1-1-2012

Contributions to Robust Methods: Modified Rank Covariance Matrix and Spatial-EM Algorithm

Kai Yu

University of Mississippi

Recommended Citation

Yu, Kai, "Contributions to Robust Methods: Modified Rank Covariance Matrix and Spatial-EM Algorithm" (2012). *Electronic Theses and Dissertations*. 1438.
https://egrove.olemiss.edu/etd/1438

Contributions to Robust Methods: Modified Rank Covariance Matrix and Spatial-EM Algorithm

Kai Yu

A dissertation submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy
Mathematics
Concentration in Statistics

The University of Mississippi
2012

Abstract

Classical multivariate statistical inference methods including multivariate analysis of variance, principal component analysis, factor analysis, canonical correlation analysis are based on sample covariance matrix. Those moment-based techniques are optimal (most efficient) under the normality distributional assumption. They are, however, extremely sensitive to outlying observations, susceptible to small perturbation in data and poor in the efficiency for heavy-tailed distributions. A straightforward treatment is to replace the sample covariance matrix with a robust one. Visuri *et al.* (2000) proposed a technique for robust covariance matrix estimation based on different notions of multivariate sign and rank. Among them, the spatial rank based covariance matrix estimator that utilizes a robust scale estimator (MRCM) is especially appealing due to its high robustness, computational ease and good efficiency. In this dissertation, properties of the estimator on orthogonal equivariance under any distribution and affine equivariance under elliptically symmetric distributions have been established. The major robustness properties of the estimator are studied by the *breakdown point* and *influence function* analysis. More specifically, the finite sample breakdown point is obtained and the upper bound of the finite sample breakdown point can be achieved by a proper choice of univariate robust scale estimator. The influence functions for eigenvalues and eigenvectors of the estimator are derived. They are found to be bounded under some mild assumptions. Moreover, empirical comparisons to popular robust MCD, M and S estimators show that MRCM has a competitive performance on efficiency as well as robustness.

With rapid advances in information technology, data have been becoming huge in size and complex in structure. A single elliptical distribution is no longer sufficient to model such data. This motivates a generalization of our notion of MRCM to mixture models. In this dissertation, we propose a robust Spatial-EM algorithm for estimating parameters in the mixture model. Rather than using sample covariance matrix in each M-step, Spatial-EM ingeniously implements MRCM to enhance stability and robustness of the estimation procedure. Analyzing the log-likelihood function, the proposed one is found to be closely related to the maximum likelihood estimator (MLE) of Kotz type mixture model. Comparing with the direct MLE, Spatial-EM has advantages in computation ease as well as stability.

Applications of Spatial-EM to data mining become natural. We illustrate procedures how to use Spatial-EM for supervised and unsupervised learning problems. More specifically, robust clustering and outlier detection methods based on Spatial-EM have been proposed. We adopt the outlier detection to taxonomic research on fish species novelty discovery. UCI Wisconsin diagnostic breast cancer data and Yeast cell cycle data are used for clustering analysis. Comparing with the regular EM and many other existing methods such as X-EM and SVM, Spatial-EM demonstrates its competitive classification power and high robustness.

Dedication

This work is dedicated to my parents, Yaoqing Yu and Shaoping Li,
who have fully supported me since my day one.
Without them there was never a chance.
It is also dedicated to my wife, Qianru Lin,
who always gives me endless love and encouragement.

Acknowledgments

I am grateful to my advisor Dr. Xin Dang for her tireless guidance, support, and inspiration over these years. This work would not have been possible without Dr. Dang's generous help and patient encouragement. She constantly gave me moral support and guided me through in different matters regarding the research. Her enthusiasm and sparkling ideas on research deeply impressed me. In part due to her influence, I am getting more and more in love with my research and I want to devote my future to it. I am thankful to Dr. Yixin Chen, for the fruitful discussions and helpful advice through the years of this research. He kindly provided me the data for analysis and always pointed me in valuable directions when I encountered problems. I would also like to express my appreciation to Dr. Hanxiang Peng, who gave me the lectures of fundamental statistics courses during my years in Ph.D. program. His solid knowledge and thorough explanation of subjects inspired me a lot.

I would like to thank my committee members Dr. Iwo Labuda, the chair of the Department of Mathematics, Dr. William Staton, Professor of Mathematics, and Dr. Ali Al-Sharadqah, Visiting Professor of Mathematics, for serving as members of the examining committee. Especially thanks for Dr. Labuda's support in all aspects regarding to my research and study in the department, for Dr. Staton's entertaining and informative math courses that solidify my profession, and for immensely useful suggestions from Dr. Al-Sharadqah on my dissertation.

Last but no the least, great appreciation goes to all my teachers, fellow graduate students and friends for their incessant support and encouragement. They make my stay at Olemiss rich and colorful. This has been the most treasured memory in my life.

Table of Contents

List of Figures

Figure Number Page

List of Tables

Table Number Page

Chapter 1

Introduction

1.1 Statistical Functionals

Let $\boldsymbol{X}_1$,...,$\boldsymbol{X}_n$ be a random sample from a population with probability distribution function $F \in \mathbb{R}^p$. Assume that F belongs to some family $\mathcal{F}$ of distribution functions but is unknown. One is usually interested in estimating some quantities related to F, denoted by the parameter $\boldsymbol{\theta}$. In fact, $\boldsymbol{\theta}$ can be viewed as a functional $T(F)$. Then, an intuitive estimator of $\boldsymbol{\theta}$. with sample data $\{\boldsymbol{X}_1, ..., \boldsymbol{X}_n\}$ is simply $T(F_n)$, Here F_n is the *empirical distribution function* such that for a set $A \subseteq \mathbb{R}^P$,

$$F_n(A) = \frac{1}{n}\sum_{i=1}^{n} I[\boldsymbol{x}_i \in A],$$

where $I[\cdot]$ is the indicator function which is 0 when $I[False]$ and is 1 when $I[True]$. F_n uniformly puts probability mass $1/n$ on each point $\boldsymbol{x}_i$, therefore it has *empirical probability mass function* $f_n(\boldsymbol{x}_i) = 1/n, i = 1, ..., n$.

Usually, a parameter $\boldsymbol{\theta}$ can be characterized by various functionals. For instance, in unimodal symmetric distributions, the center of symmetry can be represented by one of the following: the expected value, median or mode of a given distribution. Some functionals are explicitly defined in terms of F. For instance, the expected value of random variable X is a functional of F defined by $E_F(X) = T(F) := \int x dF(x)$, then $\hat{\theta} = T(F_n) = 1/n \sum_{i=1}^{n} x_i = \bar{x}$. However, some other functionals $T(F)$ are implicitly defined as a root of a system of equations or as a solution of a minimization (maximization) problem. For instance, the median of X, $\mathrm{Med}(X) = T(F)$ is a solution of the equation $F(X) = 1/2$. The maximal likelihood estimator of $\boldsymbol{\theta}$ is a solution maximizing the likelihood function based on F, etc. Estimator of $\boldsymbol{\theta}$ is usually obtained as an empirical functional based on F_n. That is, given $\boldsymbol{\theta} := T(F)$, we estimate $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}} = T(F_n)$. Estimators based on distinct functionals T have different statistical properties regarding to *efficiency*, *robustness* and computational complexity. The robustness of a statistical functional is discussed in detail in Section 1.2. The concept of Statistical efficiency is introduced in Section 1.3. A few example of affine equivariant estimators are listed in Section 1.4. In this paper, we propose a new robust statistical approach that improves existing methods in terms of above mentioned properties.

1.2 Characteristics of Robustness

Every statistical approach builds upon a set of explicit and implicit assumptions. Any result coming out of the approach relies on these assumptions. The outcomes that are not influenced much by (little) changes in assumptions are called them *robust*.

Statisticians noticed the sensitivity of statistical procedures to deviations from model assumptions. It took a long time to develop the concept of robustness from various points of view. Meanwhile, the issue prompted researchers from different areas to determine the most important and most reasonable assumptions for the underlying model. In order to give an feeling of robustness and what the effect of an outlier can be, let us consider the following example.

Example 1 *Consider a sorted random sample of X, $\mathbb{X} = \{2, 3, 5, 7, 7, 9, 11, 12, 13, 16, \mathbf{100}\}$, it is obvious that the value 100 is far away from the majority of data distribution and therefore can be thought as an outlier. Different estimators of location and scale can be used. Two columns of values represent the estimates without or with (bold face font) the outlier 100. The values of sample mean and standard deviation change significantly with the presence of the outlier. However, the other estimators remain stable. So we can roughly consider mode, median, IQR and MAD are robust estimators, but sample mean and sample standard deviation are not.*

Location			Scale		
Sample Mean	8.5	**16.82**	Standard Deviation	4.53	**27.92**
Mode	7	**7**	IQR=$\|Q_3 - Q_1\|$	6.25	**6.5**
Median	8	**9**	MAD	5.19	**5.93**

In fact, the interpretation of robustness can involve many different aspects. From a technical point of view, Hampel *et al.* (1986, p.6) provided the following definition of statistical robustness.

> In a broad informal sense, robust statistics is a body of knowledge, partly formalized into "theories of robustness" relating to deviations from idealized assumptions of statistics.

For instance, in the setting of linear models, number of discrepancies from ideal assumptions have been investigated. These include the effect of dependence of observations, heteroscedasticity, collinearity, measurement error, influential point in terms of both leverage and outliers, normality of error, etc. However, we restrict our discussion of robustness of a statistical procedure to deviation from the underlying distribution of data. More specifically, there are two types of definitions of outliers. One is nonparametric, based on some notions of distance. An outlier is considered to be the data far away from the majority of data in a predefined distance metric. The other is parametric, based on some distributions. An outlier is thought to be the data point residing in the low density region of the given distribution. Perhaps the two most popular examples that are sensitive to outlying observations are sample mean and sample standard deviation as indicated in Example 1.

Statistically, there are three basic tools to examine whether an estimator is robust or not. They are qualitative robustness, infinitesimal robustness, and quantitative robustness by breakdown points. We illustrate their properties in the univariate case and focus on robustness of location and scale estimators in this chapter. The generalizations to the multivariate case would be studied in the following chapters.

Before going further, the meaning of measure of location and scale (scatter in the multivariate case) are explained here. For univariate distribution $F(X)$, the basic requirement for a quantity θ

to be a measure of location is the *scale and location equivariance*. That is, for any number a and $b \in \mathbb{R}$, θ must satisfy

$$\theta(aX + b) = a\theta(X) + b. \tag{1.1}$$

A quantity σ is a measure of scale, e.g., population standard deviation, if it is *scale equivariant* i.e. for $a > 0$,

$$\sigma(aX) = a\sigma(X) \tag{1.2}$$

and *location and sign invariant*, i.e., for any number b,

$$\sigma(X + b) = \sigma(X) \tag{1.3}$$

$$\sigma(-X) = \sigma(X). \tag{1.4}$$

Affine equivariance is similar to the situation of a change of unit, For example, when the unit changes, then the values of θ and σ changes.

1.2.1 Qualitative Robustness

Qualitative robustness can be understood as the continuity of a function $f(x)$ (not necessary the probability density function). For example, if $f(x) = 0$ for $x \geq 1$, but $f(x) = 1000$ for any $x < 1$, then the function $f(x)$ is not continuous at $x = 1$. It is considered to be not "robust" in the sense that a small change in x around 1 would cause a big change in $f(x)$. The notion of continuity of function can be easily extended to robustness of a functional in this way. That is, we say a functional (estimator) is robust if small changes of function can only cause small change in the functional.

The Prohorov metric on the function space $\mathcal{F}$ was first proposed to measure the neighborhoods of a function. It is also used to define the continuity of a functional. Other suitable metrics on $\mathcal{F}$ such as the Kolmogorov metric are employed most often nowadays due to their theoretical convenience.

The Kolomogorov metric is defined as $D_K(F, G_n) := \sup_x |F(x) - G_n(x)|$. A functional T is qualitative robust if for any distribution $G_n(x)$, $D_K(F, G_n) \to 0$ as $n \to 0$ implies $T(G_n) \to T(F)$. For example, the mean μ of a distribution, denoted by a location functional $T(F)$, can not be qualitative robust, because one can always find a sequence of distributions $H_n(x)$, such that $D_K(F, H_n) \to 0$ as $n \to 0$ but $T(H_n) \nrightarrow \mu$. Details are given by Staudte & Sheather (1990, p.66).

The continuity assumption of the statistical functional plays a central role in this paper. It is a necessary condition for the existence of the influence function, which we will discuss in the following sections.

1.2.2 Infinitesimal Robustness by Influence Function

The second robustness concept is the infinitesimal robust. It can be regarded to the differentiability of a functional. A better understanding of infinitesimal robustness is to think of the differentiability of a function $f(x)$, (again, not necessarily the probability density function). Suppose we want to know what constraints should be imposed on $f(x)$ so that small change of x would not lead to a large change in $f(x)$. One may require more rigorous condition such that $f(x)$ is differentiable and its derivative is small or at least bounded. In the context of statistical functional $T(F)$, the "derivative" we consider here is called the *influence function*.

The following definitions of functional derivative are shown to be useful in describing the robustness.

Definition 2 *Let F and G be two cdf's that belong to $\mathcal{F}$. The functional T is differentiable in the Gâteau sense in F in the direction of G, if there exists the limit*

$$T'_G(F) = \lim_{t \to 0+} \frac{T(F + t(G - F)) - T(F)}{t}.$$

$T'_G(F)$ is called the Gâteau derivative T in F in the direction of G.

Definition 3 *Let $\mathcal{X}$ be a separable and complete metric space with metric d, denoted $\mathcal{B}$ the σ-field of its Borel subsets. $F \in \mathcal{F}$, the system of all probability measures on space $(\mathcal{X}, \mathcal{B})$. The functional*

T is differentiable in F in the Frêchet sense, if there exists a linear functional $L_F(G-F)$ such that

$$\lim_{t\to 0}\frac{T(F+t(G-F))-T(F)}{t}=L_F(G-F)$$

uniformly for $G\in\mathcal{F}$, $d(F,G)\le C$ for any fixed $C\in(0,\infty)$. The linear functional $L_F(G-F)$ is called the Frêchet derivative of functional T in F in the direction G.

Remark 4

a *If we denote $\phi(t)=T((1-t)F+t(G))$, where $0\le t\le 1$, the Gâteau derivative $T'_G(F)$ equals the ordinary right derivative of $\phi'(0+)$, It is obvious that differentiability of T in the Frêchet sense implies its differentiability in Gâteau sense, and Gâteau derivative $T'_G(F)=L_F(G-F)$.*

b *If T is differentiable in the Frêchet sense, by Riesz Representation theorem, there exists a funciton $h:\mathcal{X}\mapsto\mathbb{R}$ such that*

$$T'_G(F)=L_F(G-F)=\int_{\mathcal{X}}hd(G-F).$$

Specifically, define Dirac Probability measure $\Delta_{\boldsymbol{x}}=1$ *to the singleton set* $\{\boldsymbol{x}\}$, 0 *elsewhere,*

$$T'_{\Delta_{\boldsymbol{x}}}(F)=\int_{\mathcal{X}}hd(\Delta_{\boldsymbol{x}}-F)=h(x)-\int_{\mathcal{X}}hdF.$$

Furthermore, abbreviating $T'_{\Delta_{\boldsymbol{x}}}(F)$ to $T'_{\boldsymbol{x}}(F)$ we have,

$$\mathbb{E}_F(T'_{\boldsymbol{x}}(F))=\int_{\mathcal{X}}[h(\boldsymbol{x})-\int_{\mathcal{X}}hdF]dF=\mathbf{0}. \tag{1.5}$$

This equality is useful when we discuss influence function of any statistical functional.

c *If denote $\phi(t)=T((1-t)F+tF_n)$, for $0\le t\le 1$, the Taylor expansion at u can be shown as*

$$\phi(t)=\phi(u)+\sum_{k=1}^{n-1}\frac{\phi^{(k)}(u)}{k!}(t-u)^k+\frac{\phi^{(n)}(v)}{n!}(t-u)^n,\ \ v\in[u,t].$$

In special case when $t = 1$, $u = 0+$, and by the first order Taylor expansion, we have

$$T(F_n) - T(F) = T'_{F_n}(F) + \frac{1}{2}\left[\frac{d^2}{dt^2}T(F + t(F_n - F))\right]_{t=v\in(0,1)} \tag{1.6}$$

Definition 5 *The Gâteau derivative of functional T in distribution F in the direction of $\Delta_{\boldsymbol{x}}$, $\boldsymbol{x} \in \mathcal{X}$ is called the influence function of T in F, thus*

$$IF(\boldsymbol{x}, T; F) = T'_{\boldsymbol{x}}(F) = \lim_{t\to 0+} \frac{T(F_t(\Delta_{\boldsymbol{x}})) - T(F)}{t}$$

where $F_t(\Delta_{\boldsymbol{x}}) = (1-t)F + t\Delta_{\boldsymbol{x}}$.

Note that $IF(\boldsymbol{x}; T, F)$ describes the effect of an infinitesimal contamination at the point $\boldsymbol{x}$ to functional T. One can define a *global* sensitivity of the functional T under the distribution F to be

$$\gamma^* = \sup_{\boldsymbol{x}\in\mathcal{X}} \|IF(\boldsymbol{x}, T; F)\|_I$$

where $\|\cdot\|_I$ is a proper norm defined accordingly in the range space of T. For example, it is defined to be the absolute value function if $T \in \mathbb{R}$, Euclidean norm if $T \in \mathbb{R}^d$, a matrix norm if $T \in \mathbb{R}^p \times \mathbb{R}^p$, etc.

A functional T is considered to be infinitesimal robust if it has a bounded influence function w.r.t $\boldsymbol{x}$, and therefore a finite γ^*.

Example 6 *Consider the following statistics in the univariate case where the empirical distribution is simplified as $F_n(x) = 1/n\sum_{i=1}^n I[x_i \leq x],\ x \in \mathbb{R}$.*

(a) Expected value:

Let

$$T(F) = E(X) = \int_{\mathbb{R}} xdF,$$

then

$$T(F_n) = \int_{\mathbb{R}} xdF_n = \frac{1}{n}\sum_{i=1}^{n} x_i,\ \text{denoted by } \bar{x}_n$$

Further, if

$$\begin{aligned}\phi(t) &:= T((1-t)F + t\Delta_x) \\ &= \int_{\mathbb{R}} X d((1-t)F + t\Delta_x) \\ &= (1-t)E(X) + tx\end{aligned}$$

so

$$IF(x; E(X), F) = T'_x(F) = \phi'(0+) = x - \mathrm{E}(X) \tag{1.7}$$

Then $\gamma^* = \infty$. Hence the expected value is not infinitesimal robust in that global sense.

(b) Quantiles:

For an univariate random variable X with cdf F, the qth quantile, $X_q = \inf\{x \in \mathbb{R} | F(x) \geq q\}$, where $0 < q < 1$. For example, the median of X equals $X_{.5}$. In particular, if X is continuous random variable, $X_q = F^{-1}(q)$. Let $f(x)$ be the probability density function associated with $F(x)$. The influence function of qth quantile thus has the following form,

$$IF(x; X_q, F) = \begin{cases} \dfrac{q-1}{f(X_q)}, & \textit{if } x < X_q \\ 0, & \textit{if } x = X_q \\ \dfrac{q}{f(X_q)}, & \textit{if } x > X_q. \end{cases}$$

Then γ^* is bounded. Hence the qth quantile is infinitesimal robust in the global sense.

(c) Variance:

Let

$$T(F) = \mathrm{var}(X) = \int_{\mathbb{R}} x^2 dF - (E(X))^2,$$

then

$$T(\hat{F}_n) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\bar{x}_n)^2, \text{ denoted by } \mathrm{S}_\mathrm{n}^2.$$

If

$$\begin{aligned}\phi(t) &:= T((1-t)F + t\Delta_y) \\ &= \int_{\mathbb{R}} x^2 d((1-t)F + t\Delta_y) - \Big(\int_{\mathbb{R}} x d((1-t)F + t\Delta_y)\Big)^2 \\ &= (1-t)E(X^2) + ty^2 - (1-t)^2(E(X))^2 - 2t(1-t)yE(X) - t^2y^2\end{aligned}$$

then

$$\begin{aligned}IF(y; \mathrm{var}(X), F) &= T'_y(F) = \phi'(0+) \\ &= y^2 - E(X^2) - 2xE(X) + 2(E(X))^2 \\ &= (y - E(X))^2 - \mathrm{var}(X) \qquad (1.8)\end{aligned}$$

γ^* *is unbounded in* y *by the first term in* (1.8)*, so* $\mathrm{var}(X)$ *is sensitive (not infinitesimal robust) in global sense.*

(d) Median Absolute Deviation (MAD)

The median absolute deviation ω *is defined implicitly by*

$$P(|X - X_{.5}| \leq w) = .5$$

That is, the ω *is the median of the distribution of* $|X - X_{.5}|$*, the distance between random variable* X *and its median. Its influence function is lengthy and shown in Wilcox (2005).*

$$IF(x; w, F) = \frac{\mathrm{sign}(|x - X_{.5}| - \omega) - \dfrac{f(X_{.5}+\omega) - f(X_{.5}-\omega)}{f(X_{.5})}\mathrm{sign}(x - X_{.5})}{2[f(X_{.5}+\omega) + f(X_{.5}-\omega)]}.$$

Assuming $f(X_{.5})$ *and* $2[f(X_{.5}+\omega) + f(X_{.5}-\omega)]$ *are not equal to* 0*, then MAD has a bounded* γ^* *and therefore robust in global sense. In addition, if density function of* F *is symmetric around*

0, *then* $X_{.5}$*=0, and* $f(\omega) = f(-\omega)$. *The influence function can be further written as*

$$IF(x; w, F) = \frac{\text{sign}(|x| - w)}{4f(\omega)} \tag{1.9}$$

This result is useful in our derivation of influence function of MRCM in the Chapter 4.

It is well known that the continuity is a necessary condition for differentiability of a given function. Since the qualitative robustness is defined as the continuity of given functional, the infinitesimal robustness is thus deemed to be stronger by requiring the boundedness condition on derivative of the given functional. In fact, Jurečková & Picek (2006, section 2.4) pointed out that a qualitative robust estimator can not automatically guarantee its bounded influence function and they provide a counter example.

In addition, not only being a tool to measure the robustness, the influence function is also convenient to be used for finding the asymptotic distribution of the empirical functional $T(F_n)$, see Jurečková & Picek (2006). This is due to the Central Limit Theorem and the fact that $E_F(IF(\boldsymbol{x}; T, F)) = \mathbf{0}$, as shown in (1.5),

Theorem 7 *If* T *is Frêchet differentiable, for metric space* $(\mathcal{F}, d)$ *of all probability distributions on* $(\mathcal{X}, \mathcal{B})$, $\hat{F}_n, F \in \mathcal{F}$ *satisfy condition* $\sqrt{n}d(\hat{F}_n, F) = O_F(1)$, *as* $n \to \infty$, *and* $\text{var}_F(IF(\boldsymbol{X}; T, F)) = \mathbb{E}_F(IF(\boldsymbol{X}; T, F))(IF(\boldsymbol{X}; T, F))^T$ *is positive semidefinite, then*

$$\big(\sqrt{n}(T(\hat{F}_n) - T(F))\big) \to \mathcal{N}\big(\mathbf{0}, \text{var}_F(IF(\boldsymbol{X}; T, F))\big). \tag{1.10}$$

Example 8 *(a) Expected value: If* $T(F) := E(X)$ *then* $T(F_n) = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. *By* (1.7)

$$\begin{aligned} \text{var}(IF(x; \text{E}(X), F)) &= \text{E}(IF(x; \text{E}(X), F))^2 \\ &= \text{E}(X^2) - \big(\text{E}(X)\big)^2 = \text{var}(X) \end{aligned}$$

We therefore have the following classical representation by (1.10),

$$\sqrt{n}(\bar{X} - \text{E}(X)) \to \mathcal{N}(0, \text{var}(X)). \tag{1.11}$$

(b) Variance: If $T(F) := var(X)$ then $T(F_n) = \dfrac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{X}^2$. By (1.8)

$$\mathrm{var}\big(IF(x; \mathrm{var}(X), F)\big) = E(x - E(X))^4 - (\mathrm{var}(X))^2$$

Again, by (1.10)*, we have the asymptotic distribution of S_n^2*

$$\sqrt{n}\big(S_n^2 - \mathrm{var}(X)\big) \to \mathcal{N}\Big(0, E(x - E(X))^4 - (\mathrm{var}(X))^2\Big).$$

1.2.3 Quantitative Robustness by Breakdown Point

Breakdown point is another popular tool in robustness analysis. It provides a measurement of robustness in a global sense. There are two types of breakdown points (BP): the "*addition / replacement* breakdown point". Roughly speaking, they are the minimal proportion of points to be added / replaced in the original data set so that the estimator on the new data set deviates from the original estimator beyond any boundary. We feel the replacement breakdown point more likely to describe the realistic condition. So, in this dissertation, the use of terminology breakdown point and RBP are interchangeable without a further mention. Different from the influence function indicating the effect brought by a single "bad" point, the breakdown point reflects the capacity of a statistic that how many "bad" points it can handle. Since they both return a measurement in numerical value, some people also consider the influence function as one of quantitative robustness measurements.

To give the formal definition, we start with a random sample $\mathbb{X}^{(0)} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$ and consider the corresponding $T_n(\mathbb{X}^{(0)})$ be an estimator of functional T. Suppose for this "initial" sample, we can replace any m data points by arbitrary values, even the unrealistic large quantity ∞. Denote the new sample after replacement as $\mathbb{X}^{(m)} = \{\boldsymbol{x}_1^*, ..., \boldsymbol{x}_m^*, \boldsymbol{x}_{m+1}, ...\boldsymbol{x}_n\}$, and $T_n(\mathbb{X}^{(m)})$ is the estimator of T based on the new sample. The (finite sample) breakdown point of the estimator T_n for sample $\mathbb{X}^{(0)}$ is

$$\varepsilon(T_n, \mathbb{X}^{(0)}) = \frac{m^*(\mathbb{X}^{(0)})}{n}$$

where $m^*(\mathbb{X}^{(0)})$ is the smallest integer m such that

$$\sup_{\mathbb{X}^{(m)}} \|T_n(\mathbb{X}^{(m)}) - T_n(\mathbb{X}^{(0)})\|_B = \infty$$

The asymptotic breakdown point $\tilde{\varepsilon}$ is considered when the sample size goes large, that is,

$$\tilde{\varepsilon} = \lim_{n\to\infty} \varepsilon(T_n, \mathbb{X}^{(0)}).$$

$\|\cdot\|_B$ is defined differently according to the functional T_n. If T_n is location estimator in $\mathbb{R}$ or $\mathbb{R}^p$, $\|\cdot\|_B$ would be absolute value or Euclidian norm. If T_n is scatter estimator in $\mathbb{R}^p \times \mathbb{R}^p$, $\|\cdot\|_B$ has to be specially defined. Because T_n can breakdown in either way of *explosion* (its biggest eigenvalue tends to infinity) or *implosion* (its smallest eigenvalues tends to zero). For instance, the scale estimator (correspondence of scatter estimator in the univariate case), sample variance or MAD has breakdown if it equals to 0 or ∞. We will give more details about the method to define $\|\cdot\|_B$ in the Chapter 4 when we consider the breakdown of scatter estimator in high dimensional spaces.

Example 9 *Under the univariate case, for any initial random sample $\mathbb{X}^{(0)}$:*

(a) The sample mean $\bar{X}_n$ and sample variance S_n^2 have breakdown point $\varepsilon(\bar{X}_n, \mathbb{X}^{(0)}) = \varepsilon(S_n^2, \mathbb{X}^{(0)}) = \frac{1}{n}$, so the asymptotic breakdown point $\tilde{\varepsilon} = \lim_{n\to\infty} \varepsilon(\bar{X}_n, \mathbb{X}^{(0)}) = \lim_{n\to\infty} \varepsilon(\bar{S}_n^2, \mathbb{X}^{(0)}) = 0$

(b) The qth quantile has $\tilde{\varepsilon}$=$\min(q, 1-q)$. In particular, the sample median $x_{.5} = x_{\lceil \frac{n+1}{2} \rceil}$ (for simplicity, only consider n being odd) have breakdown point $\varepsilon(X_{.5}, \mathbb{X}^{(0)}) = \frac{n+1}{2n}$, and the asymptotic breakdown point $\tilde{\varepsilon} = 1/2$

(c) MAD has asymptotic breakdown point $\tilde{\varepsilon} = 1/2$, the highest level for an equivariant scale estimator.

1.3 Statistical Efficiency

To give a comprehensive assessment of any proposed statistical method, the notion of *statistical efficiency* has to be reported. Efficiency is a measure of optimality of an estimator. To put it simple, an estimator is said to be more efficient than the other if it needs a smaller sample size to achieve the same accuracy level under a given distribution. Here, accuracy often refers to the standard error or mean square error of the estimator. Comparing efficiency of different estimators of a parameter is often done by comparing their accuracy for a fixed sample size under a given data distribution. If the estimator is naturally presented in a matrix form, such as scatter estimators in the multivariate case, distinct criteria of accuracy may apply, see Chapter 4.

Two general approaches exist to estimate standard error (accuracy) of an estimator. One way is to develop the mathematical formula of the asymptotic standard error of the estimator. Most of time, this would require the use of Central Limit Theorem, such as (1.10). However, if the algebraic expression is too difficult to derive or the standard error of the estimator can not be written out as an explicit form, one can use *bootstrap* to estimate it. Specifically, bootstrap can be done by constructing a number of resamples of the observed dataset, each of which is obtained by random sampling with replacement on the original dataset. For each resample, one estimate is computed. Then the standard error of the estimator can further be estimated by the standard deviation of these bootstrap estimates.

In practice, the standard error of location estimator developed by the first approach often involves the an estimation of scale, e.g. (1.11). So, when seeking for an estimator of scale, robustness is considered to be more important than efficiency. For instance, using the sample standard deviation to estimate the standard error (accuracy) of measure of location (e.g., sample mean) will widen the confidence interval if outliers exist. From here we can see, a robust estimator of scale is especially important regarding to the statistical inference. Typically, in the way of choosing or constructing a robust estimator, we hope to balance the following two situations: (1) The standard error would not be inflated due to outliers or the heavy-tailed distributions; (2) The standard error is not too large comparing to the most efficient estimators under the target distribution like Gaussian distribution. An interesting discussion on how to achieve this goal can be found by Wilcox (2005, p.57). In the context, a trimming parameter of sample trimmed mean is tuning to balance both properties. Furthermore, in Section 4.3, we would show the relative statistical efficiency of our proposed method by considering of this two situations simultaneously. Previous sections are served as an introduciton of necessary conpets with simple examples on the univariate case. From now on, we will focus out discussion on the multivaraite case and robust scatter estimators.

1.4 Robust Affine Equivariant Estimator for Multivariate Model

Similar to the parameters location and scale in the univariate case, i.e. (1.1)- (1.4), the location and scatter parameters in multivariate case should possess the same properties. The generalization of these characteristics from one dimension to high dimensions is called *affine equivariance*. They have different meanings respectively. Assume $\boldsymbol{X} \in \mathbb{R}^p$, for measure of location, affine equivariance requires that for any $p \times p$ non-singular matrix $\boldsymbol{A}$ and p-vector $\boldsymbol{b}$, if T_n is a location estimator,

$$T_n(\boldsymbol{AX}_1 + \boldsymbol{b}, ..., \boldsymbol{AX}_n + \boldsymbol{b}) = \boldsymbol{A}T_n(\boldsymbol{X}_1, ..., \boldsymbol{X}_n) + \boldsymbol{b}.$$

For measure of scatter, affine equivariance requires that for the same matrix $\boldsymbol{A}$, if V_n is scatter estimator,

$$V_n(\boldsymbol{AX}_1 + \boldsymbol{b}, ..., \boldsymbol{AX}_n + \boldsymbol{b}) = AV_n(\boldsymbol{X}_1, ..., \boldsymbol{X}_n)A^T.$$

Donoho & Gasko (1992, p.1811) indicated that the breakdown point has an upper bound $(n - p + 1)/(2n - p + 1)$ for any affine equivariant estimator. The scatter estimator we shall propose in the next chapter is designed to preserve the affine equivariance and attain the highest possible breakdown point in the same time under the assumption of elliptically symmetric distributions.

In the multivariate case, the classical sample mean (componentwise average) and sample covariance matrix are examples of affine equivariant location and scatter estimator. Applied researchers commonly use them to characterize the distribution or make statistical inferences without a careful check. However, like their univariate counterparts, they are extremely sensitive to outliers with an unbounded influence function and low breakdown point (asymptotically 0). Before the decade of 1960's, methods for dealing with outliers were ad hoc. Not until Tukey wrote a paper in 1960 discussing the contaminated normal distribution, did statisticians gather around to address those technical issues. Moreover, the rapid growth of computational power provides a huge support for developing more sophisticated methods to deal with high-dimensional noisy data sets.

Many alternatives were proposed to replace the classical methods that are sensitive to outliers. Variant robust affine equivariant estimators have been discussed about their robustness, statistical efficiency, and computational complexity. We will briefly review some of them in the following.

1.4.1 Minimum Volume Ellipsoid and Minimum Covariance Determinant

Rousseeuw (1985) consecutively introduced the Minimum Volume Ellipsoid (MVE) and Minimum Covariance Determinant (MCD) estimators. MVE location estimator is the center of the smallest regular ellipsoid containing h data points (out of n). The scatter estimator $\mathbf{\Sigma}$ is then defined by the ellipsoid shape matrix. MCD is a variant of the MVE. Instead of considering the ellipsoid, the objective is to find the h observations (out of n) whose classical covariance matrix has the lowest determinant. The MCD location and scatter estimator are then given by the center and covariance matrix on this collection of h observations. h is set to be between $n/2$ and n, and usually taken to be $n/2+1$ or rounded to the nearest integer to cover half of the data points. Both methods are making decision on tightly clustered data in order to reduce effects from outliers. They have the a same break down point $(n-h)/n$, but the MVE is less attractive by its lower statistical efficiency compared to MCD, see Davies (1992).

The major drawback of both MVE and MCD is their computational complexity. It is generally difficult to exhaust all the subsets containing half of data when sample size n is large. Rousseeuw & Leroy (1987) proposed a basic resampling algorithm to approximate the MVE, called MINVOL. Rocke & Woodruff (1993) then find an algorithms combining the resampling principle with other heuristic search techniques. In contrast to MVE, the computation time of MCD is greatly reduced by Rousseeuw & Driessen (1998). They proved a theorem called *C-step* and further proposed an extremely fast algorithm, even in high dimensions, to compute the MCD. The use of MCD becomes popular since then. In addition, Croux & Haesbroeck (1999) studied the bounded influence function of MCD scatter estimator and compute the asymptotic variances of its elements based on the similar idea as (1.10).

1.4.2 M-estimator

The class of M-estimators was introduced by Huber (1964) from the estimation of a univariate location parameter. The name "M-estimator" comes from "generalized maximum likelihood", which can be a solution of a minimization problem

$$\sum_{i=1}^{n} \rho(\boldsymbol{X}_i, \boldsymbol{\theta}) := \min \quad \text{with respect to } \boldsymbol{\theta} \in \boldsymbol{\Theta}$$

where $\rho(\cdot,\cdot)$ is a properly chosen function.

Maronna (1976) was the first one to define M-estimators for multivariate location $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. Huber & Ronchetti (2009) extended Maronna's definition and defined the simultaneous M-estimator as solutions of following system of equations:

$$\begin{aligned} &\tfrac{1}{n}\textstyle\sum_{i=1}^{n} v_1(d_i)(\boldsymbol{X}_i - \boldsymbol{\mu}) = \boldsymbol{0} \\ &\tfrac{1}{n}\textstyle\sum_{i=1}^{n} \{v_2(d_i^2)(\boldsymbol{X}_i - \boldsymbol{\mu})(\boldsymbol{X}_i - \boldsymbol{\mu})^T - v_3(d_i)\boldsymbol{\Sigma}\} = \boldsymbol{0}, \end{aligned} \tag{1.12}$$

where $d_i = \sqrt{(\boldsymbol{X}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{X}_i - \boldsymbol{\mu})}$ is the Mahalanobis distance. Maronna (1976) showed the existence and uniqueness of the solution above for the special case $v_3 = 1$. Huber & Ronchetti (2009) also studies the robustness of these estimators by showing the breakdown point, which is typically at most $1/(p+1)$, where p is the dimension. The influence function is bounded when v_1 and v_2 are suitably chosen. Therefore, from the prospective of breakdown, M-estimator becomes less robust to outlier in higher dimensions. But from the prospective of influence function, M-estimator is infinitesimal robust.

1.4.3 S-estimator

S-estimator was first introduced in a regression context by Rousseeuw & Yohai (1984). From its direct generalization, the S-estimator of multivariate location $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ is the solution of

$$\text{Minimize } \det(\boldsymbol{\Sigma}), \tag{1.13}$$

s.t.

$$\frac{1}{n}\sum_{i=1}^{n} \rho\Big(\sqrt{(\boldsymbol{X}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{X}_i - \boldsymbol{\mu})}\Big) = b_0$$

where ρ is often chosen to be *Tukey's biweight* ρ_B function

$$\rho_B(y, c_0) = \begin{cases} \dfrac{y^2}{2} - \dfrac{y^4}{2c_0^2} + \dfrac{y^6}{6c_0^4}, & \text{for } |y| \le c_0; \\ \dfrac{c_0^2}{6}, & \text{for } |y| \ge c_0. \end{cases}$$

It is worth noting that S-estimator of location and covariance can also be viewed as a robust version of the least squared estimator if $b_0 = p$ and $\rho = y^2$. In that case, the classical sample mean and sample covariance matrix would be the solution.

S-estimator can reach the maximal breakdown point $\lfloor (n-p+1)/2 \rfloor / n$, or asymptotically 0.5, when the ratio of $\frac{6b_0}{c_0^2} = (n-p)/2n$, see Lopuhaä & Rousseeuw (1991). However, the asymptotic variance of estimators is positively proportional to c_0. Therefore, it is not possible to achieve the most efficiency and highest breakdown point at the same time. Lopuhaä (1989) further discussed existence and continuity of S-estimator to obtain the influence function. He also indicated that the S-estimator is a type of M-estimator. However, statisticians often refer M-estimators to those that have low breakdown points and are solutions of the implicit equations (1.12), with decreasing function of v_2 and constant function $v_3 = 1$. The S-estimators, instead, are associated with totally different $v_2(\cdot)$ and $v_3(\cdot)$, and have a high breakdown point. Nevertheless, the asymptotic behaviors and influence functions of M-estimators and S-estimators are the same.

1.5 Dissertation Overview

The rest of this dissertation will be formalized in the following way. Chapter 2 will briefly go over two types of sign and rank concepts, the marginal sign and rank and Oja sign and rank functions. Chapter 3 will discuss the spatial sign and rank, which is different from Chapter 2, the detailed explanation of spatial rank function, spatial rank covariance matrix (RCM) and a modified spatial rank covariance matrix (MRCM) will be present. Chapter 4 will give the theoretical robustness results of MRCM in terms of the influence function and breakdown point. Further, finite sample efficiency is also showed by comparing with other robust methods. In Chapter 5, the mixture model and the regular EM algorithm are reviewed. The problems associated with the regular EM algorithm are summarized. Some existing procedures are suggested to resolve them. In particular, the undue influence of outliers in mixture of Gaussian model is discussed in detail. In Chapter 6, a robust Spatial-EM algorithm that integrates the notion of MRCM is proposed to solve the problem brought by outliers. The algorithm is also analyzed from the likelihood point of view. Finally, in Chapter 7, experiments of using the robust Spatial-EM on outlier detection and clustering are implemented. Results are compared to the regular EM and some existing techniques in statistical learning or data mining. Chapter 8 will give the concluding remarks and possible future work.

Chapter 2

Two Types of Sign and Rank Covariance Matrices

2.1 Preliminaries

Classical multivariate statistical inference methods including multivariate analysis of variance, principal component analysis, factor analysis, canonical correlation analysis are based on the covariance matrix. Those moment-based techniques, e.g., sample mean and sample covariance matrix, are optimal (most efficient) under the normality distributional assumption. They are, however, extremely sensitive to outlying observations, and poor in the efficiency for heavy-tailed distributions. A straightforward treatment is to replace the sample covariance matrix with a robust one. A variety of robust estimators of scatter matrix have been reviewed in previous chapter. Besides from those M-estimators, S-estimators, MCD-estimators mentioned before, sign and rank covariance estimates (Visuri *et al.*, 2000) received more attention recently. In this chapter, we focus on one branch of these estimators.

Sign and rank functions defined in the univariate case are simple and intuitive. However, when comes to multivariate case, the correspondences become interesting. They can be defined differently with different forms with different points of view. This idea also applies to the related sign and rank covariance matrices. Chapter 2 would give a brief review of these scatter estimators. They are marginal and Oja sign and rank covariance matrices. After that, in Chapter 3 we would focus on the discussion of the spatial sign and rank covariance matrices and a modified version of spatial rank covariance matrix.

2.2 Elliptical Models and Role of Covariance

Before continue, it would be wise to declare the general model assumption we mainly interested in and explain why covariance matrix (scatter parameter) is so important. We know that for the multivariate Gaussian distribution, the covariance matrix determines the shape of distribution, i.e. whether the density region is flat and sparse or concentrated. In fact, for a broader family called elliptical models, in which a multivariate Gaussian distribution is just a special case, the covariance matrix plays the same vital role.

A distribution is *elliptical* or *elliptically symmetric* if it has a density of the form

$$f(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \{det(\boldsymbol{\Sigma})\}^{-1/2} h\{(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\}, \tag{2.1}$$

for some $\boldsymbol{\mu} \in \mathbb{R}^p$, a positive definite symmetric $p \times p$ matrix $\boldsymbol{\Sigma}$, and a nonnegative function h with $\int_0^\infty t^{p/2-1} h(t) dt < \infty$ independent to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. The parameter $\boldsymbol{\mu}$ is the symmetric center of the distribution and it equals the first moment $E\boldsymbol{X}$ if it exists, while the scatter parameter $\boldsymbol{\Sigma}$ is proportional to the covariance matrix Cov($\boldsymbol{X}$) when it exists. In the case of the multivariate t distribution with degrees of freedom $\nu > 0$, h in (2.1) is of the form $h(t) = c(\nu, p)(1+t/\nu)^{-(p+\nu)/2}$, where $c(\nu, p)$ is the normalization constant. For $\nu > 2$, the covariance matrix Cov$(\boldsymbol{X}) = \nu/(\nu - 2)\boldsymbol{\Sigma}$. For $\nu = 1$, it is called p-variate Cauchy distribution. It has very heavy (fat) tails so that even the first moment doesn't exist. When $\nu \to \infty$, it yields the Gaussian distribution with $h(t) = (2\pi)^{-p/2} e^{-t/2}$, such that all moments exist and Cov$(\boldsymbol{X}) = \boldsymbol{\Sigma}$.

Bensmail & Celeux (1996) presented that the (theoretical) covariance matrix can be eigendecomposed in the form

$$\boldsymbol{\Sigma} = \lambda \boldsymbol{U} \boldsymbol{C} \boldsymbol{U}^T,$$

where $\boldsymbol{U}$ is the matrix of eigenvectors, $\boldsymbol{C}$ is the diagonal matrix with normalized eigenvalues c_i's such that $\prod_{i=1}^p c_i = det(\boldsymbol{C}) = 1$ and λ^p is the Wilks generalized variance. λ, $\boldsymbol{C}$ and $\boldsymbol{U}$ are described as *scale*, *shape* and *orientation* respectively. The idea of our modified rank covariance matrix is originated from here. It is believed that if we can robustly estimate the eigenvector matrix and eigenvalues separately, we can have a robust estimator of covariance matrix.

In general, the diagonal matrix $\mathbf{\Lambda} = \lambda \boldsymbol{C}$ is the usual eigenvalue matrix. Wilks generalized variance, one of the "global" measure of the multivariate scatter, is just the geometrical mean of the eigenvalues to the power of p, which also equals $det(\mathbf{\Sigma})$. Another "global" measurement of the multivariate scatter is the sum of eigenvalues $trace(\mathbf{\Sigma}) = trace(\mathbf{\Lambda})$. Keep in mind that these measurements also provide ways in defining the breakdown of scatter estimators.

As shown in the Section 1.4, affine equivariance of a scatter estimator is an important property that any robust scatter estimator should strive for. Under an elliptical distribution assumption, this turns to be especially significant. We shall see in the following sections, some of the robust scatter estimators do not fully possess the affine equivariance. This is again the reason why we want to propose a modified version of them.

2.3 Marginal Sign and Rank Covariance Matrices

The marginal sign function in high dimensions can be componentwisely generalized from the univariate case. Recall that in the case $x \in \mathbb{R}$, the sign function $sign(x)$ takes value $1, 0$ or -1 as $x > 0$, $x = 0$ or $x < 0$. The (sample) sign and rank functions associated with a random sample $\{x_1, ..., x_n\}$ are defined below

$$S(x, F_n) = sign(x - Med(x_1, ..., x_n)),$$

and

$$R(x, F_n) = ave\{S(x - x_i)\} = \frac{1}{n}\sum_{i=1}^{n} sign(x - x_i).$$

Note that $R(x)$ is in fact the derivative of criterion function $ave\{|x - x_i|\}$. Notation ave is the average taking on the index i. In here, it is equivalent to $\frac{1}{n}\sum_{i=1}^{n}$.

For p-variate data set $\mathbb{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$, where $\boldsymbol{x}_i = [x_{i1}, ..., x_{ip}]^T \in \mathbb{R}^p$, consider the objective functions

$$H_1(\boldsymbol{x}) = \|\boldsymbol{x}\|_1 = |x_1| +, ..., +|x_p|,$$

and

$$D_1(\boldsymbol{x}) = ave\{\|x - x_i\|_1\}.$$

One can define the marginal sign function $\boldsymbol{S}_1(\boldsymbol{x})$ and marginal rank function $\boldsymbol{R}_1(\boldsymbol{x})$ as the gradient of the above objective functions,

$$\boldsymbol{S}_1(\boldsymbol{x}) = \nabla_{\boldsymbol{x}} H_1(\boldsymbol{x}),$$

$$\boldsymbol{R}_1(\boldsymbol{x}) = \nabla_{\boldsymbol{x}} D_1(\boldsymbol{x}).$$

So the $\boldsymbol{S}_1(\boldsymbol{x}) = [sign(x_1), ..., sign(x_p)]^T$. The vector of marginal rank of $\boldsymbol{x}$ is $\boldsymbol{R}_1(\boldsymbol{x}) = ave\{\boldsymbol{S}_1(\boldsymbol{x} - \boldsymbol{x}_i)\}$. The marginal sign of $\boldsymbol{x}$ is $\boldsymbol{S}_1(\boldsymbol{x} - \boldsymbol{M}_1(\mathbb{X}))$, where $\boldsymbol{M}_1(\mathbb{X})$ is the marginal median (also called the componentwise median) which minimizes the criterion function $D_1(\boldsymbol{x})$. The marginal median also satisfies the equality

$$\boldsymbol{R}_1\left(\boldsymbol{M}_1(\mathbb{X})\right) = \mathbf{0}.$$

The corresponding (sample) sign covariance matrix (SCM) is defined as

$$SCM_1 = ave\{(\boldsymbol{S}_1\left(\boldsymbol{x}_i - \boldsymbol{M}_1(\mathbb{X})\right)\left(\boldsymbol{S}_1\left(\boldsymbol{x}_i - \boldsymbol{M}_1(\mathbb{X})\right)^T\},$$

and (sample) rank covariance matrix (RCM) is defined as

$$RCM_1 = ave\{\boldsymbol{R}_1(\boldsymbol{x}_i)\boldsymbol{R}_1^T(\boldsymbol{x}_i)\}.$$

Visuri *et al.* (2000) illustrated that the marginal sign and rank covariance matrices are scale invariant (re-scaling the coordinates does not change the values of the matrices). Simply put, they can not reserve the orientation and shape (eigenvectors and eigenvalues) of the original geometry (distribution) of data cloud. Also, they lack the efficiency under the Gaussian model. We thus are not interested in this type of scatter estimators even though their computation is relatively simple.

2.4 Oja Sign and Rank Covariance Matrices

The volume of p-variate simplex determined by $\boldsymbol{x}$ and p observations with indices $i_1 < i_2 < ... < i_p$ is

$$\frac{1}{p!} abs \left\{ det \begin{pmatrix} 1 & ... & 1 & 1 \\ \boldsymbol{x}_{i_1} & ... & \boldsymbol{x}_{i_p} & \boldsymbol{x} \end{pmatrix} \right\}.$$

Consider the objective functions, Visuri *et al.* (2000)

$$H_2(\boldsymbol{x}) = ave \left\{ abs \left\{ det \begin{pmatrix} 1 & 1 & ... & 1 & 1 \\ \mathbf{0} & \boldsymbol{x}_{i_1} & ... & \boldsymbol{x}_{i_{p-1}} & \boldsymbol{x} \end{pmatrix} \right\} \right\},$$

and

$$D_2(\boldsymbol{x}) = ave \left\{ abs \left\{ det \begin{pmatrix} 1 & ... & 1 & 1 \\ \boldsymbol{x}_{i_1} & ... & \boldsymbol{x}_{i_p} & \boldsymbol{x} \end{pmatrix} \right\} \right\}.$$

The Oja sign and rank functions, $\boldsymbol{S}_2(\boldsymbol{x})$ and $\boldsymbol{R}_2(\boldsymbol{x})$ are defined as the gradient functions as follows,

$$\boldsymbol{S}_2(\boldsymbol{x}) = \nabla_{\boldsymbol{x}} H_2(\boldsymbol{x}),$$

$$\boldsymbol{R}_2(\boldsymbol{x}) = \nabla_{\boldsymbol{x}} D_2(\boldsymbol{x}).$$

Oja median (see Oja, 1983), $\boldsymbol{M}_2(\mathbb{X})$, minimizes $D_2(\boldsymbol{x})$, which is the solution of $\boldsymbol{R}_2(\boldsymbol{x}) = \mathbf{0}$. The Oja sign covariance matrix SCM_2 (w.r.t. the Oja median $\boldsymbol{M}_2(\mathbb{X})$) and Oja rank covariance matrix RCM_2 can be constructed in a similar way as SCM_1 and RCM_1. By introducing the notion of simplex, both statistics are granted with "affine equivariance in the sense". That is, if $\boldsymbol{x}_i^* = \boldsymbol{A}\boldsymbol{x}_i + \boldsymbol{b}$, where $\boldsymbol{A}$ is a non-singular $p \times p$ matrix, and $\boldsymbol{b}$ is a p-variate vector, the Oja sign and rank covariance matrix on the transformed data satisfies

$$SCM_2^* = \boldsymbol{A}^* SCM_2 \boldsymbol{A}^{*T},$$

and

$$RCM_2^* = \boldsymbol{A}^* RCM_2 \boldsymbol{A}^{*T},$$

where $\boldsymbol{A}^* = abs(det(\boldsymbol{A}))(\boldsymbol{A}^{-1})^T$. See Hettmansperger *et al.* (1998). If $\boldsymbol{A}$ is orthogonal, then $\boldsymbol{A}^* = \boldsymbol{A}$, and diagonal $\boldsymbol{A} = diag(a_1, ..., a_p)$ with all positive entries, then $\boldsymbol{A}^* = diag(\frac{\prod_i a_i}{a_1}, ... \frac{\prod_i a_i}{a_p})$. This means the SCM_2 and RCM_2 carry the orientation (eigenvectors) and shape (eigenvalues) of the data distribution.

A series of papers investigate covariance matrices based on Oja sign and rank functions. For example, a regression model with coefficient estimated by sign covariance matrix was dealt by Ollila *et al.* (2002). Ollila *et al.* (2003) discuss the same issue base on Oja rank covariance matrix Ollila *et al.* (2004) explored the influence function and efficiency of the Oja rank covariance matrix. It has high efficiency under the multivariate Gaussian distribution and superior performance in heavy-tailed distributions. However, the "affine equivariance in the sense" is still different from the traditional definition of affine equivariance we give in Section 1.4. They are not robust in the usual sense. The influence function is unbounded, and the breakdown point is 0. Even being worse in practice, it needs $\binom{n}{p}$ iterations for just calculating a single rank $\boldsymbol{R}_2(\boldsymbol{x})$. Thus for data with dimension higher than 5 or reasonable large sample size, the time taken on computation is already prohibitively long.

Chapter 3

Modified Spatial Rank Covariance Matrix and Equivariance

3.1 Introduction

Comparing to the methods in Chapter 2, spatial sign and rank covariance matrices are more attractive due to its computational ease, robustness and statistical efficiency. But they are only orthogonally equivariant. In order to gain fully affine equivariance property, one approach is to utilize transformation-retransformation(TR) technique, which serves as standardization of multivariate data. More details can be found in Serfling (2010). The well-known scatter functional of Tyler (1987) is a TR version of spatial sign covariance matrix. Use the same idea, Dümbgen (1998) considered symmetrized TR spatial sign covariance matrix. Oja & Randles (2004) constructed nonparametric tests based on TR spatial rank covariance matrix. Indeed, the above mentioned TR scatter functionals are of the form of M-estimator, hence inherit the pitfall from M-estimator. The breakdown point is disappointingly low in high dimensions. Dümbgen & Tyler (2005) studied the breakdown properties of those estimators: the breakdown point for Tyler's estimator is $1/p$ and Dümbgen's estimator is $1 - \sqrt{1 - 1/p} \in (1/2p, 1/p)$, where p is the size of dimension. K-step versions of the above estimators are studied in Croux *et al.* (2010) and Taskinen *et al.* (2010). A related approach is a spatial trimming technique used by Mazumder & Serfling (2010), in which a scatter estimator is obtained based on the trimmed data with the TR version spatial outlyingness less than some threshold. The robustness depends on the value of threshold, the trimming parameter. With authors' suggestion on the parameter, the breakdown point is $1/(p + 2)$.

In this chapter, we use a different approach to obtain affine equivariance property of spatial sign and rank covariance matrices under elliptical models (see (2.1)) without sacrifice of robustness. Suggestion of modifying structure for different rank covariances is proposed by Visuri, Oja and Koivunen (2000). The basic idea is to take advantage of the fact that the spatial sign and rank functions preserve directional information but lose some measure on distance. Consequently, eigenvectors of the spatial sign and rank covariance matrices are able to capture principal components (orientation) of a data cloud (or underlying distribution), but eigenvalues no longer reflect the true variation on those directions. This is the result of Marden (1999). Our strategy is to replace each eigenvalue with an univariate scale estimator on the corresponding direction such that it depicts the proper variability. This is also the approach that Visuri *et al.* (2000) took. For consideration of robustness, the univariate scale functional must also be robust, e.g. MAD (median of absolute deviation). The spatial rank covariance matrix is in favor over the spatial sign covariance matrix because it is more statistically efficient. In addition, there is no initial location estimator needed for computing rank vectors. We call the resulting covariance matrix the modified spatial rank covariance matrix (MRCM). In the next chapter, we will study the robustness properties of MRCM by the influence function and the breakdown point. The finite sample breakdown point is also obtained. We show that the finite sample breakdown point can attain the upper bound by a proper choice of univariate scale estimator. The influence functions of eigenvalues and eigenvectors of the covariance matrix are derived and found to be bounded.

3.2 Spatial Sign, Spatial rank and Spatial depth

One can create the spatial sign or rank covariance matrix similar to the idea of the sign and rank covariance matrix quoted in Chapter 2. Nevertheless, different from the simple way as the marginal sign function taking the componentwise sign or the complex way as the Oja sign based on the notion of simplex. We consider the following two objective functions,

$$H(\boldsymbol{x}) = \|\boldsymbol{x}\| = \sqrt{x_1^2 + ... + x_p^2},$$

and

$$D(\boldsymbol{x}) = ave\{\|\boldsymbol{x} - \boldsymbol{x}_i\|\}.$$

The spatial sign function and the spatial rank function are defined as the gradient of them,

$$\boldsymbol{S}(x) = \nabla_{\boldsymbol{x}} H(\boldsymbol{x}),$$

$$\boldsymbol{R}(x) = \nabla_{\boldsymbol{x}} D(\boldsymbol{x}).$$

So the spatial sign function $\boldsymbol{S}(\boldsymbol{x}) = \boldsymbol{x}/\|\boldsymbol{x}\|$ ($\boldsymbol{S}(\boldsymbol{0}) = \boldsymbol{0}$). In fact, the spatial sign can be viewed as the unit vector in the direction of $\boldsymbol{x}$. The spatial sign of $\boldsymbol{x}$ with respect to (w.r.t.) a random sample $\mathbb{X} = \{\boldsymbol{X}_1, ..., \boldsymbol{X}_n\}$ is $\boldsymbol{S}(\boldsymbol{x}, F_n) = \boldsymbol{S}(\boldsymbol{x} - M(\mathbb{X}))$, where $M(\mathbb{X})$ is the spatial median. The (sample) spatial rank is thus derived accordingly

$$\boldsymbol{R}(\boldsymbol{x}, F_n) = ave\{\boldsymbol{S}(\boldsymbol{x} - \boldsymbol{x}_i)\} = \frac{1}{n}\sum_{i=1}^{n} \frac{\boldsymbol{x} - \boldsymbol{x}_i}{\|\boldsymbol{x} - \boldsymbol{x}_i\|}.$$

Follow by the definition of spatial sign and spatial rank, the spatial median $\boldsymbol{M}(\mathbb{X})$ is a point that minimizes the expected Euclidean norm $D(\boldsymbol{x})$. It can be defined as a solution of

$$\boldsymbol{R}\left(\boldsymbol{x}, F_n\right) = \boldsymbol{0}.$$

Small (1990) took this from the other perspective and defined it to be the point with maximal ("deepest") depth $Depth(\boldsymbol{x}, F_n)$. Vardi & Zhang (2000) called it L_1 *median*. Regardless, they are all the solution of the same equation $\|\boldsymbol{R}(\boldsymbol{x}, F_n)\| = 0$.

Notice that the spatial sign and rank functions are conceptually similar to the marginal sign and rank functions and the Oja sign and rank functions. They are only different in terms of the objective functions. But, it would be evident that this difference from where they start makes the spatial rank the prominence.

In order to be more convenient in developing the theoretical result, we would give the population version of spatial rank function as well. If $\boldsymbol{X} \in \mathbb{R}^p$ is a random variable from a distribution with cdf F, the expected Euclidean distance from $\boldsymbol{x}$ to $\boldsymbol{X}$ is $D(\boldsymbol{x}, F) = E_F\|\boldsymbol{x} - \boldsymbol{X}\|$. The spatial median of F minimizes the criterion function D w.r.t. $\boldsymbol{x}$ The multivariate centered spatial rank function is defined as the gradient of D:

$$\boldsymbol{R}(\boldsymbol{x}, F) = \nabla_{\boldsymbol{x}} D(\boldsymbol{x}, F) = E_F \frac{\boldsymbol{x} - \boldsymbol{X}}{\|\boldsymbol{x} - \boldsymbol{X}\|} = E_F\{\boldsymbol{S}(\boldsymbol{x} - \boldsymbol{X})\}.$$

The spatial rank function of $\boldsymbol{x}$ is the *expected direction* to $\boldsymbol{x}$ from $\boldsymbol{X}$. We call it centered because the rank of a random vector from the same distribution F has expected value at $\mathbf{0}$, that is, $E_F\boldsymbol{R}(\boldsymbol{X}, F) = \mathbf{0}$. It is interesting to see, the three objective functions D_1, D_2 (see Chapter 2) and D would degenerate to the same absolute value function in the univariate case. If we consider with the population version, the marginal rank, Oja Rank and spatial rank objective functions all equal to $D(x, F) = E_F|x - X|$. Their gradient functions (rank functions) lead to the univariate rank function $R(x, F) = E_F\text{sign}(x - X) = 2F(x) - 1 \in [-1, 1]$. Therefore, the marginal median, Oja median and spatial median would coincide at the same point and be equal to the regular univariate median.

The spatial rank function has many nice properties. It characterizes the distribution F (up to a location shift) (see Koltchinskii, 1997), which means that if we know the rank function, we know the distribution (up to a location shift). Under very weak assumptions on F, $\boldsymbol{R}(\boldsymbol{x}, F)$ is a one-to-one mapping from $\boldsymbol{x} \in \mathbb{R}^p$ to a vector inside the unit ball with the magnitude $\|\boldsymbol{R}(\boldsymbol{x}, F)\| \in [0, 1]$.

Chaudhuri (1996) proposed the spatial quantile based on the inverse mapping of $\boldsymbol{R}(\boldsymbol{x}, F)$. Serfling (2002) extended the notion and defined the *spatial depth*, $Depth(\boldsymbol{x}, F)$, of point $\boldsymbol{x}$ to be $Depth(\boldsymbol{x}, F) = 1 - \|\boldsymbol{R}(\boldsymbol{x}, F)\|$. Distinct from the univariate case when scaler sample points have natural ordering from small to large, in multivariate case, there is no unique method to order the data. Historically, the most popular depths used in applied statistic science are *Tukey halfspace depth* (Tukey, 1975) and various *projection depths*. There were vast amount of well-established results based on the research of these depth functions, see Wilcox (2005). Depth of a point can be used to provide the relative position of the point regarding to a data cloud or the a population distribution. Based on that, one can perform outlier detection (see figure 3.1), and statistical inference on high dimensional data. In Chapter 7, we would propose a brand new robust outlier detection method based on the notion in here.

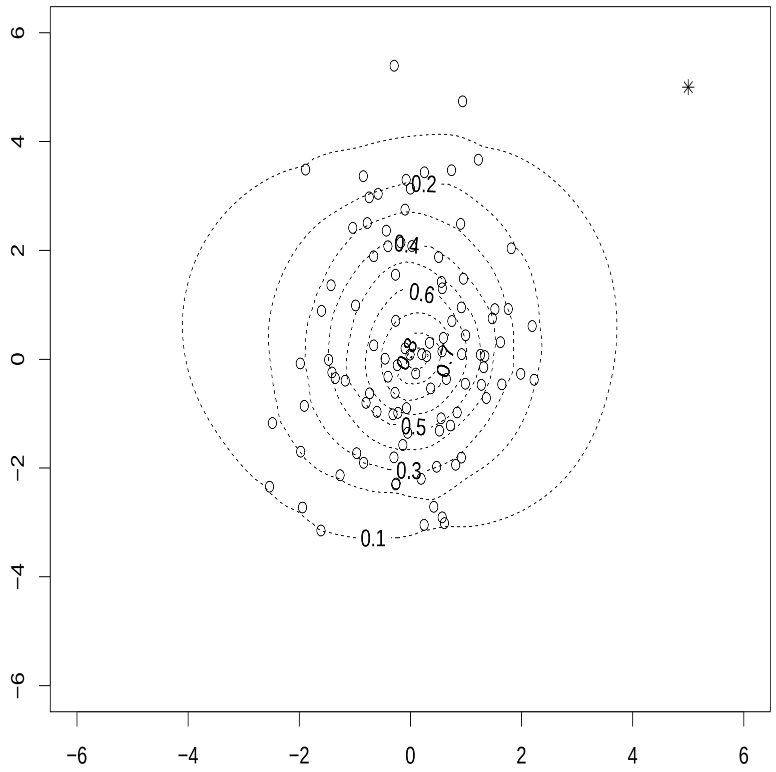


Figure 3.1. A contour plot of sample spatial depth

A contour plot of sample spatial depth based on 100 random observations (o's) from Normal Distribution. * on the upper corner is considered to be a possible outlier with $Depth(*, \mathbb{X}) = 0.0372$, which is relative low.

3.3 Spatial Rank Covariance Matrix

By convention, one can define the sample version spatial sign covariance matrix by

$$SCM = ave\{\boldsymbol{S}(\boldsymbol{X}_i - \boldsymbol{M}(\boldsymbol{X}))\boldsymbol{S}^T(\boldsymbol{X}_i - \boldsymbol{M}(\boldsymbol{X}))\},$$

and the sample version *spatial rank covariance matrix* as

$$\begin{aligned} RCM &= \boldsymbol{\Sigma}_R(F_n) = \boldsymbol{\Sigma}_R(\mathbb{X}) \\ &= ave\{\boldsymbol{R}(\boldsymbol{X}_i)\boldsymbol{R}^T(\boldsymbol{X}_i)\} \\ &= \frac{1}{n}\sum_i \boldsymbol{R}(\boldsymbol{x}_i, F_n)\boldsymbol{R}^T(\boldsymbol{x}_i, F_n) \\ &= \frac{1}{n(n-1)^2}\sum_{j,k\neq i} \boldsymbol{S}(\boldsymbol{x}_i - \boldsymbol{x}_j)\boldsymbol{S}^T(\boldsymbol{x}_i - \boldsymbol{x}_k). \end{aligned} \tag{3.1}$$

We would focus on RCM rather than SCM, since it does not require the quantity of the spatial median $\boldsymbol{M}(\boldsymbol{X})$.

If we treat the RCM as a functional of F (or random variable $\boldsymbol{X}$ for convenient), then $\boldsymbol{\Sigma}_R(F)$ or $\boldsymbol{\Sigma}_R(\boldsymbol{X})$ are defined as follows,

$$\boldsymbol{\Sigma}_R(F) = \boldsymbol{\Sigma}_R(\boldsymbol{X}) = E_F\{\boldsymbol{R}(\boldsymbol{X}, F)\boldsymbol{R}^T(\boldsymbol{X}, F)\}.$$

Since the rank is centered, the RCM is nothing but the covariance matrix of the rank of $\boldsymbol{X}$, which is $\mathrm{Cov}(\boldsymbol{R}(\boldsymbol{X}, F))$. Recall that $\|\boldsymbol{R}(\boldsymbol{X}, F)\| \leq 1$, hence the assumptions on F for existence of $\boldsymbol{\Sigma}_R$ are much weaker than the ones for existence of $\mathrm{Cov}(\boldsymbol{X})$.

From the last term in the (3.1), the computational complexity of $\boldsymbol{\Sigma}_R(F_n)$ seems to be $O(n^3)$. However, utilizing the middle term to compute $\boldsymbol{\Sigma}_R(F_n)$ needs only $O(n^2)$ computing time. It is worth noting that $\boldsymbol{\Sigma}_R(F_n)$ is asymptotically equivalent to a matrix-valued U-statistic with the kernel of size 3, hence the convergence of the sample version to the population one can be established by the practice of U-statistic theory.

As any spatial procedure, spatial signs and spatial ranks are orthogonally equivariant in the sense that for any $p \times p$ orthogonal matrix $\boldsymbol{O}$ ($\boldsymbol{O}^T = \boldsymbol{O}^{-1}$), p-dimensional vector $\boldsymbol{b}$ and nonzero scaler c, letting $\boldsymbol{x}^* = c\boldsymbol{O}\boldsymbol{x} + \boldsymbol{b}$ and $\boldsymbol{X}^* = c\boldsymbol{O}\boldsymbol{X} + \boldsymbol{b}$ with the distribution $F_{\mathbf{X}^*}$,

$$\boldsymbol{S}(\boldsymbol{x}^*) = \text{sign}(c)\boldsymbol{O}\boldsymbol{S}(\boldsymbol{x}), \quad \text{and} \quad \boldsymbol{R}(\boldsymbol{x}^*, F_{\mathbf{X}^*}) = \text{sign}(c)\boldsymbol{O}\boldsymbol{R}(\boldsymbol{x}, F_{\mathbf{X}}).$$

Therefore, $\boldsymbol{\Sigma}_R(F)$ is orthogonally equivariant, meaning that,

$$\boldsymbol{\Sigma}_R(\boldsymbol{X}^*) = c^2\boldsymbol{O}\boldsymbol{\Sigma}_R(\boldsymbol{X})\boldsymbol{O}^T.$$

Orthogonal equivariance ensures that under rotation, translation and homogeneous scale change, the quantities are transformed accordingly. It has the same property as the much more complicated Oja rank covariance matrix at this point. However, it does not allow heterogeneous scale changes. The above equations do not hold for a general $p \times p$ nonsingular matrix $\boldsymbol{A}$. Hence, they are not fully affine equivariant. In order to achieve fully affine equivariance, we strengthen the assumption by confining our focus to the family of all elliptically symmetric distributions, which is the most frequently used assumption in practice. See Section 2.2.

The key result from Marden (1999) provides the fundamental to modify spatial rank covariance matrix such that it is affine equivariant under elliptical models. In fact, we may remove the assumption from Marden that requires the existence of the covariance matrix. The result and proofs are still valid with only difference being the interpretation of eigenvalues.

Lemma 10 *(Marden, 1999) If $\boldsymbol{X}$ is elliptically distributed from F with the scatter parameter $\boldsymbol{\Sigma}$ having the spectral decomposition $V\Lambda V^T$, then $\boldsymbol{\Sigma}_R(F) = \boldsymbol{V}\boldsymbol{\Lambda}_R\boldsymbol{V}^T$, where $\boldsymbol{\Lambda}_R$ is the diagonal matrix of eigenvalues of $\boldsymbol{\Sigma}_R$.*

The lemma insures that the same orthogonal matrix V diagonalize $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_R$. In other words, spatial rank covariance matrix $\boldsymbol{\Sigma}_R$ has the same eigenvectors as $\boldsymbol{\Sigma}$. For any eigenvector $\boldsymbol{v}$ as a column of $\boldsymbol{V}$, the corresponding eigenvalue λ in diagonal position of $\boldsymbol{\Lambda}$ is the measure of variability of $\boldsymbol{v}^T\boldsymbol{X}$. If the scatter parameter $\boldsymbol{\Sigma}$ exists, the eigenvalue λ is proportional to the variance of $\boldsymbol{v}^T\boldsymbol{X}$. Therefore, instead of using eigenvector of sample covariance $\mathrm{Cov}(\boldsymbol{X})$ to estimate the eigenvector of $\boldsymbol{\Sigma}$, one can use the one of RCM $\boldsymbol{\Sigma}_R(\boldsymbol{X})$. The entries of diagonal matrix $\boldsymbol{\Lambda}$ can then be estimated by a scale estimation on the projection of $\boldsymbol{v}^T\boldsymbol{X}$.

3.4 Modified RCM

The way we construct the modified spatial rank covariance matrix MRCM (sample version), denoted as $\tilde{\boldsymbol{\Sigma}}(F_n)$ or $\tilde{\boldsymbol{\Sigma}}(\mathbb{X})$, is as follows.

1 Compute the sample spatial covariance matrix $\boldsymbol{\Sigma}_R(F_n)$ using (3.1).

2 Construct eigenvector estimates. Find the corresponding eigenvector estimates $\mathrm{s}_1, \mathrm{s}_2, ..., \mathrm{s}_p$ by the spectral decomposition of $\boldsymbol{\Sigma}_R(F_n)$, denoted by the matrix S. That is, $S = [\mathrm{s}_1, \mathrm{s}_2, ..., \mathrm{s}_p]$

3 Find scale estimates (eigenvalues, principal values) of $\mathbb{X}$ on directions of s_i's, by using an univariate robust scale estimator σ. Let $\hat{\lambda}_i = \{\sigma(\boldsymbol{s}_i^T\mathbb{X})\}^2$ and denote $\hat{\Lambda} = \mathrm{diag}(\hat{\lambda}_1, ..., \hat{\lambda}_p)$. Here, we take σ to be the median absolute deviation (MAD).

4 The scatter estimate is $\tilde{\boldsymbol{\Sigma}}(F_n) = S\hat{\Lambda}S^T$.

Remark 11 *MAD is a widely used robust measure of variability of a univariate sample of data. For a univariate data set $\mathbb{X} = \{x_1, ..., x_n\}$, the MAD is defined as the median of absolute deviations from the median of the sample. That is,*

$$\mathrm{MAD} = median(|x_i - median(\mathbb{X})|).$$

Its robust properties are illustrated in the Chapter 1, by the Example 1, Example 6 (d) and Example 9.

Let $F_{\boldsymbol{v}}$ be the distribution of $\boldsymbol{v}^T\boldsymbol{X}$, we may obtain the population version of MRCM by finding eigenvector $\boldsymbol{v}_i$ of $\boldsymbol{\Sigma}_R(F)$ and $\tilde{\lambda}_i = \sigma^2(F_{\boldsymbol{v}_i})$ for $i = 1, ..., p$, then $\tilde{\boldsymbol{\Sigma}}(F) = \boldsymbol{V}\tilde{\boldsymbol{\Lambda}}\boldsymbol{V}^T$, where $\boldsymbol{V} = [\boldsymbol{v}_1, ..., \boldsymbol{v}_p]$ and $\tilde{\boldsymbol{\Lambda}} = \mathrm{diag}(\tilde{\lambda}_1, .., \tilde{\lambda}_p)$.

By the way of constructing $\tilde{\mathbf{\Sigma}}$, an immediate consequence of Lemma 10 is that $\tilde{\mathbf{\Sigma}}(F)$ is proportional to the true scatter parameter $\mathbf{\Sigma}$, that is,

$$\tilde{\mathbf{\Sigma}}(F) = c(h_F, \sigma)\mathbf{\Sigma}, \tag{3.2}$$

where $c(h_F, \sigma)$ is a constant only depends on h function of the distribution F and the choice of univariate scale functional σ. For example, in here, by taking $\sigma = MAD$, if F is the multivariate Gaussian distribution, $c(h_F, \sigma) = \{\Phi^{-1}(3/4)\}^2 \approx 0.455$, where Φ^{-1} is the quantile function of the standardized Gaussian distribution. If F is a t-distribution with $\nu > 2$, $c(h_F, \sigma) \approx 0.455\nu/(\nu - 2)$. If F is the Cauchy distribution, $c(h_F, \sigma) = 1$. We would show in later, for the sake to obtain the highest possible breakdown point, we will use the MAD$_k$, a variation of MAD as a substitution of σ to estimate the scale and shape of $\mathbf{\Sigma}$. Its definition and further discussion are given in Section 4.2.

Theorem 12 *Under an elliptical distribution F with scatter parameter $\mathbf{\Sigma}$, $\tilde{\mathbf{\Sigma}}(F)$ (or $\tilde{\mathbf{\Sigma}}(\boldsymbol{X})$) is an affine equivariant scatter functional.*

Proof of theorem 12:

The proof of the affine equivariance of $\tilde{\mathbf{\Sigma}}$ is straightforward. Let $\boldsymbol{X}$ be a random vector elliptically distributed from F with h and scatter parameter $\mathbf{\Sigma}$, then $\boldsymbol{X}^* = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}$ has the elliptical distribution with the same h and scatter parameter $\boldsymbol{A}\mathbf{\Sigma}\boldsymbol{A}^T$. So by (3.2), we have

$$\tilde{\mathbf{\Sigma}}(\boldsymbol{X}^*) = c(h_F, \sigma)A\mathbf{\Sigma}A^T = c(h_F, \sigma)Ac(h_F, \sigma)^{-1}\tilde{\mathbf{\Sigma}}(\boldsymbol{X})A^T = A\tilde{\mathbf{\Sigma}}(\boldsymbol{X})A^T.$$

□

Therefore, the modified spatial rank covariance matrix is affine equivariant under elliptical models. For any distribution, it is orthogonally equivariant, the property inherited from the spatial rank covariance matrix.

Theorem 13 *For any p-variate random vector $\boldsymbol{X}$, $\tilde{\boldsymbol{\Sigma}}(F)$ (or $\tilde{\boldsymbol{\Sigma}}(\boldsymbol{X})$) is orthogonally equivariant.*

Proof of Theorem 13:

Let $\boldsymbol{X}^* = cO\boldsymbol{X} + \boldsymbol{b}$ for any orthogonal matrix $\boldsymbol{O}$, p-vector $\boldsymbol{b}$ and nonzero scalar c. Let the spectral decomposition of $\boldsymbol{\Sigma}_R(\boldsymbol{X})$ be $\boldsymbol{U}\boldsymbol{\Lambda}_R\boldsymbol{U}^T$. By the orthogonal equivariance of spatial rank covariance matrix, we have

$$\boldsymbol{\Sigma}_R(\boldsymbol{X}^*) = c^2\boldsymbol{O}\boldsymbol{\Sigma}_R(\boldsymbol{X})\boldsymbol{O}^T = c^2\boldsymbol{O}\boldsymbol{U}\boldsymbol{\Lambda}_R\boldsymbol{U}^T\boldsymbol{O}^T = c^2(\boldsymbol{O}\boldsymbol{U})\boldsymbol{\Lambda}_R(\boldsymbol{O}\boldsymbol{U})^T,$$

so the eigenvector matrix of $\boldsymbol{\Sigma}_R(\boldsymbol{X}^*)$ is OU, which is $[\boldsymbol{O}\boldsymbol{u}_1, ..., \boldsymbol{O}\boldsymbol{u}_p]$. Then for each scale estimate $\tilde{\lambda}_i(\boldsymbol{X}^*) = \sigma^2((\boldsymbol{O}\boldsymbol{u}_i)^T\boldsymbol{X}^*) = \sigma^2((\boldsymbol{O}\boldsymbol{u}_i)^T(cO\boldsymbol{X} + \boldsymbol{b})) = c^2\sigma^2(\boldsymbol{u}_i^T O^T O\boldsymbol{X}) = c^2\sigma^2(\boldsymbol{u}_i^T\boldsymbol{X}) = c^2\tilde{\lambda}_i(\boldsymbol{X})$ for $i = 1, ..., p$. Let $\tilde{\boldsymbol{\Lambda}}(\boldsymbol{X}) = \text{diag}(\tilde{\lambda}_1, ...\tilde{\lambda}_p)$, by the construction of $\tilde{\boldsymbol{\Sigma}}$, we have

$$\tilde{\boldsymbol{\Sigma}}(\boldsymbol{X}^*) = (\boldsymbol{O}\boldsymbol{U})c^2\tilde{\boldsymbol{\Lambda}}(\boldsymbol{X})(\boldsymbol{O}\boldsymbol{U})^T = c^2\boldsymbol{O}\tilde{\boldsymbol{\Sigma}}(\boldsymbol{X})\boldsymbol{O}^T.$$

□

3.5 More on Affine Equivariance

Under elliptical symmetric distributions, MRCM is affine equivariant and proportional to the scatter parameter. Indeed, besides the the class of elliptical distributions, the MRCM can be shown as affine equivariance in a broader class of distributions. For example, if $\boldsymbol{X}$ have an exchangeable and symmetric distribution. That is, $\boldsymbol{X}$ and $\boldsymbol{D}\boldsymbol{J}\boldsymbol{X}$ have the same distribution for any permutation matrix $\boldsymbol{J}$ and any diagonal matrix $\boldsymbol{D}$ with diagonal elements ± 1, MRCM $\tilde{\boldsymbol{\Sigma}}(F_{\boldsymbol{X}^*}) = \boldsymbol{A}\tilde{\boldsymbol{\Sigma}}(F_{\boldsymbol{X}})\boldsymbol{A}^T$, where $\boldsymbol{X}^* = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}$. Elliptical symmetric distributions belong to this class because the distribution of the corresponding $\boldsymbol{X}$ is independent and symmetric in each component.

Under the assumption of data distributed in this class, we may obtain the affine equivariant location estimator using the spatial median by the transformation and retransformation technique.

More specifically, one can find the spatial median of the transformed data $(\tilde{\mathbf{\Sigma}}^{-1/2}(\mathbb{X}^*)\boldsymbol{x}_1^*, \cdots, \tilde{\mathbf{\Sigma}}^{-1/2}(\mathbb{X}^*)\boldsymbol{x}_n^*)$, denoted as M_s, then retransform it back to the original coordinate system, i.e., $\tilde{\mathbf{\Sigma}}^{1/2}(\mathbb{X}^*)M_s$ gives an affine equivariant location estimator.

Tyler *et al.* (2009) presented a general method for exploring high dimensional data based on the spectrum decomposition of one scatter estimator matrix relative to the other. For a distribution which lies outside of the above mentioned class, different scatter statistics may estimate different quantities of underly distributions, hence their methods may reveal interesting features in data structure. Our MRCM, an easy-computed high breakdown point scatter matrix, certainly deserves further investigation in application to their method and other multivariate methods.

Chapter 4

Properties of MRCM

In the this chapter, we would conduct robustness analysis on the MRCM through two approaches: influence function and breakdown point. In addition, we would study the finite sample efficiency among different methods.

The MRCM is robust locally in terms of the influence function and highly robust globally in terms of the sample breakdown point. We derived the influence functions for the eigenvectors and eigenvalues of the MRCM, then the influence function for the MRCM. They are bounded under the assumption that the scatter parameter has distinct eigenvalues. A generalization to multiple eigenvalues is possible as in Tanaka (1988). The breakdown point attains the upper bound by the a choice of robust univariate scale functional to be $\sigma = MAD_k$ with some optimal values for k. Comparing with the other high breakdown point estimators such as the MCD, the S-estimator and the projection based estimator, our MRCM is easy to compute with the complexity $O(n^2 + p^3)$. Even for large data sets in high dimensions, using MRCM is still practical. Also, MRCM is highly statistical efficient under Gaussian distribution and heavy-tailed distributions.

4.1 Infinitesimal Robust on Influence Function

As shown in the Chapter 1, the influence function is a Gâteau derivative of functional. For any fixed point $\boldsymbol{x} \in \mathbb{R}^p$, let the *$\varepsilon$-contamination distribution* at F be $F_\varepsilon = (1-\varepsilon)F + \varepsilon\Delta_{\boldsymbol{x}}$. Then the influence function of a functional $T(\cdot)$ at the given distribution F is given by

$$IF(\boldsymbol{x}, T; F) = \lim_{\varepsilon \to 0^+} \frac{T(F_\varepsilon) - T(F)}{\varepsilon} = \left.\frac{\partial T(F_\varepsilon)}{\partial \varepsilon}\right|_{\varepsilon=0}.$$

The influence function measures the effect on T of infinitesimal point mass contamination of the distribution F. Clearly, it is desired to be small or at least bounded. A functional T with a bounded influence function is regarded as infinitesimal robust, see Section 1.2.2.

Lemma 14 *For any random vector $\boldsymbol{X}$ with distribution F in $\mathbb{R}^p$, the influence functions of the spatial rank and RCM are given by*

$$IF(\boldsymbol{x}, \boldsymbol{R}(\boldsymbol{X}, F); F) = \boldsymbol{S}(\boldsymbol{x} - \boldsymbol{X}) - \boldsymbol{R}(\boldsymbol{X}, F),$$

and

$$\begin{aligned} IF(\boldsymbol{x}, \boldsymbol{\Sigma}_R; F) &= E_F\boldsymbol{S}(\boldsymbol{X} - \boldsymbol{x})\boldsymbol{S}(\boldsymbol{X} - \boldsymbol{x})^T + \boldsymbol{R}(\boldsymbol{x}, F)\boldsymbol{R}(\boldsymbol{x}, F)^T \\ &\quad -2\boldsymbol{\Sigma}_R(F) - E_F IF(\boldsymbol{x}, \boldsymbol{R}(\boldsymbol{X}, F); F) IF(\boldsymbol{x}, \boldsymbol{R}(\boldsymbol{X}, F); F)^T. \end{aligned}$$

Proof of Lemma 14: We have

$$\begin{aligned} IF(\boldsymbol{x}, \boldsymbol{R}(\boldsymbol{X}, F); F) &= \frac{\partial}{\partial\varepsilon}\boldsymbol{R}\big(\boldsymbol{X}, (1-\varepsilon)F + \varepsilon\Delta_{\boldsymbol{x}}\big)\Big|_{\varepsilon=0} \\ &= \frac{\partial}{\partial\varepsilon}\left[(1-\varepsilon)\boldsymbol{R}(\boldsymbol{X}, F) + \varepsilon\frac{\boldsymbol{X} - \boldsymbol{x}}{\|\boldsymbol{X} - \boldsymbol{x}\|}\right]\Bigg|_{\varepsilon=0} \\ &= \boldsymbol{S}(\boldsymbol{X} - \boldsymbol{x}) - \boldsymbol{R}(\boldsymbol{X}, F). \end{aligned}$$

Because

$$\begin{aligned} IF(\boldsymbol{x}, \boldsymbol{\Sigma}_R; F) &= \frac{\partial}{\partial\varepsilon}E_{F_\varepsilon}\{\boldsymbol{R}(\boldsymbol{X}, F_\varepsilon)\boldsymbol{R}(\boldsymbol{X}, F_\varepsilon)^T\}\Big|_{\varepsilon=0} \\ &= \frac{\partial}{\partial\varepsilon}E_{F_\varepsilon}\{[(1-\varepsilon)\boldsymbol{R}(\boldsymbol{X}, F) + \varepsilon\boldsymbol{S}(\boldsymbol{X} - \boldsymbol{x})] \\ &\quad [(1-\varepsilon)\boldsymbol{R}(\boldsymbol{X}, F) + \varepsilon\boldsymbol{S}(\boldsymbol{X} - \boldsymbol{x})]^T\}\Big|_{\varepsilon=0} \\ &= \frac{\partial}{\partial\varepsilon}(1-\varepsilon)E_F\mathbf{M} + \varepsilon I[\boldsymbol{X} = \boldsymbol{x}]\mathbf{M}\Big|_{\varepsilon=0} \end{aligned}$$

where $I[A]$ is the indicator function being 1 when A is true or 0 otherwise, and

$$\mathrm{M} = [(1-\varepsilon)\boldsymbol{R}(\boldsymbol{X}, F) + \boldsymbol{S}(\boldsymbol{X}-\boldsymbol{x})][(1-\varepsilon)\boldsymbol{R}(\boldsymbol{X}, F) + \boldsymbol{S}(\boldsymbol{X}-\boldsymbol{x})]^T.$$

Simplify further, we have

$$\begin{aligned} IF(\boldsymbol{x}, \boldsymbol{\Sigma}_R; F) &= -3\boldsymbol{\Sigma}_R(F) + E_F\boldsymbol{S}(\boldsymbol{X}-\boldsymbol{x})\boldsymbol{R}(\boldsymbol{X}, F)^T + E_F\boldsymbol{R}(\boldsymbol{X}, F)\boldsymbol{S}(\boldsymbol{X}-\boldsymbol{x})^T \\ &\quad +\boldsymbol{R}(\boldsymbol{x}, F)\boldsymbol{R}(\boldsymbol{x}, F)^T \\ &= -3\boldsymbol{\Sigma}_R(F) - E_F[\boldsymbol{S}(\boldsymbol{X}-\boldsymbol{x}) - \boldsymbol{R}(\boldsymbol{X}, F)][\boldsymbol{S}(\boldsymbol{X}-\boldsymbol{x}) - \boldsymbol{R}(\boldsymbol{X}, F)]^T \\ &\quad +E_F\boldsymbol{S}(\boldsymbol{X}-\boldsymbol{x})\boldsymbol{S}(\boldsymbol{X}-\boldsymbol{x})^T + \boldsymbol{\Sigma}_R(F) + \boldsymbol{R}(\boldsymbol{x}, F)\boldsymbol{R}(\boldsymbol{x}, F)^T \\ &= E_F\boldsymbol{S}(\boldsymbol{X}-\boldsymbol{x})\boldsymbol{S}(\boldsymbol{X}-\boldsymbol{x})^T + \boldsymbol{R}(\boldsymbol{x}, F)\boldsymbol{R}(\boldsymbol{x}, F)^T - 2\boldsymbol{\Sigma}_R(F) \\ &\quad -E_F IF(\boldsymbol{x}, \boldsymbol{R}(\boldsymbol{X}, F); F)\mathrm{IF}(\boldsymbol{x}, \boldsymbol{R}(\boldsymbol{X}, F); F)^T. \end{aligned}$$

□

Remark 15 (i) The influence function of the rank function $\boldsymbol{R}(\boldsymbol{X}, F)$ is bounded with $\sup_{\boldsymbol{x}} \|\mathrm{IF}(\boldsymbol{x}, \boldsymbol{R}(\boldsymbol{X}, F); F)\| = \|\boldsymbol{R}(\boldsymbol{X}, F)\| + 1 < 2$ and the supremum is achieved at $\boldsymbol{x} = \boldsymbol{X} + c\boldsymbol{R}(\boldsymbol{X}, F)$, where c is any positive scalar.

(ii) The influence function for the RCM is bounded due to the boundedness of the spatial sign function, rank function and influence function of the rank function. In here, we say a matrix to be bounded if all of its elements are bounded.

(iii) The IF of the RCM for a spherically symmetrical distribution F can be obtained from Sirkiä *et al.* (2009), in which it was derived through the U-theory. However, the result can not be extended to an elliptical distribution by using Lemma 1 of Croux & Haesbroeck (2000), since RCM is not affine equivariant.

In order to give the influence function of MRCM $\tilde{\boldsymbol{\Sigma}}$, we need the following lemma that provided by Croux & Haesbroeck (2000).

Lemma 16 *(Croux & Haesbroeck, 2000) Let $S : \mathcal{F} \to SPD(p)$ be a statistical functional and F a p-dimensional distribution. Suppose that $IF(\boldsymbol{x}, S; F)$ exists. Denote $\boldsymbol{v}_1, ..., \boldsymbol{v}_p$ and $\lambda_1, ..., \lambda_p$ the eigenvectors and eigenvalues of $S(F)$. Then the influence functions of $\boldsymbol{v}_j$ and λ_j are given by*

$$IF(\boldsymbol{x}, \lambda_j; F) = \boldsymbol{v}_j^T IF(\boldsymbol{x}, S; F)\boldsymbol{v}_j,$$

$$IF(\boldsymbol{x}, \boldsymbol{v}_j; F) = \sum_{k \neq j}^{p} \frac{1}{\lambda_j - \lambda_k}\{\boldsymbol{v}_k^T \mathrm{IF}(\boldsymbol{x}, S; F)\boldsymbol{v}_j\}\boldsymbol{v}_k.$$

The modified rank covariance matrix is determined by the eigenvectors of RCM and robust scale estimator of the univariate projection on each eigenvector. For $\sigma = \mathrm{MAD}$, we conduct the perturbation analysis for eigenvalues and eigenvectors of $\tilde{\boldsymbol{\Sigma}}(F)$. Their influence functions are given by the following theorem.

Theorem 17 *Let $\tilde{\boldsymbol{\Sigma}}(F)$ be the modified spatial rank covariance functional on an elliptical distribution F with $\sigma = \mathrm{MAD}$. Suppose the spatial rank covariance matrix $\boldsymbol{\Sigma}_R(F)$ has distinct eigenvalues $\lambda_1 > ... > \lambda_p > 0$ and the corresponding eigenvectors $\boldsymbol{v}_1, ..., \boldsymbol{v}_p$. Denote $\tilde{\lambda}_j$ and $\tilde{\boldsymbol{v}}_j$ as the j^{th} eigenvalue and corresponding eigenvector of $\tilde{\boldsymbol{\Sigma}}(F)$, respectively. Then the influence functions of $\tilde{\boldsymbol{v}}_j$ and $\tilde{\lambda}_j$ ($j = 1, ..., p$) are given by*

$$IF(\boldsymbol{x}, \tilde{\boldsymbol{v}}_j; F) = \sum_{k \neq j}^{p} \frac{1}{\lambda_j - \lambda_k}\{\boldsymbol{v}_k^T IF(\boldsymbol{x}, \boldsymbol{\Sigma}_R; F)\boldsymbol{v}_j\}\boldsymbol{v}_k$$

and

$$IF(\boldsymbol{x}, \tilde{\lambda}_j; F) = \frac{1}{4h(1)}\mathbf{sgn}^T(|\boldsymbol{v}_j \circ \boldsymbol{x}| - \mathbf{1}_p)IF(\boldsymbol{x}, \tilde{\boldsymbol{v}}_j; F), \tag{4.1}$$

where $\boldsymbol{a} \circ \boldsymbol{b}$ is the component-wise product of $\mathbf{a}$ and $\mathbf{b}$, $|\boldsymbol{a}| = (|a_1|, ..., |a_p|)^T$ is the component-wise absolute value, $\mathbf{1_p}$ is the p-vector with all entries 1, and $\mathbf{sgn}(\boldsymbol{a})$ is the component-wise sign vector and equal to $(\mathrm{sgn}(a_1), ..., \mathrm{sgn}(a_p))^T$.

Proof of Theorem 17:

Since $\tilde{\boldsymbol{v}}_j = \boldsymbol{v}_j$ for $j = 1, ..., d$, the influence function of $\tilde{\boldsymbol{v}}_j$ is directly followed from Lemma 16 with $S = \boldsymbol{\Sigma}_R$.

Treating median and MAD as simultaneous M-estimators as in Page 135 of Huber & Ronchetti (2009), it is easy to prove that for any unit vector $\boldsymbol{u} \in R^d$ under an elliptical model

$$\mathrm{IF}(\boldsymbol{u}^T\boldsymbol{x}, \mathrm{MAD}; F_{\boldsymbol{u}}) = \frac{\mathrm{sgn}(|\boldsymbol{u}^T\boldsymbol{x}| - 1)}{4h(1)}. \tag{4.2}$$

Notice that the similar equality is shown in (1.9). Now $\tilde{\lambda}_j = \mathrm{MAD}(F_{\boldsymbol{v}_j})$, where $F_{\boldsymbol{v}}$ is the distribution of $\boldsymbol{v}^T\boldsymbol{X}$. By the chain rule of vector derivatives,

$$\frac{\partial \tilde{\lambda}_j(F_\varepsilon)}{\partial \varepsilon} = \frac{\partial \tilde{\lambda}_j(F_\varepsilon)}{\partial \boldsymbol{v}_j(F_\varepsilon)} \frac{\partial \boldsymbol{v}_j(F_\varepsilon)}{\partial \varepsilon}, \tag{4.3}$$

where $F_\varepsilon = (1-\varepsilon)F + \varepsilon\Delta\boldsymbol{x}$. The evaluation at $\varepsilon = 0$ of the second derivative on the right hand side of (4.3) is the column vector that is the influence function of $\tilde{\boldsymbol{v}}_j$. To be more specific, denote $\boldsymbol{v}_j = (v_{j1}, v_{j2}, ..., v_{jp})^T$. The evaluation of the first derivative at $\varepsilon = 0$ is the row vector whose i^{th} element is the influence function of MAD at $(0, ..., 0, v_{ji}, 0, ..., 0)^T\boldsymbol{x}$. By (4.2), the influence function of eigenvalue of $\tilde{\boldsymbol{\Sigma}}$ follows.

□

Remark 18 (i) The boundedness of IF of RCM implies the boundedness of influence functions for the eigenvectors of $\tilde{\boldsymbol{\Sigma}}(F)$.

(ii) With a robust choice of univariate scale estimator (MAD), the influence functions for eigenvalues of $\tilde{\boldsymbol{\Sigma}}(F)$ are also kept bounded.

(iii) The RCM provides an immediate application to robust principle component analysis for dimension reduction.

(iv) The result is obtained under the assumption of distinct eigenvalues. A generalization to multiple eigenvalues is possible as in Tanaka (1988).

(v) The assumption on elliptical symmetry is not necessary. However, for a general model, the representations of IF for MAD and eigenvectors may be in length.

Based on the influence functions of eigenvalues and eigenvectors, we are able to derive the influence function for our modified spatial rank covariance matrix, which is given in the following theorem.

Theorem 19 *For an elliptical distribution F, let the eigenvalues and eigenvectors of $\boldsymbol{\Sigma}_R(F)$ be $\lambda_1 > ... > \lambda_p > 0$ and $\boldsymbol{v}_1, ..., \boldsymbol{v}_p$ respectively. Then the influence function of $\tilde{\boldsymbol{\Sigma}}$ at F is given by*

$$IF(\boldsymbol{x}, \tilde{\boldsymbol{\Sigma}}; F) = IF(\boldsymbol{x}, \boldsymbol{\Sigma}_R; F) + \sum_{j=1}^{p} a_j(\boldsymbol{x}) \boldsymbol{v}_j \boldsymbol{v}_j^T,$$

where $a_j(\boldsymbol{x}) = \dfrac{1}{4h(1)} \sum_{k \neq j}^{p} \dfrac{1}{\lambda_j - \lambda_k} \boldsymbol{v}_k^T IF(\boldsymbol{x}, \boldsymbol{\Sigma}_R; F) \boldsymbol{v}_j \mathbf{sign}^T(|\boldsymbol{v}_j \circ \boldsymbol{x}| - \mathbf{1}_p) \boldsymbol{v}_k - \boldsymbol{v}_j^T IF(\boldsymbol{x}, \boldsymbol{\Sigma}_R; F) \boldsymbol{v}_j$

Proof of Theorem 19:

By comparing to the first order Taylor expansion in (1.6), the results of Theorem 17 imply that

$$\tilde{\lambda}_j(F_\varepsilon) = \tilde{\lambda}_j(F) + \varepsilon \mathrm{IF}(\boldsymbol{x}, \tilde{\lambda}_j; F) + O(\varepsilon^2),$$

and

$$\boldsymbol{v}_j(F_\varepsilon) = \boldsymbol{v}_j(F) + \varepsilon \mathrm{IF}(\boldsymbol{x}, \boldsymbol{v}_j; F) + O(\varepsilon^2).$$

Then

$$\begin{aligned}
\tilde{\boldsymbol{\Sigma}}(F_\varepsilon) &= \sum_{j=1}^{p} \tilde{\lambda}_j(F_\varepsilon) \boldsymbol{v}_j(F_\varepsilon) \boldsymbol{v}_j^T(F_\varepsilon) \\
&= \sum_{j=1}^{p} \{\tilde{\lambda}_j(F) \boldsymbol{v}_j(F) \boldsymbol{v}_j^T(F) + \varepsilon \mathrm{IF}(\boldsymbol{x}, \tilde{\lambda}_j; F) \boldsymbol{v}_j(F) \boldsymbol{v}_j^T(F) \\
&\quad + \varepsilon \tilde{\lambda}_j(F) \mathrm{IF}(\boldsymbol{x}, \boldsymbol{v}_j; F) \boldsymbol{v}_j^T(F) + \varepsilon \tilde{\lambda}_j(F) \boldsymbol{v}_j(F) \mathrm{IF}(\boldsymbol{x}, \boldsymbol{v}_j; F)^T\} + O(\varepsilon^2).
\end{aligned}$$

Hence,

$$\text{IF}(\boldsymbol{x}, \tilde{\boldsymbol{\Sigma}}; F) = \sum_{j=1}^{p} \{\text{IF}(\boldsymbol{x}, \tilde{\lambda}_j; F)\boldsymbol{v}_j(F)\boldsymbol{v}_j^T(F) + \tilde{\lambda}_j(F)[\text{IF}(\boldsymbol{x}, \boldsymbol{v}_j; F)\boldsymbol{v}_j^T(F) + \boldsymbol{v}_j(F)\text{IF}(\boldsymbol{x}, \boldsymbol{v}_j; F)^T]\}. \tag{4.4}$$

The summation of the last two terms is

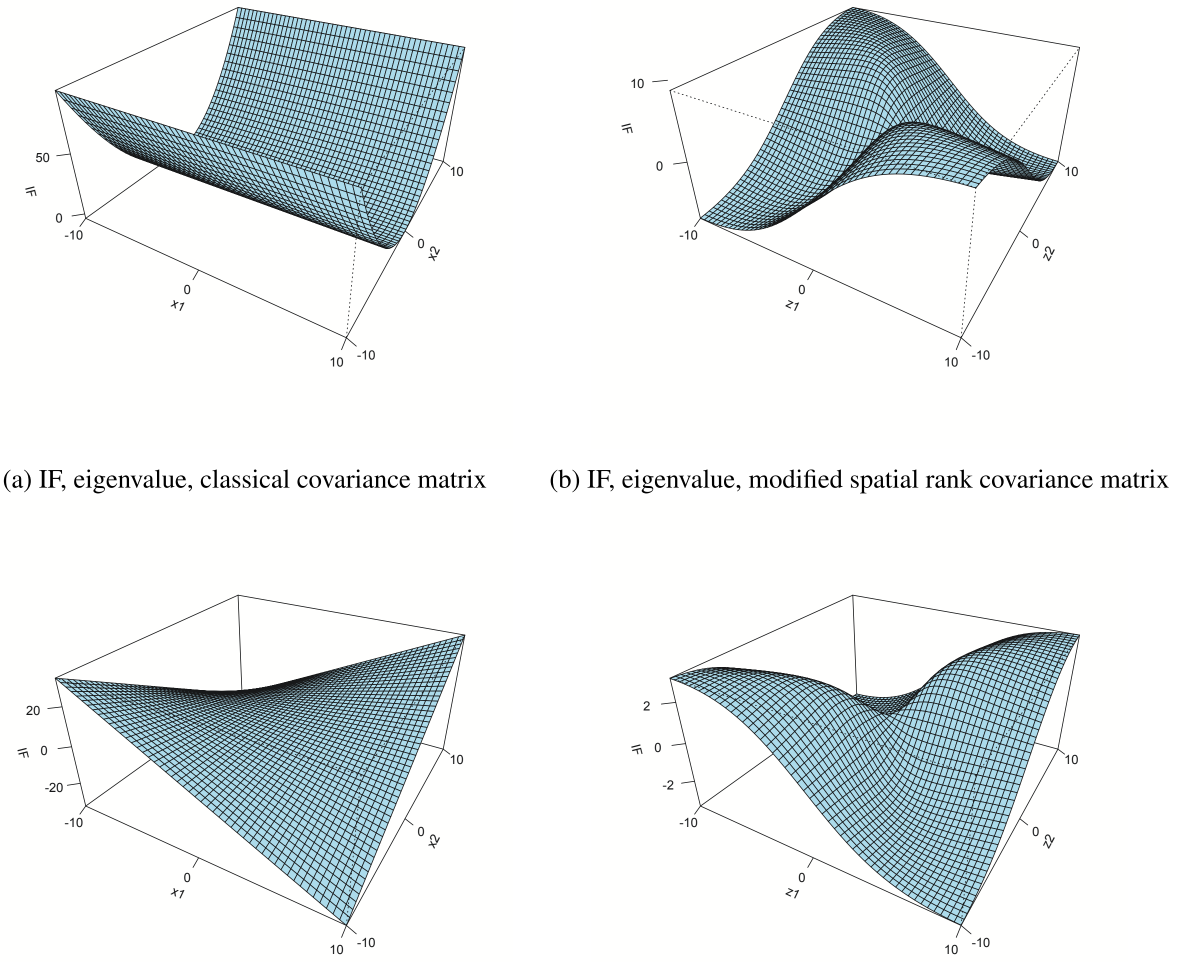
$$\sum_{j=1}^{p}\sum_{k\neq j}^{p} \boldsymbol{v}_j^T \text{IF}(\boldsymbol{x}, \boldsymbol{\Sigma}_R; F)\boldsymbol{v}_k\boldsymbol{v}_k\boldsymbol{v}_j^T = \text{IF}(\boldsymbol{x}, \boldsymbol{\Sigma}_R; F) - \sum_{j=1}^{p} \boldsymbol{v}_j^T \text{IF}(\boldsymbol{x}, \boldsymbol{\Sigma}_R; F)\boldsymbol{v}_j\boldsymbol{v}_j\boldsymbol{v}_j^T.$$

Plugging the influence functions of eigenvectors into the first term of (4.4) in the right side yields the stated expressions.

□

Remark 20 (i) $a_j(\boldsymbol{x})$ is bounded in $\boldsymbol{x}$, therefore the influence function of $\tilde{\Sigma}(F)$ is bounded.
(ii) Even though MRCM is affine equivariant under elliptical distributions, its influence function could not be written as the form in the result of Lemma 1 in Croux & Haesbroeck (2000). This is due to the construction way of MRCM. It is based on RCM, which is the covariance matrix of nonlinearly transformed ranks.

For demonstration and comparison, we compute and plot the influence functions of eigenvalues and eigenvectors for our MRCM, as well as for the classical covariance matrix. At $F = N(\mathbf{0}, \text{diag}(1, 4))$, both scatter functionals and RCM have the same eigenvectors $\boldsymbol{v}_1 = (0, 1)^T$ and $\boldsymbol{v}_2 = (1, 0)^T$. Since IF's of $\boldsymbol{v}_1$ for both functionals are of the form $b(\boldsymbol{x})\boldsymbol{v}_2$, hence the second components are always zero. We plot the first components of IF for $\boldsymbol{v}_1$ in Fig. 4.1 (c) and (p). The curve for our MRCM is saddle-shaped and bounded. Note that if we turn the curve upside down or rotate the curve 90 degree, we obtain the curve of the second component of IF for $\boldsymbol{v}_2$ since it equals the negative of the first component of IF of $\boldsymbol{v}_1$. Surprisingly, the influence function of the largest eigenvalue for MRCM is the first component of IF for $\boldsymbol{v}_1$ multiplying a factor $-1/4h(1) = -2.5898$. The curve of IF of $\tilde{\lambda}_1$ is plotted in Fig. 4.1 (b). The component-wise sign function in the formula (4.1) doesn't play any role in this case because the first component of $\boldsymbol{v}_1$ is 0 and the second component of IF for $\boldsymbol{v}_1$ is also 0. However, for a general distribution, we shall anticipate that the curve of IF for eigenvalue has more jumps, hence more local valleys and peaks than the curve of IF for eigenvector because of the component-wise sign function. As expected, the curves for our MRCM are kept bounded, while unbounded for the sample covariance matrix.



(a) IF, eigenvalue, classical covariance matrix

(b) IF, eigenvalue, modified spatial rank covariance matrix

(c) IF, eigenvector, classical covariance matrix

(d) IF, eigenvector, modified spatial rank covariance matrix

Table 4.1. Influence functions of the classical covariance matrix and modified spatial rank covariance matrix

Influence functions of (a) the largest eigenvalue for the classical covariance matrix, (b) the largest eigenvalue for the modified spatial rank covariance matrix, (c) the first component of the eigenvector corresponding to the largest eigenvalue for the classical covariance matrix and (d) the first component of the eigenvector corresponding to the largest eigenvalue for the modified spatial rank covariance matrix at $F = N(\mathbf{0}, \mathrm{diag}(1, 4))$.

4.2 Quantitative Robustness on Breakdown Point

The influence function measures the infinitesimal robustness of a functional $T(F)$, while the breakdown point captures the quantitative robustness of estimator $T(F_n)$, see Section 1.2.3. For scale estimators, it can break down in two ways, either become arbitrarily large (explosion) or become arbitrarily close to zero (implosion). In terms of scatter estimators, they can break down if one of its eigenvalues approaches ∞ or 0.

So the breakdown point of an scatter estimator T on a random sample $\mathbb{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$ is defined as

$$\varepsilon(T, \mathbb{X}) = \min\{\frac{m}{n} : \sup_{\mathbb{X}_m} |\text{det}\{T(\mathbb{X})T(\mathbb{X}_m)^{-1}\} + \text{det}\{T(\mathbb{X})^{-1}T(\mathbb{X}_m)\}| = \infty\},$$

where $\mathbb{X}_m$ denotes a contaminated sample resulting from replacing m points of $\mathbb{X}$ with arbitrary values.

Before we derive the finite sample breakdown point for the MRCM, a variation of MAD, MAD$_k$ is better introduced in here. Only in this section, the univariate scale estimator σ, which is used to obtain MRCM, is taken to be the MAD$_k$. This can lead to a slightly higher breakdown point and an elegant theoretical result. Similar ideas were adopted by several authors in the literature, for example Tyler (1994), Gather & Hilker (1997) and Zhou & Dang (2010).

MAD$_k$ can be roughly considered as the k/nth percentile of $|x_i - median|$. Specifically, Let $\mathbb{X} = \{x_1, ..., x_n\}$ be a random sample of n points in $\mathbb{R}$ with ordered values $x_{(1)} \leq \cdots \leq x_{(n)}$.

$$\text{MAD}_k(\mathbb{X}) = median_k(|x_1 - median(\mathbb{X})|, ..., |x_n - median(\mathbb{X})|),$$

where $median_k(\mathbb{X}) = (x_{(\lfloor (n+k)/2 \rfloor)} + x_{(\lfloor (n+k+1)/2 \rfloor)})/2, \quad 1 \leq k \leq n$, and $\lfloor a \rfloor$ is the greatest integer smaller or equal to a. The regular median and MAD correspond to median$_k$ and MAD$_k$ with $k = 1$, respectively. A simple example of MAD$_k$ is shown as follows.

Example 21 *Given the sorted data* $\mathbb{X} = \{8, 8, 12, 15, 17, 19, 21\}$, $\text{median}(\mathbb{X}) = 15$*, the sorted absolute deviations are*

$$\begin{array}{ccccccc} \{\,0 & 2 & 3 & 4 & 6 & 7 & 7\,\} \\ & & & \uparrow & \uparrow & \uparrow & \\ & & & \text{MAD}_1 & \text{MAD}_3 & \text{MAD}_5 & \end{array}$$

Theorem 22 *For any $p-$variate random sample $\mathbb{X} = \{\boldsymbol{z}_1, ..., \boldsymbol{z}_n\}$, let $\sigma = MAD_k$ and $c_1(\mathbb{X})$ be the maximum number of points of $\mathbb{X}$ in any $(p-1)$-dimensional hyperplane. If $n > 2c_1(\mathbb{X}) - k + 1$, then*

$$\varepsilon(\tilde{\boldsymbol{\Sigma}}, \mathbb{X}) = \begin{cases} \lfloor (n - 2c_1(\mathbb{X}) + k + 1)/2 \rfloor / n & \textit{if } 1 \leq k \leq c_1(\mathbb{X}) \\ \lfloor (n - k + 2)/2 \rfloor / n & \textit{if } c_1(\mathbb{X}) + 1 \leq k \leq n. \end{cases}$$

Proof of Theorem 22:

Let $\varepsilon^*(\sigma, \mathbb{X})$ represent the uniform finite sample breakdown point of σ at $\mathbb{X}$ as defined by Tyler (1994) when all univariate projections of the data are considered. That is,

$$\varepsilon^*(\sigma, \mathbb{X}) = \min_m \{\frac{m}{n} : \sup_{\|\boldsymbol{u}\|=1} \sup_{\mathbb{X}_m} \{\sigma(\boldsymbol{u}^T\mathbb{X})\sigma(\boldsymbol{u}^T\mathbb{X}_m)^{-1} + \sigma(\boldsymbol{u}^T\mathbb{X})^{-1}\sigma(\boldsymbol{u}^T\mathbb{X}_m)\} = \infty\}.$$

Let $\varepsilon(\sigma, \boldsymbol{u}^T\mathbb{X})$ represent the finite sample breakdown point for σ for the projected data in direction $\boldsymbol{u}$.

The estimator $\tilde{\boldsymbol{\Sigma}}(\mathbb{X})$ breaks down only if $\sigma = MAD_k$ breaks down for some direction $\boldsymbol{u}$. Since MAD_k can be exploded ($\to \infty$) or imploded ($\to 0$), the breakdown point of $\tilde{\boldsymbol{\Sigma}}(\mathbb{X})$ is determined by two quantities corresponding to the explosion and implosion of MAD_k, respectively.

According to Lemma 1 in Gather & Hilker (1997), for $k \in [1, c_2(\boldsymbol{u}^T\mathbb{X})]$, MAD_k in direction $\boldsymbol{u}$ will implode with the breakdown point being $\lfloor (n - 2c_2(\boldsymbol{u}^T\mathbb{X}) + k + 1)/2 \rfloor / n$, where $c_2(\boldsymbol{u}^T\mathbb{X})$ represents the maximum number of data points on the hyperplane orthogonal to the direction $\boldsymbol{u}$. If $k \in [c_2(\boldsymbol{u}^T\mathbb{X}) + 1, n]$, the finite sample explosive breakdown points for MAD_k in direction $\boldsymbol{u}$ is $\lfloor (n - k + 2/2 \rfloor / n$. Tyler (1994) states that $\varepsilon^*(\sigma, \mathbb{X}) \leq \inf_{\boldsymbol{u}} \varepsilon(\sigma, \boldsymbol{u}^T\mathbb{X})$ and equality holds if $\sigma(\boldsymbol{u}^T\mathbb{X})$

is a continuous function of $\boldsymbol{u}$. This is the case if σ is MAD$_k$. Also note that $c_2(\boldsymbol{u}^T\mathbb{X}) \leq c_1(\mathbb{X})$ for any $\boldsymbol{u}$ with equality holding for some $\boldsymbol{u}$. So we have

$$\varepsilon^*(\text{MAD}_k, \mathbb{X}) = \begin{cases} \lfloor (n - 2c_1(\mathbb{X}) + k + 1)/2 \rfloor / n & \text{if } 1 \leq k \leq c_1(\mathbb{X}) \\ \lfloor (n - k + 2)/2 \rfloor / n & \text{if } c_1(\mathbb{X}) + 1 \leq k \leq n. \end{cases}$$

The proof will be finished if we show that $\varepsilon(\tilde{\boldsymbol{\Sigma}}, \mathbb{X}) = \varepsilon^*(\text{MAD}_k, \mathbb{X})$. First, $\varepsilon(\tilde{\boldsymbol{\Sigma}}, \mathbb{X}) \leq \varepsilon^*(\text{MAD}_k, \mathbb{X})$ due to orthogonal equivariance of $\tilde{\boldsymbol{\Sigma}}$. This follows by noting that if $m = \lfloor (n - 2c_1(\mathbb{X}) + k + 1)/2 \rfloor$ and the replacements all lie in the same plane as $c_1(\mathbb{X})$ data points reside, then MAD$_k$ equals to 0 for the univariate projection orthogonal to that plane. By the orthogonal equivariance of $\tilde{\boldsymbol{\Sigma}}$, there exists orthogonal matrix $\boldsymbol{O}$ such that $\tilde{\boldsymbol{\Sigma}}(\boldsymbol{O}\mathbb{X}_m)$ has an eigenvalue 0. Hence $0 = \det\{\tilde{\boldsymbol{\Sigma}}(\boldsymbol{O}\mathbb{X}_m)\} = \det\{\boldsymbol{O}\tilde{\boldsymbol{\Sigma}}(\mathbb{X}_m)\boldsymbol{O}^T\} = \det(\boldsymbol{O})\det\{\tilde{\boldsymbol{\Sigma}}(\mathbb{X}_m)\}\det(\boldsymbol{O}^T) = \det\{\tilde{\boldsymbol{\Sigma}}(\mathbb{X}_m)\}$ and $\tilde{\boldsymbol{\Sigma}}$ implodes. Similarly, when MAD$_k$ explodes, our estimator explodes.

One the other hand, we also have $\varepsilon(\tilde{\boldsymbol{\Sigma}}, \mathbb{X}) \geq \varepsilon^*(\text{MAD}_k, \mathbb{X})$. Suppose $\varepsilon_m = \varepsilon(\tilde{\boldsymbol{\Sigma}}, \mathbb{X}) < \varepsilon^*(\text{MAD}_k, \mathbb{X})$. Then for all ε_m-corrupted data sets $\mathbb{X}_m$ and for all unit directions $\boldsymbol{u}$, there exist σ_0 and σ_1 such that $0 < \sigma_0 < MAD_k(\boldsymbol{u}^T\mathbb{X}) < \sigma_1 < \infty$. This implies that for all $\mathbb{X}_m$, $0 < \sigma_0 < \lambda_j(\tilde{\boldsymbol{\Sigma}}(\mathbb{X}_m)) < \sigma_1 < \infty$ for all $j = 1, ..., p$, where λ_j is an eigenvalue of $\tilde{\boldsymbol{\Sigma}}(\mathbb{X}_m)$. Hence $\tilde{\boldsymbol{\Sigma}}$ does not break down at ε_m, contradicting to the definition of $\varepsilon(\tilde{\boldsymbol{\Sigma}}, \mathbb{X})$.

□

Remark 23 (i) The breakdown point of $\tilde{\Sigma}$ depends only on the sample size n and $c_1(\mathbb{X})$, but is independent of other configurations of $\mathbb{X}$. For the breakdown point as a quantitative robustness measurement of estimators, this 'sample-free' property is definitely desirable.

(ii) The optimal choices of k are $c_1(\mathbb{X})$ or $c_1(\mathbb{X}) + 1$ so that $\varepsilon(\tilde{\Sigma}, \mathbb{X}) = \lfloor (n - c_1(\mathbb{X}) + 1)/2 \rfloor / n$ attains the maximum value. This is the upper bound of breakdown point for any affine equivariant scatter estimators, see Tyler (1994). Although MRCM is only orthogonally invariant in general, it is affine equivariant for elliptical models. Clearly, this attainability is preferred.

(iii) If $\mathbb{X}$ is in general position, that is, $c_1(\mathbb{X}) = d$, then the breakdown point of $\tilde{\Sigma}$ equals $\lfloor (n - d +$

$1)/2\rfloor/n$ when $k = d$, reaching the upper bound given by Davies (1987).

(iv) Theorem 22 focuses on the case $\sigma = \mathrm{MAD}_k$. The result, however, can be extended to any scale estimator with the same breakdown point as MAD_k.

(v) On the discussion of Davies & Gather (2005), Tyler mentioned an example to construct a high breakdown point covariance matrix estimator based on the sample covariance matrix by replacing the eigenvalues of the sample covariance matrix with robust variances for the sample principle component variables. Then the resulting covariance matrix has a high breakdown point. His intention was to call a reasonable concept of breakdown on principle component vectors that are some estimators in a compact set. It is out of scope of this paper to do so. However, it is worthwhile to note that MRCM has different robust properties comparing to that example. It has bounded influence functions for eigenvectors. Such property is closely related to the bounded breakdown function concept proposed by He & Simpson (1992).

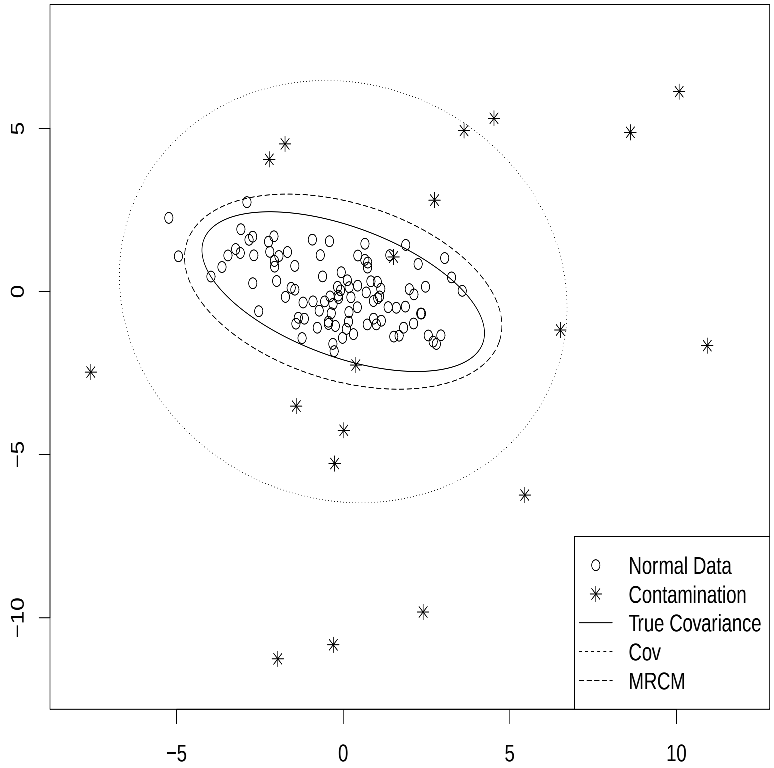


Figure 4.1. Comparison between sample covariance estimator and MRCM on on contaminated Gaussian distribution

This is the comparison of the sample covariance estimator and MRCM. Normal data are generated from $N\Big((0,0)^T, \big((3,-.9)^T;(-.9,1)^T\big)\Big)$ of size 100. 20% of contamination points uniformly distributed on $[-10,15] \times [-15,10]$ are added. Sample covariance estimate is severely distorted by the contamination. However, the MRCM returns a reasonable estimate of true covariance matrix

4.3 Finite Sample Efficiency

In this section, we consider the statistical efficiency of MRCM. Although the influence functions derived in Section 4.1 are useful in the calculation of asymptotic variance by the theorem 7, it is extremely challenge to obtain explicit expression of spatial rank covariance matrix under non-standard Gaussian distributions, even numerical integration becomes difficult. We insist to use Monte-Carlo simulation for efficiency comparison with other scatter estimators.

The following estimators are considered:

RCM The function *spatial.rank* in the R package ICSNP is used for computing the spatial rank vector. It is only orthogonally equivalent, hence RCM is only suitable for spherically symmetric distributions.

Mest Tyler's M-estimator is obtained by the function *tyler.shape* in the R package ICSNP. For a consideration of computational simplicity and robustness, the location vector is specified as the spatial median computed by the function *spatial.median*.

Mcd Minimum covariance determinant estimator is computed by the R package rrcov. The MCD method looks for the h observations (out of n) whose classical covariance matrix has the lowest possible determinant. Then MCD scatter estimator is the covariance matrix based on those h observations. h value is set to be default value $n/2$ of the function *CovMcd* which leads the breakdown point close to $1/2$.

Sest Re-weighted S-estimator (Sest) is calculated by the R package riv using Tukey bi-weighted ρ function with $c = 2.661$ for $d = 2$ and $c = 4.652$ for $d = 5$. Such c values provided as output of the function *slc* yield the breakdown point close to $1/2$.

Cov Non-robust sample covariance matrix.

As performance criteria for matrices, two quantities are used. One is the condition number of $\boldsymbol{\Sigma}^{-1}\boldsymbol{V}$, where $\boldsymbol{\Sigma}$ is the true scatter matrix and $\boldsymbol{V}$ is one of the above mentioned estimators. It is the ratio of the largest eigenvalue to the smallest eigenvalue of $\boldsymbol{\Sigma}^{-1}\boldsymbol{V}$. A good estimator $\boldsymbol{V}$ estimates $\boldsymbol{\Sigma}$ well such that $\boldsymbol{\Sigma}^{-1}\boldsymbol{V}$ is close to the identity matrix, hence the mean of log condition number (MLCN) of $\boldsymbol{\Sigma}^{-1}\boldsymbol{V}$ is expected to be close to 0. Such a criterion was also utilized in Maronna & Yohai (1995) and Gervini (2003). The other one is called the ADD, the angle difference in the direction of the first eigenvectors to measure the accuracy on estimating the first principle component, that is, $\cos^{-1}(|\boldsymbol{v}_1^T\hat{\boldsymbol{v}}_1|)$, where $\boldsymbol{v}_1$ is the first eigenvector of the theoretical scatter matrix $\boldsymbol{\Sigma}$ and $\hat{\boldsymbol{v}}_1$ is the first eigenvector of the scatter estimator $\boldsymbol{V}$.

We generate $M = 200$ samples with two sample sizes $n = 50$ and $n = 200$ from each of the following three scenarios:

Case I Standard Gaussian distributions with contaminations on shifted locations. i.e., $(1-\varepsilon)\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{p\times p})+\varepsilon\mathcal{N}(\mathbf{10}_p, \boldsymbol{\Sigma}_{p\times p})$, where $\mathbf{10}_p$ is the p-vector with all elements 10, $\boldsymbol{\Sigma}_{p\times p} = \mathrm{diag}(4, \mathbf{1}_{p-1}^T)$, ε to be 0, 0.1, 0.2 and $p = 2, 5$.

Case II Heavy-tailed $t_\nu(\mathbf{0}, \boldsymbol{\Sigma}_{p\times p})$ distributions for different degrees of freedom $\nu = 1, 3, 5$ and ∞ with dimension $p = 2$ and $p = 5$. Note that $\nu = \infty$ corresponds to Case I with $\varepsilon = 0$, the standard Gaussian distributions.

Case III Normal mixtures with contaminations on rotation. i.e. $(1-\varepsilon)\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{p\times p})+\varepsilon\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}^*_{p\times p})$, where $\boldsymbol{\Sigma}_{p\times p} = \mathrm{diag}(4, \mathbf{1}_{p-1}^T)$ and $\boldsymbol{\Sigma}^*_{p\times p} = 10 \times \mathrm{diag}(\mathbf{1}_{p-1}^T, 4)$. We take $\varepsilon = 0, 0.05, 0.1$.

		$\varepsilon = 0$		$\varepsilon = 0.1$		$\varepsilon = 0.2$	
		$p = 2$	$p = 5$	$p = 2$	$p = 5$	$p = 2$	$p = 5$
$n = 50$	MRCM(RE)	0.69	0.83	1.06	1.27	1.17	1.38
	RCM(RE)	0.37	0.91	0.55	1.44	0.62	1.50
	Mest(RE)	0.70	0.84	1.06	1.33	1.24	1.40
	Mcd(RE)	0.55	0.68	0.87	1.09	1.02	1.15
	Sest(RE)	0.56	0.86	0.89	1.35	1.04	1.39
	Cov(MLCN)	0.25	0.67	0.38	1.04	0.43	1.10
$n = 200$	MRCM(RE)	0.73	0.82	1.23	1.34	1.16	1.47
	RCM(RE)	0.19	0.57	0.30	0.91	0.30	0.97
	Mest(RE)	0.74	0.85	1.14	1.37	1.13	1.47
	Mcd(RE)	0.62	0.84	1.04	1.32	0.97	1.43
	Sest(RE)	0.62	0.88	1.02	1.41	1.01	1.50
	Cov(MLCN)	0.13	0.33	0.21	0.52	0.21	0.57

Table 4.2. Mean of log condition numbers for contaminated Gaussian distribution

Mean of log condition numbers (MLCN) of the sample covariance matrix (Cov) and relative efficiencies (RE) of other estimators relative to Cov under $F = (1-\varepsilon)\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{p\times p})+\varepsilon\mathcal{N}(10\mathbf{1}_p, \mathbf{\Sigma}_{p\times p})$, where $\mathbf{\Sigma}_{p\times p} = \text{diag}(4, \mathbf{1}_{p-1}^T)$.

		$\nu = 1$		$\nu = 3$		$\nu = 5$		$\nu = \infty$	
		$p = 2$	$p = 5$	$p = 2$	$p = 5$	$p = 2$	$p = 5$	$p = 2$	$p = 5$
$n = 50$	MRCM(RE)	4.23	3.93	1.36	1.36	1.00	1.12	0.69	0.83
	RCM(RE)	3.56	4.63	0.99	1.72	0.72	1.40	0.37	0.91
	Mest(RE)	4.78	4.12	1.37	1.40	1.01	1.08	0.70	0.84
	Mcd(RE)	3.26	2.50	0.99	0.95	0.73	0.77	0.55	0.68
	Sest(RE)	3.49	3.05	1.08	1.23	0.79	1.00	0.56	0.86
	Cov(MLCN)	2.49	4.77	0.70	1.60	0.51	1.29	0.25	0.67
$n = 200$	MRCM(RE)	8.77	7.85	1.63	1.83	1.10	1.16	0.73	0.82
	RCM(RE)	3.56	6.77	0.62	1.58	0.40	1.01	0.19	0.57
	Mest(RE)	10.12	8.32	1.66	1.95	1.10	1.17	0.74	0.85
	Mcd(RE)	7.19	5.25	1.26	1.49	0.86	0.93	0.62	0.84
	Sest(RE)	7.69	6.41	1.41	1.76	0.94	1.11	0.62	0.88
	Cov(MLCN)	2.41	4.57	0.42	1.05	0.28	0.64	0.13	0.33

Table 4.3. Mean of log condition numbers for $t_\nu-$distributions

Mean of log condition numbers (MLCN) for Cov and relative efficiencies (RE) for other estimators relative to Cov under t_ν-distributions with $\mu = \mathbf{0}, \mathbf{\Sigma} = \text{diag}(4, \mathbf{1}_{p-1}^T)$.

		$\varepsilon = 0$		$\varepsilon = 0.05$		$\varepsilon = 0.1$	
		$p = 2$	$p = 5$	$p = 2$	$p = 5$	$p = 2$	$p = 5$
$n = 100$	MRCM(RE)	0.93	0.96	4.66	3.48	7.64	5.32
	Mest(RE)	0.65	0.82	3.96	3.26	7.98	5.36
	Mcd(RE)	0.56	0.72	3.71	3.09	8.39	5.14
	Sest(RE)	0.55	0.87	3.77	3.56	8.06	6.04
	Cov($\cos^{-1}(\boldsymbol{v}_1^T \hat{\boldsymbol{v}}_1)$)	2.9°	7.1°	21.1°	28.8°	42.9°	51.3°
$n = 400$	MRCM(RE)	0.97	0.95	5.59	3.90	12.19	7.70
	Mest(RE)	0.76	0.83	5.36	3.66	11.91	8.05
	Mcd(RE)	0.62	0.77	5.22	3.68	13.14	8.71
	Sest(RE)	0.58	0.86	5.29	3.94	13.04	9.25
	Cov($\cos^{-1}(\boldsymbol{v}_1^T \hat{\boldsymbol{v}}_1)$)	2.1°	5.2°	18.8°	23.6°	47.5°	54.0°

Table 4.4. Mean of angle differences in the direction for contaminated Gaussian distribution

Mean of angle differences in the direction (MADD) for Cov and relative efficiencies for other estimators relative to Cov under $F = (1-\varepsilon)\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{p\times p}) + \varepsilon\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}^*_{p\times p})$, where $\mathbf{\Sigma}_{p\times p} = \mathrm{diag}(4, \mathbf{1}^T_{p-1})$, and $\mathbf{\Sigma}^*_{p\times p} = 10 \times \mathrm{diag}(\mathbf{1}^T_{p-1}, 4)$.

In the first two cases, we check the efficiency and robustness on the eigenvalues of estimators, hence use the mean of log(cond) (MLCN) as criterion to measure non-sphericity of $\boldsymbol{\Sigma}^{-1}\boldsymbol{V}$. The finite sample relative efficiency (RE) of estimator $\boldsymbol{V}$ is obtained by the ratio of MLCN of sample covariance to that of $\boldsymbol{V}$. Reported in Tables 4.2 and 4.3 are MLCN for Cov and RE for other estimators. In the third case, contaminations have totally different orientation and we are interested in estimating the first principle component. Hence the mean of angle difference in the direction between the first eigenvector of the true scatter matrix $\boldsymbol{\Sigma}$ and that of $\boldsymbol{V}$ (MADD) are computed, and the relative efficiency of $\boldsymbol{V}$ is the ratio of MADD of Cov to that of $\boldsymbol{V}$. Results of MADD of Cov and RE's of other estimators are listed in Table 4.4. RCM is skipped since it yields the exactly same results as the MRCM.

For the non-standard Gaussian distribution, RCM performs poorly due to its failure on estimating eigenvalues of the true covariance matrix. The MRCM seems to be better than other estimators, especially in estimating orientation of data clouds with the efficiencies at least 0.93. Efficiency on shape (eigenvalues) of MRCM is lower than it on the orientation, while the other estimators have almost the same RE on the shape as it on the orientation. This phenomenon can be explained by the separated steps in the construction of MRCM and relatively low efficiency on MAD$_k$. In the contaminated Gaussian models, the MRCM has a comparable or better performance comparing with other robust estimators. MRCM and Tyler M-estimator have very similar behaviors since both of them are based on spatial procedures with some treatments for affine equivariance property. Without a surprise, Tyler M-estimator is superior to MRCM and others under the heavy-tailed t-distribution of the degree of freedom $\nu = 1$, since it is the limiting form of MLE for scatter as $\nu \to 0$. In summary, MRCM has a competitive performance on efficiency as well as robustness.

Chapter 5

Mixture Model and EM Algorithm

In Section 2.2, we reviewed the unimodal multivariate elliptical models. The MRCM is shown to be a robust estimator of the scatter parameter $\mathbf{\Sigma}$ with many nice properties. However, if data have a more complex structure and come from the mixture of distributions (with "multi-mode"), all the methods we described earlier may not be sufficient. In order to give a robust method in estimating the parameters of mixture models, it is highly desired to extend the notion of MRCM to a broader class of distributions. Starting form this chapter, we would like to develop a robust method by modifying the traditional EM algorithm on the assumption of mixture of elliptical model. Actually, multivariate mixture of elliptical distribution has been widely used for various purposes in data mining and statistical learning. As the size of data getting larger and larger in terms of sample size and dimension, outlying observations are naturally inhabited in the data set. People in robust statistical society and computer science start to pay more and more attention to develop numerous methods in solving such problems. The main goal of robust methods in data mining and statistical learning is to capture the most important information from those noisy data. We would give several examples as part of application to show how we use the proposed method in performing outlier detection, clustering and classification at the end of this dissertation.

In this chapter, we would review the EM algorithm in general, discuss its contributions in estimating parameters of mixture models. In addition, we would point out the drawbacks of the tradition EM algorithm in different aspects and introduce some way to improve it based on the recent research done in this field.

5.1 Review of Regular EM Algorithm

Dempster *et al.* (1977) first advocated a unified algorithm, called the *Expectation-Maximization (EM)* Algorithm for deriving the maximum likelihood estimates from "incomplete" data. Rubin (1991) regarded the EM algorithm as one of the methodologies for solving incomplete data problems sequentially based on a complete data framework. During the past 3 decades, EM algorithm has been applied to almost all fields where statistical analysis is required, including medical science, engineering, sociology and business intelligence.

Suppose we consider a random sample of variable $\boldsymbol{Z} \in \mathbb{R}^p$ with density function $f(\boldsymbol{Z}|\boldsymbol{\theta})$. For each observation of $\boldsymbol{Z}$, we have observed value, denoted by variable $\boldsymbol{X}$. Besides that, we have missing value denoted by $\boldsymbol{Y}$. Thus, $\boldsymbol{Z} = (\boldsymbol{X}, \boldsymbol{Y})$. $\boldsymbol{X}$ and $\boldsymbol{Y}$ are assumed to be mutually independent. One can find out the *maximum likelihood estimate* (MLE) of $\boldsymbol{\theta}$ base on the *observed likelihood* function $L(\boldsymbol{\theta}; \mathbb{X}) = \prod_{i=1}^n f(\boldsymbol{x}_i|\boldsymbol{\theta})$. The EM algorithm was first proposed to maximize the observed likelihood $L(\boldsymbol{\theta}; \boldsymbol{X})$ by using the complete likelihood $L(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{Y})$.

The EM algorithm iteratively updates the estimate of $\boldsymbol{\theta}$. The following basic identity gives the way to shift the maximization problem from observed likelihood to complete likelihood. Assuming $\hat{\boldsymbol{\theta}}^{(t)}$, t-stage (iteration) estimate of $\boldsymbol{\theta}$, is in the same parameter space of $\boldsymbol{\theta}$, then

$$\begin{aligned}
\log L(\boldsymbol{\theta}; \boldsymbol{X}) &= \int \log f(\boldsymbol{X}|\boldsymbol{\theta}) f(\boldsymbol{Y}|\hat{\boldsymbol{\theta}}^{(t)}, \boldsymbol{X}) d\boldsymbol{Y} \qquad (5.1) \\
&= \int \log \frac{f(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{\theta})}{f(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{X})} f(\boldsymbol{Y}|\hat{\boldsymbol{\theta}}^{(t)}, \boldsymbol{X}) d\boldsymbol{Y} \\
&= \int \log f(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{\theta}) f(\boldsymbol{Y}|\hat{\boldsymbol{\theta}}^{(t)}, \boldsymbol{X}) d\boldsymbol{Y} - \int \log f(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\theta}) f(\boldsymbol{Y}|\hat{\boldsymbol{\theta}}^{(t)}, \boldsymbol{X}) d\boldsymbol{Y} \\
&= \mathrm{E}[\log L(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{Y})|\hat{\boldsymbol{\theta}}^{(t)}, \boldsymbol{X}] - \mathrm{E}[\log f(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{X})|\hat{\boldsymbol{\theta}}^{(t)}, \boldsymbol{X}] \qquad (5.2)
\end{aligned}$$

Iteratively applying the two following steps:

- E-step: Calculate the first term of (5.2), the conditional expectation of $\boldsymbol{Y}$ by the given density

$f(\boldsymbol{Y}|\hat{\boldsymbol{\theta}}^{(t)}, \boldsymbol{X})$, and denote

$$Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(t)}) = E_{\boldsymbol{Y}|\hat{\boldsymbol{\theta}}^{(t)},\boldsymbol{X}}[\log L(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{Y})],$$

- M-step: Update $\boldsymbol{\theta}$ by maximizing $Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(t)})$ at stage t, denoted by $\hat{\boldsymbol{\theta}}^{(t+1)}$

$$\hat{\boldsymbol{\theta}}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(t)}).$$

Dempster *et al.* (1977) proved that the set of likelihood values $\{L(\hat{\boldsymbol{\theta}}^{(t)}; \boldsymbol{X})\}_{t=1,2,...}$ converges to some value L^*, by showing $L(\hat{\boldsymbol{\theta}}^{(t+1)}; \boldsymbol{X}) \geq L(\hat{\boldsymbol{\theta}}^{(t)}; \boldsymbol{X})$. But the convergence of $\{L(\hat{\boldsymbol{\theta}}^{(t)}; \boldsymbol{X})\}_{t=1,2,...}$ does not automatically guarantee the convergence of $\hat{\boldsymbol{\theta}}^{(t)}$. Wu (1983) further proved the convergence of EM iterates $\hat{\boldsymbol{\theta}}^{(t)}$ by requiring more stringent regularity conditions. In many practical applications, in particular, with the assumption of the mixture of Gaussian model that we mainly focus on, L^* will be a local maximum of likelihood values.

5.2 Mixture of Gaussian Models

Definition 24 *The pdf $f(\boldsymbol{x})$ of mixture distributions with K number of components is of the form*

$$f(\boldsymbol{x}) = \sum_{j=1}^{K} \tau_j f_j(\boldsymbol{x}), \quad \sum_{j=1}^{K} \tau_j = 1, \quad \tau_j \geq 0, \;\; j = 1, 2, ..., K,$$

where $f_j(\boldsymbol{x})$ is pdf of the jth component, and τ_j is the proportion weight.

We can further define a K mixture of elliptical models of $f(\boldsymbol{x})$, based on the definition of elliptical distribution in Section 2.2. That is , for $j = 1, 2, ..., K$,

$$f_j(\boldsymbol{x}) = \{det(\boldsymbol{\Sigma})\}^{-1/2} h\{(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\}.$$

In particular, the the mixture of Gaussian distribution is obtained when pdf f_j is associated with $h(t) = (2\pi)^{-p/2}e^{-t/2}$, therefore,

$$f_j(\boldsymbol{x}) = N(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) = (\frac{1}{2\pi})^{p/2}\frac{1}{\sqrt{|\boldsymbol{\Sigma}_j|}}\exp\big(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)\big).$$

The family of mixture of elliptical models contains a quite rich collection of distributions. Perhaps the most widely used is the mixture of Gaussian distribution. Other than that, mixture of t_v and Laplace distributions are also used in modeling the data distribution with heavy tails. Later on, we would introduce a distribution called the mixture of Kotz type distribution. It has a similar form as Gaussian distribution, and the parameters estimated by the tradition EM algorithm has an expression very close to the one from our proposed method.

EM algorithm is found to be very effective in estimating the parameters of mixture models. In particular, if the likelihood function belongs to *exponential family*, all the estimates are formulated in closed form in each iteration step, see Sundberg (1972). Examples of exponential family include the Gaussian, Exponential, Gamma, Chi-squared, etc, while it does not include the Cauchy and Laplace distributions (with mean not equals to zero). The Kotz type distribution that we will see later is one of generalizations of Laplace distribution in a multivariate case.

Many authors (McLachlan & Krishnan, 1997) focus on mixtures of multivariate Gaussian distributions (or sometimes refer to the mixture of Gaussian model) by using EM algorithm to estimate the set of parameters. If $\boldsymbol{X}$ comes form the mixture of Gaussian model, its log-likelihood function is

$$\log L(\boldsymbol{\theta}; \mathbb{X}) = \sum_{i=1}^{n}\log\left(\sum_{j=1}^{K}\tau_j f(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\right)$$

For many models, if the likelihood function is convex, the Newton-Raphson method is used to solve for the maximum in usual. Nevertheless, in this case, the Hessian matrix used in Newton method involves second derivatives of logarithm of summation, which is difficult to derived analytically. Even worse, numerically it might be closed to a singular matrix. So, the Newton method would not be applicable. People have to find a more reliable solution as an alternative. The EM algorithm, however, comes to be the best fit in such case. One can maximize the likelihood function by introducing the "unobserved variable" $\boldsymbol{Y}$, and iteratively update $\hat{\boldsymbol{\theta}}$ until it converges.

Specifically, for the mixture of Gaussian model, let $\mathbb{X} = \{\boldsymbol{X}_1, ..., \boldsymbol{X}_n\}$ be a collection of random samples of $\boldsymbol{X} \in \mathbb{R}^p$ from mixture model f. We know that the probability density of jth component, f_j is parameterized by $f_j(\boldsymbol{x}) = N(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, $j = 1, 2, ..., K$. Then, one can retrieve the mixture model by finding the parameters of components $\{\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\}_{j=1}^K$ and the weight factors $\{\tau_j\}_{j=1}^K$. To keep the consistence of notations used in (5.1), we combine them as a set of parameters $\boldsymbol{\theta} = \{\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, \tau_j\}_{j=1}^K$. Then the MLE of $\boldsymbol{\theta}$, denoted by $\hat{\boldsymbol{\theta}}$, is solved by maximizing the likelihood function of $\boldsymbol{X}$ (or the log likelihood function base on the random sample $\mathbb{X}$), such that,

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \log L(\boldsymbol{\theta}; \mathbb{X}) = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log \left(\sum_{j=1}^{K} \tau_j f_j(\boldsymbol{x}_i|\boldsymbol{\theta}) \right).$$

In the language used in describing the EM algorithm, the random variable $\boldsymbol{X}$ is the "observed data". We can "complete" the data by introducing the variable $\boldsymbol{Y}$, which is the "unobserved data". For each $\boldsymbol{X}_i$, define a latent variable $\boldsymbol{Y}_i = (Y_{1,i}, Y_{2,i}, ..., Y_{K,i})^T$, such that $Y_{j,i} = 1[\boldsymbol{X}_i \sim f_j]$, which is the indicator function such that $Y_{j,i} = 1$ if $\boldsymbol{X}_i$ is taken from jth component of mixture, 0, otherwise. So, in the E-step

$$\begin{aligned} Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(t)}) &= E_{\boldsymbol{Y}|\hat{\boldsymbol{\theta}}^{(t)}, \boldsymbol{X}} [\log L(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{Y})] \\ &= E_{\boldsymbol{Y}|\hat{\boldsymbol{\theta}}^{(t)}, \boldsymbol{X}} \sum_{i=1}^{n} \sum_{j=1}^{K} Y_{j,i} \log \tau_j f_j(\boldsymbol{x}_i) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{K} E(Y_{j,i}|\hat{\boldsymbol{\theta}}^{(t)}, \boldsymbol{x}_i)(\log \tau_j + \log f_j(\boldsymbol{x}_i)). \end{aligned} \tag{5.3}$$

In (5.3), we need the conditional expectation of $Y_{j,i}$ given $\boldsymbol{X} = \boldsymbol{x}_i$ and $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^{(t)}$, $E(Y_{j,i}|\hat{\boldsymbol{\theta}}^{(t)}, \boldsymbol{x}_i)$. Because $Y_{j,i}$ is Bernoulli distributed, it is equal to $P(Y_{j,i} = 1|\hat{\boldsymbol{\theta}}^{(t)}, \boldsymbol{x}_i)$.

If we define variable $T_{j,i}^{(t)} = P(Y_{j,i} = 1|\hat{\boldsymbol{\theta}}^{(t)}, \boldsymbol{x}_i)$, and use marginal distribution $P(Y_{j,i} = 1) =$

$\tau_j^{(t)}$, for $j = 1, ..., K$, then the estimate of $T_{j,i}^{(t)}$ (the conditional probability) can be determined by *Bayes theorem*, that is,

$$T_{j,i}^{(t)} = \frac{\tau_j^{(t)} f(\boldsymbol{x}_i | \boldsymbol{\mu}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)})}{\sum_{j=1}^{K} \tau_j^{(t)} f(\boldsymbol{x}_i | \boldsymbol{\mu}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)})}. \tag{5.4}$$

Therefore, in M-step, maximize $Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(t)})$ w.r.t. $\boldsymbol{\tau}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$, we have for each $j = 1, ..., K$,

$$\tau_j^{(t+1)} = \frac{\sum_{i=1}^{n} T_{j,i}^{(t)}}{\sum_{j=1}^{K} \sum_{i=1}^{n} T_{j,i}^{(t)}} = \frac{1}{n} \sum_{i=1}^{n} T_{j,i}^{(t)}. \tag{5.5}$$

If we denote

$$w_{j,i}^{(t)} = \frac{T_{j,i}^{(t)}}{\sum_{i=1}^{n} T_{j,i}^{(t)}}, \tag{5.6}$$

then

$$\boldsymbol{\mu}_j^{(t+1)} = \sum_{i=1}^{n} w_{j,i}^{(t)} \boldsymbol{x}_i, \tag{5.7}$$

$$\boldsymbol{\Sigma}_j^{(t+1)} = \sum_{i=1}^{n} w_{j,i}^{(t)} (\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(t+1)})(\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(t+1)})^T. \tag{5.8}$$

5.3 Limitations of Regular EM on Mixture Model

As an iterative algorithm seeking for the maximum of the likelihood function, the EM algorithm with the mixture of Gaussian model suffers from some problems. This method is sensitive to (1) the number of components, (2) initial values and (3) outliers in sample. Issues (1) and (3) are stem from the usage of likelihood function in general.

5.3.1 Number of components

Similar to a model selection problem, the issue (1) can be categorized into two cases. First, in the situation that the number of components has physical meaning given by the nature of the problem, we don't want it to be varied but rather fixed. In other situation, even though the data are known being collected from a C number of groups, Chances are the data may not come from a mixture of C component Gaussian distributions. For instance, one of the group is constituted by more than one Gaussian components. Or, two or more groups are so closed that it might be better to combine as one single Gaussian component. So a criterion of splitting or combining components has to be considered. One way to select the number of components is cross validation. Please refer to our experiment on fish species data on in Chapter 7. Besides that, some other techniques based on model selection were proposed. Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are set to maximize the likelihood function plus some penalty terms on the model complexity, see Schwarz (1978). Medasani & Krishnapuram (1997) used an agglomerative method to choose the number of components. Zhang & Cheung (2006) proposed the X-EM algorithm to perform model selection by fading the redundant components out from a density mixture, meanwhile estimate the model parameters appropriately. Vetrov *et al.* (2010) proposed different methods on determination of the number of components by other criteria. It is worth to point out, because both AIC and BIC are methods based on likelihood function, they are easily influenced by existence of outlying observations in the sample. Issue (3) still be unsolved.

5.3.2 Initial Values

Since EM algorithm is an iterative algorithm, the initial value often affects the convergence of the algorithm, and whether it converges to global maximum or a local one. In practice, the issue (2) is mainly caused by the initial location parameters. This problem can be alleviated by multiple *random restarts*. One compares the values of likelihood functions on multiple runs with different initial values, and choose the one with the highest likelihood value. Herein, location estimates from K-Means (MacQueen, 1967) and Fuzzy C-Means (Bezdek, 1981), are frequently used to initialize the location of components.

5.3.3 Outliers in Sample

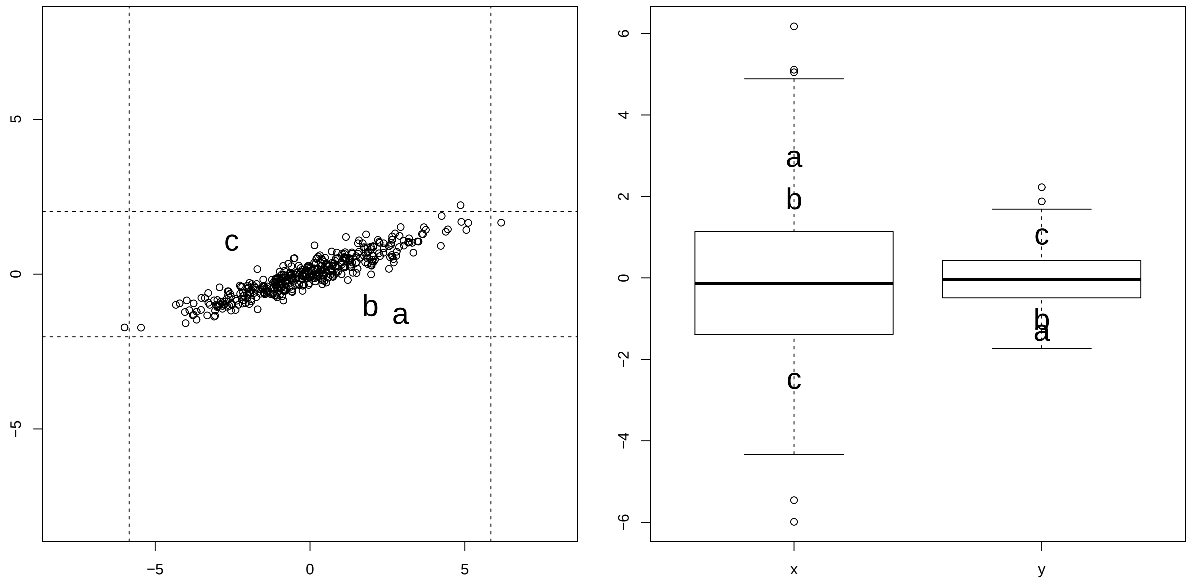
As the data being analyzed coming from a wider and wider field in recent years, the issue (3), outlying observation in the sample, draws more and more attention in statistics society. Some of applied researchers use EM algorithm on mixture of Gaussian model without careful check of the data with presence of outliers. Part of the reason is due to the hardness of detecting outlying observations in high dimensional spaces. For example, in the univariate case, outliers can be detected by traditional methods such as box-whisker plot. However, this method on each feature of p features is not helpful to find out outliers in high dimensions, see Figure (5.1). Marginal checking of each features for outlyingness is inadequate. Statistician usually used the *hat matrix* $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ and *leverage* $\boldsymbol{x}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}$ to identify the outlier in one multivariate elliptical distribution. Kutner *et al.* (2005), chapter 10, gave a through summary to identify the outliers base on the hat matrix and leverage values. It is not difficult to imagine that outliers are even harder to be identified if they are hidden in the mixture of multivariate distributions. This is the problem we are trying to tackle in this dissertation.

Since the outliers are the focal point in this dissertation, we would briefly review some existing methods that are robust to outlier with the EM algorithm on the mixture model as follows.

First, perhaps the most intuitive way to deal with outliers is to add the number of components. It assumes that outliers can be explained by one or more mixture components, see Banfield & Raftery (1993). However, adding the number of components may increase the complexity of the model and therefore create an overfitting model. An extreme case can be thought as fitting k outliers in the training set with k components. It is obvious that this model would be useless for prediction. Further, increasing the number of components may lead the model difficult to interpret, and lost the potential physical meaning. For example, in breast cancer diagnostic data, see Chapter 7, patients are naturally categorized into two groups (components) being benign, or malignant. Two Gaussian components are used to model the data and provide a clustering on whether or not the patient is malignant. Adding extra components to explain outliers would destroy the interpretation of the model.

Rather than adding components, one may use distributions with heavier tails than Gaussian distribution to model the mixture components, e.g., mixture of t_ν-distributions or Kotz type distributions, see Lange *et al.* (1989), and Shoham (2002). Assumption on heavy-tailed distributions often provide robustness against outlier in statistical analysis. For the mixture of t_v-distributions, degree of freedom ν controls how heavy the of distribution tails are. For instance, t_1 is the Cauchy distribution which doesn't even have the first moment, and t_∞ is the Gaussian distribution. The EM algorithm based on assumptions of these distributions is deemed to be more sensitive to outliers with large value of ν , and more robust with smaller value of ν. I personally consider this method is more appropriate than the first approach to overcome the influences from outliers. It is more general and interpretable. Many of researches have been done on using mixture of t_ν distributions. However, because the problem involves an estimation of ν, no explicit form of the estimators of location and scatter parameters can be obtained. For other heavy-tailed distributions, e.g. the Kotz type distribution (a generalized Laplace distribution in the multivariate case), location and scatter estimates in each M-step have to be numerically solved by the other algorithm, which dramatically increases the computation complexity.

Third, use robust methods in M-step in EM algorithm. Tadjudin & Landgrebe (2000) uses a robust M-estimators defined by theψ-function (Huber, 1964) as a hybrid of Gaussian distribution with a Laplacian tail. Medasani *et al.* (1998) uses least trimmed squared estimators. We know that infinitesimal robust M- estimators, see Section 1.4.2, are obtained by minimizing some given $\rho(\cdot)$ functions that have an odd and bounded derivative function $\psi(\cdot)$. Qin & Priebe (2012) adopted the similar idea to maximize the L_q likelihood such that $\hat{\boldsymbol{\theta}}_{MLqE} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^n L_q(f(\boldsymbol{x}_i; \boldsymbol{\theta}))$, where $L_q(u) = (u^{1-q} - 1)/(1-q)$ and $q > 0$. Our proposed robust EM algorithm belongs to this category. It will be evident that our rank-based location estimator in EM is a robust $M-$estimator. And these spatial rank based estimates of location and scatter are also shown to closely relate to the MLE of mixture of Kotz type distributions.



(a) (b)

Figure 5.1. Outlying observations in 2-dimensional space

(a) is the data generated by 2-dimensional multivariate Gaussian distribution with $\boldsymbol{\mu} = (0,0)^T$, $\boldsymbol{\Sigma} = ((4,1.3)^T,(1.3,.5)^T)^T$. The points a, b and c are consider to be outliers in the sample. (b) is the box-and-whisker plot for data in (a). Obviously, none of the point a, b or c is detected as an outlier by marginal checking.

5.4 Sensitivity of Location and Scatter Estimates

Regular EM algorithm returns decent estimates of parameters when data is distributed as mixture of Gaussian. The algorithm gains its fame on applications in Statistical/Machine Learning society. Chandola *et al.* (2007) gave a comprehensive review on it.

In the previous sections, we keep saying that regular EM algorithm is weak in overcoming problems from presence of the outlying observations. In here, we simply explain the reason why. By revisiting the expression of those location (5.4) and scale (5.4) estimates in M-step,

$$\boldsymbol{\mu}_j^{(t+1)} = \sum_{i=1}^{n} w_{j,i}^{(t)} \boldsymbol{x}_i,$$

$$\boldsymbol{\Sigma}_j^{(t+1)} = \sum_{i=1}^{n} w_{j,i}^{(t)} (\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(t+1)})(\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(t+1)})^T.$$

The similar formulation to the location estimator sample mean and scatter estimator sample covariance causes the problem. They are not infinitesimal robust in terms of influence function. The unboundness in $\boldsymbol{x}_i$ would draw the estimates to be severely biased from the true parameters. This is illustrated in the Figure 5.2. The "normal" data o's are from a same mixture of two component Gaussian distribution on both graphes. The elliptical covariance contours are estimated by regular EM Algorithm on both cases with same initial condition. However, with the single contamination point * in the upper right corner in (b), one of covariance estimates is stretched drastically in the corresponding direction.

If we look at the quantitative robustness of estimates of regular EM algorithm on mixture Gaussian model by RBP, both the location (mean), and scatter (covariance) estimates have RBP equal to 1/n, which is extremely low. In spite of the fact that it is impractical to collect a data point equals to "∞" in a sample, the small value of RBP implies that even tiny amount of large points deviate from the "normal" data points could exert undue influence on the estimates. Again, Figure 5.2 explains the phenomenon.

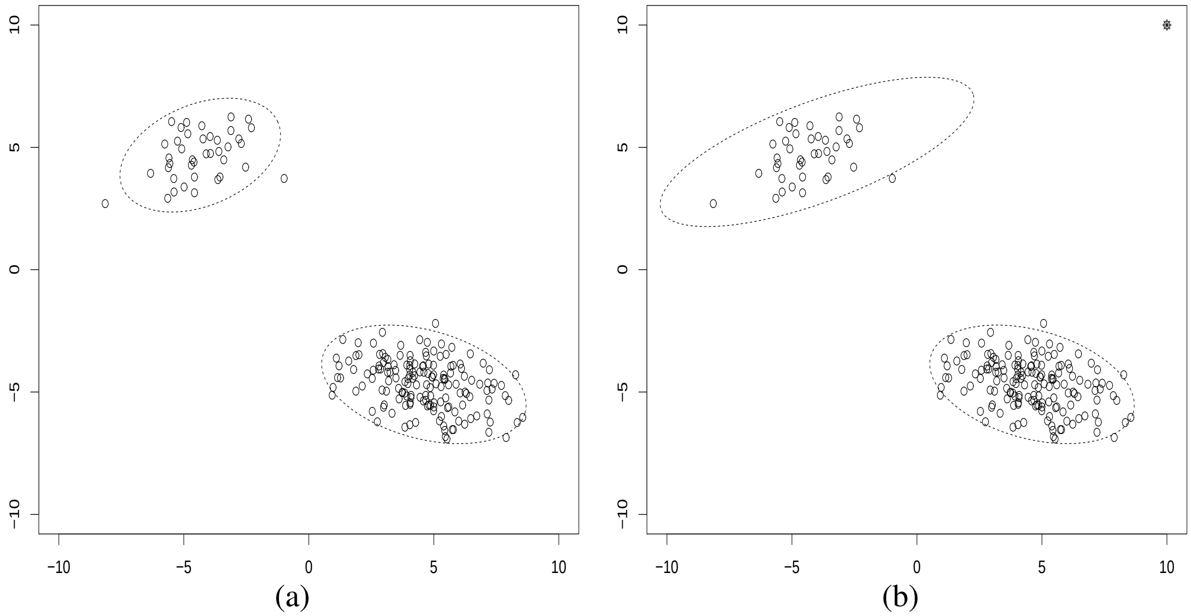


Figure 5.2. Sensitivity of regular EM algorithm to single contamination point

“Normal” data o’s are from a same mixture of 2-component Gaussian distribution on both graphes. The elliptical covariance contours are estimated by regular EM Algorithm on both cases with same initial condition. However, with the single contamination point * in the upper corner in (b), one of covariance estimates is stretched drastically. Therefore any further inferences based on it would be badly impacted.

Chapter 6

Spatial EM

The idea to strengthen robustness of regular EM algorithm on mixture of Gaussian model comes from the closed forms of $\boldsymbol{\mu}_j$ and Σ_j in M-step. It is shown that the estimates are not robust there. We want to solve the problem at the place where it is. Can we use robust estimators to estimate the location and scatter parameters of different components? Would that be reasonable and applicable? In this chapter, first of all, we discuss why the spatial median and MRCM is chosen. Second, we show how they imbed into the EM algorithm. Third, we explore some connection between our method with the mixture of Kotz type distribution from the point of likelihood function.

6.1 Why Spatial Median and MRCM

As noticed in the previous chapter, if we can use some robust estimators to estimate the location and scatter parameters for different components, the sensitivity problem to the outlying observation of EM algorithm can be solved. Our choices of these robust estimators are the spatial median and MRCM we proposed in Chapter 3.

Similar to the robust location and scale estimators, median and MAD, in univariate case, one of the extension in high dimensional spaces are are the spatial median and MRCM. We give the definition and way to construct above statistics in Chapter 3. Remember that we define the spatial rank to be the expectation of spatial sign $\boldsymbol{S}(\boldsymbol{x}-\boldsymbol{X})$, such that

$$\boldsymbol{R}(\boldsymbol{x}, F) = E_F\{\boldsymbol{S}(\boldsymbol{x}-\boldsymbol{X})\} = E_F\frac{\boldsymbol{x}-\boldsymbol{X}}{\|\boldsymbol{x}-\boldsymbol{X}\|},$$

where $\| \cdot \|$ is the Euclidean norm. Then spatial median $\boldsymbol{M}(\boldsymbol{X})$ is defined to be the solution of $\|\boldsymbol{R}(\boldsymbol{x}, F)\| = \mathbf{0}$. It is shown in Lemma 14, for any $\boldsymbol{X}$, the associated rank $\boldsymbol{R}(\boldsymbol{X}, F)$ is infinitesimal robust. Part of the reason is that the influence of every data point is being standardized by its Euclidean norm. Every single point is mapped to a unit vector but the information of direction is reserved. Thus even the extremist can only influence the spatial median by just one unit in the corresponding direction. In fact, spatial median can also be treated as a type of M-estimator. It is obtained by minimizing the loss function $\rho(\boldsymbol{X} - \boldsymbol{\mu}) = E_F\|\boldsymbol{\mu} - \boldsymbol{X}\|$ w.r.t. $\boldsymbol{\mu}$, which has a componentwise odd and bounded derivative $\psi(\boldsymbol{X} - \boldsymbol{\mu}) = E_F\boldsymbol{S}(\boldsymbol{\mu} - \boldsymbol{X})$. Therefore, it is proven that the spatial median has the RBP asymptotically $1/2$ and a bounded influence function (Jurečková & Picek, 2006).

In Chapter 4, the MRCM is shown to be a robust affine equivariant scatter estimator of elliptical models. It ensures that the quantity transformed accordingly under rotation, translation (shift center) and even heterogenous scale changes of data. Not only that, it has a bounded influence function. The RBP can attain the upper bound by properly chosen MAD$_k$. It is statistically relatively efficient. Comparing to other robust scatter estimators we mention in the Section 1.4 and Chapter 2, the MRCM is easy to compute, well balanced between the statistical robustness and efficiency. We therefore hope to extend the notions of the spatial median and MRCM to regular EM for mixture models.

However, in the mixture of elliptical models, the calculation of MRCM for each component is not as simple as it seems. In order to illustrate our thinking process, it is better to step back into the way of constructing the MRCM and see how we adjust it to the mixture model.

MRCM looks complicated to discuss by the way it formulated. However, it would be easier if we separate our viewpoints into looking at direction and looking at scaler on that direction.

Robustness on directions of MRCM is shielded, because MRCM is a function of bounded spatial ranks. Eigenvectors of MRCM are identical to those of RCM, and thanks to Marden (1999), they preserve the orientation of the original elliptical model. For the scaler, the robust estimator MAD that measures the dispersion of projection on those directions (eigenvectors) would again prevent the outliers from interfering. When it comes to mixture of elliptical distributions, data that are used to construct the component MRCM should belong to the corresponding the component. That is, in each iteration, when construct the MRCM for each component on the mixture of elliptical models, both the direction and scaler estimates should be emphasized more by the data points from that component, but ignore or reduce the influences of the data points from other components. It seems redundant to say that, since we are estimating the scatter of each component rather than the whole data. But still, as it would be seen later, this is the trickiest part in the Spatial-EM algorithm.

6.2 Spatial-EM Algorithm on Mixture of Gaussian model

The concept to strengthen the robustness of regular EM algorithm on mixture of Gaussian model comes from the closed forms of $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ in the M-step. The basic idea of *Spatial-EM* is to replace the estimators in M-step by the spatial median and MRCM.

Following the notations used in the Section 5.2, the implementation of Spatial-EM is as follows.

Algorithm 25 *Spatial-EM Algorithm*

```
{Initialization}
  Use K-means for the centers $\{\boldsymbol{\mu}_j^{(0)}\}_{i=1}^{K}$,
  $\boldsymbol{\Sigma}_j^{(0)} = \boldsymbol{I}_{p\times p}$ (identity matrix),
  $\tau_j^{(0)} = 1/K$, for $j = 1, ..., K$
Do Until $\boldsymbol{\mu}_j^{(t)}$, $\boldsymbol{\Sigma}_j^{(t)}$ and $\tau_j^{(t)}$'s converge for all $j$
    $t = 1$
    For $j = 1$ To $K$
E-Step:
            Calculate $T_{j,i}^{(t)}$ by equation (5.4)
M-Step:
        Update $\tau_j^{(t)}$'s by equation (5.5)
        Define $w_{j,i}^{(t)}$ as equation (5.6)
        Find weighted spatial median $\boldsymbol{\mu}_j^{(t+1)}$ (Refer to Algorithm 26)
        Find weighted MRCM $\tilde{\boldsymbol{\Sigma}}_j^{(t+1)}$ (Refer to Algorithm 27)
    End
    $t = t + 1$
End
```

Obviously, we need the following two functions for the jth component spatial median and MRCM,

Algorithm 26 *Function for weighted spatial median $\boldsymbol{\mu}_j^{(t+1)}$*

```
Input $\{\boldsymbol{x}_i\}_{i=1}^n$, $w_{j,i}^{(t)}$
For $\ell = 1$ To $n$
    $\boldsymbol{R}_j^{(t+1)}(\boldsymbol{x}_\ell) = \sum_{i=1}^n w_{j,i}^{(t)} \boldsymbol{S}(\boldsymbol{x}_\ell - \boldsymbol{x}_i)$
End
$\boldsymbol{\mu}_j^{(t+1)} = \arg\min_{\boldsymbol{x}_\ell} \|\boldsymbol{R}_j^{(t+1)}(\boldsymbol{x}_\ell)\|$
Output $\boldsymbol{\mu}_j^{(t+1)}$
```

Algorithm 27 *Function for weighted MRCM $\tilde{\boldsymbol{\Sigma}}_j^{(t+1)}$*

```
Input $\boldsymbol{R}_j^{(t)}(\boldsymbol{x}_i)$,$T_{j,i}^{(t)}$, $w_{j,i}^{(t)}$, $\boldsymbol{\mu}_j^{(t+1)}$
Compute $j$th component RCM
    $\boldsymbol{\Sigma}_{R,j}^{(t+1)} = \sum_{i=1}^n w_{j,i}^{(t)} \big(\boldsymbol{R}_j^{(t)}(\boldsymbol{x}_i)\big)\big(\boldsymbol{R}_j^{(t)}(\boldsymbol{x}_i)\big)^T$
Find eigenvectors $\boldsymbol{u}_{j,m}$'s of $\boldsymbol{\Sigma}_{R,j}^{(t+1)}$.
For $m = 1$ To $p$
    Generate projected sequence
    $\{T_{j,i}^{(t)} \boldsymbol{u}_{j,m}^T (\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(t+1)})\}_{i=1,..,n}$
    Sort the set $\{T_{j,i}^{(t)} \boldsymbol{u}_{j,m}^T (\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(t+1)})\}_{i=1,2,..,n}$,
    take away the $\lceil n(1 - \tau_j^{(t+1)}) \rceil$ smallest values,
        then denote the new collection by
        $\{T_{j,i_k}^{(t)} \boldsymbol{u}_{j,m}^T (\boldsymbol{x}_{i_k} - \boldsymbol{\mu}_j^{(t+1)})\}_{i_k}$
        $\hat{\lambda}_{j,m} = \textit{MAD}\ (\{T_{j,i_k}^{(t)} \boldsymbol{u}_{j,m}^T (\boldsymbol{x}_{i_k} - \boldsymbol{\mu}_j^{(t+1)})\}_{i_k})$
End
$\hat{\boldsymbol{\Lambda}}_j = \textit{diag}(\hat{\lambda}_{j,1}, ..., \hat{\lambda}_{j,p})$
MRCM $\tilde{\boldsymbol{\Sigma}}_j^{(t+1)} = \boldsymbol{U}_j \hat{\Lambda}_j \boldsymbol{U}_j^T$
Output $\tilde{\boldsymbol{\Sigma}}_j^{(t+1)}$
```

In fact, different criteria can be used as the stoping rule of the algorithm. Especially, for Algorithm 25, if a stopping rule involves the converge of vectors and matrices, it would be time consuming to check one by one. In practice, We can just simply set the algorithm to stop when $\tau_j^{(t)}$ get converged for all j.

6.2.1 On M-Step

There are several places worth to be noted in M-step.

First is the way to update $\boldsymbol{\mu}_j^{(t+1)}$. In the algorithm 26, we use the spatial median by convolving $w_{j,i}$ in the definition of spatial ranks for component j. For simplicity, we confine our search in the pool of sample points. That is, replace

$$\boldsymbol{\mu}_j^{(t+1)} = \boldsymbol{x} \text{ such that } \left\| \boldsymbol{R}_j^{(t)}(\boldsymbol{x}) \right\| = \left\| \sum_{i=1}^{n} w_{j,i}^{(t)} \boldsymbol{S}(\boldsymbol{x} - \boldsymbol{x}_i) \right\| = \boldsymbol{0} \tag{6.1}$$

by

$$\boldsymbol{\mu}_j^{(t+1)} = \arg\min_{\boldsymbol{x}_i} \left\| \boldsymbol{R}_j^{(t+1)}(\boldsymbol{x}_i) \right\|.$$

This would save a great amount of computational time and works fine when the sample size is large enough.

Second, in the way of defining MRCM for a certain component at the tth iteration , we need to set up a weighted RCM in algorithm 27. It is not hard to see, for the points that can be well clustered into different components, $T_{j,i}^{(t)}$ would be either close to 1 or 0. It is similar to a binary classification on whether a point belongs to jth component or not. So the factor $w_{j,i}^{(t)} = T_{j,i}^{(t)} / \sum_{i=1}^{n} T_{j,i}^{(t)}$, can provide a proper weight to average the elements that belongs to the jth component. As the iteration goes on, the jth component RCM would be finally stands out by "picking" the correct ranks using $w_{j,i}^{(t)}$.

Third, construction of MRCM becomes trickier when applying MAD to the projection data on each eigenvector of RCM. As shown on the step 5 in algorithm 27, we first centralize the data by shifting toward the respective spatial medians $\boldsymbol{\mu}_j^{(t+1)}$, then multiply the factor $T_{j,i}^{(t)}$'s and generate the whole sequence of $\{T_{j,i}\boldsymbol{s}_{j,m}^T(\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(t+1)})\}_{i=1,..,n}$. Again, because each $T_{j,i}^{(t)}$ would play as a classifier and degenerate to 0 if ith data point does not belong to jth component. The sequence above thus contains quite a lot of small values (probably sufficiently close to 0) that indicates the corresponding points do not belong to the component j. Therefore, we shall omit the smallest $\lceil n(1-\tau_j^{(t+1)})\rceil$ number of such values, then apply MAD on the rest set of projection data. Various experiments show that this is a reasonable robust scale estimator on all eigen-directions for mixture of Gaussian distributions. Taking the normality consistency into account, the consistent factor $3/4$th quantile of Gaussian distribution, $\Phi^{-1}(3/4) \approx 1.4862$, is multiplied to all the MAD estimates. In this way, if data is Gaussian distributed, MAD is an unbiased estimator of the true standard deviation.

6.2.2 More on M-Step

It is interesting to compare the M-step between regular EM and Spatial-EM. In fact, for regular EM, location and scatter estimates can be viewed as the M-estimators by minimizing the objective function $\rho_1(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, which is negatively proportional to the log-likelihood function modeled by a mixture of Gaussian w.r.t. $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$, sequentially.

$$\rho_1(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) = \sum_{i=1}^{n} T_{j,i}\{\log|\boldsymbol{\Sigma}_j| + (\boldsymbol{x}_i - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_j)\}. \tag{6.2}$$

Because

$$\frac{\partial \rho_1}{\partial \boldsymbol{\mu}_j} = -2 \sum_{i=1}^{n} T_{j,i} (\boldsymbol{x}_i - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}^{-1},$$

$$\frac{\partial \rho_1}{\partial \boldsymbol{\Sigma}_j^{-1}} = \sum_{i=1}^{n} T_{j,i} \boldsymbol{\Sigma}_j - \sum_{i=1}^{n} T_{j,i} (\boldsymbol{x}_i - \boldsymbol{\mu}_j)(\boldsymbol{x}_i - \boldsymbol{\mu}_j)^T$$

are both unbounded in $\boldsymbol{x}$, they are not considered to be robust from the theory of M-estimator. Therefore, being a robust method, we have to change from maximizing the likelihood of mixture Gaussian model to other possible mixture models that might have a bounded differentiate w.r.t. parameters.

Simply put, we can change the assumption from Gaussian mixture to some other elliptical distributions to improve the robustness. However, what exactly the model likelihood that Spatial-EM is trying to maximize would be difficult to answer. By investigating one type of heavier tail elliptical distribution, we discover some amusing similarities with Spatial-EM. This is the *Kotz type distribution*. It has a heavier tail regions than that of multivariate Gaussian. Its density function is given by, see Fang & Anderson (1990),

$$g(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = c_p |\boldsymbol{\Sigma}|^{-1/2} \exp\{-[(\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu})]^{1/2}\}, \tag{6.3}$$

where location $\boldsymbol{\mu} \in \mathbb{R}^p$, scatter $\boldsymbol{\Sigma}$ is a positive definite symmetric $p \times p$ matrix, and $c_p = \frac{\Gamma(p/2)}{(2\pi)^{p/2}\Gamma(p)}$. It obviously falls into the family of elliptical symmetric distribution. Moreover, those estimators in Spatial EM have closed relationship with MLE of mixture of Kotz type distribution.

Suppose that data come from a mixture of Kotz type distribution, one can obtain the MLE by EM algorithm. Maximizing the Q function in (5.3) w.r.t $\boldsymbol{\mu}_j$ and Σ_j would be equivalent to minimizing the following objective function,

$$\rho_2(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) = \sum_{i=1}^{n} T_{j,i}\big(\frac{1}{2}\ln|\boldsymbol{\Sigma}_j| + \sqrt{(\boldsymbol{x}_i - \boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_j)}\,\big).$$

Its first derivatives are

$$\begin{aligned}\frac{\partial\rho_2}{\partial\boldsymbol{\mu}_j} &= \sum_{i=1}^{n} T_{j,i}\frac{(\boldsymbol{x}_i - \boldsymbol{\mu}_j)^T}{\|(\boldsymbol{x}_i - \boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}_j^{-1/2}\|}\boldsymbol{\Sigma}_j^{-1} \\ &= \sum_{i=1}^{n} T_{j,i}\boldsymbol{S}((\boldsymbol{x}_i - \boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}_j^{-1/2})\boldsymbol{\Sigma}_j^{-1/2}, \end{aligned} \tag{6.4}$$

$$\frac{\partial\rho_2}{\partial\boldsymbol{\Sigma}_j^{-1}} = \sum_{i=1}^{n} T_{j,i}\boldsymbol{\Sigma}_j - \sum_{i=1}^{n} T_{j,i}\frac{(\boldsymbol{x}_i - \boldsymbol{\mu}_j)(\boldsymbol{x}_i - \boldsymbol{\mu}_j)^T}{\sqrt{(\boldsymbol{x}_i - \boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_j)}}. \tag{6.5}$$

It is not hard to see, the influence of $\boldsymbol{x}_i$ in both derivatives is either bounded by spatial sign $\boldsymbol{S}(\cdot)$ in (6.4) or bounded by “standardization” in (6.5). $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ thus are considered to be robust in theory of M-estimator.

Further, set the above two derivatives to zero, we can solve for the $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$. They have the similar form as what we defined weighted spatial median (6.1) and weighted MRCM in algorithm 27.

$$\sum_{i=1}^{n} w_{j,i}\boldsymbol{S}((\boldsymbol{x}_i - \boldsymbol{\mu}_j)\boldsymbol{\Sigma}_j^{-1/2}) = \boldsymbol{0}$$

$$\boldsymbol{\Sigma}_j = \sum_{i=1}^{n} w_{j,i}\frac{(\boldsymbol{x}_i - \boldsymbol{\mu}_j)(\boldsymbol{x}_i - \boldsymbol{\mu}_j)^T}{\sqrt{(\boldsymbol{x}_i - \boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_j)}} \tag{6.6}$$

Rao (1988) proposed a way to solve this generalized spatial median $\boldsymbol{\mu_j}$ and rank covariance matrix $\boldsymbol{\Sigma_j}$ problem. Plungpongpun & Naik (2008) presented an algorithm to compute them. First initialize $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{j}}$ and then solve $\hat{\mu}_j$ by minimizing

$$\sum_{i=1}^{n} w_{j,i}\sqrt{(\boldsymbol{x}_i - \boldsymbol{\mu}_j)^T\hat{\boldsymbol{\Sigma}}_j^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_j)}, \tag{6.7}$$

denoted it as $\hat{\boldsymbol{\mu}}$. Then update $\hat{\boldsymbol{\Sigma}}_j$ sequentially

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{j}} = \sum_{i=1}^{n} w_{j,i} \frac{(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{\boldsymbol{j}})(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{\boldsymbol{j}})^T}{\sqrt{(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{\boldsymbol{j}})^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{j}}^{-\boldsymbol{1}} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{\boldsymbol{j}})}} \tag{6.8}$$

until they converge.

Different from Spatial EM, the above estimates of location and covariance require the data transformation w.r.t. $\boldsymbol{\Sigma}^{-1/2}$ prior to each calculation of "median" $\boldsymbol{\mu}$ and "RCM" $\boldsymbol{\Sigma}_R$. In Particular, for location parameter $\boldsymbol{\mu}_j$, instead of minimizing (6.7), the Spatial EM attempts to minimize $\sum_{i=1}^{n} w_{j,i}\sqrt{(\boldsymbol{x}_i - \boldsymbol{\mu}_j)^T(\boldsymbol{x} - \boldsymbol{\mu}_j)}$. These two versions of spatial medians are not generally identical. However, they are the same if $\boldsymbol{\Sigma} = c\boldsymbol{I}$.

The EM algorithm with mixture of Kotz type distribution is theoretically tractable, but it is not computationally ease. As suggested by Rao, an order to solve for the $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{j}}$, an inner iteration has to be done in each M-step. It would significantly increase the computation complexity. Other than that, involving $\boldsymbol{\Sigma}_j^{-1}$ in denominator of (6.6) not only causes more computational time ($O(p^3)$ needed to find the $\boldsymbol{\Sigma}^{-1}$) but also makes the algorithm unstable to converge. Imagine if there is an extreme outlier in jth component, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{j}}$ tends to be inflated and makes the denominator of (6.8) small, hence result in even more inflation of $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{j}}$ in the next run. Therefore, the risk of getting large estimate of $\boldsymbol{\Sigma}_j$ piles up if extreme outlier exists. Because of the cut-off effect brought by MAD, it is clear that scatter parameter estimated by Spatial-EM is more robust in the sense of RBP than that of the mixture of Kotz type distribution In fact, for these two similar versions of rank covariance matrices, we would like to give more detailed discussion and experiment in the future.

Chapter 7

Application on Statistical Learning

7.1 Introduction

The problems from science and industry constantly challenge the field of Statistics. The statisticians in early days mainly dealt with the problems from agricultural and industrial experiments, which were relatively small in scope. With emerging of advanced computers and flooding flow of information, statistical problems becomes more large in size and more difficult in complexity. Huge amounts of data are generated from different areas. It is the statistician's job to extract important trends and patterns from the data, and tell the story form the data. This process is called *learning from data*, see Hastie *et al.* (2009)

Provided by a survey of Chandola *et al.* (2007), based on types of input data, labeled or unlabeled, problems can be categorized into supervised or unsupervised. In supervised learning, the goal is to predict the value of outcome variable based on a number of input measure. In unsupervised learning, there is no outcome measure, and the goal is to describe the association and pattern among the input measures. The EM algorithm with assumption of the mixture of Gaussian model has been used for supervised and unsupervised statistical learning purposes. Since, the spatial-EM algorithm is robust against the outlier. It is straightforward to use spatial-EM to build a robust outlier detection model.

Using EM to perform the outlier detection can be viewed as unsupervised or supervised classification problem. In the unsupervised setting, the training sample contains normal instances and outliers without class labels. EM builds a mixture model on the whole sample, then identify outliers in the sample based on the distribution. This is out experiment in Section 7.2.2. In the supervised setting, we construct the mixture model by EM only on the normal instances. Different from the unsupervised learning, the class label information (normal) is used. Outlier is identified if it largely deviates from the mixture model. Section 7.2.3 would give a detailed explanation on this unsupervised outlier detector.

Further, the nature of mixture of Gaussian model also give a chance of using spatial EM algorithm to perform a clustering. By a given number of components, one can use spatial-EM to as a robust method to establish the mixture model. The pattern of the data can be recognized by clustering data into different components.

In this chapter, we mainly focus on the application (experiment) of Spatial-EM on outlier detection and clustering. In the Sections 7.2- 7.2.3, outliers in a mixture of Gaussian model are defined. Two experiments to detect outliers are illustrated. One is on synthetic data, the other is on real data set (fish data). In the fish data experiment, a heuristic robust method to estimate the number of components of the mixture model is also proposed. The results of Spatial-EM are compared to the regular EM and other methods. In the Sections 7.3-7.3.3, clustering with the mixture model is briefly reviewed. Two experiments of real data set, UCI Wisconsin Diagnostic Breast Cancer Data and Yeast Cell Cycle Data are used to evaluate the model performance of Spatial-EM, regular EM and some other existing unsupervised and supervised learning techniques.

7.2 Outlier Detection

7.2.1 Outlyingness and Two Types Errors

As seen in Chapter 1 example 1 and Section 5.3.3, an outlier is an observation that is numerically distinct from the rest of data. There are various different ways to define an outlier. Outlier determination is often associated with a threshold parameter depending on different notions of outlyingness measure. An observation with outlyingness beyond the threshold is claimed as outlier. Before we continue, it is necessary to clarify the definition of underlying outliers in the mixture of Gaussian model.

It is well known that if p-variate random vector $\boldsymbol{X}$ is distributed as $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the square of Mahalanobis distance $(\boldsymbol{X} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu})$ has a $\chi^2_{(p)}$ distribution. Based on this, we can easily define the region of outlier in the following way. Assume the $p-$variate random vector $\boldsymbol{X}$ has distribution function (cdf) F, which is a mixture of K component Gaussian distributions, in which each component is distributed as $\boldsymbol{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ with weight τ_j, $j = 1, ..., K$. Let $\xi_j = (\boldsymbol{x} - \boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_j)$, and G be cdf of $\chi^2_{(p)}$. We then define the outlyingness function to be

$$H(\boldsymbol{x}) = 1 - \sum_{j=1}^{K} \tau_j(1 - G(\xi_j)). \tag{7.1}$$

For a given $\epsilon \in (0, 1)$, if $1 - G(\xi_j) < \epsilon$, then $\boldsymbol{x}$ is considered to be an outlier to the jth component. For the case of more than one components, if $\sum_{j=1}^{K} \tau_j(1-G(\boldsymbol{\xi}_j)) < \epsilon$, $\boldsymbol{x}$ is considered being an outlier to the mixture of Gaussian model. If $\boldsymbol{X}$ is assumed from mixture of Gaussian distribution, the probability of classifying a random sample point $\boldsymbol{x}$ as an outlier would be ϵ. From the frequentist point of view, the theoretical Type-I error (probability of misclassifying a normal point as an outlier) would be less than ϵ.

Be aware that $H(\boldsymbol{x}) \in (0, 1)$ is not a cumulative distribution function. $H(\boldsymbol{x})$ measures the outlyingness of $\boldsymbol{x}$ to the underlying mixture Gaussian model. A potential outlier has a large value of $H(\boldsymbol{x})$ close to 1. For any fixed threshold ϵ, point $\boldsymbol{x}$ is categorized as an outlier if $H(\boldsymbol{x}) > 1 - \epsilon$, normal, otherwise.

In order to evaluate performance of an outlier identifier, the sensitivity and specificity are need to be calculated. They are related to two types of errors, which are called the false negative FN (type-II error or masking effect) and false positive FP (type-I error or swamping effect). In the context of outlier detection problem, if a point is a true outlier, we usually call it condition positive. In the contrary, if the point belongs to the normal group, we call it condition negative. Specifically, the probability of type-II error, P_{err2}, and type-I error, P_{err1}, can be formulated as

$$\begin{aligned} P_{err1} &= \text{Prob}(\text{predicted as outlier} \mid \text{data is normal}), \\ P_{err2} &= \text{Prob}(\text{predicted as normal} \mid \text{data is outlier}). \end{aligned} \tag{7.2}$$

In addition, P_{err1} can be estimated by the false positive rate (FPR) and P_{err2} can be estimated by the false negative rate (FNR) in the sample.

Relations to the sensitivity and specificity are showed as follows.

$$specificity = 1 - \hat{P}_{err1} = 1 - FPR,$$
$$sensitivity = 1 - \hat{P}_{err2} = 1 - FNR.$$

A perfect outlier detection method would have 100% sensitivity (i.e. classify all the outlying points as outlier) and 100% specificity(i.e. classify all the normal points as normal). However, any detection mechanism will possess a minimum error bound known as the Bayes error rate, see Fukunaga (1990). By our definition of the outlier in mixture of Gaussian model with $H(\boldsymbol{x})$, ϵ is a tuning parameter that determines the region of potential outliers. It thus controls the underlying P_{err1} and P_{err2}. If ϵ is small, the region of potential outliers is small. It is more likely to commit the type-II error and therefore have a large value of P_{err2} but a small value of P_{err1}. On the opposite, if ϵ is large, the region of potential outlier is large. Then, it is more likely to have a large value of P_{err1} but small value of P_{err2}. Since the purpose of outlier detection is to take more serious attention to potential outlying observations, in usual, ϵ is chosen to control the type-II error, the false negative to be small. For example, if one intends to maintain the upper bound of theoretical $P_{err2} \leq 0.05$, we can simply set $\epsilon = 0.05$.

In order to give a better understanding of different terminologies that are commonly used in statistical learning society. The classification table is showed in Table 7.2.1.

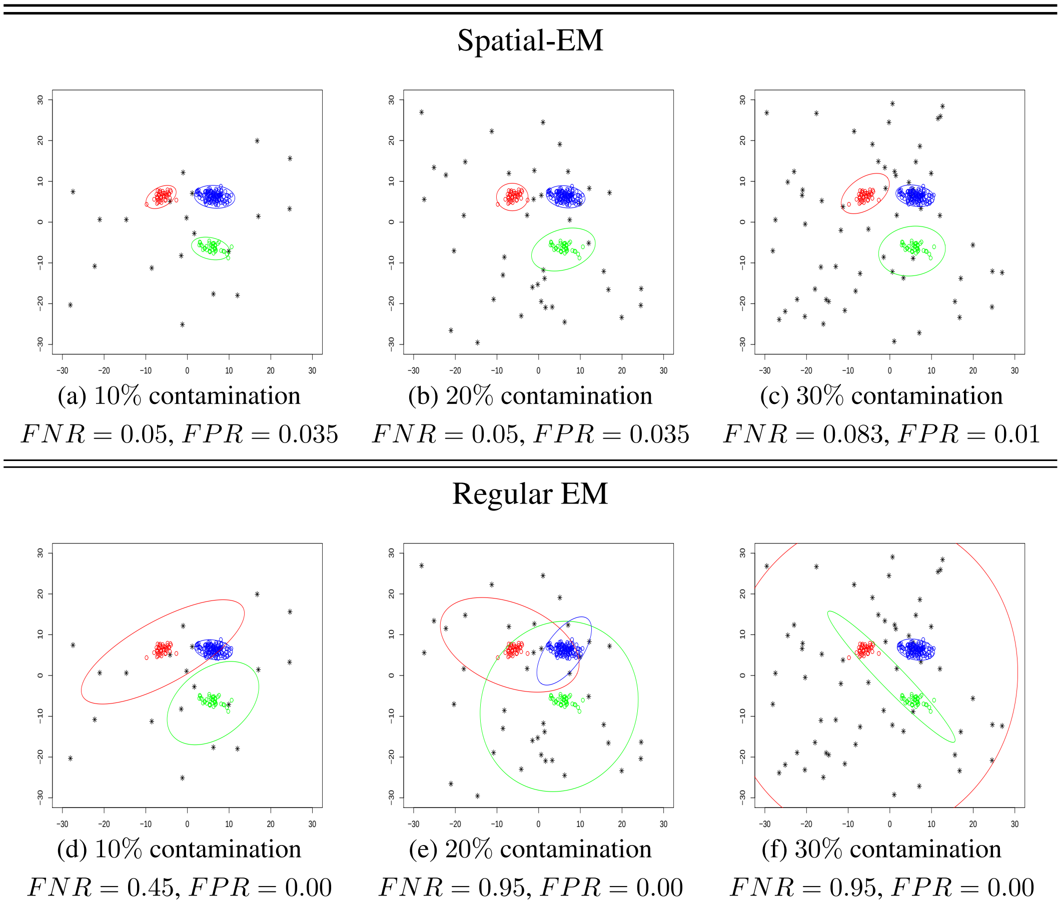
		Condition Positive	Condition Negative	
Predict Outcome	Predict Outcome Positive	**True Positive**	**False Positive** (Type I error)	Positive predictive value= # of True Positive / # of Predict Outcome Positive
	Predict Outcome Negative	**False Negative** (Type II error)	**True Negative**	Negative predictive value = # of True Negative / # of Predict Outcome Negative
		Sensitivity= # of True Positive / # of Condition Positive	Specificity = # of True Negative / # of Condition Negative	

Table 7.1. Classification Terminologies

7.2.2 Synthetic Data

Figure 7.1 illustrates the performance of Spatial-EM and regular EM on a synthetic data set of 3 separable Gaussian components. They are generated from $\boldsymbol{N}((-6,6)^T, [(2,.5)^T, (.5,1)^T])$, $\boldsymbol{N}((6,-6)^T, [(3,-.5)^T, (-.5,1)^T])$, and $\boldsymbol{N}((6,6)^T, (4,-.3)^T, (-.3,1)^T)$, size 40, 40 and 120 respectively. In increasing proportions 10%, 20% and 30% of contamination uniformly distributed on span $[-30,30] \times [-30,30]$ are added to the data.

Here, we set the $\epsilon = 0.05$, which means that the theoretical probability of type-II error is controlled to be less than 5%. It is shown in the plots that with the proportion of contamination increases, the observed type-II error rates (FNR) of both methods increase. However, regular EM algorithm can not maintain the level of type-II error. Even under 10% contamination, FNR of the regular EM increase to 0.45. The Spatial-EM algorithm, however, well control the FNR as 0.05 for contamination proportion up to 20%. The FNR of Spatial-EM only slightly increase to 0.08 when contamination level reaches 30%. This example demonstrates the unreliability of regular EM to outliers and robustness of Spatial-EM to outliers.



Spatial-EM

(a) 10% contamination $FNR = 0.05$, $FPR = 0.035$

(b) 20% contamination $FNR = 0.05$, $FPR = 0.035$

(c) 30% contamination $FNR = 0.083$, $FPR = 0.01$

Regular EM

(d) 10% contamination $FNR = 0.45$, $FPR = 0.00$

(e) 20% contamination $FNR = 0.95$, $FPR = 0.00$

(f) 30% contamination $FNR = 0.95$, $FPR = 0.00$

Figure 7.1. Comparison between Spatial-EM and regular EM on Mixture of 3-component Gaussian distributions

Mixture of 3-components Gaussian distributed data are represented by o's, with contamination points * uniformly distributed in the plot box. $\epsilon = 0.05$ is used to define a potential outlier. Elliptic curves are 95% contours of estimated covariance matrices. So, points are deemed to be normal if they are inside the contours, otherwise, be outliers. With the same initial conditions of EM Algorithm, Spatial-EM exhibits highly robustness capability against outliers, while regular EM Algorithm fails to predict the region of the normal data.

7.2.3 Taxonomic Research on Fish Species Novelty Discovery

In the field of Biological Taxonomy, an individual is different from others in one way or another. Scientists would put those who are similar into a cluster and define them as a species. If an individual is significantly different from the existing species, one might consider it to be a new species. Therefore, novelty species discovery problem on taxonomic study can be viewed as an outlier detection problem. Base on type of input data, labeled or unlabeled, outlier detection problem can be categorized into supervised or unsupervised learning algorithm Chandola *et al.* (2007).

As mentioned in Section 7.1. We can use the EM algorithm to construct an supervised outlier detector. It takes only the normal instances in the training phase and models them as a mixture of Gaussian distributions. In the detection phase, one can define those data points in the low density regions of the model as outliers or novelty species. The $H(\boldsymbol{x})$ we define in (7.1) can be used for this purpose.

Roberts (1999) and Yamanishi *et al.* (2004) applied a similar approach as this learning scheme in their application of outlier detection problems. They defined *extreme values* w.r.t. the normal (non-outlier) mixture models as outliers. Following in the same manner, the performance of Spatial EM algorithm is evaluated by being applied to an experiment on the real data below. We would also give a comparison with regular EM algorithm, the kernelized spatial depth (KSD) from Chen *et al.* (2009), and single Gaussian model at the end of this section.

Data set and training scheme

This data set consists of 989 specimens from Tulane University Museum of Natural History (TUMNH). There are 10 species that include 128 *Carpiodes carpio*, 297 *Carpiodes cyprinus*, 172 *Carpiodes velifer*, 42 *Hypenteilum nigricans*, 36 *Pantosteus discobolus*, 53 *Campostoma olibolepis*, 39 *Cyprinus carpio*, 60 *Hybopsis storeriana*, 76 *Notropis petersoni*, and 86 *Luxilus zonatus*. For each species, 12 features are generated by using 15 landmarks, which are those biologically definable points along the body outline, see Chen *et al.* (2009) for a detailed description.

The picture of these 10 species is shown in the table **??** at the end of Section 7.2.3. In order to unify the measurement on each feature, we standardize each feature by subtracting the corresponding mean and dividing the standard deviation.

In this experiment, we treated one of the 10 species as a “undiscovered” species and the other 9 species as known. Our experiment is to model those 9 known species as a mixture of Gaussian model. Hopefully, the “undiscovered” fish is so different from the known species and thus can be considered as an outlier or novelty species. This concept of discovering new species is also very common in taxonomy study.

For instance, to do experiment on the species Carpiodes carpio, we use the data set consisted of the other 9 species, which have 861 observations with 12 features, to construct a mixture of Gaussian model. In Table 7.2 (a), the box-plot indicates that a considerable number of outliers exist. In addition, it looks like the model can not be explained by just one single multivariate Gaussian component. We therefore need to find out the number of mixture components w.r.t. the data cloud.

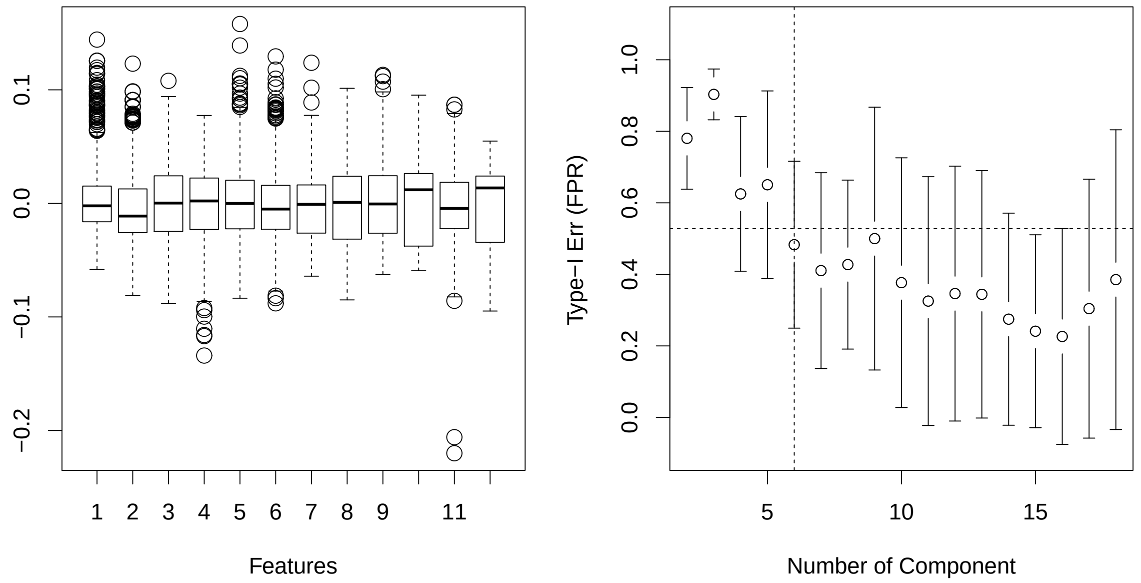
One can use various off-the-shelf methods to guess the number of components in mixture Gaussian model. It is exactly the same problem as issue (1) that we described in Section 5.3. However, those methods for choosing number of components do not take outlying effect into consideration in general. We have no idea which one is better fit into the robust condition we are thinking about. So a simple method is proposed next.

We would like to employ a more heuristic method to predict the number of components by multiple uses of spatial-EM (or regular EM). It is a similar idea as 10-fold cross-validation. To be more specific, first, we call the 9 species fish data as the training sample. The 1 species of fish being hold out for evaluation is called the test set. For each value of k, the number of components between 1 to K, the training sample is randomly split even into 10 folds. 9/10 of the training sample used to build the mixture model is called the training set. The other 1/10 of the training sample used for validate the model is called the validation set. For each different fold, the spatial-EM algorithm is run on the training set. So, we have 10 runs in total. After every run, use the H

function, (7.1), with $\epsilon = 0.05$ to determine the outlying observation on the validation set. Since, by assumption, the fishes in training sample are all considered as normal instances. $\hat{P}_{err1}$, the FPR on validation set, can be calculated for every single run. Based on these 10 values of $\hat{P}_{err1}$, the mean and standard deviation are reported. Hastie *et al.* (2009), page 216, proposed a rule called "one-standard-error" to select the number of parameters for regression model base on MSE. We find it similar to apply for selecting the number of component of mixture model. we choose the most parsimonious model whose mean $\hat{P}_{err1}$ is no more than on one standard error (deviation) above the mean $\hat{P}_{err1}$ of the best model. As shown in Table 7.2 (b), the best model that yields the smallest mean $\hat{P}_{err1}$ has 16 components, and the 6-component model is the one we choose eventually.

After the component size is determined, we use the whole training sample to establish the mixture model by both version of EM algorithms. In order to overcome the randomness of the initial location parameters for starting the spatial-EM (or regular EM) algorithm, we repeat the training process 20 times with different initial parameters and report the mean and standard deviations of $\hat{P}_{err1}$ (FPR) and $\hat{P}_{err2}$ (FNR) for each fish species in Table 7.3.

Remember in the way of detecting an outlier, ϵ can be used to seek the balance between P_{err1} and P_{err2}. In order to compare different outlier detection methods under a same scenario, ϵ is then chosen to satisfy $\hat{P}_{err1} \approx \hat{P}_{err2}$. Chen *et al.* (2009) did the similar experiment on these data set using the kernalized spatial depth (KSD). The basic idea of KSD is to evaluate the spatial depth in feature space induced by a positive definite kernel. A point with KSD below a threshold is claimed as an outlier. The FPR was reported in his paper when it equals FNR.



(a) Box-Plot (b) One-Standard-Error Rule

Table 7.2. (a)Box-Plot of fish species sets 2 ∼10; (b)One-Standard-Error rule for choosing number of components

(a) is the box-plot of 12 features of the fish data with 9 species without the Carpiodes carpio. Features number 1, 2, 5, 6, are obviously skewed distributed. So one Gaussian distribution is not sufficient to model the data. (b) 68% C.I.'s of the P_{err1} are plotted vs the number of components. One-Standard-Error rule choose the most parsimony model (the one with the smallest number of components) whose mean of $\hat{P}_{err1}$ is no more than one standard deviation above the one of the best model. Here the chosen component size is 6.

Results

After the mixture model is finally estimated. We can evaluate the model performance by showing the $\hat{P}_{err1}$ and $\hat{P}_{err2}$ based on the training sample and test set. Note that we would use all of the 989 observations and ground truth of species labels to calculate the error rates. For example, when the training sample are data from species sets 2-10, the false positive error occurs if the fish from the training sample is classified as a novelty species (outlier), and the false negative error occurs if the fish from the species set 1 is identified as a known species (normal) that belongs to the species in training sample. FPR ($\hat{P}_{err1}$) and FNR ($\hat{P}_{err2}$) can be calculated by the formulas in Table 7.2.1.

Table 7.3 illustrates the comparison among Spatial-EM and regular EM under assumption of the mixture of Gaussian model, KSD (Chen *et al.*, 2009), and simple outlier detector based on robust Mahalanobis distance, MCD covariance matrix, see the introduction from Section 1.4, is used to estimate the covariance matrix. The mean value of FNR ($\hat{P}_{err2}$) from 20 runs with different initial parameters are reported with the standard deviations inside parenthesis. Here, the mean FNR is shown when it equals to the mean FPR ($\hat{P}_{err1}$). The numbers of components that are determined by one-standard-error rule are also reported inside the brackets.

The spatial-EM detects most of the "unknown" species as outliers with high sensitivity and specificity. For instance, the sensitivity of *Carpiodes carpio* is 0.740, and the risk for making type-II error is 0.260; the sensitivity of *Carpiodes cyprinus* is 0.819, and the risk for making type-II error is 0.181, etc. Most interestingly, the results of Spatial-EM outperform KSD in all the species except the first species, *Carpiodes carpio*. Outlier detection based on robust Mahalanobis distance with the assumption of single Gaussian distribution performs the worst in here. This is also the reason why we want to provide a robust parameter estimation process (Spatial-EM) for mixture model in our paper.

Comparing the result between regular EM and KSD, we notice there are 6 out of 10 species with lower $\hat{P}_{err2}$ (higher sensitivity) than KSD. The main reason EM beating KSD in here is due to the flexibility of the mixture Gaussian model. KSD is a nonparametric technique, which usually has lass assumptions. As pointed by out by Chen *et al.* (2009), the KSD is weak to deal with the "masking" effect. It refers to the case that the "unknown" species in the middle is surrounded by "known" species. However, by using mixture of Gaussian model, such masking effect can be well explained by adding more components that are surround ringlike distributed. Masking effect seems to appear seriously in the last 3 fish species *Hybopsis storeriana*, *Notropis petersoni*, and *Luxilus zonatus*. The KSD can hardly detects the "new" species, hence high FNR's are reported. However, mixture model assumption can moderately alleviate this problem.

Spatial-EM has better performance than the regular EM. It outperforms the regular EM in 6 out of 10 species in terms of the mean value of $\hat{P}_{err2}$ (FNR). Moreover, it is impractical to use regular EM for detecting outlier in general, because its high variance of FNR compare to spatial-EM. Large standard deviation indicates that the prediction is not stable. With the random initial parameters for EM algorithm, it means that there are large portion of fish being detected as novelty in one run but known species in another run. For example, both EM algorithms has the similar FNR around 0.35 for detecting the novelty of *Luxilus zonatus*. The estimated number of Gaussian components by Spatial-EM is also similar to the regular EM, (i.e. 6 and 5). However, just because the randomness of initial location inputs for the algorithms, the standard deviation produced by regular EM is 0.427, about 5 times more than Spatial-EM, which is 0.086. Theoretically, we need to collect a lot more observations with "normal" data in training sample to have a reliable detector based on the regular EM. It is believed that the potential outliers in data cause us to choose more number of components in mixture model when we train data by the regular EM. So, the spatial-EM is much more statistical efficient and reliable to be used in the outlier detection problem. Also, Spatial-EM tends to choose simpler model than the regular EM. The number of components selected by the method described in early section for the Spatial-EM is smaller than the one for the regular EM. Usually, too complicated models overfit data with poor generalization performance. This also explain the large variance of the regular EM.

Unknown Species	Spatial-EM	regular EM	KSD	Single Gaussian
Carpiodes carpio	0.260 [6](0.04)	0.303 [9](0.289)	0.234	0.408
Carpiodes cyprinus	0.181 [8](0.114)	0.212 [11](0.230)	0.209	0.245
Carpiodes velifer	0.11 [5](0.009)	0.095 [9](0.131)	0.180	0.144
Hypentelium nigricans	0.007 [5](0.011)	0.006 [11](0.011)	0.071	0.0538
Pantosteus discobolus	0.042 [5](0.065)	0.083 [9](0.091)	0.056	0.091
Campostoma oligolepis	0.151 [8](0.065)	0.138 [12](0.289)	0.208	0.385
Cyprinus carpio	0.001 [7](0.001)	0.019 [12](0.034)	0.051	0.0473
Hybopsis storeriana	0.294 [7](0.033)	0.371 [14](0.403)	0.367	0.320
Notropis petersoni	0.318 [7](0.154)	0.181 [10](0.159)	0.487	0.355
Luxilus zonatus	0.324 [6](0.086)	0.388 [5](0.427)	0.512	0.460

Table 7.3. Results of Fish Species Novelty Discovery

Experiment on Spatial-EM and regular EM are repeated 20 times with the respective number of components and random input of locations $\boldsymbol{\mu}_j$'s. Contents are the $\hat{P}_{err2}$ (FNR) based on the test set which is equal to $\hat{P}_{err1}$ (FPR) based on the training sample. "[m]" means the number of Gaussian components in the mixture model chosen by one-standard-error rule. "(n)" means the standard deviation of $\hat{P}_{err2}$. For instance, first entry 0.260 reflects the mean of probabilities of misclassifying Carpiodes carpio as an existing specimen, when using Spatial EM. FPR that misclassifying the existing specimen as a novelty is also set to equal 0.260 by using Spatial EM. The values associate with KSD are presented here, see Chen *et al.* (2009)

7.3 Clustering

7.3.1 Clustering Methods and Confusion Matrix

Clustering is to partition a set of objects into groups (clusters) so that the objects in the same cluster are more similar to each others in some sense than those in the other groups. During the process of modeling, no group label information is used. So, clustering is considered as the unsupervised learning in general. The notion of clustering depends on a particular objective. There are different ways to make clustering possible. Here are a few typical examples of cluster models:

- Connectivity model: hierarchical clustering based on the distance connectivity.
- Centroid model: K-means algorithm, each cluster represented by a single mean vector.
- Graph-based model: a cluster is considered as a *clique*, i.e., a subset of nodes such that every two nodes in it are connected by an edge.
- Density models: clusters are connected dense regions in the data space.
- Distribution models: clusters are modeled by a mixture of distributions.

Our experiments of clustering shown the next are based on the distribution model. We assume the data set is generated by sampling from mixture of Gaussian distribution. The number of components (clusters) are given according to the nature of problems. For any data point, which cluster it should belong to can be estimated by the posterior probability $P(Y_{j,i} = 1|\boldsymbol{\theta}, \boldsymbol{x}_i)$, which is denoted by $T_{j,i}$ in Section 5.2, equation (5.4). Specifically, by the Bayes theorem,

$$P(Y_{j,i} = 1|\boldsymbol{\theta}, \boldsymbol{x}_i) = \frac{\tau_j f_j(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\sum_{j=1}^{K} \tau_j f_j(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.$$

The clustering rule is to assign the point $\boldsymbol{x}_i$ into the cluster with the highest posterior probability, that is, $\arg\max_j T_{j,i}$. In practice, since we assume the mixture of Gaussian model, $f(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ is monotonically decreasing w.r.t. the square value of Mahalanobis distance $\xi_{j,i} = (\boldsymbol{x}_i - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1} (\boldsymbol{x}_i -$

$\boldsymbol{\mu}_j)$. In order to save computational time on clustering, we can assign the point to the jth cluster which associates with the smallest value of $\xi_{j,i}$.

After the clustering result is produced, we can evaluate the clustering method based on the sample of data with the known class label information. The data we choose for experiments are the sets of data consisting of only the pre-classified items. As an unsupervised learning, we just do not use the class label when we train the model. The *benchmark set* used in our experiment is the same training sample with class labels. It can be thought of as a gold standard for evaluation. The contingence table based on the benchmark set with predicted cluster ID and true class labels can be used to recognize the association between them. This is well explained in the work flow shown as the Figure 7.2. Finally, the confusion matrix (matching matrix) can be calculated to show how different a cluster is from the gold standard (true) cluster.

An example of confusion matrix is show in Table 7.4. The predicted class c matches the actual class C with 11 correct predictions but 2 mistakes. The predicted class a matching the true class A yields 5 correct predictions but 3 mistakes. The predicted class b associated with actual class B has 6 correct predictions but 5 mistakes.

In the ideal case, the non-diagonal entries of the matrix is equal to 0, which means no misclassification is done by the clustering model. However, this seldom happens. Base on the confusion matrix, the FPR and FNR of a given class can be calculated with different clustering algorithms for comparison.

		Actual class		
		A	B	C
Predict class	A	**5**	3	0
	B	4	**6**	1
	C	0	2	**11**

Table 7.4. Confusion Matrix

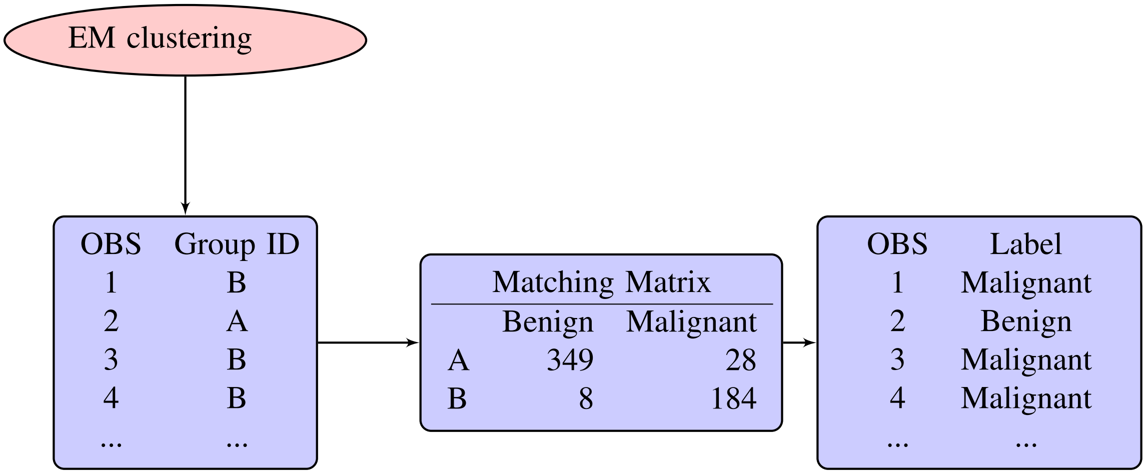
7.3.2 UCI Wisconsin Diagnostic Breast Cancer Data

This is the Breast Cancer Wisconsin (Diagnostic) Data Set. The data set can be downloaded from `http://archive.ics.uci.edu/ml/datasets`. This two-group (benign and malignant) data set is used for breast cancer diagnosis analysis (Mangasarian *et al.*, 1995). There are 569 observations. No information about the known benign and malignant label are used when training the clustering model. With the same setup as the experiment done by Fraley & Raftery (2002), we use two features *mean texture* and *extreme area*. The scatter plot of these two features in Figure 7.3 shows a considerable overlap between benign and malignant patients. Our goal is to partition patients into two groups, denoted by group A and group B, without using the known benign or malignant label information of each patient. Every patient in the sample is clustered either to group A or B based on the highest posterior probability. In order to evaluate the model performance, after each data point is clustered into two groups, we "match" the cluster ID by comparing to the known label information (benign or malignant) from the same training sample to get the highest contingency, shown in the Figure 7.2.

In health care application, the malignant patient should get the most attention in the clinical practice. We define the malignant as positive effect and benign as negative effect as usual. According to Table 7.2.1, the probabilities of false positive and false negative and their estimated probabilities would be redefined as follows,

$$P_{err1} = \text{Prob}(\text{predicted as malignant} \mid \text{patient is benign}),$$
$$P_{err2} = \text{Prob}(\text{predicted as benign} \mid \text{patient is malignant}).$$

All the observations form the data set are used to construct the mixture model. The FPR ($\hat{P}_{err1}$) and FNR ($\hat{P}_{err2}$) are calculated upon the same set of sample. The performance is compared between regular and Spatial EM algorithm, see the Figure 7.3. The resulting spatial-EM has the FNR= 0.1320 slightly smaller than the one of the regular EM. Moreover the FPR= 0.0224 of Spatial EM is just around $1/3$ of that of regular EM. Therefore, in this example, the Spatial-EM totally beats up the regular EM in terms of sensitivity and specificity. In fact, a medical screen test that maintains the same level of FNR and much smaller FPR can prevent patients from spending more time and money on the follow-up diagnostic procedure.



OBS	Group ID
1	B
2	A
3	B
4	B
...	...

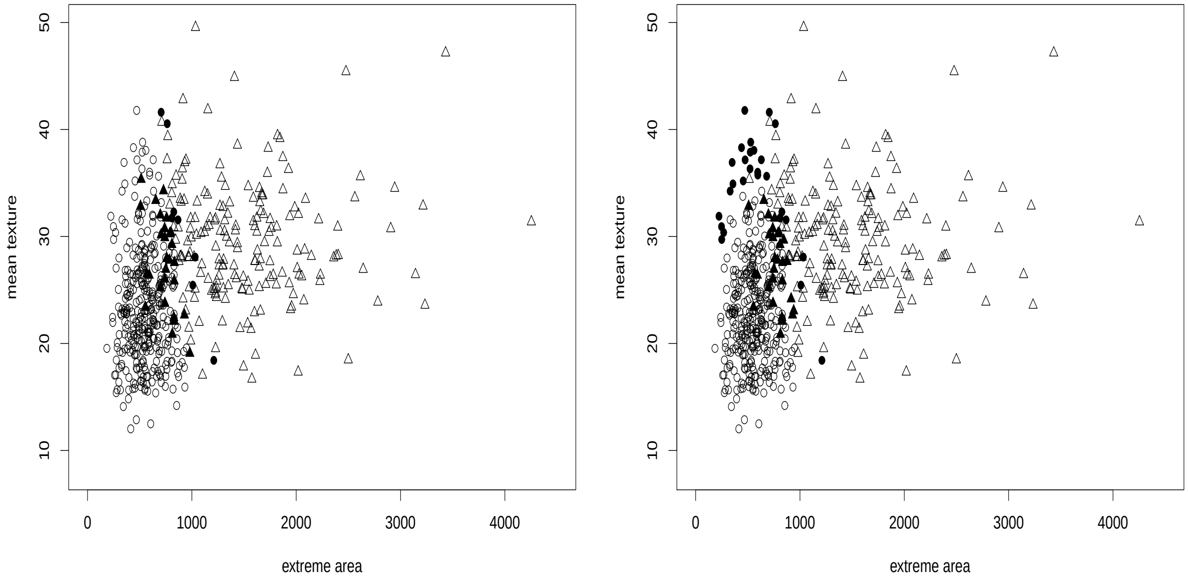
Matching Matrix

	Benign	Malignant
A	349	28
B	8	184

OBS	Label
1	Malignant
2	Benign
3	Malignant
4	Malignant
...	...

Figure 7.2. Work flow of using EM algorithm on clustering

This is the work flow by using the spatial-EM algorithm with mixture of Gaussian model to identify whether or not a potential breast cancer patient is predicted as malignant. By measuring the Mahalanobis distance w.r.t. the cluster, the observation (OBS) is assigned into group ID: A or B. To match the group ID with the true label, the matching matrix is generated. Most of the benign patients are clustered into group A, and most of the malignant patients are clustered into group B. So A=Benign, B=malignant



(a) Spatial-EM
$\hat{P}_{err1} = 0.0224$, $\hat{P}_{err2} = 0.1320$

(b) Regular EM
$\hat{P}_{err1} = 0.0756$, $\hat{P}_{err2} = 0.1368$

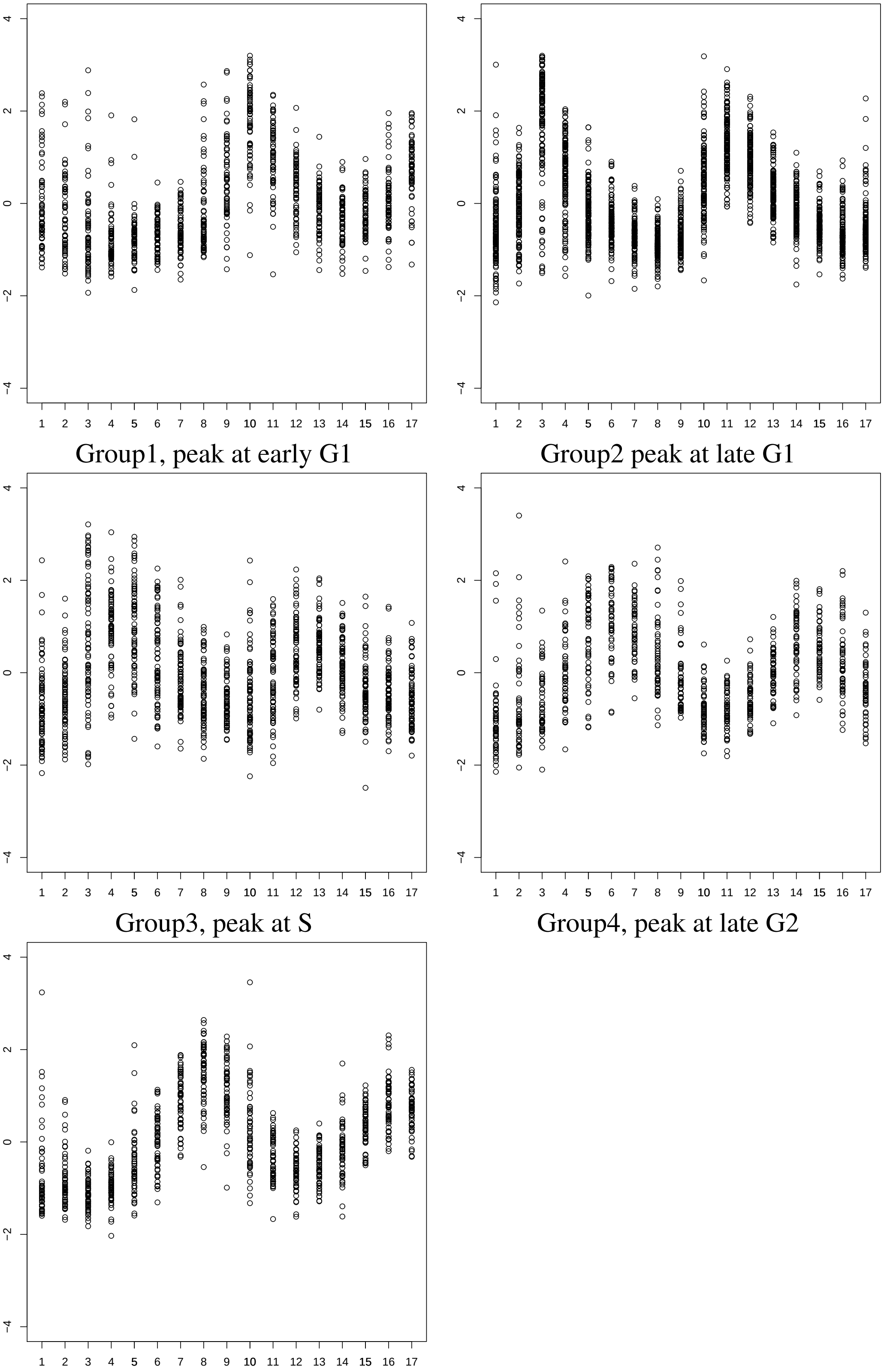
Figure 7.3. Results of Clustering on UCI Wisconsin Diagnostic Breast Cancer Data

A Projection of the UCI Wisconsin Diagnostic Breast Cancer Data on feature mean texture v.s. extreme area. o and $\triangle$ represent a patient being benign and malignant respectively. Closed symbols represent misclassifications. $\hat{P}_{err1}$ denotes FPR as the estimated probability of an underlying malignant patient being diagnosed as benign. $\hat{P}_{err2}$ denotes the FNR as the estimated probability of a benign person being diagnosed as a malignant patient.

7.3.3 Yeast Cell Cycle Data

The data set was used for experiment in Zhang & Cheung (2006), who proposed the X-EM algorithm to automatically choose the number of components and estimate the parameters simultaneously. It is mentioned in the Section 5.3 as a method that intends to overcome one of limitations of regular EM algorithm. The data set is microarray data of the yeast cell cycle. It can be downloaded from `http://www.cs.washington.edu/homes/kayee/model`. The yeast cell cycle data (Cho *et al.*, 1998) showed fluctuation of expression levels of genes over two cycles (17 time points). Zhang & Cheung (2006) used a subset of data consisting of 384 genes, whose expression levels peak at different time points corresponding to the five phases of cell cycles, see the Table 7.5. Group 1 has 67 genes whose expression level reach peak at early G1. Group 2 has 135 genes whose expression level peak at late G1. Group 3 has 75 genes whose expression level reach peak at S. Group 4 has 52 genes whose expression level reach peak at late G2. Group 5 has 55 genes whose expression level peak at M. X-EM successfully choose the number of components to be 5. Gene clusters by X-EM showed some grouping patterns of the cell cycle phases. For the given number of 5 Gaussian components, they compared the performance of their algorithm with regular EM (Reg EM) as well as the other two supervised classification methods: the supervised cluster analysis (SCA) by Qu & Xu (2004) and the support vector machines (SVM) by Brown *et al.* (2000). Here, we run the spatial-EM with 5-component mixture of Gaussian model on the same data set and compare our results to all the methods conducted in Zhang & Cheung (2006).

We consider the positive effect occur if a gene belongs to the given cell division phase. The model performance are measured based on four indices: false positive (FP), false negative (FN), true positive (TP), true negative (TN), see Table 7.6. The total error defined as FP+FN is shown in Table 7.7 . It is showing that the Spatial-EM outperforms all of those 4 methods in terms of the total error. The regular EM has high FPR and FNR. It is interesting to see that even the two supervised learning methods that use the label information can not beat spatial-EM. It can be seen that the X-EM has a relative high FNR, but spatial-EM well balances both FP and FN. That is probably due to the robustness of spatial-EM making the parameter estimation much more accurate.



Group1, peak at early G1

Group2 peak at late G1

Group3, peak at S

Group4, peak at late G2

Group5, peak at M

Table 7.5. The genes expression levels of the five classes in the Yeast Cell Cycle Data. 384 genes are plot with 17 time point measurements.

Cell division phase	Mehtods	FP	FN	TP	TN
Early G1 (67 genes)	**Spatial-EM**	**20**	**17**	**50**	**297**
	X-EM	11	24	43	306
	Reg EM	50	12	55	267
	SCA	21	21	46	296
	SVM	38	10	57	279
Late G1 (135 genes)	**Spatial-EM**	**32**	**18**	**117**	**217**
	X-EM	13	54	81	236
	Reg EM	28	40	95	221
	SCA	24	35	100	225
	SVM	43	10	125	206
S (75 genes)	**Spatial-EM**	**13**	**42**	**33**	**296**
	X-EM	10	47	28	299
	Reg EM	33	49	26	276
	SCA	37	36	39	272
	SVM	72	18	57	237
G2 (52 genes)	**Spatial-EM**	**17**	**17**	**35**	**315**
	X-EM	13	22	30	319
	Reg EM	28	41	11	304
	SCA	18	29	23	314
	SVM	46	5	47	286
M (55 genes)	**Spatial-EM**	**19**	**7**	**48**	**310**
	X-EM	12	26	29	317
	Reg EM	38	42	13	291
	SCA	19	8	47	310
	SVM	47	2	53	282

Table 7.6. Performance of five methods of Yeast Cell Cycle Microarray Data

Performance of five methods of Yeast Cell Cycle Microarray Data.The model performance are measured based on four indices: false positive (FP), false negative (FN), true positive (TP) and true negative (TN).

Methods	FP	FN	FP+FN
Spatial-EM	**101**	**101**	**202**
X-EM	59	173	232
Reg EM	177	184	361
SCA	119	129	248
SVM	246	45	291

Table 7.7. Comparison of total error rates of the five methods on Yeast Cell Cycle Microarray Data

Comparison of the total error rates of the five methods on Yeast Cell Cycle Microarray Data. Spatial-EM outperforms the other methods in terms of total error.

Chapter 8

Concluding Remarks and Future Work

A series of robust parameter estimation procedures for single distribution and the mixture of distributions were developed in this work. The previous chapters presented the theory and implementation of these methods. All the algorithms are coded in R/S+ and C language. The contribution of this work to the robust statistics can be summarized as two parts from this dissertation.

In the first half of this dissertation, we studied the robustness properties of the modified spatial rank covariance matrix (MRCM) proposed by Visuri *et al.* (2000). It is infinitesimal robust in terms of the influence function and quantitative robust in terms of finite sample breakdown point. We derived the influence functions for the eigenvector and eigenvalues of the MRCM, then the influence function for the MRCM. They are bounded under the assumption that the scatter parameter has distinct eigenvalues. The breakdown point attains the upper bound by the choice of robust univariate scale functional $\sigma = \mathrm{MAD}_k$ with some optimal values for k. Comparing with other high breakdown point estimators such as the MCD, the S-estimators and the projection based estimators, our MRCM is easy to compute with the complexity $O(n^2 + p^3)$. Even for large data set in high dimensions, using MRCM is still practical. Also, MRCM is highly statistical efficient under unimodal Gaussian distribution and other heavy-tailed distributions. It is also shown, under elliptical symmetric distribution, MRCM is affine equivariant and proportional to the scatter parameter.

In the second half, in order to make use of the MRCM in a broader variety of distributions, we extend the notion of MRCM from unimodal elliptical distributions to the multimode mixture of elliptical distributions with the help of EM algorithm. In particular, we mainly focus on the mixture of Gaussian distribution. Based on the spatial rank and MRCM, we proposed a novel Spatial-EM algorithm to estimate the component location and scatter parameters. The Spatial-EM is shown to be highly robust in parameter estimation with the contaminated sample versus the regular EM. Moreover, form the likelihood point of view, the estimates of Spatial-EM have similar form as those done by the regular EM under a mixture of heavy-tailed distribution (Kotz type distribution). But Spatial-EM is more practical to implement regarding to the computation issue.

Most importantly, by inventing a brand new robust parameter estimation process for mixture models, the problem of outlier effects that hinder the data analysis from using statistical learning on noisy data set is well resolved. Learning schemes including outlier detection and clustering by employing the Spatial-EM are also illustrated in this dissertation. It has better results in terms of sensitivity and specificity comparing to some other existing supervised or unsupervised learning techniques in the experiments we have done.

Future Work

In this dissertation, we modified the spatial rank covariance matrix by taking MAD of the projection data on the eigenvectors. As pointed out earlier, this was a method mentioned in Visuri *et al.* (2000). However, despite the loss of the affine equivariance, RCM is approximately proportional to the scatter parameter. Therefore, the true scatter parameter of an elliptical distribution may be estimated by a constant $\hat{\lambda}$ times RCM, where the $\hat{\lambda}$ is an estimate of the Wilks generalized variance described in Section 2.2. If this is reasonable, the computational time on calculating this version of MRCM can be reduced. In addition, the concept is easy to be conducted on mixture model by the same way as Spatial-EM, and therefore decrease the computation complexity in total. It is believed that some other modified RCM versions are waiting to develop. They want to maintain the same level of estimation accuracy but make a faster computation possible.

The other future work we are interested is the development of a systematic way in choosing number of component with data set contains outliers. We have done some of simulation work on using AIC or BIC with Spatial-EM to choose an optimal number of component in mixture model. It is worthy to derive some more theoretical results and do more experiments based on Spatial-EM.

Bibliography

BANFIELD, J. D. & RAFTERY, A. E. (1993) "Model-based Gaussian and non-Gaussian clustering." *Biometrics*, Vol. 49(3), pp. 803–821.

BENSMAIL, H. & CELEUX, G. (1996) "Regularized Gaussian discriminant analysis through eigenvalue decomposition." *J. Amer. Statist. Assoc.*, Vol. 91(436), pp. 1743–1748.

BEZDEK, J. C. (1981) *Pattern recognition with fuzzy objective function algorithms*. Plenum Press, New York With a foreword by L. A. Zadeh, Advanced Applications in Pattern Recognition.

BROWN, M. P. S.; GRUNDY, W. N.; LIN, D.; CRISTIANINI, N.; SUGNET, C. W.; FUREY, T. S.; ARES, M.; & HAUSSLER, D. (2000) "Knowledge-based analysis of microarray gene expression data by using support vector machines." *Proceedings of the National Academy of Sciences*, Vol. 97(1), pp. 262–267.

CHANDOLA, V.; BANERJEE, A.; & KUMAR, V. (2007) "Outlier detection: a survey." University of Minnesota, Technical report.

CHAUDHURI, P. (1996) "On a geometric notion of quantiles for multivariate data." *J. Amer. Statist. Assoc.*, Vol. 91(434), pp. 862–872.

CHEN, Y.; DANG, X.; PENG, H.; & JR., H. L. B. (2009) "Outlier Detection with the Kernelized Spatial Depth Function." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 31, pp. 288–305.

CHO, R. J.; CAMPBELL, M. J.; WINZELER, E. A.; STEINMETZ, L.; CONWAY, A.; WODICKA, L.; WOLFSBERG, T. G.; GABRIELIAN, A. E.; LANDSMAN, D.; LOCKHART, D. J.; & DAVIS, R. W. (1998) "A genome-wide transcriptional analysis of the mitotic cell cycle.." *Molecular cell*, Vol. 2(1), pp. 65–73.

CROUX, C.; DEHON, C.; & YADINE, A. (2010) "The k-step spatial sign covariance matrix." *Adv. Data Anal. Classif.*, Vol. 4(2-3), pp. 137–150.

CROUX, C. & HAESBROECK, G. (1999) "Influence function and efficiency of the minimum covariance determinant scatter matrix estimator." *J. Multivariate Anal.*, Vol. 71(2), pp. 161–190.

CROUX, C. & HAESBROECK, G. (2000) "Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies." *Biometrika*, Vol. 87(3), pp. 603–618.

DAVIES, L. (1992) "The asymptotics of Rousseeuw's minimum volume ellipsoid estimator." *Ann. Statist.*, Vol. 20(4), pp. 1828–1843.

DAVIES, P. L. (1987) "Asymptotic Behaviour of S-Estimates of Multivariate Location Parameters and Dispersion Matrices." *Annals of Statistics*, Vol. 15(3), pp. pp. 1269–129.

DAVIES, P. L. & GATHER, U. (2005) "Breakdown and groups." *Ann. Statist.*, Vol. 33(3), pp. 977–1035 With discussions and a rejoinder by the authors.

DEMPSTER, A. P.; LAIRD, N. M.; & RUBIN, D. B. (1977) "Maximum likelihood from incomplete data via the EM algorithm." *J. Roy. Statist. Soc. Ser. B*, Vol. 39(1), pp. 1–38 With discussion.

DONOHO, D. & HUBER, P. J. (1983) "The notion of breakdown point." In *A Festschrift for Erich L. Lehmann*, Wadsworth Statist./Probab. Ser., pages 157–184. Wadsworth, Belmont, CA.

DONOHO, D. L. & GASKO, M. (1992) "Breakdown properties of location estimates based on halfspace depth and projected outlyingness." *Ann. Statist.*, Vol. 20(4), pp. 1803–1827.

DÜMBGEN, L. (1998) "On Tyler's M-functional of scatter in high dimension." *Ann. Inst. Statist. Math.*, Vol. 50(3), pp. 471–491.

DÜMBGEN, L. & TYLER, D. E. (2005) "On the breakdown properties of some multivariate M-functionals." *Scand. J. Statist.*, Vol. 32(2), pp. 247–264.

Fang, K. T. & Anderson, T. W., editors (1990) *Statistical inference in elliptically contoured and related distributions*. Allerton Press, New York.

FRALEY, C. & RAFTERY, A. E. (2002) "Model-based clustering, discriminant analysis, and density estimation." *J. Amer. Statist. Assoc.*, Vol. 97(458), pp. 611–631.

FUKUNAGA, K. (1990) *Introduction to statistical pattern recognition*. Computer Science and Scientific Computing. Academic Press Inc., Boston, MA, second edition.

GATHER, U. & HILKER, T. (1997) "A note on Tyler's modification of the MAD for the Stahel-Donoho estimator." *Ann. Statist.*, Vol. 25(5), pp. 2024–2026.

GERVINI, D. (2003) "A robust and efficient adaptive reweighted estimator of multivariate location and scatter." *J. Multivariate Anal.*, Vol. 84(1), pp. 116–144.

HAMPEL, F. R.; RONCHETTI, E. M.; ROUSSEEUW, P. J.; & STAHEL, W. A. (1986) *Robust statistics*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York The approach based on influence functions.

HASTIE, T.; TIBSHIRANI, R.; & FRIEDMAN, J. (2009) *The elements of statistical learning*. Springer Series in Statistics. Springer, New York, second edition Data mining, inference, and prediction.

He, X. & Simpson, D. G. (1992) “Robust direction estimation.” *Ann. Statist.*, Vol. 20(1), pp. 351–369.

Hettmansperger, T. P.; Möttönen, J.; & Oja, H. (1998) “Affine invariant multivariate rank tests for several samples.” *Statist. Sinica*, Vol. 8(3), pp. 785–800.

Huber, P. J. (1964) “Robust estimation of a location parameter.” *Ann. Math. Statist.*, Vol. 35, pp. 73–101.

Huber, P. J. & Ronchetti, E. M. (2009) *Robust statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons Inc., Hoboken, NJ, second edition.

Jurečková, J. & Picek, J. (2006) *Robust statistical methods with R*. Chapman & Hall/CRC, Boca Raton, FL.

Koltchinskii, V. I. (1997) “M-estimation, convexity and quantiles.” *Ann. Statist.*, Vol. 25(2), pp. 435–477.

Kutner, M. H.; Nachtsheim, C. J.; Neter, J.; & Li, W. (2005) *Applied Linear Statistical Models*. McGraw Hill, 5th edition.

Lange, K. L.; Little, R. J. A.; & Taylor, J. M. G. (1989) “Robust statistical modeling using the t distribution.” *J. Amer. Statist. Assoc.*, Vol. 84(408), pp. 881–896.

Lopuhaä, H. P. (1989) “On the relation between S-estimators and M-estimators of multivariate location and covariance.” *Ann. Statist.*, Vol. 17(4), pp. 1662–1683.

Lopuhaä, H. P. & Rousseeuw, P. J. (1991) “Breakdown points of affine equivariant estimators of multivariate location and covariance matrices.” *Ann. Statist.*, Vol. 19(1), pp. 229–248.

MacQueen, J. (1967) “Some methods for classification and analysis of multivariate observations.” In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66)*, pages Vol. I: Statistics, pp. 281–297. Univ. California Press, Berkeley, Calif.

Mangasarian, O. L.; Street, W. N.; & Wolberg, W. H. (1995) “Breast cancer diagnosis and prognosis via linear programming.” *OPERATIONS RESEARCH*, Vol. 43, pp. 570–577.

Marden, J. I. (1999) “Some robust estimates of principal components.” *Statist. Probab. Lett.*, Vol. 43(4), pp. 349–359.

Maronna, R. A. (1976) “Robust M-estimators of multivariate location and scatter.” *Ann. Statist.*, Vol. 4(1), pp. 51–67.

Maronna, R. A. & Yohai, V. J. (1995) “The behavior of the Stahel-Donoho robust multivariate estimator.” *J. Amer. Statist. Assoc.*, Vol. 90(429), pp. 330–341.

Mazumder, S. & Serfling, R. (2010) “Spatial trimming, with applications to robustify sample spatial quantile and outlyingness functions, and to construct a new robust scatter estimator.” preprint.

McLachlan, G. J. & Krishnan, T. (1997) *The EM algorithm and extensions*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York A Wiley-Interscience Publication.

Medasani, S.; ; Medasani, S.; Krishnapuram, R.; & Krishnapuram, P. R. (1998) "Categorization of Image Databases for Efficient Retrieval Using Robust Mixture Decomposition.".

Medasani, S. & Krishnapuram, R. (1997) "Determination of the number of components in Gaussian mixtures using agglomerative clustering." In *Neural Networks,1997., International Conference on*, volume 3, pages 1412 –1417 vol.3. IEEE.

Oja, H. (1983) "Descriptive statistics for multivariate distributions." *Statist. Probab. Lett.*, Vol. 1(6), pp. 327–332.

Oja, H. & Randles, R. H. (2004) "Multivariate nonparametric tests." *Statist. Sci.*, Vol. 19(4), pp. 598–605.

Ollila, E.; Croux, C.; & Oja, H. (2004) "Influence function and asymptotic efficiency of the affine equivariant rank covariance matrix." *Statist. Sinica*, Vol. 14(1), pp. 297–316.

Ollila, E.; Oja, H.; & Hettmansperger, T. P. (2002) "Estimates of regression coefficients based on the sign covariance matrix." *J. R. Stat. Soc. Ser. B Stat. Methodol.*, Vol. 64(3), pp. 447–466.

Ollila, E.; Oja, H.; & Koivunen, V. (2003) "Estimates of regression coefficients based on lift rank covariance matrix." *J. Amer. Statist. Assoc.*, Vol. 98(461), pp. 90–98.

Plungpongpun, K. & Naik, D. N. (2008) "Multivariate Analysis of Variance Using a Kotz Type Distribution." In *e World Congress on Engineering 2008*, volume 2.

Qin, Y. & Priebe, C. (2012) "Maximum Lq-Likelihood Estimation via the Expectation Maximization Algorithm: A Robust Estimation of Mixture Models." *Journal of the American Statistical Association*, Vol. .

Qu, Y. & Xu, S. (2004) "Supervised cluster analysis for microarray data based on multivariate Gaussian mixture." *Bioinformatics*, Vol. 20(12), pp. 1905–1913.

Rao, C. R. (1988) "Methodology based on the L_1-norm, in statistical inference." *Sankhyā Ser. A*, Vol. 50(3), pp. 289–313.

Roberts, S. (1999) "Novelty detection using extreme value statistics." *Vision, Image and Signal Processing, IEE Proceedings -*, Vol. 146(3), pp. 124 –129.

Rocke, D. M. & Woodruff, D. L. (1993) "Computation of robust estimates of multivariate location and shape." *Statist. Neerlandica*, Vol. 47(1), pp. 27–42.

ROUSSEEUW, P. (1985) "Multivariate estimation with high breakdown point." In *Mathematical statistics and applications, Vol. B (Bad Tatzmannsdorf, 1983)*, pages 283–297. Reidel, Dordrecht.

ROUSSEEUW, P. & YOHAI, V. (1984) "Robust regression by means of S-estimators." In *Robust and nonlinear time series analysis (Heidelberg, 1983)*, volume 26 of *Lecture Notes in Statist.*, pages 256–272. Springer, New York.

ROUSSEEUW, P. J. & DRIESSEN, K. V. (1998) "A Fast Algorithm for the Minimum Covariance Determinant Estimator." *Technometrics*, Vol. 41, pp. 212–223.

ROUSSEEUW, P. J. & LEROY, A. M. (1987) *Robust regression and outlier detection*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York.

RUBIN, D. B. (1991) "EM and beyond." *Psychometrika*, Vol. 56(2), pp. 241–254.

SCHWARZ, G. (1978) "Estimating the dimension of a model." *Ann. Statist.*, Vol. 6(2), pp. 461–464.

SERFLING, R. (2002) "A depth function and a scale curve based on spatial quantiles." In *Statistical data analysis based on the L_1-norm and related methods (Neuchâtel, 2002)*, Stat. Ind. Technol., pages 25–38. Birkhäuser, Basel.

SERFLING, R. (2010) "Equivariance and invariance properties of multivariate quantile and related functions, and the role of standardisation." *J. Nonparametr. Stat.*, Vol. 22(7), pp. 915–936.

SHOHAM, S. (2002) "Robust clustering by deterministic agglomeration EM of mixtures of multivariate t-distributions." *Pattern Recognition*, Vol. 35(5), pp. 1127 – 1142.

SIRKIÄ, S.; TASKINEN, S.; OJA, H.; & TYLER, D. E. (2009) "Tests and estimates of shape based on spatial signs and ranks." *J. Nonparametr. Stat.*, Vol. 21(2), pp. 155–176.

SMALL, C. G. (1990) "A Survey of Multidimensional Medians." *International Statistical Review / Revue Internationale de Statistique*, Vol. 58(3), pp. pp. 263–277.

STAUDTE, R. G. & SHEATHER, S. J. (1990) *Robust estimation and testing*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York A Wiley-Interscience Publication.

SUNDBERG, R. (1972) "Maximum likelihood theory and applications for distributions generated when observing a function of an exponential family variable." Dissertaion.

TADJUDIN, S. & LANDGREBE, D. A. (2000) "Robust parameter estimation for mixture model." *IEEE Trans. Geoscience and Remote Sensing, Vol*, Vol. 38, pp. 439–445.

TANAKA, Y. (1988) "Sensitivity analysis in principal component analysis:influence on the subspace spanned by principal components.." *Communications in Statistics - Theory and Methods*, Vol. 17(9), pp. 3157–3175.

TASKINEN, S.; SIRKIÄ, S.; & OJA, H. (2010) "k-step shape estimators based on spatial signs and ranks." *J. Statist. Plann. Inference*, Vol. 140(11), pp. 3376–3388.

TUKEY, J. W. (1975) "Mathematics and the picturing of data." In *Proceedings of the International Congress of Mathematicians (Vancouver, B. C., 1974), Vol. 2*, pages 523–531. Canad. Math. Congress, Montreal, Que.

TYLER, D. E. (1987) "A distribution-free M-estimator of multivariate scatter." *Ann. Statist.*, Vol. 15(1), pp. 234–251.

TYLER, D. E. (1994) "Finite sample breakdown points of projection based multivariate location and scatter statistics." *Ann. Statist.*, Vol. 22(2), pp. 1024–1044.

VARDI, Y. & ZHANG, C.-H. (2000) "The Multivariate L1-Median and Associated Data Depth." *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 97(4), pp. pp. 1423–1426.

VETROV, D. P.; KROPOTOV, D. A.; & OSOKIN, A. A. (2010) "Automatic determination of the numbers of components in the EM algorithm for the restoration of a mixture of normal distributions." *Zh. Vychisl. Mat. Mat. Fiz.*, Vol. 50(4), pp. 770–783.

VISURI, S.; KOIVUNEN, V.; & OJA, H. (2000) "Sign and rank covariance matrices." *J. Statist. Plann. Inference*, Vol. 91(2), pp. 557–575 Prague Workshop on Perspectives in Modern Statistical Inference: Parametrics, Semi-parametrics, Non-parametrics (1998).

WILCOX, R. R. (2005) *Introduction to robust estimation and hypothesis testing*. Academic Press Inc., San Diego, CA.

WU, C.-F. J. (1983) "On the convergence properties of the EM algorithm." *Ann. Statist.*, Vol. 11(1), pp. 95–103.

YAMANISHI, K.; TAKEUCHI, J.-I.; WILLIAMS, G.; & MILNE, P. (2004) "On-Line Unsupervised Outlier Detection Using Finite Mixtures with Discounting Learning Algorithms." *Data Mining and Knowledge Discovery*, Vol. 8, pp. 275–300 10.1023/B:DAMI.0000023676.72185.7c.

ZHANG, Z. & CHEUNG, Y. (2006) "On Weight Design of Maximum Weighted Likelihood and an Extended EM Algorithm.." *IEEE Trans. Knowl. Data Eng.*, Vol. 18(10), pp. 1429–1434.

ZHOU, W. & DANG, X. (2010) "Projection based scatter depth functions and associated scatter estimators." *J. Multivariate Anal.*, Vol. 101(1), pp. 138–153.

Vita

Kai Yu was born in Guangzhou, China, in 1984. He received his B.Sc. degree in Financial Mathematics from Guangzhou University, China, in 2007.

From 2007 to 2012, he worked as a teaching assistant (graduate instructor) at the Department of Mathematics in the University of Mississippi, where he pursued his Ph.D. degree. His research interests include Nonparametric and Robust Multivariate Analysis, Expectation-Maximization algorithm, Statistical Learning and Data Mining, Experimental Design, and Survival Analysis.