Hence, differentiating $H(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with respect to $\boldsymbol{\Sigma}^{-1}$ and equating the derivative to zero, we have the second estimating equation

$$\begin{aligned}&\tfrac{1}{n}\textstyle\sum_{i=1}^{n} \phi_p^{\beta}(\boldsymbol{X}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma})\ \frac{\partial \log \phi_p(\boldsymbol{X}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}^{-1}} = e_0 \\ &\implies \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \phi_p^{\beta}(\boldsymbol{X}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma})\tfrac{1}{2}(\boldsymbol{M} - (2\boldsymbol{S}_i - Diag(\boldsymbol{S}_i))) = e_0,\end{aligned} \tag{2.18}$$

where $\boldsymbol{S}_i = (\boldsymbol{X}_i - \boldsymbol{\mu})(\boldsymbol{X}_i - \boldsymbol{\mu})^{'}$ and

$$\begin{aligned}e_0 &= \frac{\partial}{\partial \boldsymbol{\Sigma}^{-1}} \left[\frac{1}{1+\beta} \int \phi_p^{1+\beta}(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})\ d\boldsymbol{x}\right] \\ &= \int \phi_p^{1+\beta}(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \frac{\partial \log\ \phi_p(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}^{-1}}\ d\boldsymbol{x} \\ &= \frac{\boldsymbol{M}}{2} \int \phi_p^{1+\beta}(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})\ d\boldsymbol{x} - \int (\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^{'} \phi_p^{1+\beta}(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})\ d\boldsymbol{x} \\ &+ \frac{1}{2} \int Diag(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^{'} \phi_p^{1+\beta}(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})\ d\boldsymbol{x}.\end{aligned}$$

Now using the facts that,

$$\phi_p^{1+\beta}(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{p\beta}{2}} |\boldsymbol{\Sigma}|^{\frac{\beta}{2}} (1+\beta)^{\frac{p}{2}}} \phi_p(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}_0) \tag{2.19}$$

with $\boldsymbol{\Sigma}_0 = \frac{1}{1+\beta}\boldsymbol{\Sigma}$ and $(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^{'}$ follows a p-dimensional Wishart distribution with scale parameter $\boldsymbol{\Sigma}_0$ and 1 degree of freedom for a random vector $\boldsymbol{X}$ which follows $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}_0)$,

$$\begin{aligned}e_0 &= \frac{\boldsymbol{M}}{2} \frac{1}{(2\pi)^{\frac{p\beta}{2}} |\boldsymbol{\Sigma}|^{\frac{\beta}{2}} (1+\beta)^{\frac{p}{2}}} - \frac{\boldsymbol{M}}{2} \frac{1}{1+\beta} \frac{1}{(2\pi)^{\frac{p\beta}{2}} |\boldsymbol{\Sigma}|^{\frac{\beta}{2}} (1+\beta)^{\frac{p}{2}}} \\ &= \frac{\boldsymbol{M}\beta}{2(2\pi)^{\frac{p\beta}{2}} |\boldsymbol{\Sigma}|^{\frac{\beta}{2}} (1+\beta)^{\frac{p+2}{2}}}.\end{aligned}$$

Precisely, the second estimating equation in Equation (2.18) becomes,

$$\frac{1}{n}\sum_{i=1}^{n}\phi_p^{\beta}(\boldsymbol{X}_i,\boldsymbol{\mu},\boldsymbol{\Sigma})\frac{1}{2}(\boldsymbol{M}-(2\boldsymbol{S}_i-Diag(\boldsymbol{S}_i)))=\frac{\boldsymbol{M}\beta}{2(2\pi)^{\frac{p\beta}{2}}|\boldsymbol{\Sigma}|^{\frac{\beta}{2}}(1+\beta)^{\frac{p+2}{2}}}$$

$$\implies \frac{1}{n}\sum_{i=1}^{n}\phi_p^{\beta}(\boldsymbol{X}_i,\boldsymbol{\mu},\boldsymbol{\Sigma})(\boldsymbol{M}-(2\boldsymbol{S}_i-Diag(\boldsymbol{S}_i)))=c_0\boldsymbol{M}$$

which is algebraically equivalent to

$$\frac{1}{n}\sum_{i=1}^{n}\phi_p^{\beta}(\boldsymbol{X}_i,\boldsymbol{\mu},\boldsymbol{\Sigma})(\boldsymbol{\Sigma}-(\boldsymbol{X}_i-\boldsymbol{\mu})(\boldsymbol{X}_i-\boldsymbol{\mu})')=c_0\boldsymbol{\Sigma}, \tag{2.20}$$

where $c_0=\beta(2\pi)^{-\frac{p\beta}{2}}|\boldsymbol{\Sigma}|^{-\frac{\beta}{2}}(1+\beta)^{-\frac{p+2}{2}}$. Now, by expanding the multivariate normal density $\phi_p(\boldsymbol{X}_i,\boldsymbol{\mu},\boldsymbol{\Sigma})$ in Equations (2.17) and (2.20), we get the following simplified versions of the aforesaid estimating equations.

$$\frac{1}{n}\sum_{i=1}^{n}e^{-\frac{\beta}{2}(\boldsymbol{X}_i-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_i-\boldsymbol{\mu})}(\boldsymbol{X}_i-\boldsymbol{\mu})=0,$$

$$\frac{1}{n}\sum_{i=1}^{n}e^{-\frac{\beta}{2}(\boldsymbol{X}_i-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_i-\boldsymbol{\mu})}(\boldsymbol{\Sigma}-(\boldsymbol{X}_i-\boldsymbol{\mu})(\boldsymbol{X}_i-\boldsymbol{\mu})')=\frac{\beta}{(1+\beta)^{\frac{p}{2}+1}}\boldsymbol{\Sigma}.$$

This completes derivation of the estimating equations.

2.7.3 Proof of Theorem 2.2

Since the cluster assignments and the estimates $\hat{\boldsymbol{\mu}}_j$, $\hat{\boldsymbol{\Sigma}}_j$ are known, the values of the assignment functions $Z(\cdot,\cdot)$ are also known. Thus, maximizing the empirical objective function with respect to $(\pi_1,\pi_2,\ldots,\pi_k)$ is equivalent to the optimization problem:

$$\textbf{maximize}\quad \sum_{j=1}^{k}n_j\log\ \pi_j, \qquad \textbf{subject to}\quad \sum_{j=1}^{k}\pi_j=1.$$

Now the optimal choice of π_j can be directly obtained by optimizing the Lagrangian,

$$l(\pi_1,\pi_2,\ldots,\pi_k,\lambda)=\sum_{j=1}^{k}n_j\log\ \pi_j-\lambda\left(\sum_{j=1}^{k}\pi_j-1\right).$$

2.7.4 Derivations of the Systems of Equations defining the Maximum Pseudo β-Likelihood Functional (MPLF$_\beta$)

Consider the set-up and notation of Section 2.3.2. We now present the derivation of the system of equations defining the MPLF$_\beta$ $\boldsymbol{\theta}(P)$, given in system (2.13) of equations.

Let us first note that, under $p = 1, k = 2$ and the true distribution function P, our objective function for the MPLF$_\beta$ functional is given by

$$L_\beta(\boldsymbol{\theta}, P) = E_P \left[\sum_{j=1}^{2} Z_j(X, \boldsymbol{\theta}) \left[\log\ \pi_j + \frac{1}{\beta} f^\beta(X, \mu_j, \sigma_j^2) - \frac{1}{1+\beta} \int f^{1+\beta}(x, \mu_j, \sigma_j^2)\, dx \right] \right], \tag{2.21}$$

where $f(\cdot, \mu_j, \sigma_j^2)$ is the pdf of univariate normal distribution with mean μ_j and variance σ_j^2 for $j = 1, 2$, P is the true but unknown distribution function and hence the parameter as a functional can be written as,

$$\boldsymbol{\theta}_\beta(P) = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}_C}{\operatorname{argmax}}\ L_\beta(\boldsymbol{\theta}, P). \tag{2.22}$$

In our case, the assignment functions can be presented as,

$$\begin{aligned} &Z_1(X, \boldsymbol{\theta}) = I(D_1(X, \boldsymbol{\theta}) \geq D_2(X, \boldsymbol{\theta})) \text{ and} \\ &Z_2(X, \boldsymbol{\theta}) = 1 - Z_1(X, \boldsymbol{\theta}). \end{aligned}$$

Note that $D_1(X, \boldsymbol{\theta})$ and $D_2(X, \boldsymbol{\theta})$ are bell-shaped convex functions. Hence,

$$\begin{aligned} &D_1(X, \boldsymbol{\theta}) \geq D_2(X, \boldsymbol{\theta}) \\ &\implies \frac{(X-\mu_2)^2}{\sigma_2^2} - \frac{(X-\mu_1)^2}{\sigma_1^2} \geq 2 \log\ \frac{\pi_2 \sigma_1}{\pi_1 \sigma_2} \\ &\implies a_1 X^2 + a_2 X + a_3 \geq 0 \end{aligned}$$

for suitable values of a_1, a_2 and a_3. Since, a quadratic equation can have at most two real roots, the aforesaid inequality leads to the following possibilities,

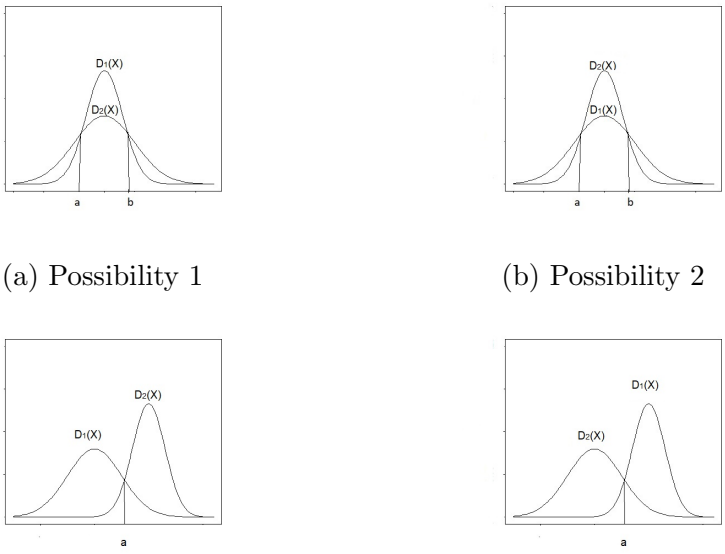
Possibility 1: $X \in (a, b)$ if $a_1 < 0$.

Possibility 2: $X \in (-\infty, a) \cup (b, \infty)$ if $a_1 > 0$.

Possibility 3: $X \in (-\infty, a)$ or $X \in (a, \infty)$ if $a_1 = 0$.

Let us carefully note that, the constants a and b depend on the parameter $\boldsymbol{\theta}$. So, we have to consider these additional functionals $a = a(P)$ and $b = b(P)$ in order to derive

the influence function of $\boldsymbol{\theta}(P)$. The above possibilities can be graphically observed in Figure 2.7.



(a) Possibility 1

(b) Possibility 2

(c) Possibility 3

(d) Possibility 3

Figure 2.7: Different possibilities.

We will first derive the influence function in possibility:1; the derivation of the same for the remaining possibilities are similar. To do that, let us first study the mathematical relationships among these functionals. The proportion functional satisfies,

$$\pi_1(P) + \pi_2(P) = 1.$$

From the plots in Figure 2.7, the functional $\boldsymbol{\theta}_\beta(P)$ satisfies,

$$\begin{aligned} D_1(a, \boldsymbol{\theta}_\beta(P)) &= D_2(a, \boldsymbol{\theta}_\beta(P)), \\ D_1(b, \boldsymbol{\theta}_\beta(P)) &= D_2(b, \boldsymbol{\theta}_\beta(P)). \end{aligned} \tag{2.23}$$

Next let us simplify $L_\beta(\boldsymbol{\theta}, P)$ under $k = 2$ and $p = 1$.

$$
\begin{aligned}
L_\beta(\boldsymbol{\theta}, P) &= E_P\left[\sum_{j=1}^{2} Z_j(X, \boldsymbol{\theta})\left[\log\ \pi_j + \frac{1}{\beta} f^\beta(X, \mu_j, \sigma_j^2) - \frac{1}{1+\beta}\int f^{1+\beta}(x, \mu_j, \sigma_j^2)\, dx\right]\right] \\
&= E_P\left[I(D_1(X, \boldsymbol{\theta}) > D_2(X, \boldsymbol{\theta}))\left[\log\ \pi_1 + \frac{1}{\beta} f^\beta(X, \mu_1, \sigma_1^2) - \frac{1}{1+\beta}\int f^{1+\beta}(x, \mu_1, \sigma_1^2)\, dx\right]\right] \\
&\quad + E_P\left[I(D_2(X, \boldsymbol{\theta}) > D_1(X, \boldsymbol{\theta}))\left[\log\ \pi_2 + \frac{1}{\beta} f^\beta(X, \mu_2, \sigma_2^2) - \frac{1}{1+\beta}\int f^{1+\beta}(x, \mu_2, \sigma_2^2)\, dx\right]\right] \\
&= E_P\left[I(X \in (a, b))\left[\log\ \pi_1 + \frac{1}{\beta} f^\beta(X, \mu_1, \sigma_1^2) - \frac{1}{1+\beta}\int f^{1+\beta}(x, \mu_1, \sigma_1^2)\, dx\right]\right] \\
&\quad + E_P\left[I(X \notin (a, b))\left[\log\ \pi_2 + \frac{1}{\beta} f^\beta(X, \mu_2, \sigma_2^2) - \frac{1}{1+\beta}\int f^{1+\beta}(x, \mu_2, \sigma_2^2)\, dx\right]\right] \\
&= \left[(P(b) - P(a))\log\ \pi_1 + \frac{1}{\beta}\int_a^b f^\beta(x, \mu_1, \sigma_1^2) p(x)\, dx - \frac{P(b) - P(a)}{1+\beta}\int f^{1+\beta}(x, \mu_1, \sigma_1^2)\, dx\right] \\
&\quad + \left[(1 - P(b) + P(a))\log\ \pi_2 + \frac{1}{\beta}\int_{x \notin (a,b)} f^\beta(x, \mu_2, \sigma_2^2) p(x)\, dx\right. \\
&\left. - \frac{1 - P(b) + P(a)}{1+\beta}\int f^{1+\beta}(x, \mu_2, \sigma_2^2)\, dx\right]
\end{aligned}
$$

with $p(x)$ as the true density corresponding to the true distribution function $P(x)$. Now using (2.19),

$$
\int f^{1+\beta}(x, \mu_j, \sigma_j^2)\, dx = \frac{1}{(2\pi)^{\frac{\beta}{2}} \sigma_j^\beta (1+\beta)^{\frac{1}{2}}},\ j = 1, 2.
$$

Hence,

$$
\begin{aligned}
L_\beta(\boldsymbol{\theta}, P) = &\left[(P(b) - P(a))\log\ \pi_1 + \frac{1}{\beta}\int_a^b f^\beta(x, \mu_1, \sigma_1^2) p(x)\, dx - \frac{P(b) - P(a)}{(2\pi)^{\frac{\beta}{2}} \sigma_1^\beta (1+\beta)^{\frac{3}{2}}}\right] \\
&+ \left[(1 - P(b) + P(a))\log\ \pi_2 + \frac{1}{\beta}\int_{x \notin (a,b)} f^\beta(x, \mu_2, \sigma_2^2) p(x)\, dx - \frac{1 - P(b) + P(a)}{(2\pi)^{\frac{\beta}{2}} \sigma_2^\beta (1+\beta)^{\frac{3}{2}}}\right].
\end{aligned}
$$

To find the estimator of μ_j and σ_j^2, we require differentiating the above with respect

to the respective parameters. Differentiating $L_\beta(\boldsymbol{\theta}, P)$ with respect to μ_1 and μ_2 gives,

$$\int_a^b f^\beta(x, \mu_1, \sigma_1^2)(x - \mu_1)p(x)\, dx = 0,$$
$$\int_{x \notin (a,b)} f^\beta(x, \mu_2, \sigma_2^2)(x - \mu_2)p(x)\, dx = 0.$$

And differentiating $L_\beta(\boldsymbol{\theta}, P)$ with respect to σ_1^2 and σ_2^2 gives,

$$\int_a^b f^\beta(x, \mu_1, \sigma_1^2) \left(\frac{(x - \mu_1)^2}{2\sigma_1^2} - 1 \right) p(x)\, dx + \frac{\beta(P(b) - P(a))}{2(2\pi)^{\frac{\beta}{2}} (\sigma_1^2)^{1+\frac{\beta}{2}} (1+\beta)^{\frac{3}{2}}} = 0,$$
$$\int_{x \notin (a,b)} f^\beta(x, \mu_2, \sigma_2^2) \left(\frac{(x - \mu_2)^2}{2\sigma_2^2} - 1 \right) p(x)\, dx + \frac{\beta(1 - P(b) + P(a))}{2(2\pi)^{\frac{\beta}{2}} (\sigma_2^2)^{1+\frac{\beta}{2}} (1+\beta)^{\frac{3}{2}}} = 0.$$

Additionally,

$$\pi_1(P) = \int_{a(P)}^{b(P)} p(x)\, dx$$

with $p(x)$ being the density function corresponding to $P(x)$. So, the functional $\boldsymbol{\theta}_\beta(P) = \Big(\pi_1(P), \pi_2(P), \mu_1(P), \mu_2(P),\ \sigma_1^2(P), \sigma_2^2(P)\Big)$ can be implicitly described through the following system of equations.

$$\begin{aligned}
&\pi_1(P) = \textstyle\int_{a(P)}^{b(P)} p(x)\, dx,\\
&\pi_1(P) + \pi_2(P) = 1,\\
&D_1(c, \boldsymbol{\theta}_\beta(P)) = D_2(c, \boldsymbol{\theta}_\beta(P)) \text{ for } c = a(P) \text{ and } b(P),\\
&\textstyle\int_a^b f^\beta(x, \mu_1, \sigma_1^2)(x - \mu_1)p(x)\, dx = 0,\\
&\textstyle\int_{x \notin (a,b)} f^\beta(x, \mu_2, \sigma_2^2)(x - \mu_2)p(x)\, dx = 0,\\
&\textstyle\int_a^b f^\beta(x, \mu_1, \sigma_1^2) \left(\frac{(x-\mu_1)^2}{2\sigma_1^2} - 1 \right) p(x)\, dx + \frac{\beta(P(b)-P(a))}{2(2\pi)^{\frac{\beta}{2}} (\sigma_1^2)^{1+\frac{\beta}{2}} (1+\beta)^{\frac{3}{2}}} = 0,\\
&\textstyle\int_{x \notin (a,b)} f^\beta(x, \mu_2, \sigma_2^2) \left(\frac{(x-\mu_2)^2}{2\sigma_2^2} - 1 \right) p(x)\, dx + \frac{\beta(1-P(b)+P(a))}{2(2\pi)^{\frac{\beta}{2}} (\sigma_2^2)^{1+\frac{\beta}{2}} (1+\beta)^{\frac{3}{2}}} = 0
\end{aligned}$$

which is exactly the same system described in system (2.13) of equations.

2.7.5 Derivation of the Influence Functions

Consider the set-up and notation of Section 2.3.2. Recall, we have assumed that, $\frac{M}{m} < c$ and $m > c_1$. In case of $p = 1$,

$$\frac{M}{m} = \frac{\max\{\sigma_1^2(P), \sigma_2^2(P)\}}{\min\{\sigma_1^2(P), \sigma_2^2(P)\}} < c.$$

The strict inequality is assumed to confirm that, the same constraint also holds in case of contaminated distribution, that is,

$$\frac{M_\epsilon}{m_\epsilon} = \frac{\max\{\sigma_1^2(P_\epsilon), \sigma_2^2(P_\epsilon)\}}{\min\{\sigma_1^2(P_\epsilon), \sigma_2^2(P_\epsilon)\}} < c,$$

for some small enough $\epsilon > 0$ (recall P_ϵ from Section 2.3.2). If the aforesaid eigenvalue ratio constraint does not hold for the contaminated distribution, it is not possible to derive the influence functions of our estimators which are derived under the same constraint. Similarly, the second inequality ($m > c_1$) confirms the fact that the non-singularity constraint also holds under contamination.

Now, to derive the influence functions of our estimators, let us introduce the following notations. Our functionals are $(\pi_1(P), \pi_2(P), a(P), b(P), \mu_1(P), \mu_2(P), \sigma_1^2(P), \sigma_2^2(P))$ which satisfy the system (2.13) of equations and suppose $IF(\boldsymbol{\theta}_\beta, P, y) = \big(IF(\pi_1, P, y), IF(\pi_2, P, y), IF(a, P, y), IF(b, P, y), IF(\mu_1, P, y), IF(\mu_2, P, y), IF(\sigma_1^2, P, y), IF(\sigma_2^2, P, y)\big)'$ be the vector of influence functions of the aforesaid functionals. Also let $\boldsymbol{\theta}_\epsilon, a_\epsilon$ and b_ϵ are the contaminated versions of $\boldsymbol{\theta}_\beta(P), a(P)$ and $b(P)$ respectively. Then, we have

$$\begin{aligned}\pi_{1\epsilon} &= \int_{a_\epsilon}^{b_\epsilon} dP_\epsilon(x) \\ &= (1-\epsilon)\int_{a_\epsilon}^{b_\epsilon} p(x)\, dx + \epsilon I(y \in (a_\epsilon, b_\epsilon)).\end{aligned}$$

Hence,

$$\begin{aligned}IF(\pi_1, P, y) &= \frac{\partial \pi_{1\epsilon}}{\partial \epsilon}\Big|_{\epsilon=0} \\ &= [P(a_0) - P(b_0)] + [p(b_0)IF(b, P, y) - p(a_0)IF(a, P, y)] + I(y \in (a_0, b_0)).\end{aligned}$$

The equation,

$$\pi_1(P) + \pi_2(P) = 1$$

gives

$$IF(\pi_1, P, y) + IF(\pi_2, P, y) = 0.$$

Next let us recall (Equation (2.23)) that,

$$D_1(a_\epsilon, \boldsymbol{\theta}_\epsilon) = D_2(a_\epsilon, \boldsymbol{\theta}_\epsilon)$$
$$\text{and}$$
$$D_1(b_\epsilon, \boldsymbol{\theta}_\epsilon) = D_2(b_\epsilon, \boldsymbol{\theta}_\epsilon).$$

Differentiating the above equations with respect to ϵ at 0 gives,

$$2\left[\frac{IF(\pi_1, P, y)}{\pi_{10}} - \frac{IF(\pi_2, P, y)}{\pi_{20}}\right] + \left[\frac{IF(\sigma_2^2, P, y)}{\sigma_{20}^2}\left(1 - \frac{(a_0 - \mu_{20})^2}{\sigma_{20}^2}\right) - \frac{IF(\sigma_1^2, P, y)}{\sigma_{10}^2}\left(1 - \frac{(a_0 - \mu_{10})^2}{\sigma_{10}^2}\right)\right]$$

$$= 2IF(a, P, y)\left[\frac{(a_0 - \mu_{10})}{\sigma_{10}^2} - \frac{(a_0 - \mu_{20})}{\sigma_{20}^2}\right] + 2\left[\frac{(a_0 - \mu_{20})IF(\mu_2, P, y)}{\sigma_{20}^2} - \frac{(a_0 - \mu_{10})IF(\mu_1, P, y)}{\sigma_{10}^2}\right]$$

and

$$2\left[\frac{IF(\pi_1,P,y)}{\pi_{10}}-\frac{IF(\pi_2,P,y)}{\pi_{20}}\right]+\left[\frac{IF(\sigma_2^2,P,y)}{\sigma_{20}^2}\left(1-\frac{(b_0-\mu_{20})^2}{\sigma_{20}^2}\right)\right.$$
$$\left.-\frac{IF(\sigma_1^2,P,y)}{\sigma_{10}^2}\left(1-\frac{(b_0-\mu_{10})^2}{\sigma_{10}^2}\right)\right]$$
$$=2IF(b,P,y)\left[\frac{(b_0-\mu_{10})}{\sigma_{10}^2}-\frac{(b_0-\mu_{20})}{\sigma_{20}^2}\right]+2\left[\frac{(b_0-\mu_{20})IF(\mu_2,P,y)}{\sigma_{20}^2}\right.$$
$$\left.-\frac{(b_0-\mu_{10})IF(\mu_1,P,y)}{\sigma_{10}^2}\right].$$

Let us observe that (from system (2.13) of equations),

$$\int_{a_\epsilon}^{b_\epsilon} f^\beta(x,\mu_{1\epsilon},\sigma_{1\epsilon}^2)(x-\mu_{1\epsilon})\,dP_\epsilon(x)=0 \text{ implies}$$
$$(1-\epsilon)\int_{a_\epsilon}^{b_\epsilon} f^\beta(x,\mu_{1\epsilon},\sigma_{1\epsilon}^2)(x-\mu_{1\epsilon})p(x)\,dx+\epsilon f^\beta(y,\mu_{1\epsilon},\sigma_{1\epsilon}^2)(y-\mu_{1\epsilon})I(y\in(a_\epsilon,b_\epsilon))=0.$$

Differentiating the above with respect to ϵ at 0 gives,

$$-\int_a^b f^\beta(x,\mu_1,\sigma_1^2)(x-\mu_1)p(x)\,dx+\frac{\partial}{\partial\epsilon}\int_{a_\epsilon}^{b_\epsilon} f^\beta(x,\mu_{1\epsilon},\sigma_{1\epsilon}^2)(x-\mu_{1\epsilon})p(x)\,dx\Big|_{\epsilon=0}$$
$$+f^\beta(y,\mu_1,\sigma_1^2)(y-\mu_1)I(y\in(a,b))=0.$$

To evaluate the middle term in the above equation we use the Leibniz integral rule (Intermediate Calculus (1985) [118]) as follows.

$$\frac{\partial}{\partial\epsilon}\int_{a_\epsilon}^{b_\epsilon} f^\beta(x,\mu_{1\epsilon},\sigma_{1\epsilon}^2)(x-\mu_{1\epsilon})p(x)\,dx\Big|_{\epsilon=0}=f^\beta(b,\mu_1,\sigma_1^2)(b-\mu_1)p(b)$$
$$-f^\beta(a,\mu_1,\sigma_1^2)(a-\mu_1)p(a)+\int_a^b\frac{\partial}{\partial\epsilon}f^\beta(x,\mu_{1\epsilon},\sigma_{1\epsilon}^2)(x-\mu_{1\epsilon})p(x)\Big|_{\epsilon=0}dx.$$

The calculation of $\frac{\partial}{\partial\epsilon}f^\beta(x,\mu_{1\epsilon},\sigma_{1\epsilon}^2)(x-\mu_{1\epsilon})p(x)\Big|_{\epsilon=0}$ is straightforward and it can be easily observed that, this integration will be a linear combination of $IF(\mu_1,P,y)$ and $IF(\sigma_1^2,P,y)$. The rest of the linear equations can be similarly derived as of those in the system (2.13) of equations. The derivation of these influence function in other possibilities (Figure 2.7) are similar and hence omitted.

These calculations finally lead to the following system of linear equations defining the required influence function $IF(\boldsymbol{\theta}_\beta, P, y)$.

$$A_\beta(\boldsymbol{\theta}_0, a_0, b_0) IF(\boldsymbol{\theta}_\beta, P, y) = B_\beta(y, \boldsymbol{\theta}_0, a_0, b_0),$$

where $\boldsymbol{\theta}_0 = (\pi_{10}, \pi_{20}, \mu_{10}, \mu_{20}, \sigma_{10}^2, \sigma_{20}^2), a_0, b_0$ are the values of $\boldsymbol{\theta}$, a and b, respectively, at the true model and

$$\begin{aligned}
&B_\beta(y, \boldsymbol{\theta}_0, a_0, b_0) \\
&= \Bigg(P(a_0) - P(b_0) + I(a_0 < y < b_0), 0, 0, 0, \int_{a_0}^{b_0} (x - \mu_{10}) f^\beta(x, \mu_{10}, \sigma_{10}^2) p(x)\, dx - \\
&(y - \mu_{10}) f^\beta(y, \mu_{10}, \sigma_{10}^2) I(a_0 < y < b_0), \int_{x \notin (a_0, b_0)} (x - \mu_{20}) f^\beta(x, \mu_{20}, \sigma_{20}^2) p(x)\, dx \\
&- (y - \mu_{20}) f^\beta(y, \mu_{20}, \sigma_{20}^2) I(y \notin (a_0, b_0)), \int_{a_0}^{b_0} f^\beta(x, \mu_{10}, \sigma_{10}^2) \left(\frac{(x - \mu_{10})^2}{\sigma_{10}^2} - 1 \right) p(x)\, dx \\
&- f^\beta(y, \mu_{10}, \sigma_{10}^2) \left(\frac{(y - \mu_{10})^2}{\sigma_{10}^2} - 1 \right) I(a_0 < y < b_0), \\
&\int_{x \notin (a_0, b_0)} f^\beta(x, \mu_{20}, \sigma_{20}^2) \left(\frac{(x - \mu_{20})^2}{\sigma_{20}^2} - 1 \right) p(x)\, dx \\
&- f^\beta(y, \mu_{20}, \sigma_{20}^2) \left(\frac{(y - \mu_{20})^2}{\sigma_{20}^2} - 1 \right) I(y \notin (a_0, b_0)) \Bigg).
\end{aligned}$$

The 8×8 matrix $A_\beta(\boldsymbol{\theta}_0, a_0, b_0)$ has the j-th row as A_{j*}, for $j = 1, \ldots, 8$, where

$$A_{1*} = (1, 0, p(a_0), -p(b_0), 0, 0, 0, 0)\,,$$

$$A_{2*} = (1, 1, 0, 0, 0, 0, 0, 0)\,,$$

$$A_{3*} = \left(\frac{2}{\pi_{10}}, \frac{-2}{\pi_{20}}, -2\left(\frac{(a_0 - \mu_{10})}{\sigma_{10}^2} - \frac{(a_0 - \mu_{20})}{\sigma_{20}^2} \right), 0, \frac{(a_0 - \mu_{10})}{\sigma_{10}^2}, -\frac{(a_0 - \mu_{20})}{\sigma_{20}^2}, \right.$$
$$\left. \frac{(a_0 - \mu_{10})^2}{\sigma_{10}^4} - \frac{1}{\sigma_{10}^2}, \frac{1}{\sigma_{20}^2} - \frac{(a_0 - \mu_{20})^2}{\sigma_{20}^4} \right),$$

$$A_{4*} = \left(\frac{2}{\pi_{10}}, \frac{-2}{\pi_{20}}, 0, -2\left(\frac{(b_0 - \mu_{10})}{\sigma_{10}^2} - \frac{(b_0 - \mu_{20})}{\sigma_{20}^2} \right), \frac{(b_0 - \mu_{10})}{\sigma_{10}^2}, -\frac{(b_0 - \mu_{20})}{\sigma_{20}^2}, \right.$$
$$\left. \frac{(b_0 - \mu_{10})^2}{\sigma_{10}^4} - \frac{1}{\sigma_{10}^2}, \frac{1}{\sigma_{20}^2} - \frac{(b_0 - \mu_{20})^2}{\sigma_{20}^4} \right),$$

$$A_{5*} = \left(0, 0, -f^\beta(a_0, \mu_{10}, \sigma_{10}^2)(a_0 - \mu_{10})p(a_0), f^\beta(b_0, \mu_{10}, \sigma_{10}^2)(b_0 - \mu_{10})p(b_0), C_3, 0, C_5, 0 \right),$$

$$A_{6*} = \left(0, 0, -f^\beta(a_0, \mu_{20}, \sigma_{20}^2)(a_0 - \mu_{20})p(a_0), f^\beta(b_0, \mu_{20}, \sigma_{20}^2)(b_0 - \mu_{20})p(b_0), 0, C_4, 0, C_6 \right)$$

$$A_{7*} = \left(0, 0, \frac{\beta p(a_0)}{C_0 \sigma_{10}^{2+\beta}} - f^\beta(a_0, \mu_{10}, \sigma_{10}^2) \left(\frac{(a_0 - \mu_{10})^2}{\sigma_{10}^2} - 1 \right) p(a_0), -\frac{\beta p(b_0)}{C_0 \sigma_{10}^{2+\beta}} \right.$$
$$+ f^\beta(b_0, \mu_{10}, \sigma_{10}^2 \left(\frac{(b_0 - \mu_{10})^2}{\sigma_{10}^2} - 1 \right) p(b_0), \frac{-2C_1}{\sigma_{10}^2 + 2C_5}, 0, C_8 - C_7$$
$$\left. + (P(a_0) - P(b_0)) \frac{\beta}{C_0 \sigma_{10}^{(4+2\beta)}}, 0 \right),$$

$$A_{8*} = \left(0, 0, \frac{\beta p(a_0)}{C_0 \sigma_{20}^{2+\beta}} - f^\beta(a_0, \mu_{20}, \sigma_{20}^2) \left(\frac{(a_0 - \mu_{20})^2}{\sigma_{20}^2} - 1 \right) p(a_0), -\frac{\beta p(b_0)}{C_0 \sigma_{20}^{2+\beta}} \right.$$
$$+ f^\beta(b_0, \mu_{20}, \sigma_{20}^2 \left(\frac{(b_0 - \mu_{20})^2}{\sigma_{20}^2} - 1 \right) p(b_0), 0, \frac{2C_2}{\sigma_{10}^2 - 2C_6}, 0, C_9 - C_{10}$$
$$\left. + (1 - P(a_0) + P(b_0)) \frac{\beta}{C_0 \sigma_{20}^{(4+2\beta)}} \right)$$

and

$$
\begin{aligned}
C_0 &= \frac{\beta}{2(2\pi)^{\beta/2}(1+\beta)^{1.5}},\\
C_1 &= \int_{a_0}^{b_0} (x-\mu_{10}) f^{\beta}(x,\mu_{10},\sigma_{10}^2) p(x)\, dx,\\
C_2 &= \int_{x\notin(a_0,b_0)} (x-\mu_{20}) f^{\beta}(x,\mu_{20},\sigma_{20}^2) p(x)\, dx,\\
C_3 &= \int_{a_0}^{b_0} f^{\beta}(x,\mu_{10},\sigma_{10}^2) \left(\frac{(x-\mu_{10})^2}{\sigma_{10}^2}-1\right) p(x)\, dx,\\
C_4 &= \int_{x\notin(a_0,b_0)} f^{\beta}(x,\mu_{20},\sigma_{20}^2) \left(\frac{(x-\mu_{20})^2}{\sigma_{20}^2}-1\right) p(x)\, dx,\\
C_5 &= \int_{a_0}^{b_0} \frac{1}{2} p(x) f^{\beta}(x,\mu_{10},\sigma_{10}^2) \left(\frac{(x-\mu_{10})^3}{\sigma_{10}^4}-\frac{x-\mu_{10}}{\sigma_{10}^2}\right)\, dx,\\
C_6 &= \int_{x\notin(a_0,b_0)} \frac{1}{2} p(x) f^{\beta}(x,\mu_{20},\sigma_{20}^2) \left(\frac{(x-\mu_{20})^3}{\sigma_{20}^4}-\frac{x-\mu_{20}}{\sigma_{20}^2}\right)\, dx,\\
C_7 &= \int_{a_0}^{b_0} p(x) f^{\beta}(x,\mu_{10},\sigma_{10}^2) \frac{(x-\mu_{10})^2}{\sigma_{10}^4}\, dx,\\
C_8 &= \int_{a_0}^{b_0} p(x) f^{\beta}(x,\mu_{10},\sigma_{10}^2) \left(\frac{(x-\mu_{10})^2}{\sigma_{10}^2}-1\right)^2 \frac{1}{2\sigma_{10}^2}\, dx,\\
C_9 &= \int_{x\notin(a_0,b_0)} p(x) f^{\beta}(x,\mu_{20},\sigma_{20}^2) \frac{(x-\mu_{20})^2}{\sigma_{20}^4}\, dx,\\
C_{10} &= \int_{x\notin(a_0,b_0)} p(x) f^{\beta}(x,\mu_{20},\sigma_{20}^2) \left(\frac{(x-\mu_{20})^2}{\sigma_{20}^2}-1\right)^2 \frac{1}{2\sigma_{20}^2}\, dx.
\end{aligned}
$$

2.7.6 Influence of a single contamination on the MDPDEs

To illustrate the effect of moving a single sample observation too far from the original data cloud on the resulting MDPDEs of a mixture normal model, we have taken a random sample of size 99 and dimension 2 from a 3-component normal mixture with component means $(0,0)^t$, $(5,5)^t$, $(-5,-5)^t$, dispersions I_2 (for all the components) and weights 0.33, 0.33 and 0.34. Now, we contaminate this sample with a single additional observation $(\delta,\delta)^t$, with δ running from -15 to 15 in intervals of 0.5, and calculate the (summed) squared L_2 norms of the biases of the component mean estimates (corresponding to $\beta = 0.1$). This summed squared norm bias is plotted in Figure 2.8 against

δ. It may be observed that as δ becomes too large or too small, in which case the additional observation becomes a clearly incongruent observation in relation to the rest of the data, the effect of this outlier actually vanishes, and the summed squared norm of the biases settles down on what one would have had (represented by the horizontal dashed line) if that observation was just not there in the sample. This is a consequence of the fact that for very distant observations, the strong density-power downweighting applied by the minimum DPD method makes the contribution of the corresponding term practically vanish.

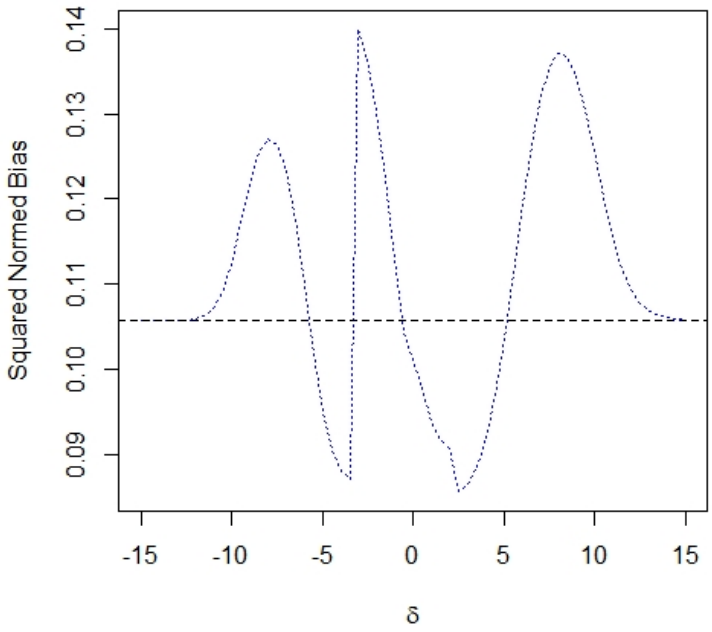


Figure 2.8: The summed squared bias of the MDPDEs of the component means from contaminated samples (blue) and the horizontal red line corresponds to the bias of the MDPDEs of the component means based on the original sample.

2.7.7 Bias and Mean Squared Errors of the Cluster Means

The bias (L_2 norm) and the mean squared errors of the cluster means for pure, uniformly (from chi-squared) contaminated, uniformly (from annulus) contaminated, outlying cluster contaminated and datasets with differentially dispersed clusters are provided in Tables 2.9, 2.10, 2.11, 2.12 and 2.13, respectively.

		MPLE$_\beta$				TCLUST		TKMEANS		KMEDOIDS	MCLUST
p	$\boldsymbol{\Sigma}$	$\beta=0$	$\beta=0.1$	$\beta=0.3$	$\beta=0.5$	$\alpha=0.0$	$\alpha=0.05$	$\alpha=0.0$	$\alpha=0.05$		
2	$\boldsymbol{I}_2$	0.014	0.016	0.019	0.021	0.014	0.019	0.019	0.021	0.014	0.030
		(0.018)	(0.018)	(0.019)	(0.022)	(0.018)	(0.021)	(0.019)	(0.024)	(0.043)	(0.018)
	$3\boldsymbol{I}_2$	0.030	0.037	0.054	0.063	0.032	0.094	0.048	0.058	0.043	0.029
		(0.076)	(0.077)	(0.081)	(0.087)	(0.078)	(0.091)	(0.068)	(0.087)	(0.133)	(0.060)
	$5\boldsymbol{I}_2$	0.086	0.110	0.163	0.194	0.181	0.142	0.168	0.060	0.126	0.082
		(0.279)	(0.379)	(0.296)	(0.310)	(0.219)	(0.320)	(0.154)	(0.178)	(0.288)	(0.146)
4	$\boldsymbol{I}_4$	0.024	0.022	0.021	0.023	0.031	0.029	0.023	0.026	0.097	0.028
		(0.035)	(0.037)	(0.043)	(0.053)	(0.037)	(0.044)	(0.034)	(0.041)	(0.444)	(0.036)
	$3\boldsymbol{I}_4$	0.040	0.040	0.043	0.046	0.042	0.057	0.053	0.059	0.131	0.044
		(0.105)	(0.110)	(0.132)	(0.163)	(0.110)	(0.121)	(0.113)	(0.132)	(1.395)	(0.117)
	$5\boldsymbol{I}_4$	0.067	0.071	0.081	0.092	0.075	0.050	0.094	0.077	0.205	0.043
		(0.230)	(0.230)	(0.257)	(0.307)	(0.222)	(0.225)	(0.227)	(0.252)	(2.423)	(0.197)
6	$\boldsymbol{I}_6$	0.025	0.026	0.028	0.035	0.038	0.031	0.035	0.040	0.193	0.030
		(0.055)	(0.057)	(0.069)	(0.090)	(0.054)	(0.061)	(0.055)	(0.061)	(1.612)	(0.056)
	$3\boldsymbol{I}_6$	0.066	0.067	0.069	0.072	0.056	0.063	0.066	0.073	0.307	0.066
		(0.164)	(0.169)	(0.203)	(0.167)	(0.184)	(0.195)	(0.174)	(0.201)	(4.923)	(0.170)
	$5\boldsymbol{I}_6$	0.046	0.048	0.053	0.063	0.057	0.091	0.059	0.073	0.386	0.062
		(0.253)	(0.281)	(0.341)	(0.443)	(0.281)	(0.309)	(0.285)	(0.311)	(7.821)	(0.270)
8	$\boldsymbol{I}_8$	0.040	0.039	0.043	0.055	0.044	0.045	0.044	0.051	0.162	0.043
		(0.073)	(0.077)	(0.100)	(0.141)	(0.074)	(0.083)	(0.075)	(0.085)	(3.202)	(0.075)
	$3\boldsymbol{I}_8$	0.058	0.059	0.068	0.081	0.073	0.074	0.054	0.056	0.490	0.086
		(0.207)	(0.217)	(0.277)	(0.384)	(0.218)	(0.233)	(0.217)	(0.242)	(10.196)	(0.210)
	$5\boldsymbol{I}_8$	0.087	0.084	0.087	0.101	0.088	0.078	0.085	0.088	0.563	0.100
		(0.381)	(0.396)	(0.504)	(0.711)	(0.382)	(0.431)	(0.369)	(0.415)	(16.049)	(0.344)
10	$\boldsymbol{I}_{10}$	0.054	0.054	0.057	0.065	0.043	0.041	0.040	0.048	0.324	0.041
		(0.088)	(0.094)	(0.129)	(0.201)	(0.089)	(0.100)	(0.090)	(0.100)	(5.240)	(0.095)
	$3\boldsymbol{I}_{10}$	0.058	0.058	0.067	0.089	0.062	0.085	0.081	0.085	0.626	0.062
		(0.258)	(0.272)	(0.375)	(0.600)	(0.272)	(0.310)	(0.260)	(0.299)	(15.654)	(0.259)
	$5\boldsymbol{I}_{10}$	0.09	0.090	0.103	0.128	0.116	0.115	0.071	0.070	0.720	0.082
		(0.498)	(0.523)	(0.704)	(1.112)	(0.492)	(0.498)	(0.482)	(0.531)	(26.962)	(0.418)

Table 2.9: Estimated bias and mean squared errors (within parentheses) for pure datasets.

2.7.8 Optimal choice of β

The optimal selection of the β parameter in the minimum DPD estimation has been explored primarily in simple estimation scenarios by, among others, Warwick and Jones (2005) [155]. They essentially evaluated the performance of the estimator through its asymptotic summed mean squared error (MSE)

$$E\left\{\left(\hat{\boldsymbol{\theta}}_\beta-\boldsymbol{\theta}^*\right)^T\left(\hat{\boldsymbol{\theta}}_\beta-\boldsymbol{\theta}^*\right)\right\}=n^{-1}\mathrm{tr}\left\{\boldsymbol{J}_\beta^{-1}\left(\boldsymbol{\theta}_\beta^g\right)\boldsymbol{K}_\beta\left(\boldsymbol{\theta}_\beta^g\right)\boldsymbol{J}_\beta^{-1}\left(\boldsymbol{\theta}_\beta^g\right)\right\}+\left(\boldsymbol{\theta}_\beta^g-\boldsymbol{\theta}^*\right)^T\left(\boldsymbol{\theta}_\beta^g-\boldsymbol{\theta}^*\right), \tag{2.24}$$

where $\boldsymbol{\theta}_\beta^g$ is the best fitting parameter value and $\boldsymbol{\theta}^*$ is the actual target parameter. If the true unknown density belongs to the model family, $\boldsymbol{\theta}^*=\boldsymbol{\theta}_\beta^g$, the detailed expressions of $\boldsymbol{J}_\beta(\boldsymbol{\theta})$ and $\boldsymbol{K}_\beta(\boldsymbol{\theta})$, the matrices involved in the asymptotic covariance matrix of

		MPLE$_\beta$				TCLUST		TKMEANS		KMEDOIDS	MCLUST
p	$\boldsymbol{\Sigma}$	$\beta=0$	$\beta=0.1$	$\beta=0.3$	$\beta=0.5$	$\alpha=0.1$	$\alpha=0.15$	$\alpha=0.1$	$\alpha=0.15$		
2	$\boldsymbol{I}_2$	0.101	0.041	0.013	0.014	0.022	0.041	0.016	0.018	0.077	0.016
		(0.076)	(0.036)	(0.023)	(0.024)	(0.024)	(0.026)	(0.020)	(0.022)	(0.107)	(0.025)
	$3\boldsymbol{I}_2$	0.106	0.076	0.039	0.039	0.044	0.027	0.068	0.078	0.093	0.056
		(0.154)	(0.129)	(0.095)	(0.091)	(0.100)	(0.089)	(0.094)	(0.097)	(0.501)	(0.086)
	$5\boldsymbol{I}_2$	0.159	0.110	0.061	0.051	0.126	0.144	0.129	0.066	0.056	0.194
		(0.309)	(0.275)	(0.245)	(0.261)	(0.352)	(0.310)	(0.253)	(0.249)	(1.193)	(0.228)
4	$\boldsymbol{I}_4$	0.058	0.028	0.027	0.030	0.020	0.033	0.033	0.028	0.120	0.028
		(0.060)	(0.037)	(0.042)	(0.052)	(0.046)	(0.051)	(0.039)	(0.044)	(0.471)	(0.040)
	$3\boldsymbol{I}_4$	0.061	0.031	0.033	0.043	0.044	0.046	0.065	0.060	0.166	0.051
		(0.206)	(0.127)	(0.119)	(0.143)	(0.124)	(0.145)	(0.119)	(0.132)	(1.464)	(0.119)
	$5\boldsymbol{I}_4$	0.114	0.092	0.081	0.080	0.054	0.073	0.062	0.079	0.295	0.071
		(0.352)	(0.272)	(0.256)	(0.301)	(0.252)	(0.274)	(0.253)	(0.265)	(2.685)	(0.256)
6	$\boldsymbol{I}_6$	0.049	0.037	0.039	0.046	0.037	0.033	0.037	0.036	0.152	0.032
		(0.173)	(0.065)	(0.078)	(0.099)	(0.069)	(0.072)	(0.066)	(0.072)	(1.697)	(0.061)
	$3\boldsymbol{I}_6$	0.066	0.054	0.057	0.063	0.067	0.059	0.048	0.054	0.394	0.061
		(0.310)	(0.216)	(0.247)	(0.307)	(0.187)	(0.207)	(0.187)	(0.222)	(5.234)	(0.186)
	$5\boldsymbol{I}_6$	0.097	0.094	0.102	0.116	0.088	0.073	0.068	0.067	0.367	0.074
		(0.438)	(0.364)	(0.421)	(0.532)	(0.314)	(0.367)	(0.312)	(0.335)	(7.889)	(0.315)
8	$\boldsymbol{I}_8$	0.094	0.030	0.033	0.038	0.038	0.039	0.044	0.034	0.284	0.051
		(0.220)	(0.088)	(0.112)	(0.156)	(0.092)	(0.088)	(0.088)	(0.089)	(3.513)	(0.092)
	$3\boldsymbol{I}_8$	0.086	0.062	0.061	0.080	0.050	0.073	0.089	0.085	0.452	0.082
		(0.363)	(0.260)	(0.336)	(0.467)	(0.257)	(0.280)	(0.278)	(0.282)	(10.232)	(0.257)
	$5\boldsymbol{I}_8$	0.138	0.113	0.118	0.133	0.109	0.197	0.103	0.103	0.655	0.085
		(0.525)	(0.411)	(0.510)	(0.715)	(0.411)	(0.444)	(0.429)	(0.454)	(17.489)	(0.413)
10	$\boldsymbol{I}_{10}$	0.066	0.040	0.046	0.055	0.042	0.046	0.054	0.051	0.267	0.046
		(0.308)	(0.107)	(0.140)	(0.213)	(0.111)	(0.113)	(0.108)	(0.110)	(5.576)	(0.104)
	$3\boldsymbol{I}_{10}$	0.086	0.068	0.078	0.103	0.089	0.086	0.088	0.088	0.474	0.075
		(0.437)	(0.306)	(0.401)	(0.613)	(0.318)	(0.319)	(0.329)	(0.345)	(16.821)	(0.291)
	$5\boldsymbol{I}_{10}$	0.122	0.115	0.129	0.168	0.085	0.090	0.119	0.118	0.819	0.112
		(0.645)	(0.537)	(0.720)	(1.101)	(0.545)	(0.570)	(0.497)	(0.554)	(27.373)	(0.511)

Table 2.10: Estimated bias and mean squared errors (within parentheses) for uniformly (chi-squared method) contaminated datasets.

the minimum DPD estimator with tuning parameter β, can be found in Basu et al. (1998), Warwick and Jones (2005) [155] or Basak et al. (2021) [9]. Now, there are two unknown quantities in Equation (2.24), namely, $\boldsymbol{\theta}_\beta^g$ and $\boldsymbol{\theta}^*$. In practice, Warwick and Jones (2005) [155] suggested $\boldsymbol{\theta}_\beta^g$ to be replaced by $\hat{\boldsymbol{\theta}}_\beta$, the MDPDE of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$ to be replaced by some suitable robust pilot estimator $\hat{\boldsymbol{\theta}}^P$. Later, Basak et al. (2021) suggested an iterative procedure to find the optimal choice of β. This is basically a refinement of the procedure proposed by Warwick and Jones (2005) [155] which chooses the optimal parameter estimate at any particular stage as the pilot estimate for the next stage and continues the process till convergence.

Any of the approaches above can be adapted to the clustering situation considered by us. However, there is one issue involved here. The literature has so far considered

		MPLE$_\beta$				TCLUST		TKMEANS		KMEDOIDS	MCLUST
p	$\boldsymbol{\Sigma}$	$\beta=0$	$\beta=0.1$	$\beta=0.3$	$\beta=0.5$	$\alpha=0.1$	$\alpha=0.15$	$\alpha=0.1$	$\alpha=0.15$		
2	$\boldsymbol{I}_2$	0.794	0.035	0.038	0.040	0.072	0.022	0.051	0.021	0.135	0.014
		(0.810)	(0.021)	(0.023)	(0.026)	(0.059)	(0.023)	(0.030)	(0.027)	(0.125)	(0.020)
	$3\boldsymbol{I}_2$	0.697	0.174	0.061	0.076	0.514	0.060	0.123	0.061	0.237	0.064
		(0.749)	(0.255)	(0.085)	(0.093)	(1.741)	(0.109)	(0.098)	(0.082)	(0.448)	(0.077)
	$5\boldsymbol{I}_2$	0.778	0.311	0.254	0.293	1.962	0.247	0.300	0.053	0.429	0.070
		(1.151)	(0.513)	(0.393)	(0.402)	(37.672)	(0.439)	(0.319)	(0.212)	(0.863)	(0.173)
4	$\boldsymbol{I}_4$	0.057	0.026	0.028	0.030	0.048	0.026	0.054	0.019	0.118	0.031
		(0.225)	(0.040)	(0.045)	(0.054)	(0.046)	(0.045)	(0.053)	(0.051)	(0.516)	(0.044)
	$3\boldsymbol{I}_4$	0.131	0.078	0.069	0.076	0.093	0.069	0.047	0.043	0.141	0.045
		(0.317)	(0.132)	(0.145)	(0.171)	(0.132)	(0.134)	(0.129)	(0.137)	(1.485)	(0.122)
	$5\boldsymbol{I}_4$	0.219	0.094	0.083	0.088	0.105	0.067	0.086	0.059	0.237	0.084
		(0.524)	(0.269)	(0.246)	(0.277)	(0.248)	(0.292)	(0.241)	(0.262)	(2.275)	(0.248)
6	$\boldsymbol{I}_6$	0.050	0.045	0.049	0.056	0.036	0.041	0.036	0.036	0.136	0.037
		(0.250)	(0.061)	(0.073)	(0.095)	(0.066)	(0.073)	(0.065)	(0.069)	(1.545)	(0.063)
	$3\boldsymbol{I}_6$	0.096	0.059	0.069	0.082	0.059	0.068	0.069	0.080	0.322	0.062
		(0.357)	(0.185)	(0.229)	(0.297)	(0.188)	(0.207)	(0.185)	(0.213)	(4.988)	(0.195)
	$5\boldsymbol{I}_6$	0.127	0.069	0.089	0.107	0.082	0.080	0.110	0.107	0.357	0.095
		(0.435)	(0.287)	(0.344)	(0.453)	(0.342)	(0.352)	(0.346)	(0.383)	(8.597)	(0.319)
8	$\boldsymbol{I}_8$	0.059	0.045	0.047	0.050	0.034	0.032	0.049	0.035	0.191	0.042
		(0.305)	(0.081)	(0.105)	(0.146)	(0.087)	(0.091)	(0.089)	(0.086)	(3.279)	(0.083)
	$3\boldsymbol{I}_8$	0.091	0.072	0.076	0.075	0.081	0.067	0.090	0.092	0.489	0.065
		(0.394)	(0.247)	(0.304)	(0.415)	(0.248)	(0.277)	(0.248)	(0.266)	(10.136)	(0.243)
	$5\boldsymbol{I}_8$	0.088	0.093	0.107	0.127	0.102	0.100	0.084	0.083	0.512	0.097
		(0.548)	(0.429)	(0.543)	(0.771)	(0.392)	(0.458)	(0.411)	(0.465)	(16.497)	(0.426)
10	$\boldsymbol{I}_{10}$	0.076	0.042	0.051	0.064	0.048	0.046	0.055	0.051	0.305	0.048
		(0.279)	(0.109)	(0.148)	(0.231)	(0.119)	(0.119)	(0.113)	(0.116)	(5.410)	(0.010)
	$3\boldsymbol{I}_{10}$	0.093	0.064	0.073	0.098	0.069	0.079	0.093	0.100	0.539	0.080
		(0.451)	(0.289)	(0.384)	(0.592)	(0.330)	(0.331)	(0.304)	(0.343)	(16.047)	(0.303)
	$5\boldsymbol{I}_{10}$	0.101	0.090	0.088	0.115	0.098	0.115	0.099	0.102	0.836	0.112
		(0.612)	(0.492)	(0.641)	(0.998)	(0.507)	(0.560)	(0.571)	(0.601)	(27.712)	(0.513)

Table 2.11: Estimated bias and mean squared errors (within parentheses) for uniformly (from annulus) contaminated datasets.

the estimation of an optimal β parameter for one sample of real data. Our situation is a little more complex than that. Here one has to estimate the parameters of each of the k populations, separately, and then again this process has to be repeated in each iteration. Thus, tuning parameter selection has to be performed separately for each population at each iteration. Therefore it does not make sense to talk about a single optimal value of β, as these cases may result in distinct optimals. Otherwise there is no conceptual difficulty in optimal tuning parameter selection in our clustering problem. However, we believe that the implementation of this should be reserved for a future work, given that the level of refinement is quite intricate.

		MPLE$_\beta$				TCLUST		TKMEANS		KMEDOIDS	MCLUST
p	$\boldsymbol{\Sigma}$	$\beta=0$	$\beta=0.1$	$\beta=0.3$	$\beta=0.5$	$\alpha=0.1$	$\alpha=0.15$	$\alpha=0.1$	$\alpha=0.15$		
2	$\boldsymbol{I}_2$	4.870	0.028	0.029	0.030	2.232	0.025	1.424	0.021	20.888	22.152
		(23.97)	(0.018)	(0.021)	(0.024)	(47.236)	(0.028)	(27.977)	(0.027)	(457.109)	(497.0417)
	$3\boldsymbol{I}_2$	4.876	4.487	0.062	0.070	10.536	0.427	3.187	0.053	21.928	21.762
		(24.141)	(20.599)	(0.087)	(0.094)	(233.780)	(10.392)	(68.016)	(0.093)	(482.016)	(482.281)
	$5\boldsymbol{I}_2$	9.712	10.096	0.990	0.566	21.899	20.053	8.837	0.064	26.925	21.938
		(145.236)	(158.673)	(28.353)	(0.685)	(485.105)	(461.605)	(191.516)	(0.189)	(491.365)	(486.658)
4	$\boldsymbol{I}_4$	6.860	0.025	0.030	0.037	2.661	0.030	3.249	0.030	29.363	31.221
		(47.878)	(0.040)	(0.045)	(0.054)	(77.778)	(0.050)	(98.839)	(0.052)	(906.572)	(987.839)
	$3\boldsymbol{I}_4$	6.881	6.310	0.055	0.059	17.412	0.051	5.772	0.048	30.596	30.730
		(47.980)	(44.003)	(0.131)	(0.155)	(547.829)	(0.145)	(172.106)	(0.144)	(941.274)	(959.393)
	$5\boldsymbol{I}_4$	6.974	7.079	0.089	0.093	28.472	21.841	6.983	0.047	30.767	30.856
		(49.457)	(51.252)	(0.239)	(0.283)	(892.367)	(689.382)	(217.568)	(0.243)	(951.243)	(968.882)
6	$\boldsymbol{I}_6$	8.757	0.033	0.038	0.041	5.511	0.033	0.993	0.036	36.762	38.104
		(109.060)	(0.064)	(0.076)	(0.097)	(204.188)	(0.067)	(31.582)	(0.068)	(1393.215)	(1472.076)
	$3\boldsymbol{I}_6$	8.343	6.886	0.063	0.065	15.997	0.067	1.784	0.054	37.819	38.054
		(75.015)	(66.517)	(0.208)	(0.275)	(617.990)	(0.195)	(61.955)	(0.196)	(1444.099)	(0.172)
	$5\boldsymbol{I}_6$	7.859	9.443	0.320	0.128	30.467	13.672	6.331	0.076	37.696	0.061
		(75.932)	(90.913)	(6.551)	(0.495)	(976.222)	(526.805)	(238.451)	(0.356)	(1432.805)	(1470.286)
8	$\boldsymbol{I}_8$	14.642	0.042	0.054	0.059	9.995	0.044	4.623	0.042	42.513	43.710
		(215.859)	(0.086)	(0.108)	(0.150)	(430.428)	(0.088)	(188.793)	(0.087)	(1872.28)	(1937.060)
	$3\boldsymbol{I}_8$	14.509	4.952	0.095	0.117	13.260	0.061	7.295	0.069	43.359	43.846
		(214.399)	(61.377)	(0.307)	(0.429)	(591.215)	(0.272)	(321.262)	(0.249)	(1903.009)	(1951.052)
	$5\boldsymbol{I}_8$	14.555	11.787	0.099	0.114	33.417	5.136	6.360	0.103	43.435	43.567
		(213.653)	(143.740)	(0.551)	(0.758)	(1191.519)	(225.505)	(279.669)	(0.571)	(1909.009)	(1926.959)
10	$\boldsymbol{I}_{10}$	16.282	0.047	0.057	0.070	5.188	0.045	5.211	0.048	45.457	48.981
		(267.572)	(0.103)	(0.138)	(0.218)	(242.153)	(0.117)	(248.064)	(0.109)	(2232.301)	(2432.958)
	$3\boldsymbol{I}_{10}$	16.293	7.749	0.076	0.092	18.751	0.084	5.091	0.078	48.279	49.325
		(268.224)	(116.158)	(0.396)	(0.616)	(935.354)	(0.328)	(236.564)	(0.408)	(2370.414)	(2466.566)
	$5\boldsymbol{I}_{10}$	16.357	14.035	0.119	0.163	25.640	5.964	14.062	0.110	48.664	48.745
		(272.519)	(231.636)	(0.620)	(1.092)	(1281.557)	(298.761)	(698.405)	(0.550)	(2408.094)	(2411.771)

Table 2.12: Estimated bias and mean squared errors (within parentheses) for outlying cluster contaminated datasets.

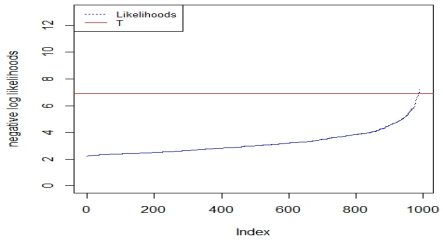
2.7.9 Determination of T

We choose the optimal value of T following the maximal-gap philosophy. To understand this, we present the plots of the negative log likelihood values of the sample observations along with the choice of $-\log(T)$ in Figure 2.9. The figure clearly indicates how the cut-offs correspond to maximal gaps. If there is no contamination in the data, normality of the clusters suggests that the concentration of the sample observations will be more in the central region and less in the low probability zones which are distant from the cluster means. Thus, the maximum gap is expected to be found in the low probability zones containing a few extreme observations. Hence, in case of no contamination, the maximal-gap strategy will, in most cases, find the threshold far away from the central region and the number of declared outliers will be small, as one would ideally expect.

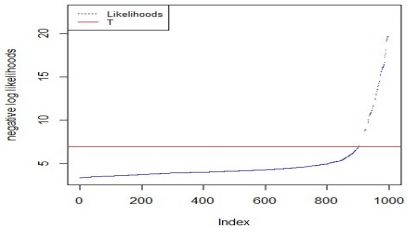
		MPLE$_\beta$				TCLUST		TKMEANS		KMEDOIDS	MCLUST
Type	p	$\beta = 0$	$\beta = 0.1$	$\beta = 0.3$	$\beta = 0.5$	$\alpha = 0$	$\alpha = 0.05$	$\alpha = 0$	$\alpha = 0.05$		
Pure	2	0.018 (0.027)	0.019 (0.027)	0.026 (0.030)	0.032 (0.034)	0.025 (0.027)	0.027 (0.034)	0.108 (0.044)	0.025 (0.054)	0.064 (0.075)	0.015 (0.030)
Pure	6	0.057 (0.084)	0.057 (0.087)	0.058 (0.107)	0.061 (0.141)	0.049 (0.105)	0.058 (0.123)	0.044 (0.098)	0.045 (0.115)	0.170 (2.422)	0.037 (0.086)
						α =0.10	α =0.15	α =0.10	α =0.15		
Uniform (chi-squared) Contaminated	2	0.136 (0.134)	0.094 (0.098)	0.060 (0.041)	0.042 (0.043)	0.030 (0.042)	0.047 (0.040)	0.135 (0.064)	0.062 (0.061)	0.102 (0.353)	0.042 (0.047)
Uniform (chi-squared) Contaminated	6	0.077 (0.251)	0.114 (0.115)	0.053 (0.121)	0.066 (0.154)	0.073 (0.125)	0.055 (0.123)	0.049 (0.113)	0.044 (0.130)	0.249 (2.639)	0.023 (0.101)
Uniform (Annulus) Contaminated	2	0.586 (0.966)	0.820 (0.940)	0.018 (0.034)	0.025 (0.037)	0.142 (0.093)	0.048 (0.047)	0.185 (0.106)	0.030 (0.062)	0.290 (0.285)	0.032 (0.040)
Uniform (Annulus) Contaminated	6	0.087 (0.357)	0.037 (0.107)	0.044 (0.126)	0.049 (0.156)	0.040 (0.107)	0.045 (0.116)	0.057 (0.112)	0.050 (0.123)	0.193 (2.653)	0.061 (0.102)
Outlying Cluster	2	4.911 (24.337)	4.538 (20.865)	0.018 (0.040)	0.017 (0.044)	6.0677 (129.004)	0.037 (0.050)	5.318 (110.232)	0.064 (0.057)	21.599 (467.394)	21.505 (462.341)
Outlying Cluster	6	8.622 (83.850)	6.970 (70.038)	0.039 (0.127)	0.047 (0.158)	9.707 (358.676)	0.048 (0.129)	6.810 (257.010)	0.074 (0.128)	36.805 (1389.584	37.312 (1392.408)

Table 2.13: Estimated bias and mean squared errors (within parentheses) for datasets with differentially dispersed clusters.

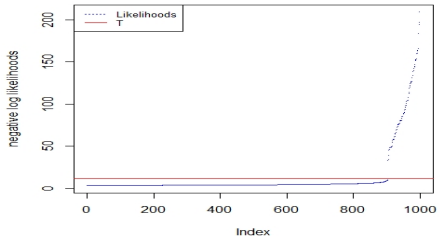
As we have mentioned in the main manuscript earlier, there could be situations where the performance of the maximal-gap strategy may not be fully satisfactory. Using the t-th largest gap (rather than the overall largest) may provide a way around in such cases but the optimal choice of this T may itself pose a major challenge. All in all, the choice of the tuning parameter T remains a difficult problem for which a fully satisfactory solution, covering all different cases, is not yet available.



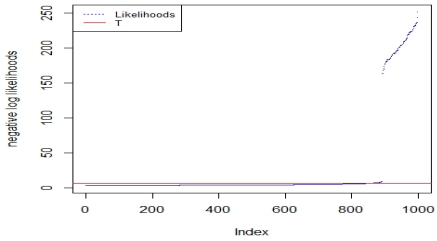
(a) Pure Data



(b) Uniformly contaminated (χ^2 type) data



(c) Annulus contaminated Data



(d) Outlying cluster Contaminated Data

Figure 2.9: Plots of the negative log likelihoods of sample observations with the optimal choice of $-\log(T)$ as outlier thresholds.

.

Chapter 3

Theoretical Properties of the Maximum Pseudo β-Likelihood Estimation

3.1 Introduction

The MPLE$_\beta$ algorithm has been developed in Chapter 2 for performing robust estimation in the multivariate normal mixture set-up with subsequent robust clustering and anomaly detection. The proposed method is validated with simulation experiments and real data applications (comparative study with the existing popular robust clustering methods). The robustness of the algorithm is assessed through the influence function analysis. However, theoretical validation of the proposed methodology is required to understand the correctness of the algorithm. We present some important theoretical properties of the MPLE$_\beta$ method in the present chapter. In particular, we establish the existence and consistency of the parameter estimates obtained from the MPLE$_\beta$ algorithm in case of the multivariate normal mixture models under a set of sufficient conditions. Before going to prove these theoretical properties, let us recall the probabilistic set-up of the MPLE$_\beta$ method once again. As earlier, let, $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n\}$ be a random sample drawn from a p-dimensional multivariate normal mixture distribution having k components with true unknown probability measure F. The corresponding true unknown PDF is modelled by the family $\{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, with,

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{j=1}^{k} \pi_j \phi_p(\boldsymbol{x}, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j), \tag{3.1}$$

where $\boldsymbol{\theta} = (\pi_1, \pi_2, \ldots, \pi_k, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \ldots, \boldsymbol{\Sigma}_k)$ is the parameter vector of interest. Let us also recall the empirical objective function (same as Equation (2.5))

$$L_\beta(\boldsymbol{\theta}, F_n) = E_{F_n}\left[\sum_{j=1}^{k} Z_j(\boldsymbol{X}, \boldsymbol{\theta}) \left[\log\ \pi_j + \frac{1}{\beta}\phi_p^\beta(\boldsymbol{X}, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) - \frac{1}{1+\beta}\int \phi_p^{1+\beta}(\boldsymbol{x}, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\, d\boldsymbol{x}\right]\right], \tag{3.2}$$

which is maximized with respect to $\boldsymbol{\theta}$ over the restricted parameter space $\boldsymbol{\Theta}_C$ (Equation (2.8)) assuming the ER and NS constraints. The theoretical objective function (same as Equation (2.6))

$$L_\beta(\boldsymbol{\theta}, F) = E_F\left[\sum_{j=1}^{k} Z_j(\boldsymbol{X}, \boldsymbol{\theta})\left[\log\ \pi_j + \frac{1}{\beta}\phi_p^\beta(\boldsymbol{X}, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) - \frac{1}{1+\beta}\int \phi_p^{1+\beta}(\boldsymbol{x}, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\, d\boldsymbol{x}\right]\right] \tag{3.3}$$

is now crucial to consider for the formal presentation of the theoretical results.

The remainder of the chapter is organized as follows. Section 3.2 provides the detailed proofs of the two main results on existence and consistency. Section 3.3 provides simulation experiments where our method is compared with other alternatives in terms of bias and mean squared errors of the component mean and covariance matrix estimates. The application of our method is further illustrated in Section 3.4 on two real datasets. Section 3.5 contains a few relevant lemmas.

3.2 Theoretical Results

Let us first state the following technical assumptions which will be crucial to prove some of our results.

Assumption 3.1. *For the true (unknown) density function* $f = \sum_{j=1}^{k} \pi_j^0 \phi_p(\cdot, \boldsymbol{\mu_j^0}, \boldsymbol{\Sigma_j^0})$ *(true probability measure* $F = \sum_{j=1}^{k} \pi_j^0 F_j(\cdot, \boldsymbol{\mu_j^0}, \boldsymbol{\Sigma_j^0})$*), let us assume that,* $\max\limits_{1\le j\le k} \pi_j^0 \ge (1+k^0)\frac{\beta}{(1+\beta)^{1+\frac{p}{2}}}$ *for some* $k^0 > 0$.

Assumption 3.2. $\min\limits_{1\le j\le k} \pi_j \ge \pi_{min} > 0$, *i.e., none of the cluster components are empty.*

Remark 3.1. *If* $\pi_j = 0$, *i.e., the* j*-th component is empty, it will have no role in the data generation process. As the number of components is finite, it is therefore reasonable to assume that each component probability is bounded away from zero. Following Assumption 3.2 and the NS constraint, we have*

$$-\infty < \log\ \pi_{min} \le \log\ \pi_j < 0 \text{ and } 0 < \frac{1}{|\boldsymbol{\Sigma}_j|} \le \frac{1}{c_1^p} < \infty,\ \forall j. \tag{3.4}$$

We first present the result on the existence of the optimizers (both empirical and population versions) and then the consistency of the resulting estimators under the multivariate normal mixture model.

Theorem 3.1 (Existence)**.** *Let P be a probability distribution (either F or F_n in our case) and Assumption 3.1 is satisfied by F. Then, there exists $\boldsymbol{\theta} \in \boldsymbol{\Theta}_C$ that maximizes*

$$L_\beta(\boldsymbol{\theta}, P) = E_P \left[\sum_{j=1}^{k} Z_j(\boldsymbol{X}, \boldsymbol{\theta}) \left[\log\ \pi_j + \frac{1}{\beta} \phi_p^\beta(\boldsymbol{X}, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) - \frac{1}{1+\beta} \int \phi_p^{1+\beta}(\boldsymbol{x}, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\, d\boldsymbol{x} \right] \right]$$

under the ER and NS constraints (for a sufficiently large sample size n in case of F_n).

Proof. Let $\{\boldsymbol{\theta}^r\} = \{(\pi_1^r, \pi_2^r, \ldots, \pi_k^r, \boldsymbol{\mu}_1^r, \boldsymbol{\mu}_2^r, \ldots, \boldsymbol{\mu}_k^r, \boldsymbol{\Sigma}_1^r, \boldsymbol{\Sigma}_2^r, \ldots, \boldsymbol{\Sigma}_k^r)\}$ be a sequence in $\boldsymbol{\Theta}_C$ such that,

$$\lim_{r\to\infty} L_\beta(\boldsymbol{\theta}^r, P) = \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_C} L_\beta(\boldsymbol{\theta}, P). \tag{3.5}$$

Let us assume, without loss of generality, that, $\pi_1^0 = \max\limits_{1\le j\le k} \pi_j^0$. Let, $\boldsymbol{\theta}^a = (\pi_1^a, \pi_2^a, \ldots, \pi_k^a, \boldsymbol{\mu}_1^a, \boldsymbol{\mu}_2^a, \ldots, \boldsymbol{\mu}_k^a, \boldsymbol{\Sigma}_1^a, \boldsymbol{\Sigma}_2^a, \ldots, \boldsymbol{\Sigma}_k^a) \in \boldsymbol{\Theta}_C$ such that,

$$\pi_j^a = 1\ ,\ \boldsymbol{\mu}_j^a = \boldsymbol{\mu}_j^0,\ \boldsymbol{\Sigma}_j^a = c^{'}\boldsymbol{\Sigma}_j^0 \text{ for } j = 1,$$

where $c^{'}$ is a positive real; choice of $c^{'}$ will be fixed appropriately. This implies that,

$$Z_j(\boldsymbol{X}, \boldsymbol{\theta}^a) = \begin{cases} 1, & \text{if } j = 1 \\ 0, & \text{if } 2 \le j \le k. \end{cases}$$

Hence,

$$\lim_{r\to\infty} L_\beta(\boldsymbol{\theta}^r, P) = \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_C} L_\beta(\boldsymbol{\theta}, P) \ge L_\beta(\boldsymbol{\theta}^a, P).$$

If P is the true unknown distribution function (i.e., F), then,

$$\begin{aligned} L_\beta(\boldsymbol{\theta}^a, P) &= E_F\left(\frac{1}{\beta}\phi_p^\beta(\boldsymbol{X}, \boldsymbol{\mu}_1^0, c^{'}\boldsymbol{\Sigma}_1^0)\right) - \frac{1}{1+\beta}\int \phi_p^{1+\beta}(\boldsymbol{x}, \boldsymbol{\mu}_1^0, c^{'}\boldsymbol{\Sigma}_1^0)\, d\boldsymbol{x} \\ &\ge \pi_1^0 E_{F_1}\left(\frac{1}{\beta}\phi_p^\beta(\boldsymbol{X}, \boldsymbol{\mu}_1^0, c^{'}\boldsymbol{\Sigma}_1^0)\right) - \frac{1}{1+\beta}\int \phi_p^{1+\beta}(\boldsymbol{x}, \boldsymbol{\mu}_1^0, c^{'}\boldsymbol{\Sigma}_1^0)\, d\boldsymbol{x} \\ &= \frac{1}{\beta(2\pi)^{\frac{p\beta}{2}}|c^{'}\boldsymbol{\Sigma}_1^0|^{\frac{\beta}{2}}(1+\frac{\beta}{c^{'}})^{\frac{p}{2}}}\left[\pi_1^0 - \frac{\beta}{(1+\beta)^{1+\frac{p}{2}}}\left(1+\frac{\beta}{c^{'}}\right)^{\frac{p}{2}}\right] \text{ (after some algebra).} \end{aligned}$$

Now, the positivity of the aforesaid term can be achieved by taking a large enough $c^{'}$

(as a consequence of Assumption 3.1), and thus

$$L_\beta(\boldsymbol{\theta}^a, P) \geq 0.$$

But if P is the empirical distribution F_n (n represents the sample size),

$$L_\beta(\boldsymbol{\theta}^a, P) = \frac{1}{n\beta}\sum_{i=1}^{n}\phi_p^\beta(\boldsymbol{X}_i, \boldsymbol{\mu_1^0}, c'\boldsymbol{\Sigma_1^0}) - \frac{1}{1+\beta}\int \phi_p^{1+\beta}(\boldsymbol{x}, \boldsymbol{\mu_1^0}, c'\boldsymbol{\Sigma_1^0})\, d\boldsymbol{x}.$$

The positivity of the above quantity can be established by an application of the strong law of large numbers (SLLN) assuming a moderately large sample size n followed by the aforesaid argument in case of $P = F$. So, the sequence $\{\boldsymbol{\theta}^r\}$ satisfies,

$$\lim_{r\to\infty} L_\beta(\boldsymbol{\theta}^r, P) \geq 0. \tag{3.6}$$

Since $(\pi_1^r, \pi_2^r, \ldots, \pi_k^r) \in [0,1]^k$ and $[0,1]^k$ is a compact set in $\mathbb{R}^k$, the sequence $\{\boldsymbol{\theta}^r\}$ has a subsequence $\{\boldsymbol{\theta}^r\}^l$ such that $\{\pi_1^r, \pi_2^r, \ldots, \pi_k^r\}^l$ is convergent. To simplify the notation, we will denote this subsequence $\{\boldsymbol{\theta}^r\}^l$ as the original sequence $\{\boldsymbol{\theta}^r\}$.

Hence the sequence must satisfy the following properties.

1. For the proportion sequence $\{\pi_1^r, \pi_2^r, \ldots, \pi_k^r\}$,

$$\pi_j^r \to \pi_j \in [0,1] \text{ for } 1 \leq j \leq k. \tag{3.7}$$

2. For the mean sequence $\{\boldsymbol{\mu}_1^r, \boldsymbol{\mu}_2^r, \ldots, \boldsymbol{\mu}_k^r\}$,

$$\boldsymbol{\mu}_j^r \to \boldsymbol{\mu}_j \in \mathbb{R}^p \text{ for } 1 \leq j \leq g \text{ and } ||\boldsymbol{\mu}_j^r|| \to \infty \text{ for } g+1 \leq j \leq k \text{ for some } 0 \leq g \leq k, \tag{3.8}$$

   where $||\cdot||$ denotes the standard L_2 norm.

3. Finally, the dispersion sequence $\{\boldsymbol{\Sigma}_1^r, \boldsymbol{\Sigma}_2^r, \ldots, \boldsymbol{\Sigma}_k^r\}$ must satisfy exactly one of the following conditions. Either,

$$\boldsymbol{\Sigma}_j^r \to \boldsymbol{\Sigma}_j \in \mathbb{R}^{p\times p} \text{ for } 1 \leq j \leq k, \tag{3.9}$$

or,

$$M_r \to \infty, \tag{3.10}$$

or,

$$m_r \to 0, \tag{3.11}$$

where M_r and m_r are the largest and the smallest elements of the set of eigenvalues of $\boldsymbol{\Sigma}_1^r, \boldsymbol{\Sigma}_2^r, \ldots, \boldsymbol{\Sigma}_k^r$, respectively.

Now, by Lemma 3.4, presented in Section 3.5, the condition in (3.8) holds for $g = k$ in case of component means (if $\pi_j > 0$ for all $j = 1, 2, \ldots, k$ in (3.7)) and (3.9) holds for the component covariance matrices. Thus, if $\pi_j > 0$ for all $j = 1, 2, \ldots, k$ in (3.7), then, the choice of the optimizer is obvious. But, if $\pi_j > 0$ for $j = 1, 2, \ldots, g$ for some $1 \leq g < k$, and $\pi_j = 0$ for $j > g$, then take $\pi_j = \lim_{r\to\infty} \pi_j^r$ for $j = 1, 2, \ldots, g$ and $\pi_j = 0$ for $j > g$, $\boldsymbol{\mu}_j = \lim_{r\to\infty} \boldsymbol{\mu}_j^r$, $\boldsymbol{\Sigma}_j = \lim_{r\to\infty} \boldsymbol{\Sigma}_j^r$ for $j = 1, 2, \ldots, g$ and $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ arbitrarily (satisfying the ER and NS constraints) for $j > g$. These values will provide the maximizer of the objective function in consideration. This completes the proof of Theorem 3.1. □

Our next theorem provides consistency properties of the MPLE$_\beta$ (that is, the consistency of the sample version (optimizer in Equation (3.2)) to the population version (optimizer in Equation (3.3)). To achieve this, we need uniqueness of the maximizer of (3.3) under the ER and NS constraints and the following prerequisite result in order to establish the consistency.

Theorem 3.2 (Corollary 3.2.3, van der Vaart and Wellner (1996) [151])**.** *Let M_n be a stochastic process indexed by a metric space $\boldsymbol{\Theta}$ and let $M : \boldsymbol{\Theta} \to \mathbb{R}$ be a deterministic function. Suppose the following conditions hold.*

1. *Suppose that $||M_n - M||_{\boldsymbol{\Theta}} \to 0$ in probability.*
2. *There exists a point $\boldsymbol{\theta}_0$ such that, $M(\boldsymbol{\theta}_0) > \sup_{\boldsymbol{\theta}\notin G} M(\boldsymbol{\theta})$ for every open set G that contains $\boldsymbol{\theta}_0$.*
3. *Suppose a sequence $\hat{\boldsymbol{\theta}}_n$ satisfies $M_n(\hat{\boldsymbol{\theta}}_n) > \sup_{\boldsymbol{\theta}} M_n(\boldsymbol{\theta}) - o_p(1)$.*

Then, $\hat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}_0$ in probability.

Theorem 3.3 (Consistency)**.** *Suppose that $\boldsymbol{\theta}_0$ be the unique maximizer of (3.3) subject to the ER and NS constraints and suppose Assumptions 3.1 and 3.2 hold. Then, if $\hat{\boldsymbol{\theta}}_n$ is a maximizer of (3.2) based on a sample of size n, we have $\hat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}_0$ in probability as $n \to \infty$.*

Proof. In order to derive the estimators, we had maximized the objective function $L_\beta(\boldsymbol{\theta}, F_n)$ which is not differentiable with respect to the parameter $\boldsymbol{\theta}$. Hence, the standard Taylor series expansion approach (used to derive the asymptotics of maximum likelihood estimators) may not work for this problem. Thus, we are going to use the modern empirical process tools (Theorem 3.2) to establish weak consistency.

Following the notations of Theorem 3.2, we have, $M_n(\boldsymbol{\theta}) = L_\beta(\boldsymbol{\theta}, F_n) = E_{F_n}(m_{\boldsymbol{\theta}}(\boldsymbol{X}))$ and $M(\boldsymbol{\theta}) = E_F(m_{\boldsymbol{\theta}}(\boldsymbol{X}))$ with

$$m_{\boldsymbol{\theta}}(\boldsymbol{X}) = \sum_{j=1}^{k} Z_j(\boldsymbol{X}, \boldsymbol{\theta}) \left[\log\ \pi_j + \frac{1}{\beta} \phi_p^\beta(\boldsymbol{X}, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) - \frac{1}{1+\beta} \int \phi_p^{1+\beta}(\boldsymbol{x}, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\, d\boldsymbol{x} \right].$$

The third condition of Theorem 3.2 is satisfied due to Theorem 3.1. The second condition of Theorem 3.2 is also satisfied due to the assumption of the existence of a unique maximizer of the theoretical objective function in Equation (3.3). Thus, we have to check the first condition only. To verify this, we need a Glivenko-Cantelli (GC) property (van der Vaart and Wellner (1996) [151]) of the class $\mathcal{F} = \{m_{\boldsymbol{\theta}}(\boldsymbol{X}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}_C\}$.

To do that, first let us observe the fact (van der Vaart and Wellner (1996) [151]) that any appropriately measurable Vapnik-Červonenkis (VC) class is Glivenko-Cantelli (GC) provided its envelope function is integrable with respect to the underlying probability measure (the true probability measure F in our case). Hence, it is enough to show that, $\mathcal{F}$ is VC with an integrable envelope function.

To establish that $\mathcal{F}$ is GC, we will follow the methodologies developed in Section 2.6 of van der Vaart and Wellner (1996) [151] and Kosorok (2008) [88]. Let us observe the following facts.

- The functions $(\boldsymbol{X} - \boldsymbol{\mu}_j)' \boldsymbol{\Sigma}_j^{-1} (\boldsymbol{X} - \boldsymbol{\mu}_j)$ are polynomials of degree 2. Hence, these functions together form a finite dimensional vector space and hence is VC.
- The function $\phi(x) = e^{-x}$ is monotone and continuous hence $\phi \circ \mathcal{G}$ is VC if $\mathcal{G}$ is VC.

- The collection $\{Z_j(\boldsymbol{X}, \boldsymbol{\theta}) = 1\}$ can be obtained through polynomials of degree 2 and hence is VC. Thus, the collection $\{Z_j(\boldsymbol{X}, \boldsymbol{\theta})\}$ (as indicators of the sets $\{Z_j(\boldsymbol{X}, \boldsymbol{\theta}) = 1\}$) is VC.

- Suppose $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_k$ be GC classes of functions on the probability measure P and ϕ is a continuous function from $\mathbb{R}^k$ to $\mathbb{R}$. Then $\mathcal{H} = \phi(\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_k)$ is GC with respect to the true probability measure provided that $\mathcal{H}$ has an integrable envelope function (Kosorok (2006) [88], Wellner (2012) [63]).

The first observation implies that the collection of functions $\{(\boldsymbol{X} - \boldsymbol{\mu}_j)' \boldsymbol{\Sigma}_j^{-1} (\boldsymbol{X} - \boldsymbol{\mu}_j)\}$ is VC. Now, the second observation implies that $\{\phi_p^{\beta}(\boldsymbol{x}, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\}$ is VC with envelope function $(2\pi c_1)^{-\frac{p\beta}{2}}$ (by the NS constraint, see Remark 3.1) which is integrable with respect to the true probability measure and hence is GC. Similarly, the third observation implies that the collection of functions $\{Z_j(\boldsymbol{X}, \boldsymbol{\theta})\}$ is VC (with envelope 1 which is again integrable) and hence is GC.

Now the fact that $\mathcal{F}$ is GC can be concluded by the fourth observation with $\mathcal{F}_{1j} = \{Z_j(\boldsymbol{X}, \boldsymbol{\theta}),\ \boldsymbol{\theta} \in \boldsymbol{\Theta}_C\}$, $\mathcal{F}_{2j} = \{\log\ \pi_j + \frac{1}{\beta}\phi_p^{\beta}(\boldsymbol{X}, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) - \frac{1}{1+\beta}\int \phi_p^{1+\beta}(\boldsymbol{x}, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\, d\boldsymbol{x},\ \boldsymbol{\theta} \in \boldsymbol{\Theta}_C\}$ for $j = 1, \ldots,\ k$ and $\phi(x_{11}, \ldots,\ x_{1k},\ x_{21}, \ldots,\ x_{2k}) = \sum_{j=1}^{k} x_{1j}x_{2j}$. The integrability (with respect to the true probability measure) of the envelope function of $\phi(\mathcal{F}_{11}, \ldots,\ \mathcal{F}_{1k}, \mathcal{F}_{21}, \ldots,\ \mathcal{F}_{2k})$ follows from the facts that

$$
\begin{aligned}
&Z_j(\boldsymbol{X}, \boldsymbol{\theta}) \in \{0, 1\},\ -\infty < \log \pi_{min} \leq \log \pi_j < 0, \text{ and} \\
&0 < \phi_p^{\beta}(\boldsymbol{X}, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \leq \frac{1}{(2\pi)^{\frac{p\beta}{2}} |\boldsymbol{\Sigma}|^{\frac{\beta}{2}}} \leq \frac{1}{(2\pi c_1)^{\frac{p\beta}{2}}},
\end{aligned}
$$

$\forall j$, where the last couple of inequalities follow from Assumption 3.2 and Remark 3.1. This completes the proof of Theorem 3.3. □

Remark 3.2. *We explicitly assume that the true unknown distribution (with probability measure F) belongs to the normal mixture family of distributions. However, it is observed that this assumption is only essential to prove the inequality (3.6). If we assume that the true unknown distribution satisfies inequality* (3.6)*, the results can be extended to more generalized set-ups. In case of real datasets, it is expected that the data come from distributions which are not exactly of normal mixture type. The aforesaid assumption (inequality (3.6)) can help to establish the theoretical results for distributions which lie in a certain neighbourhood of the normal mixture family. For*

example, let us consider the contaminated normal mixture distribution with density

$$f = \sum_{j=1}^{k} \pi_j^0 \phi_p(\cdot, \boldsymbol{\mu_j^0}, \boldsymbol{\Sigma_j^0}) + \pi_{k+1}^0 g(\cdot),$$

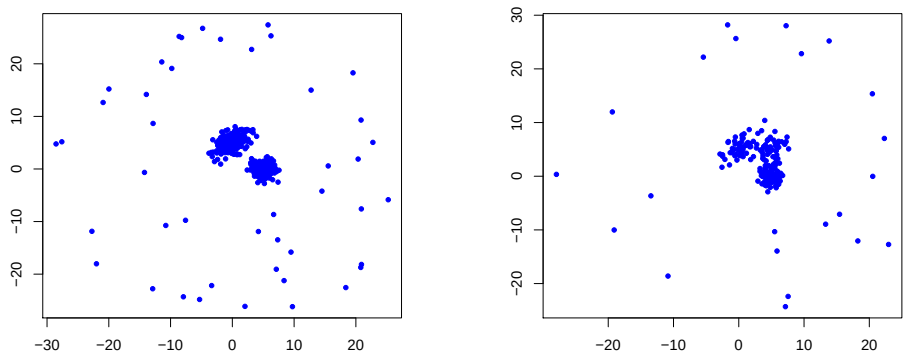
where the density g is the density of some contaminating distribution and π_{k+1}^0 is the contamination proportion. The proof of inequality (3.6) *depends only on the normal component density with the highest component weight and Assumption 3.1. Thus, the aforesaid inequality is valid in case of the aforesaid contaminated normal mixture distribution if the contaminating proportion does not exceed all of the regular component weights, i.e., $\pi_{k+1}^0 \leq max\{\pi_1^0, \ldots \pi_k^0\}$ which is quite expected in real life scenarios.*

3.3 More Simulation Experiments

We have assessed the comparative performance of the MPLE$_\beta$ method through simulation experiments in various set-ups by means of estimated misclassification rates and estimated bias and mean squared errors of the cluster mean estimates in Chapter 2. In this section, we further investigate the same via simulation (with slightly different set-ups), but this time, in terms of the estimated bias and mean squared errors of both the cluster mean and covariance estimates. To do this, multivariate normal mixture datasets are generated with different data dimensions ($p = 2,\ 6$), cluster components ($k = 2,\ 3$) and various sample sizes (n) depending on p and k. To assess the robustness of all the competing algorithms, contaminated datasets are used with 10% contaminating proportion. Component means, covariances, weights and sample sizes for different set-ups (depending on p and k) are tabulated in Table 3.1. The contaminating observations are generated using uniform distributions on p-dimensional spheres (centred at the origin and radius 30). Observations from the aforesaid distributions are generated and only those ones are used for contamination whose squared Mahalanobis distances from all of the component means exceed $10\chi^2_{0.975,p}$; $\chi^2_{0.975,p}$ being the 97.5-th percentile of the χ^2_p distributions ($p = 2, 6$). A couple of two dimensional simulated datasets following the aforesaid methodology with $k = 2$ and $k = 3$ are presented in Figure 3.1. For each of the aforesaid simulation set-ups, 100 samples are generated and the component mean and covariance estimates are obtained based on them. The average bias (L_2 normed bias) and mean squared errors are tabulated in Tables 3.2 and 3.3, for $k = 2$ and $k = 3$, respectively. As earlier, the competitors of our MPLE$_\beta$ algorithm

k	p	Means	Covariances	Weights	Sample sizes
2	2	$(5,0)^{'}$	$\boldsymbol{I}_2$	0.4	100, 200, 200
		$(0,5)^{'}$	$\boldsymbol{\Delta}$	0.5	
	6	$(5,0,0,0,0,0)^{'}$	$\boldsymbol{I}_6$	0.4	200, 300, 400
		$(0,5,0,0,0,0)^{'}$	$\begin{bmatrix} \boldsymbol{\Delta} & \boldsymbol{0}_{2\times 4} \\ \boldsymbol{0}_{4\times 2} & \boldsymbol{I}_4 \end{bmatrix}$	0.5	
3	2	$(5,0)^{'}$	$\boldsymbol{I}_2$	0.35	200, 300, 400
		$(0,5)^{'}$	$\boldsymbol{\Delta}$	0.30	
		$(5,5)^{'}$	$3\boldsymbol{I}_2$	0.25	
	6	$(5,0,0,0,0,0)^{'}$	$\boldsymbol{I}_6$	0.35	500, 750, 1000
		$(0,5,0,0,0,0)^{'}$	$\begin{bmatrix} \boldsymbol{\Delta} & \boldsymbol{0}_{2\times 4} \\ \boldsymbol{0}_{4\times 2} & \boldsymbol{I}_4 \end{bmatrix}$	0.30	
		$(5,5,0,0,0,0)^{'}$	$3\boldsymbol{I}_6$	0.25	

Table 3.1: The component means and covariances in various set-ups with $\boldsymbol{\Delta} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$. Contamination accounts for the remaining 10% weights in each case.



(a) $k = 2,\ n = 100.$ (b) $k = 3,\ n = 200.$

Figure 3.1: Structures of simulated datasets.

used here are the TCLUST [58], trimmed K-means (TKMEANS) [31], K-medoids [84] and the MCLUST [137] methods. The non-robust likelihood based method is also presented for comparison.

Let us now understand the implications of the simulation outputs presented in Tables 3.2 and 3.3. The ordinary maximum likelihood based method (i.e., MLE based) and the K-medoids algorithm produce large bias and mean squared errors due to the presence of the potential contaminating observations. But the MPLE$_\beta$ algorithm along with the trimming based methodologies (TCLUST and trimmed K-means) are

Method	Tuning Parameter	$p=2$ n	Location Bias	Location MSE	Scale Bias	Scale MSE	$p=6$ n	Location Bias	Location MSE	Scale Bias	Scale MSE
MLE		100	2.456	102.911	89.511	33525.003	200	1.490	39.583	90.632	31877.026
		200	2.043	55.030	126.075	51841.075	300	2.182	33.074	111.770	42793.568
		300	1.929	36.675	116.745	46775.057	400	2.355	21.025	119.032	42263.149
MPLE_β	0.1	100	0.057	0.131	0.216	1.006	200	0.038	0.161	0.097	1.420
		200	0.027	0.075	0.088	0.420	300	0.028	0.104	0.097	0.868
		300	0.008	0.042	0.057	0.242	400	0.020	0.083	0.095	0.693
	0.3	100	0.063	0.146	0.249	1.073	200	0.035	0.192	0.146	1.900
		200	0.028	0.079	0.132	0.456	300	0.032	0.123	0.156	1.163
		300	0.008	0.047	0.094	0.262	400	0.019	0.098	0.173	0.937
	0.5	100	0.067	0.168	0.360	1.442	200	0.035	0.255	0.167	2.798
		200	0.028	0.087	0.233	0.604	300	0.037	0.158	0.198	1.693
		300	0.010	0.053	0.185	0.352	400	0.023	0.125	0.227	1.331
TCLUST	0.1	100	1.024	59.916	5.897	577.924	200	0.676	57.445	14.757	4829.239
		200	1.020	40.727	5.009	291.974	300	0.488	80.375	20.296	6667.174
		300	0.376	16.151	2.698	110.372	400	0.969	73.178	18.696	5332.256
	0.15	100	0.014	0.151	0.478	1.426	200	0.042	0.185	0.390	2.081
		200	0.039	0.081	0.437	4.130	300	0.019	0.118	0.312	1.148
		300	0.021	0.057	0.579	0.637	400	0.031	0.079	0.321	0.891
TKMEANS	0.1	100	0.164	0.266	2.061	29.67	200	0.058	0.243	1.869	40.497
		200	0.148	0.185	1.976	19.631	300	0.062	0.154	1.395	22.100
		300	0.090	0.091	1.321	10.195	400	0.033	0.107	1.190	16.107
	0.15	100	0.044	0.184	0.936	1.774	200	0.042	0.192	0.845	1.824
		200	0.023	0.094	0.987	1.255	300	0.029	0.116	0.849	1.485
		300	0.020	0.060	1.077	1.314	400	0.018	0.095	0.840	1.297
KMEDOIDS		100	0.324	6.92	47.589	3682.42	200	0.192	2.155	39.704	2560.485
		200	0.16	0.272	47.014	2877.583	300	0.183	1.626	38.900	2159.686
		300	0.129	0.171	44.982	2496.328	400	0.173	1.472	38.518	1967.819
MCLUST		100	0.456	12.157	4.661	1385.746	200	0.215	5.015	5.961	3392.367
		200	0.030	0.075	0.066	0.352	300	0.030	0.100	0.086	0.867
		300	0.020	0.046	0.107	0.256	400	0.026	0.074	0.094	0.616

Table 3.2: Estimated bias and mean squared errors of different methods in case of $k = 2$.

successful in producing small bias and mean squared errors due to their robust natures. The MPLE_β clearly outperforms all its competitors in terms of estimated bias and mean squared errors, especially with small values of β (0.1 or 0.3). Although the MCLUST method performs closely in case of the two component set-up ($k = 2$), the same is found to be not that efficient in case of the three cluster models. It is observed that the convergence of the MCLUST method requires substantially larger sample sizes particularly for the three component set-up.

Method	Tuning Parameter	n	$p=2$ Location Bias	Location MSE	Scale Bias	Scale MSE	n	$p=6$ Location Bias	Location MSE	Scale Bias	Scale MSE
MLE		200	7.028	259.287	221.808	118324.918	500	4.167	59.609	201.924	91847.286
		300	6.083	193.402	242.414	135285.542	750	5.086	69.364	239.843	95023.254
		400	4.225	106.831	251.179	133127.074	1000	5.211	54.509	239.101	86296.694
MPLE_β	0.1	200	0.289	2.848	25.531	12830.310	500	0.131	0.345	0.483	5.235
		300	0.411	5.059	25.637	14755.109	750	0.116	0.235	0.385	3.393
		400	0.272	4.270	16.570	8395.707	1000	0.103	0.160	0.449	2.633
	0.3	200	0.174	0.474	0.252	2.974	500	0.105	0.417	0.882	8.057
		300	0.065	0.328	0.176	2.008	750	0.093	0.262	0.751	4.927
		400	0.119	0.257	0.211	1.723	1000	0.085	0.187	0.865	4.077
	0.5	200	0.168	0.504	0.229	4.048	500	0.183	0.697	1.356	13.774
		300	0.066	0.307	0.594	4.127	750	0.076	0.320	1.052	7.289
		400	0.079	0.257	0.344	2.535	1000	0.069	0.231	1.152	5.967
TCLUST	0.12	200	0.895	43.434	2.142	86.929	500	0.570	18.736	3.017	332.234
		300	1.121	34.043	1.699	18.225	750	0.432	9.539	2.777	323.290
		400	0.560	14.420	1.532	5.867	1000	0.187	0.200	1.179	4.242
	0.15	200	0.368	0.846	1.960	6.829	500	0.233	0.617	1.848	9.584
		300	0.315	0.564	2.021	5.951	750	0.239	0.364	1.844	7.487
		400	0.228	0.357	1.985	5.261	1000	0.175	0.234	1.825	6.376
TKMEANS	0.12	200	0.305	0.554	1.668	4.903	500	0.701	20.683	3.877	437.964
		300	0.268	0.363	1.569	4.044	750	0.397	0.433	1.674	4.802
		400	0.457	6.815	1.632	9.424	1000	0.428	0.395	1.673	4.416
	0.15	200	0.239	0.579	2.170	5.302	500	0.352	0.614	2.248	7.781
		300	0.157	0.356	2.044	4.636	750	0.348	0.442	2.272	6.861
		400	0.221	0.305	2.003	4.372	1000	0.387	0.385	2.271	6.420
KMEDOIDS		200	4.326	146.706	72.435	12268.229	500	1.018	34.213	54.660	6547.075
		300	4.592	147.494	72.143	11719.062	750	0.545	3.702	47.612	2929.350
		400	3.5	113.030	67.284	8937.877	1000	0.682	3.476	49.184	2948.087
MCLUST		200	2.179	98.377	6.839	1169.55	500	1.684	92.432	38.103	18314.392
		300	2.052	73.679	4.297	285.849	750	4.530	123.320	113.597	37746.933
		400	1.157	42.544	2.986	136.394	1000	4.155	50.796	166.641	56291.524

Table 3.3: Estimated bias and mean squared errors of different methods in case of $k = 3$.

3.4 Further Real Data Examples

We now illustrate our method using two more real life datasets (in addition to the datasets used in Chapter 2); one of them involves univariate data while the other is a multivariate example.

3.4.1 Univariate Case

Here we apply our method on the red blood cell sodium-lithium countertransport (SLC) dataset which has been analyzed in the past by several authors including Dudley et al.

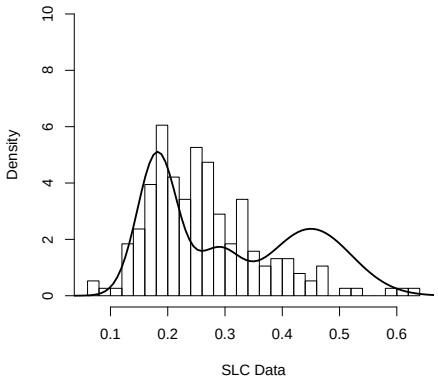
(1991) [41], Roeder (1994) [124] and Fujisawa and Eguchi (2006) [54].

Geneticists are concerned about SLC as it may correlate blood pressure and hence may be a potential factor behind hypertension. SLC is also less complicated to assess than blood pressure, because the latter is a complex quantitative trait influenced by environmental and genetic factors. The sample size is 190 and the dataset consists of 3 genotypes, namely, A_1A_1, A_1A_2, A_2A_2. We analyze these data using the ordinary maximum likelihood approach, the maximum pseudo β-likelihood approach, as well as the approach of Fujisawa and Eguchi [54]. Figure 3.2 presents the histogram of the original data with different fits overlaid, including the maximum likelihood fit, the Fujisawa-Eguchi (FE) fit with $\beta = 0.45$, the maximum pseudo β-likelihood fit with $\beta = 0.5$, all with both 3 clusters and 4 clusters. The component parameter estimates of the aforesaid models are presented in Table 3.4.

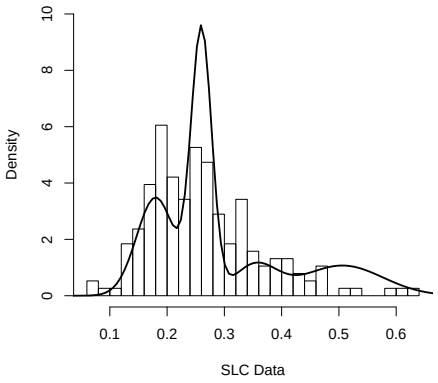
Methods	Clusters	$\hat{\omega}_1$	$\hat{\omega}_2$	$\hat{\omega}_3$	$\hat{\omega}_4$	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\mu}_4$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\sigma}_3^2$	$\hat{\sigma}_4^2$
MLE	3	0.442	0.137	0.421	-	0.182	0.288	0.450	-	0.001	0.001	0.005	-
MLE	4	0.100	0.421	0.289	0.190	0.180	0.261	0.342	0.473	0.001	0.0003	0.0008	0.005
FE ($\beta = 0.45$)	3	0.076	0.584	0.340	-	0.187	0.227	0.336	-	0.0001	0.004	0.012	-
FE ($\beta = 0.45$)	4	0.103	0.447	0.367	0.073	0.202	0.232	0.270	0.343	0.007	0.007	0.008	0.014
MPLE$_\beta$ ($\beta = 0.5$)	3	0.422	0.289	0.289	-	0.185	0.260	0.365	-	0.001	0.0004	0.004	-
MPLE$_\beta$ ($\beta = 0.5$)	4	0.421	0.289	0.153	0.137	0.185	0.261	0.330	0.422	0.0008	0.0004	0.0003	0.002

Table 3.4: Component parameter estimates for the SLC data.

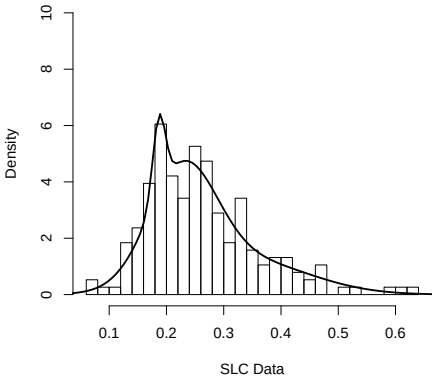
The SLC dataset was originally composed of three clusters (representing the three genotypes). The histogram of the data shows three possible modes for the three probable clusters. However, it is observed that although the first two clusters (around the first two modes in the histogram) are approximately symmetric and bell-shaped in nature, the third cluster appears to significantly deviate from symmetry with a very long right tail. This leads to the discovery of only one significant real mode by the method of maximum likelihood (with 3 clusters), together with an almost invisible (and incorrect) second mode, and an entirely inaccurate third mode which is pushed way to the right to accommodate some very large observations on the right tail. The FE method (with 3 clusters) possibly identifies the second mode, but it is far too tentative and diffused, perhaps due to its closeness with the first mode. The third mode is not at all discernible in the figure in this case, a consequence of the large estimated variance for the third component. The FE solution also appears to be substantially affected by the very large observations on the right tail. The MPLE$_\beta$ method (with 3 clusters) provides a much more improved fit compared to the previous two. The first two modes are



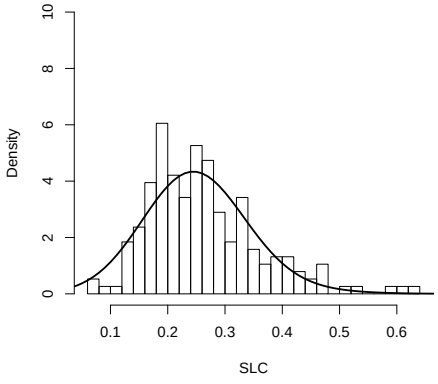
(a) MLE based fit using 3 clusters.



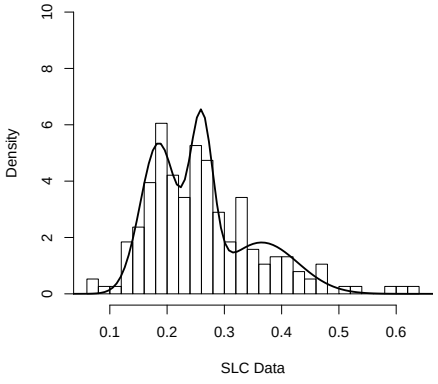
(b) MLE fit using 4 clusters.



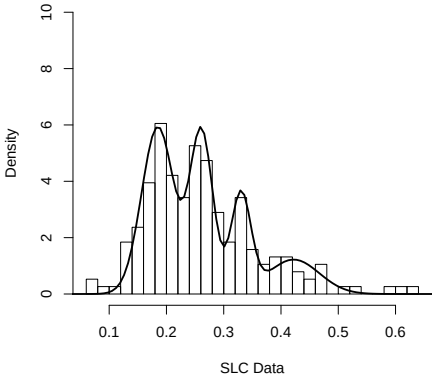
(c) FE fit using 3 clusters with $\beta = 0.45$.



(d) FE fit using 4 clusters with $\beta = 0.45$.



(e) MPLE$_\beta$ fit using 3 clusters with $\beta = 0.5$.



(f) MPLE$_\beta$ fit using 4 clusters with $\beta = 0.5$.

Figure 3.2: Fitted densities using different methods.

very accurately determined with suitable separation. The estimated third mode does not fully match the observed third mode, possibly because of the skewed pattern in the third cluster. However, unlike previous two fits, this fit clearly discounts the effect of the very large outliers to the right. Observing the skewed third cluster in the data which is representing a model misspecification, a 4-component normal mixture model has also been fitted using all of the aforesaid methods (with the aforementioned tuning parameters). The three modes are now successfully and accurately recognized by the MPLE$_\beta$ method with the fourth fitted cluster pooling the skewed and misspecified part

in the overall data. In this case also the large outliers are clearly discounted. On the other hand, the maximum likelihood method and the FE method (with 4 clusters) are still unable to fit the histogram appropriately; the former finds the modes inaccurately while the latter fails to distinguish among the different clusters. In an overall sense, it can be concluded that the MPLE$_\beta$ method (with both 3 and 4 component normal mixture models) has provided substantially improved fits to these data compared to the maximum likelihood and FE methods.

3.4.2 Multivariate Case

We now describe the utility of our method by applying it on a multivariate dataset, namely, the Thyroid Gland data[1]. The data provide information on 215 patients about the laboratory test outcomes of five medical attributes. These attributes are *(i)* $T3$-resin uptake test (in percentage, RT3U), *(ii)* total serum thyroxin as measured by the isotopic displacement method (T4), *(iii)* total serum triiodothyronine as measured by radioimmuno assay (T3), *(iv)* basal thyroid-stimulating hormone (TSH) as measured by radioimmuno assay (TSH) and *(v)* maximal absolute difference of TSH value after injection of 200 micro grams of thyrotropin releasing hormone as compared to the basal value (DTSH).

Component	MLE		MPLE$_\beta$	
Means	Original	Contaminated	Original	Contaminated
$\hat{\mu}_1$	93.194	93.190	95.781	95.786
	17.019	17.019	15.993	15.993
	4.161	4.167	3.654	3.658
	0.975	0.982	0.953	0.953
	−0.047	−0.047	−0.049	−0.044
$\hat{\mu}_2$	110.908	111	110.414	110.411
	9.156	9.132	9.006	9.007
	1.725	1.724	1.690	1.690
	1.324	1.327	1.246	1.244
	2.582	2.711	2.388	2.383
$\hat{\mu}_3$	124.577	134.039	126.936	124.291
	3.635	12.696	2.925	3.833
	1.031	6.418	0.920	1.047
	14.677	30.864	12.608	10.697
	19.596	33.601	19.173	17.727

Table 3.5: Component mean estimates for the Thyroid Gland data (MPLE$_\beta$ method with $\beta = 0.3$ and MLE) in case of the original and the contaminated datasets.

The data also reveal the actual thyroidal state of these 215 patients, i.e., whether

[1]Source: https://archive.ics.uci.edu/ml/datasets/thyroid+disease and the R package mclust [137]

they are suffering from euthyroidism (normal thyroid gland function), hypothyroidism (underactive thyroid not producing enough thyroid hormone) or hyperthyroidism (overactive thyroid producing and secreting excessive amounts of the free thyroid hormones T3 and/or thyroxine T4). Thus, we can fit a 3-component normal mixture model (with 5-dimensions) using the MPLE$_\beta$ method. Here the original cluster sizes are not equal and two of the original cluster sizes are only 30 and 35 while the data dimension is 5. For this relatively larger $\frac{p}{n}$ value, the MPLE$_\beta$ estimates for each clusters need a stable starting value. We thus used S-estimates of location and scale (Lopuhaä (1989) [97]) in this regard. An exploratory analysis of these data appear to indicate that there are no major outliers in the dataset. This is also suggested by the similarity of the component mean estimates in case of the MPLE$_\beta$ (with $\beta = 0.3$) and the non-robust maximum likelihood estimates in Table 3.5. To establish the outlier stability of the MPLE$_\beta$ method, we contaminate the original data artificially with 10 additional points which are discrepant in comparison with the original data and can be viewed as outliers. These contaminating observations are listed in Section 3.5.2. The component mean estimates of this artificially contaminated dataset by the maximum β-likelihood method (with $\beta = 0.3$) and usual likelihood based method are also presented in Table 3.5. The stability of the MPLE$_\beta$ method is immediately observed in the minimal shifts in the component mean estimates which obviously cannot be claimed for the MLEs. The variation in the estimates of the third cluster is higher than those of the other two, but here also the MPLE$_\beta$ estimator is far more stable than the MLE. The superiority of the MPLE$_\beta$ approach over the ordinary likelihood version is quite apparent, at least as far as the evidence of this example. The covariance matrix estimates for the original data and the artificially contaminated data for both likelihood based and MPLE$_\beta$ algorithms are presented in Section 3.5.3. The superiority of the MPLE$_\beta$ approach over the ordinary likelihood version can again be observed in terms of greater stability of the estimated covariance matrix elements and greater sign consistency of the same.

3.5 Appendices

3.5.1 A Required Lemma

Lemma 3.4. *Consider the set-up of Theorem 3.1, its proof and assume that $\pi_j > 0$ for $j = 1, \cdots, k$ in (3.7). Then, we have the following results under the ER and NS constraints.*

1. *$g = k$ in (3.8) for the mean sequence in the proof of Theorem 3.1.*

2. *The dispersion sequence $\{\boldsymbol{\Sigma}_1^r, \boldsymbol{\Sigma}_2^r, \ldots, \boldsymbol{\Sigma}_k^r\}$, from the proof of Theorem 3.1, only satisfies (3.9) (and not (3.10) or (3.11)).*

Proof. To prove the lemma, we need the following inequalities which hold trivially from the definitions of M_r and m_r.

I1 For $1 \leq j \leq k$ and $r \in \mathbb{N}$,

$$m_r^p \leq |\boldsymbol{\Sigma}_j^r| \leq M_r^p.$$

I2 For $1 \leq j \leq k$ and $r \in \mathbb{N}$,

$$(\boldsymbol{X} - \boldsymbol{\mu}_j^r)'(\boldsymbol{\Sigma}_j^r)^{-1}(\boldsymbol{X} - \boldsymbol{\mu}_j^r) \geq M_r^{-1}||\boldsymbol{X} - \boldsymbol{\mu}_j^r||^2.$$

I3 For $1 \leq j \leq k$ and $r \in \mathbb{N}$,

$$\begin{aligned}\frac{1}{1+\beta}\int \phi_p^{1+\beta}(\boldsymbol{x}, \boldsymbol{\mu}_j^r, \boldsymbol{\Sigma}_j^r)\, d\boldsymbol{x} &= \frac{1}{(2\pi)^{\frac{p\beta}{2}}|\boldsymbol{\Sigma}_j^r|^{\frac{\beta}{2}}(1+\beta)^{\frac{p+2}{2}}} \\ &\geq \frac{1}{(2\pi)^{\frac{p\beta}{2}} M_r^{\frac{p\beta}{2}}(1+\beta)^{\frac{p+2}{2}}}.\end{aligned}$$

Using the above inequalities, we have,

$$L_\beta(\boldsymbol{\theta}^r, P) \leq E_P\left[\sum_{j=1}^k Z_j(\boldsymbol{X}, \boldsymbol{\theta}^r)\left[\log\ \pi_j^r + \frac{1}{\beta(2\pi)^{\frac{p\beta}{2}} m_r^{\frac{p\beta}{2}}} e^{-\frac{\beta M_r^{-1}}{2}||\boldsymbol{X}-\boldsymbol{\mu}_j^r||^2} - \frac{1}{(2\pi)^{\frac{p\beta}{2}} M_r^{\frac{p\beta}{2}}(1+\beta)^{\frac{p+2}{2}}}\right]\right]. \tag{3.12}$$

Let us first prove the second part of Lemma 3.4 . Suppose that (3.10) holds. Then the eigenvalue ratio constraint implies,

$$m_r \geq \frac{M_r}{c} \to \infty.$$

These would imply,

$$\begin{aligned}
&\lim_{r\to\infty} L_\beta(\boldsymbol{\theta}^r, P) \\
&\leq \lim_{r\to\infty} E_P\left[\sum_{j=1}^{k} Z_j(\boldsymbol{X},\boldsymbol{\theta}^r)\left[\log\ \pi_j^r + \frac{1}{\beta(2\pi)^{\frac{p\beta}{2}} m_r^{\frac{p\beta}{2}}} e^{-\frac{\beta M_r^{-1}}{2}||\boldsymbol{X}-\boldsymbol{\mu}_j^r||^2} - \frac{1}{(2\pi)^{\frac{p\beta}{2}} M_r^{\frac{p\beta}{2}}(1+\beta)^{\frac{p+2}{2}}}\right]\right] \\
&\leq \lim_{r\to\infty} E_P\left[\sum_{j=1}^{k} \log\ \pi_j^r\right] < 0
\end{aligned}$$

which contradicts (3.6). Now, let us assume that (3.11) holds. But this contradicts the non-singularity constraint. Hence the dispersion sequence can only satisfy the condition in (3.9), and not the conditions (3.10) or (3.11).

To prove the first part of the Lemma (i.e., $g = k$ in (3.8)), let us observe that if $g = 0$ then, $||\boldsymbol{\mu}_j^r|| \to \infty$ and thus $e^{-||\boldsymbol{X}-\boldsymbol{\mu}_j^r||^2} \to 0$ for all $1 \leq j \leq k$. Hence, (3.12) again implies $\lim_{r\to\infty} L_\beta(\boldsymbol{\theta}^r, P) < 0$, which contradicts (3.6). Hence $g > 0$.

Next let us assume that, $1 \leq g < k$. Then *bounded convergence theorem* implies,

$$E_P\left(\sum_{j=g+1}^{k} Z_j(\boldsymbol{X},\boldsymbol{\theta}^r)\right) \to 0. \tag{3.13}$$

Now,

$$\begin{aligned}
&\operatorname*{limsup}_{r\to\infty} L_\beta(\boldsymbol{\theta}^r, P) \\
&= \operatorname*{limsup}_{r\to\infty} E_P\left[\sum_{j=1}^{k} Z_j(\boldsymbol{X},\boldsymbol{\theta}^r)\left[\log\ \pi_j^r + \frac{1}{\beta}\phi_p^\beta(\boldsymbol{X},\boldsymbol{\mu}_j^r,\boldsymbol{\Sigma}_j^r) - \frac{1}{1+\beta}\int \phi_p^{1+\beta}(\boldsymbol{x},\boldsymbol{\mu}_j^r,\boldsymbol{\Sigma}_j^r)\,d\boldsymbol{x}\right]\right] \\
&\leq \operatorname*{limsup}_{r\to\infty} E_P\left[\sum_{j=1}^{g} Z_j(\boldsymbol{X},\boldsymbol{\theta}^r)\left[\log\ \pi_j^r + \frac{1}{\beta}\phi_p^\beta(\boldsymbol{X},\boldsymbol{\mu}_j^r,\boldsymbol{\Sigma}_j^r) - \frac{1}{1+\beta}\int \phi_p^{1+\beta}(\boldsymbol{x},\boldsymbol{\mu}_j^r,\boldsymbol{\Sigma}_j^r)\,d\boldsymbol{x}\right]\right] \\
&+ \operatorname*{limsup}_{r\to\infty} E_P\left[\sum_{j=g+1}^{k} Z_j(\boldsymbol{X},\boldsymbol{\theta}^r)\left[\log\ \pi_j^r + \frac{1}{\beta}\phi_p^\beta(\boldsymbol{X},\boldsymbol{\mu}_j^r,\boldsymbol{\Sigma}_j^r) - \frac{1}{1+\beta}\int \phi_p^{1+\beta}(\boldsymbol{x},\boldsymbol{\mu}_j^r,\boldsymbol{\Sigma}_j^r)\,d\boldsymbol{x}\right]\right].
\end{aligned}$$

The second term in the right hand side of the above inequality less than equal to 0 due

to (3.13). Hence,

$$\begin{aligned}
&\limsup_{r\to\infty} L_\beta(\boldsymbol{\theta}^r, P) \\
&\leq \limsup_{r\to\infty} E_P\left[\sum_{j=1}^{g} Z_j(\boldsymbol{X},\boldsymbol{\theta}^r)\left[\log\ \pi_j^r + \frac{1}{\beta}\phi_p^\beta(\boldsymbol{X},\boldsymbol{\mu}_j^r,\boldsymbol{\Sigma}_j^r) - \frac{1}{1+\beta}\int \phi_p^{1+\beta}(\boldsymbol{x},\boldsymbol{\mu}_j^r,\boldsymbol{\Sigma}_j^r)\,d\boldsymbol{x}\right]\right] \\
&= E_P\left[\sum_{j=1}^{g} Z_j(\boldsymbol{X},\boldsymbol{\theta}^*)\left[\log\ \pi_j + \frac{1}{\beta}\phi_p^\beta(\boldsymbol{X},\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j) - \frac{1}{1+\beta}\int \phi_p^{1+\beta}(\boldsymbol{x},\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)\,d\boldsymbol{x}\right]\right]
\end{aligned}$$

where,

$$\begin{aligned}
\boldsymbol{\theta}^* &= \left(\pi_1,\pi_2,\ldots,\pi_g,\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\ldots,\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2,\ldots,\boldsymbol{\Sigma}_g\right) \\
&= \lim_{r\to\infty}\ \left(\pi_1^r,\pi_2^r,\ldots,\pi_g^r,\boldsymbol{\mu}_1^r,\boldsymbol{\mu}_2^r,\ldots,\boldsymbol{\mu}_g^r,\boldsymbol{\Sigma}_1^r,\boldsymbol{\Sigma}_2^r,\ldots,\boldsymbol{\Sigma}_g^r\right)
\end{aligned}$$

and π_j, $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ are as in (3.7), (3.8) and (3.9) respectively. Let us observe that $\sum_{j=1}^{g}\pi_j < 1$ due to the assumption that $\pi_j > 0$ for all $1 \leq j \leq k$. Motivated by this observation, we introduce the following standardized weights,

$$\pi_j^{'} = \begin{cases} \frac{\pi_j}{\sum_{j=1}^{g}\pi_j}, & \text{for } 1 \leq j \leq g \\ 0, & \text{for } j > g. \end{cases}$$

It is easy to observe that,

1. For all $1 \leq j \leq g$, $\log\ \pi_j < \log\ \pi_j^{'}$.

2. This aforesaid modification keeps the orderings of the discriminant functions $\{D_j(\boldsymbol{X},\cdot) :\ 1 \leq j \leq k\}$ invariant so that values of the assignment functions $\{Z_j(\boldsymbol{X},\cdot) :\ 1 \leq j \leq k\}$ would not change.

The aforesaid facts together imply,

$$\begin{aligned}
&\limsup_{r\to\infty} L_\beta(\boldsymbol{\theta}^r, P) \\
&\leq E_P\left[\sum_{j=1}^{g} Z_j(\boldsymbol{X},\boldsymbol{\theta}^*)\left[\log\ \pi_j + \frac{1}{\beta}\phi_p^\beta(\boldsymbol{X},\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j) - \frac{1}{1+\beta}\int \phi_p^{1+\beta}(\boldsymbol{x},\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)\ d\boldsymbol{x}\right]\right] \\
&< E_P\left[\sum_{j=1}^{g} Z_j(\boldsymbol{X},\boldsymbol{\theta}^{*'})\left[\log\ \pi_j^{'} + \frac{1}{\beta}\phi_p^\beta(\boldsymbol{X},\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j) - \frac{1}{1+\beta}\int \phi_p^{1+\beta}(\boldsymbol{x},\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)\ d\boldsymbol{x}\right]\right] \\
&= L_\beta(\boldsymbol{\theta}^{*'}, P),
\end{aligned}$$

where $\boldsymbol{\theta}^{*'} = \left(\pi_1^{'}, \pi_2^{'}, \ldots, \pi_g^{'}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \ldots, \boldsymbol{\Sigma}_g\right) \in \boldsymbol{\Theta}_C$. But this contradicts (3.5). Hence, $g = k$, completing the proof of the Lemma. □

3.5.2 Contaminating Observations Added to the Thyroid Gland Data

The contaminating observations added artificially to the Thyroid Gland data are as follows:

Units	Contaminating Observations				
1	155.704	36.535	17.451	66.078	64.804
2	156.124	35.039	20.026	67.200	66.104
3	154.383	33.625	20.428	68.302	66.031
4	156.890	34.951	20.496	66.417	65.875
5	156.572	34.154	17.831	68.719	65.724
6	155.042	35.240	18.686	68.095	65.878
7	154.823	33.750	20.586	66.143	66.802
8	153.185	36.636	19.510	66.242	68.082
9	155.616	35.583	19.511	67.386	64.267
10	155.996	36.244	20.071	66.410	64.583

Table 3.6: Contaminating observations added to the Thyroid Gland data.

3.5.3 Component Covariance Matrix Estimates for the Thyroid Gland data

The component covariance matrix estimates for the Thyroid Gland data are as follows:

Cluster	Original Data					Contaminated Data				
1	218.051	-30.978	-21.142	-1.284	0.608	218.047	-30.973	-21.148	-1.281	0.604
	-30.975	20.704	5.188	0.154	0.006	-30.978	20.704	5.182	0.154	0.007
	-21.144	5.188	5.130	0.063	0.014	-21.142	5.188	5.125	0.067	0.011
	-1.281	0.154	0.063	0.162	-0.030	-1.286	0.152	0.062	0.162	-0.031
	0.604	0.006	0.011	-0.039	0.061	0.607	0.006	0.011	-0.035	0.061
2	77.360	9.764	2.166	-0.355	-0.129	77.000	9.469	2.126	-0.325	0.783
	9.764	6.356	0.515	-0.040	-1.283	9.469	6.326	0.514	-0.047	-1.493
	2.166	0.515	0.265	-0.005	0.003	2.126	0.514	0.263	-0.006	-0.009
	-0.355	-0.04	-0.005	0.245	0.106	-0.325	-0.047	-0.006	0.243	0.138
	-0.129	-1.283	0.003	0.106	3.784	0.783	-1.493	-0.009	0.138	5.037
3	72.254	-5.453	-0.858	-12.438	-58.142	248.724	203.876	119.329	317.760	248.581
	-5.453	4.118	0.807	-14.686	11.245	203.876	219.471	126.209	341.553	320.974
	-0.858	0.807	0.285	-2.802	3.722	119.329	126.209	73.552	201.171	183.511
	-12.438	-14.686	-2.802	153.898	-6.269	317.760	341.553	201.171	669.271	490.303
	-58.142	11.245	3.722	-6.269	245.921	248.581	320.974	183.511	490.303	628.838

Table 3.7: Component covariance matrix estimates for the Thyroid Gland data in case of the maximum likelihood estimation.

Cluster	Original Data					Contaminated Data				
1	155.261	-28.989	-17.165	-0.107	0.002	155.263	-28.983	-17.165	-0.105	0.002
	-28.989	24.782	5.454	-0.233	0.251	-28.984	24.782	5.449	-0.237	0.256
	-17.165	5.454	3.845	-0.058	0.082	-17.162	5.453	3.841	-0.052	0.086
	-0.102	-0.233	-0.058	0.167	-0.021	-0.105	-0.231	-0.053	0.162	-0.020
	0.002	0.251	0.082	-0.020	0.057	0.002	0.252	0.084	-0.021	0.055
2	66.207	5.769	1.480	-0.034	2.163	65.676	5.728	1.470	-0.033	2.136
	5.769	4.493	0.393	-0.030	-0.452	5.728	4.456	0.391	-0.029	-0.436
	1.480	0.393	0.220	-0.002	0.099	1.470	0.391	0.218	-0.002	0.100
	-0.034	-0.030	-0.002	0.202	0.041	-0.033	-0.029	-0.002	0.199	0.039
	2.163	-0.452	0.099	0.041	3.232	2.136	-0.436	0.100	0.039	3.169
3	62.309	-3.334	-1.324	-5.678	-109.539	98.298	-13.328	-2.083	21.443	-69.500
	-3.334	2.852	0.734	-4.840	22.356	-13.328	6.382	1.236	-12.788	12.732
	-1.324	0.734	0.311	-1.418	6.925	-2.083	1.236	0.388	-2.276	4.817
	-5.678	-4.840	-1.418	31.978	-10.720	21.443	-12.788	-2.276	52.686	-0.003
	-109.539	22.356	6.925	-10.720	386.560	-69.500	12.732	4.817	-0.003	321.714

Table 3.8: Component covariance matrix estimates for the Thyroid Gland data in case of the minimum DPD estimation.

Chapter 4

Sequential Minimum Density Power Divergence Estimation

4.1 Introduction

We have developed an IRLS based algorithm for minimizing the DPD in case of multivariate normal models and applied this estimating algorithm to develop a robust clustering tool (MPLE$_\beta$) in Chapter 2. This methodology is broadly illustrated with its theoretical properties (in Chapter 3), simulation experiments, real data examples and an application in image reconstruction. Although this algorithm enjoys some novel theoretical properties, some computational difficulties can be observed, especially with growing data dimensions and large values of β (specially $\beta > 0.5$) which is essential to achieve a desired level of robustness. This problem is mainly associated with the aforesaid IRLS algorithm for the estimation of component means and covariance matrices. In particular, this IRLS algorithm is found to be non-covergent in some situations (for large data dimensions or large values of β) mainly due to the non-singularities of the covariance matrix iterates in intermediate steps of iteration. Observing this hindrance, we attempt to search for possible modifications of the IRLS procedure for estimating the location and scale parameters in multivariate location-scale set-up. In the present chapter, we propose a new componentwise robust estimation procedure for elliptically symmetric models in the spirit of minimizing the DPD.

Efficient estimation of covariance and correlation matrices is of prime interest in the analysis of multivariate data arising across all scientific disciplines. Many celebrated statistical tools, such as discriminant analysis (linear as well as quadratic), principal component analysis, factor analysis and cluster analysis are based on the covariance structures of different attributes; see, e.g., Mardia et al. [104]. That is why numerous estimation procedures for multivariate location and scatter have been proposed in the literature. The most eminent classical covariance estimator based on a random sample of multivariate observations, $\{\boldsymbol{X}_1, \ \boldsymbol{X}_2, \ldots, \ \boldsymbol{X}_n\}$, is the sample covariance matrix

$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})'$, where $\bar{\boldsymbol{X}}$ is the sample mean. In fact, this is the method of moments estimator of $\boldsymbol{\Sigma}$ (the population covariance matrix) which is consistent, location invariant and scale equivariant. Under normality of the data, it is also the MLE and thus asymptotically the most efficient one under standard regularity conditions. But, this estimator is not resistant to outliers as it depends on the absolute values of the deviations of sample observations from the non-robust sample mean. The asymptotic breakdown point of this estimator is indeed zero indicating its extreme non-robust nature. However, the subjective complexity of real life practical problems as well as the volume of data are increasing by the day and hence the analyses of such data are often expected to face the issues of model misspecification and sensitivity to outliers. Although non-parametric methods based on ranks, signs and depth measures provide possible solutions for non-robustness against model misspecification, they often lead to a major compromise in asymptotic efficiency along with the curse of dimensionality. Alternatively, when the majority of the data can be well-fitted by a parametric model, the effects of the outlying minority can be handled robustly, without a major loss in asymptotic efficiency, by using appropriate parametric robust procedures.

For robust estimation of covariance (or correlation) matrix, however, one cannot avoid the estimation of the corresponding location parameter. Bickel (1964) [15], and Sen and Puri (1971) [138] proposed coordinatewise median (see Small (1990) [139] for a detailed discussion on generalization of medians into higher dimensional settings) and R estimators for multivariate location estimation, respectively. Later Maronna (1976) [109] proposed the pioneering idea of simultaneous M-estimation of multivariate location and scatter for the elliptically symmetric distributions. The breakdown bounds and influence functions are calculated to show strong robustness properties of these multivariate M-estimators, although the upper bounds of their breakdown points $\left(\frac{1}{p+1}\right.$ in dimension p, Donoho (1982) [38]$\left.\right)$ decrease to zero as the dimension grows to infinity. Tyler ((1983) [148], (1987) [149], (2014) [150]) and Kent and Tyler ((1991) [86], (1996) [87]) made significant contributions to the area of M-estimations of multivariate location and scatter, which include the constrained M-estimators (Kent and Tyler (1996) [87]), distribution free M-estimators (Tyler (1987) [149]), redescending M-estimators (Kent and Tyler (1991) [86]) and some more. Theoretical properties, indicating their robustness and efficiencies, have also been discussed in some of the aforesaid works. In particular, Tyler (2014) [150] has shown that the asymptotic breakdown point of the multivariate scatter M-estimator is bounded above by $\frac{1}{p+1}$ only under certain "copla-

nar" contamination of the data. Under the absence of such "coplanar" contamination, the breakdown point of the scatter M-estimator is close to $\frac{1}{2}$ which fixes the problem of poor breakdown values of the scatter M-estimators in higher dimensions (a drawback of Maronna's method [109]). Another popular approach to construct robust estimators of location and scatter is to perform trimming on sample observations, leading to the popular minimum covariance determinant (MCD) and minimum volume ellipsoid (MVE) estimators. A detailed exposition of MCD and MVE estimators can be found in Rousseeuw and Driessen (1999) [130] and Rousseeuw (1985) [129], respectively. A major drawback of the MVE estimators is its slower convergence rate of $o(n^{-\frac{1}{3}})$. However, the MCD estimators are $n^{\frac{1}{2}}$-consistent but compromise significantly in asymptotic efficiency compared to maximum likelihood to achieve robustness (higher breakdown value) under normality assumption. Stahel (1981) [141] and Donoho (1982) [38] have independently proposed estimators of multivariate location and scatter as weighted means and weighted covariance matrices of the data where the weights are based on reasonable measures of "outlyingness". Some other recent works in the field of robust estimation of multivariate location and scatter matrix include Danilov et al. (2012) [35], who have dealt with data contamination and missing data at the same time, Maronna and Yohai (2014) [111] and Agostinelli et al. (2015) [2], who have proposed estimates of multivariate location and scatter that are resistant to both cellwise and casewise contamination, and Agostinelli and Greco (2019) [1], who have proposed a weighted likelihood based approach for estimating multivariate location and scatter. Minimum distance methodology have also been utilized to produce robust and efficient estimators of multivariate location and scatter matrices; see Basu et al. (2011) [12], Ghosh et al. (2017) [61].

One of the crucial challenges behind constructing robust estimators in multivariate set-ups is the resulting lower asymptotic efficiency along with significantly higher computational costs as dimension increases. Most existing robust estimators that enjoy high asymptotic efficiency (e.g., methods based on sophisticated iterative algorithms) are computationally expensive and, hence, it becomes really problematic to compute them for large and high-dimensional multivariate datasets. A satisfactory solution to this problem is extremely important for analyzing modern large datasets that we are frequently encountering in the twenty-first century. One possible approach to settle this issue could be the incorporation of a two-stage procedure where a highly robust but inefficient estimator is first taken as the initial choice and subsequently an efficient

estimator is produced in the second stage by solving suitable estimating equations. Gervini (2003) [60] has presented such a two-stage procedure that uses the reweighted estimators of multivariate location and scatter with data-driven weights, starting from a highly robust and fast (but not necessarily efficient) estimator at the first step. Componentwise estimation of location and scatter is another possible solution, which is computationally much simpler and often can maintain desirable robustness and asymptotic efficiency, simultaneously. Mehrotra (1995) [115], Ma and Genton (2001) [101] are two such examples where elementwise estimation of scatter matrix is done, respectively, using the modified A-estimator (Lax (1985) [90]) of scale and using a highly robust estimator of scale by bypassing the estimation of location vector.

In the present chapter, we propose a new robust estimation procedure for the multivariate location vector and the scatter matrix for the class of elliptically symmetric distributions, which is easy to compute via a suitably parallelised computational algorithm and also has significantly high asymptotic efficiency at the model distribution. In particular, the component means and variances are first estimated separately for each variable (in parallel), which are subsequently used to estimate the correlation coefficients between each pair of variables (in parallel). The covariances are then computed from the estimated variances and correlation values. In each step of estimation, we utilize nice properties, e.g., strong robustness and high efficiency, of the MDPDE. As a result, our proposed estimators, which we refer to as the sequential minimum DPD estimators (SMDPDEs) of multivariate location and scatter, also become highly robust and efficient along with computational tractability in all scenarios, particularly for large dimensional datasets. In contrast, the IRLS procedure developed in Chapter 2 does not have such computaional tractability in case of either large data dimensions or large values of β.

We present some important theoretical properties of our estimators in Sections 4.3 and 4.4 (detailed proofs are deferred to Section 4.8), such as consistency and asymptotic normality of our estimators, asymptotic positive definiteness of our covariance matrix estimator and the influence function analysis of our functionals. Asymptotic relative efficiencies of our estimators (component means, variances and correlation estimators) and the explicit forms of the influence functions of our functionals are derived under the assumption of normality. In this scenario, we observed the superiority of SMDPDEs over the ordinary MDPDEs in terms of asymptotic efficiency, especially for the location and scale estimators; in case of the correlation also the SMDPDEs

are, at the least, competitive with the ordinary MDPDEs. Simulation experiments are conducted to compare the sequential method with the (non-robust) maximum likelihood method and (robust) ordinary minimum DPD method and some non-parametric methods including the MCD and MVE algorithms (and others) which also suggest the possible superiority of the sequential method in terms of bias and mean squared errors. Obtaining numerical solutions (with small to moderate computing effort) for the estimates is guaranteed under the sequential approach. Finding the ordinary MDPDE (the simultaneous minimizer) may, on the other hand, be an uphill task, particularly under growing dimensions and larger tuning parameter β; one will occasionally come across a situation, more common in small sample sizes (relative to dimensions), where practically none of the root solving techniques will be able to find numerical solutions for the simultaneous minimization problem.

Importantly, since our newly proposed algorithm is componentwise, it can also be applied to high dimensional set-ups ($n \ll p$) for analyzing massive datasets. However, to deal with sparsity, a crucial assumption in high dimensional literature, appropriate thresholding or regularization tools are needed to be incorporated.

In summary, the following positive points stand out among the benefits of our new proposal.

- **Computational Superiority:** The principal advantage of this new sequential method over the ordinary minimum DPD estimation procedure is the huge computational success of the componentwise philosophy. It offers fully converged estimates of the location vector and scale matrix even in case of small sample sizes and large values of the tuning parameter β. The ordinary minimum DPD method, unfortunately, cannot match this performance which provides the prime motivation of this work.

- **Improved Empirical Performance:** As reflected in the simulation outputs, the new sequential method performs comparatively better than the existing methodologies (MCD, MVE, S, etc.) in terms of bias and mean squared errors, especially for the covariance matrix estimators.

- **Ease of Extension to High Dimensions:** Although we develop a sequential method for robust and efficient estimation of multivariate location and scale under higher data dimensions in this work, our method can also be applied to standard high dimensional set-up, where number of parameters is greater than the sample

size. As long as we have $4-5$ observations, our method can technically be applied to a multivariate data of any dimension; however, we recommend to apply this method with at least a sample of size around 50 to get a reasonable accuracy of the parameter estimates.

The present chapter is organized as follows. The method along with its statistical and probabilistic background is described in Section 4.2 while its theoretical properties are provided in Sections 4.3 and 4.4. The asymptotic relative efficiencies of our estimators and the detailed structures of our influence functions are presented in special cases in Section 4.5. Simulation experiments and results are discussed in Section 4.6. In Section 4.7, we apply our method on a real life dataset on credit card transactions (with some fraudulent ones) to understand the capability of our method in maintaining desired robustness even in higher dimensions. All the proofs of the aforesaid theoretical results are provided in Section 4.8.

4.2 Model Set-up and the Proposed Estimation Procedure

4.2.1 Elliptically Symmetric Distributions

Let $\boldsymbol{Z} = (Z_1,\ Z_2, \cdots,\ Z_p)'$ be an absolutely continuous random vector in $\mathbb{R}^p$ which is spherically symmetric around the origin, that is,

$$\boldsymbol{Z} \stackrel{d}{=} \boldsymbol{P}\boldsymbol{Z} \tag{4.1}$$

for any orthogonal matrix $\boldsymbol{P}$ of dimension $p \times p$. The PDF of $\boldsymbol{Z}$ is of the form $k\psi(||z||)$, $z \in \mathbb{R}^p$ for a non-negative real valued function ψ and normalizing constant $k > 0$, where $||\cdot||$ represents the L_2-norm. By choosing $\boldsymbol{P} = -\boldsymbol{I}_p$ it can be shown that $E(\boldsymbol{Z}) = \boldsymbol{0}_p$. By choosing $\boldsymbol{P}$ to be different $p \times p$ permutation matrices, it can be shown that all the component variances of the random vector $\boldsymbol{Z}$ are same (c, say). Finally, by choosing $\boldsymbol{P} = Diag(1,\ 1, \cdots,\ 1,\ -1,\ 1, \cdots,\ 1)$ where the i-th diagonal element is -1, it can be shown that $Cov(Z_i,\ Z_j) = 0$ for any $j \neq i$. Thus it gives $E(\boldsymbol{Z}) = \boldsymbol{0}_p$ and $Var(\boldsymbol{Z}) = c\boldsymbol{I}_p$. We assume $c = 1$ for standardization. Let $\boldsymbol{\mu} \in \mathbb{R}^p$ and $\boldsymbol{\Sigma}$ be a symmetric, positive definite $p \times p$ matrix. We define the random vector $\boldsymbol{X}$ as,

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Z}, \tag{4.2}$$

where $\boldsymbol{\Sigma}^{\frac{1}{2}}$ is the positive square root of $\boldsymbol{\Sigma}$. Hence, $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and $Var(\boldsymbol{X}) = \boldsymbol{\Sigma}$. Then $\boldsymbol{X}$ is said to have an elliptically symmetric distribution, with the corresponding PDF of $\boldsymbol{X}$ being given by,

$$f_{\boldsymbol{\omega}}(\boldsymbol{x}) = \frac{k}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \psi\left(\sqrt{(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}\right),\ \boldsymbol{\omega} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{4.3}$$

We refer to Kelker (1970) [85] and Chmielewski (1981) [25] for further details about spherical and elliptical distributions. The multivariate normal distribution is an example of elliptically symmetric distributions with $k = (2\pi)^{-\frac{p}{2}}$ and $\psi(x) = exp(-\frac{x^2}{2})$.

4.2.2 Our Proposed Estimation Algorithm: The Sequential MDPDE

Let us introduce the notation $\mathcal{E}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to denote the family of elliptically symmetric distributions of dimension p, as described in Section 4.2.1, with mean $\boldsymbol{\mu} = (\mu_1,\ \mu_2, \ldots,\ \mu_p)'$, covariance matrix $\boldsymbol{\Sigma} = [[\sigma_{ij}]]$ and PDF $f_{\boldsymbol{\omega}}$, as in Equation (4.3). Note that, the correlation between the j-th and k-th components is given by $\rho_{jk} = \frac{\sigma_{jk}}{\sqrt{\sigma_{jj}\sigma_{kk}}}$. Now, suppose we have a random sample of multivariate observations $\{\boldsymbol{X}_1,\ \boldsymbol{X}_2, \ldots,\ \boldsymbol{X}_n\}$ (with $\boldsymbol{X}_i = (X_{i1},\ X_{i2}, \ldots,\ X_{ip})'$ for $i = 1,\ 2, \ldots,\ n$) from an unknown probability distribution with density g (distribution G) and we model this unknown g by the $\mathcal{E}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ model family. Our objective is to estimate this $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ and subsequently, the correlation matrix $\boldsymbol{R} = [[\rho_{jk}]]$. Thus, our parameter of interest is $\boldsymbol{\omega} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and the parameter space is $\boldsymbol{\Omega} = \Big\{\boldsymbol{\omega} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}) : \boldsymbol{\mu} \in \mathbb{R}^p,\ \boldsymbol{\Sigma} \in \mathbb{R}^{p\times p},\ \boldsymbol{\Sigma} \text{ is symmetric and positive definite}\Big\}$. Let us consider the problem of estimating $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ of the model distribution $\mathcal{E}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ based on the sample $\{\boldsymbol{X}_1,\ \boldsymbol{X}_2, \ldots,\ \boldsymbol{X}_n\}$. *Now, we additionally assume that, for each $j = 1, \ldots, p$, the marginal PDF of the j-th component be f_j which is fully characterized by the parameters μ_j and σ_{jj}.* Similarly, let the joint marginal PDF of the j-th and k-th components is f_{jk} which is fully characterized by the parameters $\mu_j,\ \sigma_{jj},\ \mu_k,\ \sigma_{kk},$ and ρ_{jk} for all $1 \leq j < k \leq p$. Such assumptions often hold for common elliptically symmetric distributions including the multivariate normal distribution; in the latter case f_j and f_{jk} are univariate and bivariate normal PDFs for all $1 \leq j < k \leq p$, respectively. Analogously, let g_j (G_j) and g_{jk} (G_{jk}) be the true (unknown) marginal density (distribution) of the j-th component, $j = 1, \ldots,\ p$ and the joint density (distribution) of the j-th and k-th components, respectively.

Now, in order to propose the new componentwise algorithm for estimating $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, let us introduce some new notation by systematically reparametrizing $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

For notational simplicity, we hereonwards denote the j-th component variance σ_{jj} by σ_j^2. The parameter of interest is $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p, \boldsymbol{\rho})$ with $\boldsymbol{\theta}_j = (\theta_{j1}, \theta_{j2}) = (\mu_j, \sigma_j^2)$ for $j = 1, \ldots, p$ and $\boldsymbol{\rho} = (\rho_{jk} : 1 \leq j < k \leq p)$ where $\rho_{jk} = Cor(X_{ij}, X_{ik}) = \frac{\sigma_{jk}}{\sigma_j \sigma_k}$ for $1 \leq j < k \leq p$. The parameter space is given by $\boldsymbol{\Theta} = \{\boldsymbol{\theta} = (\mu_1, \sigma_1^2, \ldots, \mu_p, \sigma_p^2, \rho_{jk} : 1 \leq j < k \leq p) : \mu_j \in \mathbb{R}, \sigma_j^2 > 0, j = 1, \ldots, p, \rho_{jk} \in [-1, 1], 1 \leq j < k \leq p\}$. Let,

$$H_{jn}(\boldsymbol{\theta}_j) = \int f_j^{1+\beta}(x, \boldsymbol{\theta}_j)\, dx - \left(1 + \frac{1}{\beta}\right) \frac{1}{n} \sum_{i=1}^{n} f_j^{\beta}(X_{ij}, \boldsymbol{\theta}_j) = \frac{1}{n} \sum_{i=1}^{n} V_j(X_{ij}, \boldsymbol{\theta}_j), \text{ and}$$

$$\begin{aligned} &H_{jkn}(\boldsymbol{\theta}_j, \boldsymbol{\theta}_k, \rho_{jk}) \\ &= \int f_{jk}^{1+\beta}(x_1, x_2, \boldsymbol{\theta}_j, \boldsymbol{\theta}_k, \rho_{jk})\, dx_1\, dx_2 - \left(1 + \frac{1}{\beta}\right) \frac{1}{n} \sum_{i=1}^{n} f_{jk}^{\beta}(X_{ij}, X_{ik}, \boldsymbol{\theta}_j, \boldsymbol{\theta}_k, \rho_{jk}) \\ &= \frac{1}{n} \sum_{i=1}^{n} V_{jk}(X_{ij}, X_{ik}, \boldsymbol{\theta}_j, \boldsymbol{\theta}_k, \rho_{jk}), \end{aligned}$$

where $V_j(x, \boldsymbol{\theta}_j) = \int f_j^{1+\beta}(t, \boldsymbol{\theta}_j)\, dt - \left(1 + \frac{1}{\beta}\right) f_j^{\beta}(x, \boldsymbol{\theta}_j)$, $j = 1, \ldots, p$ and $V_{jk}(x_1, x_2, \boldsymbol{\theta}_j, \boldsymbol{\theta}_k, \rho_{jk}) = \int f_{jk}^{1+\beta}(u, v, \boldsymbol{\theta}_j, \boldsymbol{\theta}_k, \rho_{jk})\, du\, dv - \left(1 + \frac{1}{\beta}\right) f_{jk}^{\beta}(x_1, x_2, \boldsymbol{\theta}_j, \boldsymbol{\theta}_k, \rho_{jk})$, $1 \leq j < k \leq p$. The proposed SMDPDEs are defined as,

$$\widehat{\boldsymbol{\theta}}_{jn} = \underset{\boldsymbol{\theta}_j}{\operatorname{argmin}}\, H_{jn}(\boldsymbol{\theta}_j),\ \widehat{\rho}_{jkn} = \underset{\rho_{jk} \in [-1,1]}{\operatorname{argmin}}\, H_{jkn}(\widehat{\boldsymbol{\theta}}_{jn}, \widehat{\boldsymbol{\theta}}_{kn}, \rho_{jk}) \tag{4.4}$$

and thus $\widehat{\boldsymbol{\theta}}_n = (\widehat{\boldsymbol{\theta}}_{1n}, \ldots, \widehat{\boldsymbol{\theta}}_{pn}, \hat{\rho}_{jkn} : 1 \leq j < k \leq p)$ for $n \geq 1$. We can now summarize these estimators to describe the new sequential algorithm, as follows.

Step 1: For each $j = 1, \ldots, p$, we estimate the j-th component mean μ_j and variance σ_j^2 by minimizing the marginal DPD based objective function as

$$(\widehat{\mu}_{jn}, \widehat{\sigma}_{jn}^2) = (\hat{\theta}_{j1n}, \hat{\theta}_{j2n}) = \underset{\boldsymbol{\theta}_j}{\operatorname{argmin}}\, H_{jn}(\boldsymbol{\theta}_j). \tag{4.5}$$

It gives us the marginal MDPDEs of the mean and the variance corresponding to the j-th component.

Step 2: For any $1 \leq j < k \leq p$, we estimate the correlation coefficient of the j-th and k-th components, ρ_{jk}, by using the (marginal) MDPDEs of component means and variances obtained in Step 1. In particular,we estimate the correlation coefficient ρ_{jk} by minimizing the joint (bivariate) DPD based objective function

corresponding to the j-th and k-th components after plugging in the (marginal) MDPDEs of component means and variances, obtained in Step 1, as

$$\widehat{\rho}_{jkn} = \underset{\rho_{jk}\in[-1,1]}{\operatorname{argmin}} H_{jkn}(\hat{\boldsymbol{\theta}}_{jn}, \hat{\boldsymbol{\theta}}_{kn}, \rho_{jk}). \tag{4.6}$$

The resulting estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, as obtained from the above algorithm via $\widehat{\boldsymbol{\mu}}_n = (\widehat{\mu}_{1n},\ \widehat{\mu}_{2n},\ \ldots, \widehat{\mu}_{pn})'$ and $\widehat{\boldsymbol{\Sigma}}_n = ((\widehat{\sigma}_{jkn}))$ with $\widehat{\sigma}_{jkn} = \widehat{\sigma}_{jn}\widehat{\sigma}_{kn}\widehat{\rho}_{jkn}$ ($\widehat{\sigma}^2_{jn} = \widehat{\sigma}_{jjn}, 1 \le j \le p$), are referred to as the "sequential minimum DPD estimators" (SMDPDE). The corresponding SMDPDE of the correlation matrix is given by $\widehat{\boldsymbol{R}}_n = ((\widehat{\rho}_{jkn}))$ with $\widehat{\rho}_{jjn} = 1$ for all $j = 1, \ldots, p$. Note that, we have estimated the component means and variances at first and then use these estimates to estimate the correlation coefficients further. So, the unknown correlation coefficients cannot affect the estimation of component means and variances. Following this observation, we can expect that our estimators of component means and variances will achieve greater efficiency than the usual (multivariate) MDPDE (discussed in Chapter 2); this advantage would be further illustrated in the subsequent sections, concretely under normal model distributions.

Remark 4.1. *The optimization problem in Equation (*4.5*) does not have any closed form solution. Standard numerical procedures like Newton-Raphson, IRLS, etc, can be used to iteratively solve this optimization problem. In our numerical illustrations, we have used a suitable IRLS algorithm which is a simplified version of Algorithm* 2.1 *in Chapter 2; for general elliptical families the derivation of such iterative procedure is algebraically similar. The second constrained optimization problem in Equation (4.6) can be solved easily using any standard statistical software packages, e.g., the* optim *function in R software (R core team* (2018)*, [123]). In particular, we have utilised the* Brent *method within the* optim *function to solve the constrained optimization stated in Equation (4.6).*

Remark 4.2. *We will prove that our proposed SMDPDE of the covariance matrix is asymptotically positive definite (Theorem 4.4). However, it is to be noted that our method does not necessarily guarantee the positive definiteness of the covariance matrix estimator $\widehat{\boldsymbol{\Sigma}}_n$ in every finite sample case. This is a common problem of any componentwise procedures of covariance matrix estimation (see for example, Ma and Genton* (2001) *[101]). Some possible corrections were suggested by Rousseeuw and Molenberghs* (1993) *[132] and Higham* (2002) *[72]. In particular, the second one proposed a numerical algorithm to compute the nearest positive semi-definite correlation matrix to a given*

symmetric matrix which is approximately a correlation matrix. Such a correction needs to be incorporated in order to make the proposed SMDPDE $\widehat{\boldsymbol{\Sigma}}_n$ positive definite in practical applications (i.e., real datasets with comparatively small sample sizes) if the need arises. The output of the Higham's method (if applied on the estimated correlation matrix which is not positive definite) will be positive semi-definite (with some eigenvalues as zero). A further eigenvalue truncation step can be applied to make it positive definite which replaces the zero eigenvalues by some small positive constant. This entire two-stage procedure (Higham's method followed by the eigenvalue truncation step) can be implemented in the R software through the nearPD *function available in the* Matrix *package (Bates and Maechler* (2019) *[13]) which will be utilized as necessary.*

In the next section, we will show that the proposed SMDPDEs are consistent even if the true density (g) does not belong to the model family $\mathcal{E}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. However, in such a case, the SMDPDE will converge to the best fitting value of the parameter $\boldsymbol{\theta^g} = (\boldsymbol{\theta_1^g}, \ldots, \boldsymbol{\theta_p^g}, \rho_{jk}^g : 1 \leq j < k \leq p)$, defined as,

$$\boldsymbol{\theta_j^g} = \underset{\boldsymbol{\theta}_j}{\operatorname{argmin}}\ H_j(\boldsymbol{\theta}_j),\ \rho_{jk}^g = \underset{\rho_{jk} \in [-1,1]}{\operatorname{argmin}}\ H(\boldsymbol{\theta_j^g}, \boldsymbol{\theta_k^g}, \rho_{jk}), \tag{4.7}$$

where $H_j(\boldsymbol{\theta}_j) = E_{G_j}(V_j(X_{ij}, \boldsymbol{\theta}_j)),\ j = 1, \ldots,\ p,\ H_{jk}(\boldsymbol{\theta}_j, \boldsymbol{\theta}_k, \rho_{jk}) = E_{G_{jk}}(V_{jk}(X_{ij}, X_{ik}, \boldsymbol{\theta}_j, \boldsymbol{\theta}_k, \rho_{jk})),\ 1 \leq j < k \leq p$.

4.3 Asymptotic Properties

4.3.1 Technical Assumptions

In order to establish the asymptotic properties of our estimators, we need the following technical assumptions. The first three assumptions have been taken from Basu et al. (2011) [12] (Section 9.2.1) with little modifications.

Assumption 4.1. *There is an open subset γ_0 of the parameter space $\boldsymbol{\Theta}$ such that for almost all $(x_1, x_2) \in \mathbb{R}^2$ and all $\boldsymbol{\theta} \in \gamma_0$, the densities f_j $(1 \leq j \leq p)$ and f_{jk} $(1 \leq j < k \leq p)$ are four times differentiable with respect to $\boldsymbol{\theta}$ and the fourth partial derivatives are continuous with respect to $\boldsymbol{\theta}$.*

Assumption 4.2. *The integrals $\int f_j^{1+\beta}(x, \boldsymbol{\theta}_j)\ dx$, $\int f_{jk}^{1+\beta}(x_1, x_2, \boldsymbol{\theta}_j, \boldsymbol{\theta}_k, \rho_{jk})\ dx_1 dx_2$ and $\int f_j^{\beta}(x, \boldsymbol{\theta}_j) g_j(x)\ dx$, $\int f_{jk}^{\beta}(x_1, x_2, \boldsymbol{\theta}_j, \boldsymbol{\theta}_k, \rho_{jk}) g_{jk}(x_1, x_2)\ dx_1 dx_2$ are differentiable four times with respect to $\boldsymbol{\theta}$ for $1 \leq j \leq p$, $1 \leq j < k \leq p$ and the derivatives can be taken under the integral signs.*

Assumption 4.3. *There exist L_1 bounded functions $M^j_{k_1l_1m_1}(x)$ which satisfy,*

$$|\nabla_{k_1l_1m_1}V_j(x,\boldsymbol{\theta}_j)| < M^j_{k_1l_1m_1}(x),$$

for all $\boldsymbol{\theta}_j$, $(j=1,\cdots,p)$ and $x\in\mathbb{R}$.

Assumption 4.4. *Let, $U_{\rho_{jk}}(x_1,x_2,\rho_{jk}\mid\boldsymbol{\theta}_j,\boldsymbol{\theta}_k)=\frac{\partial log\, f_{jk}(x_1,x_2,\boldsymbol{\theta}_j,\boldsymbol{\theta}_k,\rho_{jk})}{\partial\rho_{jk}}$. Then, there exist L_1 bounded functions $M_{jk}(x_1,x_2,\rho_{jk})$ (with a finite integral) and $N_{jk}(x_1,x_2,\rho_{jk})$ which satisfy*

$$\Big|\frac{\partial}{\partial\theta_{j_1k_1}}f^{1+\beta}_{jk}(x_1,x_2,\boldsymbol{\theta}_j,\boldsymbol{\theta}_k,\rho_{jk})U_{\rho_{jk}}(x_1,x_2,\rho_{jk}|\boldsymbol{\theta}_j,\boldsymbol{\theta}_k)\Big| < M_{jk}(x_1,x_2,\rho_{jk}),$$
$$\Big|\frac{\partial}{\partial\theta_{j_1k_1}}f^{\beta}_{jk}(x_1,x_2,\boldsymbol{\theta}_j,\boldsymbol{\theta}_k,\rho_{jk})U_{\rho_{jk}}(x_1,x_2,\rho_{jk}|\boldsymbol{\theta}_j,\boldsymbol{\theta}_k)\Big| < N_{jk}(x_1,x_2,\rho_{jk}),$$

for all $\boldsymbol{\theta}\in\gamma_0$, $\rho_{jk}\in[-1,1]$ and $(x_1,x_2)\in\mathbb{R}^2$, where $(j_1,k_1)\in\{(j,1),(j,2),(k,1),(k,2)\}$, $1\le j<k\le p$.

Assumption 4.5. *There exist L_1 bounded functions $v^{jk}_1(x_1,x_2,\boldsymbol{\theta}_j,\boldsymbol{\theta}_k)$ and $v^{jk}_2(x_1,x_2,\boldsymbol{\theta}_j,\boldsymbol{\theta}_k)$ which satisfy*

$$\Big|\frac{\partial^3V_{jk}(x_1,x_2,\boldsymbol{\theta}_j,\boldsymbol{\theta}_k,\rho_{jk})}{\partial^3\rho_{jk}}\Big| < v^{jk}_1(x_1,x_2,\boldsymbol{\theta}_j,\boldsymbol{\theta}_k),$$
$$\Big|\frac{\partial^4V_{jk}(x_1,x_2,\boldsymbol{\theta}_j,\boldsymbol{\theta}_k,\rho_{jk})}{\partial^3\rho_{jk}\partial\theta_{j_1k_1}}\Big| < v^{jk}_2(x_1,x_2,\boldsymbol{\theta}_j,\boldsymbol{\theta}_k),$$

for all $\boldsymbol{\theta}\in\gamma_0$, $\rho_{jk}\in[-1,1]$ and $(x_1,x_2)\in\mathbb{R}^2$, where $(j_1,k_1)\in\{(j,1),(j,2),(k,1),(k,2)\}$, $1\le j<k\le p$.

Some remarks on these assumptions are provided in Section 6. It is important to note that, we have derived the estimator $\widehat{\boldsymbol{\theta}}_{jn}$ marginally by assuming $\{X_{1j},\ldots,X_{nj}\}$ as a univariate random sample $(1\le j\le p)$. The weak consistency of $\widehat{\boldsymbol{\theta}}_{jn}$ $(1\le j\le p)$ is already established in Basu et al. (2011) [12] (Theorem 9.2) and Basu et al. (1998) [10] (Theorem 2) under appropriate assumptions (which are covered by Assumptions 4.1-4.3). Thus, with probability tending to 1, there exists a sequence of estimators $\{\widehat{\boldsymbol{\theta}}_{jn}\}_{n=1}^{\infty}$, which satisfy,

$$\widehat{\boldsymbol{\theta}}_{jn}\xrightarrow{p}\boldsymbol{\theta}^{\boldsymbol{g}}_j,\ 1\le j\le p. \tag{4.8}$$

So, assuming the existence of these estimators, it remains to prove the consistency of

$\widehat{\rho}_{jkn}$, defined as in Equation (4.6). Then, utilizing consistency of these estimators, we will establish their asymptotic normality.

Since the proposed estimating algorithm is componentwise, it is sufficient to prove the theoretical results under bivariate case ($p = 2$) at first and then extend the results to general p-dimensional cases ($p \geq 3$).

4.3.2 Properties of the estimators in Bivariate Case ($p = 2$)

As we have decided in the last paragraph, we now present the consistency and asymptotic normality of the SMDPDEs only for the bivariate case. So, here $p = 2$ and hence the parameters are $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ and ρ_{12} only. But for further notational simplicity, we will drop the suffix "12" from the required notations and expressions needed as the only possible pair of components is the pair of the first and second components under $p = 2$. Precisely, we simply denote $\rho = \rho_{12}$, $\widehat{\rho}_n = \widehat{\rho}_{12n}$, $\rho^g = \rho^g_{12}$, $f = f_{12}$, $H_n = H_{n12}$, $H = H_{12}$ and $V = V_{12}$. In fact, these notations will also be used in Sections 4.4 (the influence function analysis under $p = 2$) and 4.5 (the normal model family).

Now, by the aforesaid notations, $\widehat{\rho}_n$ can be described as the solution of,

$$\frac{\partial H_n(\widehat{\boldsymbol{\theta}}_{1n}, \widehat{\boldsymbol{\theta}}_{2n}, \rho)}{\partial \rho} = 0. \tag{4.9}$$

Theorem 4.1. *[Consistency] Under Assumptions 4.1-4.5, Equation (4.9) has a consistent sequence of solution* $\{\hat{\rho}_n\}$ *with probability tending to* 1.

Our next job is to establish the asymptotic normality of $\widehat{\boldsymbol{\theta}}_n$ to study the asymptotic variances (thus asymptotic efficiencies) of the individual SMDPDEs. To state and prove the result, we need to introduce the following expressions. Let us define the random vector

$$\boldsymbol{\zeta} = \Big(\frac{\partial\, V_1(X_{11}, \boldsymbol{\theta}_1)}{\partial\, \theta_{11}}, \frac{\partial\, V_1(X_{11}, \boldsymbol{\theta}_1)}{\partial\, \theta_{12}}, \frac{\partial\, V_2(X_{12}, \boldsymbol{\theta}_2)}{\partial\, \theta_{21}}, \frac{\partial\, V_2(X_{12}, \boldsymbol{\theta}_2)}{\partial\, \theta_{22}}, \frac{\partial\, V(X_{11}, X_{12}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \rho)}{\partial\, \rho} \Big) \Big|'_{\boldsymbol{\theta} = \boldsymbol{\theta}^g}$$

and its covariance matrix is given by $\boldsymbol{\Gamma_0} = Var_g(\boldsymbol{\zeta})$. Let us also define the matrix $\boldsymbol{B}$ as,

$$\boldsymbol{B} = \begin{bmatrix} b_{11}^{(1)} & b_{12}^{(1)} & 0 & 0 & 0 \\ b_{21}^{(1)} & b_{22}^{(1)} & 0 & 0 & 0 \\ 0 & 0 & b_{11}^{(2)} & b_{12}^{(2)} & 0 \\ 0 & 0 & b_{21}^{(2)} & b_{22}^{(2)} & 0 \\ e_{11} & e_{12} & e_{21} & e_{22} & a \end{bmatrix},$$

where, $b_{jk}^{(1)} = E_g\left(\frac{\partial^2 V_1(X_{11},\boldsymbol{\theta}_1)}{\partial\,\theta_{1j}\,\partial\,\theta_{1k}}\right)\Big|_{\boldsymbol{\theta}_1=\boldsymbol{\theta}_1^{\boldsymbol{g}}}$, $b_{jk}^{(2)} = E_g\left(\frac{\partial^2 V_2(X_{12},\boldsymbol{\theta}_2)}{\partial\,\theta_{2j}\,\partial\,\theta_{2k}}\right)\Big|_{\boldsymbol{\theta}_2=\boldsymbol{\theta}_2^{\boldsymbol{g}}}$,
$e_{jk} = E_g\left(\frac{\partial^2 V(\boldsymbol{X}_1,\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\rho)}{\partial\,\theta_{jk}\,\partial\,\rho}\right)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{\boldsymbol{g}}}$ for $j,k=1,2$ and $a = E_g\left(\frac{\partial^2 V(\boldsymbol{X}_1,\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\rho)}{\partial\,\rho^2}\right)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{\boldsymbol{g}}}$.

Theorem 4.2. *[Asymptotic Normality] Under Assumptions 4.1-4.5,* $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^g)$ *converges in distribution to a normal distribution with zero mean and covariance matrix* $\boldsymbol{\Gamma} = \boldsymbol{B}^{-1}\boldsymbol{\Gamma}_0\boldsymbol{B}^{-1'}$.

4.3.3 Properties of the Estimators under General Multivariate Set-up ($p \geq 3$)

Our next task is to extend the previous results in general p-dimensional scenarios. Under the general p-dimensional case, the parameter $\boldsymbol{\theta} = (\boldsymbol{\theta}_1,\ \boldsymbol{\theta}_2, \ldots,\ \boldsymbol{\theta}_p,\ \rho_{12}, \ldots,\ \rho_{1p},\ \rho_{23}, \cdots,\ \rho_{2p}, \ldots,\ \rho_{p-1,p})$ is $P = \frac{p^2+3p}{2}$-dimensional. Now we have the following result.

Theorem 4.3. *[Multivariate Consistency and Asymptotic Normality] For the general p-dimensional case, and under Assumptions 4.1-4.5, the estimator* $\widehat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}^{\boldsymbol{g}}$ *and* $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^{\boldsymbol{g}}) \xrightarrow{d} N(\mathbf{0}, \widetilde{\boldsymbol{\Gamma}})$, *where* $\widetilde{\boldsymbol{\Gamma}} = \widetilde{\boldsymbol{B}}^{-1}\widetilde{\boldsymbol{\Gamma}}_{\mathbf{0}}\widetilde{\boldsymbol{B}}^{-1'}$.

Here, the matrix $\widetilde{\boldsymbol{\Gamma}_{\mathbf{0}}}$ is the covariance matrix of the random vector

$$\Bigg(\frac{\partial\, V_1(X_{11},\boldsymbol{\theta}_1)}{\partial\,\theta_{11}}, \frac{\partial\, V_1(X_{11},\boldsymbol{\theta}_1)}{\partial\,\theta_{12}}, \cdots, \frac{\partial\, V_p(X_{1p},\boldsymbol{\theta}_p)}{\partial\,\theta_{p1}}, \frac{\partial\, V_p(X_{p2},\boldsymbol{\theta}_p)}{\partial\,\theta_{p2}},$$
$$\frac{\partial\, V_{12}((X_{11},X_{12}),\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\rho_{12})}{\partial\,\rho_{12}}, \cdots, \frac{\partial\, V_{p-1,p}((X_{1,p-1},X_{1p}),\boldsymbol{\theta}_{p-1},\boldsymbol{\theta}_p,\rho_{p-1,p})}{\partial\,\rho_{p-1,p}}\Bigg)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{\boldsymbol{g}}}'$$

and

$$\widetilde{\boldsymbol{B}} = \begin{bmatrix}
b_{11}^{(1)} & b_{12}^{(1)} & 0 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & & \\
b_{21}^{(1)} & b_{22}^{(1)} & 0 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & & & & \\
0 & 0 & b_{11}^{(2)} & b_{12}^{(2)} & \cdots & \cdots & \cdots & \cdots & & & & & \\
0 & 0 & b_{21}^{(2)} & b_{22}^{(2)} & \cdots & \cdots & \cdots & \cdots & & & & & \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & & & & \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & & & & \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & & & & \\
0 & 0 & \cdots & \cdots & \cdots & b_{11}^{(p-1)} & b_{12}^{(p-1)} & \cdots & \cdots & & & & \\
0 & 0 & \cdots & \cdots & \cdots & b_{21}^{(p-1)} & b_{22}^{(p-1)} & \cdots & \cdots & & & & \\
0 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & b_{11}^{(p)} & b_{12}^{(p)} & \cdots & \cdots & & \\
0 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & b_{21}^{(p)} & b_{22}^{(p)} & \cdots & \cdots & & \\
e_{12}^{(1)} & e_{12}^{(2)} & e_{12}^{(3)} & e_{12}^{(4)} & 0 & 0 & \cdots & 0 & \cdots & a_{12} & 0 & \cdots & \cdots \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & 0 & \cdots & \cdots & \cdots & e_{p-1,p}^{(1)} & e_{p,p-1}^{(2)} & e_{p-1,p}^{(3)} & e_{p-1,p}^{(4)} & 0 & \cdots & \cdots & a_{p-1,p}
\end{bmatrix}.$$

with $b_{k_1 l_1}^{(j)} = E_{g_j}\left(\frac{\partial^2 V_j(X_{1j},\boldsymbol{\theta}_j)}{\partial\,\theta_{jk_1}\,\partial\,\theta_{jl_1}}\right)\Big|_{\boldsymbol{\theta}_j=\boldsymbol{\theta}_j^{\boldsymbol{g}}}$ for $k_1 = l_1 = 1, 2$,

$e_{jk}^{(l)} = E_{g_{jk}}\left(\frac{\partial^2 V_{jk}((X_{1j},X_{1k}),\boldsymbol{\theta}_j,\boldsymbol{\theta}_k,\rho_{jk})}{\partial\,\theta_{j1}\,\partial\,\rho_{jk}}\right)\Big|_{(\boldsymbol{\theta}_j,\boldsymbol{\theta}_k)=(\boldsymbol{\theta^g}_j,\boldsymbol{\theta^g}_k),\rho_{jk}=\rho_{jk}^g}$ for $l = 1$,

$e_{jk}^{(l)} = E_{g_{jk}}\left(\frac{\partial^2 V_{jk}((X_{1j},X_{1k}),\boldsymbol{\theta}_j,\boldsymbol{\theta}_k,\rho_{jk})}{\partial\,\theta_{j2}\,\partial\,\rho_{jk}}\right)\Big|_{(\boldsymbol{\theta}_j,\boldsymbol{\theta}_k)=(\boldsymbol{\theta^g}_j,\boldsymbol{\theta^g}_k),\rho_{jk}=\rho_{jk}^g}$ for $l = 2$,

$e_{jk}^{(l)} = E_{g_{jk}}\left(\frac{\partial^2 V_{jk}((X_{1j},X_{1k}),\boldsymbol{\theta}_j,\boldsymbol{\theta}_k,\rho_{jk})}{\partial\,\theta_{k1}\,\partial\,\rho_{jk}}\right)\Big|_{(\boldsymbol{\theta}_j,\boldsymbol{\theta}_k)=(\boldsymbol{\theta^g}_j,\boldsymbol{\theta^g}_k),\rho_{jk}=\rho_{jk}^g}$ for $l = 3$,

$e_{jk}^{(l)} = E_{g_{jk}}\left(\frac{\partial^2 V_{jk}((X_{1j},X_{1k}),\boldsymbol{\theta}_j,\boldsymbol{\theta}_k,\rho_{jk})}{\partial\,\theta_{k2}\,\partial\,\rho_{jk}}\right)\Big|_{(\boldsymbol{\theta}_j,\boldsymbol{\theta}_k)=(\boldsymbol{\theta^g}_j,\boldsymbol{\theta^g}_k),\rho_{jk}=\rho_{jk}^g}$ for $l = 4$ and

$a_{jk} = E_{g_{jk}}\left(\frac{\partial^2 V_{jk}((X_{1j},X_{1k}),\boldsymbol{\theta}_j,\boldsymbol{\theta}_k,\rho_{jk})}{\partial\,\rho_{jk}^2}\right)\Big|_{(\boldsymbol{\theta}_j,\boldsymbol{\theta}_k)=(\boldsymbol{\theta^g}_j,\boldsymbol{\theta^g}_k),\rho_{jk}=\rho_{jk}^g}$, $j < k = 1, \ldots, p$.

Our next result is to establish the asymptotic positive definiteness of our covariance matrix estimator $\widehat{\boldsymbol{\Sigma}}_n = ((\widehat{\sigma}_{nij}))_{i,j=1}^p$. The true covariance matrix is $\boldsymbol{\Sigma^g} = ((\sigma_{ij}^g))_{i,j=1}^p$. It will be established as a consequence of Theorem 4.1 under a specific assumption. We assume the following.

Assumption 4.6. *The minimum eigenvalue of* $\boldsymbol{\Sigma^g}$ *is bounded away from* 0*, that is, there exists a positive constant* c *such that* $\lambda_{(1)}^g \geq c$ *where* $\lambda_{(1)}^g$ *is the minimum eigenvalue of* $\boldsymbol{\Sigma^g}$*.*

Theorem 4.4. *[Asymptotic Positive Definiteness] Under Assumptions 4.1–4.6, the estimator* $\widehat{\boldsymbol{\Sigma}}_n$ *is positive definite with probability tending to* 1 *as* $n \to \infty$*.*

Detailed mathematical derivations of all the aforesaid theoretical results are provided in Section 4.8.

4.4 Influence Functions of the SMDPDEs

As noted previously, it is enough to study the properties of the SMDPDE in the bivariate case as our procedure is indeed based on all possible bivariate combinations from given multivariate data; in particular, the same is true for the influence function as well. Let us use the same notations as in Section 4.3.2. To derive the influence functions, we have to study the algebraic relationships among the statistical functionals corresponding to our SMDPDEs (i.e., the corresponding best fitting parameters, observed as functionals of the true density g), viz., θ_{11}^g, θ_{12}^g, θ_{21}^g, θ_{22}^g, and ρ^g (i.e., μ_1^g, $\sigma_1^{g^2}$, μ_2^g, $\sigma_2^{g^2}$ and ρ^g, respectively). By definition, θ_{jk}^g $(j,k=1,2)$ can be derived by solving,

$$\frac{\partial}{\partial\,\theta_{jk}}\left[\int f_j^{1+\beta}(x,\boldsymbol{\theta}_j)\,dx-\left(1+\frac{1}{\beta}\right)\int f_j^{\beta}(x,\boldsymbol{\theta}_j)g_j(x)\,dx\right]=0 \text{ for } j,k=1,2.$$

Since the model density f belongs to a location-scale family, $\int f_j^{1+\beta}(x,\boldsymbol{\theta}_j)\,dx$ is independent of θ_{j1} for $j=1,2$. So, the aforesaid equation boils down to

$$\int f_j^{\beta}(x,\boldsymbol{\theta}_j)U_{\theta_{j1}}(x,\boldsymbol{\theta}_j)g_j(x)\,dx=0,\ j=1,2, \tag{4.10}$$

and

$$\left[\int f_j^{1+\beta}(x,\boldsymbol{\theta}_j)U_{\theta_{j2}}(x,\boldsymbol{\theta}_j)\,dx-\int f_j^{\beta}(x,\boldsymbol{\theta}_j)U_{\theta_{j2}}(x,\boldsymbol{\theta}_j)g_j(x)\,dx\right]=0,\ j=1,2, \tag{4.11}$$

where $U_{\theta_{jk}}(x,\boldsymbol{\theta}_j)=\frac{\partial}{\partial\,\theta_{jk}}\log f_j(x,\boldsymbol{\theta}_j)$ for $j,k=1,2$.
The functional ρ^g can be derived by solving,

$$\begin{aligned}&\Bigg[\int f^{1+\beta}(x_1,x_2,\boldsymbol{\theta_1^g},\boldsymbol{\theta_2^g},\rho)U_\rho(x_1,x_2,\boldsymbol{\theta_1^g},\boldsymbol{\theta_2^g},\rho)\,dx_1dx_2-\\&\int f^{\beta}(x_1,x_2,\boldsymbol{\theta_1^g},\boldsymbol{\theta_2^g},\rho)U_\rho(x_1,x_2,\boldsymbol{\theta_1^g},\boldsymbol{\theta_2^g},\rho)g(x_1,x_2)\,dx_1dx_2\Bigg]=0,\end{aligned} \tag{4.12}$$

where $U_\rho(x_1, x_2, \boldsymbol{\theta}_1^{\boldsymbol{g}}, \boldsymbol{\theta}_2^{\boldsymbol{g}}, \rho) = \frac{\partial}{\partial \rho} \log\, f(x_1, x_2, \boldsymbol{\theta}_1^{\boldsymbol{g}}, \boldsymbol{\theta}_2^{\boldsymbol{g}}, \rho)$. To derive the influence functions, our first step is to contaminate the true distribution at a prefixed point in $\mathbb{R}^2$. In particular, the contaminated distribution function is given by $G_\epsilon = (1-\epsilon)G + \epsilon \wedge_{\boldsymbol{y}}$ where $\boldsymbol{y} = (y_1, y_2)$ is the point of contamination. Note that, the marginals corresponding to G_ϵ would be $G_{j\epsilon} = (1-\epsilon)G_j + \epsilon \wedge_{y_j}$ for $j = 1, 2$. Let us also assume that $\theta_{jk\epsilon}$ for $j, k = 1, 2$ and ρ_ϵ are the best fitting parameters corresponding to the contaminated distribution, i.e., G_ϵ. Now, by definition, the influence functions of the functionals can be derived by solving the Equation System (4.13).

$$
\begin{aligned}
&\frac{\partial}{\partial \epsilon}\left[\int f_1^\beta(x, \boldsymbol{\theta}_{1\epsilon}) U_{\theta_{11\epsilon}}(x, \boldsymbol{\theta}_{1\epsilon}) dG_{1\epsilon}(x)\right]\Big|_{\epsilon=0} = 0,\\
&\frac{\partial}{\partial \epsilon}\left[\int f_1^{1+\beta}(x, \boldsymbol{\theta}_{1\epsilon}) U_{\theta_{12\epsilon}}(x, \boldsymbol{\theta}_{1\epsilon})\, dx - \int f_1^\beta(x, \boldsymbol{\theta}_{1\epsilon}) U_{\theta_{12\epsilon}}(x, \boldsymbol{\theta}_{1\epsilon}) dG_{1\epsilon}(x)\right]\Big|_{\epsilon=0} = 0,\\
&\frac{\partial}{\partial \epsilon}\left[\int f_2^\beta(x, \boldsymbol{\theta}_{2\epsilon}) U_{\theta_{21\epsilon}}(x, \boldsymbol{\theta}_{2\epsilon}) dG_{2\epsilon}(x) = 0\right]\Big|_{\epsilon=0} = 0,\\
&\frac{\partial}{\partial \epsilon}\left[\int f_2^{1+\beta}(x, \boldsymbol{\theta}_{2\epsilon}) U_{\theta_{22\epsilon}}(x, \boldsymbol{\theta}_{2\epsilon})\, dx - \int f_2^\beta(x, \boldsymbol{\theta}_{2\epsilon}) U_{\theta_{22\epsilon}}(x, \boldsymbol{\theta}_2) dG_{2\epsilon}(x)\right]\Big|_{\epsilon=0} = 0,\\
&\frac{\partial}{\partial \epsilon}\Bigg[\int f^{1+\beta}(x_1, x_2, \boldsymbol{\theta}_{1\epsilon}, \boldsymbol{\theta}_{2\epsilon}, \rho_\epsilon) U_{\rho_\epsilon}(x_1, x_2, \boldsymbol{\theta}_{1\epsilon}, \boldsymbol{\theta}_{2\epsilon}, \rho_\epsilon)\, dx_1 dx_2\\
&- \int f^\beta(x_1, x_2, \boldsymbol{\theta}_{1\epsilon}, \boldsymbol{\theta}_{2\epsilon}, \rho_\epsilon) U_{\rho_\epsilon}(x_1, x_2, \boldsymbol{\theta}_{1\epsilon}, \boldsymbol{\theta}_{2\epsilon}, \rho_\epsilon) dG_\epsilon(x_1, x_2)\Bigg]\Big|_{\epsilon=0} = 0.
\end{aligned}
\tag{4.13}
$$

Further simplification of the system (4.13) is not possible in general for any arbitrary elliptically symmetric probability model but can be done in specific cases. The simplified algebraic forms of the aforesaid influence functions are derived for the normal case in Section 4.5.2 where a specific example is also presented pictorially to illustrate its behaviour indicating robustness.

4.5 Example: Normal Model Family

4.5.1 Asymptotic Relative Efficiencies

Now we study asymptotic relative efficiencies (ARE, with respect to the MLEs) of our estimators in case of normal models. We assume that the true distribution is bivariate normal with component means 0 and 0, component variances 1 and 1 and correlation coefficient ρ. We will consider seven different values of ρ, namely, -0.7, -0.5, -0.3 ,0, 0.3, 0.5 and 0.7 representing the true correlation coefficient. This range of the correlation coefficient is used to observe the nature of the ARE of the correlation estimator under both high and low correlation structures.

From Basu et al. (2011) [12], it follows that the asymptotic variance of $\sqrt{n}\widehat{\theta}_{j1}$

is $\left(1+\frac{\beta^2}{1+2\beta}\right)^{\frac{3}{2}}\theta_{j2}^g$ and $\left(1+\frac{\beta^2}{1+2\beta}\right)^{2}\theta_{j2}^g$ $(j=1,2)$, respectively for the SMDPDE and MDPDE. Clearly, the ARE of the sequential method is higher than that of the ordinary method, at least for the component mean estimators. Explicit calculations of the asymptotic variances of the component variance and correlation estimators in case of both SMDPDE and MDPDE are extremely cumbersome. Thus, the algebraic forms of these asymptotic variances are provided in Section 4.8.5. The AREs (in percentage) of our estimators (MDPDEs and SMDPDEs) are tabulated in Tables 4.1 and 4.2.

Estimators	Methods	β				
		0	0.1	0.3	0.5	0.7
Mean	SMDPDE	100.000	98.717	92.081	83.822	75.700
	MDPDE	100.000	98.328	89.606	78.989	68.966
Variance	SMDPDE	100.000	97.561	85.507	73.046	63.452
	MDPDE	100.000	97.135	83.368	69.300	58.207

Table 4.1: AREs (in percentage) of component mean and variance estimators.

	β				
ρ^g	0.0	0.1	0.3	0.5	0.7
-0.7	100(100)	97.378(97.378)	84.967(84.691)	70.845(70.270)	59.361(57.269)
-0.5	100(100)	97.574(97.574)	84.789(84.917)	70.375(70.287)	58.161(57.332)
-0.3	100(100)	97.527(97.527)	84.663(84.836)	69.992(70.229)	57.222(57.261)
0.0	100(100)	97.561(97.561)	84.890(84.890)	70.225(70.225)	57.274(57.274)
0.3	100(100)	97.527(97.527)	84.663(84.836)	69.992(70.229)	57.222(57.261)
0.5	100(100)	97.574(97.574)	84.789(84.917)	70.375(70.287)	58.161(57.332)
0.7	100(100)	97.378(97.378)	84.967(84.691)	70.845(70.270)	59.361(57.269)

Table 4.2: AREs (in percentage) of correlation estimators for the SMDPDE; the same for the corresponding MDPDEs are given in parentheses.

4.5.2 Influence Functions

After simplifying and solving the system (4.13) in case of bivariate normal model family, we have the following influence functions of the respective functionals.

$$IF(\theta_{j1}, y_j, g_j) = (\theta^g_{j2})^{\frac{\beta}{2}}(1+\beta)^{\frac{3}{2}}(2\pi)^{\frac{\beta}{2}}(y_j - \theta^g_{j1}) f_j^{\beta}(y_j, \boldsymbol{\theta^g_j}),\ j = 1, 2,$$

$$IF(\theta_{j2}, y_j, g_j) = \frac{f_j^{\beta}(y_j, \boldsymbol{\theta^g_j})(-\theta^g_{j2} + (y_j - \theta^g_{j1})^2) - \int f_j^{1+\beta}(x, \boldsymbol{\theta^g_j})(-\theta^g_{j2} + (x - \theta^g_{j1})^2)\, dx}{\frac{1}{2}\int f_j^{1+\beta}(x, \boldsymbol{\theta^g_j}) \left(\frac{(x-\theta^g_{j1})^2}{\theta^g_{j2}} - 1\right)^2 dx}, \qquad j = 1, 2,$$

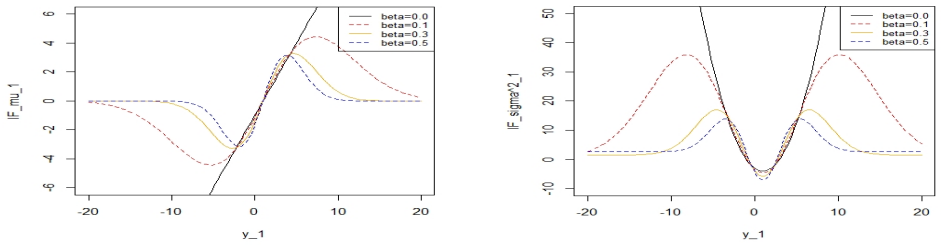
$$IF(\rho, y_1, y_2, g) = \frac{f^{\beta}(y_1, y_2, \boldsymbol{\theta^g_1}, \boldsymbol{\theta^g_2}, \rho^g) U_{\rho^g}(y_1, y_2, \boldsymbol{\theta^g_1}, \boldsymbol{\theta^g_2}, \rho^g)}{\int B(x_1, x_2) f^{\beta}(x_1, x_2, \boldsymbol{\theta^g_1}, \boldsymbol{\theta^g_2}, \rho^g) U_{\rho^g}(x_1, x_2, \boldsymbol{\theta^g_1}, \boldsymbol{\theta^g_2}, \rho^g)\, dx_1 dx_2}$$
$$- \frac{\int (1 + A(x_1, x_2, y_1, y_2)) f^{\beta}(x_1, x_2, \boldsymbol{\theta^g_1}, \boldsymbol{\theta^g_2}, \rho^g) U_{\rho^g}(x_1, x_2, \boldsymbol{\theta^g_1}, \boldsymbol{\theta^g_2}, \rho^g)\, dx_1 dx_2}{\int B(x_1, x_2) f^{\beta}(x_1, x_2, \boldsymbol{\theta^g_1}, \boldsymbol{\theta^g_2}, \rho^g) U_{\rho^g}(x_1, x_2, \boldsymbol{\theta^g_1}, \boldsymbol{\theta^g_2}, \rho^g)\, dx_1 dx_2},$$

where, detailed algebraic expressions for $A(x_1, x_2, y_1, y_2)$, $B(x_1, x_2)$ and the integrals in the aforesaid influence functions are provided in Section 4.8.6.

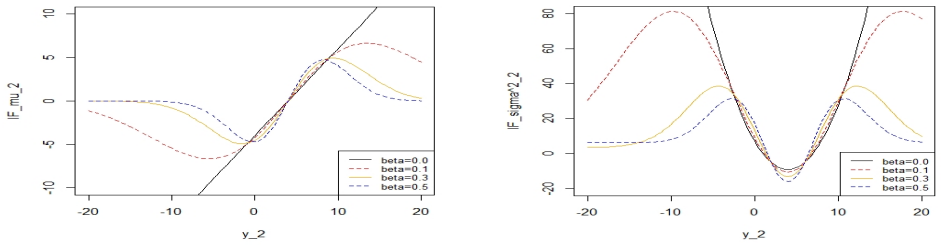
Let us observe that the functions $IF(\theta_{j1}, y_j, g_j)$ and $IF(\theta_{j2}, y_j, g_j)$ are linear in $(y_j - \theta^g_{j1}) f_j^{\beta}(y_j, \boldsymbol{\theta^g_j})$, $(y_j - \theta^g_{j1})^2 f_j^{\beta}(y_j, \boldsymbol{\theta^g_j})$ and $f_j^{\beta}(y_j, \boldsymbol{\theta^g_j})$ for $j = 1, 2$. Since each of these functions are bounded in y_j, the boundedness of $IF(\theta_{j1}, y_j, g_j)$ and $IF(\theta_{j2}, y_j, g_j)$ for $j = 1, 2$ are trivial. By observing

$$U_{\rho^g}(y_1, y_2, \boldsymbol{\theta^g_1}, \boldsymbol{\theta^g_2}, \rho^g) = \frac{\rho^g}{1 - (\rho^g)^2} - \frac{\rho^g \left[\frac{(y_1 - \theta^g_{11})^2}{\theta^g_{12}} + \frac{(y_2 - \theta^g_{21})^2}{\theta^g_{22}}\right] - (1 + (\rho^g)^2) \frac{(y_1 - \theta^g_{11})(y_2 - \theta^g_{21})}{\sqrt{\theta^g_{12}\theta^g_{22}}}}{(1 - (\rho^g)^2)^2},$$

it is easy to establish that $IF(\rho, y_1, y_2, g)$ is a linear function of $IF(\theta_{j1}, y_j, g_j)$, $IF(\theta_{j2}, y_j, g_j)$ for $j = 1, 2$ and $(y_j - \theta^g_{j1})^2 f^{\beta}(y_1, y_2, \boldsymbol{\theta^g_1}, \boldsymbol{\theta^g_2}, \rho^g)$, $f^{\beta}(y_1, y_2, \boldsymbol{\theta^g_1}, \boldsymbol{\theta^g_2}, \rho^g)$ for $j = 1, 2$ and $(y_1 - \theta^g_{11})(y_2 - \theta^g_{21}) f^{\beta}(y_1, y_2, \boldsymbol{\theta^g_1}, \boldsymbol{\theta^g_2}, \rho^g)$. The boundedness of $IF(\rho, y_1, y_2, g)$ is now easily followed by the boundedness of the aforesaid components of $IF(\rho, y_1, y_2, g)$. To see the behaviour explicitly, we present a special case pictorially where the model family is bivariate normal and the true distribution is also bivariate normal with component means $\theta^g_{11} = 1$, $\theta^g_{21} = 4$, component variances $\theta^g_{12} = 4$, $\theta^g_{22} = 9$ and correlation $\rho^g = 0.5$. The influence functions of the component means and variances are presented in Figure 4.1 and the influence functions (for different values of β including the maximum likelihood case) of the correlation are presented in Figure 4.2.



(a) $IF(\theta_{11}, y_1, g_1)$ (b) $IF(\theta_{12}, y_1, g_1)$



(c) $IF(\theta_{21}, y_2, g_2)$ (d) $IF(\theta_{22}, y_2, g_2)$

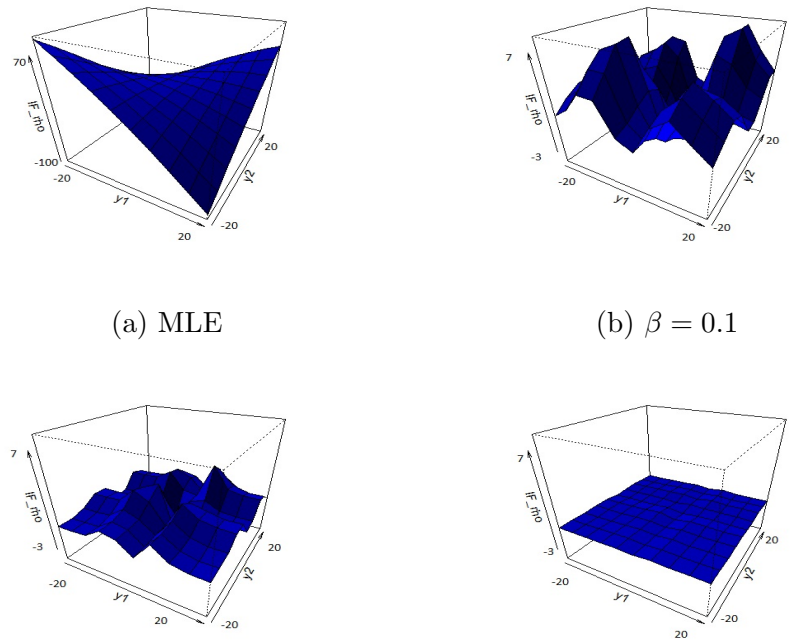
Figure 4.1: Influence functions of component means and variances.

4.6 Simulation Experiments

4.6.1 Experimental Set-up and Performance Measures

We now assess the performance of our method along with the other existing methods under the multivariate normal set-up. Here we assume that the true distribution is also multivariate normal, i.e., the true distribution belongs to the model family. To carry out our simulation experiments, we simulate 100 random samples of size $n = 1000$ from multivariate normal distributions of dimension p with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Different choices of p, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are taken to vary the simulation set-ups. Data from 2, 5, 10, 20 and 30 dimensions are simulated with mean vector $\mathbf{0}$ and two types of covariance structures are used, viz., diagonal and a special kind of non-diagonal structure. Identity matrices of appropriate dimensions are taken as diagonal covariance matrices. For the non-diagonal choice, we consider the following $p \times p$ matrix (with $p_1 = \left[\frac{p}{2}\right]$ and $p_2 = p - p_1$).

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{U}(0.7) & \mathbf{0}_{p_1 \times p_2} \\ \mathbf{0}_{p_2 \times p_1} & \boldsymbol{I}_{p_2} \end{bmatrix},$$



(a) MLE (b) $\beta = 0.1$

(c) $\beta = 0.3$ (d) $\beta = 0.5$

Figure 4.2: Influence functions of correlation coefficients.

where $\boldsymbol{U}(\rho) = [[u_{ij}]]_{i,j=1}^{p_1}$ with $u_{ij} = \rho^{|i-j|}$ for $1 \leq i, j \leq p_1$ and $-1 \leq \rho \leq 1$. This structure is known as block-banded covariance matrix. But for $p = 2$ (only), we take $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 1 \end{bmatrix}$.

We are going to evaluate two measures of accuracy, viz., L_2 bias and mean squared error (MSE) for the mean and covariance matrix estimators, separately. Suppose, $\widehat{\boldsymbol{\mu}}_i$ and $\widehat{\boldsymbol{\Sigma}}_i$ be the estimators of the mean vector and the covariance matrix from the i-th replication, $1 \leq i \leq 100$. We calculate the L_2 bias and mean squared errors as follows:

$$\text{Bias of mean estimators} = \left|\left|\frac{1}{100}\sum_{i=1}^{100}\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}\right|\right|_2, \quad \text{MSE of mean estimators} = \frac{1}{100}\sum_{i=1}^{100}\left|\left|\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}\right|\right|_2^2,$$

$$\text{Bias of covariance matrix estimators} = \left|\left|\frac{1}{100}\sum_{i=1}^{100}\widehat{\boldsymbol{\Sigma}}_i - \boldsymbol{\Sigma}\right|\right|_F,$$

$$\text{MSE of covariance matrix estimators} = \frac{1}{100}\sum_{i=1}^{100}\left|\left|\widehat{\boldsymbol{\Sigma}}_i - \boldsymbol{\Sigma}\right|\right|_F^2,$$

Dimension (p)	Different Methods	Location Vector Bias	Location Vector MSE	Scatter Matrix Bias	Scatter Matrix MSE
2	MLE	0.005	0.002	0.006	0.007
	SMDPDE ($\beta = 0.1$)	0.004	0.002	0.007	0.007
	SMDPDE ($\beta = 0.3$)	0.004	0.002	0.007	0.008
	SMDPDE ($\beta = 0.5$)	0.004	0.002	0.006	0.009
	MCD	0.003	0.002	0.019	0.013
	MVE	0.004	0.002	0.022	0.016
	GK	0.003	0.002	0.355	0.133
	MM	0.006	0.002	0.006	0.008
	S	0.004	0.003	0.016	0.015
5	MLE	0.009	0.005	0.013	0.029
	SMDPDE ($\beta = 0.1$)	0.009	0.005	0.014	0.029
	SMDPDE ($\beta = 0.3$)	0.009	0.005	0.017	0.034
	SMDPDE ($\beta = 0.5$)	0.008	0.006	0.020	0.041
	MCD	0.005	0.005	0.022	0.041
	MVE	0.007	0.006	0.040	0.052
	GK	0.005	0.006	0.341	0.149
	MM	0.005	0.005	0.012	0.032
	S	0.005	0.006	0.013	0.039
10	MLE	0.013	0.010	0.028	0.110
	SMDPDE ($\beta = 0.1$)	0.013	0.010	0.028	0.119
	SMDPDE ($\beta = 0.3$)	0.013	0.011	0.029	0.130
	SMDPDE ($\beta = 0.5$)	0.013	0.012	0.031	0.156
	MCD	0.008	0.010	0.038	0.131
	MVE	0.007	0.011	0.049	0.158
	GK	0.009	0.012	0.323	0.223
	MM	0.011	0.010	0.032	0.118
	S	0.011	0.010	0.033	0.123
20	MLE	0.011	0.020	0.062	0.425
	SMDPDE ($\beta = 0.1$)	0.010	0.020	0.063	0.435
	SMDPDE ($\beta = 0.3$)	0.010	0.022	0.067	0.497
	SMDPDE ($\beta = 0.5$)	0.010	0.022	0.073	0.598
	MCD	0.013	0.020	0.082	0.489
	MVE	0.015	0.022	0.085	0.520
	GK	0.015	0.023	0.317	0.555
	MM	0.014	0.020	0.066	0.441
	S	0.014	0.019	0.066	0.434
30	MLE	0.027	0.030	0.177	0.932
	SMDPDE ($\beta = 0.1$)	0.027	0.030	0.183	0.955
	SMDPDE ($\beta = 0.3$)	0.031	0.036	0.204	1.074
	SMDPDE ($\beta = 0.5$)	0.031	0.035	0.225	1.328
	MCD	0.039	0.031	0.217	1.094
	MVE	0.040	0.033	0.222	1.141
	GK	0.041	0.034	0.382	1.146
	MM	0.035	0.032	0.203	0.991
	S	0.035	0.032	0.200	0.958

Table 4.3: Estimated bias and mean squared errors in case of diagonal covariance structures under pure data.

where $||\cdot||_2$ is the L_2-norm and $||\cdot||_F$ is the Frobenius norm of a matrix.

To study the robustness and efficiency of our method, we consider both pure as well as contaminated datasets in each of the aforesaid set-ups (depending on data dimension and covariance structure). As we have discussed earlier, the pure datasets are

Dimension (p)	Different Methods	Location Vector Bias	Location Vector MSE	Scatter Matrix Bias	Scatter Matrix MSE
2	MLE	2.806	7.939	71.435	5136.492
	SMDPDE ($\beta = 0.1$)	0.001	0.003	0.012	0.007
	SMDPDE ($\beta = 0.3$)	0.002	0.003	0.054	0.011
	SMDPDE ($\beta = 0.5$)	0.002	0.003	0.102	0.021
	MCD	0.006	0.003	0.408	0.186
	MVE	0.006	0.003	0.408	0.188
	GK	0.014	0.003	0.297	0.096
	MM	0.004	0.003	0.310	0.109
	S	0.004	0.004	0.313	0.119
5	MLE	4.497	20.410	180.956	32986.971
	SMDPDE ($\beta = 0.1$)	0.010	0.006	0.031	0.034
	SMDPDE ($\beta = 0.3$)	0.010	0.006	0.098	0.049
	SMDPDE ($\beta = 0.5$)	0.010	0.007	0.175	0.081
	MCD	0.008	0.006	0.345	0.176
	MVE	0.007	0.007	0.349	0.191
	GK	0.010	0.007	0.262	0.108
	MM	0.005	0.006	0.406	0.214
	S	0.007	0.006	0.408	0.222
10	MLE	6.281	39.757	357.579	128617.184
	SMDPDE ($\beta = 0.1$)	0.007	0.011	0.057	0.130
	SMDPDE ($\beta = 0.3$)	0.007	0.012	0.144	0.175
	SMDPDE ($\beta = 0.5$)	0.006	0.013	0.250	0.259
	MCD	0.009	0.012	0.319	0.273
	MVE	0.008	0.013	0.324	0.296
	GK	0.014	0.013	0.260	0.201
	MM	0.008	0.012	0.549	0.477
	S	0.009	0.012	0.550	0.482
20	MLE	8.971	81.329	721.061	524251.729
	SMDPDE ($\beta = 0.1$)	0.015	0.024	0.090	0.500
	SMDPDE ($\beta = 0.3$)	0.016	0.026	0.210	0.640
	SMDPDE ($\beta = 0.5$)	0.017	0.028	0.362	0.889
	MCD	0.0131	0.026	0.330	0.709
	MVE	0.014	0.026	0.321	0.721
	GK	0.017	0.028	0.248	0.570
	MM	0.015	0.023	0.732	1.198
	S	0.015	0.023	0.731	1.189
30	MLE	11.077	123.377	1089.409	1191872.749
	SMDPDE ($\beta = 0.1$)	0.040	0.034	0.225	1.101
	SMDPDE ($\beta = 0.3$)	0.043	0.035	0.328	1.395
	SMDPDE ($\beta = 0.5$)	0.046	0.038	0.494	1.896
	MCD	0.033	0.037	0.379	1.392
	MVE	0.030	0.037	0.377	1.402
	GK	0.034	0.039	0.327	1.196
	MM	0.033	0.036	0.922	2.234
	S	0.033	0.036	0.919	2.202

Table 4.4: Estimated bias and mean squared errors in case of diagonal covariance structures under contaminated data.

simulated from multivariate normal distribution and the contaminated datasets are simulated by generating observations from a mixture normal distribution. The main component of this mixture distribution is multivariate normal with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}$ and the contaminating component is a multivariate normal with mean

Dimension (p)	Different Methods	Location Vector Bias	Location Vector MSE	Scatter Matrix Bias	Scatter Matrix MSE
2	MLE	0.004	0.002	0.006	0.006
	SMDPDE ($\beta = 0.1$)	0.004	0.002	0.006	0.007
	SMDPDE ($\beta = 0.3$)	0.004	0.002	0.005	0.008
	SMDPDE ($\beta = 0.5$)	0.004	0.003	0.004	0.009
	MCD	0.004	0.002	0.02	0.012
	MVE	0.001	0.002	0.008	0.013
	GK	0.005	0.002	0.434	0.195
	MM	0.004	0.003	0.006	0.009
	S	0.006	0.004	0.008	0.017
5	MLE	0.004	0.004	0.029	0.032
	SMDPDE ($\beta = 0.1$)	0.004	0.005	0.029	0.034
	SMDPDE ($\beta = 0.3$)	0.004	0.005	0.03	0.039
	SMDPDE ($\beta = 0.5$)	0.004	0.006	0.03	0.046
	MCD	0.013	0.005	0.019	0.043
	MVE	0.013	0.006	0.042	0.06
	GK	0.014	0.006	0.364	0.166
	MM	0.007	0.005	0.018	0.029
	S	0.007	0.006	0.021	0.035
10	MLE	0.007	0.01	0.036	0.115
	SMDPDE ($\beta = 0.1$)	0.006	0.01	0.037	0.118
	SMDPDE ($\beta = 0.3$)	0.006	0.01	0.04	0.136
	SMDPDE ($\beta = 0.5$)	0.006	0.011	0.044	0.163
	MCD	0.006	0.011	0.037	0.133
	MVE	0.006	0.012	0.05	0.156
	GK	0.006	0.012	0.428	0.305
	MM	0.011	0.009	0.027	0.118
	S	0.011	0.009	0.026	0.122
20	MLE	0.014	0.02	0.066	0.443
	SMDPDE ($\beta = 0.1$)	0.014	0.02	0.064	0.455
	SMDPDE ($\beta = 0.3$)	0.015	0.021	0.066	0.525
	SMDPDE ($\beta = 0.5$)	0.016	0.023	0.072	0.631
	MCD	0.018	0.021	0.073	0.498
	MVE	0.019	0.022	0.081	0.527
	GK	0.018	0.022	0.45	0.672
	MM	0.019	0.021	0.070	0.455
	S	0.019	0.020	0.070	0.447
30	MLE	0.032	0.032	0.191	0.973
	SMDPDE ($\beta = 0.1$)	0.032	0.032	0.195	0.999
	SMDPDE ($\beta = 0.3$)	0.034	0.033	0.209	1.142
	SMDPDE ($\beta = 0.5$)	0.036	0.037	0.227	1.365
	MCD	0.036	0.032	0.203	1.114
	MVE	0.037	0.034	0.211	1.136
	GK	0.037	0.035	0.494	1.269
	MM	0.039	0.032	0.213	1.071
	S	0.038	0.032	0.209	1.037

Table 4.5: Estimated bias and mean squared errors in case of non-diagonal covariance structures under pure data.

vector $(20, 20, \ldots, 20)$ and identity covariance matrix. The mixing proportions are 0.9 and 0.1, so that, the data represent 10% contamination of the pure model. In each case, we are considering three values of the tuning parameter β, namely, 0.1, 0.3 and 0.5. From efficiency considerations, we are not going to take higher values of β. We

Dimension (p)	Different Methods	Location Vector Bias	Location Vector MSE	Scatter Matrix Bias	Scatter Matrix MSE
2	MLE	2.829	8.071	71.912	5205.541
	SMDPDE ($\beta = 0.1$)	0.005	0.002	0.009	0.007
	SMDPDE ($\beta = 0.3$)	0.005	0.002	0.063	0.012
	SMDPDE ($\beta = 0.5$)	0.006	0.003	0.125	0.025
	MCD	0.002	0.003	0.505	0.272
	MVE	0.003	0.003	0.507	0.279
	GK	0.011	0.003	0.29	0.092
	MM	0.006	0.002	0.392	0.167
	S	0.009	0.004	0.392	0.173
5	MLE	4.434	19.877	178.341	32072.472
	SMDPDE ($\beta = 0.1$)	0.009	0.005	0.034	0.037
	SMDPDE ($\beta = 0.3$)	0.01	0.006	0.107	0.055
	SMDPDE ($\beta = 0.5$)	0.011	0.006	0.191	0.091
	MCD	0.002	0.006	0.379	0.201
	MVE	0.003	0.006	0.388	0.22
	GK	0.011	0.006	0.285	0.124
	MM	0.008	0.005	0.452	0.256
	S	0.008	0.006	0.454	0.266
10	MLE	6.265	39.664	357.366	128759.552
	SMDPDE ($\beta = 0.1$)	0.025	0.012	0.043	0.14
	SMDPDE ($\beta = 0.3$)	0.025	0.013	0.157	0.191
	SMDPDE ($\beta = 0.5$)	0.024	0.014	0.298	0.3
	MCD	0.012	0.012	0.405	0.337
	MVE	0.01	0.012	0.411	0.352
	GK	0.009	0.013	0.311	0.24
	MM	0.009	0.011	0.706	0.697
	S	0.009	0.011	0.708	0.705
20	MLE	8.858	79.225	712.878	512080.801
	SMDPDE ($\beta = 0.1$)	0.012	0.022	0.08	0.529
	SMDPDE ($\beta = 0.3$)	0.013	0.024	0.236	0.683
	SMDPDE ($\beta = 0.5$)	0.014	0.026	0.442	0.986
	MCD	0.013	0.024	0.395	0.76
	MVE	0.014	0.024	0.39	0.775
	GK	0.013	0.025	0.349	0.641
	MM	0.016	0.024	0.952	1.579
	S	0.016	0.024	0.951	1.571
30	MLE	11.061	123.419	1088.910	1193694.140
	SMDPDE ($\beta = 0.1$)	0.031	0.035	0.231	1.116
	SMDPDE ($\beta = 0.3$)	0.033	0.037	0.393	1.450
	SMDPDE ($\beta = 0.5$)	0.034	0.039	0.645	2.073
	MCD	0.042	0.037	0.441	1.487
	MVE	0.045	0.037	0.436	1.487
	GK	0.045	0.040	0.419	1.331
	MM	0.034	0.035	1.207	2.952
	S	0.034	0.035	1.202	2.907

Table 4.6: Estimated bias and mean squared errors in case of non-diagonal covariance structures under contaminated data.

present the output of the simulation experiments in Tables 4.3, 4.4, 4.5 and 4.6 which involve our method along with the ordinary maximum likelihood estimators (MLE, corresponds to $\beta = 0$ case), MCD, MVE, orthogonalized Gnanadesikan-Kettenring (GK) (see Maronna and Zamar (2002), [112]), MM estimators of location and scale

(MM) (originally proposed by Yohai (1987) [158] in regression set-up and later Tatsuoka and Tyler (2000) [143] developed the multivariate location-scale version) and S-estimators of location and scale (Rousseeuw and Yohai (1984) [134], Ruppert (1992) [135]). Appropriate R-packages [145, 81, 102] are used for computing these estimators.

4.6.2 Discussion of Simulation Results

The simulations that we have performed are quite extensive, and it is necessary to clearly pinpoint what the salient features of these numbers are. In the following we describe these features.

1. It can be trivially observed that, for contaminated datsets, the SMDPDEs (as well as its robust competitors) are much better than the MLEs both in terms of bias and MSEs. But the MLEs are a little better for almost all the robust methods in case of pure datasets. In particular, the bias and MSEs of the SMDPDEs are slightly more than those of the MLEs in case of lower β values for pure datasets and this difference is greater in magnitude for higher β values as well as for most of the other robust competitors of our estimator.

2. For pure datasets and diagonal covariance structures, the SMDPDEs are the best in terms of MSEs (for both location and scale estimation) among the robust methods most of the time; however, they have marginally higher bias (especially, in case of location vectors) in higher dimensions. But for the corresponding non-diagonal cases, the lack of unbiasedness of the SMDPDEs disappears with a little loss in MSEs. In case of pure datasets, both the MM and the S estimators maintain very close competitions with the SMDPDEs.

3. However, for contaminated datasets, the SMDPDEs are undoubtedly the best among all the robust alternatives, especially with lower values of β (specifically 0.1). The pairwise GK method improved a lot in case of contaminated datasets for covariance estimation and it is better than the MCD, MVE, MM and S estimators both in terms of bias and MSEs.

4. None of the SMDPDEs of the covariance matrices have been found to be non-positive definite throughout our entire simulation exercise.

Data Type	Covariance Structure	Dimension (p)	β	SMDPDE Bias	SMDPDE MSE	MDPDE Bias	MDPDE MSE
Pure	Diagonal	2	0.1	0.004	0.002	0.005	0.002
			0.3	0.004	0.002	0.004	0.002
			0.5	0.004	0.002	0.004	0.002
		5	0.1	0.009	0.005	0.009	0.005
			0.3	0.009	0.005	0.008	0.006
			0.5	0.008	0.006	0.007	0.008
		10	0.1	0.013	0.01	0.012	0.01
			0.3	0.013	0.011	0.011	0.014
			0.5	0.013	0.012	0.011	0.02
	Non-Diagonal	2	0.1	0.004	0.002	0.004	0.002
			0.3	0.004	0.002	0.004	0.002
			0.5	0.004	0.003	0.004	0.003
		5	0.1	0.004	0.005	0.005	0.005
			0.3	0.004	0.005	0.006	0.006
			0.5	0.004	0.006	0.007	0.008
		10	0.1	0.006	0.01	0.007	0.01
			0.3	0.006	0.01	0.009	0.013
			0.5	0.006	0.011	0.011	0.02
Contaminated	Diagonal	2	0.1	0.001	0.003	0.001	0.003
			0.3	0.002	0.003	0.001	0.003
			0.5	0.002	0.003	0.001	0.003
		5	0.1	0.01	0.006	0.009	0.006
			0.3	0.01	0.006	0.01	0.007
			0.5	0.01	0.007	0.01	0.008
		10	0.1	0.007	0.011	0.008	0.012
			0.3	0.007	0.012	0.009	0.014
			0.5	0.006	0.013	0.012	0.02
	Non-Diagonal	2	0.1	0.005	0.002	0.005	0.002
			0.3	0.005	0.002	0.006	0.002
			0.5	0.006	0.003	0.006	0.003
		5	0.1	0.009	0.005	0.011	0.005
			0.3	0.01	0.006	0.013	0.007
			0.5	0.011	0.006	0.015	0.008
		10	0.1	0.025	0.012	0.026	0.013
			0.3	0.025	0.013	0.026	0.017
			0.5	0.024	0.014	0.027	0.024

Table 4.7: Estimated bias and mean squared errors of mean estimators for the sequential and ordinary minimum DPD methods.

4.6.3 Comparison of SMDPDE with Usual MDPDE

A comparative study on bias and mean squared errors of the mean, variance and correlation estimators based on MDPDEs and SMDPDEs are presented separately in Tables 4.7, 4.8 and 4.9, respectively, in case of lower data dimensions. Let us discuss

Data Type	Covariance Structure	Dimension (p)	β	SMDPDE Bias	SMDPDE MSE	MDPDE Bias	MDPDE MSE
Pure	Diagonal	2	0.1	0.006	0.004	0.006	0.004
			0.3	0.007	0.005	0.006	0.005
			0.5	0.006	0.006	0.005	0.006
		5	0.1	0.006	0.01	0.007	0.01
			0.3	0.007	0.011	0.007	0.013
			0.5	0.008	0.013	0.009	0.017
		10	0.1	0.008	0.021	0.009	0.022
			0.3	0.009	0.024	0.01	0.03
			0.5	0.01	0.028	0.013	0.046
	Non-Diagonal	2	0.1	0.005	0.004	0.005	0.004
			0.3	0.005	0.005	0.005	0.005
			0.5	0.004	0.006	0.005	0.006
		5	0.1	0.014	0.011	0.015	0.011
			0.3	0.016	0.012	0.018	0.014
			0.5	0.017	0.015	0.022	0.018
		10	0.1	0.019	0.021	0.018	0.021
			0.3	0.02	0.024	0.013	0.028
			0.5	0.02	0.028	0.014	0.044
Contaminated	Diagonal	2	0.1	0.012	0.005	0.012	0.005
			0.3	0.054	0.008	0.051	0.008
			0.5	0.102	0.017	0.09	0.015
		5	0.1	0.028	0.011	0.027	0.011
			0.3	0.097	0.022	0.077	0.019
			0.5	0.174	0.046	0.11	0.03
		10	0.1	0.046	0.024	0.045	0.025
			0.3	0.139	0.046	0.093	0.043
			0.5	0.247	0.094	0.107	0.067
	Non-Diagonal	2	0.1	0.008	0.004	0.008	0.004
			0.3	0.051	0.008	0.048	0.007
			0.5	0.1	0.016	0.088	0.014
		5	0.1	0.031	0.013	0.03	0.013
			0.3	0.098	0.024	0.08	0.023
			0.5	0.174	0.048	0.113	0.035
		10	0.1	0.03	0.026	0.029	0.026
			0.3	0.126	0.046	0.081	0.041
			0.5	0.234	0.092	0.097	0.061

Table 4.8: Estimated bias and mean squared errors of variance estimators for the sequential and ordinary minimum DPD methods.

the implications of these comparative studies in the following points.

1. For the mean estimators (Table 4.7), it can be observed that both MDPDEs and SMDPDEs have almost similar bias and MSEs in lower dimension ($p = 2$). But as p increases to 5 or 10, the SMDPDEs tend to have less bias and MSE, especially

Data Type	Covariance Structure	Dimension (p)	β	SMDPDE Bias	SMDPDE MSE	MDPDE Bias	MDPDE MSE
Pure	Diagonal	2	0.1	0.002	0.001	0.002	0.001
			0.3	0.001	0.001	0.001	0.001
			0.5	0.001	0.002	0.001	0.002
		5	0.1	0.009	0.01	0.009	0.01
			0.3	0.011	0.011	0.013	0.012
			0.5	0.013	0.014	0.017	0.017
		10	0.1	0.019	0.046	0.019	0.047
			0.3	0.02	0.053	0.022	0.066
			0.5	0.021	0.064	0.026	0.105
	Non-Diagonal	2	0.1	0.001	0.001	0.001	0.001
			0.3	0.001	0.001	0.001	0.001
			0.5	0.001	0.001	0.001	0.001
		5	0.1	0.016	0.009	0.016	0.009
			0.3	0.015	0.011	0.015	0.012
			0.5	0.015	0.013	0.016	0.016
		10	0.1	0.019	0.041	0.018	0.042
			0.3	0.021	0.048	0.023	0.059
			0.5	0.025	0.058	0.034	0.096
Contaminated	Diagonal	2	0.1	0.003	0.001	0.003	0.001
			0.3	0.004	0.002	0.004	0.002
			0.5	0.005	0.002	0.005	0.002
		5	0.1	0.009	0.011	0.009	0.011
			0.3	0.01	0.013	0.011	0.014
			0.5	0.011	0.015	0.014	0.019
		10	0.1	0.023	0.052	0.024	0.053
			0.3	0.025	0.059	0.029	0.073
			0.5	0.028	0.071	0.035	0.115
	Non-Diagonal	2	0.1	0.001	0.001	0.001	0.001
			0.3	0.001	0.001	0.001	0.001
			0.5	0.001	0.001	0.001	0.001
		5	0.1	0.006	0.01	0.005	0.01
			0.3	0.006	0.012	0.004	0.013
			0.5	0.006	0.014	0.004	0.017
		10	0.1	0.019	0.045	0.02	0.047
			0.3	0.022	0.052	0.026	0.066
			0.5	0.026	0.063	0.033	0.103

Table 4.9: Estimated bias and mean squared errors of correlation estimators for the sequential and ordinary minimum DPD methods.

under higher values of β (0.5, in particular) for pure as well as contaminated datasets.

2. In case of the variance estimators (Table 4.8), the bias and MSEs of both SMDPDEs and MDPDEs are similar in lower dimension but as dimension increases

Data Type	Covariance Structure	Dimension (p)	β	SMDPDE Bias	SMDPDE MSE	MDPDE Bias	MDPDE MSE
Pure	Diagonal	2	0.1	0.007	0.007	0.007	0.007
			0.3	0.007	0.008	0.006	0.008
			0.5	0.006	0.009	0.005	0.01
		5	0.1	0.014	0.029	0.015	0.03
			0.3	0.017	0.034	0.02	0.037
			0.5	0.02	0.041	0.025	0.05
		10	0.1	0.028	0.113	0.029	0.117
			0.3	0.029	0.13	0.033	0.161
			0.5	0.031	0.156	0.039	0.257
	Non-Diagonal	2	0.1	0.006	0.007	0.006	0.007
			0.3	0.005	0.008	0.006	0.008
			0.5	0.004	0.009	0.005	0.01
		5	0.1	0.029	0.034	0.03	0.034
			0.3	0.03	0.039	0.032	0.042
			0.5	0.03	0.046	0.035	0.054
		10	0.1	0.037	0.118	0.035	0.121
			0.3	0.04	0.136	0.037	0.168
			0.5	0.044	0.163	0.05	0.27
Contaminated	Diagonal	2	0.1	0.012	0.007	0.012	0.007
			0.3	0.054	0.011	0.052	0.011
			0.5	0.102	0.021	0.09	0.019
		5	0.1	0.031	0.034	0.03	0.034
			0.3	0.098	0.049	0.079	0.049
			0.5	0.175	0.081	0.111	0.071
		10	0.1	0.057	0.13	0.057	0.134
			0.3	0.144	0.175	0.102	0.197
			0.5	0.25	0.259	0.119	0.313
	Non-Diagonal	2	0.1	0.009	0.007	0.009	0.007
			0.3	0.063	0.012	0.059	0.011
			0.5	0.125	0.025	0.108	0.022
		5	0.1	0.034	0.037	0.033	0.038
			0.3	0.107	0.055	0.087	0.056
			0.5	0.191	0.091	0.124	0.08
		10	0.1	0.043	0.14	0.043	0.143
			0.3	0.157	0.191	0.105	0.207
			0.5	0.298	0.3	0.127	0.32

Table 4.10: Estimated bias and mean squared errors of covariance matrix estimators for the sequential and ordinary minimum DPD methods.

the SMDPDEs tend to have more bias for pure datasets and more bias and MSEs for contaminated datasets with higher β values (0.3 and 0.5, in particular).

3. In case of correlation estimators (Table 4.9), both the SMDPDEs and MDPDEs are comparable in terms of bias and MSEs under lower dimension but as the

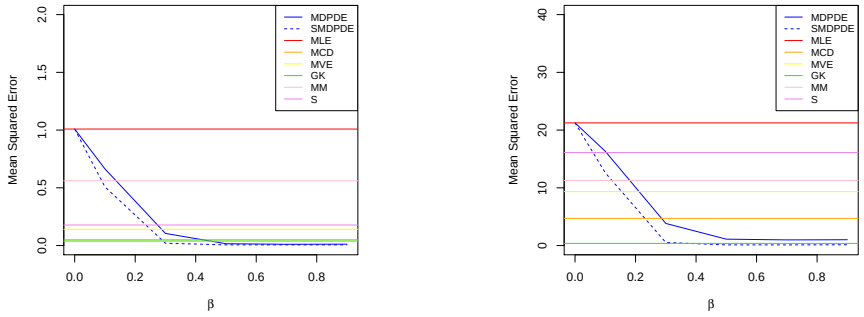
dimension grows up, the SMDPDEs become more accurate (in terms of bias and MSEs) under higher values of β under both pure and contaminated datasets.

4. Now, for the entire covariance matrix estimators (Table 4.10), both SMDPDEs and MDPDEs perform quite similarly for smaller values of β. As β increases to 0.3 or 0.5, the MDPDEs become slightly less biased, although their MSEs still remain comparable.

4.6.4 Cellwise Contamination

So far, we have only considered casewise (or rowwise) contaminations in the aforesaid simulation experiments, where some of the sample observations (rows of the data matrix) are perturbed. However, cellwise (columnwise) contamination does arise in real life datasets where some of the components (columns of the data matrix) are perturbed. This results in the contamination of a comparatively larger proportion of the sample observations, at least partially. Discarding or downweighting all those partially contaminated observations may result in a significant loss of information which affects the efficiency of the resulting estimators. Since the proposed methodology in this work is sequential in nature, it is intuitively expected that this procedure can tackle the cellwise contaminated datasets more efficiently as compared to the ordinary minimum DPD method. We simulate 100 datasets (sample size 1000, dimension 4) where the first 600 observations are simulated from a standard 4-dimensional normal distribution. The remaining 400 observations are divided into 4 subsets S_1, S_2, S_3 and S_4. The subset S_i is generated from a 4-dimensional normal distribution with mean vector $5\boldsymbol{e}_i$ and identity covariance, where $\boldsymbol{e}_i$ is the i-th canonical vector in $\mathbb{R}^4$, $1 \leq i \leq 4$. That is, the true value of the mean vector and covariance matrix are $(0,0,0,0)^{'}$ and $\boldsymbol{I}_4$, respectively and the last 400 observations are contaminated cellwise. We obtain the estimates of the mean vector and the covariance matrix based on these 100 samples using the ordinary and sequential minimum DPD methods and their competitors (considered in the simulation experiments) and present the mean squared errors (separately for the mean vector and the covariance matrix) for different values of β in Figure 4.3. The superiority of the sequential method can clearly be observed over all its competitors with the ordinary minimum DPD and GK methods being the closest ones. It should also be noted that the sequential method achieved lower mean squared errors for smaller values of β as compared to the ordinary minimum DPD method.

It is one of our future plans to investigate the performances of both ordinary and



(a) Mean vector (b) Covariance matrix

Figure 4.3: Mean squared errors of the estimates of mean vector and covariance matrix for different values of β.

sequential minimum DPD methods in case of cellwise contamination in a detailed manner.

4.6.5 Scalability of the Proposed SMDPDE

We have already realized the fact that the sequential minimum DPD estimation is computationally far more efficient than the ordinary minimum DPD estimation in the

Methods		p						
	β	2	5	10	20	30	40	50
Number of Parameters		5	20	65	230	495	860	1325
SMDPDE	0.1	100 %	100 %	100%	100%	100%	100%	100%
	0.3	100 %	100 %	100%	100%	100%	100%	100%
	0.5	100 %	100 %	100%	100%	100%	100%	100%
	0.7	100 %	100 %	100%	100%	100%	100%	100%
MDPDE	0.1	100 %	100 %	100%	100%	100%	92%	83%
	0.3	100 %	100 %	100%	94%	81%	32%	5%
	0.5	100 %	100 %	78%	46%	0%	0%	0%
	0.7	100 %	100 %	28%	0%	0%	0%	0%

Table 4.11: Empirical convergence rates of the indicated methods for a sample size of $n = 2000$.

sense that the existence of the SMDPDE is computationally guaranteed for almost all possible combinations of n, p and β unlike the ordinary MDPDE. Especially, in higher dimensions, the algorithm used to obtain the ordinary MDPDE may fail to converge. The empirical convergence rates of the ordinary and sequential minimum

DPD methods are shown in Table 4.11 where this fact is clearly borne out. In fact for large data dimensions and large values of β, the minimum DPD method practically never leads to convergence. Convergence rates of the other competitors are found to be perfect like the sequential minimum DPD method.

4.7 Credit Card Transactions Data

We now apply the sequential minimum DPD method on a real data containing information about credit card transactions by European cardholders on two particular days of September, 2013; some of these transactions were fraudulent. The dataset[1] consists of 28 features (first 28 principal components of the original dataset which is confidential; the original dataset has been collected and analysed during a research collaboration of Worldline and the Machine Learning Group[2] of ULB (Université Libre de Bruxelles) on big data mining and fraud detection) along with the elapsed times (from the first transaction), transaction amount and original transaction labels (i.e., genuine or fraudulent). This dataset has been analyzed in recent years using various machine learning tools [33, 34, 32]. For time and space complexity, we analyze a subset of this dataset with the first 10000 transactions of which only 38 are fraudulent and the remaining 9962 are genuine. We consider these fraudulent transactions as outlying observations present in the dataset (as observed in Figure 4.4). But the proportion of outliers is only 0.0038 which implies the dataset can almost be regarded as a noise-free one and no robust method is expected to show its supremacy through drastic differences in comparison with the traditional likelihood based procedure if applied on this dataset. To make the contamination stronger, we perturb another 362 observations which represent genuine transactions. This is done following two approaches. Let us note that, we only have to choose a sub-sample of size 362 of genuine observations (transactions) out of a total of 9962 genuine observations (transactions). Either we can choose this sub-sample randomly (without replacement) or we may choose those 362 genuine observations which are mostly concentrated around their central tendency. That is, if $\tilde{\boldsymbol{\mu}}$ is the sample componentwise median of the 9962 genuine observations, then we may choose those 362 genuine observations whose distances from $\tilde{\boldsymbol{\mu}}$ are the least.

We now apply our method along with the likelihood estimation procedure to fit the

[1]Source: https://www.kaggle.com/mlg-ulb/creditcardfraud/version/3

[2]http://mlg.ulb.ac.be

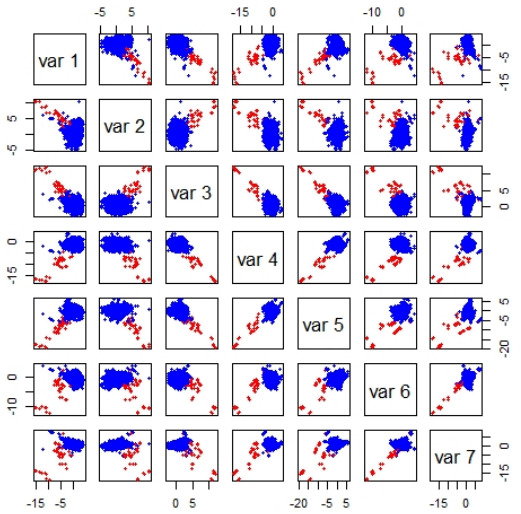


Figure 4.4: Pairwise scatter plots of some of the components with blue points as genuine observations and red as fraudulent ones.

Resampling Type	Method	Difference between estimated mean vectors	Difference between estimated covariance matrices
Most Concentrated	MLE	2.007	50.584
	SMDPDE ($\beta = 0.1$)	0.268	1.506
	SMDPDE ($\beta = 0.3$)	0.109	0.378
	SMDPDE ($\beta = 0.5$)	0.353	1.678
Random	MLE	2.013	50.857
	SMDPDE ($\beta = 0.1$)	0.740	10.616
	SMDPDE ($\beta = 0.3$)	0.280	2.213
	SMDPDE ($\beta = 0.5$)	0.211	1.516

Table 4.12: L_2 differences between estimated mean vectors and covariance matrices based on the genuine sub-samples (size 362) and the contaminated sub-samples (size 400).

aforesaid sub-samples using multivariate normal distribution (observing (Figure 4.4) an approximate elliptic nature of the overall dataset). To understand the robustness of our method, we first consider the noise-free sub-samples of the data with 362 genuine observations and then the contaminated sub-samples of size 400 where 38 fraudulent transactions were added to the previous noise-free sub-samples of size 362. We observe the differences (L_2 distances) between the estimated mean vectors (and the estimated

covariance matrices) based on the noise-free sub-samples and the contaminated sub-samples. To model either the noise-free sub-samples or the contaminated sub-samples, we need 28-dimensional normal distributions (comprising 434 parameters) but we only have samples of size 362 or 400. Thus, none of the robust alternatives to our method (i.e., the ordinary minimum DPD method, MCD, MVE etc.) used in the simulation experiments can be successful in producing the estimates of mean vector and covariance matrix for possible singularity. The L_2 differences between the estimated mean vectors (and the estimated covariance matrices) based on the noise-free sub-samples and the contaminated sub-samples by the likelihood method and the sequential minimum DPD method are presented in Table 4.12. These differences for our method are much less than those in case of the likelihood method which establish the superiority of our method in terms of robustness. Also, this application have shown the applicability of our method in case of such a higher dimensional dataset where the data dimension is 28, so that, we need to estimate 434 unknown parameters which is indeed greater than the sample size 400. For this particular data example, Higham's algorithm (followed by an eigenvalue truncation step, as discussed in Remark 4.2) is utilized to find the nearest positive definite correlation matrices (of dimension 28×28) to the estimated correlation matrices by our method.

4.8 Appendices

4.8.1 Proof of Theorem 4.1

Proof. We follow the same approach which was taken by Basu et al. (2011) [12]. Our plan is to show that for all sufficiently small ϵ, $H_n(\widehat{\boldsymbol{\theta}}_{1n}, \widehat{\boldsymbol{\theta}}_{2n}, \rho^g) < H_n(\widehat{\boldsymbol{\theta}}_{1n}, \widehat{\boldsymbol{\theta}}_{2n}, \rho)$ for all points $\rho \in Surface(Q_\epsilon)$ with probability tending to 1 where Q_ϵ is a sphere of radius $\epsilon > 0$ with center at ρ^g. Hence, $H_n(\widehat{\boldsymbol{\theta}}_{1n}, \widehat{\boldsymbol{\theta}}_{2n}, \rho)$ has a local minima in the interior of Q_ϵ. Now, let us observe the fact that at a local minimum, the Equation (4.9) must be satisfied. Thus for any sufficiently small $\epsilon > 0$, Equation (4.9) has a solution $\widehat{\rho}_{n\epsilon}$ depending on ϵ with probability tending to 1 as $n \to \infty$. To execute our plan, let us consider the following Taylor series expansion of $H_n(\widehat{\boldsymbol{\theta}}_{1n}, \widehat{\boldsymbol{\theta}}_{2n}, \rho)$ around $\rho = \rho^g$:

$$H_n(\widehat{\boldsymbol{\theta}}_{1n}, \widehat{\boldsymbol{\theta}}_{2n}, \rho) = H_n(\widehat{\boldsymbol{\theta}}_{1n}, \widehat{\boldsymbol{\theta}}_{2n}, \rho^g) + (\rho - \rho^g) \frac{\partial H_n(\widehat{\boldsymbol{\theta}}_{1n}, \widehat{\boldsymbol{\theta}}_{2n}, \rho)}{\partial \rho}\Big|_{\rho=\rho^g} + \frac{(\rho-\rho^g)^2}{2!} \frac{\partial^2 H_n(\widehat{\boldsymbol{\theta}}_{1n}, \widehat{\boldsymbol{\theta}}_{2n}, \rho)}{\partial \rho^2}\Big|_{\rho=\rho^g}$$
$$+ \frac{(\rho-\rho^g)^3}{3!} \frac{\partial^3 H_n(\widehat{\boldsymbol{\theta}}_{1n}, \widehat{\boldsymbol{\theta}}_{2n}, \rho)}{\partial \rho^3}\Big|_{\rho=\rho^*} = H_n(\widehat{\boldsymbol{\theta}}_{1n}, \widehat{\boldsymbol{\theta}}_{2n}, \rho^g) + S_1 + S_2 + S_3 \tag{4.14}$$

for some ρ^* lying between ρ and ρ^g. Now let us observe that,

$$\frac{1}{1+\beta}\frac{\partial H_n(\widehat{\boldsymbol{\theta}}_{1n},\widehat{\boldsymbol{\theta}}_{2n},\rho)}{\partial\rho}\Big|_{\rho=\rho^g} = \int f^{1+\beta}(\boldsymbol{x},\widehat{\boldsymbol{\theta}}_{1n},\widehat{\boldsymbol{\theta}}_{2n},\rho^g)U_\rho(\boldsymbol{x},\rho^g|\widehat{\boldsymbol{\theta}}_{1n},\widehat{\boldsymbol{\theta}}_{2n})\,d\boldsymbol{x} \qquad (4.15)$$
$$-\frac{1}{n}\sum_{i=1}^{n} f^{\beta}(\boldsymbol{X}_i,\widehat{\boldsymbol{\theta}}_{1n},\widehat{\boldsymbol{\theta}}_{2n},\rho^g)U_\rho(\boldsymbol{X}_i,\rho^g|\widehat{\boldsymbol{\theta}}_{1n},\widehat{\boldsymbol{\theta}}_{2n}),$$

where $U_\rho(\boldsymbol{x},\rho \mid \widehat{\boldsymbol{\theta}}_{1n},\widehat{\boldsymbol{\theta}}_{2n}) = \frac{\partial \log f(\boldsymbol{x},\widehat{\boldsymbol{\theta}}_{1n},\widehat{\boldsymbol{\theta}}_{2n},\rho)}{\partial\rho}$.
Let, $M(\boldsymbol{x},\widehat{\boldsymbol{\theta}}_{1n},\widehat{\boldsymbol{\theta}}_{2n},\rho^g) = f^{1+\beta}(\boldsymbol{x},\widehat{\boldsymbol{\theta}}_{1n},\widehat{\boldsymbol{\theta}}_{2n},\rho^g)U_\rho(\boldsymbol{x},\rho^g|\widehat{\boldsymbol{\theta}}_{1n},\widehat{\boldsymbol{\theta}}_{2n})$ and $N(\boldsymbol{X}_i,\widehat{\boldsymbol{\theta}}_{1n},\widehat{\boldsymbol{\theta}}_{2n},\rho^g) = f^{\beta}(\boldsymbol{X}_i,\widehat{\boldsymbol{\theta}}_{1n},\widehat{\boldsymbol{\theta}}_{2n},\rho^g)U_\rho(\boldsymbol{X}_i,\rho^g|\widehat{\boldsymbol{\theta}}_{1n},\widehat{\boldsymbol{\theta}}_{2n})$. Thus,

$$\frac{1}{1+\beta}\frac{\partial H_n(\widehat{\boldsymbol{\theta}}_{1n},\widehat{\boldsymbol{\theta}}_{2n},\rho)}{\partial\rho}\Big|_{\rho=\rho^g} = \int M(\boldsymbol{x},\widehat{\boldsymbol{\theta}}_{1n},\widehat{\boldsymbol{\theta}}_{2n},\rho^g)\,d\boldsymbol{x} - \frac{1}{n}\sum_{i=1}^{n} N(\boldsymbol{X}_i,\widehat{\boldsymbol{\theta}}_{1n},\widehat{\boldsymbol{\theta}}_{2n},\rho^g). \qquad (4.16)$$

Now, we are going to consider the first order Taylor series expansions of the functions $M(\boldsymbol{x},\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\rho^g)$ with respect to $(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$ and around $(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) = (\boldsymbol{\theta}_1^{\boldsymbol{g}},\boldsymbol{\theta}_2^{\boldsymbol{g}})$ and $\frac{1}{n}\sum_{i=1}^{n} N(\boldsymbol{X}_i,\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\rho^g)$ with respect to $(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$ and around $(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) = (\boldsymbol{\theta}_1^{\boldsymbol{g}},\boldsymbol{\theta}_2^{\boldsymbol{g}})$ and evaluate the same functions at $(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) = (\widehat{\boldsymbol{\theta}}_{1n},\widehat{\boldsymbol{\theta}}_{2n})$.

$$\int M(\boldsymbol{x},\widehat{\boldsymbol{\theta}}_{1n},\widehat{\boldsymbol{\theta}}_{2n},\rho^g)\,d\boldsymbol{x} = \int M(\boldsymbol{x},\boldsymbol{\theta}_1^{\boldsymbol{g}},\boldsymbol{\theta}_2^{\boldsymbol{g}},\rho^g)\,d\boldsymbol{x}$$
$$+\sum_{j,k}(\widehat{\theta}_{njk}-\theta_{jk}^g)\int \frac{\partial M(\boldsymbol{x},\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\rho^g)}{\partial\theta_{jk}}\Big|_{(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)=(\boldsymbol{\theta}_1^*,\boldsymbol{\theta}_2^*)},$$

where $(\boldsymbol{\theta}_1^*,\boldsymbol{\theta}_2^*)$ lies between $(\widehat{\boldsymbol{\theta}}_{1n},\widehat{\boldsymbol{\theta}}_{2n})$ and $(\boldsymbol{\theta}_1^{\boldsymbol{g}},\boldsymbol{\theta}_2^{\boldsymbol{g}})$ and

$$\frac{1}{n}\sum_{i=1}^{n} N(\boldsymbol{X}_i,\widehat{\boldsymbol{\theta}}_{1n},\widehat{\boldsymbol{\theta}}_{2n},\rho^g) = \frac{1}{n}\sum_{i=1}^{n} N(\boldsymbol{X}_i,\boldsymbol{\theta}_1^{\boldsymbol{g}},\boldsymbol{\theta}_2^{\boldsymbol{g}},\rho^g)$$
$$+\sum_{j,k}(\widehat{\theta}_{njk}-\theta_{jk}^g)\frac{1}{n}\sum_{i=1}^{n}\frac{\partial N(\boldsymbol{X}_i,\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\rho^g)}{\partial\theta_{jk}}\Big|_{(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)=(\boldsymbol{\theta}_1^{**},\boldsymbol{\theta}_2^{**})},$$

where $(\boldsymbol{\theta}_1^{**},\boldsymbol{\theta}_2^{**})$ lies between $(\widehat{\boldsymbol{\theta}}_{1n},\widehat{\boldsymbol{\theta}}_{2n})$ and $(\boldsymbol{\theta}_1^{\boldsymbol{g}},\boldsymbol{\theta}_2^{\boldsymbol{g}})$. Now by weak law of large numbers (WLLN) and Assumption 4.4,

$$\frac{1}{n}\sum_{i=1}^{n} N(\boldsymbol{X}_i, \boldsymbol{\theta}_1^{\boldsymbol{g}}, \boldsymbol{\theta}_2^{\boldsymbol{g}}, \rho^g) \xrightarrow{p} E_g\left(N(\boldsymbol{X}_1, \boldsymbol{\theta}_1^{\boldsymbol{g}}, \boldsymbol{\theta}_2^{\boldsymbol{g}}, \rho^g)\right),$$

$$\begin{aligned}\Big|\frac{1}{n}\sum_{i=1}^{n} \frac{\partial N(\boldsymbol{X}_i, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \rho^g)}{\partial \theta_{jk}}\Big|_{(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)=(\boldsymbol{\theta}_1^{**},\boldsymbol{\theta}_2^{**})}\Big| &< \frac{1}{n}\sum_{i=1}^{n}\Big|\frac{\partial N(\boldsymbol{X}_i, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \rho^g)}{\partial \theta_{jk}}\Big|_{(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)=(\boldsymbol{\theta}_1^{**},\boldsymbol{\theta}_2^{**})}\Big| \\ &< \frac{1}{n}\sum_{i=1}^{n} N_{jk}(\boldsymbol{X}_i, \rho^g) \xrightarrow{p} E_g(N_{jk}(\boldsymbol{X}_1, \rho^g)) < \infty.\end{aligned}$$

We also have $\left|\int \frac{\partial M(\boldsymbol{x},\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\rho^g)}{\partial \theta_{jk}}\Big|_{(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)=(\boldsymbol{\theta}_1^{*},\boldsymbol{\theta}_2^{*})}\, d\boldsymbol{x}\right| < \int\left|\frac{\partial M(\boldsymbol{x},\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\rho^g)}{\partial \theta_{jk}}\Big|_{(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)=(\boldsymbol{\theta}_1^{*},\boldsymbol{\theta}_2^{*})}\right| d\boldsymbol{x} < \int M_{jk}(\boldsymbol{x}, \rho_g)\, d\boldsymbol{x} < \infty$. From (4.8), we have $\widehat{\theta}_{njk} \xrightarrow{p} \theta_{jk}^g$ for $j, k = 1, 2$. Now using the aforesaid observations, we can write from Equation (4.16) that,

$$\begin{aligned}\frac{\partial H_n(\widehat{\boldsymbol{\theta}}_{1n}, \widehat{\boldsymbol{\theta}}_{2n}, \rho)}{\partial \rho}\Big|_{\rho=\rho^g} &\xrightarrow{p} \int M(\boldsymbol{x}, \boldsymbol{\theta}_1^{\boldsymbol{g}}, \boldsymbol{\theta}_2^{\boldsymbol{g}}, \rho^g)\, d\boldsymbol{x} - E_g\left(N(\boldsymbol{X}_1, \boldsymbol{\theta}_1^{\boldsymbol{g}}, \boldsymbol{\theta}_2^{\boldsymbol{g}}, \rho^g)\right) \\ &= \frac{\partial H(\boldsymbol{\theta}_1^{\boldsymbol{g}}, \boldsymbol{\theta}_2^{\boldsymbol{g}}, \rho)}{\partial \rho}\Big|_{\rho=\rho^g} = 0\end{aligned}$$

as $\rho^g = \underset{\rho}{\operatorname{argmin}}\ H(\boldsymbol{\theta}_1^{\boldsymbol{g}}, \boldsymbol{\theta}_2^{\boldsymbol{g}}, \rho)$. Thus, $\frac{\partial H_n(\widehat{\boldsymbol{\theta}}_{1n},\widehat{\boldsymbol{\theta}}_{2n},\rho)}{\partial \rho}\Big|_{\rho=\rho^g} \xrightarrow{p} 0$ as $n \to \infty$. Since $\rho \in Surface(Q_\epsilon)$, $|\rho - \rho^g| = \epsilon$ and by the fact that $\frac{\partial H_n(\widehat{\boldsymbol{\theta}}_{1n},\widehat{\boldsymbol{\theta}}_{2n},\rho)}{\partial \rho}\Big|_{\rho=\rho^g} \xrightarrow{p} 0$, we have

$$|S_1| < \epsilon^3 \tag{4.17}$$

with probability tending to 1. Our next agenda is to handle the term S_2 in the right hand side of Equation (4.14). Let us note that,

$$\begin{aligned}&\tfrac{1}{1+\beta}\tfrac{\partial^2 H_n(\widehat{\boldsymbol{\theta}}_{1n},\widehat{\boldsymbol{\theta}}_{2n},\rho)}{\partial^2 \rho}\Big|_{\rho=\rho^g} = (1+\beta)\textstyle\int f^{1+\beta}(\boldsymbol{x}, \widehat{\boldsymbol{\theta}}_{1n}, \widehat{\boldsymbol{\theta}}_{2n}, \rho^g) U_\rho^2(\boldsymbol{x}, \rho^g | \widehat{\boldsymbol{\theta}}_{1n}, \widehat{\boldsymbol{\theta}}_{2n})\, d\boldsymbol{x} \\ &+ \textstyle\int f^{1+\beta}(\boldsymbol{x}, \widehat{\boldsymbol{\theta}}_{1n}, \widehat{\boldsymbol{\theta}}_{2n}, \rho^g) \frac{\partial\, U_\rho(\boldsymbol{x},\rho|\widehat{\boldsymbol{\theta}}_{1n},\widehat{\boldsymbol{\theta}}_{2n})}{\partial \rho}\Big|_{\rho=\rho^g} d\boldsymbol{x} \\ &- \tfrac{\beta}{n}\textstyle\sum_{i=1}^{n} f^{\beta}(\boldsymbol{X}_i, \widehat{\boldsymbol{\theta}}_{1n}, \widehat{\boldsymbol{\theta}}_{2n}, \rho^g) U_\rho^2(\boldsymbol{X}_i, \rho^g | \widehat{\boldsymbol{\theta}}_{1n}, \widehat{\boldsymbol{\theta}}_{2n}) \\ &- \tfrac{1}{n}\textstyle\sum_{i=1}^{n} f^{\beta}(\boldsymbol{X}_i, \widehat{\boldsymbol{\theta}}_{1n}, \widehat{\boldsymbol{\theta}}_{2n}, \rho^g) \frac{\partial\, U_\rho(\boldsymbol{X}_i,\rho|\widehat{\boldsymbol{\theta}}_{1n},\widehat{\boldsymbol{\theta}}_{2n})}{\partial \rho}\Big|_{\rho=\rho^g}.\end{aligned} \tag{4.18}$$

Now, by using the same first order Taylor series expansion trick as in case of S_1, we can prove (after some algebra) that,

$$\frac{\partial^2 H_n(\widehat{\boldsymbol{\theta}}_{1n}, \widehat{\boldsymbol{\theta}}_{2n}, \rho)}{\partial^2 \rho}\Big|_{\rho=\rho^g} \xrightarrow{p} \frac{\partial^2 H(\boldsymbol{\theta}_1^{\boldsymbol{g}}, \boldsymbol{\theta}_2^{\boldsymbol{g}}, \rho)}{\partial^2 \rho}\Big|_{\rho=\rho^g} > 0$$

as ρ^g is the minimizer of $H(\boldsymbol{\theta}_1^{\boldsymbol{g}}, \boldsymbol{\theta}_2^{\boldsymbol{g}}, \rho)$. Thus for any $\rho \in Surface(Q_\epsilon)$, with probability tending to 1,

$$S_2 > c\epsilon^2 \tag{4.19}$$

for some constant $c > 0$. Now it only remains to take care of the third term S_3 in the right hand side of Equation (4.14). Let us observe that,

$$\begin{aligned}
\frac{\partial^3 H_n(\widehat{\boldsymbol{\theta}}_{1n}, \widehat{\boldsymbol{\theta}}_{2n}, \rho)}{\partial^3 \rho}\Big|_{\rho=\rho^*} &= \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^3 V(\boldsymbol{X}_i, \widehat{\boldsymbol{\theta}}_{1n}, \widehat{\boldsymbol{\theta}}_{2n}, \rho)}{\partial^3 \rho}\Big|_{\rho=\rho^*} \\
&= \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^3 V(\boldsymbol{X}_i, \boldsymbol{\theta}_1^{\boldsymbol{g}}, \boldsymbol{\theta}_2^{\boldsymbol{g}}, \rho)}{\partial^3 \rho}\Big|_{\rho=\rho^*} \\
&+ \sum_{j,k} (\widehat{\theta}_{njk} - \theta_{jk}^g) \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^4 V(\boldsymbol{X}_i, \widehat{\boldsymbol{\theta}}_{1n}, \widehat{\boldsymbol{\theta}}_{2n}, \rho)}{\partial^3 \rho \partial \theta_{jk}}\Big|_{\rho=\rho^*, (\boldsymbol{\theta}_1 \boldsymbol{\theta}_2)=(\boldsymbol{\theta}_1^\dagger, \boldsymbol{\theta}_2^\dagger)}.
\end{aligned}$$

We utilize Assumption 4.5 at this point to show that S_3 is finite in absolute sense with probability tending to 1. By Assumption 4.5 and WLLN, we have,

$$\begin{aligned}
\left| \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^3 V(\boldsymbol{X}_i, \boldsymbol{\theta}_1^{\boldsymbol{g}}, \boldsymbol{\theta}_2^{\boldsymbol{g}}, \rho)}{\partial \rho^3}\Big|_{\rho=\rho^*} \right| &< \frac{1}{n}\sum_{i=1}^{n} \left| \frac{\partial^3 V(\boldsymbol{X}_i, \boldsymbol{\theta}_1^{\boldsymbol{g}}, \boldsymbol{\theta}_2^{\boldsymbol{g}}, \rho)}{\partial \rho^3}\Big|_{\rho=\rho^*} \right| \\
&< \frac{1}{n}\sum_{i=1}^{n} v_1(\boldsymbol{X}_i, \boldsymbol{\theta}_1^{\boldsymbol{g}}, \boldsymbol{\theta}_2^{\boldsymbol{g}}) \xrightarrow{p} E_g v_1(\boldsymbol{X}_1, \boldsymbol{\theta}_1^{\boldsymbol{g}}, \boldsymbol{\theta}_2^{\boldsymbol{g}}) = d(\text{Say}) \\
&< \infty,
\end{aligned}$$

$$\begin{aligned}
&\left| \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^4 V(\boldsymbol{X}_i, \widehat{\boldsymbol{\theta}}_{1n}, \widehat{\boldsymbol{\theta}}_{2n}, \rho)}{\partial \rho^3 \partial \theta_{jk}} \Big|_{\rho=\rho^*, (\boldsymbol{\theta}_1 \boldsymbol{\theta}_2)=(\boldsymbol{\theta}_1^\dagger, \boldsymbol{\theta}_2^\dagger)} \right| \\
&< \frac{1}{n} \sum_{i=1}^{n} \left| \frac{\partial^4 V(\boldsymbol{X}_i, \widehat{\boldsymbol{\theta}}_{1n}, \widehat{\boldsymbol{\theta}}_{2n}, \rho)}{\partial \rho^3 \partial \theta_{jk}} \Big|_{\rho=\rho^*, (\boldsymbol{\theta}_1 \boldsymbol{\theta}_2)=(\boldsymbol{\theta}_1^\dagger, \boldsymbol{\theta}_2^\dagger)} \right| \\
&< \frac{1}{n} \sum_{i=1}^{n} v_2(\boldsymbol{X}_i, \boldsymbol{\theta}_1^{\boldsymbol{g}}, \boldsymbol{\theta}_2^{\boldsymbol{g}}) \xrightarrow{p} E_g v_2(\boldsymbol{X}_1, \boldsymbol{\theta}_1^{\boldsymbol{g}}, \boldsymbol{\theta}_2^{\boldsymbol{g}}) \\
&< \infty
\end{aligned}$$

for $j, k = 1, 2$. From (4.8), we have $\widehat{\theta}_{njk} \xrightarrow{p} \theta_{jk}^g$ for $j, k = 1, 2$. Thus, for $\rho \in Surface(Q_\epsilon)$,

$$|S_3| = \left| \frac{(\rho - \rho^g)^3}{3!} \right| \left| \frac{\partial^3 H_n(\widehat{\boldsymbol{\theta}}_{1n}, \widehat{\boldsymbol{\theta}}_{2n}, \rho)}{\partial \rho^3} \Big|_{\rho=\rho^*} \right| < d\epsilon^3, \text{ for some constant } d > 0. \quad (4.20)$$

Now, from (4.14), (4.17), (4.19) and (4.20), we have

$$H_n(\widehat{\boldsymbol{\theta}}_{1n}, \widehat{\boldsymbol{\theta}}_{2n}, \rho^g) - H_n(\widehat{\boldsymbol{\theta}}_{1n}, \widehat{\boldsymbol{\theta}}_{2n}, \rho) = -S_1 - S_2 - S_3 \leq |S_1| - S_2 + |S_3| < \epsilon^3 - c\epsilon^2 + d\epsilon^3,$$

which is less than 0 if $\epsilon < \frac{c}{1+d}$. Hence, for any sufficiently small ϵ, we have $H_n(\widehat{\boldsymbol{\theta}}_{1n}, \widehat{\boldsymbol{\theta}}_{2n}, \rho^g)$ $< H_n(\widehat{\boldsymbol{\theta}}_{1n}, \widehat{\boldsymbol{\theta}}_{2n}, \rho)$ for all $\rho \in Surface(Q_\epsilon)$ with probability tending to 1. This implies, with probability tending to 1, there exists a sequence $\{\widehat{\rho}_{n\epsilon}\} \in Interior(Q_\epsilon)$ which minimizes $H_n(\widehat{\boldsymbol{\theta}}_{1n}, \widehat{\boldsymbol{\theta}}_{2n}, \rho)$ for all sufficiently small ϵ.

Let us define $\widehat{\rho}_n^*$ to be the minimizer which is closest to ρ^g among all $\widehat{\rho}_{n\epsilon}$ for any sufficiently small ϵ. Thus $\widehat{\rho}_n^* \in Interior(Q_\epsilon)$ for all small ϵ with probability tending to 1 as $n \to \infty$. Then,

$$P(|\widehat{\rho}_n^* - \rho^g| < \epsilon) = P(\widehat{\rho}_n^* \in Interior(Q_\epsilon)) \to 1$$

as $n \to \infty$ for all sufficiently small $\epsilon > 0$. This completes the proof of the existence of a consistent sequence of minimizer of the function $H_n(\widehat{\boldsymbol{\theta}}_{1n}, \widehat{\boldsymbol{\theta}}_{2n}, \rho)$. □

4.8.2 Proof of Theorem 4.2

Proof. Here also we are going to use the same Taylor series expansion trick as we did in Equation (4.14).

$$\frac{\partial H_n(\widehat{\boldsymbol{\theta}}_{1n}, \widehat{\boldsymbol{\theta}}_{2n}, \rho)}{\partial \rho}\Big|_{\rho=\widehat{\rho}_n} = \frac{\partial H_n(\widehat{\boldsymbol{\theta}}_{1n}, \widehat{\boldsymbol{\theta}}_{2n}, \rho)}{\partial \rho}\Big|_{\rho=\rho^g} + (\widehat{\rho}_n - \rho^g)\frac{\partial^2 H_n(\widehat{\boldsymbol{\theta}}_{1n}, \widehat{\boldsymbol{\theta}}_{2n}, \rho)}{\partial \rho^2}\Big|_{\rho=\rho^g} + \frac{(\widehat{\rho}_n - \rho^g)^2}{2!}\frac{\partial^3 H_n(\widehat{\boldsymbol{\theta}}_{1n}, \widehat{\boldsymbol{\theta}}_{2n}, \rho)}{\partial \rho^3}\Big|_{\rho=\rho^{**}}$$

for some ρ^{**} lies between $\widehat{\rho}_n$ and ρ^g. Hence, from Equation (4.9), we can rewrite the above as,

$$-\sqrt{n}\frac{\partial H_n(\widehat{\boldsymbol{\theta}}_{1n}, \widehat{\boldsymbol{\theta}}_{2n}, \rho)}{\partial \rho}\Big|_{\rho=\rho^g} = \sqrt{n}(\widehat{\rho}_n - \rho^g)\Big[\frac{\partial^2 H_n(\widehat{\boldsymbol{\theta}}_{1n}, \widehat{\boldsymbol{\theta}}_{2n}, \rho)}{\partial \rho^2}\Big|_{\rho=\rho^g} + \frac{(\widehat{\rho}_n - \rho^g)}{2!}\frac{\partial^3 H_n(\widehat{\boldsymbol{\theta}}_{1n}, \widehat{\boldsymbol{\theta}}_{2n}, \rho)}{\partial \rho^3}\Big|_{\rho=\rho^{**}}\Big]. \tag{4.21}$$

Now, let us consider the term in the left hand side of Equation (4.21). We can expand it as,

$$\begin{aligned}
&\sqrt{n}\frac{\partial H_n(\widehat{\boldsymbol{\theta}}_{1n}, \widehat{\boldsymbol{\theta}}_{2n}, \rho)}{\partial \rho}\Big|_{\rho=\rho^g} \\
&= \sqrt{n}\frac{\partial H_n(\boldsymbol{\theta}_1^{\boldsymbol{g}}, \boldsymbol{\theta}_2^{\boldsymbol{g}}, \rho)}{\partial \rho}\Big|_{\rho=\rho^g} + \sum_{j,k}\sqrt{n}(\widehat{\theta}_{njk} - \theta^g_{jk})\frac{\partial^2 H_n(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \rho)}{\partial \rho\, \partial \theta_{jk}}\Big|_{(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)=(\boldsymbol{\theta}_1^{\boldsymbol{g}},\boldsymbol{\theta}_2^{\boldsymbol{g}}),\rho=\rho^g} \\
&+ \sum_{j,k}\sum_{j',k'}\sqrt{n}(\widehat{\theta}_{njk} - \theta^g_{jk})(\widehat{\theta}_{nj'k'} - \theta^g_{j'k'})\frac{\partial^3 H_n(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \rho)}{\partial \rho\, \partial \theta_{jk}\, \partial \theta_{j'k'}}\Big|_{(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)=(\boldsymbol{\theta}_1^{\boldsymbol{a}*},\boldsymbol{\theta}_2^{\boldsymbol{a}*}),\rho=\rho^g} \\
&= \sqrt{n}\frac{\partial H_n(\boldsymbol{\theta}_1^{\boldsymbol{g}}, \boldsymbol{\theta}_2^{\boldsymbol{g}}, \rho)}{\partial \rho}\Big|_{\rho=\rho^g} + \sum_{j,k}\sqrt{n}(\widehat{\theta}_{njk} - \theta^g_{jk})\Bigg[\frac{\partial^2 H_n(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \rho)}{\partial \rho\, \partial \theta_{jk}}\Big|_{(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)=(\boldsymbol{\theta}_1^{\boldsymbol{g}},\boldsymbol{\theta}_2^{\boldsymbol{g}}),\rho=\rho^g} \\
&+ \sum_{j',k'}(\widehat{\theta}_{nj'k'} - \theta^g_{j'k'})\frac{\partial^3 H_n(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \rho)}{\partial \rho\, \partial \theta_{jk}\, \partial \theta_{j'k'}}\Big|_{(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)=(\boldsymbol{\theta}_1^{\boldsymbol{a}*},\boldsymbol{\theta}_2^{\boldsymbol{a}*}),\rho=\rho^g}\Bigg].
\end{aligned}$$

Hence, Equation (4.21) can be rewritten as,

$$
\begin{aligned}
&-\sqrt{n}\frac{\partial H_n(\boldsymbol{\theta}_1^{\boldsymbol{g}},\boldsymbol{\theta}_2^{\boldsymbol{g}},\rho)}{\partial\rho}\Big|_{\rho=\rho^g} = \\
&\textstyle\sum_{j,k}\sqrt{n}(\widehat{\theta}_{njk}-\theta_{jk}^g)\left[\frac{\partial^2 H_n(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\rho)}{\partial\rho\,\partial\theta_{jk}}\Big|_{(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)=(\boldsymbol{\theta}_1^{\boldsymbol{g}},\boldsymbol{\theta}_2^{\boldsymbol{g}}),\rho=\rho^g}\right. \\
&\left.+\textstyle\sum_{j',k'}(\widehat{\theta}_{nj'k'}-\theta_{j'k'}^g)\frac{\partial^3 H_n(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\rho)}{\partial\rho\,\partial\theta_{jk}\,\partial\theta_{j'k'}}\Big|_{(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)=(\boldsymbol{\theta}_1^{a*},\boldsymbol{\theta}_2^{a*}),\rho=\rho^g}\right]+ \\
&\sqrt{n}(\widehat{\rho}_n-\rho^g)\left[\frac{\partial^2 H_n(\widehat{\boldsymbol{\theta}}_{1n},\widehat{\boldsymbol{\theta}}_{2n},\rho)}{\partial\rho^2}\Big|_{\rho=\rho^g}+\frac{(\widehat{\rho}_n-\rho^g)}{2!}\frac{\partial^3 H_n(\widehat{\boldsymbol{\theta}}_{1n},\widehat{\boldsymbol{\theta}}_{2n},\rho)}{\partial\rho^3}\Big|_{\rho=\rho^{**}}\right].
\end{aligned}
\tag{4.22}
$$

Next, let us consider the estimators $\widehat{\boldsymbol{\theta}}_{1n}$ and $\widehat{\boldsymbol{\theta}}_{2n}$ which were derived separately before finding $\widehat{\rho}_n$. To do that, let us first introduce the following notations. Let $H_{1n}^j(\boldsymbol{\theta}_1) = \frac{\partial\, H_{1n}(\boldsymbol{\theta}_1)}{\partial\theta_{1j}}, H_{2n}^j(\boldsymbol{\theta}_2) = \frac{\partial\, H_{2n}(\boldsymbol{\theta}_2)}{\partial\theta_{2j}}$ for $j = 1, 2$. Similarly, $H_{1n}^{jk}(\boldsymbol{\theta}_1)$, $H_{2n}^{jk}(\boldsymbol{\theta}_2)$ and $H_{1n}^{jkl}(\boldsymbol{\theta}_1)$, $H_{2n}^{jkl}(\boldsymbol{\theta}_2)$ denote the second and third order partial derivatives of $H_{1n}(\boldsymbol{\theta}_1)$ and $H_{2n}(\boldsymbol{\theta}_2)$, respectively. Let us consider the following Taylor series expansion,

$$
H_{1n}^j(\widehat{\boldsymbol{\theta}}_{1n}) = H_{1n}^j(\boldsymbol{\theta}_1^{\boldsymbol{g}}) + \sum_k(\widehat{\theta}_{1kn}-\theta_{1k}^g)H_{1n}^{jk}(\boldsymbol{\theta}_1^{\boldsymbol{g}}) + \frac{1}{2!}\sum_{k,l}(\widehat{\theta}_{1kn}-\theta_{1k}^g)(\widehat{\theta}_{1ln}-\theta_{1l}^g)H_{1n}^{jkl}(\boldsymbol{\theta}_1^*)
$$

for some $\boldsymbol{\theta}_1^*$ lying between $\widehat{\boldsymbol{\theta}}_{1n}$ and $\boldsymbol{\theta}_1^{\boldsymbol{g}}$. But from Equation (4.4), we have $H_{1n}^j(\widehat{\boldsymbol{\theta}}_{1n}) = 0$. Thus the aforesaid equation can be rewritten as,

$$
-\sqrt{n}H_{1n}^j(\boldsymbol{\theta}_1^{\boldsymbol{g}}) = \sum_k\sqrt{n}(\widehat{\theta}_{1kn}-\theta_{1k}^g)\left[H_{1n}^{jk}(\boldsymbol{\theta}_1^{\boldsymbol{g}}) + \frac{1}{2!}\sum_l(\widehat{\theta}_{1ln}-\theta_{1l}^g)H_{1n}^{jkl}(\boldsymbol{\theta}_1^*)\right]. \tag{4.23}
$$

Similarly, we have

$$
-\sqrt{n}H_{2n}^j(\boldsymbol{\theta}_2^{\boldsymbol{g}}) = \sum_k\sqrt{n}(\widehat{\theta}_{2kn}-\theta_{2k}^g)\left[H_{2n}^{jk}(\boldsymbol{\theta}_2^{\boldsymbol{g}}) + \frac{1}{2!}\sum_l(\widehat{\theta}_{2ln}-\theta_{2l}^g)H_{2n}^{jkl}(\boldsymbol{\theta}_2^*)\right] \tag{4.24}
$$

for some $\boldsymbol{\theta}_2^*$ lying between $\widehat{\boldsymbol{\theta}}_{2n}$ and $\boldsymbol{\theta}_2^{\boldsymbol{g}}$. Now let us consider System (4.25) of linear equations which is derived by assembling Equations (4.22), (4.23) and (4.24).

$$\begin{aligned}
&\textstyle\sum_k \sqrt{n}(\widehat{\theta}_{1kn} - \theta_{1k}^g) b_{jkn}^{(1)} = -\sqrt{n} H_{1n}^j(\boldsymbol{\theta}_{\mathbf{1}}^{\boldsymbol{g}}),\ j = 1, 2, \\
&\textstyle\sum_k \sqrt{n}(\widehat{\theta}_{2kn} - \theta_{2k}^g) b_{jkn}^{(2)} = -\sqrt{n} H_{2n}^j(\boldsymbol{\theta}_2^{\boldsymbol{g}}),\ j = 1, 2, \\
&\textstyle\sum_{j,k} \sqrt{n}(\widehat{\theta}_{njk} - \theta_{jk}^g) e_{jkn} + \sqrt{n}(\widehat{\rho}_n - \rho^g) a_n = -\sqrt{n} \frac{\partial H_n(\boldsymbol{\theta}_1^{\boldsymbol{g}}, \boldsymbol{\theta}_2^{\boldsymbol{g}}, \rho)}{\partial \rho}\Big|_{\rho=\rho^g},
\end{aligned} \tag{4.25}$$

where $b_{jkn}^{(1)} = \left[H_{1n}^{jk}(\boldsymbol{\theta}_1^{\boldsymbol{g}}) + \frac{1}{2!}\sum_l (\widehat{\theta}_{1ln} - \theta_{1l}^g) H_{1n}^{jkl}(\boldsymbol{\theta}_1^*)\right]$, $b_{jkn}^{(2)} = \left[H_{2n}^{jk}(\boldsymbol{\theta}_2^{\boldsymbol{g}}) + \frac{1}{2!}\sum_l (\widehat{\theta}_{2ln} - \theta_{2l}^g) H_{2n}^{jkl}(\boldsymbol{\theta}_2^*)\right]$, $e_{jkn} = \left[\frac{\partial^2 H_n(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \rho)}{\partial \rho\, \partial \theta_{jk}}\Big|_{(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = (\boldsymbol{\theta}_1^{\boldsymbol{g}}, \boldsymbol{\theta}_2^{\boldsymbol{g}}), \rho = \rho^g} + \sum_{j',k'} (\widehat{\theta}_{nj'k'} - \theta_{j'k'}^g) \frac{\partial^3 H_n(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \rho)}{\partial \rho\, \partial \theta_{jk}\, \partial \theta_{j'k'}}\Big|_{(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = (\boldsymbol{\theta}_1^{\boldsymbol{a}*}, \boldsymbol{\theta}_2^{\boldsymbol{a}*}), \rho = \rho^g}\right]$ for $j, k = 1, 2$ and $a_n = \left[\frac{\partial^2 H_n(\widehat{\boldsymbol{\theta}}_{1n}, \widehat{\boldsymbol{\theta}}_{2n}, \rho)}{\partial \rho^2}\Big|_{\rho=\rho^g} + \frac{(\widehat{\rho}_n - \rho^g)}{2!} \frac{\partial^3 H_n(\widehat{\boldsymbol{\theta}}_{1n}, \widehat{\boldsymbol{\theta}}_{2n}, \rho)}{\partial \rho^3}\Big|_{\rho=\rho^{**}}\right]$. Now, by following the proof of Theorem 9.2(b) of Basu et al. (2011) [12], we can show that, $b_{jkn}^{(l)} \xrightarrow{p} b_{jk}^{(l)}$ for $j, k, l = 1, 2$ (recall the elements of the matrix $\boldsymbol{B}$ (Theorem 4.2)). Using similar kind of arguments and assumption we have made to prove theorem 4.1, it can be proved that, $e_{jkn} \xrightarrow{p} e_{jk}$, $j, k = 1, 2$ and $a_n \xrightarrow{p} a$ as $n \to \infty$.

By the central limit theorem (CLT), we have

$$\sqrt{n}\left(-\sqrt{n}H_{1n}^1(\boldsymbol{\theta}_1^{\boldsymbol{g}}), -\sqrt{n}H_{1n}^2(\boldsymbol{\theta}_1^{\boldsymbol{g}}), -\sqrt{n}H_{2n}^1(\boldsymbol{\theta}_2^{\boldsymbol{g}}), -\sqrt{n}H_{2n}^2(\boldsymbol{\theta}_2^{\boldsymbol{g}}), -\sqrt{n}\frac{\partial H_n(\boldsymbol{\theta}_1^{\boldsymbol{g}}, \boldsymbol{\theta}_2^{\boldsymbol{g}}, \rho)}{\partial \rho}\Big|_{\rho=\rho^g}\right) \xrightarrow{d} N(0, \Gamma_0).$$

Now, using the aforesaid observations and Lemma 4.1 of Lehmann (1983) [92], we have

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^{\boldsymbol{g}}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{B}^{-1}\boldsymbol{\Gamma}_{\mathbf{0}}\boldsymbol{B}^{-1'}),$$

where $\widehat{\boldsymbol{\theta}}_n = (\widehat{\theta}_{11n}, \widehat{\theta}_{12n}, \widehat{\theta}_{21n}, \widehat{\theta}_{22n}, \widehat{\rho}_n)$ and $\boldsymbol{\theta}^{\boldsymbol{g}} = (\theta_{11}^g, \theta_{12}^g, \theta_{21}^g, \theta_{22}^g, \rho^g)$. □

4.8.3 Proof of Theorem 4.3

Proof. According to our algorithm, we first derive $\widehat{\mu}_{jn}$ and $\widehat{\sigma}_{jn}^2$ marginally from each of the p components. Thus by Theorem 9.1 of Basu et al. (2011) [12], we have, $\hat{\mu}_{jn} \xrightarrow{p} \mu_j^g$ and $\widehat{\sigma}_{jn}^2 \xrightarrow{p} \sigma^{g2}_{jn}$ as $n \to \infty$. In the next step, our algorithm finds the estimates $\widehat{\rho}_{jkn}$ for each pair of components separately. By Theorem 4.1, we have, $\widehat{\rho}_{jkn} \xrightarrow{p} \rho_{jk}^g$. For a fixed

$\epsilon > 0$,

$$
\begin{aligned}
P[||\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta^g}||_2 > \epsilon] = P[||\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta^g}||_2^2 > \epsilon^2] &= P\left[\sum_{l=1}^{P}(\widehat{\boldsymbol{\theta}}_{nl} - \boldsymbol{\theta^g}_l)^2 > \epsilon^2\right] \\
&\leq \sum_{l=1}^{P} P\left[(\widehat{\boldsymbol{\theta}}_{nl} - \boldsymbol{\theta^g}_l)^2 > \frac{\epsilon^2}{P}\right] \\
&\qquad \text{(by Boole's inequality)} \\
&\to 0 \text{ as } n \to \infty
\end{aligned}
$$

and this is true for all $\epsilon > 0$. Thus, $\widehat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta^g}$. Now the asymptotic normality of $\widehat{\boldsymbol{\theta}}_n$ can be similarly proved but the form of the asymptotic covariance matrix of $\sqrt{n}\widehat{\boldsymbol{\theta}}_n$ is quite large which is provided just after stating Theorem 4.3. The proof is exactly the same as the proof of Theorem 4.2. □

4.8.4 Proof of Theorem 4.4

Proof. Let, $S_n = \{\boldsymbol{x} = (x_1, x_2, \cdots, x_p) \in \mathbb{R}^p : \boldsymbol{x}^{'}\widehat{\boldsymbol{\Sigma}}_n\boldsymbol{x} > 0 \text{ and } \boldsymbol{x}^{'}\boldsymbol{x} = 1\}$. It is enough to show that $P(S_n) \to 1$ as $n \to \infty$.

Since $\boldsymbol{\Sigma^g}$ is positive definite, $\boldsymbol{x}^{'}\boldsymbol{\Sigma^g}\boldsymbol{x} > 0 \ \forall \ \boldsymbol{x} \neq 0$. By considering the spectral decomposition of the symmetric matrix $\boldsymbol{\Sigma^g}$, we have

$$
\boldsymbol{x}^{'}\boldsymbol{\Sigma^g}\boldsymbol{x} \geq \lambda^g_{(1)} \geq c. \tag{4.26}
$$

From Theorem 4.1, we have, $\widehat{\sigma}_{ijn} \xrightarrow{p} \sigma^g_{ij}$ as $n \to \infty, \ 1 \leq i, j \leq p$. Let us fix an $\epsilon \in (0, \frac{c}{p})$. So,

$$
\begin{aligned}
|\widehat{\sigma}_{ijn} - \sigma^g_{ij}| \leq \epsilon \implies & |x_i x_j \widehat{\sigma}_{ijn} - x_i x_j \sigma^g_{ij}| = |x_i||x_j||\widehat{\sigma}_{ijn} - \sigma^g_{ij}| \leq |x_i||x_j|\epsilon \\
\implies & x_i x_j \widehat{\sigma}_{ijn} \geq x_i x_j \sigma^g_{ij} - |x_i||x_j|\epsilon
\end{aligned}
$$

with probability tending to 1 for all $1 \leq i, j \leq p$ (here $|x_i|$ is the i-th component of the

unit vector $\boldsymbol{x}$). Now,

$$\begin{aligned}
&\boldsymbol{x}^{'}\widehat{\boldsymbol{\Sigma}}_n\boldsymbol{x}\\
&=\sum_{i,j}x_ix_j\widehat{\sigma}_{ijn}\geq\sum_{i,j}x_ix_j\sigma_{ij}^{g}-\epsilon\sum_{i,j}|x_i||x_j|\\
&=\boldsymbol{x}^{'}\boldsymbol{\Sigma}^{\boldsymbol{g}}\boldsymbol{x}-\epsilon\left(\sum_{i=1}^{p}|x_i|\right)^2\geq c-p\epsilon\text{ , by (4.26) and Cauchy-Schwarz inequality}\\
&>0
\end{aligned}$$

with probability tending to 1 for all unit vector $\boldsymbol{x}(\neq 0)\in\mathbb{R}^p$. Thus $P(S_n)\to 1$ as $n\to\infty$. □

4.8.5 Asymptotic Variances of the SMDPDE and MDPDE Variance and Correlation Estimators

Let us first introduce a set of algebraic expressions, namely, $d_{22}=\frac{\beta^2+2}{4(2\pi)^{\beta/2}(1+\beta)^{\frac{3}{2}}\sigma_1^{\beta+4}}$, $d_{44}=\frac{\beta^2+2}{4(2\pi)^{\beta/2}(1+\beta)^{\frac{3}{2}}\sigma_2^{\beta+4}}$, $E_{12}=\rho(1+\beta^2)+\rho\beta\Big(1-\beta-\frac{2\rho^4-4\rho^2+2}{(1+\beta)(1-\rho^2)^2}\Big)$, $e_1=-\frac{E_{12}}{2\sigma_1^2(\sigma_1\sigma_2)^{\beta}(2\pi)^{\beta}(1-\rho^2)^{1+\frac{\beta}{2}}(1+\beta)}$, $e_2=-\frac{E_{12}}{2\sigma_2^2(\sigma_1\sigma_2)^{\beta}(2\pi)^{\beta}(1-\rho^2)^{1+\frac{\beta}{2}}(1+\beta)}$, $F=1+\rho^2(1+\beta)-\frac{\beta(2\rho^6-3\rho^4+1)}{(1+\beta)(1-\rho^2)^2}$ and $a=\frac{F}{(\sigma_1\sigma_2)^{\beta}(2\pi)^{\beta}(1-\rho^2)^{2+\frac{\beta}{2}}(1+\beta)}$. Let us consider the matrix

$$\boldsymbol{B}=\begin{bmatrix}d_{22}&0&0\\0&d_{44}&0\\e_1&e_2&a\end{bmatrix}.$$

Let us also define, $\gamma_{22}=\frac{1}{\sqrt{1+2\beta}}\left[\frac{1}{2}+\frac{3}{2(1+2\beta)^2}-\frac{1}{1+2\beta}\right]-\frac{\beta^2}{2(1+\beta)^3}$, $\Gamma_{22}=\frac{(1+\beta)^2}{2(2\pi)^{\beta}\sigma_1^{2\beta+4}}\gamma_{22}$, $\Gamma_{44}=\frac{(1+\beta)^2}{2(2\pi)^{\beta}\sigma_2^{2\beta+4}}\gamma_{22}$, $\gamma_{24}=\frac{1}{\sqrt{(\beta+1)^2-(\beta\rho)^2}}\left[\left(1-\frac{1+\beta(1-\rho^2)}{(\beta+1)^2-(\beta\rho)^2}\right)^2+\frac{2\rho^2}{((\beta+1)^2-(\beta\rho)^2)^2}\right]-\frac{\beta^2}{(1+\beta)^3}$, $\Gamma_{24}=\frac{(1+\beta)^2\gamma_{24}}{4(2\pi)^{\beta}(\sigma_1\sigma_2)^{\beta+2}}$, $\Gamma_{25}=\frac{\beta^2\rho}{2(2\pi)^{\frac{3\beta}{2}}\sigma_1^{2\beta+2}\sigma_2^{\beta}(1-\rho^2)^{1+\frac{\beta}{2}}(1+\beta)^{\frac{3}{2}}}\left[1-\frac{2(1+\beta)^2}{(1+2\beta)^{\frac{3}{2}}}\right]$, $\Gamma_{45}=\frac{\beta^2\rho}{2(2\pi)^{\frac{3\beta}{2}}\sigma_2^{2\beta+2}\sigma_1^{\beta}(1-\rho^2)^{1+\frac{\beta}{2}}(1+\beta)^{\frac{3}{2}}}\left[1-\frac{2(1+\beta)^2}{(1+2\beta)^{\frac{3}{2}}}\right]$, $\gamma_{55}=\frac{\rho^2}{1+2\beta}+\frac{1-3\rho^4+2\rho^6}{(1-\rho^2)^2(1+2\beta)^3}-\frac{2\rho^2}{(1+2\beta)^2}-$

$\frac{\beta^2\rho^2}{(1+\beta)^4}$ and $\Gamma_{55} = \frac{(1+\beta)^2\gamma_{55}}{(2\pi)^{2\beta}(\sigma_1\sigma_2)^{2\beta}(1-\rho^2)^{\beta+2}}$. Let us consider the matrix

$$\boldsymbol{\Gamma}_0 = \begin{bmatrix} \Gamma_{22} & \Gamma_{24} & \Gamma_{25} \\ \Gamma_{24} & \Gamma_{44} & \Gamma_{45} \\ \Gamma_{25} & \Gamma_{45} & \Gamma_{55} \end{bmatrix}.$$

For the SMDPDE method, the asymptotic variance of $\sqrt{n}\widehat{\rho}_n$ is the third diagonal element of the matrix $\boldsymbol{B}^{-1}\boldsymbol{\Gamma}_0\boldsymbol{B}^{-1'}$.

Similarly, for the asymptotic variance of the ordinary MDPDE correlation estimates, let us define, $J_{22} = \frac{1}{4\sigma_1^4(2\pi)^\beta(\sigma_1\sigma_2)^\beta(1-\rho^2)^{\frac{\beta}{2}}(1+\beta)^2}\left[\beta^2+\frac{2-\rho^2}{1-\rho^2}\right]$, $J_{44} = \frac{1}{4\sigma_2^4(2\pi)^\beta(\sigma_1\sigma_2)^\beta(1-\rho^2)^{\frac{\beta}{2}}(1+\beta)^2}\left[\beta^2+\frac{2-\rho^2}{1-\rho^2}\right]$, $J_{55} = \frac{1}{(\sigma_1\sigma_2)^\beta(2\pi)^\beta(1-\rho^2)^{2+\frac{\beta}{2}}(1+\beta)}\left[\rho^2(1+\beta)+\frac{2\rho^6-3\rho^4+1}{(1+\beta)(1-\rho^2)^2}-2\rho^2\right]$,
$J_{24} = \frac{1}{4(\sigma_1\sigma_2)^{\beta+2}(2\pi)^\beta(1-\rho^2)^{\frac{\beta}{2}}}\left[1-\frac{2}{1+\beta}+\frac{1-2\rho^2}{(1-\rho^2)(1+\beta)^2}\right]$, $J_{25} = \frac{\rho}{2\sigma_1^2(\sigma_1\sigma_2)^\beta(2\pi)^\beta(1-\rho^2)^{1+\frac{\beta}{2}}(1+\beta)}\left[1-\beta-\frac{2}{1+\beta}\right]$ and
$J_{45} = \frac{\rho}{2\sigma_2^2(\sigma_1\sigma_2)^\beta(2\pi)^\beta(1-\rho^2)^{1+\frac{\beta}{2}}(1+\beta)}\left[1-\beta-\frac{2}{1+\beta}\right]$. Let us consider the matrix

$$\boldsymbol{J} = \begin{bmatrix} J_{22} & J_{24} & J_{25} \\ J_{24} & J_{44} & J_{45} \\ J_{25} & J_{45} & J_{55} \end{bmatrix}.$$

Let us also define, $K_{22} = \frac{1}{4\sigma_1^4(2\pi)^{2\beta}(\sigma_1\sigma_2)^{2\beta}(1-\rho^2)^\beta}\left[\frac{(1+\beta)^2}{(1+2\beta)^3}\left(4\beta^2+\frac{2-\rho^2}{1-\rho^2}\right)-\left(\frac{\beta}{1+\beta}\right)^2\right]$,
$K_{44} = \frac{1}{4\sigma_2^4(2\pi)^{2\beta}(\sigma_1\sigma_2)^{2\beta}(1-\rho^2)^\beta}\left[\frac{(1+\beta)^2}{(1+2\beta)^3}\left(4\beta^2+\frac{2-\rho^2}{1-\rho^2}\right)-\left(\frac{\beta}{1+\beta}\right)^2\right]$, $K_{55} = \frac{\gamma_{55}}{(2\pi)^{2\beta}(\sigma_1\sigma_2)^{2\beta}(1-\rho^2)^{2+\beta}}$,
$K_{24} = \frac{1}{4\sigma_1^2(2\pi)^{2\beta}(\sigma_1\sigma_2)^{2\beta}(1-\rho^2)^\beta}\left[\frac{(1+\beta)^2}{(1+2\beta)^3}\left(4\beta^2-\frac{\rho^2}{1-\rho^2}\right)-\left(\frac{\beta}{1+\beta}\right)^2\right]$,
$K_{25} = \frac{\rho}{2\sigma_1^2(2\pi)^{2\beta}(\sigma_1\sigma_2)^{2\beta}(1-\rho^2)^{1+\beta}}\left[\frac{(1+\beta)^2}{(1+2\beta)^2}\left(1-2\beta-\frac{2}{1+2\beta}\right)+\left(\frac{\beta}{1+\beta}\right)^2\right]$ and
$K_{45} = \frac{\rho}{2\sigma_2^2(2\pi)^{2\beta}(\sigma_1\sigma_2)^{2\beta}(1-\rho^2)^{1+\beta}}\left[\frac{(1+\beta)^2}{(1+2\beta)^2}\left(1-2\beta-\frac{2}{1+2\beta}\right)+\left(\frac{\beta}{1+\beta}\right)^2\right]$. Let us consider the matrix

$$\boldsymbol{K} = \begin{bmatrix} K_{22} & K_{24} & K_{25} \\ K_{24} & K_{44} & K_{45} \\ K_{25} & K_{45} & K_{55} \end{bmatrix}.$$

For the ordinary MDPDE method, the asymptotic variances of $\sqrt{n}\widehat{\sigma}_{1n}^2$, $\sqrt{n}\widehat{\sigma}_{2n}^2$ and $\sqrt{n}\widehat{\rho}_n$ are the first, second and third diagonal elements of the matrix $\boldsymbol{J}^{-1}\boldsymbol{K}\boldsymbol{J}^{-1}$, re-

spectively.

4.8.6 Algebraic Details of the Influence Functions

The expressions for $A(x_1, x_2, y_1, y_2)$ and $B(x_1, x_2)$ are given by,

$$A(x_1, x_2, y_1, y_2) = -\frac{1}{2}\left[\frac{IF(\theta_{12}, y_1, g_1)}{\theta_{12}^g} + \frac{IF(\theta_{22}, y_2, g_2)}{\theta_{22}^g}\right] - \frac{1}{2(1-(\rho^g)^2)}\frac{\partial z_\epsilon}{\partial \epsilon}\Big|_{\epsilon=0} \text{ and}$$

$$B(x_1, x_2) = \frac{\rho^g}{(1-(\rho^g)^2)} - \frac{z^g\rho^g}{(1-(\rho^g)^2)^2} + \frac{1}{(1-(\rho^g)^2)}\frac{(x_1-\theta_{11}^g)(x_2-\theta_{21}^g)}{\sqrt{\theta_{12}^g\theta_{22}^g}},$$

where, $z_\epsilon = \frac{(x_1-\theta_{11\epsilon})^2}{\theta_{12\epsilon}} + \frac{(x_2-\theta_{21\epsilon})^2}{\theta_{22\epsilon}} - \frac{2\rho_\epsilon(x_1-\theta_{11\epsilon})(x_2-\theta_{21\epsilon})}{\sqrt{\theta_{12\epsilon}\theta_{12\epsilon}}}$ and $z^g = z_\epsilon|_{\epsilon=0}$. After some algebraic manipulations, it can be found that,

$$\int f_j^{1+\beta}(x, \boldsymbol{\theta}_j)(-\theta_{j2} + (x-\theta_{j1})^2)\, dx = \frac{-\beta}{(2\pi)^{\frac{\beta}{2}}(\theta_{j2})^{\frac{\beta}{2}-1}(1+\beta)^{\frac{3}{2}}},$$

$$\int f_j^{1+\beta}(x, \boldsymbol{\theta}_j)\left(\frac{(x-\theta_{j1})^2}{\theta_{j2}} - 1\right)^2 dx = \frac{\beta^2+2}{(2\pi)^{\frac{\beta}{2}}(\theta_{j2})^{\frac{\beta}{2}}(1+\beta)^{\frac{5}{2}}},$$

$$\int (1 + A(x_1, x_2, y_1, y_2)) f^\beta(x_1, x_2, \boldsymbol{\theta}, \rho) U_\rho(x_1, x_2, \boldsymbol{\theta}, \rho)\, dx_1 dx_2$$

$$= \int f^\beta(x_1, x_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \rho) U_\rho(x_1, x_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \rho)\, dx_1 dx_2$$

$$+ \int A(x_1, x_2, y_1, y_2) f^\beta(x_1, x_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \rho) U_\rho(x_1, x_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \rho)\, dx_1 dx_2, \text{ where}$$

$$\int f^\beta(x, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \rho) U_\rho(x_1, x_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \rho)\, dx_1 dx_2 = \frac{\rho\beta}{(2\pi)^\beta(\theta_{12}\theta_{22})^{\frac{\beta}{2}}(1-\rho^2)^{1+\frac{\beta}{2}}(1+\beta)^2},$$

$$\int A(x_1, x_2, y_1, y_2) f^\beta(x_1, x_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \rho) U_\rho(x_1, x_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \rho)\, dx_1 dx_2$$

$$= \frac{-\rho(1+\beta^2)}{2(2\pi)^\beta(\theta_{12}\theta_{22})^{\frac{\beta}{2}}(1-\rho^2)^{1+\frac{\beta}{2}}(1+\beta)^3}\Big[\frac{IF(\theta_{12}, y_1, g_1)}{\theta_{12}} + \frac{IF(\theta_{22}, y_2, g_2)}{\theta_{22}}\Big],$$

$$\int B(x_1, x_2) f^\beta(x_1, x_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \rho) U_\rho(x_1, x_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \rho)\, dx_1 dx_2$$

$$= \frac{1}{(2\pi)^\beta(\theta_{12}\theta_{22})^{\frac{\beta}{2}}(1-\rho^2)^{2+\frac{\beta}{2}}(1+\beta)}\Big[\rho^2 + \frac{1-3\rho^4+2\rho^6}{(1-\rho^2)^2(1+\beta)^2} - \frac{2\rho^2}{1+\beta}\Big].$$

Chapter 5

On One-step Estimation using Density Power Reweighting

5.1 Introduction

The principal objective of this thesis is to develop sophisticated robust statistical tools for estimating multivariate location and scale in case of elliptically symmetric probability models and apply them to solve various problems in the domain of pattern recognition and machine learning, e.g., clustering, classification, anomaly detection (outlier and fraud detection schemes), image processing, etcetera. The problems related to the robust estimation of multivariate location and scale have been studied for long and thus a vast literature on this topic is already available. In fact, we have developed two different methodologies for robust estimation in multivariate location-scale set-ups in this thesis so far. The IRLS based methodology (described in Chapter 2) is found to be computationally problematic in case of larger values of either the data dimension or the DPD tuning parameter β. This problem is resolved by proposing the SMDPDE in Chapter 4. However, the latter algorithm (i.e., the sequential procedure) also requires a substantial computational effort in case of larger data dimensions (although the algorithmic convergence is assured). Additionally, the positive definiteness of the covariance matrix estimates is assured only under certain assumptions.

Affine equivariance is another desirable property of location and scale estimators apart from the asymptotic and robustness properties. To understand the notion, let $\{\boldsymbol{X_1}, \ldots, \boldsymbol{X_n}\}$ be a random sample from some unknown distribution with $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ as the unknown location and scale parameters, respectively. Let, $\{\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n}\}$ be an affine transformation of the first sample, i.e., $\boldsymbol{Y}_i = \boldsymbol{A}\boldsymbol{X_i} + \boldsymbol{b}$, $1 \leq i \leq n$, for a non-singular positive definite matrix $\boldsymbol{A}$. The estimators $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ are said to be affine

equivariant if

$$\hat{\boldsymbol{\mu}}(\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n}) = \boldsymbol{A}\hat{\boldsymbol{\mu}}(\boldsymbol{X_1}, \ldots, \boldsymbol{X_n}) + \boldsymbol{b},$$
$$\hat{\boldsymbol{\Sigma}}(\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n}) = \boldsymbol{A}\hat{\boldsymbol{\Sigma}}(\boldsymbol{X_1}, \ldots, \boldsymbol{X_n})\boldsymbol{A}^{'}.$$

Computational complexity is another challenge in the practical implementations of various robust estimation procedures. Especially, this complexity becomes prominent as the dimension of data increases which we have already experienced in case of computing SMDPDEs. Thus, in the era of big data, implementing robust algorithms on multivariate datasets with large dimensions or high dimensional datasets is a big technical challenge. Either this difficulty comes from non-convergence (mainly due to singularity of matrices) or severe time complexity. So, robust algorithms with high efficiency along with smooth convergence and low computational costs are of prime importance for real life applications. Precisely, ideal robust estimators of location and scale should have (i) high breakdown values, (ii) bounded influence functions, (iii) high asymptotic efficiency, (iv) affine equivariance and (v) low computational costs associated with smooth convergence.

M-estimation (Section 1.6, Chapter 1) is one of the popular paradigms for deriving robust estimators. In particular, Maronna (1974) [108] and (1976) [109] proposed the idea of simultaneous M-estimation of location and scale in multivariate set-ups which provides location and scale estimators $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$, respectively, by solving the following system of equations, namely,

$$\frac{1}{n}\sum_{i=1}^{n} u_1(d_i)(\boldsymbol{X_i} - \boldsymbol{\mu}) = \boldsymbol{0},$$
$$\frac{1}{n}\sum_{i=1}^{n} u_2(d_i)(\boldsymbol{X_i} - \boldsymbol{\mu})(\boldsymbol{X_i} - \boldsymbol{\mu})^{'} = \boldsymbol{\Sigma},$$

where $\{\boldsymbol{X_1}, \ldots, \boldsymbol{X_n}\}$ is a random sample from an unknown distribution which is modelled by a location-scale family $\{f_{(\boldsymbol{\mu},\boldsymbol{\Sigma})} : \boldsymbol{\mu} \in \mathbb{R}^p, \boldsymbol{\Sigma} \in SPD(p)\}$ ($SPD(p)$ is the class of all $p \times p$ real-valued, symmetric, positive definite matrices) of distributions, $d_i = \sqrt{(\boldsymbol{X_i} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X_i} - \boldsymbol{\mu})}$ is the Mahalanobis distance of the i-th sample observation from the location vector $\boldsymbol{\mu}$. Under certain theoretical assumptions on the functions u_1 and u_2, these estimators possess strong theoretical properties like $\sqrt{n}$-consistency, asymptotic normality and bounded influence functions. However, one major drawback

of the scale estimate $\hat{\boldsymbol{\Sigma}}$ is that its breakdown point is at most $\frac{1}{p+1}$, p being the data dimension. Thus, $\hat{\boldsymbol{\Sigma}}$ (as well as $\hat{\boldsymbol{\mu}}$) becomes less robust (in terms of experiencing a reduced breakdown value) as the data dimension increases. Another difficulty in the implementation of this estimation procedure is its increasing computational complexity with growing data dimensions. Thus, an alternative of this simultaneous M-estimation of multivariate location-scale was keenly required. Several approaches have been proposed in order to combine M-estimators with high breakdown values and bounded influence functions. Bickel (1975) [14] (later Davies (1992) [36]) proposed a procedure in the context of linear models which initiates with a $\sqrt{n}$-consistent estimator and perform the first-step of the Gauss-Newton iteration of the corresponding M-estimating equation. In this work, it was also established that the resulting "one-step" estimator may improve the rate of convergence (as compared to that of the initial estimator) and has the same asymptotic distribution as the original M-estimator (i.e., the fully converged solution of the aforesaid M-estimating equation) under suitable assumptions in the context of linear models. It is, however, important (for the computation of both the original M-estimator and the aforesaid one-step M-estimator) to initiate the process of iteration with a highly robust estimator which is (i) easy to compute, (ii) has high breakdown point (close to $\frac{1}{2}$ is preferable) and (iii) has bounded influence function. A fair bit of research has been done on one-step M-estimators in the context of linear models by mimicking the "great" idea of Bickel (1975) [14]. For example, Rousseeuw and Leroy (2005) [131] (first published in 1987) proposed the idea of a two-step regression procedure where in the first step, least median of squares (LMS) regression is used to estimate the parameters and subsequently, another ordinary least square estimation is performed after discarding those sample observations whose corresponding residual values (found using the LMS estimates of first step) exceed some reasonable cut-off value. This procedure produces robust estimates of the regression parameters with high breakdown values, although, the rate of convergence remains the same as that of the initial estimator, i.e., LMS estimates in this case (He and Portnoy (1992) [70]). Later, Welsh and Ronchetti (2002) [156] provided a unified treatment of different types of one-step M-estimation procedures in the regression framework. They have presented a comparative study of the joint effects of different methods (Newton-Raphson, scoring and iteratively reweighted least squares) and different initial estimators on the one-step regression M-estimates.

The idea of one-step M-estimation was later imported from the regression frame-

work to multivariate framework for estimating the unknown location and scale parameters in elliptically symmetric models. In fact, Rousseeuw and Leroy (2005) [131] (first published in 1987) also proposed a similar two-step estimation procedure in the multivariate context, where in the first step the MVE estimators are used to estimate the unknown location and scale, and then the sample mean and covariance matrix of those observations are taken whose Mahalanobis distances (calculated based on the aforesaid MVE estimators) are less than a certain threshold. In general, the high breakdown point of an estimate is counterbalanced by its asymptotic efficiency in case of corrupted samples. One possible way to bypass this problem is to first identify the "good" observations in the sample. Once these good observations are recognized, classical estimators, constructed by the aforesaid good observations, may achieve high breakdown and efficiency simultaneously. This can be done by calculating the weighted sample mean and covariance matrix of the entire sample in such a way that the good observations get higher weights and the bad observations get less weights. Lopuhaä and Rousseeuw (1991) [98] provided a detailed discussion on this topic where the authors have shown that for the weighted mean and covariance estimators, the breakdown point of the initial estimators are preserved under certain assumptions. On another note, this work established a link between the idea of breakdown point and large deviations which perhaps gives an alternative interpretation of breakdown point apart from being a measure of robustness. Later, Lopuhaä (1999) [100] worked further in this connection and established the asymptotic properties of reweighted estimators of multivariate location and scale parameters. Specifically, this work established that reweighted estimators converge at the same rate as the initial estimators. Moreover, if smoothed S-estimators are used as initial estimators, the reweighted estimators will be $\sqrt{n}$-consistent and asymptotic normal and these estimators will be able to bypass the trade-off between efficiency and high breakdown property. This work has also shown the connection between the reweighted mean and covariance matrix estimators with the one-step M-estimators.

Our objective in this chapter is to study the one-step versions of the original minimum DPD estimators which constitute a new class of robust, efficient and computationally tractable estimators of location and scale in univariate and multivariate set-ups. We first study the exact one-step versions of the original minimum DPD estimators using four different iterative algorithms, namely, Newton-Raphson (NR), gradient descent (GD), iteratively reweighted least squares (IRLS) and Fisher's scoring (FS) methods

with different highly robust initializations in Section 5.2. Consistency and asymptotic normality of these estimators are established for the exact one-step Newton-Raphson, gradient descent and Fisher's scoring methods along with their influence function analyses. Detailed mathematical proofs of these results are provided in Section 5.6. These methods are illustrated with simulation experiments and two real data examples. A generalization of the exact one-step IRLS procedure is further studied in details in Section 5.3. Theoretical properties of these generalized estimators are discussed following Lopuhaä and Rousseeuw (1991) [98] and Lopuhaä (1999) [100]. This generalized procedure is illustrated with multivariate simulation experiments and an application to a classification problem related to prediction on survival of heart failure patients. Simulation experiments are described in Section 5.4 and real data examples (including the prediction on survival of heart failure patients example) are described in Section 5.5. Although our focus is on estimating location and scale in case of elliptically symmetric probability models, some other examples (including Weibull and shifter Gompertz models) are also considered in the real data applications.

5.2 One-step Minimization of the Density Power Divergence

Let, $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n\}$ be a random sample from an unknown probability distribution with PDF g having CDF G. Suppose, this unknown density g is modelled by a parametric family of densities $\mathcal{F}_{\boldsymbol{\Theta}} = \{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$. Let us recall that the minimum DPD estimator (i.e., MDPDE) of the parameter $\boldsymbol{\theta}$ is obtained by minimizing

$$\overline{D}_\beta(\boldsymbol{\theta}) = \int f_{\boldsymbol{\theta}}^{1+\beta}(\boldsymbol{x})\, d\boldsymbol{x} - \left(1 + \frac{1}{\beta}\right) \frac{1}{n} \sum_{i=1}^{n} f_{\boldsymbol{\theta}}^{\beta}(\boldsymbol{X}_i) \tag{5.1}$$

with respect to $\boldsymbol{\theta}$. We first consider the one-step minimization of $\overline{D}_\beta(\boldsymbol{\theta})$ using four different iterations in the following subsection

5.2.1 Different One-step Iterations

Suppose $\hat{\boldsymbol{\theta}}_0$ is the initial (highly robust) choice for $\boldsymbol{\theta}$ in the minimization of $\overline{D}_\beta(\boldsymbol{\theta})$. Let us also denote the gradient vector and the Hessian matrix of the objective function $\overline{D}_\beta(\boldsymbol{\theta})$ by $\nabla \overline{D}_\beta(\boldsymbol{\theta})$ and $\nabla^2 \overline{D}_\beta(\boldsymbol{\theta})$, respectively. Then, the different one-step updates are as follows:

(i) **One-Step Newton-Raphson (NR) Update:** The one-step Newton-Raphson

estimator is defined as

$$\hat{\boldsymbol{\theta}}_{NR} = \hat{\boldsymbol{\theta}}_0 - \nabla^2 \overline{D}_\beta(\hat{\boldsymbol{\theta}}_0)^{-1} \nabla \overline{D}_\beta(\hat{\boldsymbol{\theta}}_0). \tag{5.2}$$

This method is a second order method as it requires the Hessian matrix.

(ii) **One-Step Gradient Descent (GD) Update:** The one-step gradient descent estimator is defined as

$$\hat{\boldsymbol{\theta}}_{GD} = \hat{\boldsymbol{\theta}}_0 - \gamma \nabla \overline{D}_\beta(\hat{\boldsymbol{\theta}}_0), \tag{5.3}$$

where γ is the step-size. Typically, a large value of γ will lead to divergence, whereas, a small value of γ makes the convergence rate slower. So, an optimal choice of γ is required. We follow a grid-search method to choose γ appropriately.

(iii) **One-Step IRLS Update:** Suppose the gradient vector $\nabla \overline{D}_\beta(\boldsymbol{\theta})$ can be expressed as $\nabla \overline{D}_\beta(\boldsymbol{\theta}) = \sum_{i=1}^{n} w(\boldsymbol{X}_i, \boldsymbol{\theta})(\boldsymbol{X}_i - \boldsymbol{\theta})$, where $w(\boldsymbol{X}_i, \boldsymbol{\theta})$, $i = 1, \ldots, n$, are data driven weights. In that case, the estimating equation $\nabla \overline{D}_\beta(\boldsymbol{\theta}) = \mathbf{0}$ becomes $\sum_{i=1}^{n} w(\boldsymbol{X}_i, \boldsymbol{\theta})(\boldsymbol{X}_i - \boldsymbol{\theta}) = \mathbf{0}$. If we estimate the weights $w(\boldsymbol{X}_i, \boldsymbol{\theta})$ by $w(\boldsymbol{X}_i, \hat{\boldsymbol{\theta}}_0)$ and replace the weights with these estimated weights in the estimating equation, then we can solve for $\boldsymbol{\theta}$ from the modified estimating equation (with estimated weights) in closed form as:

$$\hat{\boldsymbol{\theta}}_{IRLS} = \frac{\sum_{i=1}^{n} w(\boldsymbol{X}_i, \hat{\boldsymbol{\theta}}_0) \boldsymbol{X}_i}{\sum_{i=1}^{n} w(\boldsymbol{X}_i, \hat{\boldsymbol{\theta}}_0)}. \tag{5.4}$$

We call $\hat{\boldsymbol{\theta}}_{IRLS}$ as the one-step IRLS estimator of $\boldsymbol{\theta}$. We will see later that the one-step IRLS location estimators have the aforesaid algebraic form. However, the one-step IRLS scale estimators will have a slightly different form. It will be illustrated through the normal and Cauchy models. Thus, the one-step IRLS estimators do not follow any explicit algebraic form in general unlike the other one-step estimators.

(iv) **One-Step Fisher's scoring (FS) Update:** The one-step Fisher's scoring es-

timator is defined as

$$\hat{\boldsymbol{\theta}}_{FS} = \hat{\boldsymbol{\theta}}_0 - \left(E_{f_{\hat{\boldsymbol{\theta}}_0}} \nabla^2 \overline{D}_\beta(\hat{\boldsymbol{\theta}}_0)\right)^{-1} \nabla \overline{D}_\beta(\hat{\boldsymbol{\theta}}_0), \tag{5.5}$$

where the usual Hessian (as in the Newton-Raphson method) is replaced by its expectation with respect to $f_{\hat{\boldsymbol{\theta}}_0}$, the model density corresponding to $\hat{\boldsymbol{\theta}}_0$.

5.2.2 Asymptotic Properties

The one-step Newton-Raphson, gradient descent and Fisher's scoring updates admit specific algebraic representations unlike the one-step IRLS update. These explicit forms help us to derive the asymptotic properties of the aforesaid one-step estimators. We now present the asymptotic results (consistency and asymptotic normality) related to the one-step Newton-Raphson, gradient descent and Fisher's scoring estimators. A generalized form of the one-step IRLS update will be studied in Section 5.3 following Lopuhaä (1999) [100].

Let us introduce the following representation of

$$\overline{D}_\beta(\boldsymbol{\theta}) = \frac{1}{n} \sum_{k=1}^{n} V(\boldsymbol{X}_k, \boldsymbol{\theta}), \tag{5.6}$$

where $V(\boldsymbol{x}, \boldsymbol{\theta}) = \int f_{\boldsymbol{\theta}}^{1+\beta}(\boldsymbol{u})\, d\boldsymbol{u} - \left(1 + \frac{1}{\beta}\right) f_{\boldsymbol{\theta}}^{\beta}(\boldsymbol{x})$. To establish consistency of the one-step estimators, we need the following technical assumptions. In addition, Theorem 5.1 is essential for establishing the consistency of the one-step estimators. For simplicity, the initial estimator is denoted by $\hat{\boldsymbol{\theta}}_0$, although it depends on the sample size n. Let us also assume that $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta} \subseteq \mathbb{R}^m$ is the true (unknown) value of the parameter $\boldsymbol{\theta}$ and the true unknown PDF $g \in \mathcal{F}_{\boldsymbol{\theta}}$, so that, $g = f_{\boldsymbol{\theta}_0}$, in particular. Detailed proofs of all of Theorems 5.1, 5.2 and 5.3 are provided in Section 5.6.

Assumption 5.1. *The integral* $\int \left| f_{\boldsymbol{\theta}}^{\beta-1}(\boldsymbol{x}) \frac{\partial}{\partial\, \theta_j} f_{\boldsymbol{\theta}}(\boldsymbol{x}) \frac{\partial}{\partial\, \theta_i} f_{\boldsymbol{\theta}}(\boldsymbol{x}) \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} d\boldsymbol{x} < \infty$, *for all* $1 \leq i, j \leq m$.

Assumption 5.2. *There is an open subset* γ_0 *of the parameter space* $\boldsymbol{\Theta}$ *such that for almost all* $\boldsymbol{x}$ *and all* $\boldsymbol{\theta} \in \gamma_0$, *the model density* $f_{\boldsymbol{\theta}}$ *is three times differentiable with respect to* $\boldsymbol{\theta}$ *and the third partial derivatives are continuous with respect to* $\boldsymbol{\theta}$.

Assumption 5.3. *The integrals* $\int f_{\boldsymbol{\theta}}^{1+\beta}(x)\, dx$ *and* $\int f_{\boldsymbol{\theta}}^{\beta}(x) g(x)\, dx$ *are differentiable three times with respect to* $\boldsymbol{\theta}$ *and the derivatives can be taken under the integral signs.*

Assumption 5.4. *There exist real-valued functions $U_{j'ji}$, such that,*

$$\Big|\frac{\partial^3}{\partial\,\theta_{j'}\partial\,\theta_j\partial\,\theta_i}V(\boldsymbol{x},\boldsymbol{\theta})\Big| \leq U_{j'ji}(\boldsymbol{x}),$$

$\forall\ \boldsymbol{x} \in \mathbb{R}^p$ *and* $\boldsymbol{\theta} \in B_\epsilon(\boldsymbol{\theta}_0)$ *for some* $\epsilon > 0$ *with* $E_{f_{\boldsymbol{\theta}_0}}(U_{j'ji}(\boldsymbol{X})) < \infty\ \forall\ j', j, i.$

Some remarks on these assumptions are provided in Section 6.

Theorem 5.1. *Under the aforesaid assumptions and the consistency of* $\hat{\boldsymbol{\theta}}_0$ *(to* $\boldsymbol{\theta}_0$*),*

$$\nabla\overline{D}_\beta(\hat{\boldsymbol{\theta}}_0) \xrightarrow{p} \mathbf{0},\ \textit{as}\ n \to \infty. \tag{5.7}$$

The consistency of the one-step Newton Raphson, gradient-descent and Fisher's scoring estimators can now be established as a consequence of Theorem 5.1.

Theorem 5.2. *Under the consistency of the initial estimator* $\hat{\boldsymbol{\theta}}_0$ *and Assumptions 5.1-5.4,*

$$\hat{\boldsymbol{\theta}}_{NR} \xrightarrow{p} \boldsymbol{\theta}_0,\ \hat{\boldsymbol{\theta}}_{GD} \xrightarrow{p} \boldsymbol{\theta}_0,\ \textit{and}\ \hat{\boldsymbol{\theta}}_{FS} \xrightarrow{p} \boldsymbol{\theta}_0 \quad \textit{as}\ n \to \infty,$$

where $\hat{\boldsymbol{\theta}}_{NR}$, $\hat{\boldsymbol{\theta}}_{GD}$ *and* $\hat{\boldsymbol{\theta}}_{FS}$ *are the one-step updates corresponding to Equations (5.2), (5.3) and (5.5), respectively.*

To establish asymptotic normality of $\hat{\boldsymbol{\theta}}_{NR}$, $\hat{\boldsymbol{\theta}}_{GD}$ and $\hat{\boldsymbol{\theta}}_{FS}$, we need one additional assumption on the initial estimator $\hat{\boldsymbol{\theta}}_0$.

Assumption 5.5. *The initial estimator* $\hat{\boldsymbol{\theta}}_0$ *admits the following asymptotic expansion:*

$$\hat{\theta}_{0j} - \theta_{0j} = \frac{1}{n}\sum_{k=1}^{n} Z_j(\boldsymbol{X}_k, \boldsymbol{\theta}_0) + o_p(n^{-\frac{1}{2}}),\ 1 \leq j \leq m, \tag{5.8}$$

where $E_{f_{\boldsymbol{\theta}_0}}(Z_j(\boldsymbol{X}_k, \boldsymbol{\theta}_0)) = 0$ *and* $V_{f_{\boldsymbol{\theta}_0}}(Z_j(\boldsymbol{X}_k, \boldsymbol{\theta}_0)) < \infty$*, for all* $j = 1, \ldots, m$*.*

Remark 5.1. *The algebraic form presented in Equation (5.8) is required to prove the asymptotic normality of* $\sqrt{n}(\hat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}_0) = \sqrt{n}\frac{1}{n}\sum_{k=1}^{n} Z_j(\boldsymbol{X}_k, \boldsymbol{\theta}_0) + o_p(1)$*, where the first term converges weakly to a normal distribution (by the central limit theorem) and the second term converges to* 0 *in probability. Thus, the asymptotic normality of* $\sqrt{n}(\hat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}_0)$ *follows from Slutsky's Theorem.*

Let us now present the asymptotic normality of $\hat{\boldsymbol{\theta}}_{NR}$, $\hat{\boldsymbol{\theta}}_{GD}$ and $\hat{\boldsymbol{\theta}}_{FS}$ in the following Theorem.

Theorem 5.3 (Asymptotic Normality)**.** *If Assumptions 5.1-5.5 hold and the initial estimator $\hat{\boldsymbol{\theta}}_0$ is consistent (to $\boldsymbol{\theta}_0$), then*

$$\begin{aligned}
\sqrt{n}(\hat{\boldsymbol{\theta}}_{NR} - \boldsymbol{\theta}_0) &\xrightarrow{d} N_m(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta}_0)),\\
\sqrt{n}(\hat{\boldsymbol{\theta}}_{GD} - \boldsymbol{\theta}_0) &\xrightarrow{d} N_m(\mathbf{0}, \boldsymbol{\Sigma}_{GD}(\boldsymbol{\theta}_0)),\\
\sqrt{n}(\hat{\boldsymbol{\theta}}_{FS} - \boldsymbol{\theta}_0) &\xrightarrow{d} N_m(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta}_0)),
\end{aligned}$$

where $\boldsymbol{\Sigma}(\boldsymbol{\theta}_0) = \boldsymbol{F}^{-1}\boldsymbol{G}(\boldsymbol{\theta}_0)\boldsymbol{F}^{-1}$, $\boldsymbol{\Sigma}_{GD}(\boldsymbol{\theta}_0) = Var(\boldsymbol{H}(\boldsymbol{X}_1, \boldsymbol{\theta}_0))$, $\boldsymbol{F} = [[f_{ij}]]_{i,j=1}^m$ *with* $f_{ij} = (1+\beta)\int f_{\boldsymbol{\theta}_0}^{\beta-1}(\boldsymbol{x}) \frac{\partial}{\partial \theta_j} f_{\boldsymbol{\theta}}(\boldsymbol{x})\Big|_{\boldsymbol{\theta}_0} \frac{\partial}{\partial \theta_i} f_{\boldsymbol{\theta}}(\boldsymbol{x})\Big|_{\boldsymbol{\theta}_0} d\boldsymbol{x}$, $\boldsymbol{G}(\boldsymbol{\theta}_0) = Var(\nabla V(\boldsymbol{X}_1, \boldsymbol{\theta}_0))$ *and* $\boldsymbol{H}(\boldsymbol{X}_k, \boldsymbol{\theta}_0) = (\boldsymbol{I}_m - \gamma \boldsymbol{F})\boldsymbol{Z}(\boldsymbol{X}_k, \boldsymbol{\theta}_0) - \gamma \nabla V(\boldsymbol{X}_k, \boldsymbol{\theta}_0)$, $\boldsymbol{Z}(\boldsymbol{X}_k, \boldsymbol{\theta}_0) = (Z_1(\boldsymbol{X}_k, \boldsymbol{\theta}_0), \ldots, Z_m(\boldsymbol{X}_k, \boldsymbol{\theta}_0))'$. *The matix $\boldsymbol{F}$ is assumed to be positive definite.*

Remark 5.2. *Theorems 5.1 and 5.3 are specially significant in the literature of minimum DPD based inference. In particular, Theorem 5.1 ensures that any consistent estimator $\hat{\boldsymbol{\theta}}_0$ (for $\boldsymbol{\theta}_0$) will asymptotically solve the minimum DPD estimating equation $\nabla \overline{D}_\beta(\boldsymbol{\theta}) = 0$ if the true unknown distribution belongs to the model family under the stated assumptions. Moreover, Theorem 5.3 ensures that the one-step Newton-Raphson, Fisher's scoring and the original minimum DPD estimators are asymptotically equivalent and the asymptotic distribution does not depend on the initial estimator. But, the same conclusion cannot be drawn for the one-step gradient descent estimator.*

5.2.3 Influence Function Analyses

To understand the robustness of the one-step estimators (except the one-step IRLS one), we hereby present the boundedness of the influence functions of the functionals corresponding to the aforesaid one-step estimators. Firstly, we derive the influence function of the one-step gradient descent estimator. To do that, let us observe that,

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_{GD} &= \hat{\boldsymbol{\theta}}_0 - \gamma \nabla \overline{D}_\beta(\hat{\boldsymbol{\theta}}_0)\\
&= \hat{\boldsymbol{\theta}}_0 - \frac{\gamma}{n}\sum_{i=1}^n \nabla V(\boldsymbol{X}_i, \hat{\boldsymbol{\theta}}_0).
\end{aligned}$$

Let us define the one-step gradient descent functional as,

$$\boldsymbol{\theta}_{GD}(G) = \boldsymbol{\theta}_I(G) - \gamma \int \nabla V(\boldsymbol{x}, \boldsymbol{\theta}_I(G)) g(\boldsymbol{x}) d\boldsymbol{x}, \tag{5.9}$$

where G is the true CDF (and g is the true unknown PDF) and $\boldsymbol{\theta}_I(G)$ is the functional corresponding to the initial estimator $\hat{\boldsymbol{\theta}}_0$. For ease of understanding, we consider the j-th component of System (5.9), i.e.,

$$\theta_{j_{GD}}(G) = \theta_{j_I}(G) - \gamma \int \nabla_j V(\boldsymbol{x}, \boldsymbol{\theta}_I(G)) g(\boldsymbol{x}) d\boldsymbol{x}. \tag{5.10}$$

In the above, the interchange of the integral and differential may be justified by Assumption 5.3. To derive the influence function of $\boldsymbol{\theta}_{GD}(G)$, we need to consider the contaminated distribution with CDF $G_\epsilon = (1-\epsilon)G + \epsilon \wedge_{\boldsymbol{y}}$, where $\wedge_{\boldsymbol{y}}$ is the CDF of the degenerate distribution at the point mass $\boldsymbol{y}$, respectively. We assume that the derivatives of the following functionals (at the contaminated distribution with CDF G_ϵ) with respect to ϵ can be taken under the integral signs. The influence function of $\theta_{j_{GD}}(G)$ (from Equation (5.10)) is defined as

$$\begin{aligned}
&IF(\theta_{j_{GD}}, G, \boldsymbol{y}) \\
&= \frac{\partial}{\partial \epsilon} \theta_{j_{GD}}(G_\epsilon)\Big|_{\epsilon=0} \\
&= \frac{\partial}{\partial \epsilon} \theta_{jI}(G_\epsilon)\Big|_{\epsilon=0} - \gamma \frac{\partial}{\partial \epsilon} \int \nabla_j V(\boldsymbol{x}, \boldsymbol{\theta}_I(G_\epsilon)) dG_\epsilon(\boldsymbol{x})\Big|_{\epsilon=0} \\
&= IF(\theta_{jI}, G, \boldsymbol{y}) - \gamma \frac{\partial}{\partial \epsilon} \int \nabla_j V(\boldsymbol{x}, \boldsymbol{\theta}_I(G_\epsilon))(1-\epsilon) g(\boldsymbol{x}) d\boldsymbol{x}\Big|_{\epsilon=0} - \gamma \frac{\partial}{\partial \epsilon} \big[\epsilon \nabla_j V(\boldsymbol{y}, \boldsymbol{\theta}_I(G_\epsilon))\big]\Big|_{\epsilon=0} \\
&= IF(\theta_{jI}, G, \boldsymbol{y}) - \gamma \int \frac{\partial}{\partial \epsilon} \nabla_j V(\boldsymbol{x}, \boldsymbol{\theta}_I(G_\epsilon))\Big|_{\epsilon=0} g(\boldsymbol{x}) d\boldsymbol{x} + \gamma \int \nabla_j V(\boldsymbol{x}, \boldsymbol{\theta}_I(G)) g(\boldsymbol{x}) d\boldsymbol{x} - \gamma \nabla_j V(\boldsymbol{y}, \boldsymbol{\theta}_I(G)) \\
&= IF(\theta_{jI}, G, \boldsymbol{y}) - \gamma \int \Big\langle \frac{\partial}{\partial \boldsymbol{\theta}} \nabla_j V(\boldsymbol{x}, \boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_I(G)}, IF(\boldsymbol{\theta_I}, G, \boldsymbol{y}) \Big\rangle g(\boldsymbol{x}) d\boldsymbol{x} \\
&+ \gamma \int \nabla_j V(\boldsymbol{x}, \boldsymbol{\theta}_I(G)) g(\boldsymbol{x}) d\boldsymbol{x} - \gamma \nabla_j V(\boldsymbol{y}, \boldsymbol{\theta}_I(G)) \text{ (applying the chain rule)} \\
&= IF(\theta_{jI}, G, \boldsymbol{y}) - \gamma \int \Big\langle \frac{\partial}{\partial \boldsymbol{\theta}} \nabla_j V(\boldsymbol{x}, \boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_I(G)}, IF(\boldsymbol{\theta_I}, G, \boldsymbol{y}) \Big\rangle g(\boldsymbol{x}) d\boldsymbol{x} - \gamma \nabla_j V(\boldsymbol{y}, \boldsymbol{\theta}_I(G)) \\
&+ \gamma \int \nabla_j V(\boldsymbol{x}, \boldsymbol{\theta}_I(G)) g(\boldsymbol{x}) d\boldsymbol{x},
\end{aligned}$$

which is a linear combination of the influence function vector $IF(\boldsymbol{\theta_I}, G, \boldsymbol{y})$ of the functional $\boldsymbol{\theta}_I(G)$ and the function $\nabla V(\boldsymbol{y}, \boldsymbol{\theta}_I(G))$. Assuming that the initial estimators

have bounded influence functions, the one-step gradient descent functional $\boldsymbol{\theta}_{GD}(G)$ has bounded influence function vector if $\nabla V(\boldsymbol{y}, \boldsymbol{\theta}_I(G))$ is bounded as a function of the contamination $\boldsymbol{y}$.

Now, let us derive the influence functions of the functionals corresponding to the one-step Newton-Raphson and Fisher's scoring estimators. Specifically, the one-step Newton-Raphson estimator is given by

$$\hat{\boldsymbol{\theta}}_{NR} = \hat{\boldsymbol{\theta}}_0 - \nabla^2 \overline{D}_\beta(\hat{\boldsymbol{\theta}}_0)^{-1} \nabla \overline{D}_\beta(\hat{\boldsymbol{\theta}}_0) = \hat{\boldsymbol{\theta}}_0 - \hat{\boldsymbol{\theta}}_H, \tag{5.11}$$

where $\hat{\boldsymbol{\theta}}_H$ satisfies

$$[\nabla^2 \overline{D}_\beta(\hat{\boldsymbol{\theta}}_0)]\hat{\boldsymbol{\theta}}_H = \nabla \overline{D}_\beta(\hat{\boldsymbol{\theta}}_0). \tag{5.12}$$

Let, $\boldsymbol{\theta}_I(G)$, $\boldsymbol{\theta}_{NR}(G)$ and $\boldsymbol{\theta}_H(G)$ are the functionals corresponding to $\hat{\boldsymbol{\theta}}_0$, $\hat{\boldsymbol{\theta}}_{NR}$ and $\hat{\boldsymbol{\theta}}_H$, respectively. Then, Equation (5.11) implies that the aforesaid functionals should satisfy

$$\boldsymbol{\theta}_{NR}(G) = \boldsymbol{\theta}_I(G) - \boldsymbol{\theta}_H(G), \tag{5.13}$$

so that, the influence functions of the functionals in Equation (5.13) satisfy,

$$IF(\boldsymbol{\theta}_{NR}, G, \boldsymbol{y}) = IF(\boldsymbol{\theta}_I, G, \boldsymbol{y}) - IF(\boldsymbol{\theta}_H, G, \boldsymbol{y}). \tag{5.14}$$

So, it is enough to determine the algebraic form of $IF(\boldsymbol{\theta}_H, G, \boldsymbol{y})$ to understand the behaviour of $IF(\boldsymbol{\theta}_{NR}, G, \boldsymbol{y})$ as a function of the contamination $\boldsymbol{y}$. To do that, let us first observe that,

$$\nabla \overline{D}_\beta(\hat{\boldsymbol{\theta}}_0) = \frac{1}{n}\sum_{i=1}^{n} \nabla V(\boldsymbol{X}_i, \hat{\boldsymbol{\theta}}_0), \text{ and } \nabla^2 \overline{D}_\beta(\hat{\boldsymbol{\theta}}_0) = \frac{1}{n}\sum_{i=1}^{n} \nabla^2 V(\boldsymbol{X}_i, \hat{\boldsymbol{\theta}}_0).$$

Thus, from Equation (5.12), the functional $\boldsymbol{\theta}_H(G)$ satisfies

$$\left[\int \nabla^2 V(\boldsymbol{x}, \boldsymbol{\theta}_I(G)) dG(\boldsymbol{x})\right] \boldsymbol{\theta}_H(G) = \int \nabla V(\boldsymbol{x}, \boldsymbol{\theta}_I(G)) dG(\boldsymbol{x}). \tag{5.15}$$

For ease of understanding, we consider the j-th component of System (5.15)

$$\sum_{k=1}^{m} \theta_{kH}(G) \int \nabla^2_{jk} V(\boldsymbol{x}, \boldsymbol{\theta}_I(G)) dG(\boldsymbol{x}) = \int \nabla_j V(\boldsymbol{x}, \boldsymbol{\theta}_I(G)) dG(\boldsymbol{x}). \tag{5.16}$$

Now, to find the influence function of $\boldsymbol{\theta}_H(G)$ (i.e., $IF(\boldsymbol{\theta}_H, G, \boldsymbol{y})$), we put $G = G_\epsilon$ in Equation (5.15) and evaluate the derivative of both sides with respect to ϵ at $\epsilon = 0$ which imply

$$\frac{\partial}{\partial \epsilon} \left[\sum_{k=1}^{m} \theta_{kH}(G_\epsilon) \int \nabla^2_{jk} V(\boldsymbol{x}, \boldsymbol{\theta}_I(G_\epsilon)) dG_\epsilon(\boldsymbol{x}) \right] \Big|_{\epsilon=0} = \frac{\partial}{\partial \epsilon} \int \nabla_j V(\boldsymbol{x}, \boldsymbol{\theta}_I(G_\epsilon)) dG_\epsilon(\boldsymbol{x}) \Big|_{\epsilon=0}.$$

After some algebra, it can be shown that

$$\begin{aligned}
&\sum_{k=1}^{m} IF(\theta_{kH}, G, \boldsymbol{y}) \int \nabla^2_{jk} V(\boldsymbol{x}, \boldsymbol{\theta}_I(G)) g(\boldsymbol{x}) d\boldsymbol{x} \\
&+ \sum_{k=1}^{m} \theta_{kH}(G) \int \Big\langle \frac{\partial}{\partial \boldsymbol{\theta}} \nabla^2_{jk} V(\boldsymbol{x}, \boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_I(G)}, IF(\boldsymbol{\theta_I}, G, \boldsymbol{y}) \Big\rangle g(\boldsymbol{x}) d\boldsymbol{x} \\
&+ \sum_{k=1}^{m} \theta_{kH}(G) \int \nabla^2_{jk} V(\boldsymbol{x}, \boldsymbol{\theta}_I(G)) (\Delta_{\boldsymbol{y}} - g(\boldsymbol{x})) d\boldsymbol{x} \\
&= \int \Big\langle \frac{\partial}{\partial \boldsymbol{\theta}} \nabla_j V(\boldsymbol{x}, \boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_I(G)}, IF(\boldsymbol{\theta_I}, G, \boldsymbol{y}) \Big\rangle g(\boldsymbol{x}) d\boldsymbol{x} \\
&+ \int \nabla_j V(\boldsymbol{x}, \boldsymbol{\theta}_I(G)) (\Delta_{\boldsymbol{y}} - g(\boldsymbol{x})) d\boldsymbol{x}, \ \forall \ j = 1, \ldots, m.
\end{aligned}$$

The aforesaid form essentially depicts the fact that $IF(\boldsymbol{\theta}_H, G, \boldsymbol{y})$ can be obtained by solving a linear system

$$\boldsymbol{A}(\boldsymbol{\theta_I(G)}) IF(\boldsymbol{\theta}_H, G, \boldsymbol{y}) = \boldsymbol{b}(\boldsymbol{\theta}_I(G), \boldsymbol{\theta}_H(G), \boldsymbol{y}),$$

where the vector in the right hand side is a linear combination of $IF(\boldsymbol{\theta}_I, G, \boldsymbol{y})$, $\nabla_{jk} V(\boldsymbol{y}, \boldsymbol{\theta}_I(G))$ and $\nabla_j V(\boldsymbol{y}, \boldsymbol{\theta}_I(G))$, $\forall j, k$. This fact establishes the boundedness of $IF(\boldsymbol{\theta}_H, G, \boldsymbol{y})$ (and hence the boundedness of $IF(\boldsymbol{\theta}_{NR}, G, \boldsymbol{y})$) assuming the boundedness of $IF(\boldsymbol{\theta}_I, G, \boldsymbol{y})$ and the functions $\nabla^2 V(\boldsymbol{y}, \boldsymbol{\theta}_I(G))$ and $\nabla V(\boldsymbol{y}, \boldsymbol{\theta}_I(G))$ in $\boldsymbol{y}$. The boundedness of $IF(\boldsymbol{\theta}_{FS}, G, \boldsymbol{y})$, the influence function of the one-step Fisher's scoring functional, can be similarly established as in case of the one-step Newton-Raphson functional $\boldsymbol{\theta}_{NR}(G)$.

5.2.4 Example: The Normal Model Family

Let us now illustrate the aforesaid theoretical results through the univariate normal models. In particular, we assume the univariate standard normal model with $\mu_0 = 0$ and $\sigma_0 = 1$.

(i) Influence Functions: The exact algebraic forms of the influence functions of the one-step Newton-Raphson, gradient descent and Fisher's scoring funtionals have already been derived in Section 5.2.3. These influence functions are shown to be bounded under certain conditions. In particular, the influence functions of the aforesaid functionals (with S-initialization) along with that of the maximum likelihood functional in case of the univariate standard normal model are presented in Figure 5.1; the boundedness of these curves can be trivially observed.

To bypass the complex calculations of the influence functions (even in case of univariate normal models which require the influence functions of the initial functionals), we evaluate the influence functions numerically and plot them in Figure 5.1. To do that, we first generate a random sample (say $\boldsymbol{X}_0$) of size 999 from the standard normal distribution (the true distribution in this case) and let $\hat{\theta}$ be the one-step estimator (Newton-Raphson, gradient descent or Fisher's scoring with S-initialization) of the unknown parameter θ (either μ or σ^2 in the univariate normal scenario) based on $\boldsymbol{X}_0$. In the next step, $\boldsymbol{X}_0$ is contaminated with the single point mass y (resulting in 1000 observations in the dataset) and let this contaminated sample be denoted by $\boldsymbol{X}$. Suppose, $\hat{\theta}(y)$ is the one-step estimator (Newton-Raphson, gradient descent or Fisher's scoring with S-initialization) of θ based on the contaminated sample $\boldsymbol{X}$. We do it for $y \in \{-20, -19.5, \ldots, 19.5, 20\}$, separately and the influence function $IF(\theta(G), G, y)$ of the functional $\theta(G)$ (functional corresponding to the parameter θ) at the point of contamination y should be numerically evaluated by $IF_0(y) = \frac{\hat{\theta}(y) - \hat{\theta}}{0.001}$ (since the contaminating proportion is 0.001). But, we must make a small adjustment to $IF_0(y)$ following the fact that influence functions should have zero mean (Basu et al. (2011) [12]). In particular, we center the term $IF_0(y)$ with its empirical mean $\overline{IF} = \frac{1}{81} \sum_{y \in \{-20, -19.5, \ldots, 19.5, 20\}} IF_0(y)$ and obtain the influence function $IF(\theta(G), G, y)$

of the functional $\theta(G)$ numerically by

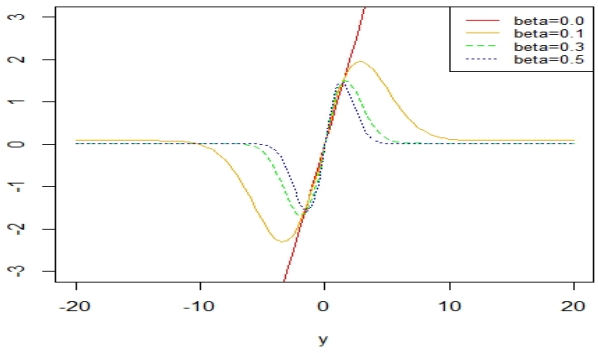
$$\begin{aligned} IF(\theta(G), G, y) &= IF_0(y) - \overline{IF} \\ &= \frac{\hat{\theta}(y) - \hat{\theta}}{0.001} - \overline{IF}. \end{aligned}$$

(ii) Asymptotic Relative Efficiency: The asymptotic normality of the one-step Newton-Raphson, gradient descent and Fisher's scoring estimators have been established in Section 5.2.2 under certain technical assumptions. We now present the asymptotic relative efficiencies of the aforesaid estimators in case of the univariate standard normal model.

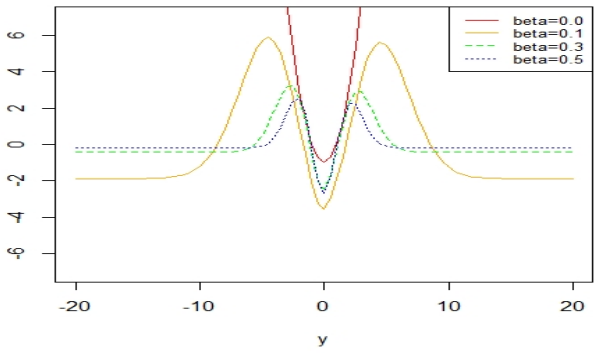
Parameter	Estimator	$\beta = 0.1$	$\beta = 0.3$	$\beta = 0.5$
$\sqrt{n}\hat{\mu}_n$	NR	98.717	92.081	83.822
	GD	90.992	73.746	50.942
	FS	98.717	92.081	83.822
$\sqrt{n}\hat{\sigma}_n$	NR	97.466	85.470	73.099
	GD	75.188	62.422	58.617
	FS	97.466	85.470	73.099

Table 5.1: Asymptotic relative efficiencies (in percentage) of the mean and standard deviation estimators (properly scaled) at the $N(0, 1)$ model.

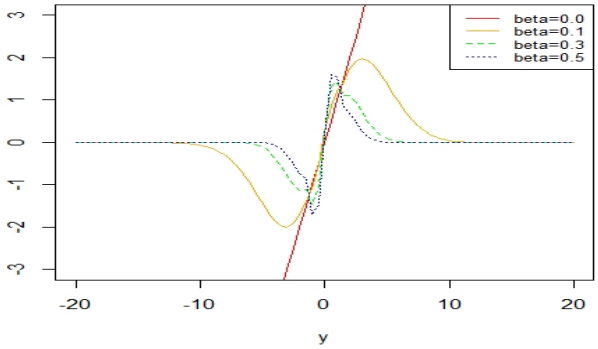
As we have seen in Theorem 5.3, the one-step Newton-Raphson and Fisher's scoring estimators are asymptotically normal with the asymptotic covariance matrix same as that of the corresponding (fully converged) MDPDE under certain assumptions. The expressions for these asymptotic variances in case of univariate normal models are provided in Basu et al. (2011) [12]. However, the asymptotic variances of the one-step gradient descent estimators are difficult to obtain in closed forms. Thus we have obtained these numerically based on a large sample. The asymptotic relative efficiencies of the one-step Newton-Raphson, Fisher's scoring estimators and the empirical asymptotic relative efficiencies of the one-step gradient descent estimators are presented in Table 5.1. It can be trivially understood that the one-step Newton-Raphson, Fisher's scoring and the fully converged minimum DPD estimators are much more efficient than the one-step gradient descent estimators in terms of these asymptotic relative efficiencies.



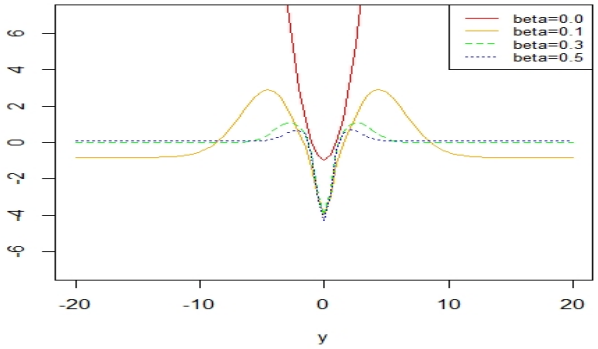
(a) One-step NR mean



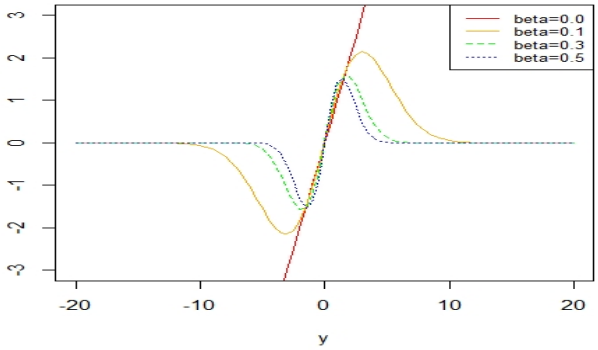
(b) One-step NR variance



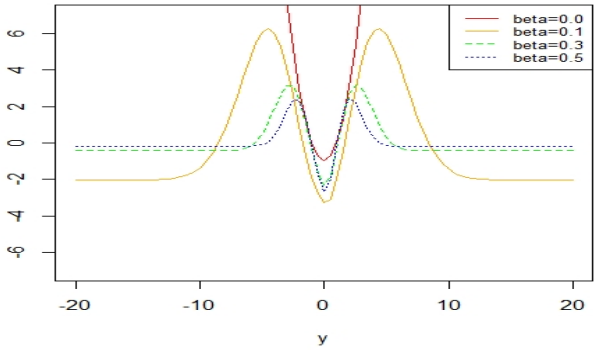
(c) One-step GD mean



(d) One-step GD variance



(e) One-step FS mean



(f) One-step FS variance

Figure 5.1: Influence curves of different one-step functionals and the minimum DPD functionals in case of the standard normal model.

5.2.5 One-step Estimators: Specific Examples

Now, let us present the aforesaid one-step updates for some of the well-known probability models explicitly.

(i) Univariate Normal Model: For the $N(\mu, \sigma^2)$ model (with model density $f(\cdot, \mu, \sigma^2)$), we have the parameter vector $\boldsymbol{\theta} = (\mu, \sigma^2)$. For the DPD objective function $\overline{D}_\beta(\boldsymbol{\theta}) = \overline{D}_\beta(\mu, \sigma^2)$, the gradient vector is $\nabla \overline{D}_\beta(\boldsymbol{\theta}) = \left(\frac{\partial \overline{D}_\beta(\mu,\sigma^2)}{\partial \mu}, \frac{\partial \overline{D}_\beta(\mu,\sigma^2)}{\partial \sigma^2}\right)'$ and the Hessian matrix is $\nabla^2 \overline{D}_\beta(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial^2 \overline{D}_\beta(\mu,\sigma^2)}{\partial^2 \mu} & \frac{\partial^2 \overline{D}_\beta(\mu,\sigma^2)}{\partial \mu \partial \sigma^2} \\ \frac{\partial^2 \overline{D}_\beta(\mu,\sigma^2)}{\partial \mu \partial \sigma^2} & \frac{\partial^2 \overline{D}_\beta(\mu,\sigma^2)}{\partial^2 \sigma^2} \end{bmatrix}$ Detailed expressions of the gradient vector and Hessian matrix are provided in Section 5.6 and these are essential to implement the one-step Newton-Raphson update. For the one-step Fisher's scoring update, we need to compute the expected Hessian matrix with respect to the model density $f(\cdot, \mu, \sigma^2)$ at $\mu = \hat{\mu}_0$, $\sigma = \hat{\sigma}_0$. The expected Hessian matrix is given by $\begin{bmatrix} \frac{1}{\sigma^{\beta+2}(2\pi)^{\frac{\beta}{2}}\sqrt{1+\beta}} & 0 \\ 0 & \frac{\beta^2+2}{4\sigma^{\beta+4}(2\pi)^{\frac{\beta}{2}}(1+\beta)^{\frac{3}{2}}} \end{bmatrix}$ at $\mu = \hat{\mu}_0$ and $\sigma = \hat{\sigma}_0$; the off-diagonal terms become 0 as the integrands are odd functions of $(x - \mu)$.

However, the one-step IRLS update was derived in multivariate normal context in Equations (2.9) and (2.10) of Chapter 2 (originally derived in Chakraborty et al. (2023) [22]). In case of univariate normal (i.e., data dimension 1), the one-step IRLS updates of μ and σ^2 are as follows:

$$\hat{\mu} = \frac{\sum_{i=1}^n e^{-\frac{\beta}{2}\left(\frac{X_i - \hat{\mu}_0}{\hat{\sigma}_0}\right)^2} X_i}{\sum_{i=1}^n e^{-\frac{\beta}{2}\left(\frac{X_i - \hat{\mu}_0}{\hat{\sigma}_0}\right)^2}}, \; \hat{\sigma}^2 = \frac{\sum_{i=1}^n e^{-\frac{\beta}{2}\left(\frac{X_i - \hat{\mu}_0}{\hat{\sigma}_0}\right)^2} (X_i - \hat{\mu})^2}{\sum_{i=1}^n e^{-\frac{\beta}{2}\left(\frac{X_i - \hat{\mu}_0}{\hat{\sigma}_0}\right)^2} - \frac{n\beta}{(1+\beta)^{\frac{3}{2}}}}. \tag{5.17}$$

It is easy to observe that, the one-step IRLS updates of the mean and covariance matrix estimates have a physical interpretation of being weighted mean and covariance matrix of the sample where the weights (which decrease as the observations become more extreme) are calculated using the initial estimates. Motivated by this observation, we will propose a new class of generalized weight based one-step estimators for location and scale in Section 5.3.

(ii) Cauchy Model: Now, let us assume the model to be Cauchy(μ, σ) (i.e., $\boldsymbol{\theta} = (\mu, \sigma)'$) distribution, i.e., the Cauchy distribution with location parameter μ and scale

parameter σ with PDF $f(x,\mu,\sigma) = \frac{1}{\pi\sigma\left[1+\left(\frac{x-\mu}{\sigma}\right)^2\right]},\ x \in (-\infty,\infty)$. Unlike the normal model, it is not possible to derive a closed form expression of the first integral term of the DPD, i.e., $\int f^{1+\beta}(x,\mu,\sigma)\,dx = \frac{1}{(\pi\sigma)^\beta} E\left[\frac{1}{(1+Z^2)^\beta}\right]$, where Z is a standard Cauchy random variable. We work with the Monte Carlo estimate of $E\left[\frac{1}{(1+Z^2)^\beta}\right]$ to bypass this difficulty. Just like the normal model, we need to calculate the gradient vector $\nabla\overline{D}_\beta(\boldsymbol{\theta}) = \left(\frac{\partial\overline{D}_\beta(\mu,\sigma)}{\partial\mu}, \frac{\partial\overline{D}_\beta(\mu,\sigma)}{\partial\sigma}\right)'$, and the Hessian matrix $\nabla^2\overline{D}_\beta(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial^2\overline{D}_\beta(\mu,\sigma)}{\partial^2\mu} & \frac{\partial^2\overline{D}_\beta(\mu,\sigma)}{\partial\mu\partial\sigma} \\ \frac{\partial^2\overline{D}_\beta(\mu,\sigma)}{\partial\mu\partial\sigma} & \frac{\partial^2\overline{D}_\beta(\mu,\sigma)}{\partial^2\sigma} \end{bmatrix}$ of the DPD objective function $\overline{D}_\beta(\boldsymbol{\theta}) = \overline{D}_\beta(\mu,\sigma)$. Detailed expressions of the elements of the gradient vector and Hessian matrix are provided in Section 5.6. Using these expressions, the gradient vector and the Hessian matrix can be calculated explicitly and these expressions are sufficient to derive the Newton-Raphson and the gradient descent algorithms. Since the Cauchy distribution is heavy tailed and the expectation and variance do not exist, we are not going to implement the Fisher's scoring algorithm in this case. The one-step IRLS updates of μ and σ are as follows.

$$\hat{\mu} = \frac{\sum_{i=1}^n w_i X_i}{\sum_{i=1}^n w_i},\ \hat{\sigma}^2 = \frac{\sum_{i=1}^n w_i (X_i-\hat{\mu})^2}{\sum_{i=1}^n w_i - \frac{n\beta}{(1+\beta)} E\left[\frac{1}{(1+Z^2)^\beta}\right]}, \tag{5.18}$$

where $w_i = \frac{1}{\left[1+\left(\frac{X_i-\hat{\mu}_0}{\hat{\sigma}_0}\right)^2\right]^{\beta+1}}$ for $i = 1,2,\ldots,n$.

(iii) Weibull Model: For the Weibull(k,λ) model, we assume the model density to be

$$f(x,k,\lambda) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^k},\ x \geq 0,$$

where $k > 0$ and $\lambda > 0$ are the shape and scale parameters, respectively. Elements of the corresponding gradient vector and Hessian matrix are provided in Section 5.6 which are necessary to construct the one-step estimators.

(iv) Shifted Gompertz Model: For the shifted Gompertz (SG) model, we assume the model density to be

$$f(x,\sigma,\eta) = \sigma e^{-\sigma x} e^{-\eta e^{-\sigma x}}(1+\eta(1-e^{-\sigma x})),\ x \geq 0,$$

where $\sigma > 0$ and $\eta > 0$ are the unknown scale and shape parameters, respectively. Ele-

ments of the corresponding gradient vector and Hessian matrix are provided in Section 5.6 which are necessary to construct the one-step estimators.

(v) Multivariate Normal Model: For the multivariate normal model, the one-step IRLS estimators of the location vector $\boldsymbol{\mu}$ and the scatter matrix $\boldsymbol{\Sigma}$ can be explicitly found in Equations (2.9) and (2.10) (Chapter 2). But for the rest of the one-step methods, we require a reparametrization of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ to construct a single parameter vector. In particular, we consider the parameter vector $\boldsymbol{\theta} = (\mu_1, \ldots, \mu_p, \sigma_1^2, \ldots, \sigma_p^2, \rho_{12}, \ldots \rho_{p-1,1})$, where μ_j and σ_j^2 are the respective mean and variance of the j-th component and ρ_{jk} is the correlation coefficient between the j-th and k-th components. But, the gradient vector and Hessian matrix, i.e., $\nabla \overline{D}_\beta(\boldsymbol{\theta})$ and $\nabla^2 \overline{D}_\beta(\boldsymbol{\theta})$ are not straightforward to obtain for this reparametrization as the objective function $\overline{D}_\beta(\boldsymbol{\theta})$ is highly nonlinear in the component means, variances and correlations in case of $p > 2$. Thus, we obtain the individual components of the gradient vector and the Hessian matrix numerically as:

$$\frac{\partial}{\partial \theta_j} \overline{D}_\beta(\boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_0} \approx \frac{\overline{D}_\beta(\hat{\boldsymbol{\theta}}_0 + \epsilon \boldsymbol{e}_j) - \overline{D}_\beta(\hat{\boldsymbol{\theta}}_0)}{\epsilon},$$
$$\frac{\partial^2}{\partial \theta_j \partial \theta_k} \overline{D}_\beta(\boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_0} \approx \frac{\overline{D}_\beta(\hat{\boldsymbol{\theta}}_0 + \epsilon(\boldsymbol{e}_j + \boldsymbol{e}_k)) - \overline{D}_\beta(\hat{\boldsymbol{\theta}}_0 + \epsilon e_j) - \overline{D}_\beta(\hat{\boldsymbol{\theta}}_0 + \epsilon e_k) + \overline{D}_\beta(\hat{\boldsymbol{\theta}}_0)}{\epsilon^2},$$

for a small prespecified constant $\epsilon > 0$. Here, $\boldsymbol{e}_j$ is the j-th canonical vector of dimension $\frac{p(p+3)}{2}$, for $1 \leq j \leq \frac{p(p+3)}{2}$.

5.2.6 Initial Estimators

We have seen the asymptotic results of the one-step Newton-Raphson, gradient descent and Fisher's scoring estimators so far. Although the asymptotic properties of the one-step Newton-Raphson and Fisher's scoring estimators do not depend on the choice of the initialization (under certain assumptions), but those of the one-step gradient descent estimator depends on the initialization. In fact, one may expect the performances of these one-step estimators (including the one-step IRLS one) should depend on both the choice of the initial estimator and the type of iteration used in small sample cases.

We utilize two types of estimators of location and scale as initial choices, namely, trimmed sample moment based estimators and quantile based estimators. The normal model is a location-scale model with finite moments. Thus, we may use both moment based as well as quantile based robust initial estimators for the location (i.e., mean)

and scale (i.e., standard deviation) parameters. But for the Cauchy model, even the first population moment (i.e., population mean) does not exist. We thus use quantile based initial estimators only in this case. In particular, for the univariate normal model we consider four different choices of initialization:

- **Sample Median and Median Absolute Deviation**: Here we use the sample median $\tilde{X}$ and a constant multiple of the sample median absolute deviation (MAD)= median$\{|X_i - \tilde{X}|, i = 1, \ldots, n\}$ (the constant is multiplied to make the estimator consistent), viz., 1.4826 MAD to estimate the population mean μ and population standard deviation σ, respectively.

- **Trimmed sample Mean and Variance:** Here we are taking a trimmed version of the sample mean and sample variance with an auto-driven trimming proportion following the metrical trimming approach and the standard χ^2 rule for outlier detection. Suppose $\{X_1, \ldots, X_n\}$ be the sample, $\tilde{X}$ be the sample median, $s =$ 1.4826 MAD (MAD being the median absolute deviation of the sample) and $d_i = \left(\frac{X_i - \tilde{X}}{s}\right)$ for $1 \leq i \leq n$. If d_i^2 is greater than the 97.5-th percentile of the $\chi^2(1)$- distribution, we trim the observation X_i from the sample. The mean and standard deviation of the remaining observations are used as initial choices of μ and σ, respectively.

- **Minimum Covariance Determinant Estimator**: Here we use the minimum covariance determinant (MCD) estimators of mean and covariance matrix (Rousseeuw (1985) [129]) which are basically trimmed sample mean and covariance matrix but the trimming is performed on the basis of minimizing the determinant of the covariance matrix (in the one-dimensional case it is equal to the variance) of possible subsamples.

- **S-Estimator of Location and Scale**: In this case, we use the S-estimator of location and scale which was first proposed by Rousseeuw and Yohai (1984) [134] in the regression set-up and later studied by Davies (1987) [37] and Lopuhaä (1989) [97] in the multivariate location-scale estimation set-up.

For the Cauchy model, we use the following initial estimators:

- **Sample Median and Median Absolute Deviation**: Here we use the sample median $\tilde{X}$ to estimate the location parameter μ and the sample median absolute deviation to estimate the scale parameter σ, initially.

- **Hodges-Lehmann Estimator:** Here also we use the sample median $\tilde{X}$ to estimate the location parameter μ, but to estimate the scale parameter σ, we use the Hodges-Lehmann estimator which is given by $\log\,\hat{\sigma} = \frac{1}{2}\text{median}(\log|(X_i - \tilde{X})(X_j - \tilde{X})|)$ where $1 \leq i < j \leq n$. (see Kravchuk and Pollett (2012) [89] for details).

In addition, for the Weibull model, we use the following highly robust initial estimators which were considered in Olive (2006) [117] and later used in Boudt et al. (2011) [16].

- For the shape parameter k and scale parameter λ, we use

$$\hat{\lambda}_0 = e^{\hat{\mu}}, \text{ and } \hat{k}_0 = \frac{1}{\hat{\sigma}}, \text{where} \tag{5.19}$$
$$\hat{\sigma} = 1.3037\ \text{MAD}(\log(X_i)), \text{ and } \hat{\mu} = \text{median}\ (\log(X_i),\ 1 \leq i \leq n) - \hat{\sigma}\log\log 2$$

  for initialization.

For the shifted Gompertz model, we consider the maximum likelihood estimators of the model parameters as initial estimators where the maximum likelihood estimators are obtained after removing the outlying observations present (using exploratory data analysis) in the datasets.

5.3 Generalization of the One-step IRLS Estimation for Elliptically Symmetric Models

Exact one-step estimation based on the Newton-Raphson, gradient descent, IRLS and Fisher's scoring iterations have been developed in Section 5.2. Although exact algebraic forms of these one-step estimators can be derived in case of the Newton-Raphson, gradient descent and Fisher's scoring methods, the one-step IRLS method does not convey any general algebraic form. However, we have derived the exact one-step IRLS estimators of location and scale parameters in normal and Cauchy cases. From Equations (5.17) and (5.18), the exact one-step IRLS location estimators are found to be weighted sample means and the exact one-step squared-scale or variance estimators are found to be weighted sample variance type estimators; these cannot be considered as weighted sample variance estimators as they are of the form $\frac{\sum_{i=1}^{n} w_i(X_i - \hat{\mu}_0)^2}{\sum_{i=1}^{n} w_i - c}$, where c is a non-zero constant. However, these scale estimators can be modified in a way so that they become constant multiples of weighted sample variance which have been

extensively studied from theoretical viewpoints (Lopuhaä and Rousseeuw (1991) [98], Lopuhaä (1999) [100], Croux and Haesbroeck (1999) [29], Hubert et al. (2018) [80]) in the past years. In contrast, the other exact one-step estimators do not follow such statistical intuition in general.

Motivated by these observations, we are going to develop and study a new form of robust estimators for the location and scale parameters of elliptically symmetric models. These new estimators are based on the generalization of the exact one-step (following IRLS) minimum DPD estimators of location and scale in case of normal models. To develop the generalized one-step estimation, let us consider the $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ model. As derived in Chapter 2 (Equations (2.9) and (2.10)), the minimum DPD estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ based on the random sample $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n\}$ can be obtained by solving the following system of equations:

$$\begin{aligned} &\tfrac{1}{n}\textstyle\sum_{i=1}^{n} \phi_p^{\beta}(\boldsymbol{X}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma})(\boldsymbol{X}_i - \boldsymbol{\mu}) = \boldsymbol{0}, \\ &\tfrac{1}{n}\textstyle\sum_{i=1}^{n} \phi_p^{\beta}(\boldsymbol{X}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma})(\boldsymbol{\Sigma} - (\boldsymbol{X}_i - \boldsymbol{\mu})(\boldsymbol{X}_i - \boldsymbol{\mu})') = c_0\boldsymbol{\Sigma}, \end{aligned} \tag{5.20}$$

where $c_0 = \beta(2\pi)^{-\frac{p\beta}{2}}|\boldsymbol{\Sigma}|^{-\frac{\beta}{2}}(1+\beta)^{-\frac{p+2}{2}}$. If $\hat{\boldsymbol{\mu}}_0$ and $\hat{\boldsymbol{\Sigma}}_0$ are some robust and consistent initial estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, respectively, exact one-step IRLS estimators of the same can be expressed as

$$\hat{\boldsymbol{\mu}} = \frac{\sum_{i=1}^{n} w(\boldsymbol{X}_i, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\Sigma}}_0)\boldsymbol{X}_i}{\sum_{i=1}^{n} w(\boldsymbol{X}_i, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\Sigma}}_0)}, \; \hat{\boldsymbol{\Sigma}} = \frac{\sum_{i=1}^{n} w(\boldsymbol{X}_i, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\Sigma}}_0)(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}})'}{\sum_{i=1}^{n} w(\boldsymbol{X}_i, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\Sigma}}_0) - \frac{n\beta}{(1+\beta)^{1+\frac{p}{2}}}} \tag{5.21}$$

by simplifying the system (5.20), where, $w(\boldsymbol{X}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = e^{-\frac{\beta}{2}(\boldsymbol{X}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})}$.

Remark 5.3. *Several things are to be noted in System (5.21). The location estimator $\hat{\boldsymbol{\mu}}$ is exactly the weighted sample mean with weights $w(\boldsymbol{X}_i, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\Sigma}}_0)$ and the form of the scale estimator $\hat{\boldsymbol{\Sigma}}$ is close to that of the weighted sample covariance matrix with weights $w(\boldsymbol{X}_i, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\Sigma}}_0)$; the little difference is due to the presence of the constant $\frac{n\beta}{(1+\beta)^{1+\frac{p}{2}}}$ in the denominator. The weight function is decreasing in the squared Mahalanobis distance $(\boldsymbol{X} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu})$, so that the observations in the central region of the data cloud get greater importance compared to the extreme observations. This kind of weights are ideal to produce robust estimators of location and scale as the resulting estimators are less affected by outliers due to the aforesaid downweighting of the extreme observations.*

Although the scale estimator $\hat{\boldsymbol{\Sigma}}$ is not exactly a weighted sample covariance matrix,

it can be thought of as a constant multiple of the weighted sample covariance matrix under certain assumptions. Let us first heuristically justify this fact. It is easy to rewrite $\hat{\boldsymbol{\Sigma}}$ as

$$\begin{aligned}\hat{\boldsymbol{\Sigma}} &= \frac{\sum_{i=1}^n w(\boldsymbol{X}_i, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\Sigma}}_0)(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}})'}{\sum_{i=1}^n w(\boldsymbol{X}_i, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\Sigma}}_0) - \frac{n\beta}{(1+\beta)^{1+\frac{p}{2}}}} \\ &= \frac{\sum_{i=1}^n w(\boldsymbol{X}_i, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\Sigma}}_0)(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}})'}{\sum_{i=1}^n w(\boldsymbol{X}_i, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\Sigma}}_0)} \frac{\sum_{i=1}^n w(\boldsymbol{X}_i, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\Sigma}}_0)}{\sum_{i=1}^n w(\boldsymbol{X}_i, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\Sigma}}_0) - \frac{n\beta}{(1+\beta)^{1+\frac{p}{2}}}} \\ &= \frac{\sum_{i=1}^n w(\boldsymbol{X}_i, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\Sigma}}_0)(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}})'}{\sum_{i=1}^n w(\boldsymbol{X}_i, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\Sigma}}_0)} \frac{\frac{1}{n}\sum_{i=1}^n w(\boldsymbol{X}_i, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\Sigma}}_0)}{\frac{1}{n}\sum_{i=1}^n w(\boldsymbol{X}_i, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\Sigma}}_0) - \frac{\beta}{(1+\beta)^{1+\frac{p}{2}}}},\end{aligned}$$

where the first fraction is the weighted sample covariance matrix with weights $w(\boldsymbol{X}_i, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\Sigma}}_0)$ and the second fraction is a function of the mean of the weights $\frac{1}{n}\sum_{i=1}^n w(\boldsymbol{X}_i, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\Sigma}}_0)$. Now, the sample is assumed to be modelled by a $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution and let us assume (for simplicity) that the true unknown probability distribution belongs to the model family (i.e., the true unknown distribution is also a p-variate normal with mean $\boldsymbol{\mu}_0$ and covariance $\boldsymbol{\Sigma}_0$; $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ being the true (unknown) values of the parameters). Assuming a large sample size, we may approximate

$$\begin{aligned}\frac{\frac{1}{n}\sum_{i=1}^n w(\boldsymbol{X}_i, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\Sigma}}_0)}{\frac{1}{n}\sum_{i=1}^n w(\boldsymbol{X}_i, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\Sigma}}_0) - \frac{\beta}{(1+\beta)^{1+\frac{p}{2}}}} &\approx \frac{\frac{1}{n}\sum_{i=1}^n w(\boldsymbol{X}_i, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)}{\frac{1}{n}\sum_{i=1}^n w(\boldsymbol{X}_i, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) - \frac{\beta}{(1+\beta)^{1+\frac{p}{2}}}} \\ &\approx \frac{E_{N_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)}(w(\boldsymbol{X}_1, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0))}{E_{N_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)}(w(\boldsymbol{X}_1, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)) - \frac{\beta}{(1+\beta)^{1+\frac{p}{2}}}}\end{aligned} \tag{5.22}$$

by virtue of consistency of the initial estimators and the weak law of large numbers (WLLN). The right hand side of (5.22) is just a constant, so that, the one-step estimator is approximately a scalar multiple of the weighted sample covariance matrix with weights $w(\boldsymbol{X}_i, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\Sigma}}_0)$.

Motivated by the above heuristic discussion, we propose a new robust estimation procedure for estimating the location and scale parameters in case of elliptically symmetric models. This estimation procedure is based on weighted sample means and variances with data-driven weights which have close connection with the minimum DPD estimation at least under normality.

5.3.1 Model Assumptions and Parameter Estimates

Let us assume that the random sample $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n\}$ is modelled by an elliptically symmetric model with PDF

$$f(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = |\boldsymbol{\Sigma}|^{-\frac{1}{2}} h\left((\boldsymbol{x}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right),\ \boldsymbol{x} \in \mathbb{R}^p, \tag{5.23}$$

where $\boldsymbol{\mu} \in \mathbb{R}^p$ is the location vector and $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$ is the scale matrix which is symmetric and positive definite. Multivariate Normal, t, logistic and many other distributions are members of this family.

Let us observe from system (5.21) that for the multivariate normal model with PDF $\phi_p(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{p}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$, the data-driven weights were $w(\boldsymbol{X}_i, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\Sigma}}_0) = e^{-\frac{\beta}{2}(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_0)' \hat{\boldsymbol{\Sigma}}_0^{-1}(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_0)} \propto \phi_p^{\beta}(\boldsymbol{X}_i, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\Sigma}}_0)$ (with proportionality constant $(2\pi)^{\frac{p\beta}{2}} |\hat{\boldsymbol{\Sigma}}_{\boldsymbol{0}}|^{\frac{\beta}{2}}$). Here, we generalize these weights through the density power weights, namely, $w(\boldsymbol{X}_i, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\Sigma}}_0) = h^{\beta}((\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_0)' \hat{\boldsymbol{\Sigma}}_0^{-1}(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_0)) \propto f^{\beta}(\boldsymbol{X}_i, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\Sigma}}_0)$ (the proportionality constant being $|\hat{\boldsymbol{\Sigma}}_0|^{\frac{\beta}{2}}$) and propose our new estimates of location and scale based on these weights. We need some further notations (following Lopuhaä (1999) [100]) to formally define our estimators. Let $\boldsymbol{M}_n$ and $\boldsymbol{V}_n$ be the initial robust (high breakdown) estimators of the location $\boldsymbol{\mu}$ and scale $\boldsymbol{\Sigma}$, respectively, i.e., we denote $\hat{\boldsymbol{\mu}}_0$ by $\boldsymbol{M}_n$ and $\hat{\boldsymbol{\Sigma}}_0$ by $\boldsymbol{V}_n$ from hereon. Thus, the data-driven weights based on these initial estimators are given by $w(\boldsymbol{X}_i, \boldsymbol{M}_n, \boldsymbol{V}_n) = h^{\beta}((\boldsymbol{X}_i - \boldsymbol{M}_n)' \boldsymbol{V}_n^{-1}(\boldsymbol{X}_i - \boldsymbol{M}_n))$ and we define the following weighted sample mean and covariance matrix based on the aforesaid weights:

$$\begin{aligned}
\boldsymbol{T}_n &= \frac{\sum_{i=1}^n w(\boldsymbol{X}_i, \boldsymbol{M}_n, \boldsymbol{V}_n) \boldsymbol{X}_i}{\sum_{i=1}^n w(\boldsymbol{X}_i, \boldsymbol{M}_n, \boldsymbol{V}_n)} = \frac{\sum_{i=1}^n h^{\beta}\big(d^2(\boldsymbol{X}_i, \boldsymbol{M}_n, \boldsymbol{V}_n)\big) \boldsymbol{X}_i}{\sum_{i=1}^n h^{\beta}(d^2(\boldsymbol{X}_i, \boldsymbol{M}_n, \boldsymbol{V}_n))}, \\
\boldsymbol{C}_n &= \frac{\sum_{i=1}^n w(\boldsymbol{X}_i, \boldsymbol{M}_n, \boldsymbol{V}_n)(\boldsymbol{X}_i - \boldsymbol{T}_n)(\boldsymbol{X}_i - \boldsymbol{T}_n)'}{\sum_{i=1}^n w(\boldsymbol{X}_i, \boldsymbol{M}_n, \boldsymbol{V}_n)} = \frac{\sum_{i=1}^n h^{\beta}\big(d^2(\boldsymbol{X}_i, \boldsymbol{M}_n, \boldsymbol{V}_n)\big)(\boldsymbol{X}_i - \boldsymbol{T}_n)(\boldsymbol{X}_i - \boldsymbol{T}_n)'}{\sum_{i=1}^n h^{\beta}(d^2(\boldsymbol{X}_i, \boldsymbol{M}_n, \boldsymbol{V}_n))},
\end{aligned} \tag{5.24}$$

where $d^2(\boldsymbol{X}_i, \boldsymbol{M}_n, \boldsymbol{V}_n) = (\boldsymbol{X}_i - \boldsymbol{M}_n)' \boldsymbol{V}_n^{-1}(\boldsymbol{X}_i - \boldsymbol{M}_n)$ is the squared Mahalanobis distance of $\boldsymbol{X}_i$ from $\boldsymbol{M}_n$ with respect to the scale $\boldsymbol{V}_n$. Following the intuition behind (5.22), it should be noted that the actual scale estimator should be a scalar multiple of the aforesaid weighted sample variance. The weight function in our set-up solely depends on the function h in the elliptical density $f(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$. Several assumptions on the weight function w (in the general set-up) and the function h have been assumed in Lopuhaä and Rousseeuw (1991) [98] and Lopuhaä (1999) [100] in order to ensure high breakdown, consistency and asymptotic normality of the weighted statistics $\boldsymbol{T}_n$ and $\boldsymbol{C}_n$ defined above. It is enough to assume the required regularity conditions on the

function h in this work as the weight function w explicitly depends on h in our set-up.

Remark 5.4. *A typical choice of the weight function is* $w(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = I(d^2(\boldsymbol{X}_i, \boldsymbol{M}_n, \boldsymbol{V}_n) \leq c)$ *which has been popularly utilized in many of the past literatures including Lopuhaä and Rousseeuw* (1991) *[98], Lopuhaä* (1999) *[100], Rousseeuw and van Zomeren* (1990) *[133], Gervini* (2003) *[60], Hubert et al.* (2018) *[80] and Croux and Haesbroeck* (1999) *[29]. In particular, the last four of the aforesaid works were developed to improve the MCD estimators of location and scale in terms of efficiency. In fact, the existing implementation for evaluating MCD based location and scale estimates in the* R *software (*covMcd *function in the package* robustbase *[102]) allows an optional reweighting step with this weight function. This weight function essentially cleans the data by detecting and deleting the sample observations whose Mahalanobis distances exceed a certain threshold and then calculate the ordinary sample mean and covariance based on the cleaned data.*

5.3.2 Regularity Conditions

In order to study the asymptotic and robustness properties of our proposed estimators, we need certain prerequisite conditions on the function h which automatically provide the required assumptions on the weight function. The following assumptions are made on the function h for the aforesaid objectives:

(H1) $h: \mathbb{R}^p \to \mathbb{R}$ is continuously differentiable.

(H2) The model density $f(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ admits a finite fourth moment, i.e.,

$$\int (\boldsymbol{x}'\boldsymbol{x})^2 h(\boldsymbol{x}'\boldsymbol{x})\, d\boldsymbol{x} < \infty.$$

(H3) h is bounded and non-increasing as a function of the squared Mahalanobis distance $d^2(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$.

The condition $(H3)$ is required to maintain a high breakdown of the weighted estimators as it guarantees downweighting of the extreme observations. In general, the weight function w requires some further regularity conditions in order to preserve the high breakdown of the initial estimates as well as for establishing asymptotic properties of the weighted estimators. Let us first look into those required conditions and then examine to what extent these can be established using $(H1)$, $(H2)$ and $(H3)$. In

general, the weight function $w[0,\infty) \to [0,\infty)$ (here w is considered as the function of the squared Mahalanobis distance) is required to satisfy

(W1) w is bounded and non-increasing.

(W2) There exists a constant C_1, such that, $w(y) = 0$ for $y \in [C_1, \infty)$.

(W3) w is of bounded variation, and continuous almost everywhere on $[0,\infty)$.

The conditions $(W1)$ and $(W2)$ have been assumed by Lopuhaä and Rousseeuw (1991) [98] for high breakdown preservation, while $(W3)$ is assumed by Lopuhaä (1999) [100] for establishing the asymptotic properties. In our case, $w\left(d^2(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})\right) = h^{\beta}\left(d^2(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})\right)$ for $\beta > 0$. So, the boundedness and monotonicity (i.e., non-increasing nature) of our weight function can be directly established using $(H3)$. Thus, our weight function satisfies $(W1)$. But the later conditions ($(W2)$ and $(W3)$) will not be followed from $(H1)$, $(H2)$ and $(H3)$, because by the non-negativity of the function h, our weight function $w\left(d^2(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})\right) = h^{\beta}\left(d^2(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})\right)$ will always be positive. Thus in this case, a finite constant C_1, as indicated in $(W2)$, will not exist. On the other hand, the notion of bounded variation is usually defined on closed intervals. Thus, our choice of the weight function does not satisfy $(W2)$ and $(W3)$ in general. So, we need to modify our weight function in such a way that the aforesaid prerequisites can be satisfied by the same. We define our modified weight function

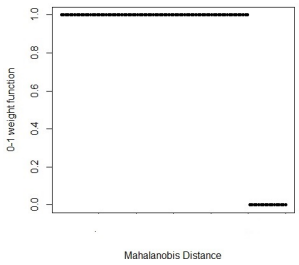
$$w_1\left(d^2(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})\right) = w\left(d^2(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})\right) I\left(w\left(d^2(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})\right) \geq c\right) \tag{5.25}$$

for some prespecified constant c. This modified weight function is defined in spirit of the traditional $0-1$ weight function (Remark 5.4); the thresholding is performed on $w\left(d^2(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})\right) = h^{\beta}\left(d^2(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})\right)$. The modified weight function w_1 clearly satisfies all of $(W1)$, $(W2)$ and $(W3)$. Intuitively, the modified weight function w_1 should not differ much from w, at least numerically. The value of w and w_1 are equal for large or moderate values of w and w_1 becomes 0, when w is small enough (equivalently, large values of the Mahalanobis distance), i.e., sufficiently small values of w are replaced by 0. Our weight function is a little bit advantageous compared to the traditional $0-1$ weight function. Both of these functions perform hard thresholding, but the amount of jump in case of our modified weight function w_1 is much smaller than that of the $0-1$ weight function (jump of constant height 1). Figure 5.2 represents this pictorially. The advantage of this phenomenon is that even if an extreme observation is erroneously

included in the cleaned sample, it will be given peripheral weightage by our modified weight function w_1 in contrast with the traditional $0-1$ weight function which assigns the same weight to all the observations. By $(W1)$, the modified weight function can also be written as

$$\begin{aligned} w_1\left(d^2(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{\Sigma})\right) &= w\left(d^2(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{\Sigma})\right) I(w\left(d^2(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{\Sigma})\right) \geq c) \\ &= w\left(d^2(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{\Sigma})\right) I(d^2(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{\Sigma}) \leq c^*) \end{aligned} \tag{5.26}$$

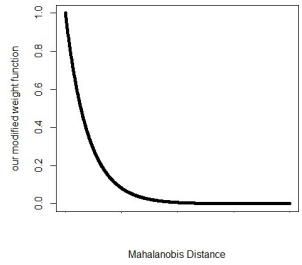
for some constant c^*.



(a) $0-1$ weight function

(b) w_1 with $c = 0.1$

(c) w_1 with $c = 0.02$



(d) w_1 with $c = 0.001$

Figure 5.2: The $0-1$ weight function versus our modified weight function w_1 (with $h(x) = e^{-0.5x}$).

So, we work with this weight function from hereon and the modified weighted sample mean and covariance matrix based on the new weight function are defined as

$$\begin{aligned} \boldsymbol{T}_{1n} &= \frac{\sum_{i=1}^{n} w_1(\boldsymbol{X}_i,\boldsymbol{M}_n,\boldsymbol{V}_n)\boldsymbol{X}_i}{\sum_{i=1}^{n} w_1(\boldsymbol{X}_i,\boldsymbol{M}_n,\boldsymbol{V}_n)}, \\ \boldsymbol{C}_{1n} &= \frac{\sum_{i=1}^{n} w_1(\boldsymbol{X}_i,\boldsymbol{M}_n,\boldsymbol{V}_n)(\boldsymbol{X}_i-\boldsymbol{T}_{1n})(\boldsymbol{X}_i-\boldsymbol{T}_{1n})'}{\sum_{i=1}^{n} w_1(\boldsymbol{X}_i,\boldsymbol{M}_n,\boldsymbol{V}_n)}. \end{aligned} \tag{5.27}$$

5.3.3 Breakdown and Asymptotic Properties

Now, we are in a position to discuss the robustness (in terms of breakdown) and asymptotic properties (consistency and asymptotic normality) of $\boldsymbol{T}_{1n}$ and $\boldsymbol{C}_{1n}$. For the result on breakdown, we follow Theorem 5.1 of Lopuhaä and Rousseeuw (1991) [98]. Our relevant result is given in the following Theorem.

Theorem 5.4. *Suppose the initial robust estimators $\boldsymbol{M}_n$ and $\boldsymbol{V}_n$ of location and scale are affine equivariant and their corresponding robust ellipsoid $E(\boldsymbol{M}_n, \boldsymbol{V}_n, c_0)$ (defined as $\{\boldsymbol{x} : d^2(\boldsymbol{x}, \boldsymbol{M}_n, \boldsymbol{V}_n) \leq c_0^2\}$) satisfies*

$$|\{i : \ \boldsymbol{X}_i \in E(\boldsymbol{M}_n, \boldsymbol{V}_n, c_0)\}| \geq \left[\frac{n+p+1}{2}\right] \tag{5.28}$$

for any random sample $\boldsymbol{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n\}$ from a p-dimensional elliptical distribution with $n \geq p+1$, where $|S|$ is the cardinality of the set S (see Equation (5.1) *of Lopuhaä and Rousseeuw* (1991) *[98] for more details). Under* $(W1)$ *and* $(W2)$*,*

$$min\{\epsilon^*(\boldsymbol{T}_{1n}, \boldsymbol{X}), \epsilon^*(\boldsymbol{C}_{1n}, \boldsymbol{X})\} \geq min\{\epsilon^*(\boldsymbol{M}_n, \boldsymbol{X}), \epsilon^*(\boldsymbol{V}_n, \boldsymbol{X})\}, \tag{5.29}$$

where $\epsilon^*(\boldsymbol{S}_n, \boldsymbol{X})$ is the breakdown point of the estimator $\boldsymbol{S}_n$ based on a random sample $\boldsymbol{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n\}$ of size n.

Here, the weight function w_1 satisfies $(W1)$ and $(W2)$ and the proof of Theorem 5.4 thus follows from that of Theorem 5.1 of Lopuhaä and Rousseeuw (1991) [98]. Some remarks are necessary in order to understand the pertinence of the aforesaid breakdown result.

Remark 5.5. *The inequality in (5.29) directly shows the preservation of the breakdown behaviour of the initial estimators $\boldsymbol{M}_n$ and $\boldsymbol{V}_n$ in the sense that the minimum of the breakdown points of the weighted sample mean $\boldsymbol{T}_{1n}$ and covariance $\boldsymbol{C}_{1n}$, calculated with the data-driven weights based on $\boldsymbol{M}_n$ and $\boldsymbol{V}_n$ is no less than the minimum of the breakdown values of the initial estimators.*

Remark 5.6. *As described in Lopuhaä and Rousseeuw* (1991) *[98], typical choices of the initial estimates $\boldsymbol{M}_n$ and $\boldsymbol{V}_n$ include the minimum volume ellipsoid estimates of location and scale (Rousseeuw* (1985) *[129]) and the S-estimates of location and scale (Lopuhaä* (1989) *[97]). If we assume $n \gg p$, then $\left[\frac{n+p+1}{2}\right] \approx \frac{n}{2}$ and the inequality in (5.28) ensures that at least half of the sample observations must be used to construct*

the initial estimators. This is crucial to maintain the high breakdown of the resulting estimates. If the MCD estimators are used as initial estimators, that must also be constructed using at least half of the sample observations.

Our next agenda is to discuss the consistency of our one-step estimators $\boldsymbol{T}_{1n}$ and $\boldsymbol{C}_{1n}$ generated by the modified weight function w_1 which satisfies $(W3)$ in particular. We follow the approach of Lopuhaä (1999) [100] in this regard. In particular, our result can be established as a consequence of Theorem 4.1 and Corollary 4.1 of Lopuhaä (1999) [100]. Before stating the result formally, we need to define the following constants:

$$
\begin{aligned}
c_0 &= \tfrac{2\pi^{\frac{p}{2}}}{\Gamma\left(\frac{p}{2}\right)} \int_0^\infty \tfrac{2}{p} w_1(r^2) h^{'}(r^2) r^{p+1}\, dr = \tfrac{2\pi^{\frac{p}{2}}}{\Gamma\left(\frac{p}{2}\right)} \int_0^{\sqrt{c^*}} \tfrac{2}{p} h^{\beta}(r^2) h^{'}(r^2) r^{p+1}\, dr, \\
c_1 &= \tfrac{2\pi^{\frac{p}{2}}}{\Gamma\left(\frac{p}{2}\right)} \int_0^\infty w_1(r^2) h(r^2) r^{p-1}\, dr = \tfrac{2\pi^{\frac{p}{2}}}{\Gamma\left(\frac{p}{2}\right)} \int_0^{\sqrt{c^*}} h^{1+\beta}(r^2) r^{p-1}\, dr, \\
c_2 &= \tfrac{2\pi^{\frac{p}{2}}}{\Gamma\left(\frac{p}{2}\right)} \int_0^\infty w_1(r^2) \left(h(r^2) + \tfrac{2}{p} h^{'}(r^2) r^2 \right) r^{p-1}\, dr \\
&= \tfrac{2\pi^{\frac{p}{2}}}{\Gamma\left(\frac{p}{2}\right)} \int_0^{\sqrt{c^*}} h^{\beta}(r^2) \left(h(r^2) + \tfrac{2}{p} h^{'}(r^2) r^2 \right) r^{p-1}\, dr, \\
c_3 &= \tfrac{2\pi^{\frac{p}{2}}}{\Gamma\left(\frac{p}{2}\right)} \int_0^\infty \tfrac{1}{p} w_1(r^2) h(r^2) r^{p+1}\, dr = \tfrac{2\pi^{\frac{p}{2}}}{\Gamma\left(\frac{p}{2}\right)} \int_0^{\sqrt{c^*}} \tfrac{1}{p} h^{1+\beta}(r^2) r^{p+1}\, dr, \\
c_4 &= \tfrac{2\pi^{\frac{p}{2}}}{\Gamma\left(\frac{p}{2}\right)} \int_0^\infty w_1(r^2) \left[\tfrac{r^2}{p} h(r^2) + \tfrac{2r^4}{p(p+2)} h^{'}(r^2) \right] r^{p+1}\, dr \\
&= \tfrac{2\pi^{\frac{p}{2}}}{\Gamma\left(\frac{p}{2}\right)} \int_0^{\sqrt{c^*}} h^{\beta}(r^2) \left[\tfrac{r^2}{p} h(r^2) + \tfrac{2r^4}{p(p+2)} h^{'}(r^2) \right] r^{p+1}\, dr,
\end{aligned}
\tag{5.30}
$$

where the constant c^* in the upper limits of the integrations has been defined in Equation (5.26). We assume that the true unknown distribution belongs to the model family, i.e., the random sample $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n\}$ is distributed with the PDF $f(\cdot, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, where $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ are the true values of the unknown location $\boldsymbol{\mu}$ and scale $\boldsymbol{\Sigma}$, respectively.

Theorem 5.5. *Let $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n\}$ be a random sample from an elliptic distribution with PDF $f(\cdot, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, location $\boldsymbol{\mu}_0$ and scale $\boldsymbol{\Sigma}_0$. If the initial estimators $\boldsymbol{M}_n$ and $\boldsymbol{V}_n$ converge in probability to $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$, respectively, and the function h satisfies $(H1)$, $(H2)$ and $(H3)$, then the one-step estimators $\boldsymbol{T}_{1n}$ and $\boldsymbol{C}_{1n}$ converge to $\boldsymbol{\mu}_0$ and $\frac{c_3}{c_1}\boldsymbol{\Sigma}_0$, respectively, as $n \to \infty$.*

It is trivial to observe that although the generalized one-step location estimator $\boldsymbol{T}_{1n}$ is consistent for the location parameter $\boldsymbol{\mu}$, the corresponding scale estimator is not. In fact, a scalar multiple of the weighted sample variance, viz., $\frac{c_1}{c_3}\boldsymbol{C}_{1n}$ is consistent for $\boldsymbol{\Sigma}$. We refer to the multiplier $\frac{c_1}{c_3}$ as the "consistency factor". In case of the multivariate normal model, we have heuristically seen (in (5.22)) that the exact one-step

(IRLS) estimator of the scale (covariance matrix in case of normality) is asymptotically equivalent to a scalar multiplie (provided in (5.22)) of the sample weighted covariance matrix with weight function w. We now show that this constant is indeed the same consistency factor $\frac{c_1}{c_3}$ in a special case.

Theorem 5.6. *If $c^* = \infty$ (equivalently, $c = 0$ which ensures $w = w_1$), the consistency factor $\frac{c_1}{c_3}$ is equal to the constant on the right hand side of (5.22) under normality.*

This relationship illustrates the link between the exact one-step (IRLS) estimation and our newly proposed generalized one-step estimation at least under normality. The proof of Theorem 5.6 is provided in Section 5.6.

The asymptotic normality of our proposed estimators need to be discussed now. Lopuhaä (1999) [100] (Theorem 5.1 in [100]) established the asymptotic normality of $\boldsymbol{T}_{1n}$ and $\boldsymbol{C}_{1n}$ with smoothed S-initialization. Assuming the regularity conditions assumed in Lopuhaä (1999) [100] and (1997) [99], this result states the following.

Theorem 5.7. *If the generalized one-step estimators $\boldsymbol{T}_{1n}$ and $\boldsymbol{C}_{1n}$ are constructed using smoothed S-initialization, then,*

$$\begin{array}{l} \sqrt{n}\,(\boldsymbol{T}_{1n} - \boldsymbol{\mu}) \xrightarrow{d} N(\mathbf{0}, \alpha\Sigma), \text{ and} \\ \sqrt{n}\left(\boldsymbol{C}_{1n} - \frac{c_3}{c_1}\boldsymbol{\Sigma}\right) \xrightarrow{d} N(\mathbf{0}, \overline{\boldsymbol{\Sigma}}), \end{array} \tag{5.31}$$

where $\overline{\boldsymbol{\Sigma}} = \sigma_1(\boldsymbol{I} + \boldsymbol{D}_{p,p})(\boldsymbol{\Sigma} \bigotimes \boldsymbol{\Sigma}) + \sigma_2 vec(\boldsymbol{\Sigma}) vec(\boldsymbol{\Sigma})$ and they are asymptotically independent.

The values of α, σ_1, σ_2 and $\boldsymbol{D}_{p,p}$ can be found in Lopuhaä (1999) [100] (Section 5, Page-1652).

5.4 Simulation Experiments

We have proposed the exact one-step minimization of the DPD with different iterative procedures and various initializations in Section 5.2 and the asymptotic and robustness properties of the resulting estimators have been extensively discussed. Following the salient statistical interpretation of the exact one-step IRLS estimators in case of normal and Cauchy models, it has also been generalized for elliptically symmetric probability models in Section 5.3 and the theoretical properties of these generalized one-step

estimators have been discussed following Lopuhaä and Rousseeuw (1991) [98] and Lopuhaä (1999) [100]. We now assess the comparative performances of these newly proposed estimators along with the initializations and other weighted estimators through simulation experiments.

5.4.1 Simulation Experiments in Univariate Set-ups

To understand the relative strengths of the exact one-step estimators proposed in Section 5.2, we simulate random samples from univariate normal and Cauchy distributions at first. Since the Cauchy distribution is heavy-tailed in nature, random samples from Cauchy distribution are expected to carry some extreme observations. However, the normal distribution is not heavy-tailed. Thus, to observe the resistant effects of the proposed exact one-step estimators (against outliers), we work with both pure and contaminated normal distributions along with the Cauchy distribution. For the pure normal set-up, 1000 random samples of sizes $n = 30,\ 100$ are generated from the $N(0, 1)$ distribution. For the contaminated normal set-up, we simulate 1000 random samples of sizes $n = 30,\ 100$ from a normal mixture distribution with a $N(0, 1)$ component with mixing proportion 90% and another $N(30, 1)$ component with mixing proportion 10%. Basically, we are contaminating the $N(0, 1)$ distribution with some distant observations from $N(30, 1)$ distribution and the contaminating proportion is 10%. For the Cauchy distribution, we simulate 1000 random samples of sizes $n = 30,\ 100$ from the Cauchy$(0, 1)$ distribution. We present the estimated absolute bias and variances of different exact one-step estimators (with different initializations and iterations), fully converged minimum DPD estimators (with different initializations) and the initial estimators in Tables 5.2, 5.3 and 5.4. Let us now briefly discuss the salient features of the simulation outputs.

- It is evident from each of the Tables 5.2, 5.3 and 5.4 that our exact one-step updates are definitely improving the performances of the initial estimators except the Newton-Raphson method (in some cases, especially for the Cauchy set-up).

- The absolute bias and variances of different exact one-step estimators are not similar (especially in case of $n = 30$) for different initializations. In case of contaminated normal samples, the trimmed mean and variance estimators become the best initialization scheme in terms of absolute bias and variances with S-estimators being a close competitor. In case of the Cauchy samples, the Hodges-

n	Initial Estimator	β	Parameter	Accuracy Measure	NR	GD	IRLS	FS	MDPDE	Initial
30	Median and MAD	0.1	Mean	Absolute Bias	0.002	0.003	0.002	0.002	0.002	0.006
				Variance	0.033	0.033	0.033	0.033	0.032	0.046
			Variance	Absolute Bias	0.110	0.050	0.041	0.037	0.028	0.002
				Variance	0.084	0.116	0.071	0.071	0.069	0.171
	Trimmed Mean, Variance	0.1	Mean	Absolute Bias	0.002	0.002	0.002	0.002	0.002	0.005
				Variance	0.033	0.033	0.033	0.033	0.032	0.039
			Variance	Absolute Bias	0.110	0.068	0.060	0.047	0.028	0.203
				Variance	0.079	0.050	0.071	0.072	0.069	0.084
	MCD Estimators	0.1	Mean	Absolute Bias	0.002	0.002	0.002	0.002	0.005	0.002
				Variance	0.033	0.033	0.033	0.033	0.032	0.044
			Variance	Absolute Bias	0.039	0.019	0.032	0.024	0.028	0.017
				Variance	0.070	0.072	0.069	0.070	0.069	0.081
	S-Estimators	0.1	Mean	Absolute Bias	0.006	0.006	0.002	0.003	0.002	0.008
				Variance	0.038	0.039	0.035	0.034	0.032	0.097
			Variance	Absolute Bias	0.111	0.010	0.036	0.023	0.028	0.063
				Variance	0.096	0.077	0.071	0.084	0.069	0.112
100	Median and MAD	0.1	Mean	Absolute Bias	0.0001	0.0001	0.0001	0.0001	0.0002	0.0003
				Variance	0.010	0.010	0.010	0.010	0.010	0.015
			Variance	Absolute Bias	0.050	0.023	0.011	0.009	0.007	0.002
				Variance	0.022	0.035	0.021	0.021	0.020	0.058
	Trimmed Mean, Variance	0.1	Mean	Absolute Bias	0.0002	0.0001	0.0001	0.0001	0.0001	0.0003
				Variance	0.010	0.010	0.010	0.010	0.010	0.012
			Variance	Absolute Bias	0.062	0.061	0.027	0.015	0.007	0.169
				Variance	0.023	0.014	0.021	0.021	0.020	0.027
	MCD Estimators	0.1	Mean	Absolute Bias	0.0002	0.001	0.0001	0.0001	0.0002	0.001
				Variance	0.010	0.010	0.010	0.010	0.010	0.014
			Variance	Absolute Bias	0.010	0.003	0.008	0.005	0.008	0.002
				Variance	0.020	0.021	0.021	0.020	0.020	0.023
	S-Estimators	0.1	Mean	Absolute Bias	0.003	0.003	0.004	0.003	0.0001	0.005
				Variance	0.010	0.011	0.010	0.010	0.010	0.031
			Variance	Absolute Bias	0.043	0.001	0.010	0.011	0.007	0.025
				Variance	0.021	0.026	0.020	0.021	0.020	0.035

Table 5.2: Estimated absolute bias and variances of various exact one-step estimates in case of pure normal samples.

Lehmann scale estimators are slightly better than the median absolute deviation in terms of estimated absolute bias and variances. These observations definitely suggest the explicit dependence of the exact one-step updates on their respective initialization schemes. But for $n = 100$, the estimated variances of the exact one-step Newton-Raphson and Fisher's scoring updates and those of the original

n	Initial Estimator	β	Parameter	Accuracy Measure	NR	GD	IRLS	FS	MDPDE	Initial
30	Median and MAD	0.3	Mean	Absolute Bias	0.008	0.093	0.035	0.015	0.009	0.162
				Variance	0.038	0.052	0.039	0.038	0.037	0.064
			Variance	Absolute Bias	0.024	0.345	0.033	0.05	0.011	0.362
				Variance	0.148	0.468	0.105	0.129	0.095	0.494
	Trimmed Mean, Variance	0.1	Mean	Absolute Bias	0.012	0.011	0.011	0.012	0.009	0.021
				Variance	0.036	0.035	0.036	0.036	0.035	0.04
			Variance	Absolute Bias	0.081	0.064	0.047	0.049	0.024	0.158
				Variance	0.082	0.06	0.077	0.077	0.076	0.087
	MCD Estimators	0.3	Mean	Absolute Bias	0.009	0.01	0.009	0.009	0.009	0.01
				Variance	0.036	0.038	0.037	0.036	0.037	0.042
			Variance	Absolute Bias	0.005	0.489	0.069	0.026	0.011	0.542
				Variance	0.116	0.383	0.103	0.096	0.095	0.372
	S-Estimators	0.3	Mean	Absolute Bias	0.017	0.005	0.001	0.0004	0.009	0.004
				Variance	0.044	0.052	0.045	0.042	0.037	0.089
			Variance	Absolute Bias	0.004	0.236	0.031	0.051	0.011	0.259
				Variance	0.119	0.283	0.094	0.099	0.095	0.24
100	Median and MAD	0.3	Mean	Absolute Bias	0.002	0.075	0.023	0.001	0.003	0.142
				Variance	0.012	0.016	0.013	0.012	0.012	0.021
			Variance	Absolute Bias	0.061	0.307	0.076	0.073	0.039	0.336
				Variance	0.032	0.109	0.032	0.031	0.028	0.117
	Trimmed Mean, Variance	0.1	Mean	Absolute Bias	0.002	0.002	0.002	0.002	0.004	0.008
				Variance	0.012	0.012	0.012	0.012	0.011	0.013
			Variance	Absolute Bias	0.022	0.044	0.006	0.009	0.006	0.100
				Variance	0.025	0.019	0.024	0.024	0.023	0.029
	MCD Estimators	0.3	Mean	Absolute Bias	0.003	0.003	0.003	0.003	0.003	0.003
				Variance	0.012	0.012	0.012	0.012	0.012	0.012
			Variance	Absolute Bias	0.045	0.511	0.106	0.047	0.039	0.563
				Variance	0.029	0.101	0.031	0.027	0.028	0.098
	S-Estimators	0.1	Mean	Absolute Bias	0.002	0.001	0.0004	0.0003	0.004	0.001
				Variance	0.012	0.014	0.012	0.012	0.011	0.029
			Variance	Absolute Bias	0.024	0.245	0.014	0.037	0.006	0.306
				Variance	0.029	0.069	0.023	0.025	0.023	0.079

Table 5.3: Estimated absolute bias and variances of various exact one-step estimates in case of contaminated normal samples.

(fully converged) minimum DPD estimators are close to each other (in most of the cases) which supports the fact that the asymptotic variances of the exact one-step Newton-Raphson, Fisher's scoring and the original minimum DPD estimators are same. The observation is prominent particularly in case of the pure normal set-up.

n	Initial Estimator	β	Parameter	Accuracy Measure	NR	GD	IRLS	MDPDE	Initial
30	Median and MAD	0.1	Location	Absolute Bias	0.008	0.003	0.002	0.002	0.001
				Variance	0.084	0.083	0.081	0.079	0.091
			Scale	Absolute Bias	0.046	0.028	0.02	0.004	0.028
				Variance	0.29	0.081	0.078	0.072	0.085
	Median and HL Estimator	0.1	Location	Absolute Bias	0.011	0.002	0.002	0.009	0.001
				Variance	0.152	0.083	0.081	0.069	0.091
			Scale	Absolute Bias	0.008	0.007	0.004	0.002	0.003
				Variance	1.941	0.076	0.077	0.074	0.079
100	Median and MAD	0.1	Location	Absolute Bias	0.001	0.001	0.001	0.001	0.001
				Variance	0.02	0.022	0.021	0.021	0.024
			Scale	Absolute Bias	0.03	0.011	0.007	0.003	0.01
				Variance	0.027	0.024	0.023	0.021	0.027
	Median and HL Estimator	0.1	Location	Absolute Bias	0.001	0.001	0.001	0.005	0.001
				Variance	0.02	0.022	0.021	0.022	0.024
			Scale	Absolute Bias	0.005	0.001	0.001	0.002	0.004
				Variance	0.025	0.021	0.021	0.021	0.021

Table 5.4: Estimated absolute bias and variances of various exact one-step estimates in case of Cauchy samples.

- The existence of cross-effects between different iterations and initializations on the exact one-step updates can clearly be seen for the pure and contaminated normal and the Cauchy outputs. In particular, the Newton-Raphson method is the most unstable one with different initial estimators in case of the contaminated normal random samples. In turn, the Fisher's scoring update is quite more stable than the Newton-Raphson update. It is probably due to the non-smooth nature of the second derivative of the objective function which is used in the Newton-Raphson method but its expectation with respect to the model density (at the initial estimators) is used in the Fisher's scoring exact one-step updates which is free of sampling fluctuations (thus expected to be smoother). On the other hand, the two first-order methods, i.e., gradient descent and IRLS (since they consist of the first derivatives of the objective function only) are more or less stable, although the effects of different initial estimators can clearly be observed. On an average, the IRLS and the Fisher's scoring exact one-step updates are the best performers.

- As expected, the fully converged minimum DPD estimators are marginally better (in terms of the estimated absolute bias and variances) than the exact one-step updates but the loss is minimal and the exact one-step updates are fairly compa-

rable with the fully converged estimates (especially the exact one-step IRLS and Fisher's scoring estimators).

We need to observe an important computational aspect at this point. Let us recall the fact that either in case of normal or in case of Cauchy models, we are estimating the location and scale parameters. Although the location parameter is unrestricted, the scale parameter is strictly positive. We have not imposed this constraint in our exact one-step methods. And if we look into the systems of estimating equations (5.2),(5.3),(5.5),(5.17),(5.18), none of them can theoretically guarantee the positive definiteness of the exact one-step scale estimators. A safeguard is required indeed to avoid this hindrance. We use a reparametrization of the scale parameter $\sigma = e^{\lambda}$ and perform the corresponding exact one-step estimation procedures if the original exact one-step scale estimates (without the reparametrization) are found to be negative. The proportion of time this does not happen (i.e., the exact one-step variance/scale estimates are found to be positive without the reparamterization) in the simulation experiments are tabulated in Tables 5.5 (for pure normal set-up), 5.6 (for contaminated normal set-up) and 5.7 (for Cauchy set-up). Clearly, there is no problem with the dif-

n	Initial Estimator	β	NR	GD	IRLS	FS
30	Median and MAD	0.1	0.938	1.000	1.000	1.000
	Trimmed Mean, Variance	0.1	1.000	1.000	1.000	1.000
	MCD Estimators	0.1	1.000	1.000	1.000	1.000
	S-Estimators	0.1	0.998	1.000	1.000	1.000
100	Median and MAD	0.1	0.997	1.000	1.000	1.000
	Trimmed Mean, Variance	0.1	1.000	1.000	1.000	1.000
	MCD Estimators	0.1	1.000	1.000	1.000	1.000
	S-Estimators	0.1	1.000	1.000	1.000	1.000

Table 5.5: Proportion of times the exact one-step pure normal variance estimates are found to be positive.

n	Initial Estimator	β	NR	GD	IRLS	FS
30	Median and MAD	0.3	0.750	1.000	1.000	1.000
	Trimmed Mean, Variance	0.1	1.000	1.000	1.000	1.000
	MCD Estimators	0.3	0.631	1.000	1.000	1.000
	S-Estimators	0.3	0.854	1.000	1.000	1.000
100	Median and MAD	0.3	0.848	1.000	1.000	1.000
	Trimmed Mean, Variance	0.1	1.000	1.000	1.000	1.000
	MCD Estimators	0.3	0.664	1.000	1.000	1.000
	S-Estimators	0.1	0.867	1.000	1.000	1.000

Table 5.6: Proportion of times the exact one-step contaminated normal variance estimates are found to be positive.

n	Initial Estimator	β	NR	GD	IRLS
30	Median and MAD	0.3	0.963	1.000	1.000
	Median and HL Scale Estimator	0.1	0.975	1.000	1.000
100	Median and MAD	0.3	1.000	1.000	1.000
	Median and HL Scale Estimator	0.1	1.000	1.000	1.000

Table 5.7: Proportion of times the exact one-step Cauchy scale estimates are found to be positive.

ferent algorithms except the Newton-Raphson method. As discussed in the simulation discussion, the Newton-Raphson method may be heavily affected by the non-smooth nature of the second derivative of the objective function which may cause the problem of both producing negative variance estimates as well as heavy fluctuations in the absolute bias and variances of the exact one-step Newton-Raphson updates.

We now present the simulation experiments in multivariate set-ups including the exact one-step gradient descent, IRLS and Fisher's scoring estimators and generalized (density power and traditional $0-1$) weight based one-step estimators along with different initial estimators.

5.4.2 Simulation Experiments in Multivariate Set-ups

Here we utilize five different initial highly robust location and scale estimators for this purpose. These include the MCD estimator, the MVE estimator, the orthogonalized Gnanadesikan-Kettenring (GK) estimator (Maronna and Zamar (2002), [112]), MM estimators of location and scale (MM, Tatsuoka and Tyler (2000) [143]) and S-estimators of location and scale (Rousseeuw and Yohai (1984) [134], Ruppert (1992) [135]). We utilize three different probabilistic set-ups for the simulation experiments. Firstly, we generate observations from pure and secondly from contaminated multivariate normal distributions of different dimensions and sample sizes to observe the comparative performances of the one-step methodologies under contamination free and outlier contaminated set-ups. Finally, observations from multivariate t-distributions of different dimensions are generated to assess the performances of the generalized weight (with density power and traditional $0-1$ weights) based estimators in case of heavy tailed set-ups. Datasets of different dimensions and sample sizes (depending on the data dimensions) are simulated from the aforesaid distributions (with 1000 replications).

For both pure and contaminated normal set-ups, we simulate datasets of dimensions $p = 2,\ 6$ and 10 with respective sample sizes $n = 100,\ 200$ and 300. For the pure

Initialization	Parameter	Accuracy	MLE	GD		IRLS		FS		MDPDE		Initial
				β		β		β		β		Estimators
		Measures		0.05	0.1	0.05	0.1	0.05	0.1	0.05	0.1	
MCD	Location	Bias	0.003	0.003	0.003	0.004	0.005	0.006	0.006	0.004	0.004	0.006
		MSE	0.02	0.024	0.024	0.02	0.02	0.024	0.024	0.02	0.02	0.024
	Scatter	Bias	0.024	0.026	0.023	0.016	0.015	0.024	0.028	0.021	0.02	0.026
		MSE	0.057	0.098	0.098	0.061	0.063	0.061	0.063	0.058	0.059	0.094
MVE	Location	Bias	0.003	0.002	0.002	0.003	0.004	0.01	0.011	0.004	0.004	0.005
		MSE	0.02	0.025	0.025	0.021	0.021	0.023	0.024	0.02	0.02	0.025
	Scatter	Bias	0.024	0.013	0.011	0.011	0.011	0.022	0.02	0.021	0.02	0.013
		MSE	0.057	0.09	0.092	0.061	0.06	0.064	0.064	0.058	0.059	0.095
GK	Location	Bias	0.003	0.004	0.004	0.004	0.004	0.009	0.011	0.004	0.004	0.006
		MSE	0.02	0.027	0.027	0.02	0.021	0.023	0.024	0.02	0.02	0.028
	Scatter	Bias	0.024	0.347	0.358	0.048	0.066	0.026	0.04	0.021	0.02	0.377
		MSE	0.057	0.169	0.18	0.058	0.063	0.065	0.064	0.058	0.059	0.203
MM	Location	Bias	0.003	0.004	0.004	0.001	0.001	0.012	0.015	0.004	0.004	0.001
		MSE	0.02	0.02	0.02	0.021	0.021	0.023	0.024	0.02	0.02	0.021
	Scatter	Bias	0.024	0.036	0.039	0.021	0.021	0.022	0.024	0.021	0.02	0.032
		MSE	0.057	0.069	0.069	0.06	0.062	0.061	0.062	0.058	0.059	0.069
S	Location	Bias	0.003	0.003	0.003	0.001	0.001	0.011	0.016	0.004	0.004	0.003
		MSE	0.02	0.032	0.032	0.021	0.022	0.023	0.023	0.02	0.02	0.035
	Scatter	Bias	0.024	0.011	0.011	0.025	0.028	0.021	0.026	0.021	0.02	0.014
		MSE	0.057	0.127	0.127	0.061	0.064	0.06	0.061	0.058	0.059	0.137

Table 5.8: Estimated bias and mean squared errors of exact one-step (GD, IRLS and FS) and fully converged minimum DPD location-scale estimators for pure normal datasets with $p = 2$.

Initialization	Parameter	Accuracy	MLE	GD		IRLS		FS		MDPDE		Initial
				β		β		β		β		Estimators
		Measures		0.05	0.1	0.05	0.1	0.05	0.1	0.05	0.1	
MCD	Location	Bias	4.275	0.002	0.002	0.002	0.002	0.012	0.012	0.008	0.008	0.001
		MSE	19.907	0.023	0.023	0.022	0.023	0.02	0.02	0.024	0.024	0.025
	Scatter	Bias	161.67	0.436	0.436	0.008	0.028	0.017	0.019	0.007	0.006	0.438
		MSE	27972.04	0.36	0.361	0.071	0.075	0.065	0.069	0.068	0.071	0.36
MVE	Location	Bias	4.275	0.006	0.006	0.008	0.008	0.008	0.008	0.008	0.008	0.009
		MSE	19.907	0.026	0.026	0.024	0.024	0.024	0.024	0.024	0.024	0.025
	Scatter	Bias	161.67	0.435	0.435	0.008	0.029	0.029	0.032	0.007	0.006	0.445
		MSE	27972.04	0.347	0.348	0.07	0.074	0.073	0.075	0.068	0.071	0.363
GK	Location	Bias	4.275	0.004	0.004	0.008	0.008	0.007	0.007	0.008	0.008	0.008
		MSE	19.907	0.029	0.03	0.024	0.024	0.024	0.024	0.024	0.024	0.03
	Scatter	Bias	161.67	0.271	0.279	0.024	0.033	0.039	0.04	0.007	0.006	0.294
		MSE	27972.04	0.138	0.145	0.068	0.072	0.07	0.071	0.068	0.071	0.162
MM	Location	Bias	4.275	0.004	0.004	0.002	0.002	0.008	0.012	0.008	0.008	0.002
		MSE	19.907	0.022	0.022	0.022	0.023	0.025	0.026	0.024	0.024	0.023
	Scatter	Bias	161.67	0.297	0.297	0.005	0.023	0.023	0.025	0.007	0.006	0.313
		MSE	27972.04	0.209	0.21	0.07	0.073	0.074	0.077	0.068	0.071	0.23
S	Location	Bias	4.275	0.003	0.003	0.002	0.003	0.012	0.012	0.008	0.008	0.005
		MSE	19.907	0.03	0.03	0.023	0.023	0.026	0.027	0.024	0.024	0.032
	Scatter	Bias	161.67	0.321	0.32	0.003	0.018	0.024	0.025	0.007	0.006	0.334
		MSE	27972.04	0.293	0.294	0.07	0.073	0.075	0.076	0.068	0.071	0.316

Table 5.9: Estimated bias and mean squared errors of exact one-step (GD, FS and IRLS) and fully converged minimum DPD location-scale estimators for contaminated normal datasets with $p = 2$.

normal set-up, datasets are generated from p-dimensional standard normal distributions of the aforesaid data dimensions and sample sizes. For the contaminated normal

Initialization	Parameter	Accuracy	MLE	GD		IRLS		FS		MDPDE		Initial
				β		β		β		β		Estimators
		Measures		0.05	0.1	0.05	0.1	0.05	0.1	0.05	0.1	
MCD	Location	Bias	0.004	0.007	0.007	0.004	0.004	0.013	0.013	0.004	0.004	0.005
		MSE	0.03	0.033	0.033	0.03	0.031	0.031	0.031	0.03	0.031	0.033
	Scatter	Bias	0.02	0.051	0.05	0.017	0.016	0.061	0.061	0.018	0.018	0.05
		MSE	0.21	0.274	0.276	0.214	0.223	0.212	0.219	0.213	0.22	0.281
MVE	Location	Bias	0.004	0.006	0.006	0.004	0.004	0.015	0.016	0.004	0.004	0.003
		MSE	0.03	0.033	0.033	0.03	0.031	0.033	0.033	0.03	0.031	0.034
	Scatter	Bias	0.02	0.03	0.031	0.019	0.018	0.066	0.068	0.018	0.018	0.035
		MSE	0.21	0.285	0.286	0.214	0.223	0.219	0.227	0.213	0.22	0.29
GK	Location	Bias	0.004	0.005	0.005	0.004	0.004	0.011	0.012	0.004	0.004	0.005
		MSE	0.03	0.036	0.037	0.03	0.032	0.033	0.034	0.03	0.031	0.037
	Scatter	Bias	0.02	0.335	0.335	0.036	0.045	0.055	0.06	0.018	0.018	0.351
		MSE	0.21	0.33	0.33	0.213	0.224	0.228	0.233	0.213	0.22	0.354
MM	Location	Bias	0.004	0.005	0.006	0.005	0.005	0.018	0.018	0.004	0.004	0.005
		MSE	0.03	0.031	0.031	0.031	0.032	0.031	0.032	0.03	0.031	0.032
	Scatter	Bias	0.02	0.02	0.02	0.016	0.016	0.042	0.041	0.018	0.018	0.017
		MSE	0.21	0.225	0.224	0.212	0.219	0.206	0.212	0.213	0.22	0.22
S	Location	Bias	0.004	0.004	0.004	0.005	0.005	0.018	0.019	0.004	0.004	0.006
		MSE	0.03	0.035	0.035	0.031	0.032	0.032	0.032	0.03	0.031	0.035
	Scatter	Bias	0.02	0.018	0.018	0.018	0.018	0.044	0.042	0.018	0.018	0.016
		MSE	0.21	0.258	0.254	0.212	0.22	0.207	0.211	0.213	0.22	0.254

Table 5.10: Estimated bias and mean squared errors of exact one-step (GD, IRLS and FS) and fully converged minimum DPD location-scale estimators for pure normal datasets with $p = 6$.

Initialization	Parameter	Accuracy	MLE	GD		IRLS		FS		MDPDE		Initial
				β		β		β		β		Estimators
		Measures		0.05	0.1	0.05	0.1	0.05	0.1	0.05	0.1	
MCD	Location	Bias	7.352	0.004	0.004	0.006	0.006	0.013	0.013	0.007	0.007	0.007
		MSE	56.37	0.035	0.036	0.033	0.034	0.034	0.035	0.033	0.034	0.035
	Scatter	Bias	483.993	0.359	0.361	0.023	0.044	0.064	0.067	0.021	0.028	0.382
		MSE	242136.3	0.49	0.491	0.242	0.255	0.243	0.256	0.236	0.247	0.513
MVE	Location	Bias	7.352	0.008	0.008	0.007	0.007	0.013	0.012	0.007	0.007	0.008
		MSE	56.37	0.034	0.034	0.033	0.034	0.034	0.035	0.033	0.034	0.034
	Scatter	Bias	483.993	0.338	0.34	0.024	0.041	0.053	0.062	0.021	0.028	0.351
		MSE	242136.3	0.461	0.461	0.238	0.25	0.245	0.251	0.236	0.247	0.47
GK	Location	Bias	7.352	0.009	0.009	0.007	0.007	0.014	0.014	0.007	0.007	0.009
		MSE	56.37	0.039	0.039	0.033	0.034	0.034	0.035	0.033	0.034	0.039
	Scatter	Bias	483.993	0.276	0.279	0.027	0.024	0.053	0.057	0.021	0.028	0.286
		MSE	242136.3	0.313	0.314	0.235	0.248	0.241	0.249	0.236	0.247	0.329
MM	Location	Bias	7.352	0.005	0.005	0.005	0.005	0.017	0.017	0.007	0.007	0.005
		MSE	56.37	0.034	0.034	0.035	0.035	0.035	0.036	0.033	0.034	0.035
	Scatter	Bias	483.993	0.412	0.411	0.014	0.033	0.061	0.064	0.021	0.028	0.419
		MSE	242136.3	0.524	0.524	0.239	0.251	0.245	0.249	0.236	0.247	0.425
S	Location	Bias	7.352	0.005	0.005	0.005	0.005	0.015	0.016	0.007	0.007	0.006
		MSE	56.37	0.036	0.036	0.035	0.036	0.036	0.036	0.033	0.034	0.038
	Scatter	Bias	483.993	0.418	0.419	0.013	0.03	0.057	0.059	0.021	0.028	0.425
		MSE	242136.3	0.563	0.563	0.239	0.251	0.245	0.249	0.236	0.247	0.564

Table 5.11: Estimated bias and mean squared errors of exact one-step (GD, IRLS and FS) and fully converged minimum DPD location-scale estimators for contaminated normal datasets with $p = 6$.

set-up, we simulate again from p-dimensional standard normal distributions but they are now contaminated with observations from another normal component with mean

Initialization	Parameter	Accuracy	MLE	GD		IRLS		FS		MDPDE		Initial
				β		β		β		β		Estimators
		Measures		0.05	0.1	0.05	0.1	0.05	0.1	0.05	0.1	
MCD	Location	Bias	0.005	0.006	0.007	0.006	0.006	0.015	0.017	0.005	0.005	0.006
		MSE	0.033	0.04	0.04	0.034	0.035	0.035	0.036	0.033	0.034	0.037
	Scatter	Bias	0.022	0.073	0.077	0.019	0.019	0.069	0.075	0.02	0.021	0.071
		MSE	0.366	0.46	0.462	0.374	0.392	0.373	0.388	0.373	0.389	0.459
MVE	Location	Bias	0.005	0.012	0.014	0.005	0.006	0.02	0.021	0.005	0.005	0.007
		MSE	0.033	0.037	0.039	0.033	0.034	0.036	0.036	0.033	0.034	0.036
	Scatter	Bias	0.022	0.047	0.047	0.021	0.02	0.077	0.075	0.02	0.021	0.035
		MSE	0.366	0.432	0.434	0.374	0.392	0.381	0.392	0.373	0.389	0.454
GK	Location	Bias	0.005	0.008	0.01	0.005	0.005	0.018	0.019	0.005	0.005	0.006
		MSE	0.033	0.04	0.04	0.034	0.035	0.036	0.036	0.033	0.034	0.04
	Scatter	Bias	0.022	0.325	0.327	0.033	0.035	0.059	0.061	0.02	0.021	0.335
		MSE	0.366	0.497	0.501	0.373	0.398	0.396	0.409	0.373	0.389	0.515
MM	Location	Bias	0.005	0.014	0.016	0.005	0.006	0.019	0.019	0.005	0.005	0.006
		MSE	0.033	0.036	0.036	0.034	0.035	0.036	0.036	0.033	0.034	0.036
	Scatter	Bias	0.022	0.041	0.04	0.022	0.021	0.055	0.057	0.02	0.021	0.021
		MSE	0.366	0.392	0.395	0.373	0.389	0.375	0.395	0.373	0.389	0.386
S	Location	Bias	0.005	0.008	0.008	0.005	0.006	0.019	0.02	0.005	0.005	0.006
		MSE	0.033	0.036	0.037	0.034	0.035	0.036	0.037	0.033	0.034	0.036
	Scatter	Bias	0.022	0.032	0.036	0.022	0.022	0.058	0.065	0.02	0.021	0.021
		MSE	0.366	0.401	0.403	0.373	0.39	0.378	0.391	0.373	0.389	0.403

Table 5.12: Estimated bias and mean squared errors of exact one-step (GD, FS and IRLS) and fully converged minimum DPD location-scale estimators for pure normal datasets with $p = 10$.

Initialization	Parameter	Accuracy	MLE	GD		IRLS		FS		MDPDE		Initial
				β		β		β		β		Estimators
		Measures		0.05	0.1	0.05	0.1	0.05	0.1	0.05	0.1	
MCD	Location	Bias	9.499	0.008	0.008	0.005	0.005	0.014	0.016	0.004	0.004	0.005
		MSE	93.146	0.038	0.038	0.038	0.039	0.039	0.039	0.037	0.038	0.039
	Scatter	Bias	808.227	0.355	0.256	0.029	0.049	0.091	0.096	0.022	0.036	0.373
		MSE	669715.2	0.704	0.705	0.423	0.446	0.451	0.465	0.416	0.44	0.714
MVE	Location	Bias	9.499	0.011	0.012	0.004	0.004	0.017	0.018	0.004	0.004	0.005
		MSE	93.146	0.039	0.039	0.037	0.038	0.035	0.036	0.037	0.038	0.039
	Scatter	Bias	808.227	0.332	0.334	0.027	0.048	0.075	0.086	0.022	0.036	0.33
		MSE	669715.2	0.641	0.643	0.418	0.44	0.415	0.442	0.416	0.44	0.652
GK	Location	Bias	9.499	0.012	0.014	0.004	0.004	0.018	0.019	0.004	0.004	0.008
		MSE	93.146	0.043	0.044	0.037	0.039	0.036	0.036	0.037	0.038	0.043
	Scatter	Bias	808.227	0.269	0.269	0.02	0.024	0.075	0.083	0.022	0.036	0.271
		MSE	669715.2	0.476	0.478	0.415	0.443	0.424	0.434	0.416	0.44	0.501
MM	Location	Bias	9.499	0.012	0.013	0.006	0.006	0.019	0.021	0.004	0.004	0.006
		MSE	93.146	0.038	0.039	0.038	0.039	0.036	0.038	0.037	0.038	0.039
	Scatter	Bias	808.227	0.524	0.525	0.033	0.054	0.079	0.092	0.022	0.036	0.539
		MSE	669715.2	0.861	0.872	0.424	0.445	0.443	0.462	0.416	0.44	0.89
S	Location	Bias	9.499	0.009	0.011	0.006	0.006	0.017	0.019	0.004	0.004	0.006
		MSE	93.146	0.037	0.037	0.038	0.039	0.037	0.038	0.037	0.038	0.04
	Scatter	Bias	808.227	0.537	0.541	0.022	0.032	0.077	0.087	0.022	0.036	0.541
		MSE	669715.2	0.893	0.897	0.414	0.426	0.443	0.455	0.416	0.44	0.908

Table 5.13: Estimated bias and mean squared errors of exact one-step (GD, IRLS and FS) and fully converged minimum DPD location-scale estimators for contaminated normal datasets with $p = 10$.

$(30, 30, \ldots, 30)'$ and covariance matrix $\boldsymbol{I}_p$; the contamination proportion being 10%. For the multivariate t set-up, datasets are simulated from t-distribution with 5 degrees

Initialization	Parameter	Accuracy Measures	MLE	Density Power Weights $\beta = 0.05$	$\beta = 0.1$	$\beta = 0.3$	$0-1$ Weights	Initial Estimators
MCD	Location	Bias	0.003	0.008	0.008	0.008	0.007	0.007
		MSE	0.02	0.02	0.02	0.023	0.023	0.024
	Scatter	Bias	0.024	0.012	0.011	0.005	0.009	0.026
		MSE	0.057	0.061	0.063	0.073	0.082	0.094
MVE	Location	Bias	0.003	0.003	0.003	0.003	0.004	0.005
		MSE	0.02	0.021	0.022	0.024	0.024	0.025
	Scatter	Bias	0.024	0.021	0.020	0.015	0.021	0.013
		MSE	0.057	0.061	0.063	0.074	0.085	0.095
GK	Location	Bias	0.003	0.003	0.003	0.003	0.003	0.006
		MSE	0.02	0.022	0.022	0.026	0.029	0.028
	Scatter	Bias	0.024	0.048	0.07	0.132	0.201	0.377
		MSE	0.057	0.062	0.066	0.091	0.131	0.203
MM	Location	Bias	0.003	0.003	0.003	0.003	0.008	0.002
		MSE	0.02	0.02	0.021	0.022	0.022	0.021
	Scatter	Bias	0.024	0.01	0.01	0.008	0.01	0.032
		MSE	0.057	0.061	0.062	0.071	0.081	0.069
S	Location	Bias	0.003	0.003	0.003	0.004	0.004	0.003
		MSE	0.02	0.021	0.021	0.025	0.024	0.035
	Scatter	Bias	0.024	0.014	0.017	0.022	0.042	0.014
		MSE	0.057	0.061	0.064	0.082	0.092	0.137

Table 5.14: Estimated bias and mean squared errors of different weighted location-scale estimators for pure normal datasets with data dimension $p = 2$.

Initialization	Parameter	Accuracy Measures	MLE	Density Power Weights $\beta = 0.05$	$\beta = 0.1$	$\beta = 0.3$	$0-1$ Weights	Initial Estimators
MCD	Location	Bias	4.275	0.003	0.003	0.002	0.003	0.001
		MSE	19.907	0.023	0.023	0.024	0.024	0.025
	Scale	Bias	161.670	0.002	0.018	0.073	0.061	0.438
		MSE	27972.040	0.067	0.070	0.090	0.093	0.360
MVE	Location	Bias	4.275	0.007	0.007	0.007	0.006	0.009
		MSE	19.907	0.021	0.021	0.022	0.022	0.025
	Scale	Bias	161.670	0.005	0.017	0.072	0.068	0.445
		MSE	27972.040	0.069	0.072	0.091	0.094	0.363
GK	Location	Bias	4.275	0.007	0.008	0.010	0.012	0.008
		MSE	19.907	0.021	0.022	0.025	0.027	0.030
	Scale	Bias	161.670	0.037	0.054	0.101	0.149	0.294
		MSE	27972.040	0.068	0.072	0.094	0.124	0.162
MM	Location	Bias	4.275	0.002	0.002	0.003	0.004	0.003
		MSE	19.907	0.023	0.023	0.025	0.024	0.023
	Scale	Bias	161.670	0.005	0.015	0.056	0.053	0.313
		MSE	27972.040	0.069	0.072	0.087	0.090	0.230
S	Location	Bias	4.275	0.002	0.002	0.004	0.003	0.005
		MSE	19.907	0.023	0.024	0.026	0.025	0.032
	Scale	Bias	161.670	0.005	0.010	0.044	0.035	0.334
		MSE	27972.040	0.069	0.072	0.093	0.095	0.316

Table 5.15: Estimated bias and mean squared errors of different weighted location-scale estimators for contaminated normal datasets with data dimension $p = 2$.

of freedom of dimensions 2, 6 and 10 with respective sample sizes $n = 100$, 200 and 300, location vector $\mathbf{0}_p$ and scale matrix $\boldsymbol{I}_p$ (p is the data dimension).

Two accuracy measures are used to assess the performances of the aforesaid estimators, namely the L_2 bias and mean squared error. The empirical estimates of the bias

Initialization (p)	Parameter	Accuracy Measures	MLE	Density Power Weights $\beta = 0.05$	$\beta = 0.1$	$\beta = 0.3$	$0-1$ Weights	Initial Estimators
MCD	Location	Bias	0.004	0.005	0.005	0.007	0.005	0.005
		MSE	0.03	0.03	0.031	0.035	0.032	0.033
	Scatter	Bias	0.02	0.017	0.016	0.016	0.025	0.05
		MSE	0.21	0.213	0.223	0.274	0.25	0.281
MVE	Location	Bias	0.004	0.006	0.006	0.006	0.006	0.004
		MSE	0.03	0.030	0.030	0.034	0.031	0.034
	Scatter	Bias	0.02	0.014	0.014	0.018	0.019	0.035
		MSE	0.21	0.214	0.222	0.273	0.249	0.290
GK	Location	Bias	0.004	0.006	0.006	0.006	0.006	0.005
		MSE	0.03	0.03	0.031	0.037	0.034	0.037
	Scatter	Bias	0.02	0.032	0.049	0.096	0.147	0.351
		MSE	0.21	0.213	0.223	0.285	0.288	0.354
MM	Location	Bias	0.004	0.005	0.005	0.007	0.006	0.005
		MSE	0.03	0.03	0.031	0.035	0.032	0.032
	Scatter	Bias	0.02	0.018	0.017	0.016	0.019	0.017
		MSE	0.21	0.211	0.217	0.257	0.242	0.22
S	Location	Bias	0.004	0.005	0.005	0.007	0.006	0.006
		MSE	0.03	0.03	0.031	0.036	0.032	0.035
	Scatter	Bias	0.02	0.019	0.02	0.019	0.028	0.016
		MSE	0.21	0.212	0.219	0.268	0.247	0.254

Table 5.16: Estimated bias and mean squared errors of different weighted location-scale estimators for pure normal datasets with data dimension $p = 6$.

Initialization	Parameter	Accuracy Measures	MLE	Density Power Weights $\beta = 0.05$	$\beta = 0.1$	$\beta = 0.3$	$0-1$ Weights	Initial Estimators
MCD	Location	Bias	7.352	0.004	0.004	0.005	0.004	0.007
		MSE	56.370	0.034	0.034	0.038	0.035	0.035
	Scale	Bias	483.993	0.014	0.018	0.069	0.026	0.382
		MSE	242136.300	0.238	0.248	0.307	0.269	0.513
MVE	Location	Bias	7.352	0.002	0.002	0.002	0.003	0.008
		MSE	56.370	0.033	0.034	0.037	0.034	0.034
	Scale	Bias	483.993	0.016	0.024	0.075	0.041	0.351
		MSE	242136.300	0.238	0.247	0.303	0.267	0.470
GK	Location	Bias	7.352	0.002	0.002	0.002	0.004	0.009
		MSE	56.370	0.033	0.035	0.041	0.038	0.039
	Scale	Bias	483.993	0.036	0.048	0.08	0.123	0.286
		MSE	242136.300	0.234	0.245	0.305	0.300	0.329
MM	Location	Bias	7.352	0.003	0.003	0.003	0.005	0.005
		MSE	56.370	0.034	0.035	0.039	0.034	0.035
	Scale	Bias	483.993	0.015	0.028	0.089	0.060	0.419
		MSE	242136.300	0.238	0.247	0.305	0.270	0.525
S	Location	Bias	7.352	0.003	0.003	0.003	0.004	0.006
		MSE	56.370	0.034	0.035	0.040	0.035	0.038
	Scale	Bias	483.993	0.015	0.026	0.084	0.046	0.425
		MSE	242136.300	0.238	0.248	0.313	0.267	0.564

Table 5.17: Estimated bias and mean squared errors of different weighted location-scale estimators for contaminated normal datasets with data dimension $p = 6$.

and mean squared errors of the estimators can be obtained in the following way. Let, $\widehat{\boldsymbol{\mu}}_i$ and $\widehat{\boldsymbol{\Sigma}}_i$ be the estimators of the location vector and the scale matrix from the i-th replication, $1 \leq i \leq 1000$. We calculate the L_2 norm of the bias and mean squared errors as: bias of location estimators $= \left|\left| \frac{1}{1000} \sum_{i=1}^{1000} \widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu} \right|\right|_2$, MSE of location estimators $=$

Initialization	Parameter	Accuracy Measures	MLE	Density Power Weights $\beta = 0.05$	$\beta = 0.1$	$\beta = 0.3$	$0-1$ Weights	Initial Estimators
MCD	Location	Bias	0.005	0.005	0.005	0.006	0.005	0.006
		MSE	0.033	0.034	0.035	0.042	0.036	0.037
	Scatter	Bias	0.022	0.018	0.017	0.028	0.028	0.071
		MSE	0.366	0.374	0.392	0.502	0.416	0.459
MVE	Location	Bias	0.005	0.008	0.007	0.008	0.008	0.007
		MSE	0.033	0.034	0.036	0.043	0.036	0.036
	Scatter	Bias	0.022	0.018	0.017	0.023	0.026	0.035
		MSE	0.366	0.375	0.393	0.502	0.419	0.454
GK	Location	Bias	0.005	0.008	0.007	0.008	0.008	0.006
		MSE	0.033	0.035	0.036	0.046	0.039	0.04
	Scatter	Bias	0.022	0.034	0.048	0.082	0.143	0.335
		MSE	0.366	0.374	0.396	0.524	0.462	0.515
MM	Location	Bias	0.005	0.006	0.006	0.007	0.006	0.006
		MSE	0.033	0.034	0.035	0.043	0.036	0.036
	Scatter	Bias	0.022	0.024	0.023	0.024	0.026	0.021
		MSE	0.366	0.372	0.388	0.49	0.411	0.386
S	Location	Bias	0.005	0.006	0.006	0.007	0.006	0.006
		MSE	0.033	0.034	0.036	0.043	0.036	0.036
	Scatter	Bias	0.022	0.024	0.024	0.024	0.028	0.021
		MSE	0.366	0.373	0.39	0.497	0.413	0.403

Table 5.18: Estimated bias and mean squared errors of different weighted location-scale estimators for pure normal datasets with data dimension $p = 10$.

Initialization	Parameter	Accuracy Measures	MLE	Density Power Weights $\beta = 0.05$	$\beta = 0.1$	$\beta = 0.3$	$0-1$ Weights	Initial Estimators
MCD	Location	Bias	9.499	0.004	0.006	0.006	0.007	0.005
		MSE	93.146	0.037	0.038	0.045	0.038	0.039
	Scale	Bias	808.227	0.019	0.028	0.085	0.028	0.373
		MSE	669715.200	0.421	0.442	0.563	0.457	0.714
MVE	Location	Bias	9.499	0.006	0.006	0.005	0.007	0.005
		MSE	93.146	0.038	0.039	0.046	0.039	0.039
	Scale	Bias	808.227	0.023	0.032	0.082	0.042	0.330
		MSE	669715.200	0.414	0.433	0.547	0.448	0.652
GK	Location	Bias	9.499	0.006	0.006	0.006	0.007	0.008
		MSE	93.146	0.038	0.04	0.05	0.043	0.043
	Scale	Bias	808.227	0.034	0.045	0.07	0.115	0.271
		MSE	669715.200	0.409	0.431	0.561	0.484	0.500
MM	Location	Bias	9.499	0.005	0.004	0.004	0.010	0.006
		MSE	93.146	0.038	0.038	0.045	0.038	0.039
	Scale	Bias	808.227	0.028	0.047	0.122	0.062	0.539
		MSE	669715.2	0.419	0.439	0.566	0.453	0.890
S	Location	Bias	9.499	0.005	0.004	0.004	0.005	0.006
		MSE	93.146	0.038	0.039	0.045	0.038	0.040
	Scale	Bias	808.227	0.027	0.045	0.120	0.063	0.541
		MSE	669715.2	0.419	0.440	0.571	0.450	0.908

Table 5.19: Estimated bias and mean squared errors of different weighted location-scale estimators for contaminated normal datasets with data dimension $p = 10$.

$\frac{1}{1000}\sum_{i=1}^{1000}\left|\left|\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}\right|\right|_2^2$, bias of scale estimators $= \left|\left|\frac{1}{1000}\sum_{i=1}^{1000}\widehat{\boldsymbol{\Sigma}}_i - \boldsymbol{\Sigma}\right|\right|_F$, MSE of scale estimators $= \frac{1}{1000}\sum_{i=1}^{1000}\left|\left|\widehat{\boldsymbol{\Sigma}}_i - \boldsymbol{\Sigma}\right|\right|_F^2$, where $||\cdot||_2$ is the L_2-norm and $||\cdot||_F$ is the Frobenius norm of a matrix. The estimated bias and mean squared errors

Initialization (p)	Parameter	Accuracy Measures	Density Power Weights $\beta = 0.05$	$\beta = 0.1$	$\beta = 0.3$	$\beta = 0.5$	$0-1$ Weights	Initial Estimators
MCD	Location	Bias	0.001	0.002	0.002	0.003	0.002	0.005
		MSE	0.029	0.027	0.026	0.027	0.031	0.031
	Scatter	Bias	0.017	0.032	0.076	0.1	0.143	0.269
		MSE	0.139	0.109	0.092	0.104	0.134	0.156
MVE	Location	Bias	0.007	0.006	0.005	0.005	0.005	0.007
		MSE	0.03	0.028	0.027	0.028	0.032	0.032
	Scatter	Bias	0.015	0.033	0.076	0.099	0.139	0.249
		MSE	0.139	0.107	0.091	0.103	0.131	0.149
GK	Location	Bias	0.007	0.006	0.005	0.006	0.005	0.006
		MSE	0.029	0.027	0.028	0.03	0.036	0.034
	Scatter	Bias	0.053	0.106	0.251	0.339	0.494	0.735
		MSE	0.13	0.102	0.132	0.188	0.327	0.574
MM	Location	Bias	0.007	0.007	0.006	0.006	0.01	0.005
		MSE	0.029	0.027	0.025	0.026	0.03	0.029
	Scatter	Bias	0.007	0.034	0.1	0.141	0.214	0.387
		MSE	0.155	0.108	0.09	0.102	0.14	0.2
S	Location	Bias	0.007	0.007	0.007	0.007	0.012	0.005
		MSE	0.029	0.026	0.026	0.029	0.033	0.035
	Scatter	Bias	0.01	0.04	0.113	0.154	0.232	0.372
		MSE	0.153	0.106	0.096	0.12	0.171	0.222

Table 5.20: Estimated bias and mean squared errors of different one-step location-scale estimators for multivariate t-datasets with data dimension $p = 2$.

Dimension (p)	Parameter	Accuracy Measures	Density Power Weights $\beta = 0.05$	$\beta = 0.1$	$\beta = 0.3$	$\beta = 0.5$	$0-1$ Weights	Initial Estimators
MCD	Location	Bias	0.007	0.005	0.007	0.008	0.008	0.005
		MSE	0.041	0.038	0.039	0.044	0.046	0.041
	Scatter	Bias	0.022	0.052	0.148	0.201	0.318	0.596
		MSE	0.441	0.323	0.314	0.391	0.461	0.579
MVE	Location	Bias	0.008	0.008	0.007	0.007	0.006	0.004
		MSE	0.041	0.038	0.038	0.043	0.046	0.043
	Scatter	Bias	0.034	0.032	0.112	0.159	0.268	0.516
		MSE	0.432	0.316	0.298	0.363	0.417	0.505
GK	Location	Bias	0.008	0.007	0.007	0.007	0.007	0.005
		MSE	0.04	0.037	0.041	0.048	0.052	0.045
	Scatter	Bias	0.027	0.117	0.328	0.46	0.743	1.16
		MSE	0.407	0.301	0.364	0.518	0.826	1.458
MM	Location	Bias	0.004	0.004	0.005	0.006	0.007	0.006
		MSE	0.04	0.036	0.036	0.041	0.043	0.037
	Scatter	Bias	0.041	0.026	0.128	0.19	0.325	0.594
		MSE	0.435	0.321	0.29	0.355	0.425	0.517
S	Location	Bias	0.004	0.004	0.005	0.006	0.006	0.006
		MSE	0.04	0.036	0.037	0.043	0.044	0.038
	Scatter	Bias	0.037	0.033	0.139	0.193	0.321	0.588
		MSE	0.43	0.317	0.305	0.386	0.446	0.533

Table 5.21: Estimated bias and mean squared errors of different one-step location-scale estimators for multivariate t-datasets with data dimension $p = 6$.

of the location and scale exact one-step gradient descent, IRLS and Fisher's scoring estimators are tabulated in Tables 5.8-5.13. Those of the generalized weighted (with the density power weights, and traditional $0-1$ weights) estimators and the initial estimators are tabulated in Tables 5.14-5.19 for pure and contaminated multivariate

Dimension (p)	Parameter	Accuracy Measures	Density Power Weights $\beta = 0.05$	$\beta = 0.1$	$\beta = 0.3$	$\beta = 0.5$	$0-1$ Weights	Initial Estimators
MCD	Location	Bias	0.004	0.004	0.005	0.006	0.005	0.007
		MSE	0.041	0.038	0.045	0.061	0.05	0.047
	Scatter	Bias	0.133	0.027	0.194	0.297	0.496	0.88
		MSE	0.813	0.554	0.605	0.92	0.836	1.125
MVE	Location	Bias	0.006	0.007	0.009	0.011	0.007	0.007
		MSE	0.041	0.038	0.043	0.057	0.049	0.046
	Scatter	Bias	0.127	0.022	0.157	0.259	0.421	0.725
		MSE	0.805	0.552	0.554	0.829	0.748	0.894
GK	Location	Bias	0.006	0.007	0.009	0.012	0.008	0.005
		MSE	0.041	0.038	0.048	0.069	0.057	0.047
	Scatter	Bias	0.072	0.11	0.401	0.622	0.991	1.482
		MSE	0.756	0.523	0.681	1.171	1.461	2.386
MM	Location	Bias	0.002	0.002	0.005	0.008	0.006	0.005
		MSE	0.042	0.039	0.044	0.059	0.049	0.04
	Scatter	Bias	0.134	0.026	0.161	0.254	0.42	0.735
		MSE	0.833	0.56	0.551	0.831	0.738	0.82
S	Location	Bias	0.002	0.003	0.005	0.008	0.006	0.005
		MSE	0.042	0.038	0.045	0.06	0.05	0.04
	Scatter	Bias	0.13	0.027	0.164	0.251	0.416	0.732
		MSE	0.828	0.557	0.564	0.852	0.747	0.828

Table 5.22: Estimated bias and mean squared errors of different one-step location-scale estimators for multivariate t-datasets with data dimension $p = 10$.

normal set-ups and in Tables 5.20-5.22 for the multivariate t-set-up.

We delineate the critical observations from the aforesaid simulation results in the following points.

1. The simulation outputs in case of pure normal datasets for the maximum likelihood, exact one-step (GD, IRLS and FS), fully converged minimum DPD and the initial estimators are tabulated in Tables 5.8, 5.10, and 5.12, while those of the generalized weight based (density power and $0-1$ weights) and the initial estimators are tabulated in Tables 5.14, 5.16 and 5.18. As expected from the angle of statistical efficiency, the maximum likelihood estimators of location and scale become the best among all the estimators in terms of the estimated bias and mean squared errors. Additionally, the exact one-step (especially the IRLS based), density power weight based and fully converged minimum DPD estimators are found to have less bias and mean squared errors as compared to both the initial as well as the $0-1$ weight based one-step estimators.

2. The simulation outputs in case of contaminated normal datasets for the maximum likelihood, exact one-step (GD, IRLS and FS), fully converged minimum DPD and the initial estimators are tabulated in Tables 5.9, 5.11 and 5.13 while those of the generalized weight based (density power and $0-1$ weights) and the

initial estimators are tabulated in Tables 5.15, 5.17 and 5.19. The maximum likelihood estimators perform in the poorest manner following its non-robust nature while all of the remaining estimators are found to be quite resistant towards contamination. All of the exact one-step and generalized weight based (density power and $0-1$ weights) estimators are found to improve the initial estimators in terms of the estimated bias and mean squared errors; the improvement being significant especially in case of scale (covariance matrix in case of multivariate normality) estimators.

3. The exact-one step gradient descent estimators are found to perform similarly to the respective initializations, whereas, the exact one-step Fisher's scoring estimators are a little more biased (although the estimated mean squared errors are comparable with exact one-step IRLS and fully converged minimum DPD estimators). Thus, it can be concluded that the exact one-step IRLS estimators have a superior performance as compared to the remaining exact one-step (GD and FS) estimators.

4. Observing the outputs of both pure and contaminated normal models, it can be aggregatively concluded that the exact one-step IRLS and the density power weight based estimators have similar precisions and they can improve either of the initial or the $0-1$ weight based and the remaining exact one-step estimators in terms of bias and mean squared errors.

5. The simulation outputs in case of the multivariate t datasets can be found in Tables 5.20, 5.21 and 5.22. Similar patterns can be observed in case of multivariate t simulation set-ups, i.e., our density power weight based method can achieve less bias and mean squared errors compared to both initial and $0-1$ weight based one-step estimators for almost all the initializations with low β values (0.1, in particular).

5.5 Real Data Examples

We now illustrate our proposed methodologies through modelling three real life datasets. In particular, we model two univariate datasets (viz., the mice lifetime dataset and breaking strength dataset) through the exact one-step methodologies and one multivariate dataset (viz., the heart failure dataset) through the generalized density power

weight based and traditional $0-1$ weight based methodologies.

5.5.1 Mice Lifetime Data

The mice lifetime dataset contains information about 38 lifetimes (in number of days) of male mice who received radiation dose of 300 rads at the age of $5-6$ weeks. The dataset was originally published in Hoel (1972) [73] and republished in Kalbfleisch and Prentice (1980) [83]. We are working on this dataset as reported in Boudt et al. (2011) [16] who analysed the same using a suitable Weibull distribution. The histogram of the dataset is presented in Figure 5.3. It is evident from the same that there are three distant observations, viz., 317, 318 and 337 towards the left tail of the dataset. We utilize the exact one-step Newton-Raphson and gradient descent procedures assuming a Weibull model and a normal model, separately.

The exact one-step Newton-Raphson (with $\beta = 0.1$) and gradient descent estimators (with $\beta = 0.3$) of the unknown parameters k and λ are derived (with the initialization described in (5.19)) assuming a Weibull model to fit this dataset. The fitted density curves with the aforesaid estimators are presented in Figure 5.3. Although these fitted densities can capture the right tail of the data accurately, the left tail could not be properly fitted because of the distant observations. Consequently, the fitted densities also deviate from the histogram around the mode. If we ignore the distant observations from the data, the remaining observations, however, obey an approximate bell-shaped pattern. Following this observation, we also fit these data with the exact one-step Fisher's scoring estimators (with $\beta = 0.3$) of μ and σ^2 based on a normal model. This density (presented in Figure 5.3) is found to fit the histogram more accurately (in comparison with the Weibull fits) by downweighting the distant observations towards the left tail.

5.5.2 Breaking Strength Data

The breaking strength data consist of the information about the breaking strengths of 64 single carbon fibres (SCF, tested under tension at gauge length of 10 mm). The dataset was originally published in Barber and Priest (1982) [6]. Aydin et al. (2018) [5] later worked on this dataset using a reparametrized version (location-scale) of the shifted Gompertz distribution. The histogram of the original dataset is presented in the left panel of Figure 5.4 which presents the positively skewed nature of this dataset. Following this, we fit the usual shifted Gompertz model with our exact one-step esti-

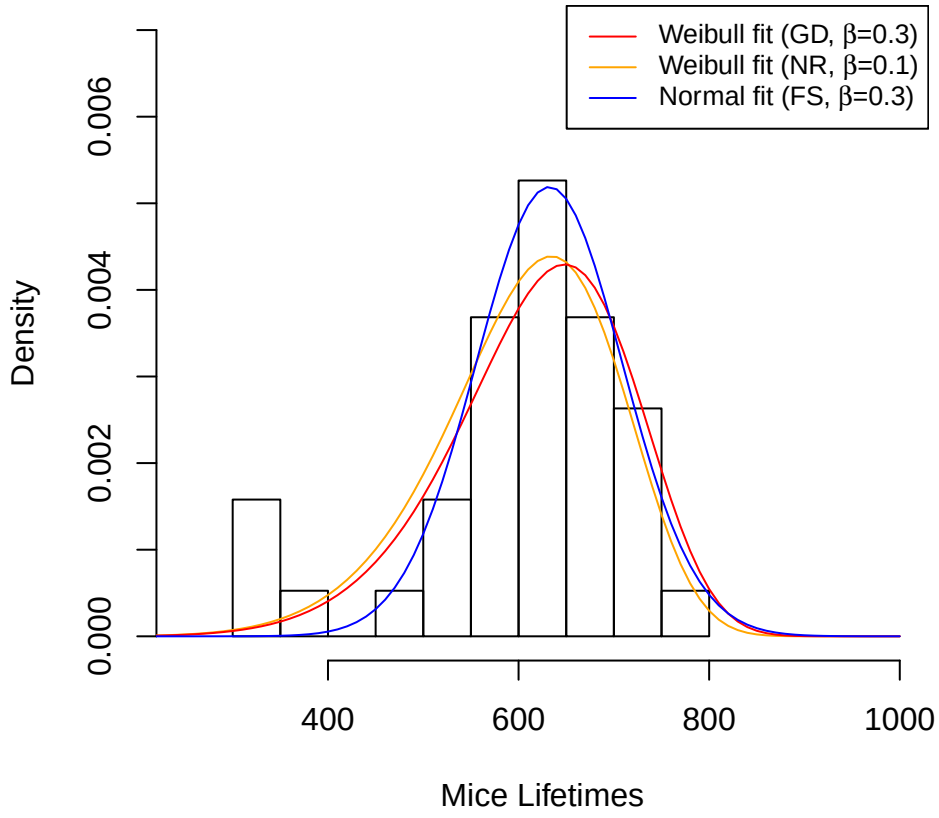
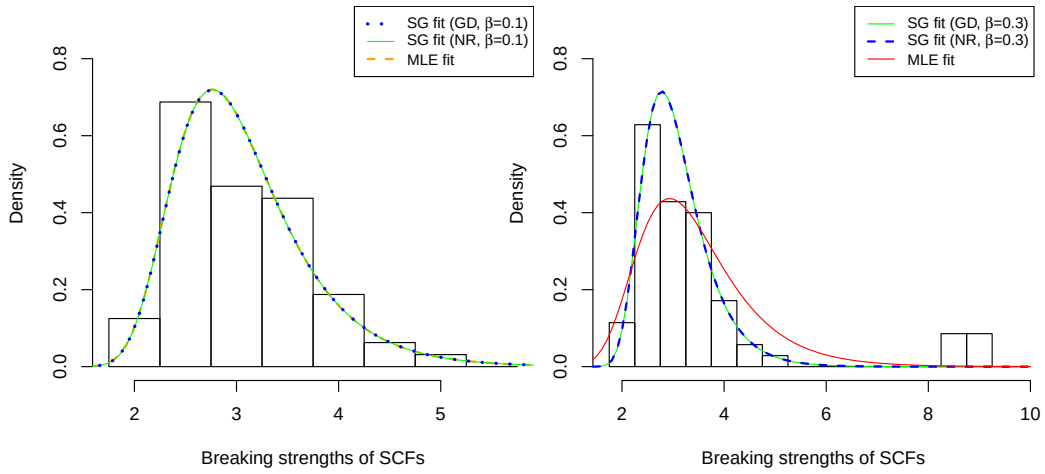


Figure 5.3: Different fits on the mice lifetime data.

mators and the usual maximum likelihood estimators on this dataset. It is also to be noted that, there is no potentially outlying observation in this dataset. To understand the robustness of our exact one-step estimators, we thus artificially contaminate the dataset with six more observations (viz., 8.461, 8.258, 8.979, 8.409, 9.128 and 8.857) and fit the shifted Gompertz model based on the exact one-step estimators and the usual maximum likelihood estimators again for this artificially contaminated dataset. For the original dataset, we first determine the maximum likelihood estimators of these parameters using the "optim" function in R software (R core team (2018) [123]). These estimators will then be used as initial estimators for deriving our exact one-step estimators for both the original and the artificially contaminated datasets and for computing the maximum likelihood estimators in case of the artificially contaminated dataset.



(a) Original data (b) Artificially contaminated data

Figure 5.4: Different density fits of the breaking strength data.

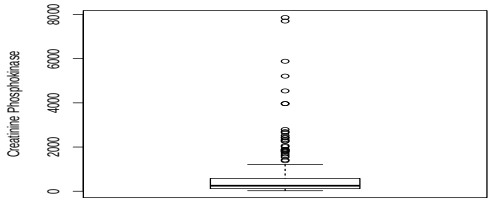
For the original data (presented in the left panel of Figure 5.4), the maximum likelihood fit and the exact one-step fits (both Newton-Raphson and gradient descent with $\beta = 0.1$) coincide with each other and each one of them can capture the shape of the histogram almost accurately. However, the maximum likelihood fit (right panel of Figure 5.4) in case of the artificially contaminated data is influenced by the contamination and consequently produces a longer right tail towards the contaminating observations. But our exact one-step estimators (both Newton-Raphson and gradient descent with $\beta = 0.3$) can capture the actual shape of the histogram successfully by effectively downweighting the outlying observations.

5.5.3 Application to Prediction on Survival of Patients with Heart Failure

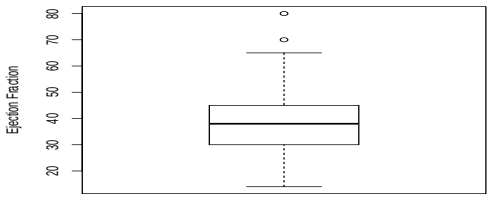
Here we illustrate our generalized weight based (density power and $0-1$ weights based) one-step methods with a multivariate real data example through predicting the survival possibilities of patients with heart failure based on certain clinical attributes, namely, creatinine phosphokinase, ejection fraction, platelets, serum creatinine and serum sodium. The original dataset[1] [26] consists of some more attributes, such as age, smoking habit, anaemic, diabetic and blood pressure status, gender and follow-up

[1]Source: https://archive.ics.uci.edu/dataset/519/heart+failure+clinical+records

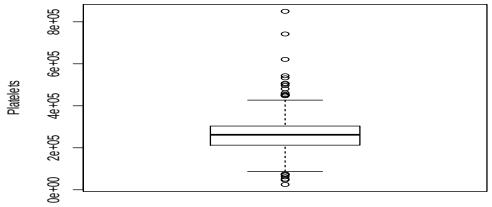
times of 299 patients. However, we have not used these attributes as our methods deal with continuous probability models. Let us first visualize the data through box plots of individual clinical attributes in Figure 5.5. The existence of outliers is evident from the box plots in Figure 5.5 which suggests the need for robust classification tools to analyse the dataset.



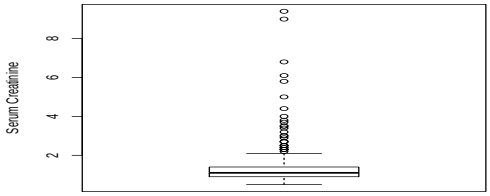
(a) Creatinine Phosphokinase



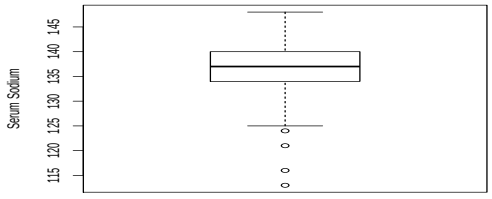
(b) Ejection Fraction



(c) Platelets



(d) Serum Creatinine



(e) Serum Sodium

Figure 5.5: Box plots of individual attributes.

Our goal is to predict whether a patient who have suffered from heart failure will survive by assessing their aforementioned five continuous clinical components. To do that, we utilize the standard Bayes classification rule with the generalized one-step estimators and the initial robust estimators of location and scale.

Precisely, we consider 100 random splits of the data into training and test subsets with 80% training and 20% test weights, respectively. Assuming joint multivariate

normality of the continuous clinical attributes, the unknown parameters of the two classes (who faced heart failures and who have not) are estimated using the generalized one-step and the robust initial estimation procedures (along with the non-robust maximum likelihood estimation) from the training subsets. These estimators are then utilized to construct the Bayes classification rule and apply it on the test subset observations and the estimated misclassification rates are computed. The average estimated misclassification rates are then obtained from the aforesaid 100 different random splits of the dataset. However, Chicco and Jurman (2020) [26] analysed this dataset and it was found that two of the clinical attributes, namely, the ejection fraction and the serum creatinine can predict the survival of patients with heart failure more accurately than the full set of covariates. Thus, we also develop the classification rule using only these two clinical features and observed the same, that is, an enhanced accuracy in classification (decreased misclassification rates). The estimated misclassification rates are presented in Table 5.23. The classification rule constructed by the density power weight based one-step estimators (with MVE initialization) are found to be the most accurate one in predicting the survival of patients with heart failure using only the ejection fraction and the serum creatinine.

Estimators	One-step Weights	β-values	Estimated Misclassification Rates	
			All the five attributes	Ejection Fraction and Serum Creatinine
MLE			0.302	0.291
MCD	Initial		0.341	0.260
	Density Power	0.1	0.295	0.243
	$0-1$		0.296	0.255
MVE	Initial		0.271	0.251
	Density Power	0.3	0.307	0.238
	$0-1$		0.298	0.253
S	Initial		0.277	0.244
	Density Power	0.1	0.292	0.245
	$0-1$		0.305	0.253

Table 5.23: Estimated misclassification rates using all the five clinical attributes as well as only ejection fraction and serum creatinine.

5.6 Appendices

5.6.1 Proof of Theorem 5.1

Proof. The proof is based on the Taylor series expansion of the individual components of the vector $\nabla\, \overline{D}_\beta(\hat{\boldsymbol{\theta}}_0)$ around the true parameter value $\boldsymbol{\theta}_0$. For $i = 1, \ldots, m$, let us

consider the following Taylor series expansions:

$$\frac{\partial}{\partial\,\theta_i}\overline{D}_\beta(\boldsymbol{\theta})\Big|_{\hat{\boldsymbol{\theta}}_0} = \frac{\partial}{\partial\,\theta_i}\overline{D}_\beta(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}_0} + \sum_{j=1}^{m}(\hat{\theta}_{0j}-\theta_{0j})\frac{\partial^2}{\partial\,\theta_j\partial\,\theta_i}\overline{D}_\beta(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}_0} + \sum_{j,j'=1}^{m}(\hat{\theta}_{0j'}-\theta_{0j'})(\hat{\theta}_{0j}-\theta_{0j})\frac{\partial^3}{\partial\,\theta_{j'}\partial\,\theta_j\partial\,\theta_i}\overline{D}_\beta(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}^{*i}} \tag{5.32}$$

, where $\boldsymbol{\theta} = (\theta_1, \ldots,\ \theta_m)'$, $\hat{\boldsymbol{\theta}}_0 = (\hat{\theta}_{01}, \ldots,\ \hat{\theta}_{0m})'$ and $\boldsymbol{\theta}^{*i}$ is a point on the line segment connecting $\hat{\boldsymbol{\theta}}_0$ and $\boldsymbol{\theta}_0$. Now,

$$\begin{aligned}
&\frac{\partial}{\partial\,\theta_i}\overline{D}_\beta(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}_0} \\
&= (1+\beta)\int f^\beta_{\boldsymbol{\theta}_0}(\boldsymbol{x})\frac{\partial}{\partial\,\theta_i}f_{\boldsymbol{\theta}}(\boldsymbol{x})\Big|_{\boldsymbol{\theta}_0}\,d\boldsymbol{x} - \frac{1+\beta}{n}\sum_{k=1}^{n} f^{\beta-1}_{\boldsymbol{\theta}_0}(\boldsymbol{X}_k)\frac{\partial}{\partial\,\theta_i}f_{\boldsymbol{\theta}}(\boldsymbol{X}_k)\Big|_{\boldsymbol{\theta}_0} \\
&\xrightarrow{p} (1+\beta)\int f^\beta_{\boldsymbol{\theta}_0}(\boldsymbol{x})\frac{\partial}{\partial\,\theta_i}f_{\boldsymbol{\theta}}(\boldsymbol{x})\Big|_{\boldsymbol{\theta}_0}\,d\boldsymbol{x} - (1+\beta)E_{f_{\boldsymbol{\theta}_0}}\left[f^{\beta-1}_{\boldsymbol{\theta}_0}(\boldsymbol{X}_1)\frac{\partial}{\partial\,\theta_i}f_{\boldsymbol{\theta}}(\boldsymbol{X}_1)\Big|_{\boldsymbol{\theta}_0}\right] \\
&= (1+\beta)\int f^\beta_{\boldsymbol{\theta}_0}(\boldsymbol{x})\frac{\partial}{\partial\,\theta_i}f_{\boldsymbol{\theta}}(\boldsymbol{x})\Big|_{\boldsymbol{\theta}_0}\,d\boldsymbol{x} - (1+\beta)\int f^\beta_{\boldsymbol{\theta}_0}(\boldsymbol{x})\frac{\partial}{\partial\,\theta_i}f_{\boldsymbol{\theta}}(\boldsymbol{x})\Big|_{\boldsymbol{\theta}_0}\,d\boldsymbol{x} = 0,
\end{aligned}$$

as $n \to \infty$, where the in probability convergence of the second term follows from the weak law of large numbers (WLLN). Thus,

$$\frac{\partial}{\partial\,\theta_i}\overline{D}_\beta(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}_0} \xrightarrow{p} 0 \tag{5.33}$$

as $n \to \infty$. To handle the second term in Equation (5.32), let us observe that for $i, j \in \{1, \ldots, m\}$,

$$
\begin{aligned}
&\frac{\partial^2}{\partial\, \theta_j \partial\, \theta_i} \overline{D}_\beta(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}_0} \\
&= \beta(1+\beta) \int f_{\boldsymbol{\theta}_0}^{\beta-1}(\boldsymbol{x}) \frac{\partial}{\partial\, \theta_j} f_{\boldsymbol{\theta}}(\boldsymbol{x})\Big|_{\boldsymbol{\theta}_0} \frac{\partial}{\partial\, \theta_i} f_{\boldsymbol{\theta}}(\boldsymbol{x})\Big|_{\boldsymbol{\theta}_0} d\boldsymbol{x} + (1+\beta) \int f_{\boldsymbol{\theta}_0}^{\beta}(\boldsymbol{x}) \frac{\partial^2}{\partial\, \theta_j \partial\, \theta_i} f_{\boldsymbol{\theta}}(\boldsymbol{x})\Big|_{\boldsymbol{\theta}_0} d\boldsymbol{x} \\
&- \frac{(1+\beta)(\beta-1)}{n} \sum_{k=1}^{n} f_{\boldsymbol{\theta}_0}^{\beta-2}(\boldsymbol{X}_k) \frac{\partial}{\partial\, \theta_j} f_{\boldsymbol{\theta}}(\boldsymbol{X_k})\Big|_{\boldsymbol{\theta}_0} \frac{\partial}{\partial\, \theta_i} f_{\boldsymbol{\theta}}(\boldsymbol{X_k})\Big|_{\boldsymbol{\theta}_0} \\
&- \frac{(1+\beta)}{n} \sum_{k=1}^{n} f_{\boldsymbol{\theta}_0}^{\beta-1}(\boldsymbol{X}_k) \frac{\partial^2}{\partial\, \theta_j \partial\, \theta_i} f_{\boldsymbol{\theta}}(\boldsymbol{X}_k)\Big|_{\boldsymbol{\theta}_0} \\
&\xrightarrow{p} \beta(1+\beta) \int f_{\boldsymbol{\theta}_0}^{\beta-1}(\boldsymbol{x}) \frac{\partial}{\partial\, \theta_j} f_{\boldsymbol{\theta}}(\boldsymbol{x})\Big|_{\boldsymbol{\theta}_0} \frac{\partial}{\partial\, \theta_i} f_{\boldsymbol{\theta}}(\boldsymbol{x})\Big|_{\boldsymbol{\theta}_0} d\boldsymbol{x} + (1+\beta) \int f_{\boldsymbol{\theta}_0}^{\beta}(\boldsymbol{x}) \frac{\partial^2}{\partial\, \theta_j \partial\, \theta_i} f_{\boldsymbol{\theta}}(\boldsymbol{x})\Big|_{\boldsymbol{\theta}_0} d\boldsymbol{x} \\
&- (1+\beta)(\beta-1) \int f_{\boldsymbol{\theta}_0}^{\beta-1}(\boldsymbol{x}) \frac{\partial}{\partial\, \theta_j} f_{\boldsymbol{\theta}}(\boldsymbol{x})\Big|_{\boldsymbol{\theta}_0} \frac{\partial}{\partial\, \theta_i} f_{\boldsymbol{\theta}}(\boldsymbol{x})\Big|_{\boldsymbol{\theta}_0} d\boldsymbol{x} - (1+\beta) \int f_{\boldsymbol{\theta}_0}^{\beta}(\boldsymbol{x}) \frac{\partial^2}{\partial\, \theta_j \partial\, \theta_i} f_{\boldsymbol{\theta}}(\boldsymbol{x})\Big|_{\boldsymbol{\theta}_0} d\boldsymbol{x} \\
&= (1+\beta) \int f_{\boldsymbol{\theta}_0}^{\beta-1}(\boldsymbol{x}) \frac{\partial}{\partial\, \theta_j} f_{\boldsymbol{\theta}}(\boldsymbol{x})\Big|_{\boldsymbol{\theta}_0} \frac{\partial}{\partial\, \theta_i} f_{\boldsymbol{\theta}}(\boldsymbol{x})\Big|_{\boldsymbol{\theta}_0} d\boldsymbol{x}
\end{aligned}
$$

which is finite by Assumption 5.1. Now, the consistency of $\hat{\boldsymbol{\theta}}_0$ (to $\boldsymbol{\theta}_0$) implies

$$
\sum_{j=1}^{m} (\hat{\theta}_{0j} - \theta_{0j}) \frac{\partial^2}{\partial\, \theta_j \partial\, \theta_i} \overline{D}_\beta(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}_0} \xrightarrow{p} 0 \tag{5.34}
$$

as $n \to \infty$. The aforesaid manipulations also establish that,

$$
\nabla^2 \overline{D}_\beta(\boldsymbol{\theta}_0) \xrightarrow{p} \boldsymbol{F}, \tag{5.35}
$$

where $\boldsymbol{F} = [[f_{ij}]]_{i,j=1}^{m}$ and $f_{ij} = (1+\beta) \int f_{\boldsymbol{\theta}_0}^{\beta-1}(\boldsymbol{x}) \frac{\partial}{\partial\, \theta_j} f_{\boldsymbol{\theta}}(\boldsymbol{x})\Big|_{\boldsymbol{\theta}_0} \frac{\partial}{\partial\, \theta_i} f_{\boldsymbol{\theta}}(\boldsymbol{x})\Big|_{\boldsymbol{\theta}_0} d\boldsymbol{x}$ which is assumed to be positive definite (statement of Theorem 5.3). Now, to tackle the third term, let us observe that,

$$
\begin{aligned}
\left| \frac{\partial^3}{\partial\, \theta_{j'} \partial\, \theta_j \partial\, \theta_i} \overline{D}_\beta(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}^{*i}} \right| &\leq \frac{1}{n} \sum_{k=1}^{n} \left| \frac{\partial^3}{\partial\, \theta_{j'} \partial\, \theta_j \partial\, \theta_i} V(\boldsymbol{X}_k, \boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{*i}} \right| \\
&\leq \frac{1}{n} \sum_{k=1}^{n} U_{j'ji}(\boldsymbol{X}_k) \xrightarrow{p} E_{f_{\boldsymbol{\theta}_0}}(U_{j'ji}(\boldsymbol{X}_k)) \\
&< \infty
\end{aligned} \tag{5.36}
$$

as a consequence of the weak law of large numbers and Assumption 5.4. While applying Assumption 5.4, we assume a sufficiently large sample size n, so that, $\hat{\boldsymbol{\theta}}_0$ (and thus $\boldsymbol{\theta}^{*i}$) is contained in $B_\epsilon(\boldsymbol{\theta}_0)$ for some $\epsilon > 0$ with high probability as a consequence of the consistency of $\hat{\boldsymbol{\theta}}_0$. The consistency of $\hat{\boldsymbol{\theta}}_0$ (to $\boldsymbol{\theta}_0$) now implies

$$\sum_{j=1}^{m}(\hat{\theta}_{0j} - \theta_{0j})\frac{\partial^3}{\partial\, \theta_{j'}\partial\, \theta_j\partial\, \theta_i}\overline{D}_\beta(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}^{*i}} \xrightarrow{p} 0 \tag{5.37}$$

as $n \to \infty$. The in probability convergences in (5.33), (5.34) and (5.37) imply $\frac{\partial}{\partial\, \theta_i}\overline{D}_\beta(\boldsymbol{\theta})\Big|_{\hat{\boldsymbol{\theta}}_0}$ $\xrightarrow{p} 0\ \forall\ i = 1, \ldots,\ m$ which conclude that, $\nabla\overline{D}_\beta(\hat{\boldsymbol{\theta}}_0) \xrightarrow{p} \mathbf{0}$ as $n \to \infty$. □

5.6.2 Proof of Theorem 5.2

Proof. Consistency of $\hat{\boldsymbol{\theta}}_{GD}$ follows immediately from the facts that

$$\hat{\boldsymbol{\theta}}_0 \xrightarrow{p} \boldsymbol{\theta}_0, \text{ and } \nabla\overline{D}_\beta(\hat{\boldsymbol{\theta}}_0) \xrightarrow{p} \mathbf{0}.$$

But, we need to show finite in probability limits of both $\nabla^2\overline{D}_\beta(\hat{\boldsymbol{\theta}}_0)$ and $E_{f_{\hat{\theta}_0}}(\nabla^2\overline{D}_\beta(\hat{\boldsymbol{\theta}}_0))$ in order to establish consistency of $\hat{\boldsymbol{\theta}}_{NR}$ and $\hat{\boldsymbol{\theta}}_{FS}$, respectively. Let us recall (5.35). One can show that,

$$\nabla^2\overline{D}_\beta(\hat{\boldsymbol{\theta}}_0) \xrightarrow{p} \boldsymbol{F} \text{ and } E_{f_{\hat{\theta}_0}}(\nabla^2\overline{D}_\beta(\hat{\boldsymbol{\theta}}_0)) \xrightarrow{p} \boldsymbol{F} \tag{5.38}$$

by considering the Taylor series expansions of each of the components of the aforesaid quantities around $\boldsymbol{\theta}_0$ with similar methodologies for establishing (5.35) and (5.37). This will establish the consistency of $\hat{\boldsymbol{\theta}}_{NR}$ and $\hat{\boldsymbol{\theta}}_{FS}$. □

5.6.3 Proof of Theorem 5.3

Proof. We start with the following Taylor series expansions:

$$\begin{aligned}\sqrt{n}\frac{\partial}{\partial\, \theta_i}\overline{D}_\beta(\boldsymbol{\theta})\Big|_{\hat{\boldsymbol{\theta}}_0} = \sqrt{n}\frac{\partial}{\partial\, \theta_i}\overline{D}_\beta(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}_0} &+ \sum_{j=1}^{m}\sqrt{n}(\hat{\theta}_{0j} - \theta_{0j})\frac{\partial^2}{\partial\, \theta_j\partial\, \theta_i}\overline{D}_\beta(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}_0} \\ &+ \sum_{j',j=1}^{m}\sqrt{n}(\hat{\theta}_{0j'} - \theta_{0j'})\frac{\partial^3}{\partial\, \theta_{j'}\partial\, \theta_j\partial\, \theta_i}\overline{D}_\beta(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}^{*i}}(\hat{\theta}_{0j} - \theta_{0j}),\end{aligned} \tag{5.39}$$

$\forall\ i = 1, \ldots, m$. Assumption 5.5 and Remark 5.1 conclude $\sqrt{n}(\hat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}_0)$ converges weakly to a normal distribution. Thus,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{0j'} - \boldsymbol{\theta}_{0j'}) = O_p(1),\ \forall j'. \tag{5.40}$$

By the consistency of $\hat{\boldsymbol{\theta}}_0$ (to $\boldsymbol{\theta}_0$), $(\hat{\boldsymbol{\theta}}_{0j} - \boldsymbol{\theta}_{0j}) = o_p(1)$ and $\left| \frac{\partial^3}{\partial\, \theta_{j'} \partial\, \theta_j \partial\, \theta_i} \overline{D}_\beta(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}^{*i}} \right| < \infty$ from (5.36). Hence the third term (in the expansion (5.39))

$$\sum_{j',j=1}^{m} \sqrt{n}(\hat{\theta}_{0j'} - \theta_{0j'}) \frac{\partial^3}{\partial\, \theta_{j'} \partial\, \theta_j \partial\, \theta_i} \overline{D}_\beta(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}^{*i}} (\hat{\theta}_{0j} - \theta_{0j}) = o_p(1). \tag{5.41}$$

As a consequence of (5.41), we have,

$$\sqrt{n} \nabla \overline{D}_\beta(\boldsymbol{\theta}) \Big|_{\hat{\boldsymbol{\theta}}_0} = \sqrt{n} \nabla \overline{D}_\beta(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}_0} + \nabla^2 \overline{D}_\beta(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}_0} \sqrt{n}(\hat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}_0) + o_p(1),$$

by representing the expansions in (5.39) in vector and matrix notations. From (5.35) and (5.40),

$$\begin{aligned} \nabla^2 \overline{D}_\beta(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}_0} \sqrt{n}(\hat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}_0) &= (\boldsymbol{F} + o_p(1)) \sqrt{n}(\hat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}_0) \\ &= \boldsymbol{F} \sqrt{n}(\hat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}_0) + o_p(1) \sqrt{n}(\hat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}_0) \\ &= \boldsymbol{F} \sqrt{n}(\hat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}_0) + o_p(1) O_p(1) = \boldsymbol{F} \sqrt{n}(\hat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}_0) + o_p(1). \end{aligned}$$

Thus,

$$\sqrt{n} \nabla \overline{D}_\beta(\boldsymbol{\theta}) \Big|_{\hat{\boldsymbol{\theta}}_0} = \sqrt{n} \nabla \overline{D}_\beta(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}_0} + \boldsymbol{F} \sqrt{n}(\hat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}_0) + o_p(1). \tag{5.42}$$

In case of the one-step Newton-Raphson update,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{NR} - \boldsymbol{\theta}_0) = \sqrt{n}(\hat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}_0) - \left(\nabla^2 \overline{D}_\beta(\boldsymbol{\theta}) \Big|_{\hat{\boldsymbol{\theta}}_0} \right)^{-1} \sqrt{n} \nabla \overline{D}_\beta(\boldsymbol{\theta}) \Big|_{\hat{\boldsymbol{\theta}}_0} \tag{5.43}$$

Following (5.38), we may write $\left(\nabla^2 \overline{D}_\beta(\boldsymbol{\theta}) \Big|_{\hat{\boldsymbol{\theta}}_0} \right)^{-1} = \boldsymbol{F}^{-1} + o_p(1)$.

Thus, Equation (5.42) implies

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\theta}}_{NR}-\boldsymbol{\theta}_0) &= \sqrt{n}(\hat{\boldsymbol{\theta}}_0-\boldsymbol{\theta}_0)-(\boldsymbol{F}^{-1}+o_p(1))\left(\sqrt{n}\nabla\overline{D}_\beta(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}_0}+\boldsymbol{F}\sqrt{n}(\hat{\boldsymbol{\theta}}_0-\boldsymbol{\theta}_0)+o_p(1)\right)\\ &= -\boldsymbol{F}^{-1}\sqrt{n}\nabla\overline{D}_\beta(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}_0}-o_p(1)\sqrt{n}\nabla\overline{D}_\beta(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}_0}-o_p(1)\boldsymbol{F}\sqrt{n}(\hat{\boldsymbol{\theta}}_0-\boldsymbol{\theta}_0)-o_p(1)\end{aligned}$$

Let us note that,

$$\begin{aligned}\sqrt{n}\nabla\overline{D}_\beta(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}_0} &= \sqrt{n}\frac{1}{n}\sum_{k=1}^{n}\nabla V(\boldsymbol{X}_k,\boldsymbol{\theta}_0)\text{ (from Equation (5.6))}\\ &\xrightarrow{d} N_m(\mathbf{0},\boldsymbol{G}(\boldsymbol{\theta}_0))\text{ (by the central limit theorem),}\end{aligned}$$

where, $\boldsymbol{G}(\boldsymbol{\theta}_0)=Var(\nabla V(\boldsymbol{X}_1,\boldsymbol{\theta}_0))$. Hence, $\sqrt{n}\nabla\overline{D}_\beta(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}_0}=O_p(1)$. Thus,

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\theta}}_{NR}-\boldsymbol{\theta}_0) &= -\boldsymbol{F}^{-1}\sqrt{n}\nabla\overline{D}_\beta(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}_0}-o_p(1)O_p(1)-o_p(1)O_p(1)-o_p(1)\\ &= -\boldsymbol{F}^{-1}\sqrt{n}\nabla\overline{D}_\beta(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}_0}-o_p(1).\end{aligned}$$

Since, $\nabla\overline{D}_\beta(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}_0}$ is a sum of identically and independently distributed random vectors, the application of the central limit theorem (along with the Slutsky's theorem) now implies that,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{NR}-\boldsymbol{\theta}_0)\xrightarrow{d}N_m(\mathbf{0},\boldsymbol{\Sigma}(\boldsymbol{\theta}_0))$$

as $n\xrightarrow{d}\infty$, where, $\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)=\boldsymbol{F}^{-1}\boldsymbol{G}(\boldsymbol{\theta}_0)\boldsymbol{F}^{-1}$. By similar argument, it can be shown that,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{FS}-\boldsymbol{\theta}_0)\xrightarrow{d}N_m(\mathbf{0},\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)).$$

For the one-step gradient descent estimator,

$$\begin{aligned}
\sqrt{n}(\hat{\boldsymbol{\theta}}_{GD} - \boldsymbol{\theta}_0) &= \sqrt{n}(\hat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}_0) - \gamma\sqrt{n}\nabla\overline{D}_\beta(\boldsymbol{\theta})\Big|_{\hat{\boldsymbol{\theta}}_0} \\
&= \sqrt{n}(\hat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}_0) - \gamma\sqrt{n}\nabla\overline{D}_\beta(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}_0} - \gamma\boldsymbol{F}\sqrt{n}(\hat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}_0) - o_p(1) \text{ (Equation (5.42))} \\
&= (\boldsymbol{I}_m - \gamma\boldsymbol{F})\sqrt{n}(\hat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}_0) - \gamma\sqrt{n}\nabla\overline{D}_\beta(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}_0} - o_p(1) \\
&= (\boldsymbol{I}_m - \gamma\boldsymbol{F})\sqrt{n}\left(\frac{1}{n}\sum_{k=1}^{n}\boldsymbol{Z}(\boldsymbol{X}_k, \boldsymbol{\theta}_0) + o_p(n^{-\frac{1}{2}})\right) - \sqrt{n}\frac{1}{n}\sum_{k=1}^{n}\gamma\nabla V(\boldsymbol{X}_k, \boldsymbol{\theta}_0) - o_p(1) \\
&= \sqrt{n}\frac{1}{n}\sum_{k=1}^{n}\left((\boldsymbol{I}_m - \gamma\boldsymbol{F})\boldsymbol{Z}(\boldsymbol{X}_k, \boldsymbol{\theta}_0) - \gamma\nabla V(\boldsymbol{X}_k, \boldsymbol{\theta}_0)\right) - o_p(1) \\
&= \sqrt{n}\frac{1}{n}\sum_{k=1}^{n}\boldsymbol{H}\left(\boldsymbol{X}_k, \boldsymbol{\theta}_0\right) - o_p(1).
\end{aligned}$$

An application of the central limit theorem (with Slutsky's theorem) concludes that,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{GD} - \boldsymbol{\theta}_0) \xrightarrow{d} N_m(\boldsymbol{0}, \boldsymbol{\Sigma}_{GD}(\boldsymbol{\theta}_0)),\ \boldsymbol{\Sigma}_{GD}(\boldsymbol{\theta}_0) = Var(\boldsymbol{H}\left(\boldsymbol{X}_1, \boldsymbol{\theta}_0\right)).$$

This completes the proof. □

5.6.4 Proof of Theorem 5.6

Proof. To understand this, let us first note that, $h(r^2) = \frac{1}{(2\pi)^{\frac{p}{2}}}e^{-\frac{r^2}{2}}$ in case of the p-dimensional normal model. Since $c^* = \infty$,

$$\begin{aligned}
\int_0^{\sqrt{c^*}} e^{-\frac{r^2}{2}(1+\beta)}r^\alpha\, dr &= \int_0^{\infty} e^{-\frac{r^2}{2}(1+\beta)}r^\alpha\, dr \\
&= \frac{1}{\sqrt{2(1+\beta)}}\left(\frac{2}{1+\beta}\right)^{\frac{\alpha}{2}}\Gamma\left(\frac{\alpha+1}{2}\right), \text{ for } \alpha > 0.
\end{aligned}$$

Putting, $\alpha = p-1$, and $\alpha = p+1$ we have

$$\begin{aligned}
c_1 &= \frac{2\pi^{\frac{p}{2}}}{\Gamma\left(\frac{p}{2}\right)}\int_0^{\infty} e^{-\frac{r^2}{2}(1+\beta)}r^{p-1}\, dr = \frac{2\pi^{\frac{p}{2}}}{\Gamma\left(\frac{p}{2}\right)}\frac{1}{\sqrt{2(1+\beta)}}\left(\frac{2}{1+\beta}\right)^{\frac{p-1}{2}}\Gamma\left(\frac{p}{2}\right), \\
c_3 &= \frac{2\pi^{\frac{p}{2}}}{p\Gamma\left(\frac{p}{2}\right)}\int_0^{\infty} e^{-\frac{r^2}{2}(1+\beta)}r^{p+1}\, dr = \frac{2\pi^{\frac{p}{2}}}{p\Gamma\left(\frac{p}{2}\right)}\frac{1}{\sqrt{2(1+\beta)}}\left(\frac{2}{1+\beta}\right)^{\frac{p+1}{2}}\Gamma\left(\frac{p}{2}+1\right),
\end{aligned}$$

so that, the consistency factor $\frac{c_1}{c_3} = 1 + \beta$.
In case of the exact one-step (IRLS) estimator of scale, the constant multiplier (upto the asymptotic extent by virtue of consistency of the initial estimators and WLLN) was

$$\frac{E_{N_p(\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)}(w(\boldsymbol{X}_1,\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0))}{E_{N_p(\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)}(w(\boldsymbol{X}_1,\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)) - \frac{\beta}{(1+\beta)^{1+\frac{p}{2}}}} = \frac{E_{N_p(\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)}(e^{-\frac{\beta}{2}(\boldsymbol{X}_1-\boldsymbol{\mu}_0)'\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{X}_1-\boldsymbol{\mu}_0)})}{E_{N_p(\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)}(e^{-\frac{\beta}{2}(\boldsymbol{X}_1-\boldsymbol{\mu}_0)'\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{X}_1-\boldsymbol{\mu}_0)}) - \frac{\beta}{(1+\beta)^{1+\frac{p}{2}}}}.$$

Now, $(\boldsymbol{X}_1 - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{X}_1 - \boldsymbol{\mu}_0)$ follows the $\chi^2(p)$ distribution, so that, $E_{N_p(\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)}(e^{-\frac{\beta}{2}(\boldsymbol{X}_1-\boldsymbol{\mu}_0)'\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{X}_1-\boldsymbol{\mu}_0)}) = (1+\beta)^{-\frac{p}{2}}$. This implies

$$\frac{E_{N_p(\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)}(e^{-\frac{\beta}{2}(\boldsymbol{X}_1-\boldsymbol{\mu}_0)'\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{X}_1-\boldsymbol{\mu}_0)})}{E_{N_p(\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)}(e^{-\frac{\beta}{2}(\boldsymbol{X}_1-\boldsymbol{\mu}_0)'\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{X}_1-\boldsymbol{\mu}_0)}) - \frac{\beta}{(1+\beta)^{1+\frac{p}{2}}}} = 1 + \beta.$$

Thus, the constant multiplier in the right hand side of (5.22) and the consistency factor $\frac{c_1}{c_3}$ are equal in case of normality under $c^* = \infty$. □

5.6.5 Elements of the Gradient Vectors and Hessian Matrices for Different Distributions

The elements of the gradient vector and Hessian matrix in case of the univariate normal model are as follows:

$$\begin{aligned}
\frac{\partial \overline{D}_\beta(\mu,\sigma^2)}{\partial \mu} &= -\frac{1+\beta}{n}\sum_{i=1}^n f^\beta(X_i,\mu,\sigma^2)\frac{X_i-\mu}{\sigma^2}, \text{ and} \\
\frac{\partial \overline{D}_\beta(\mu,\sigma^2)}{\partial \sigma^2} &= -\frac{\beta}{2\sigma^{\beta+2}(2\pi)^{\frac{\beta}{2}}\sqrt{1+\beta}} - \frac{1+\beta}{n}\sum_{i=1}^n f^\beta(X_i,\mu,\sigma^2)\frac{1}{2\sigma^2}\left[\left(\frac{X_i-\mu}{\sigma}\right)^2 - 1\right], \\
\frac{\partial^2 \overline{D}_\beta(\mu,\sigma^2)}{\partial^2 \mu} &= -\frac{1+\beta}{n\sigma^2}\left[\sum_{i=1}^n f^\beta(X_i,\mu,\sigma^2)\left(\beta\left(\frac{X_i-\mu}{\sigma}\right)^2 - 1\right)\right], \\
\frac{\partial^2 \overline{D}_\beta(\mu,\sigma^2)}{\partial \mu \partial \sigma^2} &= -\frac{1+\beta}{n\sigma^2}\left[\sum_{i=1}^n f^\beta(X_i,\mu,\sigma^2)\left(\frac{X_i-\mu}{\sigma^2}\right)\left(\frac{\beta}{2}\left(\left(\frac{X_i-\mu}{\sigma}\right)^2 - 1\right) - 1\right)\right], \\
\frac{\partial^2 \overline{D}_\beta(\mu,\sigma^2)}{\partial^2 \sigma^2} &= \frac{\beta(\beta+2)}{4\sigma^{\beta+4}(2\pi)^{\frac{\beta}{2}}\sqrt{1+\beta}} - \frac{1+\beta}{n\sigma^4}\Big[\sum_{i=1}^n f^\beta(X_i,\mu,\sigma^2)\Big[\frac{\beta}{4}\Big(\Big(\frac{X_i-\mu}{\sigma}\Big)^2 - 1\Big)^2 \\
&\quad + \Big(\frac{1}{2} - \Big(\frac{X_i-\mu}{\sigma}\Big)^2\Big)\Big]\Big].
\end{aligned}$$

For the univariate Cauchy model,

$$\frac{\partial \int f^{1+\beta}(x,\mu,\sigma)\,dx}{\partial \mu} = 0 \text{ (since this integral is free of } \mu),$$
$$\frac{\partial \int f^{1+\beta}(x,\mu,\sigma)\,dx}{\partial \sigma} = -\frac{\beta}{\pi^{\beta}\sigma^{\beta+1}} E\left[\frac{1}{(1+Z^2)^{\beta}}\right],$$
$$\frac{\partial^2 \int f^{1+\beta}(x,\mu,\sigma)\,dx}{\partial^2 \mu} = 0,\ \frac{\partial^2 \int f^{1+\beta}(x,\mu,\sigma)\,dx}{\partial \mu \partial \sigma} = 0,$$
$$\frac{\partial^2 \int f^{1+\beta}(x,\mu,\sigma)\,dx}{\partial^2 \sigma} = -\frac{\beta(\beta+1)}{\pi^{\beta}\sigma^{\beta+2}} E\left[\frac{1}{(1+Z^2)^{\beta}}\right],$$
$$\frac{\partial f^{\beta}(X,\mu,\sigma)}{\partial \mu} = \frac{2\beta}{\pi^{\beta}\sigma^{\beta+2}} \frac{X-\mu}{\left[1+\left(\frac{X-\mu}{\sigma}\right)^2\right]^{\beta+1}},\ \frac{\partial f^{\beta}(X,\mu,\sigma)}{\partial \sigma} = \frac{\beta}{\pi^{\beta}\sigma^{\beta+3}} \frac{(X-\mu)}{\left[1+\left(\frac{X-\mu}{\sigma}\right)^2\right]^{\beta+1}},$$
$$\frac{\partial^2 f^{\beta}(X,\mu,\sigma)}{\partial^2 \mu} = \frac{2\beta}{\pi^{\beta}\sigma^{\beta+2}} \frac{1}{\left[1+\left(\frac{X-\mu}{\sigma}\right)^2\right]^{\beta+2}} \left[(2\beta+1)\left(\frac{X-\mu}{\sigma}\right)^2 - 1\right],$$
$$\frac{\partial^2 f^{\beta}(X,\mu,\sigma)}{\partial \mu \partial \sigma} = \frac{2\beta}{\pi^{\beta}\sigma^{\beta+3}} \frac{(X-\mu)}{\left[1+\left(\frac{X-\mu}{\sigma}\right)^2\right]^{\beta+2}} \left[\beta\left(\frac{X-\mu}{\sigma}\right)^2 - \beta - 2\right],$$
$$\frac{\partial^2 f^{\beta}(X,\mu,\sigma)}{\partial^2 \sigma} = \frac{\beta}{\pi^{\beta}\sigma^{\beta+2}} \frac{1}{\left[1+\left(\frac{X-\mu}{\sigma}\right)^2\right]^{\beta+2}} \left[(\beta-1)\left(\frac{X-\mu}{\sigma}\right)^4 - 2(\beta+2)\left(\frac{X-\mu}{\sigma}\right)^2 + (\beta+1)\right].$$

For the Weibull model,

$$\frac{\partial f(x,k,\lambda)}{\partial k} = f(x,k,\lambda)\left[\frac{1}{k} + \log\left(\frac{x}{\lambda}\right) - \left(\frac{x}{\lambda}\right)^k \log\left(\frac{x}{\lambda}\right)\right],$$
$$\frac{\partial f(x,k,\lambda)}{\partial \lambda} = f(x,k,\lambda)\frac{k}{\lambda}\left[-1 + \left(\frac{x}{\lambda}\right)^k\right],$$
$$\frac{\partial^2 f(x,k,\lambda)}{\partial^2 k} = \frac{\partial f(x,k,\lambda)}{\partial k}\left[\frac{1}{k} + \log\left(\frac{x}{\lambda}\right) - \left(\frac{x}{\lambda}\right)^k \log\left(\frac{x}{\lambda}\right)\right]$$
$$- f(x,k,\lambda)\left[\frac{1}{k^2} + \left(\frac{x}{\lambda}\right)^k \left(\log\left(\frac{x}{\lambda}\right)\right)^2\right],$$
$$\frac{\partial^2 f(x,k,\lambda)}{\partial^2 \lambda} = \frac{\partial f(x,k,\lambda)}{\partial \lambda}\frac{k}{\lambda}\left[-1 + \left(\frac{x}{\lambda}\right)^k\right] - f(x,k,\lambda)\left[\frac{k}{\lambda^2} + (k+1)k\frac{x^k}{\lambda^{k+2}}\right], \text{ and,}$$
$$\frac{\partial^2 f(x,k,\lambda)}{\partial k \partial \lambda} = \left[-1 + \left(\frac{x}{\lambda}\right)^k\right]\left[\frac{\partial f(x,k,\lambda)}{\partial k}\frac{k}{\lambda} + \frac{f(x)}{\lambda}\right] + f(x,k,\lambda)\frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^k \log\left(\frac{x}{\lambda}\right).$$

For the shifted Gompertz model,

$$\frac{\partial f(x,\sigma,\eta)}{\partial \sigma} = f(x,\sigma,\eta)\left[\frac{1}{\sigma} - x + \eta x e^{-\sigma x} - \frac{\eta x e^{-\sigma x}}{1+\eta(1-e^{-\sigma x})}\right],$$

$$\frac{\partial f(x,\sigma,\eta)}{\partial \eta} = f(x,\sigma,\eta)\left[-e^{-\sigma x} + \frac{1-e^{-\sigma x}}{1+\eta(1-e^{-\sigma x})}\right],$$

$$\frac{\partial^2 f(x,\sigma,\eta)}{\partial^2 \sigma} = \frac{\partial f(x,\sigma,\eta)}{\partial \sigma}\left[\frac{1}{\sigma} - x + \eta x e^{-\sigma x} - \frac{\eta x e^{-\sigma x}}{1+\eta(1-e^{-\sigma x})}\right] + f(x,\sigma,\eta)\Big[-\frac{1}{\sigma^2}$$

$$-\eta x^2 e^{-\sigma x} + \frac{\eta^2 x^2 e^{-2\sigma x}}{(1+\eta(1-e^{-\sigma x}))^2}\Big],$$

$$\frac{\partial^2 f(x,\sigma,\eta)}{\partial^2 \eta} = \frac{\partial f(x,\sigma,\eta)}{\partial \eta}\left[-e^{-\sigma x} + \frac{1-e^{-\sigma x}}{1+\eta(1-e^{-\sigma x})}\right] - f(x,\sigma,\eta)\left[\frac{1-e^{-\sigma x}}{1+\eta(1-e^{-\sigma x})}\right]^2,$$

$$\frac{\partial^2 f(x,\sigma,\eta)}{\partial \sigma \partial \eta} = \frac{\partial f(x,\sigma,\eta)}{\partial \eta}\left[\frac{1}{\sigma} - x + \eta x e^{-\sigma x} - \frac{\eta x e^{-\sigma x}}{1+\eta(1-e^{-\sigma x})}\right]$$

$$+ f(x,\sigma,\eta)\left[x e^{-\sigma x} - \frac{x e^{-\sigma x}}{1+\eta(1-e^{-\sigma x})} + \frac{\eta x e^{-\sigma x}(1-e^{-\sigma x})}{(1+\eta(1-e^{-\sigma x}))^2}\right].$$

Chapter 6

Concluding Remarks and Future Plans

The principal focus of this thesis is to develop computationally efficient robust estimation tools in the multivariate location-scale set-up and to apply them to design sophisticated robust machine learning algorithms (e.g., clustering, classification and anomaly detection). We have proposed three different approaches of robust estimation following the minimum DPD philosophy in this purpose. They are illustrated with various theoretical properties (asymptotic and robustness related), simulation experiments, real data examples and applications in the domain of pattern recognition and machine learning (such as clustering, classification, anomaly detection and image reconstruction). Let us now present the concluding remarks and possible future directions of these research works.

In case of the maximum pseudo β-likelihood estimation, we have proposed an algorithm which robustly estimates the component weights, means and covariance matrices of a mixture normal model and performs data clustering and anomaly detection in the spirit of the minimum DPD philosophy. An IRLS algorithm for minimizing the DPD has been developed (for multivariate normal models) based on which the MPLE_β algorithm can robustly estimate the component means and covariance matrices. Theoretical results including existence and consistency of the estimators have been derived under certain technical assumptions. To explore the robustness of this method, we have studied the influence functions of our estimators and established the boundedness of the influence functions. Simulation studies have been presented in terms of regular misclassification error rates, proportion of undetected outliers, bias and mean squared errors of the component mean and covariance matrix estimators. On an average, satisfactory results have been obtained and our method worked has competitively or better than many of the well-known robust clustering methods in case of pure and contaminated datasets. We plan to investigate the following in our future research works.

Firstly, most of the tuning parameters in this work have been selected either subjectively, or through ad-hoc ideas. This has been briefly discussed in Section 2.2.3. This aspect of the MPLE_β method can possibly be improved, although it is a difficult

problem because of the reasons mentioned. In future, we will endeavour to build sophisticated statistical tools which would lead to the automatic selection of data-driven tuning parameters taking into account the (unknown) amount of anomaly in the data. Secondly, the asymptotic results related to the proposed estimators have primarily been considered in the context of mixtures of multivariate normals; the true unknown distribution has been assumed to belong to the model family. A possible generalization has been briefly discussed in Remark 3.2 under a specific additional assumption in Chapter 3. In the future, it will be of interest to establish the theoretical results under more general and relaxed assumptions. Thirdly, the NS constraint is crucial to prove some of the theoretical results, and at the moment a proof which bypasses this condition is not available. However, it is required only for some very rare pathological cases (discussed at the end of Section 2.2.1). In future, it will be among our primary goals to develop a proof which avoids the use of the NS constraint given the ER constraint, possibly by assuming some suitable moment conditions. Other possible future works may include the study of breakdown points of the MPLE_β estimators, extension of the MPLE_β algorithm to deal with subspace clustering and more real life applications in the domains of pattern recognition and machine learning.

In case of sequential minimum DPD estimation, we have derived a robust and asymptotically efficient method to estimate the location and scatter matrices of elliptically symmetric probability models. This estimation procedure is componentwise and thus scalable to large dimensions. Computational scalability is the main motivation behind developing this method and we hope that it will encourage practitioners to use the sequential minimum DPD method over simultaneous minimization of the DPD in large dimensions. We have established consistency and asymptotic normality of our estimators and derived the influence functions and illustrated these aspects explicitly under the assumption of normality. Finally, the simulation experiments have suggested the positive features of our newly established method.

In future, it will be interesting to apply our newly proposed sequential procedure in various machine learning, image analysis, finance and econometrics problems. This method can also be extended to high dimensional set-ups which is still a quite difficult statistical (and computational) problem. Apart from these applications, it will be interesting to study the behaviours of both the ordinary and sequential minimum DPD methods under cellwise contamination. We have provided one such example in Section 4.6.4. We plan to study the same in detail in a future work.

Finally, we have studied different one-step versions of the minimum DPD estimator and explored their theoretical properties. Four types of iterative methods have been utilized, namely, Newton-Raphson, gradient descent, IRLS and Fisher's scoring methods. The Newton-Raphson and the Fisher's scoring exact one-step estimators are shown to have the same asymptotic normal distribution as that of the original (fully converged) minimum DPD estimators under certain regularity assumptions. Additionally, these one-step estimators involve reduced computational complexity which makes it more preferable to practitioners. These methods are validated through simulation experiments and two real data examples.

The one-step Newton-Raphson, gradient descent and the Fisher's scoring estimators do not have any statistical interpretation in general, but, the one-step IRLS estimators are closely associated with the weighted sample mean and covariance matrix with data-driven weights. Motivated by this observation, we have proposed a generalized one-step methodology following the algebraic form of the exact one-step IRLS minimum DPD estimators under normality. These estimators of location and scale are weighted sample mean and a constant multiple of the weighted sample covariance matrix, respectively, with density power weights. The theoretical properties of these newly proposed estimators are discussed following Lopuhaä and Rousseeuw (1991) [98] and Lopuhaä (1999) [100]. The method has further been validated and compared with the traditional $0-1$ weight based one-step estimators and the initial robust estimators through simulation experiments and an application to prediction on survival of patients with heart failure via Bayes classification. Satisfactory results have been found in both simulation experiments and the medical application.

In future, it will interesting to apply the computationally efficient one-step methods for building up fast and robust classification, clustering and anomaly detection tools which will be helpful to deal with real life problems from various domains, including medical science, genetics and finance. We have proposed our one-step procedures in standard multivariate set-ups. We plan to extend this idea to regression and generalized linear models in future. Extension of the one-step methods into standard high-dimensional set-up with less computer intensive and highly robust initial estimators will be helpful to analyse datasets with larger dimensions as compared to their sample sizes. We hope to explore this in our future research endeavours.

We have assumed certain technical assumptions for the theoretical discussions and certain model assumptions in case of modelling the real datasets in this thesis. Some

remarks must be left in order to understand the testability of these assumptions. For deriving the theoretical properties of the proposed algorithms and estimators in this thesis, we have required to assume certain mathematical assumptions which were mostly used in past research works.

For example, in Chapter 2, the eigenvalue ratio constraint was used earlier while proving the theoretical properties of the TCLUST method (García-Escudero et al. (2008) [58]). A vivid discussion on the eigenvalue ratio and non-singularity constraints has been presented at the end of Section 2.2.1.

In Chapter 3, we have assumed Assumptions 3.1 and 3.2, both of which are technical assumptions on the true and model cluster proportions (π_j), respectively. In Assumption 3.1, it is assumed that at least one of the true cluster proportions is strictly greater than $\frac{\beta}{(1+\beta)^{1+\frac{p}{2}}}$, which is expected in practice provided all the clusters are not "too small". A discussion on Assumption 3.2 is provided in Remark 3.1.

In Chapter 4, the first three assumptions (i.e., Assumptions 4.1-4.3) were also assumed by Basu et al. (2011) [12]. We refer to Basu et al. (1998) [10] and (2011) [12] for discussions on these assumptions. Assumptions 4.4 and 4.5 are similar to Assumption 4.3 where the bounds can be derived from the boundedness of the function ψ (in Equation 4.3). Assumption 4.6 assumes sufficient positive definiteness of the true unknown covariance matrix which is quite natural to assume for avoiding non-singularity of the same.

In Chapter 5, Assumption 5.1 becomes valid if the density f and its partial derivatives with respect to the parameters are bounded above by some L_1-integrable functions, which is true at least for the normal distribution. Assumptions 5.2-5.4 were primarily assumed by Basu et al. (2011) [12]. Assumption 5.5 essentially states that the initial estimators are asymptotically normal (with a convergence rate of $o(n^{-\frac{1}{2}})$) which is required to prove the asymptotic normality of the one-step estimators. This property is satisfied by most of the robust estimators (e.g., MCD, S and many more) which can be used as initial robust estimators for the computation of one-step updates in order to achieve asymptotic normality of the one-step estimators.

In case of real data examples/applications, we have utilized certain probability models (mostly normal) to fit them. We have chosen the suitable probability models either following exploratory data analyses of these datasets or some past research works have used these probability models to fit these datasets. For example, SLC dataset (Chapter 3) was modelled by normal mixture models in Fujisawa and Eguchi (2006)

[54], the mice lifetime dataset (Chapter 5) was modelled by the Weibull distribution in Boudt et al. (2011) [16] and the breaking strength dataset (Chapter 5) was modelled by a reparametrized version of the shifted Gompertz model in Aydin et al. (2018) [5].

Publications and Preprints

Publications:

- Chakraborty, S., Basu, A., and Ghosh, A. (2023). Robust clustering with normal mixture models: A pseudo β-likelihood approach. *Econometrics and Statistics*, *Special Issue on Robustness Dedicated to Elvezio Ronchetti and Peter Rousseeuw*. DOI: https://doi.org/10.1016/j.ecosta.2023.10.004.

- Chakraborty, S., Basu, A., and Ghosh, A. (2025). A Componentwise Estimation Procedure for Multivariate Location and Scatter: Robustness, Efficiency and Scalability. *Journal of Multivariate Analysis*, 105546. DOI: https://doi.org/10.1016/j.jmva.2025.105546 (Volume: **212**, to appear in March, 2026).

Preprints:

- Chakraborty, S., Basu, A. and Ghosh, A. (2022). Existence and Consistency of the Maximum Pseudo β-Likelihood Estimators for Multivariate Normal Mixture Models. *ArXiv preprint*, **arXiv:2205.05405**.

Under Preparation:

- Chakraborty, S., Basu, A., and Ghosh, A. On One-step Estimation using Density Power Reweighting.

Publication not included in the Thesis:

- Basu, A., Chakraborty, S., Ghosh, A., and Pardo, L. (2022). Robust density power divergence based tests in multivariate analysis: A comparative overview of different approaches. Journal of Multivariate Analysis, 188, 104846. DOI: https://doi.org/10.1016/j.jmva.2021.104846.

Bibliography

[1] Agostinelli, C. and Greco, L. (2019). Weighted likelihood estimation of multivariate location and scatter. *Test*, **28(3)**, 756-784.

[2] Agostinelli, C., Leung, A., Yohai, V. J., and Zamar, R. H. (2015). Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. *Test*, **24(3)**, 441-461.

[3] Agostinelli, C. and Markatou, M. (2001). Test of hypotheses based on the weighted likelihood methodology. *Statistica Sinica*, **11(2)**, 499-514.

[4] Andrews, D. F. and Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society: Series B*, **36(1)**, 99-102.

[5] Aydin, D., Akgül, F. G., and Şenoğlu, B. (2018). Robust estimation of the location and the scale parameters of shifted Gompertz distribution. *Electronic Journal of Applied Statistical Analysis*, **11(01)**, 92-107.

[6] Bader, M. G., and Priest, A. M. (1982). Statistical aspects of fibre and bundle strength in hybrid composites. *Progress in Science and Engineering of Composites*, 1129-1136.

[7] Badsha, M. B., Mollah, M. N. H., Jahan, N. and Kurata, H. (2013). Robust complementary hierarchical clustering for gene expression data analysis by β-divergence. *Journal of bioscience and bioengineering*, **116(3)**, 397-407.

[8] Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, **49(3)**, 803–821.

[9] Basak, S., Basu, A., and Jones, M. C. (2021). On the 'optimal' density power divergence tuning parameter. *Journal of Applied Statistics*, **48(3)**, 536-556.

[10] Basu, A., Harris, I. R., Hjort, N. L. and Jones, M. C. (1998). Robust and efficient estimation by minimizing a density power divergence. *Biometrika*, **85(3)**, 549-559.

[11] Basu, A., and Lindsay, B. G. (1994). Minimum disparity estimation for continuous models: efficiency, distributions and robustness. *Annals of the Institute of Statistical Mathematics*, **46(4)**, 683-705.

[12] Basu, A., Shioya, H. and Park, C. (2011). *Statistical Inference: The Minimum Distance Approach*. Chapman and Hall/CRC.

[13] Bates, D. and Maechler, M. (2019). *Matrix: Sparse and Dense Matrix Classes and Methods. R package version* **1.2-17**. http://cran. r-project. org/package= Matrix.

[14] Bickel, P. J. (1975). One-step Huber estimates in the linear model. *Journal of the American Statistical Association*, **70(350)**, 428-434.

[15] Bickel, P. J. (1964). On some alternative estimates for shift in the p-variate one sample problem. *The Annals of Mathematical Statistics*, **35(3)**, 1079-1090.

[16] Boudt, K., Caliskan, D., and Croux, C. (2011). Robust explicit estimators of Weibull parameters. *Metrika*, **73**, 187-209.

[17] Box, G. E. (1953). Non-normality and tests on variances. *Biometrika*, **40(3/4)**, 318-335.

[18] Box, G. E. P. and Andersen, S. (1955). Permutation theory in derivation of robust criteria and the study of departures from assumption. *Journal of the Royal Statistical Society: Series B*, **17(1)**, 1-26.

[19] Cabana, E., Lillo, R. E. and Laniado, H. (2021). Multivariate outlier detection based on a robust Mahalanobis distance with shrinkage estimators. *Statistical Papers*, **62(4)**, 1583-1609.

[20] Cerioli, A., García-Escudero, L. A., Mayo-Iscar, A. and Riani, M. (2018). Finding the number of normal groups in model-based clustering via constrained likelihoods. *Journal of Computational and Graphical Statistics*, **27(2)**, 404-416.

[21] Chakraborty, S., Basu, A. and Ghosh, A. (2022). Existence and Consistency of the Maximum Pseudo β-Likelihood Estimators for Multivariate Normal Mixture Models. *ArXiv preprint*, **arXiv:2205.05405**.

[22] Chakraborty, S., Basu, A., and Ghosh, A. (2023). Robust clustering with normal mixture models: A pseudo β-likelihood approach. *Econometrics and Statistics*. https://doi.org/10.1016/j.ecosta.2023.10.004.

[23] Chakraborty, S., Basu, A., and Ghosh, A. (2024). A Componentwise Estimation Procedure for Multivariate Location and Scatter: Robustness, Efficiency and Scalability. *ArXiv preprint*, **arXiv:2410.21166**.

[24] Chen, T. L., Hsieh, D. N., Hung, H., Tu, I. P., Wu, P. S., Wu, Y. M., $\cdots$ and Huang, S. Y. (2014). γ-SUP: A clustering algorithm for cryo-electron microscopy images of asymmetric particles. *Annals of Applied Statistics*, **8(1)**, 259-285.

[25] Chmielewski, M. A. (1981). Elliptically symmetric distributions: A review and bibliography. *International Statistical Review/Revue Internationale de Statistique*, **49(1)**, 67-74.

[26] Chicco, D., and Jurman, G. (2020). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Medical Informatics and Decision Making*, **20(1)**, 1-16.

[27] Cichocki, A., & Amari, S. I. (2010). Families of alpha-beta-and gamma-divergences: Flexible and robust measures of similarities. *Entropy*, **12(6)**, 1532-1568.

[28] Cox, D. R. and D. V. Hinkley (1974). *Theoretical Statistics*. Chapman and Hall.

[29] Croux, C., and Haesbroeck, G. (1999). Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. *Journal of Multivariate Analysis*, **71(2)**, 161-190.

[30] Csiszár, I. (1963). Eine informationstheoretische Ungleichung und ihre Anwendung aufden Beweis der Ergodizität von Markoffschen Ketten. *Publ. Math. Inst. Hungar. Acad. Sci.*, **8**, 85–107.

[31] Cuesta-Albertos, J. A., Gordaliza, A. and Matrán, C. (1997). Trimmed k-means: An attempt to robustify quantizers. *The Annals of Statistics*, **25(2)**, 553-576.

[32] Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C. and Bontempi, G. (2017). Credit card fraud detection: a realistic modeling and a novel learning strategy. *IEEE transactions on neural networks and learning systems*, **29(8)**, 3784-3797.

[33] Dal Pozzolo, A., Caelen, O., Johnson, R. A. and Bontempi, G. (2015, December). Calibrating probability with undersampling for unbalanced classification. In *2015 IEEE Symposium Series on Computational Intelligence* (pp. 159-166). IEEE.

[34] Dal Pozzolo, A., Caelen, O., Le Borgne, Y. A., Waterschoot, S. and Bontempi, G. (2014). Learned lessons in credit card fraud detection from a practitioner perspective. *Expert Systems with Applications*, **41(10)**, 4915-4928.

[35] Danilov, M., Yohai, V. J., and Zamar, R. H. (2012). Robust estimation of multivariate location and scatter in the presence of missing data. *Journal of the American Statistical Association*, **107(499)**, 1178-1186.

[36] Davies, L. (1992). An efficient Fréchet differentiable high breakdown multivariate location and dispersion estimator. *Journal of Multivariate Analysis*, **40(2)**, 311-327.

[37] Davies, P. L. (1987). Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices. *The Annals of Statistics*, **15(3)**, 1269-1292.

[38] Donoho, D. L. (1982). Breakdown properties of multivariate location estimators. *Technical report*, Harvard University, Boston. URL: http://www-stat.stanford.edu/donoho/Reports/Oldies/BPMLE.pdf.

[39] Donoho, D.L. and Huber, P.J. (1983). *The notion of breakdown point.* In *"A Festschrift for Erich Lehmann" (P.J. Bickel, K. Doksum and J.L. Hodges, Jr., Eds.)*, Wadsworth, Belmont, CA, 157–184.

[40] Dua, D., and Graff, C. (2019). UCI machine learning repository (*http://archive. ics. uci. edu/ml*). Irvine, CA: University of California, School of Information and Computer Science. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **1(1)**, 1-29.

[41] Dudley, C. R., Giuffra, L. A., Raine, A. E. and Reeders, S. T. (1991). Assessing the role of APNH, a gene encoding for a human amiloride-sensitive Na+/H+ antiporter, on the interindividual variation in red cell Na+/Li+ countertransport. *Journal of the American Society of Nephrology*, **2(4)**, 937-943.

[42] Dykstra, R. L. (1983). An algorithm for restricted least squares regression. *Journal of the American Statistical Association*, **78(384)**, 837-842.

[43] Farcomeni, A., and Punzo, A. (2020). Robust model-based clustering with mild and gross outliers. *Test*, **29(4)**, 989-1007.

[44] Ferrari, D. and Paterlini, S. (2009). The maximum L_q-likelihood method: an application to extreme quantile estimation in finance. *Methodology and Computing in Applied Probability*, **11(1)**, 3-19.

[45] Ferrari, D. and Vecchia, D.L. (2012). On robust estimation via pseudo-additive information. *Biometrika*, **99(1)**, 238-244.

[46] Ferrari, D. and Yang, Y. (2007). *Estimation of tail probability via the maximum L_q-likelihood method. Technical Report* **659**, University of Minnesota, School of Statistics.

[47] Ferrari, D. and Yang, Y. (2010). Maximum L_q-likelihood estimation. *The Annals of Statistics*, **38(2)**, 753-783.

[48] Fisher, R. A. (1912). On an absolute criterion for fitting frequency curves. *Messenger of Mathematics*, **41**, 155-156.

[49] Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical transactions of the Royal Society of London Series A*, **222** (594-604), 309-368.

[50] Fisher, R. A. (1925, July). Theory of statistical estimation. In *Mathematical proceedings of the Cambridge philosophical society*, **22(5)**, 700-725. Cambridge University Press.

[51] Flury, B. and Riedwyl, H. (1988). *Multivariate Statistics: A Practical Approach.* London: Chapman & Hall.

[52] Fritz, H., Garcıa-Escudero, L.A. and Mayo-Iscar, A. (2012). tclust: An R package for a trimming approach to cluster analysis. *Journal of Statistical Software*, **47(12)**, 1-26.

[53] Fritz, H., García-Escudero, L. A., and Mayo-Iscar, A. (2013). A fast algorithm for robust constrained clustering. *Computational Statistics and Data Analysis*, **61**, 124-136.

[54] Fujisawa, H. and Eguchi, S. (2006). Robust estimation in the normal mixture model. *Journal of Statistical Planning and Inference*, **136(11)**, 3989-4011.

[55] Fujisawa, H. and Eguchi, S. (2008). Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis*, **99(9)**, 2053-2081.

[56] Gallegos, M. T. (2002). Maximum likelihood clustering with outliers. In *Classification, Clustering and Data Analysis*, Conference paper, 247-255. Springer, Berlin, Heidelberg.

[57] Gallegos, M. T. and Ritter, G. (2005). A robust method for cluster analysis. *The Annals of Statistics*, **33(1)**, 347-380.

[58] García-Escudero, L. A., Gordaliza, A., Matrán, C. and Mayo-Iscar, A. (2008). A general trimming approach to robust cluster analysis. *The Annals of Statistics*, **36(3)**, 1324-1345.

[59] García-Escudero, L. A., Gordaliza, A., Matrán, C. and Mayo-Iscar, A. (2010). A review of robust clustering methods. *Advances in Data Analysis and Classification*, **4(2-3)**, 89-109.

[60] Gervini, D. (2003). A robust and efficient adaptive reweighted estimator of multivariate location and scatter. *Journal of Multivariate Analysis*, **84(1)**, 116-144.

[61] Ghosh, A., Harris, I. R., Maji, A., Basu, A. and Pardo, L. (2017). A generalized divergence for statistical inference. *Bernoulli*, **23(4A)**, 2746-2783.

[62] Green, P. J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society: Series B (Methodological)*, **46(2)**, 149-170.

[63] Giné, E., Mason, D.M. and Wellner, J.A. eds. (2012). *High Dimensional Probability II* (**Vol. 47**). Springer Science and Business Media.

[64] Güney, Y., Tuac, Y., Özdemir, Ş., and Arslan, O. (2021). Conditional maximum L_q-likelihood estimation for regression model with autoregressive error terms. *Metrika*, **84(1)**, 47-74.

[65] Hampel, F. R. (1968). *Contributions to the Theory of Robust Estimation*. University of California, Berkeley.

[66] Hampel, F.R. (1971),A general definition of qualitative robustness, *The Annals of Mathematical Statistics*, **42(6)**, 1887–1896.

[67] Hampel, F.R. (1974), The influence curve and its role in robust estimation, *Journal of the American Statistical Association*, **69(346)**, 383–393.

[68] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986).*Robust Statistics—The Approach Based on Influence Functions*. John Wiley and Sons, New York.

[69] Hathaway, R. J. (1985). A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *The Annals of Statistics*, **13(2)**, 795-800.

[70] He, X., and Portnoy, S. (1992). Reweighted LS estimators converge at the same rate as the initial estimator. *The Annals of Statistics*, **20(4)**, 2161-2167.

[71] Henning, C. (2020). trimcluster: Cluster Analysis with Trimming. *R package* version **0.1-5**. https://CRAN.R-project.org/package=trimcluster.

[72] Higham, N.J., (2002). Computing the nearest correlation matrix—a problem from finance. *IMA journal of Numerical Analysis*, **22(3)**, pp.329-343.

[73] Hoel, D. G. (1972). A representation of mortality data by competing risks. *Biometrics*, **28(2)**, 475-488.

[74] Hu, F. (1997). The asymptotic properties of the maximum-relevance weighted likelihood estimators. *Canadian Journal of Statistics*, **25(1)**, 45-59.

[75] Hu, F. and Zidek, J. V. (2001). The relevance weighted likelihood with applications. In Empirical Bayes and Likelihood Inference, Lecture Notes in Statistics (**Vol. 148**), 211-235. Springer New York.

[76] Hu, F. and Zidek, J. V. (2002). The weighted likelihood. *Canadian Journal of Statistics*, **30(3)**, 347-371.

[77] Huber, P.J. (March, 1964). Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*. **35 (1)**, 73-101.

[78] Huber, P.J. (1965). The Behaviour of Maximum Likelihood Estimates Under Nonstandard Conditions. *Proceedings of the Fifth Berkeley Symposium*, **1**, 221-233.

[79] Huber, P. J. (2004). *Robust statistics* (**Vol. 523**). John Wiley and Sons.

[80] Hubert, M., Debruyne, M., and Rousseeuw, P. J. (2018). Minimum covariance determinant and extensions. *Wiley Interdisciplinary Reviews: Computational Statistics*, **10(3)**, e1421.

[81] Jiahui Wang, Ruben Zamar, Alfio Marazzi, Victor Yohai, Matias Salibian-Barrera, Ricardo Maronna, Eric Zivot, David Rocke, Doug Martin, Martin Maechler and Kjell Konis. (2019). *robust: Port of the S+ "Robust Library"*. R package version **0.4-18.1**.

[82] Jones, M. C., Hjort, N. L., Harris, I. R., and Basu, A. (2001). A comparison of related density-based minimum divergence estimators. *Biometrika*, **88(3)**, 865-873.

[83] Kalbfleisch, J. D., and Prentice, R. L. (2011). *The Statistical Analysis of Failure Time Data*. John Wiley and Sons.

[84] Kaufman, L. and Rousseeuw, PJ. (1987). Clustering by means of medoids. *In: Dodge Y (ed) Statistical Data Analysis Based on the L_1 Norm and Related Methods*, pp 405–416.

[85] Kelker, D. (1970). Distribution theory of spherical distributions and a location-scale parameter generalization. *Sankhyā: The Indian Journal of Statistics, Series A*, **32(4)**, 419-430.

[86] Kent, J. T. and Tyler, D. E. (1991). Redescending M-estimates of multivariate location and scatter. *The Annals of Statistics*, **19(4)**, 2102-2119.

[87] Kent, J. T. and Tyler, D. E. (1996). Constrained M-estimation for multivariate location and scatter. *The Annals of Statistics*, **24(3)**, 1346-1370.

[88] Kosorok, M.R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. New York: Springer.

[89] Kravchuk, O. Y., and Pollett, P. K. (2012). Hodges-Lehmann scale estimator for Cauchy distribution. *Communications in Statistics-Theory and Methods*, **41(20)**, 3621-3632.

[90] Lax, D. A. (1985). Robust estimators of scale: Finite-sample performance in long-tailed symmetric distributions. *Journal of the American Statistical Association*, **80(391)**, 736-741.

[91] Le Cam, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates. *Univ. Calif. Publ. in Statist.*, **1**, 277-330.

[92] Lehmann, E. L. (1983). *Theory of Point Estimation*. John Wiley and Sons, New York.

[93] Lenth, R. V. and Green, P. J. (1987). Consistency of deviance-based M estimators. *Journal of the Royal Statistical Society: Series B (Methodological)*, **49(3)**, 326-330.

[94] Lindsay, B. G. (1994). Efficiency versus robustness: the case for minimum Hellinger distance and related methods. *The Annals of Statistics*, **22(2)**, 1081-1114.

[95] Lindsay, B. G. (1995, January). Mixture models: theory, geometry and applications. In *NSF-CBMS regional conference series in probability and statistics* **(i-163)**. Institute of Mathematical Statistics and the American Statistical Association.

[96] Loève, M. (1977). *Probability Theory I. Graduate Texts in Mathematics*. Springer New York. ISBN: 9780387902104.

[97] Lopuhaä, H. P. (1989). On the relation between S-estimators and M-estimators of multivariate location and covariance. *The Annals of Statistics*, **17(4)**, 1662-1683.

[98] Lopuhaä, H. P., and Rousseeuw, P. J. (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *The Annals of Statistics*, **19(1)**, 229-248.

[99] Lopuhaä, H. P. (1997). Asymptotic expansion of S-estimators of location and covariance. *Statistica Neerlandica*, **51(2)**, 220-237.

[100] Lopuhaä, H. P. (1999). Asymptotics of reweighted estimators of multivariate location and scatter. *The Annals of Statistics*, **27(5)**, 1638-1665.

[101] Ma, Y. and Genton, M. G. (2001). Highly robust estimation of dispersion matrices. *Journal of Multivariate Analysis*, **78(1)**, 11-36.

[102] Maechler, M. et al. (2021). Package 'robustbase': *Basic Robust Statistics R package version* **0.93-6**. http://CRAN.R-project.org/package=robustbase.

[103] Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M. and Hornik, K. (2017). cluster: Cluster analysis basics and extensions (2019). *R package* version **2(3)**.

[104] Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979). *Multivariate Analysis*. Probability and Mathematical Statistics, Academic Press Inc.

[105] Markatou, M. (2000). Mixture models, robustness, and the weighted likelihood methodology. *Biometrics*, **56(2)**, 483-486.

[106] Markatou, M., Basu, A. and Lindsay, B. (1997). Weighted likelihood estimating equations: The discrete case with applications to logistic regression. *Journal of Statistical Planning and Inference*, **57(2)**, 215-232.

[107] Markatou, M., Basu, A. and Lindsay, B. G. (1998). Weighted likelihood equations with bootstrap root search. *Journal of the American Statistical Association*, **93(442)**, 740-750.

[108] Maronna, R. A. (1974). Estimación robusta de locación y dispersión multivariadas (*Doctoral dissertation, Universidad de Buenos Aires. Facultad de Ciencias Exactas y Naturales*).

[109] Maronna, R. A. (1976). Robust M-estimators of multivariate location and scatter. *The Annals of Statistics*, **4(1)**, 51-67.

[110] Maronna, R. A., Martin, R. D., Yohai, V. J., and Salibián-Barrera, M. (2019). *Robust Statistics: Theory and Methods (with R)*. John Wiley and Sons.

[111] Maronna, R. A. and Yohai, V. J. (2014). Robust estimation of multivariate location and scatter. *Wiley StatsRef: Statistics Reference Online*, 1-12.

[112] Maronna, R. A. and Zamar, R. H. (2002). Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics*, **44(4)**, 307-317.

[113] Martín, N. (2020). Rao's Score Tests on Correlation Matrices. *ArXiv preprint*, **arXiv:2012.14238**.

[114] McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. John Wiley and Sons, New York.

[115] Mehrotra, D. V. (1995). Robust elementwise estimation of a dispersion matrix. *Biometrics*, **51(4)**, 1344-1351.

[116] Notsu, A., Komori, O. and Eguchi, S. (2014). Spontaneous clustering via minimum gamma-divergence. *Neural Computation*, **26(2)**, 421-448.

[117] Olive, D. (2006). Robust estimators for transformed location scale families. Southern Illinois University. Mailcode, 4408, 62901-4408.

[118] Protter, M.H. and Charles Jr, B. (2012). *Intermediate Calculus*. Springer Science and Business Media.

[119] Punzo, A., and McNicholas, P. D. (2016). Parsimonious mixtures of multivariate contaminated normal distributions. *Biometrical Journal*, **58(6)**, 1506-1537.

[120] Qin, Y. and Priebe, C.E. (2017). Robust Hypothesis Testing Via L_q-Likelihood. *Statistica Sinica*, **27**, 1793-1813.

[121] Qin, Z. S., Damien, P. and Walker, S. (2003, November). Scale mixture models with applications to Bayesian inference. *AIP Conference Proceedings*, **690(1)**, 394-395.

[122] Rao, C. R. (1948, January). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, **44(1)**, 50-57. Cambridge University Press.

[123] R Core Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

[124] Roeder, K. (1994). A graphical technique for determining the number of components in a mixture of normals. *Journal of the American Statistical Association*, **89(426)**, 487-495.

[125] Ronchetti, E. (1982). Robust testing in linear models: the infinitesimal approach. *Dissertation ETH*, **7084**. Eidgenoessische Technische Hochschule Zuerich (Switzerland).

[126] Ronchetti, E. (1982, June). Robust alternatives to the F-test for the linear model. In Probability and Statistical Inference: *Proceedings of the 2nd Pannonian Symposium on Mathematical Statistics, Bad Tatzmannsdorf, Austria*, June 14–20, 1981 (pp. 329-342). Dordrecht: Springer Netherlands.

[127] Ronchetti, E. (1985). Robust model selection in regression. *Statistics and Probability Letters*, **3(1)**, 21-23.

[128] Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, **79(388)**, 871-880.

[129] Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications*, **8(37)**, 283-297.

[130] Rousseeuw, P. J. and Driessen, K. V. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, **41(3)**, 212-223.

[131] Rousseeuw, P. J. and Leroy, A. M. (2005). *Robust Regression and Outlier Detection*. John wiley and sons. New York.

[132] Rousseeuw, P. J. and Molenberghs, G. (1993). Transformation of non positive semidefinite correlation matrices. *Communications in Statistics–Theory and Methods*, **22(4)**, 965-984.

[133] Rousseeuw, P. J. and van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical association*, **85(411)**, 633-639.

[134] Rousseeuw, P.J. and Yohai, V. (1984). Robust regression by means of S-estimators. *In Robust and Nonlinear Time Series Analysis* (pp. 256-272). Springer, New York, NY.

[135] Ruppert, D. (1992). Computing S estimators for regression and multivariate location/dispersion. *Journal of Computational and Graphical Statistics*, **1(3)**, 253-270.

[136] Ruwet, C., García-Escudero, L. A., Gordaliza, A. and Mayo-Iscar, A. (2012). The influence function of the TCLUST robust clustering procedure. *Advances in Data Analysis and Classification*, **6(2)**, 107-130.

[137] Scrucca, L., Fop, M., Murphy, T. B., and Raftery, A. E. (2016). mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, **8(1)**, 289.

[138] Sen, P. K. and Puri, M. L. (1971). *Nonparametric Methods in Multivariate Analysis*. John Wiley and Sons, Limited.

[139] Small, C. G. (1990). A survey of multidimensional medians. *International Statistical Review*, **58(3)**, 263-277.

[140] SpaceX Satellite. Digital image, https://medium.com/swlh/using-deep-learning-semantic-segmentation-method-for-ship-detection-on-satellite-optical-imagery-ffeaae8c1ab.

[141] Stahel, W. A. (1981). Breakdown of covariance estimators. Fachgruppe für Statistik, Research Report Number **31**, *Eidgenössische Techn. Hochsch.*

[142] Sugar, C. A. and James, G. M. (2003). Finding the number of clusters in a dataset: An information-theoretic approach. *Journal of the American Statistical Association*, **98(463)**, 750-763.

[143] Tatsuoka, K. S. and Tyler, D. E. (2000). On the uniqueness of S-functionals and M-functionals under nonelliptical distributions. *The Annals of Statistics*, **28(4)**, 1219-1243.

[144] Titterington, D. M., Smith, A. F. and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Chichester, Wiley.

[145] Todorov, V. and Filzmoser, P. (2009). An object-oriented framework for robust multivariate analysis. *Journal of Statistical Software*, **32(3)**, 1-47. URL: http://www.jstatsoft.org/v32/i03/.

[146] Tukey, J.W. (1960), *A survey of sampling from contaminated distributions, Contributions to Probability and Statistics*, I. Olkin (ed.), Stanford, CA: Stanford University Press.

[147] Tukey, J. W. (1962). The future of data analysis. *Breakthroughs in Statistics: Methodology and Distribution*, 408-452. New York, NY: Springer New York.

[148] Tyler, D. E. (1983). Robustness and efficiency properties of scatter matrices. *Biometrika*, **70(2)**, 411-420.

[149] Tyler, D. E. (1987). A distribution-free M-estimator of multivariate scatter. *The Annals of Statistics*, **15(1)**, 234-251.

[150] Tyler, D. E. (2014). Breakdown properties of the M-estimators of multivariate scatter. *ArXiv preprint*, **arXiv:1406.4904**.

[151] Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer.

[152] von Mises, R. (1936). Les dois de probabilite pour les fonctions statistiques. *Ann. Inst. H. Poincare B.*, **6**, 185–212.

[153] von Mises, R. (1937). Sur les fonctions statistiques. In *Soc. Math. de France, Conference de la Reunion Internat. des Math*, Paris, France.

[154] von Mises, R. (1947). On the asymptotic distribution of differentiable statistical functions. *The Annals of Mathematical Statistics*, **18**, 309–348.

[155] Warwick, J. and Jones, M. C. (2005). Choosing a robustness tuning parameter. *Journal of Statistical Computation and Simulation*, **75(7)**, 581-588.

[156] Welsh, A. H., and Ronchetti, E. (2002). A journey in single steps: robust one-step M-estimation in linear regression. *Journal of Statistical Planning and Inference*, **103(1-2)**, 287-310.

[157] Xu, L., Xiang, S. and Yao, W. (2019). Robust maximum L_q-likelihood estimation of joint mean–covariance models for longitudinal data. *Journal of Multivariate Analysis*, **171**, 397-411.

[158] Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, **15(2)**, 642-656.