On Some Classification Methods for High Dimensional and Functional Data

by

OLUSOLA SAMUEL MAKINDE

A thesis submitted to
The University of Birmingham
for the degree of
Doctor of Philosophy

School of Mathematics
The University of Birmingham
December 2014

Abstract

Classification involves assigning an observation to one of the known groups, on the basis of a vector of measurements on each of the observations. In this study, we propose classification method based on multivariate rank. We show that this classifier is Bayes rule under suitable conditions. Multivariate ranks are not invariant under affine transformation of the data and so, the effect of deviation from property of spherical symmetry is investigated. Based on this, we construct affine invariant version of this classifier. When the distributions of competing populations have different covariance matrices, minimum rank classifier performs poorly irrespective of affine invariance. To overcome this limitation, we propose a classifier based on multivariate rank region. The asymptotic properties of this method and its associated probability of misclassification are studied. Also, we propose classifiers based on the distribution of the spatial rank and establish some theoretical results for this classification method. For affine invariant version of this method, two invariants are proposed. Many multivariate techniques fail to perform well when data are curves or functions. We propose classification method based on L_2 distance to spatial median and later generalise it to L_p distance to L_p median. The optimal choice of p is determined by cross validation of misclassification errors. The performances of our propose methods are examined by using simulation and real data set and the results are compared with the results from existing methods.

Acknowledgements

I would like to deeply thank my supervisor, Dr Biman Chakraborty, who provided me support and encouragement. I really appreciate your suggestions, comments, effort and all the time spent in helping me to make this thesis possible. Thank you very much. I will like to thank all my friends in the School of Mathematics, especially Mohammed Hussein Baragilly, Hend Gabr and Dr Vipin Michael. You are wonderful people.

I would like to express my deepest gratitude to my parents, brothers and sisters for their love and support throughout my education. I would also like to thank Apostle Paul Iyiola Adedosu, Dr Olusoga Akin Fasoranbaku and Mr Gbenga Jacob Abiodun for their support and encouragement. A big thank you to my fiancée, Olubukola Olayemi Makanjuola for her undaunted love, support and majorly, her prayers. I love you.

I would like to thank the School of Mathematics, University of Birmingham and the Federal University of Technology, Akure, Nigeria for the financial support that I received.

To all who have in one way or the other contributed to this thesis, I say thank you very much.

To almighty God.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Classification is aimed at getting maximum information about separability or distinction among classes or populations and then assigns each observation to one of these populations on the basis of a vector of measurements or features, denoted by $\mathbf{x}$, on each of the observations. It has many important applications in different fields, such as disease diagnosis in medical sciences, risk identification in finance, admission of prospective students into university based on a battery of tests, among others. An example is to classify iris flower (Fisher, 1936) from unknown group or species to any of the three known species on the basis of their attributes (See Figure 1.1). The known groups or species of iris flowers are Iris Setosa (red), Iris Versicolour (green) and Iris Virginica (black). The attributes are sepal length in cm, sepal width in cm, petal length in cm and petal width in cm.

Anderson (1984) described classification problem as the problem of statistical decision making. A good classification procedure is the one that classifies observations from unknown populations correctly. Suppose each population has a well defined distribution function, which is characterised by some location and scale parameters. Classification of observations to populations can be viewed from this characterisation in terms of shift in location and scale of each of the population distributions. In a classification problem, competing populations may have either location shift, scale shift or both (location-scale

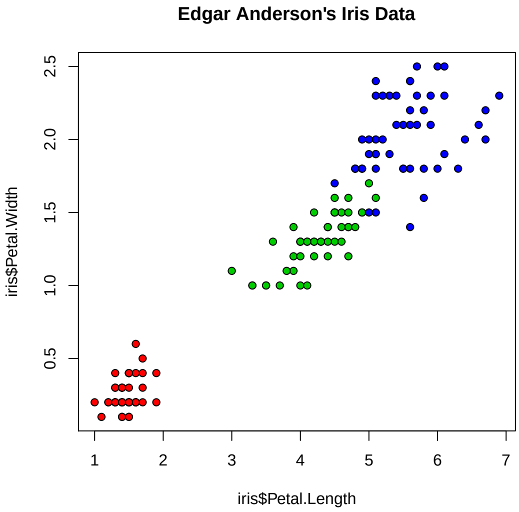


Figure 1.1: Iris data: petal width vs petal length

shift). Consider populations $\pi_j, j = 1, 2, \ldots, J$ from multivariate distributions, F_j having probability density functions f_j with prior probabilities p_j. Bayes rule, proposed in Welch (1939), is to classify each observation to the population π_j, whose posterior probability $P(\pi_j|\mathbf{x})$ is the highest. It assigns $\mathbf{x}$ to population π_k if

$$P(\pi_k|\mathbf{x}) = \max_{1 \leqslant j \leqslant J} P(\pi_j|\mathbf{x}) = \max_{1 \leqslant j \leqslant J} \frac{f_j(\mathbf{x})p_j}{\sum_{j=1}^{J} f_j(\mathbf{x})p_j}.$$

This is equivalent to assigning $\mathbf{x}$ to population π_1, in a two class problem, if

$$\frac{f_1(\mathbf{x})p_1}{f_2(\mathbf{x})p_2} > 1,$$

and to π_2 otherwise. Suppose R_1 and R_2 are regions for classifying observations to populations π_1 and π_2, having probability density functions f_1 and f_2, and prior probabilities p_1 and p_2 respectively. Classification procedure involves assigning $\mathbf{x}$ to π_1 if $\mathbf{x} \in R_1$ (i.e. $\mathbf{x}$ is in region R_1) or to π_2 if $\mathbf{x} \in R_2$ (i.e. $\mathbf{x}$ is in region R_2). Anderson (1984) described

R_1 and R_2 as the region for which

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \gtrless \frac{p_2}{p_1}$$

respectively if

$$P\left(\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} = \frac{p_2}{p_1} \;\middle|\; \pi_i\right) = 0,\ i = 1, 2.$$

Then the procedure is unique except for sets of probability zero. If $\mathbf{x}$ is on the boundary, then it belongs to either of R_1 and R_2. We assign such observation to R_1.

Wald (1944) argued that if each population has a cost, $C(i|j)$ associated with misclassifying $\mathbf{x}$ whose true population is π_j into π_i, then assign observations to the class or population that has the highest expected cost of misclassification (that is, $C(i|j)P(\pi_j|\mathbf{x})$ is highest). In a two class problem, define $c(2|1)$ as cost of misclassifying observation $\mathbf{x}$ whose true population is π_1 into population π_2 and $c(1|2)$ as cost of misclassifying observation $\mathbf{x}$ whose true population is π_2 into population π_1. Expected cost of misclassifying $\mathbf{x}$ whose true population is π_1 is $c(2|1)P(\pi_1|\mathbf{x})$. Expected cost of misclassifying $\mathbf{x}$ whose true population is π_2 is $c(1|2)P(\pi_2|\mathbf{x})$. Mathematically, Wald's proposal is to assign $\mathbf{x}$ to π_1 if

$$\frac{c(2|1)P(\pi_1|\mathbf{x})}{c(1|2)P(\pi_2|\mathbf{x})} = \frac{c(2|1)f_1(\mathbf{x})p_1}{c(1|2)f_2(\mathbf{x})p_2} \geqslant 1,$$

and to π_2 otherwise. The regions of classification are defined as

$$R_1 : \ \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geqslant \frac{c(1|2)p_2}{c(2|1)p_1} \ \text{ and } \ R_2 : \ \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \frac{c(1|2)p_2}{c(2|1)p_1}.$$

Welch (1939) showed that for any two normally distributed populations, the ratio of log likelihood functions of the two populations is the theoretical basis for building discriminant function that best classify new individuals to any of the two populations given that the prior probabilities of the populations are known. Since $f(\mathbf{x})$ and $\log_e f(\mathbf{x})$ attain their maximum values at the same value of $\mathbf{x}$, then the regions R_1 and R_2 are equivalent to

$$\begin{aligned} R_1: \quad \log_e \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} &\geqslant \log_e \frac{c(1|2)p_2}{c(2|1)p_1} \\ R_2: \quad \log_e \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} &< \log_e \frac{c(1|2)p_2}{c(2|1)p_1}. \end{aligned} \tag{1.0.1}$$

1.1 Misclassification Errors

In classifying an observation into either π_1 or π_2 with prior probabilities p_1 and p_2 respectively, either of these two errors can be made; error of misclassifying an observation $\mathbf{x}$ that is actually from π_1 into π_2 with probability $p_1 P(\mathbf{x} \in R_2|\pi_1) = p_1 P(2|1)$ or misclassifying $\mathbf{x}$ that is actually from π_2 into π_1 with probability $p_2 P(\mathbf{x} \in R_1|\pi_2) = p_2 P(1|2)$. Total probability of misclassifying an observation is the sum of probabilities that the observation comes from population π_i but does not eventually fall in the region of classification into population π_i, where $i = 1, 2$. Mathematically, the total probability of misclassification, denoted by Δ, is

$$\Delta = p_1 P(\mathbf{x} \in R_2|\pi_1) + p_2 P(\mathbf{x} \in R_1|\pi_2) = p_1 P(2|1) + p_2 P(1|2).$$

Suppose there are $J(>2)$ classes, the total probability of misclassification is

$$\Delta = \sum_{j=1}^{J} p_j P(\mathbf{x} \notin R_j|\pi_j).$$

1.2 Linear and Quadratic Classification Rules

Suppose there are two populations with equal covariance matrix (this case is referred to as location shift or homogenous scale), Fisher (1936) described the separation between these two populations to be ratio of variance between the populations to variance within the populations. This postulation leads to discriminant analysis, called Fisher's discriminant analysis. Suppose there are two populations from the same family of multivariate distributions to which observations can be classified. If these populations are normally distributed and have the same covariance matrix, the disriminant analysis is referred to as linear discriminant analysis (LDA). Similarly, if these populations are normally distributed but have different covariance matrices, the optimal rule is nonlinear and referred to as quadratic discriminant analysis (QDA). QDA can be seen as the problem of scale shift or location-scale shift, depending on whether the populations have the same location vector or not. Based on Fisher (1936), Welch (1939) and Wald (1944) showed that linear discriminant function has optimal properties for two group classification if the populations are multivariate normally distributed.

Suppose competing populations are normally distributed, it follows from equation (1.0.1) that the classification procedure is to assign $\mathbf{x}$ into π_1 if

$$\begin{aligned}&-\frac{1}{2}\log_e|\boldsymbol{\Sigma}_1|+\frac{1}{2}\log_e|\boldsymbol{\Sigma}_2|-\frac{1}{2}\mathbf{x}^T(\boldsymbol{\Sigma}_1^{-1}-\boldsymbol{\Sigma}_2^{-1})\mathbf{x}+\mathbf{x}^T(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1-\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2)\\&-\frac{1}{2}(\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1-\boldsymbol{\mu}_2^T\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2)\geqslant\log_e\left(\frac{c(1|2)p_2}{c(2|1)p_1}\right).\end{aligned}\tag{1.2.1}$$

Assign $\mathbf{x}$ into π_2 otherwise, where $(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$ and $(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)$ are pairs of mean vector and covariance matrix corresponding to the distributions of π_1 and π_2 respectively. When population covariance matrices are the same (i.e. $\boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}_2=\boldsymbol{\Sigma}$), the LHS of equation (1.2.1) becomes

$$U=\mathbf{x}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)-\frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2).\tag{1.2.2}$$

U is a linear function of vector of measurements on individual observation containing maximum information about class separability, called linear discriminant function (LDF).

Theorem 1.2.1 (Anderson, 1984) *If* $p_1 = p_2$*,* $c(2|1) = c(1|2)$ *and* $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$*, the best regions of classification corresponding to Bayes' rule are*

$$\begin{aligned} R_1 &: \mathbf{x}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \geqslant \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ R_2 &: \mathbf{x}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) < \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2). \end{aligned} \tag{1.2.3}$$

Suppose $\mathbf{x}$ is distributed as $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, $E(U) = \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, $\mathrm{var}(U) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ and $U \sim N(\frac{1}{2}c_0^2, c_0^2)$, where $c_0^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$. Similarly, if $\mathbf{x}$ is distributed as $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, then $E(U) = \frac{-1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, $\mathrm{var}(U) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ and $U \sim N(\frac{-1}{2}c_0^2, c_0^2)$. See pages 205 - 206 of Anderson (1984) for detail.

Define $p_1P(2|1)$ as probability that $\mathbf{x}$ comes from population π_1 but eventually falls in the region of classification into population π_2 and $p_2P(1|2)$ as probability that $\mathbf{x}$ comes from population π_2 but eventually falls in the region of classification into population π_1, as discussed in Section 1.1. Suppose $\mathbf{x}$ is distributed as $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, then $E[\mathbf{x}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)] = \boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, $\mathrm{var}\big(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\big) = c_0^2$ and

$$P(2|1) = P\bigg[\mathbf{x}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) < \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\big|\pi_1\bigg] = \Phi\bigg(\frac{-c_0}{2}\bigg),$$

where Φ is the cumulative distribution function of the standard normal distribution. Similarly, suppose $\mathbf{x}$ is distributed as $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, then $E[\mathbf{x}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)] = \boldsymbol{\mu}_2^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, $\mathrm{var}\big(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\big) = c_0^2$ and

$$P(1|2) = P\bigg[\mathbf{x}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \geqslant \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\big|\pi_2\bigg] = \Phi\bigg(\frac{-c_0}{2}\bigg).$$

The total probability of misclassification of $\mathbf{x}$ into either π_1 or π_2 is

$$\Delta = p_2 P(1|2) + p_1 P(2|1) = p_1 \Phi\left(\frac{-c_0}{2}\right) + p_2 \Phi\left(\frac{-c_0}{2}\right) = \Phi\left(\frac{-c_0}{2}\right) \tag{1.2.4}$$

since $p_1 + p_2 = 1$. A good classification method is the one that minimises Δ.

Now consider $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$, the regions R_1 and R_2 corresponding to Bayes' rule are

$$\begin{aligned} R_1 &: -\frac{1}{2}\mathbf{x}^T(\mathbf{\Sigma}_1^{-1} - \mathbf{\Sigma}_2^{-1})\mathbf{x} + \mathbf{x}^T(\mathbf{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \mathbf{\Sigma}_2^{-1}\boldsymbol{\mu}_2) \geqslant k^* \\ R_2 &: -\frac{1}{2}\mathbf{x}^T(\mathbf{\Sigma}_1^{-1} - \mathbf{\Sigma}_2^{-1})\mathbf{x} + \mathbf{x}^T(\mathbf{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \mathbf{\Sigma}_2^{-1}\boldsymbol{\mu}_2) < k^*, \end{aligned} \tag{1.2.5}$$

where $k^* = \frac{1}{2}(\boldsymbol{\mu}_1^T\mathbf{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^T\mathbf{\Sigma}_2^{-1}\boldsymbol{\mu}_2) + \frac{1}{2}\log_e\left(\frac{|\mathbf{\Sigma}_1|}{|\mathbf{\Sigma}_2|}\right) + \log_e\left(\frac{c(1|2)p_2}{c(2|1)p_1}\right)$. Equation (1.2.5) is quadratic in $\mathbf{x}$ when $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$. It becomes linear when $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$.

Theorem 1.2.2 (Gilbert, 1969) *Suppose π_1 and π_2 are two populations from $N(\boldsymbol{\mu}_1, \mathbf{\Sigma}_1)$ and $N(\boldsymbol{\mu}_2, \mathbf{\Sigma}_2)$ with prior probabilities p_1 and p_2 respectively, where $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$. Then the Bayes' regions of classification into π_1 and π_2 are as given in equation (1.2.5). Furthermore if $\mathbf{\Sigma}_2 = \sigma^2\mathbf{\Sigma}_1$ and $\sigma \neq 1$, then*

$$\Delta = \begin{cases} p_1 P(\chi^2_{f_1} > \frac{k}{c_1}) + p_2 P(\chi^2_{f_2} < \frac{k}{c_2}), & \textit{for } \sigma^2 > 1 \\ p_1 P(\chi^2_{f_1} < -\frac{k}{c_1}) + p_2 P(\chi^2_{f_2} > -\frac{k}{c_2}), & \textit{for } \sigma^2 < 1 \end{cases} \tag{1.2.6}$$

where

$$\begin{aligned} & k = \log_e\left(\frac{p_1}{p_2}\right) + \frac{d}{2}\log_e(\sigma^2) + \frac{[p_1 + p_2\sigma^2]U^2}{2(\sigma^2 - 1)}, \quad U^2 = \boldsymbol{\upsilon}^T\mathbf{\Sigma}\boldsymbol{\upsilon}, \\ & \boldsymbol{\upsilon} = \mathbf{A}^T\mathbf{\Sigma}^{-\frac{1}{2}}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \quad \mathbf{\Sigma} = p_1\mathbf{\Sigma}_1 + p_2\mathbf{\Sigma}_2, \quad c_i = \frac{\overline{\sigma}_i^2}{\overline{\mu}_i}, \quad f_i = \frac{\overline{\mu}_i^2}{c_i}, \quad i = 1, 2, \\ & \overline{\mu}_1 = \frac{1}{2\sigma^2}\left\{\frac{[p_1 + p_2\sigma^2]U^2}{|\sigma^2 - 1|} + d|\sigma^2 - 1|\right\}, \quad \overline{\mu}_2 = \frac{1}{2}\left\{\frac{\sigma^2[p_1 + p_2\sigma^2]U^2}{|\sigma^2 - 1|} + d|\sigma^2 - 1|\right\}, \\ & \overline{\sigma}_1^2 = \frac{1}{\sigma^2}\left\{[p_1 + p_2\sigma^2]U^2 + \frac{d(\sigma^2 - 1)^2}{2}\right\}, \quad \overline{\sigma}_2^2 = \sigma^2[p_1 + p_2\sigma^2]U^2 + \frac{d(\sigma^2 - 1)^2}{2}, \end{aligned}$$

where $\mathbf{A}$ *is the orthogonal matrix such that* $\mathbf{A}^T\boldsymbol{\Sigma}_1^{-\frac{1}{2}}\boldsymbol{\Sigma}_2(\boldsymbol{\Sigma}_1^{-\frac{1}{2}})'\mathbf{A}$ *is a diagonal matrix.*

Many researchers have worked on the estimation of probability of misclassification given that observations are from multivariate normally distributed random samples or populations, which include studies of Anderson and Bahadur (1962), Dunn (1971), Anderson (1972), Das Gupta (1972), Chang and Afifi (1974).

In practice, population quantities are unknown. So, Wald (1944) and Anderson (1984) suggested replacing the population parameters with their sample estimates for a large sample size. Suppose $\mathbf{X}_{11}, \mathbf{X}_{12}, \ldots, \mathbf{X}_{1n_1} \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $\mathbf{X}_{21}, \mathbf{X}_{22}, \ldots, \mathbf{X}_{2n_2} \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$. Let

$$\overline{\mathbf{X}}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1}\mathbf{X}_{1i},\ \overline{\mathbf{X}}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2}\mathbf{X}_{2i} \text{ and } \mathbf{S} = \frac{\sum_{k=1}^{2}(n_k-1)\mathbf{S}_k}{\sum_{k=1}^{2}(n_k-1)},\ k = 1, 2$$

be estimators of $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}$ respectively, where $\mathbf{S}_k$ is the estimate of covariance matrix of kth sample with size n_k. The empirical version of U in equation (1.2.2) is

$$T = \mathbf{x}^T\mathbf{S}^{-1}(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2) - \frac{1}{2}(\overline{\mathbf{X}}_1 + \overline{\mathbf{X}}_2)^T\mathbf{S}^{-1}(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2). \tag{1.2.7}$$

Theorem 1.2.3 (Anderson, 1984) *The limiting distribution of* T *as* $n_1 \to \infty$ *and* $n_2 \to \infty$ *is* $N(\frac{1}{2}c_0^2, c_0^2)$ *if* $\mathbf{x}$ *is distributed according to* $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ *and* $N(-\frac{1}{2}c_0^2, c_0^2)$ *if* $\mathbf{x}$ *is distributed according to* $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$*, where* $c_0 = \sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}$*.*

Krzanowski (1977) and Johnson and Wichern (2007) called this sample version of LDF, an Anderson statistic. Fisher's linear discriminant function and Anderson statistic are popular techniques in multivariate statistics. Hills (1967) pointed out that Fisher's LDF provides a useful tool for discriminating between populations under wide distributional conditions though it has a limitation that its performance may be suboptimal when populations are not multivariate normally distributed. Krzanowski (1977) reviewed the performance of Fisher's linear discriminant function when underlying assumptions are violated.

Sitgreaves (1961), Memon and Okamoto (1971) worked on the distribution of the classification statistics.

1.2.1 Numerical Example

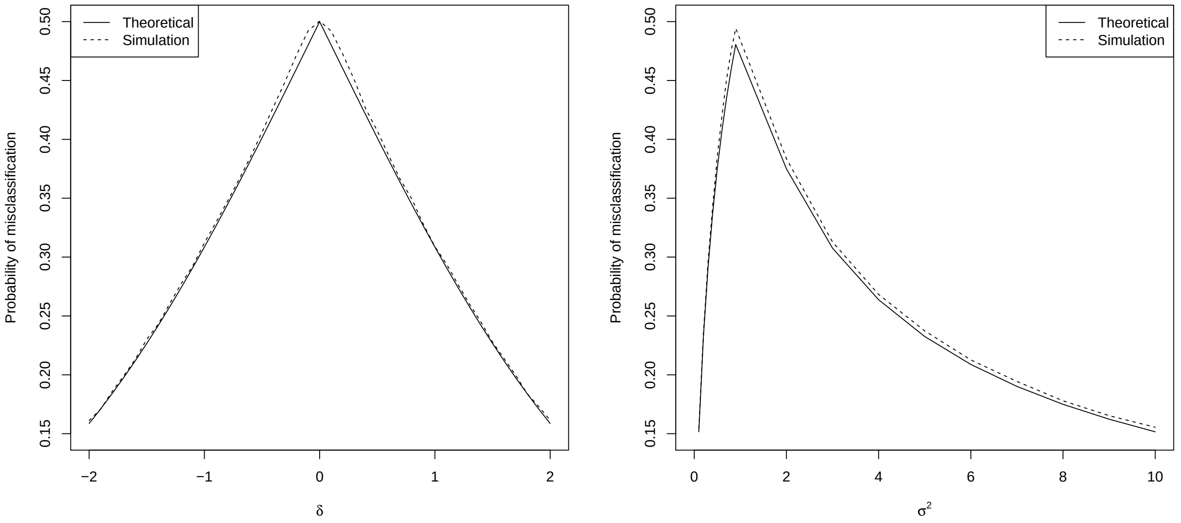
Example 1 : Normal populations with location shift

In this example, we want to compare known theoretical result with simulation result. Let π_1 and π_2 be two d-variate normal populations with mean vector and covariance matrix, $(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ respectively. Assume that the prior probabilities, p_1 and p_2 and costs of misclassifcation, $c(2|1)$ and $c(1|2)$ of π_1 and π_2 respectively, are equal. Consider

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\mu}_2 = \begin{pmatrix} \delta \\ 0 \end{pmatrix} \text{ and } \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \mathbf{I}_2,$$

where $\mathbf{I}_2$ is a 2 × 2 identity matrix. The total probability of misclassification associated with LDA is a function of non-centrality parameter δ and is $\Phi\big(\frac{-\delta}{2}\big)$.

Figure 1.2 present the comparison between theoretical probability of misclassification and empirical error rate based on simulation study. It is clearly shown in Figure 1.2(a) that the sample estimate of probability of misclassification associated with LDA is a good approximation for its population version. Figure 1.3 presents a comparison of misclassification rates among three bivariate spherically symmetric distributions for various values of δ. The distributions are bivariate normal distribution, bivariate Laplace distribution and bivariate t distribution with 3 degrees of freedom. The chance of misclassifying observations varies from one distribution to another. The chance of misclassifying observations is least in bivariate normally distributed samples and highest in bivariate Laplace distributed samples.



(a) LDA

(b) QDA

Figure 1.2: Misclassification Error: Theoretical versus Simulation

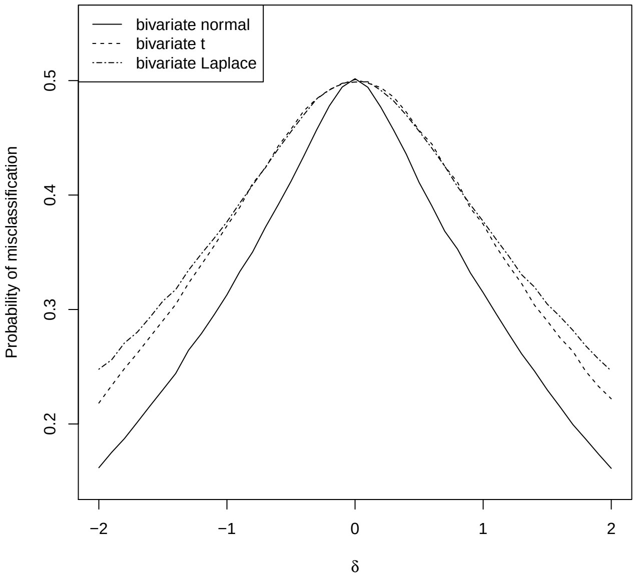


Figure 1.3: Misclassification error rates associated with LDA for spherical distributions with $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \mathbf{I}_2$ and $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$.

Example 2 : Normal populations with scale shift

Consider the set up as in Example 1 above, but take $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and $\boldsymbol{\Sigma}_1 = \mathbf{I}_2, \boldsymbol{\Sigma}_2 = \sigma^2 \mathbf{I}_2$ for $\sigma \neq 1$. For $\mathbf{x} \in \mathbb{R}^d$, if $\mathbf{x} \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, $\mathbf{x}^T\mathbf{x} \sim \chi^2_d$ and if $\mathbf{x} \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, $\mathbf{x}^T\mathbf{x} \sim \sigma^2\chi^2_d$. $f_1(\mathbf{x})/f_2(\mathbf{x}) \geqslant 1$ implies $e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)+\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)} \geqslant (|\boldsymbol{\Sigma}_1|/|\boldsymbol{\Sigma}_2|)^{1/2}$, which can be written as

$$(\mathbf{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) - (\mathbf{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) \leqslant \log_e|\boldsymbol{\Sigma}_2| - \log_e|\boldsymbol{\Sigma}_1|.$$

This gives

$$\begin{aligned} \mathbf{x}^T\mathbf{x} - \frac{1}{\sigma^2}\mathbf{x}^T\mathbf{x} &\leqslant 2\log_e(\sigma^2) - \log_e(1) \\ \left(1 - \frac{1}{\sigma^2}\right)\mathbf{x}^T\mathbf{x} &\leqslant 2\log_e(\sigma^2) \\ \left(\frac{\sigma^2 - 1}{\sigma^2}\right)\mathbf{x}^T\mathbf{x} &\leqslant 2\log_e(\sigma^2). \end{aligned}$$

For $\sigma^2 > 0$, we consider two cases. These are $\sigma^2 > 1$ and $\sigma^2 < 1$.

1. When $\sigma^2 > 1$, the region of classification is

$$R_1 : \mathbf{x}^T\mathbf{x} \leqslant \frac{2\sigma^2}{\sigma^2 - 1}\log_e \sigma^2 \ \text{ and } \ R_2 : \mathbf{x}^T\mathbf{x} > \frac{2\sigma^2}{\sigma^2 - 1}\log_e \sigma^2.$$

Then

$$\begin{aligned} P(2|1) &= P\left(\mathbf{x}^T\mathbf{x} > \frac{2\sigma^2}{\sigma^2-1}\log_e\sigma^2 \;\middle|\; \mathbf{x}^T\mathbf{x} \sim \chi^2_2\right) = 1 - F_2\left(\frac{2\sigma^2}{\sigma^2-1}\log_e\sigma^2\right), \\ P(1|2) &= P\left(\mathbf{x}^T\mathbf{x} \leqslant \frac{2\sigma^2}{\sigma^2-1}\log_e\sigma^2 \;\middle|\; \mathbf{x}^T\mathbf{x} \sim \sigma^2\chi^2_2\right) = F_2\left(\frac{2}{\sigma^2-1}\log_e\sigma^2\right) \end{aligned}$$

and Δ, probability of misclassification is

$$\Delta = \frac{1}{2}\left[1 - F_2\left(\frac{2\sigma^2}{\sigma^2-1}\log_e \sigma^2\right) + F_2\left(\frac{2}{\sigma^2-1}\log_e \sigma^2\right)\right]$$

since $p_1 = p_2 = 0.5$. Here $F_2(.)$ denotes distribution function of central Chi-square distribution with 2 degrees of freedom.

2. When $\sigma^2 < 1$, the region of classification is

$$R_1 : \mathbf{x}^T\mathbf{x} \geqslant \frac{2\sigma^2}{\sigma^2-1}\log_e \sigma^2 \text{ and } R_2 : \mathbf{x}^T\mathbf{x} < \frac{2\sigma^2}{\sigma^2-1}\log_e \sigma^2.$$

$$P(2|1) = P\left(\mathbf{x}^T\mathbf{x} < \frac{2\sigma^2}{\sigma^2-1}\log_e \sigma^2 \;\middle|\; \mathbf{x}^T\mathbf{x} \sim \chi^2_2\right) = F_2\left(\frac{2\sigma^2}{\sigma^2-1}\log_e \sigma^2\right)$$
$$P(1|2) = P\left(\mathbf{x}^T\mathbf{x} \geqslant \frac{2\sigma^2}{\sigma^2-1}\log_e \sigma^2 \;\middle|\; \mathbf{x}^T\mathbf{x} \sim \sigma^2\chi^2_2\right) = 1 - F_2\left(\frac{2}{\sigma^2-1}\log_e \sigma^2\right)$$

where $F_2(.)$ is the distribution function of central Chi-square distribution with 2 degrees of freedom. The probability of misclassification is

$$\Delta = \frac{1}{2}\left[1 + F_2\left(\frac{2\sigma^2}{\sigma^2-1}\log_e \sigma^2\right) - F_2\left(\frac{2}{\sigma^2-1}\log_e \sigma^2\right)\right]$$

since $p_1 = p_2 = 0.5$. These results are compared with empirical results based on simulation. The numerical results are presented in Figure 1.2(b).

Example 3 : Normal populations with location-scale shift

Now consider $\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $\boldsymbol{\mu}_2 = \begin{pmatrix} \delta \\ 0 \end{pmatrix}$ and $\boldsymbol{\Sigma}_1 = \mathbf{I}_2, \boldsymbol{\Sigma}_2 = \sigma^2 \mathbf{I}_2$, we use Theorem 1.2.2 and obtain

$$\Delta = \begin{cases} p_1 P(\chi^2_{f_1} > \frac{k}{c_1}) + p_2 P(\chi^2_{f_2} < \frac{k}{c_2}), & \text{for } \sigma^2 > 1 \\ p_1 P(\chi^2_{f_1} < -\frac{k}{c_1}) + p_2 P(\chi^2_{f_2} > -\frac{k}{c_2}), & \text{for } \sigma^2 < 1 \end{cases}$$

where

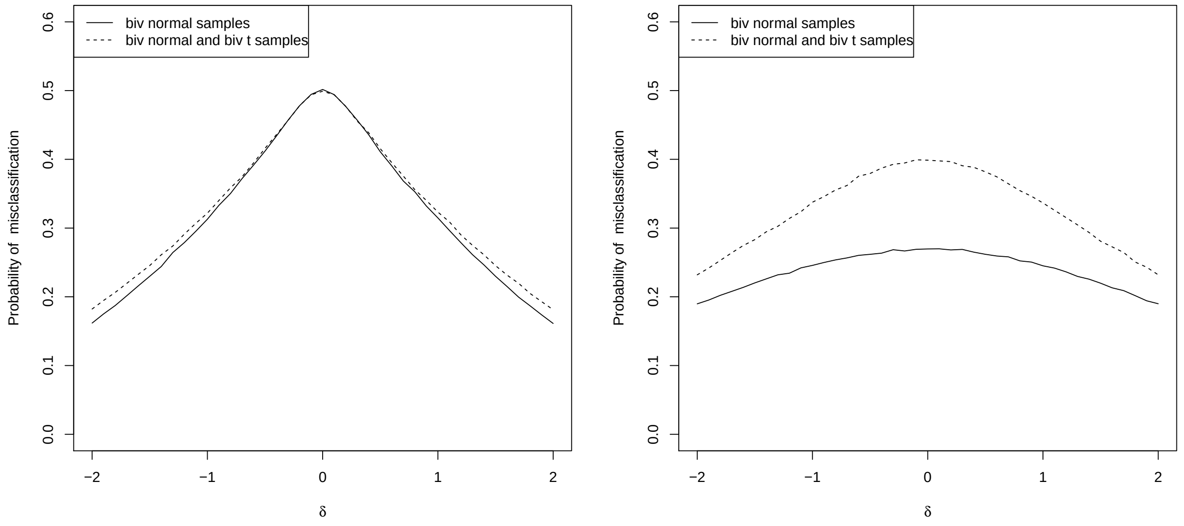
$$k = \log_e \sigma^2 + \frac{1}{4}\frac{\delta^2(\sigma^2+1)}{\sigma^2-1}, \quad c_i = \frac{\overline{\sigma}_i^2}{\overline{\mu}_i}, \quad f_i = \frac{\overline{\mu}_i^2}{c_i}, \quad i = 1, 2,$$

$$\boldsymbol{\Sigma} = \mathbf{I}_2 + \sigma^2\mathbf{I}_2, \;\; \mathbf{A} = \mathbf{I}_2, \;\; \boldsymbol{\upsilon} = \boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \;\; U^2 = \boldsymbol{\upsilon}^T\boldsymbol{\Sigma}\boldsymbol{\upsilon} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \delta^2,$$

$$\overline{\mu}_1 = \frac{1}{2\sigma^2}\Big\{\frac{\frac{1}{2}[1+\sigma^2]\delta^2}{|\sigma^2-1|} + 2|\sigma^2-1|\Big\}, \quad \overline{\mu}_2 = \frac{1}{2}\Big\{\frac{\frac{1}{2}\sigma^2[1+\sigma^2]\delta^2}{|\sigma^2-1|} + 2|\sigma^2-1|\Big\},$$

$$\overline{\sigma}_1^2 = \frac{1}{\sigma^2}\Big\{\frac{1}{2}[1+\sigma^2]\delta^2 + (\sigma^2-1)^2\Big\}, \quad \overline{\sigma}_2^2 = \frac{1}{2}\sigma^2[1+\sigma^2]\delta^2 + (\sigma^2-1)^2.$$

LDA and QDA have some inadequacies for non-normal distributions. LDA and QDA are Bayes rules under normality for location shift and location-scale shift respectively. This means that they are optimal when normality is assumed. These classifiers are not optimal when some or all the competing distributions are non-normal. To illustrate this, suppose F and G are both not multivariate normal distributions but are from the same family of distributions. We compare their misclassification rates with when F and G are both multivariate normal. The results are shown in Figure 1.3. It is clearly shown from this figure that misclassification rates for non-normal distributions is higher than that of normal distributions. Similarly, suppose $F \equiv t(3, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ is a multivariate t distribution with mean $\boldsymbol{\mu}_1$, variance $\boldsymbol{\Sigma}_1$ and 3 degrees of freedom, and $G \equiv N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, where $\boldsymbol{\mu}_1$, $\boldsymbol{\Sigma}_1$, $\boldsymbol{\mu}_2$, and $\boldsymbol{\Sigma}_2$ are as defined in examples 1 and 3 for location shift and location-scale shift respectively and $\sigma = 2$. LDA and QDA have higher misclassification rates also in this



(a) LDA

(b) QDA

Figure 1.4: Effect of normality on optimality of LDA and QDA.

case than when F and G are both multivariate normal, as shown in Figure 1.4 (a)-(b). These results confirm the optimality of LDA and QDA if all competing distributions are normally distributed.

Also, some moments of some non-normal distributions do not exist, for example multivariate Cauchy distribution. This may limit the use of LDA and QDA. Furthermore, Hubert and Van Driessen (2004) has shown that outlying training sample points affect the performance of LDA and QDA. Hence, both linear and quadratic classifiers are not robust against outliers.

1.2.2 Robust Version of Linear and Quadratic Classification Rules

Hubert and Van Driessen (2004) proposed a robust versions of LDA and QDA called robust linear discriminant analysis (RLDA) and robust quadratic discriminant analysis (RQDA) respectively. Both involve replacing the estimates of $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ in equation (1.2.1) by reweighted MCD estimator of multivariate location and scatter based on FAST-MCD algorithm of Rousseeuw and Van Driessen (1999).

1.2.3 Derivation of Theoretical Bayes Risk - Location Shift

We want to derive Bayes risk (misclassification probability associated with Bayes rule) for some competing distributions with location shift in a two-class problem. The distributions are multivariate normal distribution, multivariate t distribution with k degree of freedom and multivariate Laplace distribution.

Multivariate normal distribution

Suppose π_1 has distribution $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ with prior probability p_1 and π_2 has distribution $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ with prior probability p_2. The probability of misclassification associated with Bayes rule, denoted by Δ_B is $\Delta_B = \Phi(-\frac{c_0}{2})$, where $c_0 = \sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}$ and Φ is the distribution function of standard normal distribution. See equation (1.2.4).

Multivariate t distribution

Let $\mathbf{Z} \sim N_d(\mathbf{0}, \boldsymbol{\Sigma})$ and $U \sim \chi^2_k$ be independent, where k is the degree of freedom of Chi-squared distribution. Define

$$\mathbf{X} = \left(\mathbf{Z} \sqrt{\frac{k}{U}} \right) + \boldsymbol{\mu} \tag{1.2.8}$$

The distribution of $\mathbf{X}$ is multivariate t distribution with k degree of freedom, denoted by $t(k, \boldsymbol{\mu}, \boldsymbol{\Sigma})$. The probability density function of $\mathbf{x}$ is

$$f(\mathbf{x}) = (k\pi)^{-\frac{d}{2}} \frac{\mathit{\Gamma}(\frac{k+d}{2})}{\mathit{\Gamma}(\frac{k}{2})} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \{1 + \frac{1}{k}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\}^{-(\frac{k+d}{2})}. \tag{1.2.9}$$

Suppose π_1 has distribution $t(k, \boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ with probability density function $f_1(\mathbf{x})$ and π_2 has distribution $t(k, \boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ with probability density function $f_2(\mathbf{x})$. Let π_1 and π_2 have equal prior probabilities (that is, $p_1 = p_2 = 0.5$). Bayes rule is to assign $\mathbf{x}$ to π_1 if

$$f_1(\mathbf{x}) > f_2(\mathbf{x}),$$

which is equivalent to

$$(\mathbf{x} - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_1) < (\mathbf{x} - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_2). \tag{1.2.10}$$

This holds if the competing distributions have the same degree of freedom. Equation (1.2.11) reduces to

$$-2\mathbf{x}^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) < 0$$

and can be written as

$$T(\mathbf{z}) = \left(\mathbf{z} \sqrt{\frac{k}{u}} + \boldsymbol{\mu} \right)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \frac{1}{2} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) > 0,$$

where $\mathbf{x} = \mathbf{z}\sqrt{\frac{k}{u}} + \boldsymbol{\mu}$ and $\mathbf{z}$ is distributed as $N_d(\mathbf{0}, \boldsymbol{\Sigma})$. If $\mathbf{x}$ is from π_1, $\boldsymbol{\mu} = \boldsymbol{\mu}_1$. Similarly, if $\mathbf{x}$ is from π_2, $\boldsymbol{\mu} = \boldsymbol{\mu}_2$. Define $p_1 P(2|1)$ as probability that $\mathbf{x}$ comes from population π_1 but eventually falls in the region of classification into population π_2 and $p_2 P(1|2)$ as probability that $\mathbf{x}$ comes from population π_2 but eventually falls in the region of classification into population π_1.

$$\begin{aligned}
P(2|1) &= P\left[\sqrt{\frac{k}{u}}\mathbf{z}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)-\frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)+\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)<0\right]\\
&= P\left[\mathbf{z}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)-\frac{1}{2}\sqrt{\frac{u}{k}}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)+\sqrt{\frac{u}{k}}\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)<0\right]\\
&= P\left[\mathbf{z}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)<\frac{1}{2}\sqrt{\frac{u}{k}}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)-\sqrt{\frac{u}{k}}\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\right]\\
&= P\left[\mathbf{z}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)<\frac{-1}{2}\sqrt{\frac{u}{k}}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\right]
\end{aligned}$$

This holds because u takes values in $[0,\infty)$. For either of the population, $E\big(\mathbf{z}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\big)=\mathbf{0}$ and $\text{var}\big(\mathbf{z}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\big)=(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$, then

$$\begin{aligned}
P(2|1) &= P\left[R<\frac{\frac{-1}{2}\sqrt{\frac{u}{k}}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)}{\big((\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\big)^{1/2}}\right]\\
&= P\left[R<\frac{-1}{2}\sqrt{\frac{u}{k}}c_0\right]=\int\Phi(c_1)f_u(u)du
\end{aligned}$$

where Φ is the distribution function of the standard normal distribution, f_u is probability density function of χ_k^2 and R is a standard normal random variable. Similarly,

$$\begin{aligned}
P(1|2) &= P\left[\sqrt{\frac{k}{u}}\mathbf{z}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)-\frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)+\boldsymbol{\mu}_2^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)>0\right]\\
&= P\left[\mathbf{z}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)-\frac{1}{2}\sqrt{\frac{u}{k}}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)+\sqrt{\frac{u}{k}}\boldsymbol{\mu}_2^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)>0\right]\\
&= P\left[\mathbf{z}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)>\frac{1}{2}\sqrt{\frac{u}{k}}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)-\sqrt{\frac{u}{k}}\boldsymbol{\mu}_2^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\right]\\
&= P\left[R>\frac{\frac{1}{2}\sqrt{\frac{u}{k}}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)-\sqrt{\frac{u}{k}}\boldsymbol{\mu}_2^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)}{\big((\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\big)^{1/2}}\right]\\
&= 1-P\left[R<\frac{\frac{1}{2}\sqrt{\frac{u}{k}}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)}{\big((\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\big)^{1/2}}\right]\\
&= 1-P\left[R<\frac{1}{2}\sqrt{\frac{u}{k}}c_0\right]=1-\int\Phi(c_2)f_u(u)du
\end{aligned}$$

where

$$R = \frac{\mathbf{z}^T\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - E[\mathbf{z}^T\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]}{\big(\mathrm{var}(\mathbf{z}^T\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))\big)^{1/2}},$$

$$c_1 = \frac{-1}{2}\sqrt{\frac{u}{k}}c_0, \quad c_2 = \frac{1}{2}\sqrt{\frac{u}{k}}c_0,$$

$$c_0 = \big((\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\big)^{1/2}$$

The probability of misclassification associated with Bayes rule, denoted by Δ_B, is

$$\Delta_B = p_1P(2|1) + p_2P(1|2) = p_1\int \Phi(c_1)f_u(u)du + p_2\bigg(1 - \int \Phi(c_2)f_u(u)du\bigg) \quad (1.2.11)$$

where $p_1 + p_2 = 1$.

Multivariate Laplace distribution

Suppose the distribution of $\mathbf{X} \in \mathbb{R}^d$ is multivariate Laplace distribution $\mathcal{L}(\boldsymbol{\mu}, \mathbf{\Sigma})$, where $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ are mean and covariance of the distribution respectively. The probability density function of $\mathbf{x}$ is of the form

$$f(\mathbf{x}) \propto e^{-\sqrt{(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}}. \quad (1.2.12)$$

Without loss of generality, let $d = 2$, $r \sim \mathrm{Gamma}(d)$, $\theta \sim \mathrm{Uniform}(0, 2\pi)$, $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$. Define

$$Z_1 = r\cos\theta, \quad Z_2 = r\sin\theta, \quad \mathbf{Z} = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}.$$

Then, $\mathbf{X} = \mathbf{\Sigma}^{\frac{1}{2}}\mathbf{Z} + \boldsymbol{\mu}$ has bivariate Laplace distribution $BL(\boldsymbol{\mu}, \mathbf{\Sigma})$, where $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ are mean and covariance of the distribution respectively. It follows that $\mathbf{X} - \boldsymbol{\mu} = \mathbf{\Sigma}^{\frac{1}{2}}\mathbf{Z}$. Suppose populations π_1 and π_2 have distribution functions $BL(\boldsymbol{\mu}_1, \mathbf{\Sigma})$ and $BL(\boldsymbol{\mu}_2, \mathbf{\Sigma})$

respectively. If $\mathbf{x} \in \pi_1$, then $\mathbf{Z} = \boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{X} - \boldsymbol{\mu}_1)$,

$$\sqrt{(\mathbf{x} - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1)} = \sqrt{\mathbf{z}^T \mathbf{z}} = r \sim \text{Gamma}(d)$$

and $(\mathbf{x} - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_2) \neq r^2$ except $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, where $d = 2$. Similarly, if $\mathbf{x} \in \pi_2$, then $\mathbf{Z} = \boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{X} - \boldsymbol{\mu}_2)$,

$$\sqrt{(\mathbf{x} - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_2)} = \sqrt{\mathbf{z}^T \mathbf{z}} = r \sim \text{Gamma}(d)$$

and $(\mathbf{x} - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) \neq r^2$ except $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$. It follows that

$$\log\left(\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}\right) = -\sqrt{(\mathbf{x} - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1)} + \sqrt{(\mathbf{x} - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_2)}.$$

For $\mathbf{x}$, $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2 \in \mathbb{R}^d$ and $d \geqslant 2$, the separating hyperplane between π_1 and π_2 can be written as

$$(\mathbf{x} - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) = (\mathbf{x} - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_2).$$

This is equivalent to

$$\mathbf{x}^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

It follows that if $\mathbf{x}$ is distributed as population π_1,

$$\begin{aligned}
&\mathbf{x}^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \quad \text{implies} \\
&\mathbf{x}^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)
\end{aligned}$$

which gives

$$(\mathbf{x} - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = -\frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

This can be written as

$$\mathbf{z}^T\mathbf{\Sigma}^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = -\frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

which is the same as

$$\mathbf{z}^T\mathbf{a} = -\frac{1}{2}\mathbf{a}^T\mathbf{a}$$

where $\mathbf{a} = \mathbf{\Sigma}^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ and $\mathbf{z}$ is a standard multivariate Laplace distributed random variable. Kotz, Kozubowski and Podgorski (2001) has shown that linear combination of standard multivariate Laplace random variables has a univariate symmetric Laplace distribution $\mathcal{L}(0, \sigma_l)$ (See Proposition 5.1.1 in pp. 232). That is, $w = \mathbf{a}^T\mathbf{z}$ has a univariate Laplace distribution with mean 0 and variance σ_l, where $\sigma_l = \sqrt{\text{var}(\mathbf{a}^T\mathbf{z})}$ and $\mathbf{a}$ is a vector of constant real numbers. Similarly, $\mathbf{x}^T\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \boldsymbol{\mu}_2^T\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ gives $\mathbf{z}^T\mathbf{a} = \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ if $\mathbf{x}$ is distributed as population π_2.

Suppose $f_1(\mathbf{x}) > f_2(\mathbf{x})$, then $(\mathbf{x} - \boldsymbol{\mu}_1)^T\mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) < (\mathbf{x} - \boldsymbol{\mu}_2)^T\mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_2)$ and $\mathbf{x}^T\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) > \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)^T\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$. The probability of misclassifying $\mathbf{x}$, whose true population is π_1, into π_2 is

$$\begin{aligned}
&P(2|1) = P\big(\mathbf{x}^T\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) < \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)^T\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)|\mathbf{x} \in \pi_1\big)\\
&= P\big(\mathbf{x}^T\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \boldsymbol{\mu}_1^T\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) < \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)^T\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \boldsymbol{\mu}_1^T\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\big)\\
&= P\big(\mathbf{z}^T\mathbf{a} < -\frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\big) = P\big(w < -\frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \mu_2)\big)\\
&= F\bigg(-\frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\bigg),
\end{aligned}$$

where F is the distribution function of 1-dimensional symmetric Laplace distribution $\mathcal{L}\big(0, \sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}\big)$ with $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) > 0$. Similarly, the probability of misclassifying $\mathbf{x}$, whose true population is π_2, into π_1 is

$$
\begin{aligned}
&P(1|2) = P\big(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) > \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)|\mathbf{x} \in \pi_2\big) \\
&= P\big(\mathbf{x}^{'}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \boldsymbol{\mu}_2^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) > \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \boldsymbol{\mu}_2^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\big) \\
&= P\big(\mathbf{z}^T\mathbf{a} > \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\big) = P\big(w > \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\big) \\
&= 1 - F\bigg(\frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\bigg)
\end{aligned}
$$

where F is the distribution function of 1-dimensional symmetric Laplace distribution $\mathcal{L}\big(0, \sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}\big)$, with $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) > 0$. The Bayes probability of misclassifying of $\mathbf{x}$ into either π_1 or π_2, denoted by Δ_B, is

$$
\begin{aligned}
\Delta_B &= p_1P(2|1) + p_2P(1|2) \\
&= p_1F\bigg(- \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\bigg) + p_2\bigg[1 - F\bigg(\frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\bigg)\bigg]
\end{aligned}
$$

where $p_1 + p_2 = 1$. Suppose G is a Laplace distribution function which is symmetric about c, then $G(-c) = 1 - G(c)$ for all $c \in \mathbb{R}$. Hence

$$
\begin{aligned}
\Delta_B &= p_1F\bigg(- \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\bigg) + p_2\bigg[F\bigg(- \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\bigg)\bigg] \\
&= F\bigg(- \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\bigg)
\end{aligned}
$$

1.3 Nonparametric Classification Rules for Multivariate Data

Use of nonparametric approach for classifying observations has gained significant attention in the last two decades as it does not depend heavily on the underlying distributions. Nonparametric classification methods do not involve estimating moments of population distributions. One of their intuitive features is their robustness against outliers and extreme values. Various nonparametric classification approaches can be seen in the work of Cover and Hart (1967), Cover (1968), Vapnik (1982, 1998), Liu (1990), Cortes and Vapnik (1995), Liu, Parelius and Singh (1999), Jörnsten (2004), Ghosh and Chaudhuri (2005a, 2005b), Cui, Lin and Yang (2008), Li, Cuesta-Albertos and Liu (2012), Dutta and Ghosh (2012a, 2012b), among others.

1.3.1 Support Vector Machine for Multivariate Data

Support vector machine (SVM) is a popular method for classifying multivariate data. The foundation of Support Vector Machines (SVM) was developed by Vapnik (1982). Cortes and Vapnik (1995) upgraded this method from maximum margin idea to soft margin approach which enables the SVM to choose a boundary that splits data points as cleanly as possible, while still maximizing the distance to the nearest cleanly split data points. Suppose $(\mathbf{X}, y)$ is a pair of random variable in which y, class membership takes values in $\{-1, 1\}$ and $\mathbf{X} \in \aleph$, where $\aleph$ is a sample of training data points in $\mathbb{R}^d$, SVM aims at predicting the value of y given observed value $\mathbf{x}$. SVMs separate different classes of data by a hyperplane

$$\mathbf{w}^T\mathbf{x} + b = 0 \tag{1.3.1}$$

and the corresponding decision rule is

$$y(\mathbf{x}) = \text{sign}(\mathbf{w}^T\mathbf{x} + b), \tag{1.3.2}$$

where $\mathbf{w}$ is a finite dimensional vector to be estimated and b is a constant scalar. In order to obtain a best hyperplane, $\|\mathbf{w}\|$ is minimised subject to the decision rule.

1.3.2 Nearest neighbour rule

The k-nearest neighbour rule (k-NN) is another nonparametric method for classifying multivariate observations based on closest training observations in the data cloud. It is proposed in Cover and Hart (1967). This involves assigning an unclassified sample point to the class that is commonest amongst its k nearest neighbours, where k is a positive integer. Suppose $\mathbf{X}_{(i)}, i = 1, 2, \ldots, k$ are k nearest neighbours to $\mathbf{x}$, this classification rule is to assign $\mathbf{x}$ to the class that is commonest amongst its k nearest neighbours. k nearest neighbours are the k observations in the training sample with minimum distance from $\mathbf{x}$.

1.3.3 Depth Based Classifiers for Multivariate Data

Liu, Parelius and Singh (1999) defined data depth as a measure of the depth or centrality of d-dimensional observation $\mathbf{x}$ with respect to a multivariate data cloud or underlying multivariate distribution, F. It is denoted by $D(F, \mathbf{x})$. Data depth has some appealing characteristics. It helps to build systematic and nonparametric approach for generalising features and properties of univariate distributions to multivariate distributions. It characterises the centrality of a distribution and motivates nonparametric robust statistical methodologies. Depth functions include Mahalanobis depth (Mahalanobis, 1936), half-space depth or Tukey depth (Tukey, 1975), simplicial depth (Liu, 1990), projection depth (Donoho, 1982), Oja depth (Oja, 1983), simplicial volume depth (Zuo and Serfling, 2000a, 2000b), spatial depth (Vardi and Zhang, 2000), regression depth (Rousseeuw and Hubert, 1999). The possibility of using of data depth for classification was first raised in Liu (1990). Ghosh and Chaudhuri (2005a, 2005b) developed it into full-fledged nonparametric classification method called maximum depth classifier.

Ghosh and Chaudhuri (2005a) used half-space depth and regression depth to construct linear and nonlinear separating curves or surfaces. In those depth based methods, a finite dimensional parametric form for the separating surface is often assumed. Also, Ghosh and

Chaudhuri (2005b) proposed a nonparametric method called maximum depth classifier. Maximum depth classifier assigns observations to the population or sample for which the classifier attains its highest depth value. In a two class problem, suppose $D(F, \mathbf{x})$ and $D(G, \mathbf{x})$ are depths of $\mathbf{x}$ with respect to distributions F and G respectively. Maximum depth classifier is to assign $\mathbf{x}$ to F if

$$D(F, \mathbf{x}) \geqslant D(G, \mathbf{x}),$$

and to G otherwise. This method is fully nonparametric and readily lends itself to multiclass extension. They have shown that this classifier is the Bayes rule for the location shift problem. However, the performance of the classifier is affected by deviations from the location shift model or violation of monotonic nature of density functions. This limitation is overcome by modifying maximum depth classifier. The modified classifier is to assign $\mathbf{x}$ to population, π_j with

$$\max_{1 \leqslant j \leqslant J} p_j \theta_j \{D(F_j, \mathbf{x})\},$$

where $\theta_j\{D(F_j, \mathbf{x})\}$ is a function of $D(F_j, \mathbf{x})$, which is a depth of $\mathbf{x}$ with respect to j-th population distribution, F_j and p_j is the prior probability of j-th population. The function $\theta_j(.)$ varies from one type of depth to another. The modified method performs well when the populations differ in both location and scale in the case of elliptically symmetric distributions when the half-space depth is used. The modified method suffers computational difficulty of half-space depth when $d > 2$. Also, the method requires estimating several unknown parameters, some of which involve complicated estimation techniques (Li, Cuesta-Albertos and Liu, 2012). Similarly, on maximum depth classifier, Cui, Lin and Yang (2008) proposed maximum depth classifier based on modified projection depth as the depth function. Its result is appealing and works well only when samples are multivariate normally distributed. Dutta and Ghosh (2012a) suggested use of robust version of Mahalanobis depth and projection depth as depth functions for the purpose of maximum depth classification. Dutta and Ghosh (2012b) proposed a L_p depth classifier for l_p symmetric distributions, which assigns observations to the class for which the classifier possesses maximum L_p depth, where p is adaptively chosen using training data.

Possibility of using DD-plot, a two-dimensional representation of multivariate objects by their data depths with respect to two known classes, was raised in Liu, Parelius and Singh (1999). Li, Cuesta-Albertos and Liu (2012) proposed use of DD plot for classification. DD-classifier assigns observations to the population or sample with highest depth value. DD-classifier depends on the optimal choice of coefficient vector of polynomial function of Mahalanobis depth that minimises overall misclassification rate. The method is data driven, simple to visualise and easy to implement if the degree of polynomial is known. In practice, the degree of polynomial is unknown and its estimation involves complex optimisation, which may lead to trade off between prediction bias and prediction variance. According to Lange, Mosler and Mozharovskyi (2014), Mahalanobis depth does not reflect asymmetries of the data. Lange, Mosler and Mozharovskyi (2014) proposed DD_α-procedure based on zonoid depth and α-procedure algorithms. This method is an extension of DD-classifiers proposed in Li, Cuesta-Albertos and Liu (2012). The choice of α depends on minimiser of the average misclassification rate. The method is completely nonparametric. It uses q-dimensional depth plot to discriminate between classes in the depth space $[0,1]^q$. In case of more than two classes, several binary classifications are performed and a majority rule is applied.

1.4 Functional Classification Procedures

Functional data refer to data which consist of observed functions or curves evaluated at a finite subset of some interval. The word functional refers to the infinite dimensionality of the data. A random function denotes a random variable valued in an infinite dimension space (Ferraty and Vieu 2003, pp. 162). Functional data may be function of time or function of quantities. Definition 1.4.1 below is given in Ferraty and Vieu (2006) for functional random variable.

Definition 1.4.1 *A random variable $\mathcal{X}$ is called functional random variable if it takes values in an infinite dimensional space or functional space.*

That is, $\mathcal{X} = \{\mathcal{X}(t); t \in \mathcal{I}\}$, where $\mathcal{I} \subset \mathbb{R}$.

Definition 1.4.2 *A functional dataset $\{X_1, X_2, \ldots, X_n\}$ is a dataset generated by n identically distributed functional variables $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n$.*

Suppose $X(t_i)$ is the value of X at t_i. For $i \in \{1, 2, \ldots, d\}$, $X(t_i) = \{X(t_1), X(t_2), \ldots, X(t_d)\}$ is a finitely observed functional datum if X is a functional data generated by a real functional random variable $\mathcal{X}$.

Definition 1.4.3 *Suppose $X_i(t_{ij})$ is the value of X_i at t_{ij}. For $i \in \{1, 2, \ldots, n\}$ and $j \in \{1, 2, \ldots, d_i\}$,*

$$X_i(t_{ij}) = \big\{\{X_1(t_{11}), X_1(t_{12}), \ldots, X_1(t_{1d_1})\}, \ldots, \{X_n(t_{n1}), X_n(t_{n2}), \ldots, X_n(t_{nd_n})\}\big\}$$

is a finitely observed functional dataset if $\{X_1(t), X_2(t), \ldots, X_n(t)\}$ is a functional dataset generated by a real functional random variables $\{\mathcal{X}_1(t), \mathcal{X}_2(t), \ldots, \mathcal{X}_n(t)\}$.

An example of functional data is a population P_0 consisting of trajectories of the process $X(t) = m_0(t) + e(t)$, where $m_0(t) = 2.5|\sin(10\pi t)|$ and $e(t)$ is a Gaussian process with mean 0 and $\text{cov}(X(s), X(t)) = \exp(-|s - t|)$. Figure 1.5 gives the plot of mean function of the population P_0.

Functional data are always highly correlated (for example, micro-array data and clinical outcomes), which results in singularity of the covariance estimates of such data. In

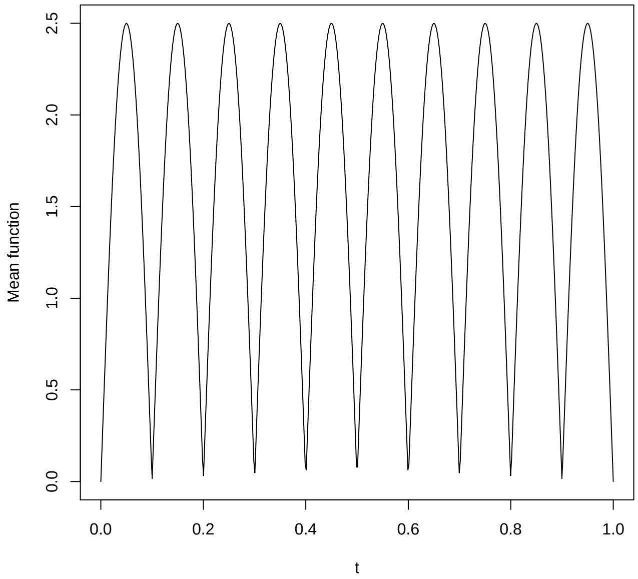


Figure 1.5: Mean function of a population of trajectories.

$\mathbb{R}^d$, the density functions allow us to easily characterize distributions, calculate probabilities and moments and define likelihood functions but for functional data, there is no simple natural way to define and calculate density functions in infinite-dimensional spaces (Cuevas, 2014). The applications of functional data are very useful in medicine, crime analysis, signal processing, chemometrics, among others. Examples of real functional data include growth data, tumors identification and differentiation data, LSVT voice rehabilitation data, among others. Functional data can exist as univariate or multivariate. We refer readers to Claeskens et al. (2014), Ferraty and Vieu (2006) and Ramsay and Silverman (2005) for detail.

In classifying functional data, different classification procedures have emerged since last two decades for functional data. These include different forms of linear discriminant analysis for functional data (Dudoit, Fridlyand and Speed, 2002; James and Hastie, 2001 and Preda, Saporta and leveder, 2007), classifiers based on kernel estimators of posterior probabilities (Hall, Poskitt and Presnell, 2001; Ferraty and Vieu, 2003), classifier based on a distance measure(Vilar and Pertega, 2004), model-based classifiers (Leng and M$\ddot{u}$ller, 2006), nearest neighbour classification rule for functional data (Biau, Bunea and Wegkamp, 2005; Cover and Hart, 1967), weighted distance approach (Alonso, Casado and Romo, 2012), Support Vector Machines (Vapnik, 1998; Cortes and Vapnik, 1995; Rossi and Villa, 2006; Li and Yu, 2008), classifiers based on shape descriptors (Epifanio, 2008), classification method based on distance to class centroid or its trimmed version(Delaigle and Hall, 2012a; L$\acute{o}$pez-Pintado and Romo, 2006; Cuesta-Albertos and Nieto-Reyes, 2010), classifiers based on functional mixed model (Zhu, Brown and Morris, 2012) and maximum depth classifiers (Cuevas, Febrero and Fraiman, 2007), among others.

1.5 Current Work

In this thesis, we propose some classification methods based on multivariate ranks and distance based rules for multivariate and functional data respectively, and study properties of each of the classifiers. In chapter two, we propose a nonparametric classification method based on multivariate rank and refer to it as minimal rank classifier. We show that it is Bayes procedure under suitable conditions. The variations in total probability of misclassification of d-dimensional observations from two classes of populations with different location vectors are considered separately as well as cases of homogenous and heterogeneous scales. It is well known that multivariate rank is not invariant under arbitrary affine transformations, so it may be affected by deviation of population distribution from spherical symmetry. Also, we investigate the performance of minimal rank classifier under location shift, accounting for the effect of deviation from spherical symmetry, scale shift and location-scale shift. Based on the effect of deviation from spherical symmetry on minimal rank classifier, we introduce a way of constructing affine invariant multivariate rank. Using the affine invariant multivariate rank, we transform the classification method to affine invariant version and study its statistical properties for location shift problem. In chapter three, we propose a classifier based on volume of rank region for location-scale shift problem and study its properties. The improved version of this method is also proposed in the chapter. Chapter four consists of nonparametric classification methods based on distribution function of outlyingness of multivariate rank and its invariants. When data are functions, many multivariate techniques fail to perform well. Classification method based on L_2 distance to functional medians are proposed and generalised into L_p distance in chapter five. Conclusion and areas of further research are presented in chapter six.

Chapter 2

Rank Classifiers for Multivariate Data

2.1 Multivariate Rank

Signs and ranks are commonly used in statistical methodology to develop methods or procedures that are independent of distribution assumptions. Use of rank for computing statistical quantities gives robust estimators (e.g. estimator for location) as they are not affected by the presence of outlying values in the data. For the univariate data, sign of $x \in \mathbb{R}$ can be defined as

$$\operatorname{sign}(x) = \begin{cases} -1, & \text{if } x < 0 \\ 0, & \text{if } x = 0 \\ 1, & \text{if } x > 0, \end{cases}$$

or equivalently,

$$\operatorname{sign}(x) = \begin{cases} \frac{x}{|x|}, & \text{if } x \neq 0 \\ 0, & \text{if } x = 0. \end{cases}$$

Univariate centred rank of x with respect to data points $X_1, X_2, \ldots, X_n$ from distribution F can be defined as

$$\text{rank}(x) = \frac{1}{n} \sum_{i=1}^{n} \text{sign}(x - X_i).$$

Following are some of the basic properties of rank(x).

1. $|\text{rank}(x)| \leqslant 1$.

2. $\text{rank}(x) = 0$ implies x is the median.

3. $|\text{rank}(x)| = 1$ implies x is an extreme point.

4. $E[\text{rank}(x)] = 2F(x) - 1$.

These properties suggest that rank(x) is not only a useful descriptive statistics, it also characterises the distribution.

Now, we want to define sign and rank functions in a multivariate set up following Chakraborty (2001). Suppose $\mathbf{x} \in \mathbb{R}^d$, then the l_p sign of $\mathbf{x}$ is

$$\text{sign}_p(\mathbf{x}) = \begin{cases} \frac{\partial}{\partial \mathbf{x}} ||\mathbf{x}||_p = \frac{v(\mathbf{x})}{||\mathbf{x}||_p^{p-1}}, & \text{if } \mathbf{x} \neq \mathbf{0} \\ \mathbf{0}, & \text{if } \mathbf{x} = \mathbf{0} \end{cases}$$

where $||\mathbf{x}||_p = \left\{ x_1^p + x_2^p + \ldots + x_d^p \right\}^{\frac{1}{p}}$ and

$$v(\mathbf{x}) = \left(\text{sign}(x_1)|x_1|^{p-1}, \text{sign}(x_2)|x_2|^{p-1}, \ldots, \text{sign}(x_d)|x_d|^{p-1} \right)^T.$$

The l_p rank of $\mathbf{x} \in \mathbb{R}^d$ with respect to data points $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n \in \mathbb{R}^d$ is defined as

$$\text{rank}_p(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \text{sign}_p(\mathbf{x} - \mathbf{X}_i).$$

When $p = 1$, $\text{sign}_1(\mathbf{x}) = \big(\text{sign}(x_1), \text{sign}(x_2), \ldots, \text{sign}(x_d)\big)^T$, the vector of coordinatewise signs and for $p = 2$,

$$\text{sign}_2(\mathbf{x}) = \frac{\mathbf{x}}{||\mathbf{x}||_2}$$

where $||.||_2$ is the Euclidean norm, $||\mathbf{y}||_2 = \left\{y_1^2 + y_2^2 + \ldots + y_d^2\right\}^{\frac{1}{2}}$. $\text{sign}_2(\mathbf{x})$ is called the spatial sign vector.

Definition 2.1.1 *Suppose $\mathbf{X} \in \mathbb{R}^d$ has a d-dimensional distribution F. The multivariate rank function of any point $\mathbf{x} \in \mathbb{R}^d$ with respect to F is defined as*

$$\text{rank}_F(\mathbf{x}) = E_F\left(\text{sign}_2(\mathbf{x} - \mathbf{X})\right) = E_F\left[\frac{\mathbf{x} - \mathbf{X}}{||\mathbf{x} - \mathbf{X}||_2}\right]. \tag{2.1.1}$$

This is also known as spatial rank vector (Möttönen and Oja, 1995). In a similar way, we can define sample version of the spatial rank vector.

Definition 2.1.2 *Suppose $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n \in \mathbb{R}^d$ is a random sample from a distribution function, F on $\mathbb{R}^d$. The spatial rank of $\mathbf{x} \in \mathbb{R}^d$ with respect to $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ is defined as*

$$\text{rank}_{F_n}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \text{sign}_2(\mathbf{x} - \mathbf{X}_i) = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbf{x} - \mathbf{X}_i}{||\mathbf{x} - \mathbf{X}_i||_2} \tag{2.1.2}$$

where $||\mathbf{x} - \mathbf{X}_i||_2 \neq \mathbf{0}$, for all $i = 1, 2, ..., n$.

If $\text{rank}_F(\mathbf{x}) = \mathbf{0}$, then $\mathbf{x}$ is the spatial median. From now on, we will use $||.||$ to denote the Euclidean norm $||.||_2$, whenever there is no scope of confusion. Let $||\text{rank}_F(\mathbf{x})||$ denotes the measure of outlyingness of $\text{rank}_F(\mathbf{x})$. $||\text{rank}_F(\mathbf{x})||$ is invariant under location shift or translation (that is, $||\text{rank}_F(\mathbf{x} + \boldsymbol{\theta})|| = ||\text{rank}_F(\mathbf{x})||$ for a constant vector $\boldsymbol{\theta}$) and under orthogonal scale transformation (that is, $||\text{rank}_F(\mathbf{A}\mathbf{x})|| = ||\text{rank}_F(\mathbf{x})||$ for an orthogonal matrix $\mathbf{A}$). Spatial rank helps determine the geometric position of points in $\mathbb{R}^d$ with respect to the data cloud, and hence can be viewed as a descriptive statistic (Guha and Chakraborty, 2013).

Suppose F is spherically symmetric and characterised by location parameter $\boldsymbol{\theta} \in \mathbb{R}^d$, $\|\text{rank}_F(\mathbf{x})\|$ increases as $\|\mathbf{x} - \boldsymbol{\theta}\|$ increases. This is stated formally by the theorem below.

Theorem 2.1.1 (Guha and Chakraborty, 2013) *If F is a spherically symmetric distribution in $\mathbb{R}^d$ with $\boldsymbol{\theta}$ as the centre of symmetry, then for any $\mathbf{x} \in \mathbb{R}^d$,*

$$\text{rank}_F(\mathbf{x}) = q(||\mathbf{x} - \boldsymbol{\theta}||)\frac{\mathbf{x} - \boldsymbol{\theta}}{||\mathbf{x} - \boldsymbol{\theta}||} \tag{2.1.3}$$

for some increasing, non-negative function q.

Following Theorem 2.1.1, we observe that $||\text{rank}_F(\mathbf{x})|| = q(||\mathbf{x} - \boldsymbol{\theta}||)$ and $\|\text{rank}_F(\mathbf{x})\|$ increases as $\|\mathbf{x} - \boldsymbol{\theta}\|$ increases. The implication of this is that smaller rank indicates more central observation and larger rank indicates extreme observation.

Chaudhuri (1996) defined spatial quantiles as vectors in $\mathbb{R}^d$ that are indexed by a vector $\mathbf{u}$ in d-dimensional unit ball. Define an open ball $B^{(d)} = \{\mathbf{u} | \mathbf{u} \in \mathbb{R}^d, ||\mathbf{u}|| < 1\}$. For any $\mathbf{u} \in B^{(d)}$ and $\mathbf{t} \in \mathbb{R}^d$, define $\Phi(\mathbf{u}, \mathbf{t}) = ||\mathbf{t}|| + < \mathbf{u}, \mathbf{t} >$, where $< .,. >$ denotes the usual Euclidean inner product. Spatial quantile corresponding to $\mathbf{u}$ and based on $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n \in \mathbb{R}^d$ is defined as

$$\widehat{Q}_n(\mathbf{u}) = \arg\min_{Q \in \mathbb{R}^d} \sum_{i=1}^{n} \Phi(\mathbf{u}, \mathbf{X}_i - Q).$$

It follows from Theorem 1.1.2 of Chaudhuri (1996) that

$$\sum_{i=1}^{n} \frac{\mathbf{X}_i - \widehat{Q}_n(\mathbf{u})}{||\mathbf{X}_i - \widehat{Q}_n(\mathbf{u})||} + n\mathbf{u} = \mathbf{0},$$

if $\widehat{Q}_n(\mathbf{u}) \neq \mathbf{X}_i$ for all $1 \leqslant i \leqslant n$. This implies

$$\mathbf{u} = \frac{1}{n} \sum_{i=1}^{n} \frac{\widehat{Q}_n(\mathbf{u}) - \mathbf{X}_i}{||\widehat{Q}_n(\mathbf{u}) - \mathbf{X}_i||}. \tag{2.1.4}$$

Serfling (2004) defined $\text{rank}_{F_n}(\mathbf{x})$ as the inverse function of the spatial quantile function, $\widehat{Q}_n(\mathbf{u})$. Mathematically, we can write equation (2.1.4) as

$$\mathbf{u} = \text{rank}_{F_n}\big(\widehat{Q}_n(\mathbf{u})\big) = \text{rank}_{F_n}(\mathbf{x})$$

and so

$$\widehat{Q}_n(\mathbf{u}) = \mathbf{x} \quad \text{implies} \;\; \text{rank}_{F_n}(\mathbf{x}) = \mathbf{u}.$$

2.2 Minimal Rank Classifier

In this study, we propose a nonparametric classification method based on spatial ranks of the d-dimensional observations with respect to multivariate data clouds. In a two class problem, let π_1 and π_2 denote two populations with distributions F and G respectively with equal prior probabilities, where F and G are absolutely continuous with respect to Lebesgue measure in $\mathbb{R}^d$. The classification rule is to assign an observation $\mathbf{x}$, into population π_1 if

$$||\text{rank}_F(\mathbf{x})|| \leqslant ||\text{rank}_G(\mathbf{x})|| \tag{2.2.1}$$

otherwise, assign $\mathbf{x}$ to population π_2. If there are $J(\geqslant 2)$ populations, then assign $\mathbf{x}$ to population π_k, $1 \leqslant k \leqslant J$ if

$$||\text{rank}_{F_k}(\mathbf{x})|| = \min_j ||\text{rank}_{F_j}(\mathbf{x})|| \tag{2.2.2}$$

where $F_1, F_2, \ldots, F_J$ are absolutely continuous distributions corresponding to J populations.

Note that $P(||\text{rank}_F(\mathbf{x})|| > ||\text{rank}_G(\mathbf{x})|| \;\mid\; \mathbf{x} \in \pi_1)$ is the probability of assigning $\mathbf{x}$ into population π_2 when true population of $\mathbf{x}$ is π_1 and $P(||\text{rank}_F(\mathbf{x})|| \leqslant ||\text{rank}_G(\mathbf{x})|| \;\mid\; \mathbf{x} \in \pi_2)$ is the probability of assigning $\mathbf{x}$ into population π_1 when true population of $\mathbf{x}$ is π_2. Then the total probability of misclassification corresponding to two populations, π_1 and π_2, denoted by Δ, is

$$\Delta = p_1 P(||\text{rank}_F(\mathbf{x})|| > ||\text{rank}_G(\mathbf{x})|| \ \mid \ \mathbf{x} \in \pi_1) + p_2 P(||\text{rank}_F(\mathbf{x})|| \leqslant ||\text{rank}_G(\mathbf{x})|| \ \mid \ \mathbf{x} \in \pi_2)$$

with prior probabilities p_1 and p_2 for π_1 and π_2 respectively. For the case of J populations with prior probabilities $p_1, p_2, \ldots, p_J$, the total probability of misclassification is

$$\Delta = \sum_{j=1}^{J} p_j P(\|\text{rank}_{F_j}(\mathbf{x})\| \ \text{ is not the minimum } \ \mid \ \mathbf{x} \in j\text{th population}).$$

2.2.1 Properties of Minimal Rank Classifier

In this section, we shall show some properties of minimal rank classifier under suitable conditions. Theorem 2.2.1 shows that minimal rank classifier is equivalent to Bayes rule under some conditions.

Theorem 2.2.1 *Let f_1 and f_2 be the probability density functions of populations, π_1 and π_2 having spherically symmetric distributions F and G in $\mathbb{R}^d$ about $\boldsymbol{\theta}_F$ and $\boldsymbol{\theta}_G$ respectively with equal prior probabilities $p_1 = p_2 = \frac{1}{2}$, then the Bayes rule is*

$$\textit{assign } \ \mathbf{x} \ \textit{ to population } \ \pi_1 \ \textit{ if } \ ||\text{rank}_F(\mathbf{x})|| \leqslant h\big(||\text{rank}_G(\mathbf{x})||\big)$$

and

$$\textit{assign } \ \mathbf{x} \ \textit{ to population } \ \pi_2 \ \textit{ if } \ ||\text{rank}_F(\mathbf{x})|| > h\big(||\text{rank}_G(\mathbf{x})||\big)$$

where h is a non-negative function.

Proof: For an absolutely continuous spherically symmetric distribution F about $\boldsymbol{\theta}_F \in \mathbb{R}^d$, the probability density function can be written as

$$f_1(\mathbf{x}) = g_1\left(\|\mathbf{x} - \boldsymbol{\theta}_F\|\right)$$

for some non-negative real-valued decreasing function g_1 and similarly, the probability density function of G can be written as

$$f_2(\mathbf{x}) = g_2\left(\|\mathbf{x} - \boldsymbol{\theta}_G\|\right)$$

for some non-negative real-valued decreasing function g_2. Then $f_1(\mathbf{x}) = g_1\left(\|\mathbf{x} - \boldsymbol{\theta}_F\|\right)$ and $f_2(\mathbf{x}) = g_2\left(\|\mathbf{x} - \boldsymbol{\theta}_G\|\right)$.

Now by Theorem 2.1.1, we know that

$$\mathrm{rank}_F(\mathbf{x}) = h_1(||\mathbf{x} - \boldsymbol{\theta}_F||)\frac{\mathbf{x} - \boldsymbol{\theta}_F}{||\mathbf{x} - \boldsymbol{\theta}_F||}$$

and

$$\mathrm{rank}_G(\mathbf{x}) = h_2(||\mathbf{x} - \boldsymbol{\theta}_G||)\frac{\mathbf{x} - \boldsymbol{\theta}_G}{||\mathbf{x} - \boldsymbol{\theta}_G||}$$

for some increasing functions h_1 and h_2. This implies $\|\mathrm{rank}_F(\mathbf{x})\| = h_1(\|\mathbf{x} - \theta_F\|)$ and $\|\mathrm{rank}_G(\mathbf{x})\| = h_2(\|\mathbf{x} - \boldsymbol{\theta}_G\|)$. Therefore, we can write $f_1(\mathbf{x}) = g_1\bigl(h_1^{-1}(\|\mathrm{rank}_F(\mathbf{x})\|)\bigr)$ and $f_2(\mathbf{x}) = g_2\bigl(h_2^{-1}(\|\mathrm{rank}_G(\mathbf{x})\|)\bigr)$. Now,

$$f_1(\mathbf{x}) \geqslant f_2(\mathbf{x}) \Leftrightarrow g_1\bigl(h_1^{-1}(\|\mathrm{rank}_F(\mathbf{x})\|)\bigr) \geqslant g_2\bigl(h_2^{-1}(\|\mathrm{rank}_G(\mathbf{x})\|)\bigr).$$

Since h_1 and h_2 are increasing functions and g_1 and g_2 are decreasing functions, $g_1 \circ h_1^{-1}$ and $g_2 \circ h_2^{-1}$ are decreasing functions. The proof is complete. □

Suppose the two populations π_1 and π_2 are separated by location only, that is, $G(\mathbf{x}) = F(\mathbf{x} - \boldsymbol{\theta})$, then we will have $g_1 = g_2$ and $h_1 = h_2$ and as a result, we get $||\text{rank}_F(\mathbf{x})|| \leqslant ||\text{rank}_G(\mathbf{x})||$ when $f_1(\mathbf{x}) \geqslant f_2(\mathbf{x})$. This result is formally stated in Corollary 2.2.1 below.

Corollary 2.2.1 *Let the populations π_1 and π_2 have spherically symmetric distributions F and G in $\mathbb{R}^d$ respectively with equal prior probabilities $p_1 = p_2 = \frac{1}{2}$ such that $G(\mathbf{x}) = F(\mathbf{x} - \boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \mathbb{R}^d$, then the Bayes rule is*

$$\begin{aligned} &\textit{assign } \mathbf{x} \textit{ to } \pi_1 \textit{ if } ||\text{rank}_F(\mathbf{x})|| \leqslant ||\text{rank}_G(\mathbf{x})|| \\ &\qquad\qquad\qquad \textit{and assign } \mathbf{x} \textit{ to } \pi_2 \textit{ otherwise.} \end{aligned}$$

If training samples $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m$ from π_1 and $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n$ from π_2 are available, then we replace the population version of the rank functions rank_F and rank_G by their empirical versions rank_{F_m} and rank_{G_n} respectively to construct the empirical classification rule, assign $\mathbf{x}$ to π_1 if $||\text{rank}_{F_m}(\mathbf{x})|| \leqslant ||\text{rank}_{G_n}(\mathbf{x})||$ and assign $\mathbf{x}$ to population π_2 otherwise.

Theorem 2.2.2 *Suppose F is a d-variate distribution function, which is absolutely continuous, then for sufficiently large n*

$$\sup_x \big| \, ||\text{rank}_{F_n}(\mathbf{x})|| - ||\text{rank}_F(\mathbf{x})|| \, \big| \overset{a.s.}{\to} 0.$$

Proof: Suppose $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ is random sample with empirical distribution function F_n, then for $j = 1, \ldots, n$, $\text{sign}(\mathbf{x} - \mathbf{X}_j)$ are independent and identically distributed and bounded. We know that $||\mathbf{a}-\mathbf{b}|| \leqslant ||\mathbf{a}||+||\mathbf{b}||$ and $||\mathbf{a}||-||\mathbf{b}|| \leqslant ||\mathbf{a}-\mathbf{b}||$. For any $\varepsilon > 0$,

$$\begin{aligned}
&\Big|\ ||\mathrm{rank}_{F_n}(\mathbf{x})|| - ||\mathrm{rank}_F(\mathbf{x})||\ \Big| > \varepsilon \\
&\Rightarrow ||\mathrm{rank}_{F_n}(\mathbf{x}) - \mathrm{rank}_F(\mathbf{x})||^2 \geqslant \Big|\ ||\mathrm{rank}_{F_n}(\mathbf{x})|| - ||\mathrm{rank}_F(\mathbf{x})||\ \Big|^2 > \varepsilon^2 \\
&\Rightarrow \sum_{i=1}^{d} \left|\mathrm{rank}_{F_n}(\mathbf{x})_i - \mathrm{rank}_F(\mathbf{x})_i\right|^2 > \varepsilon^2 \\
&\Rightarrow \ \text{at least one of } \left|\mathrm{rank}_{F_n}(\mathbf{x})_i - \mathrm{rank}_F(\mathbf{x})_i\right|^2 > \frac{\varepsilon^2}{d}
\end{aligned}$$

where $\mathrm{rank}_{F_n}(\mathbf{x})_i$ is the ith feature of $\mathrm{rank}_{F_n}(\mathbf{x})$. Therefore

$$P\left(\Big|\ ||\mathrm{rank}_{F_n}(\mathbf{x})|| - ||\mathrm{rank}_F(\mathbf{x})||\ \Big| > \varepsilon\right) \leqslant \sum_{i=1}^{d} P\left(\left|\mathrm{rank}_{F_n}(\mathbf{x})_i - \mathrm{rank}_F(\mathbf{x})_i\right|^2 > \frac{\varepsilon^2}{d}\right)$$

$$P\left(\Big|\ ||\mathrm{rank}_{F_n}(\mathbf{x})|| - ||\mathrm{rank}_F(\mathbf{x})||\ \Big| > \varepsilon\right) \leqslant \sum_{i=1}^{d} P\left(\left|\mathrm{rank}_{F_n}(\mathbf{x})_i - \mathrm{rank}_F(\mathbf{x})_i\right| > \frac{\varepsilon}{\sqrt{d}}\right)$$

By Hoeffding lemma, $P\left(\left|\mathrm{rank}_{F_n}(\mathbf{x})_i - \mathrm{rank}_F(\mathbf{x})_i\right| > \frac{\varepsilon}{\sqrt{d}}\right) \leqslant e^{-n\varepsilon^2/d}$. So,

$$P\left(\Big|\ ||\mathrm{rank}_{F_n}(\mathbf{x})|| - ||\mathrm{rank}_F(\mathbf{x})||\ \Big| > \varepsilon\right) \leqslant de^{-n\varepsilon^2/d}$$

Distribution F is defined on $\mathbb{R}^d$ for $d \geqslant 2$, then using Kiefer theorem (see Kiefer, 1961) for any $\delta > 0$,

$$P\left(\sup_x \Big|\ ||\mathrm{rank}_{F_n}(\mathbf{x})|| - ||\mathrm{rank}_F(\mathbf{x})||\ \Big| > \varepsilon\right) \leqslant ce^{-(2-\delta)n\varepsilon^2}$$

where c is a positive constant depending on δ and d only and not on F, then as $n \to \infty$,

$$P\left(\sup_x \Big|\ ||\mathrm{rank}_{F_n}(\mathbf{x})|| - ||\mathrm{rank}_F(\mathbf{x})||\ \Big| > \varepsilon\right) \to 0.$$

This implies

$$\sup_x \Big| \; ||\text{rank}_{F_n}(\mathbf{x})|| - ||\text{rank}_F(\mathbf{x})|| \; \Big| \overset{a.s.}{\to} 0 \text{ as } n \to \infty.$$

The proof is complete. □

Remark : Theorem 2.2.2. shows that $||\text{rank}_{F_n}(\mathbf{x})||$ converges to its population version almost surely.

Theorem 2.2.3 *Suppose f_1 and f_2 are the probability density functions of populations, π_1 and π_2 having spherically symmetric distributions F and G in $\mathbb{R}^d$ about $\boldsymbol{\theta}_F$ and $\boldsymbol{\theta}_G$ respectively with equal prior probabilities $p_1 = p_2 = \frac{1}{2}$, $\text{rank}_F(\mathbf{x})$ and $\text{rank}_G(\mathbf{x})$ are continuous and satisfies*

$$||\text{rank}_{F_{n_1}}(\mathbf{x})|| \overset{a.s.}{\to} ||\text{rank}_F(\mathbf{x})||, \;\; ||\text{rank}_{G_{n_2}}(\mathbf{x})|| \overset{a.s.}{\to} ||\text{rank}_G(\mathbf{x})||$$

as $min(n_1, n_2) \to \infty$, then the total probability of misclassification for the classification rule based on the training sample,

$$\begin{aligned}\Delta_n = &\frac{1}{2} P\left(||\text{rank}_{F_{n_1}}(\mathbf{x})|| > ||\text{rank}_{G_{n_2}}(\mathbf{x})|| \;\big|\; \mathbf{x} \in F\right) \\ &+ \frac{1}{2} P\left(||\text{rank}_{F_{n_1}}(\mathbf{x})|| \leqslant ||\text{rank}_{G_{n_2}}(\mathbf{x})|| \;\big|\; \mathbf{x} \in G\right)\end{aligned}$$

converges to optimal Bayes risk for sufficiently large n_1 and n_2 such that $min(n_1, n_2) \to \infty$ and $\frac{n_1}{n_1+n_2} \to \frac{1}{2}$.

Proof: For $p_1 = p_2 = \frac{1}{2}$, the total probability of misclassification for the Bayes rule is

$$\Delta = \frac{1}{2} P\left(||\text{rank}_F(\mathbf{x})|| > ||\text{rank}_G(\mathbf{x})|| \,\big|\mathbf{x} \in F\right) + \frac{1}{2} P\left(||\text{rank}_F(\mathbf{x})|| \leqslant ||\text{rank}_G(\mathbf{x})|| \,\big|\mathbf{x} \in G\right).$$

Now,

$$
\begin{aligned}
&|\Delta_n - \Delta| \\
&= \frac{1}{2}\Bigg| \int [I_{\{||\mathrm{rank}_{F_{n_1}}(\mathbf{y})||>||\mathrm{rank}_{G_{n_2}}(\mathbf{y})|| \ \big|\mathbf{y}\in F\}} - I_{\{||\mathrm{Rank}_F(\mathbf{x})||\leqslant||\mathrm{Rank}_G(\mathbf{x})|| \ \big|\mathbf{y}\in F\}}] f_1(\mathbf{y})d(\mathbf{y}) \\
&+ \int [I_{\{||\mathrm{rank}_{F_{n_1}}(\mathbf{y})||\leqslant||\mathrm{rank}_{G_{n_2}}(\mathbf{y})|| \ \big|\mathbf{y}\in G\}} - I_{\{||\mathrm{rank}_F(\mathbf{y})||\leqslant||\mathrm{rank}_G(\mathbf{y})|| \ \big|\mathbf{y}\in G\}}] f_2(\mathbf{y})d(\mathbf{y})\Bigg| \\
&\leqslant \frac{1}{2}\int \Big|I_{\{||\mathrm{rank}_{F_{n_1}}(\mathbf{y})||>||\mathrm{rank}_{G_{n_2}}(\mathbf{y})||\big|\mathbf{y}\in F\}} - I_{\{||\mathrm{rank}_F(\mathbf{y})||\leqslant||\mathrm{rank}_G(\mathbf{y})||\big|\mathbf{y}\in F\}}\Big| f_1(\mathbf{y})d(\mathbf{y}) \\
&+ \frac{1}{2}\int \Big|I_{\{||\mathrm{rank}_{F_{n_1}}(\mathbf{y})||\leqslant||\mathrm{rank}_{G_{n_2}}(\mathbf{y})||\big|\mathbf{y}\in G\}} - I_{\{||\mathrm{rank}_F(\mathbf{y})||\leqslant||\mathrm{rank}_G(\mathbf{y})||\big|\mathbf{y}\in G\}}\Big| f_2(\mathbf{y})d(\mathbf{y}).
\end{aligned}
$$

By Theorem 2.2.2 and Lebesgue dominated convergence theorem, each of the above integrals converges to zero and hence we have

$$\lim_{min(n_1,n_2)\to\infty} |\Delta_n - \Delta| \to 0.$$

□

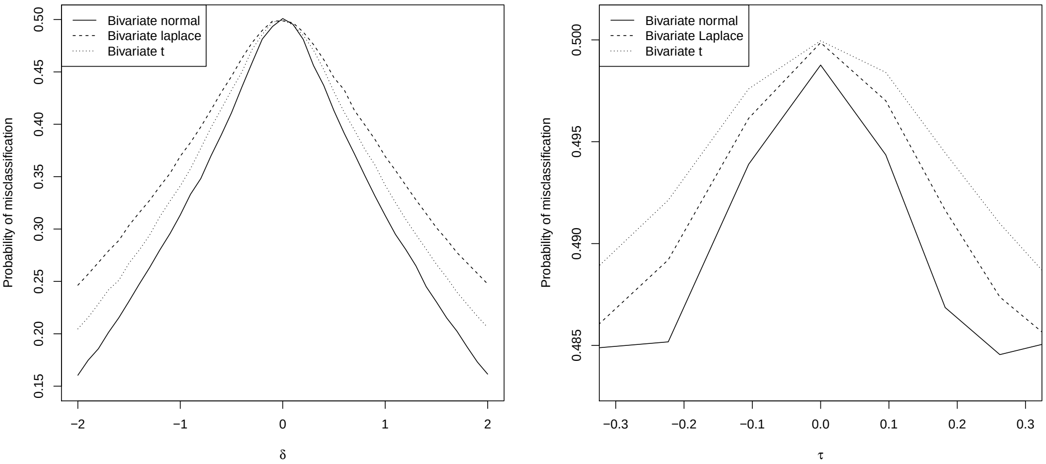
2.3 Numerical Examples

In this section, we shall simulate data for three spherically symmetric distributions and study the finite sample performance of our proposed classifier. Let π_1 and π_2 be two d-variate normal populations with mean vectors $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and covariance matrix $\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_2$, respectively. Assume that the prior probabilities of the populations are equal ($p_1 = p_2 = 0.5$) and costs of misclassifcations are also equal ($c(1|2) = c(2|1)$, as defined in Chapter one). Consider the following simulation study:

1. Generate $\mathbf{X}_1, \ldots, \mathbf{X}_n$ from π_1 and $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ from π_2.

2. Generate $\mathbf{Z}_1, \ldots, \mathbf{Z}_m$ from π_1 and $\mathbf{Z}_{m+1}, \ldots, \mathbf{Z}_{2m}$ from π_2.

3. Using the proposed classification rule, classify the observations $\mathbf{Z}_1, \ldots, \mathbf{Z}_{2m}$ and



(a) Location shift only

(b) Scale shift only

Figure 2.1: Misclassification rates associated with minimal rank classifier for 3 different families of distributions for (a) location shift only ($\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{I}_2, \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$) and (b) scale shift only ($\mathbf{\Sigma}_1 = \mathbf{I}_2, \mathbf{\Sigma}_2 = \sigma^2\mathbf{I}_2, \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, $\tau = \log_e \sigma$).

count the number of misclassified observations (say, l) and then estimate the probability of misclassification as $l/2m$.

4. Repeat the process $N = 1000$ times and compute the average of N number of estimates of misclassification probability obtained, determine the estimated total probability of misclassification.

2.3.1 Example 1: Location shift

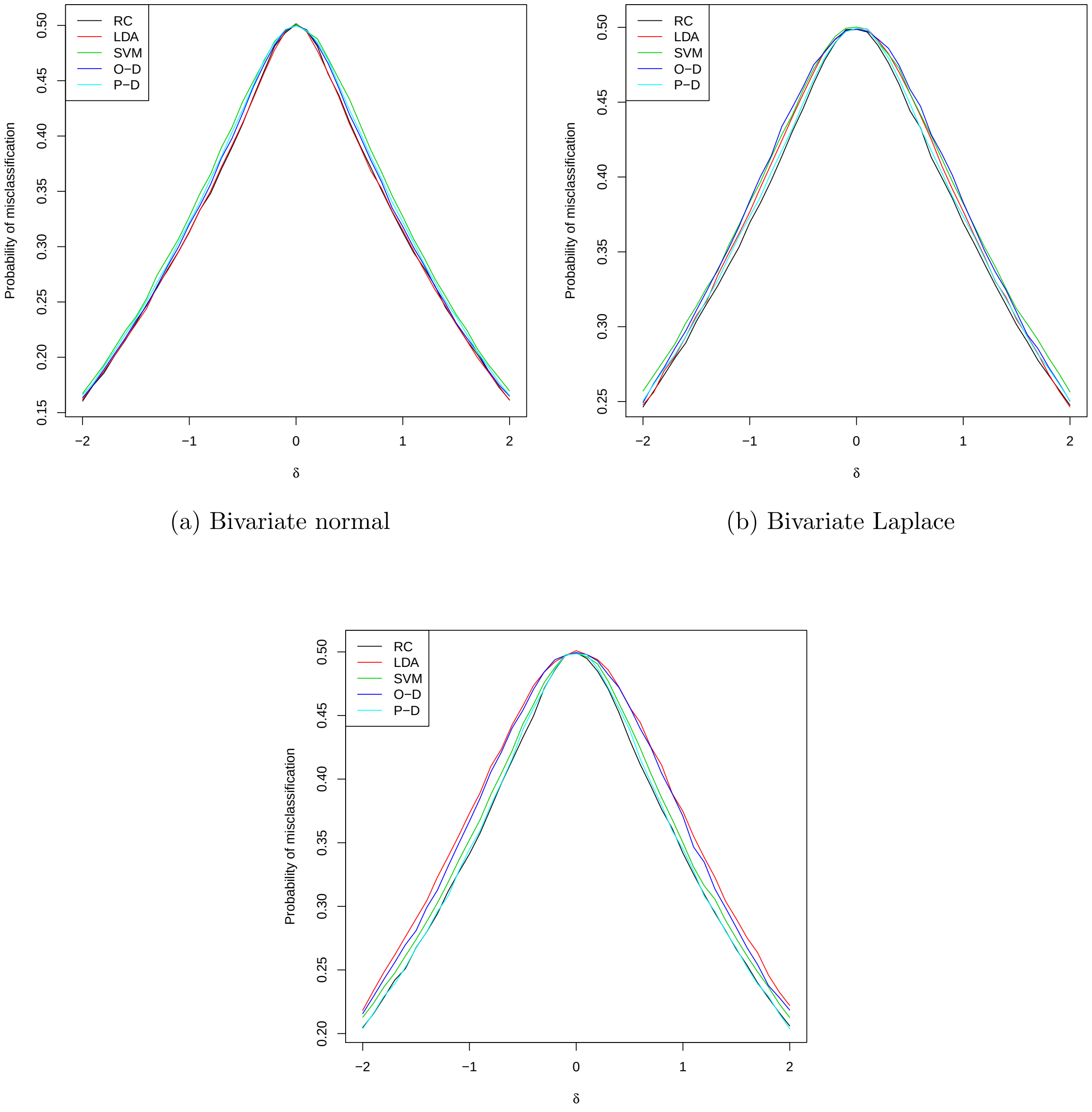
Consider

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \ \boldsymbol{\mu}_2 = \begin{pmatrix} \delta \\ 0 \end{pmatrix}, \ n = m = 100, \ \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{I}_2,$$

where $\mathbf{I}_2$ is a 2×2 identity matrix. Suppose π_1 has a distribution $N(\boldsymbol{\mu}_1, \mathbf{\Sigma}_1)$ and π_2 has a distribution $N(\boldsymbol{\mu}_2, \mathbf{\Sigma}_2)$. Also, we compare the cases when π_1 and π_2 are from bivariate Laplace($\boldsymbol{\mu}_1, \mathbf{\Sigma}_1$) and bivariate Laplace($\boldsymbol{\mu}_2, \mathbf{\Sigma}_2$) respectively and also when they have bivariate Student's t distribution with 3 degrees of freedom and location and scale parameters as described above.

1. For the location shift only, Figure 2.1(a) shows the plot of empirical misclassification rates against the non-centrality parameter $\delta = \sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}$ for three bivariate spherically symmetric distributions using minimal rank classifier (RC). Under this setting, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \mathbf{I}_2$ and $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$. It is shown that the probability distribution of samples has implication on the misclassification rate. Misclassification probability is least in bivariate normally distributed samples and highest in bivariate Laplace distributed samples among the three distributions given that the competing classes have equal scale, as seen in Figure 2.1(a). When $\delta = 0$, the distributions of $\mathbf{X}$ and $\mathbf{Y}$ are the same, and have equal chance of being correctly classified. Hence probability of misclassification at this value of δ is half. As δ goes away from 0, the distinction between the two classes becomes clearer and misclassification error decreases as $|\delta|$ increases for each of the three distributions.

2. In literature, LDA and QDA are the optimal traditional classification procedures for distributions with location shift and scale shift respectively, provided normality is assumed. Comparing the performance of minimal rank classifier with some existing methods (Fisher's LDA, support vector machine (SVM), maximum depth classifier based on Oja depth (O-D) and projection depth (P-D)), Figure 2.2 shows that minimal rank classifier competes favourably with other classifiers. In Figure 2.2(a), it is shown that minimal rank classifier compares favorably with LDA and other classifiers for bivariate normal samples when there is location shift problem. The misclassification probabilities of these classifiers are almost equivalent for each value of δ. This is similar for bivariate Laplace samples and for bivariate t samples as shown in Figure 2.2(b)-(c) respectively. It is shown in Ghosh and Chaudhuri (2005b) that classifiers based on maximum depth, for some depth functions, are Bayes rule

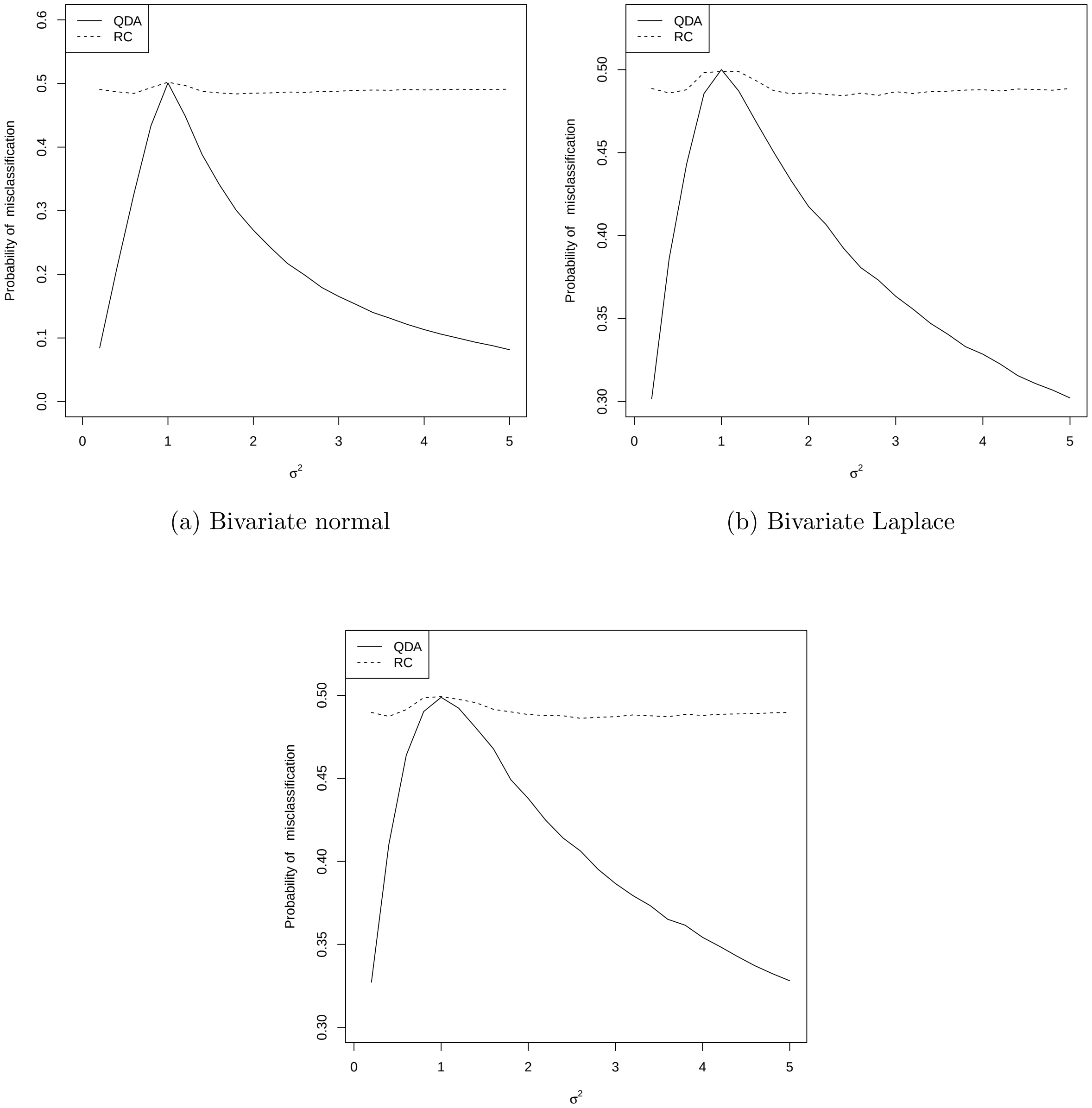


(a) Bivariate normal

(b) Bivariate Laplace

(c) Bivariate t

Figure 2.2: Comparison of minimal rank classifier with some other classifiers based on misclassification rates for distributions with $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{I}_2$ and $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$.



(a) Bivariate normal

(b) Bivariate Laplace

(c) Bivariate t

Figure 2.3: Comparison of minimal rank classifier with some other classifiers based on misclassification rates for distributions with $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$ and $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$.

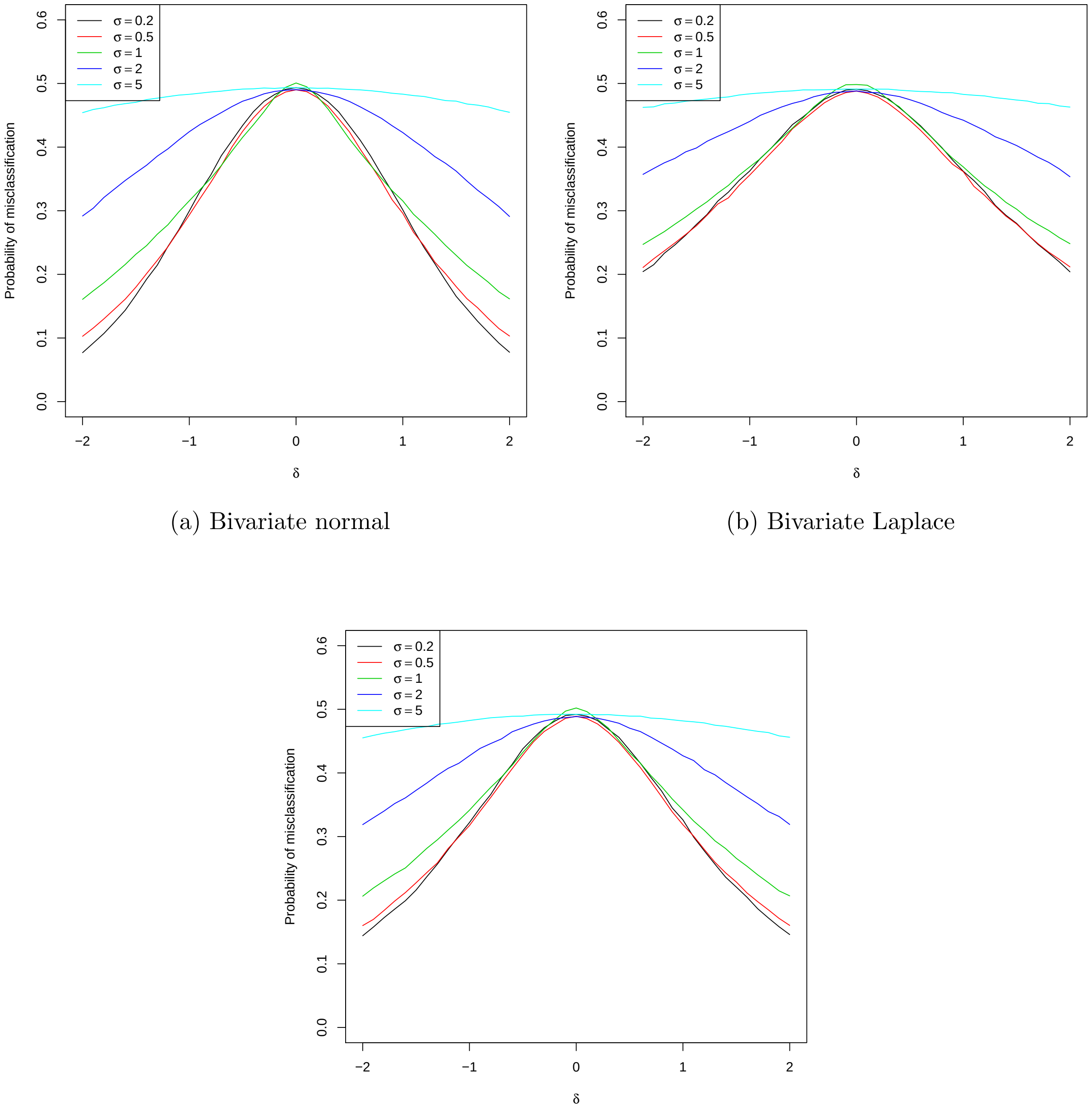
for location shift problem. Hence, all these competing classifiers are Bayes rule, except LDA which is optimal under normality.

2.3.2 Example 2: Scale shift

Suppose $\mathbf{X}$ and $\mathbf{Y}$ are bivariate normally distributed samples with $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}_1)$ and $\mathbf{Y} \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}_2)$, with sizes n_1 and n_2 respectively. We construct a separating curve using minimal rank classifier, classify m observations from each of the distributions and compute the probability of misclassification associated with it. $n_1 = n_2 = m = 100$, $d = 2$, $\boldsymbol{\Sigma}_1 = I_2$ and $\boldsymbol{\Sigma}_2 = \sigma^2 I_2$, and make plots of estimates of associated misclassification rate for homogenous scale and heterogenous scale cases. The results are thereafter compared with the result from existing methods. We repeat the simulation process for bivariate t distributed samples with 3 degrees of freedom and bivariate Laplace distributed samples. Both bivariate t distribution with 3 degree of freedom and bivariate Laplace distribution have mean vectors and covariance matrices, $(\boldsymbol{\mu}, \boldsymbol{\Sigma}_1)$ and $(\boldsymbol{\mu}, \boldsymbol{\Sigma}_2)$ for two competing samples, where $\boldsymbol{\mu}, \boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are as defined above.

1. Figure 2.1(b) plots misclassification rate against $2\log_e \sigma$ for scale shift only. Under this setting, $\boldsymbol{\Sigma}_1 = \mathbf{I}_2$, $\boldsymbol{\Sigma}_2 = \sigma^2\mathbf{I}_2$, $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$. When $\sigma = 1, \tau = 0$, the distributions of $\mathbf{X}$ and $\mathbf{Y}$ are the same, and hence have equal chance of being correctly classified and hence probability of misclassification is half as shown in Figure 2.1(b). As $2\log\sigma$ goes away from 0, distinction between the two classes becomes clearer and misclassification error decreases as $|2\log\sigma|$ increases for each of the three distributions for $\sigma > 0$. Also, error rate is highest when classes are bivariate t distributed among the three distributions. This is due to the fact that bivariate t distribution and bivariate Laplace distribution have heavier tails.

2. Figure 2.3 is a plot of misclassification probability with different values of σ in order



(a) Bivariate normal

(b) Bivariate Laplace

(c) Bivariate t

Figure 2.4: The plot of misclassification rates associated with minimal rank classifier for distributions with $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$ and $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$.

to account for the performance of RC when there is scale shift only. Comparing RC with QDA for each of the families of distributions discussed above, it can be seen in Figure 2.3(a - c) that QDA outperforms RC when the difference between the competing populations is only in their scale.

2.3.3 Example 3: Location-scale shift

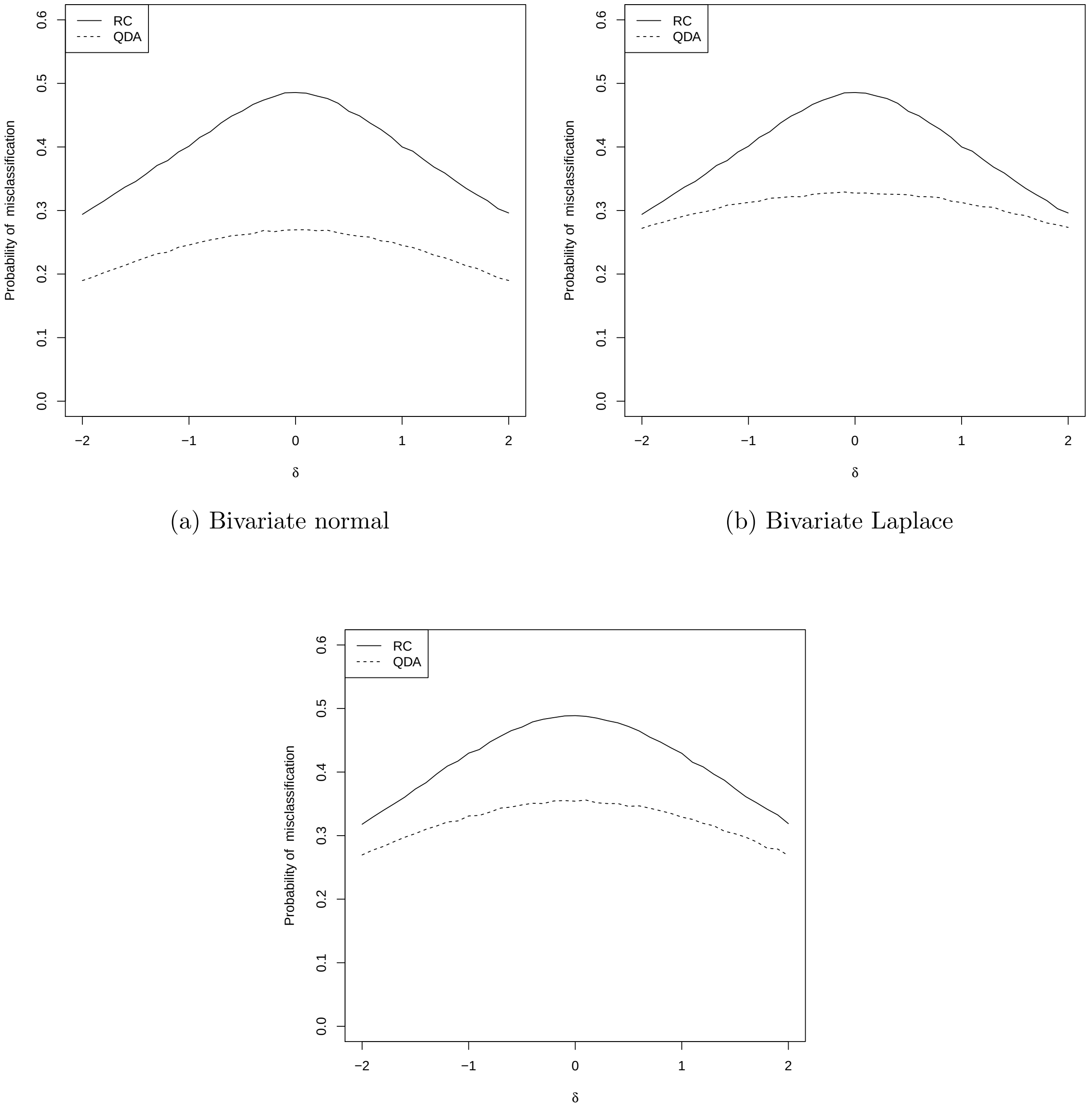
Finally, consider $\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $\boldsymbol{\mu}_2 = \begin{pmatrix} \delta \\ 0 \end{pmatrix}$ and $\boldsymbol{\Sigma}_1 = \mathbf{I}_2, \boldsymbol{\Sigma}_2 = \sigma^2\mathbf{I}_2$. Figure 2.4 gives the plot of miscassification rates against δ for different values of σ. It can be ascertained from Figure 2.4 that the misclassification rate for σ^2 and σ^{-2} at the median of any symmetric distribution are the same and increases as σ increases. Figure 2.5 gives the comparison between misclassifcation rates based on RC and QDA when the value of $\sigma = 2$ for families of three bivariate distributions. It shows that QDA outperform RC when competing populations differ in both location and scale. It can be inferred, based on these results, that RC performs poorly like LDA when the samples are from distributions with different covariance structures (i.e. $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$) irrespective of whether $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ or not.

2.4 Affine Invariant Version of Minimal Rank Classifier

2.4.1 Transformation and Re-transformation Technique

Affine invariance and symmetry

Suppose $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n \in \mathbb{R}^d$ are random variables from the same distribution. Let $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n \in \mathbb{R}^d$ be defined as $\mathbf{Y}_1 = \mathbf{A}\mathbf{X}_1 + \mathbf{b}, \mathbf{Y}_2 = \mathbf{A}\mathbf{X}_2 + \mathbf{b}, \ldots, \mathbf{Y}_n = \mathbf{A}\mathbf{X}_n + \mathbf{b}$.



(a) Bivariate normal

(b) Bivariate Laplace

(c) Bivariate t

Figure 2.5: Comparison of misclassification rates of minimal rank classifier and QDA for distributions with $\mathbf{\Sigma}_1 = \mathbf{I}_2$, $\mathbf{\Sigma}_2 = 4\mathbf{I}_2$ and $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$.

Define a statistic $T : \mathbb{R}^d \to \mathbb{R}^d$, then statistic T is affine invariant if

$$T(\mathbf{A}\mathbf{X}_1 + \mathbf{b}, \ldots, \mathbf{A}\mathbf{X}_n + \mathbf{b}) = T(\mathbf{X}_1, \ldots, \mathbf{X}_n)$$

where $\mathbf{A}$ is any $d \times d$ non-singular matrix and $\mathbf{b}$ is a d-dimensional constant vector. Non-invariance of spatial rank under arbitrary affine transformation is well known (Chakraborty and Chaudhuri, 1996; Chakraborty and Chaudhuri, 1998; Chakraborty, Chaudhuri and Oja, 1998; Serfling, 2002) and may affect the performance of any classifier based on it if the distribution of the data cloud deviates from spherical symmetry property. The distribution of a random variable $\mathbf{X}$ is said to be spherically symmetric about $\boldsymbol{\theta}$ if, for any orthogonal matrix $\mathbf{A}$,

$$\mathbf{X} - \boldsymbol{\theta} \overset{d}{=} \mathbf{A}(\mathbf{X} - \boldsymbol{\theta}).$$

The density function of any spherically symmetric distribution of a random variable $\mathbf{X}$, if it exists, is of the form $f(\mathbf{x}) \propto g\left((\mathbf{x} - \boldsymbol{\theta})^T(\mathbf{x} - \boldsymbol{\theta})\right)$ for some nonnegative scalar function $g(.)$. The distribution of a random variable $\mathbf{X}$ is said to be elliptically symmetric about $\boldsymbol{\theta}$ if there exists a $d \times d$ nonsingular matrix $\mathbf{A}$ such that $\mathbf{A}^T(\mathbf{X} - \boldsymbol{\theta})$ has a spherically symmetric distribution about $\mathbf{0}$. See Liu(1990), Liu and Singh (1993), Liu, Parelius and Singh (1999) and Serfling (2006a) for further reading on multivariate symmetry.

Need for affine invariant classifier

To demonstrate robustness of minimal rank classifier against deviation from the property of spherical symmetry (i.e. correlation among variables in the population from which the sample is drawn) using a numerical example, we use the set-up in simulation study in Section 2.3 for location shift and assume $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. Suppose the difference between competing population means is $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T = \begin{pmatrix} \delta & \rho\delta \end{pmatrix}$, it is easy to show LDA is independent of correlation ρ existing within the populations. We know that

$$(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) = \frac{1}{1-\rho^2}\begin{pmatrix} \delta & \rho\delta \end{pmatrix}\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}\begin{pmatrix} \delta \\ \rho\delta \end{pmatrix} = \frac{1}{1-\rho^2}(\delta^2-\rho^2\delta^2) = \delta^2.$$

The probability of misclassification associated with LDA for normal populations with equal covariance matrices is $\Phi\big(-\sqrt{(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)}\big) = \Phi(-\delta)$. As a result of this, misclassification rate remains constant for different values of $\rho \in [0.0, 1.0)$ as shown in Figure 2.6.

Define $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$ and $\mathbf{Y}_i = \mathbf{A}\mathbf{X}_i + \mathbf{b}$ for nonsingular matrix $\mathbf{A}$ and constant vector $\mathbf{b}$, then

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbf{y}-\mathbf{Y}_i}{||\mathbf{y}-\mathbf{Y}_i||}\right\| = \left\|\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbf{A}(\mathbf{x}-\mathbf{X}_i)}{||\mathbf{A}(\mathbf{x}-\mathbf{X}_i)||}\right\| \neq \left\|\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbf{x}-\mathbf{X}_i}{||\mathbf{x}-\mathbf{X}_i||}\right\|.$$

Table 2.1 presents the performance of minimal rank classifier for various values of ρ. The misclassification probability behaves anomalously for different values of $\rho \in [0, 1)$ irrespective class distribution. The misclassification rates are not in any specific order of ρ. The reason is that though F and G are taking more ellipsoid form as ρ increases, the classifier is being computed with respect to sphere as minimal rank classifier is based on non-invariant spatial rank under affine transformation. To overcome the problem of affine non-invariance property of spatial rank, we use transformation and re-transformation technique developed in Chakraborty (2001).

Transformation and re-transformation technique (TR)

Transformation and re-transformation methodology is a procedure involving conversion of non-equivariant and non-invariant measures under affine transformation to affine equivariant and affine invariant versions respectively, using data driven coordinate system. TR technique was developed in Chakraborty and Chaudhuri (1996) and then used to con-

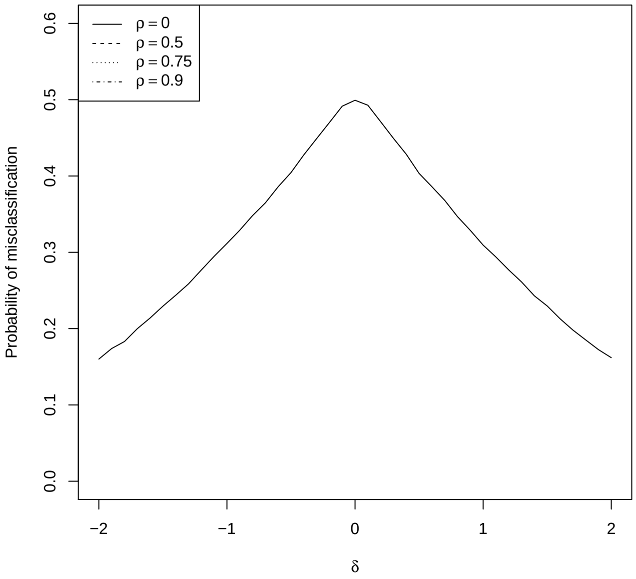


Figure 2.6: Robustness of LDA against deviation in spherical symmetry.

Table 2.1: Misclassification rates of minimal rank classifier when $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}, \mathbf{\Sigma} \neq \lambda \mathbf{I}_2$, $\lambda \in \mathbb{R}$.

			Minimal rank classifier			
Distribution	δ	LDA	$\rho = 0.00$	$\rho = 0.50$	$\rho = 0.75$	$\rho = 0.90$
Bivariate normal	0.5	0.4059	0.4132	0.4121	0.4108	0.4091
	1	0.3117	0.3137	0.3158	0.3148	0.3127
	1.5	0.2290	0.2306	0.2371	0.2362	0.2325
	2	0.1602	0.1607	0.1699	0.1691	0.1643
Bivariate Laplace	0.5	0.4361	0.4459	0.4455	0.4433	0.4415
	1	0.3577	0.3691	0.3688	0.3664	0.3645
	1.5	0.2960	0.3022	0.3027	0.3014	0.2993
	2	0.2434	0.2472	0.2503	0.2495	0.2466
Bivariate t	0.5	0.4216	0.4326	0.4310	0.4287	0.4271
	1	0.3347	0.3419	0.3420	0.3404	0.3379
	1.5	0.2618	0.2652	0.2675	0.2663	0.2636
	2	0.2018	0.2054	0.2099	0.2091	0.2066

struct an affine equivariant median. This technique was also used in Chakraborty and Chaudhuri (1998) to construct robust estimate of location; in Chakraborty, Chaudhuri and Oja (1998) to construct an affine equivariant median and angle test; in Chakraborty (2001) to construct an affine equivariant quantile; and also in Dutta and Ghosh (2012b). Consider the d-dimensional data points $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n$ in $\mathbb{R}^d$, for any $d \times d$ nonsingular matrix $\mathbf{A}$ and any $\mathbf{b} \in \mathbb{R}^d$, Chakraborty, Chaudhuri and Oja (1998) has shown that the transformation that maps $\mathbf{X}_i$ into $\mathbf{A}\mathbf{X}_i + \mathbf{b}$, where $1 \leqslant i \leqslant n$, essentially expresses the original data in terms of a new coordinate system determined by $\mathbf{A}$ and $\mathbf{b}$ with origin at $-\mathbf{A}^{-1}\mathbf{b}$. The concept is to form an appropriate data driven coordinate system and express all the data points in terms of the new coordinate system. Then compute the spatial rank of the transformed data. Define

$$S_n = \{\alpha | \alpha \subset \{1, 2, \ldots, n\} \text{ and } |\alpha| = d + 1\} \tag{2.4.1}$$

as the collection of all $d+1$ subset of $\{1, 2, ..., n\}$. For a fixed $\alpha = \{i_0, i_1, ..., i_d\} \subset S_n$, we define $\mathbf{X}(\alpha)$ to be a $d \times d$ matrix whose columns are $\mathbf{X}_{i_1} - \mathbf{X}_{i_0}$, $\mathbf{X}_{i_2} - \mathbf{X}_{i_0}$, ..., $\mathbf{X}_{i_d} - \mathbf{X}_{i_0}$. That is, one of the $d+1$ data points determines the origin and the lines joining that origin to the remaining d data point will form the coordinate system. Assuming that elements of α are naturally ordered and that $\mathbf{X}_i$'s are independent and identically distributed observations with common probability distribution, which is absolutely continuous with respect to the Lebesgue measure in $\mathbb{R}^d$, $\mathbf{X}(\alpha)$ is invertible with probability one (Chakraborty, 2001). So, $\mathbf{X}(\alpha)$ is the transformation matrix and for each $i \notin \alpha$, we transform the data set $\mathbf{X}_i$ into a new coordinate system, $\mathbf{Y}_i = \{\mathbf{X}(\alpha)\}^{-1}\mathbf{X}_i$ and compute the rank of $\mathbf{y} = \{\mathbf{X}(\alpha)\}^{-1}\mathbf{x}$.

2.4.2 Adaptive choice of α

We choose $\mathbf{X}(\alpha)$ in such a way that the columns of $\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{X}(\alpha)$ are as orthogonal as possible. That is, we want to choose $\mathbf{X}(\alpha)$ in a way that $\{\mathbf{X}(\alpha)\}^T\mathbf{\Sigma}^{-1}\mathbf{X}(\alpha)$ becomes as close as possible to a scalar multiple of identity matrix (i.e. $\lambda\mathbf{I}_d$, where $\mathbf{I}_d$ is a $d\times d$ identity matrix and λ is any positive constant). Since $\mathbf{\Sigma}$ is unknown in practice, we compute its estimate from the data. The choice of α depends on the value of α that minimises

$$v(\alpha) = \frac{\text{trace}\big(\{\mathbf{X}(\alpha)\}^T\widehat{\mathbf{\Sigma}}^{-1}\mathbf{X}(\alpha)\big)/d}{\big[\det\big(\{\mathbf{X}(\alpha)\}^T\widehat{\mathbf{\Sigma}}^{-1}\mathbf{X}(\alpha)\big)\big]^{\frac{1}{d}}} \tag{2.4.2}$$

so that $v(\alpha)$ becomes very close to 1. Obviously, once α is selected, the computation of affine invariant spatial rank is straightforward in any dimension.

2.4.3 Affine Invariant Multivariate Rank

The affine invariant spatial rank is defined as

$$\text{rank}_F(\mathbf{x}) = E_F\left(\frac{\{\mathbf{X}(\alpha)\}^{-1}[\mathbf{x}-\mathbf{X}]}{||\{\mathbf{X}(\alpha)\}^{-1}[\mathbf{x}-\mathbf{X}]||}\right). \tag{2.4.3}$$

The sample version is defined as

$$\text{rank}_{F_n}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\{\mathbf{X}(\alpha)\}^{-1}[\mathbf{x}-\mathbf{X}_i]}{||\{\mathbf{X}(\alpha)\}^{-1}[\mathbf{x}-\mathbf{X}_i]||}\right) \tag{2.4.4}$$

Steps involved in computing the affine invariant spatial rank after determining the transformation matrix are:

1. Transform every observation $\mathbf{X}_i, i = 1, ..., n$ into a new coordinate system, $\mathbf{Y}_i = \{\mathbf{X}(\alpha)\}^{-1}\mathbf{X}_i$.

2. Transform observation $\mathbf{x}$ into a new coordinate system, $\mathbf{y} = \{\mathbf{X}(\alpha)\}^{-1}\mathbf{x}$.

3. Compute rank of $\mathbf{y}$ with respect to data cloud, $\mathbf{Y}_i, i = 1, ..., n$.

Now, we want to show the affine invariance property of the transformed multivariate rank defined in equation (2.4.4) by the lemma below.

Lemma 2.4.1 *Suppose* $\mathbf{X}_i, 1 \leqslant i \leqslant n$ *is a sample on* $\mathbb{R}^d$ *having a distribution* F. *For any* $\alpha \in S_n$, $\mathrm{rank}_{F_n}(\mathbf{x})$ *defined in equation (2.4.4) is affine invariant.*

Proof: For any $d \times d$ nonsingular matrix $\mathbf{A}$, let $\mathbf{Y}_i = \mathbf{A}\mathbf{X}_i + \mathbf{b}$. Since

$$\mathbf{X}(\alpha) = [\mathbf{X}_{i_1} - \mathbf{X}_{i_0}, ..., \mathbf{X}_{i_d} - \mathbf{X}_{i_0}]$$

we have

$$\begin{aligned}
&\mathbf{Y}(\alpha) = [\mathbf{Y}_{i_1} - \mathbf{Y}_{i_0}, \mathbf{Y}_{i_2} - \mathbf{Y}_{i_0}, \dots, \mathbf{Y}_{i_d} - \mathbf{Y}_{i_0}] \\
&= [\mathbf{A}\mathbf{X}_{i_1} + \mathbf{b} - (\mathbf{A}\mathbf{X}_{i_0} + \mathbf{b}), \mathbf{A}\mathbf{X}_{i_2} + \mathbf{b} - (\mathbf{A}\mathbf{X}_{i_0} + \mathbf{b}), \dots, \mathbf{A}\mathbf{X}_{i_d} + \mathbf{b} - (\mathbf{A}\mathbf{X}_{i_0} + \mathbf{b})] \\
&= [\mathbf{A}\mathbf{X}_{i_1} - \mathbf{A}\mathbf{X}_{i_0}, \mathbf{A}\mathbf{X}_{i_2} - \mathbf{A}\mathbf{X}_{i_0}, \dots, \mathbf{A}\mathbf{X}_{i_d} - \mathbf{A}\mathbf{X}_{i_0}] \\
&= \mathbf{A}[\mathbf{X}_{i_1} - \mathbf{X}_{i_0}, \mathbf{X}_{i_2} - \mathbf{X}_{i_0}, \dots, \mathbf{X}_{i_d} - \mathbf{X}_{i_0}] = \mathbf{A}\mathbf{X}(\alpha) \qquad (2.4.5)
\end{aligned}$$

The transformed multivariate rank of a data point $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$, where $\mathbf{x} \in \mathbb{R}^d$ is

$$\begin{aligned}
\mathrm{rank}_{G_n}(\mathbf{A}\mathbf{x} + \mathbf{b}) = \mathrm{rank}_{G_n}(\mathbf{y}) &= \frac{1}{n}\sum_{i=1}^{n} \frac{\{\mathbf{Y}(\alpha)\}^{-1}[\mathbf{y} - \mathbf{Y}_i]}{||\{\mathbf{Y}(\alpha)\}^{-1}[\mathbf{y} - \mathbf{Y}_i]||} \\
&= \frac{1}{n}\sum_{i=1}^{n} \frac{\{\mathbf{A}\mathbf{X}(\alpha)\}^{-1}[(\mathbf{A}\mathbf{x} + \mathbf{b}) - (\mathbf{A}\mathbf{X}_i + \mathbf{b})]}{||\{\mathbf{A}\mathbf{X}(\alpha)\}^{-1}[(\mathbf{A}\mathbf{x} + \mathbf{b}) - (\mathbf{A}\mathbf{X}_i + \mathbf{b})]||} \\
&= \frac{1}{n}\sum_{i=1}^{n} \frac{\{\mathbf{X}(\alpha)\}^{-1}\mathbf{A}^{-1}\mathbf{A}[\mathbf{x} - \mathbf{X}_i]}{||\{\mathbf{X}(\alpha)\}^{-1}\mathbf{A}^{-1}\mathbf{A}[\mathbf{x} - \mathbf{X}_i]||} \\
&= \frac{1}{n}\sum_{i=1}^{n} \frac{\{\mathbf{X}(\alpha)\}^{-1}[\mathbf{x} - \mathbf{X}_i]}{||\{\mathbf{X}(\alpha)\}^{-1}[\mathbf{x} - \mathbf{X}_i]||} = \mathrm{rank}_{F_n}(\mathbf{x}) \qquad (2.4.6)
\end{aligned}$$

that is, $\mathrm{rank}_{G_n}(\mathbf{A}\mathbf{x} + \mathbf{b}) = \mathrm{rank}_{F_n}(\mathbf{x})$. □

Hence, the transformed multivariate rank is invariant under affine transformation. Classifier based on this transformed rank is affine invariant and can handle the problem associated with deviation from spherical symmetry.

2.4.4 Affine Invariant version of Minimum Rank Classifier

Given any two populations π_1, with $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{n_1} \in \pi_1$ and π_2, with $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_{n_2} \in \pi_2$. For the training sample in the population π_1, let $\mathbf{X}_{i_0}, \ldots, \mathbf{X}_{i_d}$ be $d+1$ observations and $\alpha = \{i_0, i_1, \ldots, i_d\}$ denotes the set of $d+1$ indices.

$$\text{rank}_F(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \frac{\{\mathbf{X}(\alpha)\}^{-1}(\mathbf{x} - \mathbf{X}_i)}{\|\{\mathbf{X}(\alpha)\}^{-1}(\mathbf{x}\mathbf{X}_i)\|}.$$

Similarly, we can define the affine invariant spatial rank function with respect to the training sample $\mathbf{Y}_1, \ldots, \mathbf{Y}_m \in \pi_2$ as

$$\text{rank}_G(\mathbf{x}) = \frac{1}{m} \sum_{i=1}^{m} \frac{\{\mathbf{Y}(\beta)\}^{-1}(\mathbf{x} - \mathbf{Y}_i)}{\|\{\mathbf{Y}(\beta)\}^{-1}(\mathbf{x} - \mathbf{Y}_i)\|}$$

where β is a set of $d+1$ indices $\{j_0, j_1, \ldots, j_d\}$ and $\mathbf{Y}(\beta)$ is the $d \times d$ matrix formed with the columns $\mathbf{Y}_{j_1} - \mathbf{Y}_{j_0}, \ldots, \mathbf{Y}_{j_d} - \mathbf{Y}_{j_0}$. The optimal transformation matrix $\mathbf{Y}(\beta)$ is obtained by minimising a similar criterion for the data $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$. Then an affine invariant version of the classification rule for any $\mathbf{x} \in \mathbb{R}^d$ can be defined as

$$\text{assign } \mathbf{x} \text{ to } \pi_1 \text{ if } \|\text{rank}_F(\mathbf{x})\| \leqslant \|\text{rank}_G(\mathbf{x})\| \tag{2.4.7}$$
$$\text{assign } \mathbf{x} \text{ to } \pi_2 \text{ otherwise.}$$

We call this classification method minimal affine invariant rank classifier (AIRC). If there are $J(\geqslant 2)$ populations, then assign $\mathbf{x}$ to population π_k, $1 \leqslant k \leqslant J$ if

$$||\text{rank}_{F_k}(\mathbf{x})|| = \min_j ||\text{rank}_{F_j}(\mathbf{x})||; \;\; j = 1, 2, ..., J \tag{2.4.8}$$

where $F_1, F_2, \ldots, F_J$ are distribution functions corresponding to J populations and $\text{rank}_{F_j}(\mathbf{x})$ is as defined in equation (2.4.3). Now, we want to show that AIRC is a Bayes rule under the conditions of Theorem 2.2.1 for elliptically symmetric distributions. This is given in Theorem 2.4.1 below.

Theorem 2.4.1 *Let f_1 and f_2 be the density functions of populations, π_1 and π_2 having elliptically symmetric distributions F and G about $\boldsymbol{\theta}_F$ and $\boldsymbol{\theta}_G$ respectively with prior probabilities p_1 and p_2 respectively from the same family of multivariate distributions such that $G(\mathbf{x}) = F(\mathbf{x} - \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a location shift in $\mathbb{R}^d$. If $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, $c(1|2) = c(2|1)$ and $p_1 = p_2$, then the Bayes rule is*

$$\begin{cases} ||\mathrm{rank}_F(\mathbf{x})|| \leqslant ||\mathrm{rank}_G(\mathbf{x})|| & \Rightarrow \textit{assign } \mathbf{x} \textit{ to population } \pi_1 \\ ||\mathrm{rank}_F(\mathbf{x})|| > ||\mathrm{rank}_G(\mathbf{x})|| & \Rightarrow \textit{assign } \mathbf{x} \textit{ to population } \pi_2, \end{cases} \tag{2.4.9}$$

$\mathrm{rank}_F(\mathbf{x})$ *and* $\mathrm{rank}_G(\mathbf{x})$ *are as defined in equation (2.4.3).*

Proof: Suppose the distribution F is absolutely continuous elliptically symmetric about $\boldsymbol{\theta}_F \in \mathbb{R}^d$, then its probability density function can be written as $f_1(\mathbf{x}) = g_1\left(\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{x} - \boldsymbol{\theta}_F)\|\right)$ for some non-negative decreasing function $g_1 : \mathbb{R} \to \mathbb{R}$ and similarly, $f_2(\mathbf{x}) = g_2\left(\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{x} - \boldsymbol{\theta}_G)\|\right)$ for some non-negative decreasing function $g_2 : \mathbb{R} \to \mathbb{R}$ for the distribution G, which is elliptically symmetric about $\boldsymbol{\theta}_G \in \mathbb{R}^d$.

Now by Theorem 2.1.1 and Lemma 2.4.1,

$$\mathrm{rank}_F(\mathbf{x}) = h_1(||\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{x} - \boldsymbol{\theta}_F)||)\frac{\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{x} - \boldsymbol{\theta}_F)}{||\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{x} - \boldsymbol{\theta}_F)||}$$

and

$$\mathrm{rank}_G(\mathbf{x}) = h_1(||\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{x} - \boldsymbol{\theta}_G)||)\frac{\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{x} - \boldsymbol{\theta}_G)}{||\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{x} - \boldsymbol{\theta}_G)||}$$

for some increasing functions h_1 and h_2. This implies

$$\|\mathrm{rank}_F(\mathbf{x})\| = h_1(\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{x} - \boldsymbol{\theta}_F)\|)$$

and

$$\|\mathrm{rank}_G(\mathbf{x})\| = h_2(\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{x}-\boldsymbol{\theta}_G)\|).$$

We can write $f_1(\mathbf{x}) = g_1\big(h_1^{-1}(\|\mathrm{rank}_F(\mathbf{x})\|)\big)$ and $f_2(\mathbf{x}) = g_2\big(h_2^{-1}(\|\mathrm{rank}_G(\mathbf{x})\|)\big)$. Now,

$$f_1(\mathbf{x}) \geqslant f_2(\mathbf{x}) \text{ gives } g_1\big(h_1^{-1}(\|\mathrm{rank}_F(\mathbf{x})\|)\big) \geqslant g_2\big(h_2^{-1}(\|\mathrm{rank}_G(\mathbf{x})\|)\big).$$

For $G(\mathbf{x}) = F(\mathbf{x}-\boldsymbol{\theta})$, we have $h_1 = h_2 = h$ (say) and $g_1 = g_2 = g$ (say). So,

$$f_1(\mathbf{x}) \geqslant f_2(\mathbf{x}) \text{ implies } g\big(h^{-1}(\|\mathrm{rank}_F(\mathbf{x})\|)\big) \geqslant g\big(h^{-1}(\|\mathrm{rank}_G(\mathbf{x})\|)\big).$$

$g \circ h^{-1}$ is decreasing function since h is increasing function and g is monotone decreasing function. Hence

$$||\mathrm{rank}_F(\mathbf{x})|| \leqslant ||\mathrm{rank}_G(\mathbf{x})||.$$

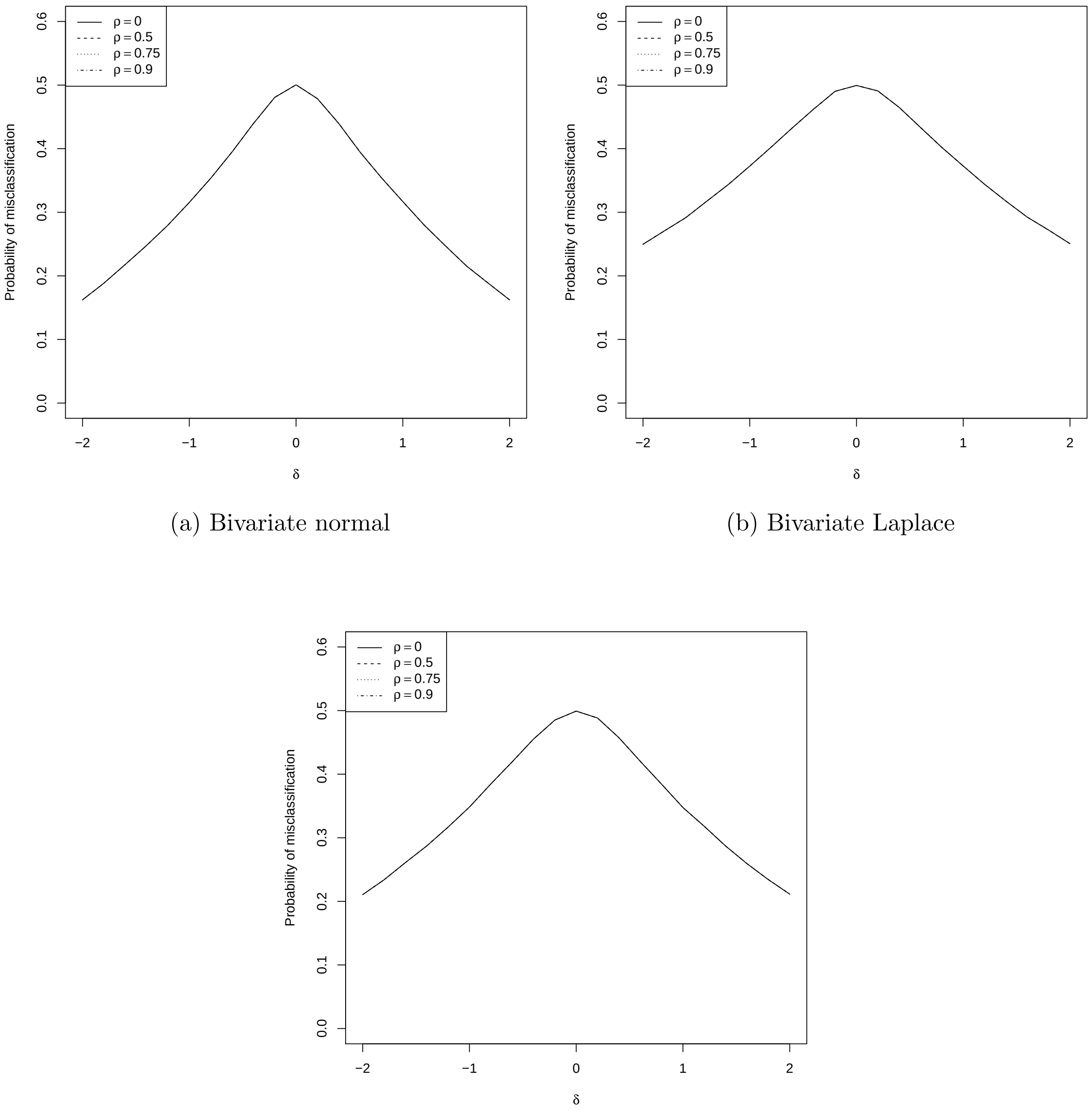
The proof is complete. □

Example: Location Shift

Consider

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \; n = 100, \; \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

and $\boldsymbol{\mu}_2$ is chosen in such a way that $(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) = \delta^2$. Figure 2.7(a)-(c) present plots of misclassification rates asociated with AIRC against non-centrality parameter δ for some values of ρ for bivariate normally distributed, bivariate Laplace distributed and bivariate t distributed (with 3 degrees of freedom) samples of size n each. These plots show that, for each value of δ, misclassification probability is the same for all values of ρ. The implication of this is that AIRC is independent of values of $\rho, \forall \rho \in [0, 1)$. This is because effect of ρ on the data has been removed when both $\mathbf{x}$ and $\mathbf{X}_i$ are premultiplied



(a) Bivariate normal

(b) Bivariate Laplace

(c) Bivariate t

Figure 2.7: Robustness of AIRC against deviation in spherical symmetry.

Table 2.2: Comparison of classifiers based on average misclassification errors for distributions with location shift.

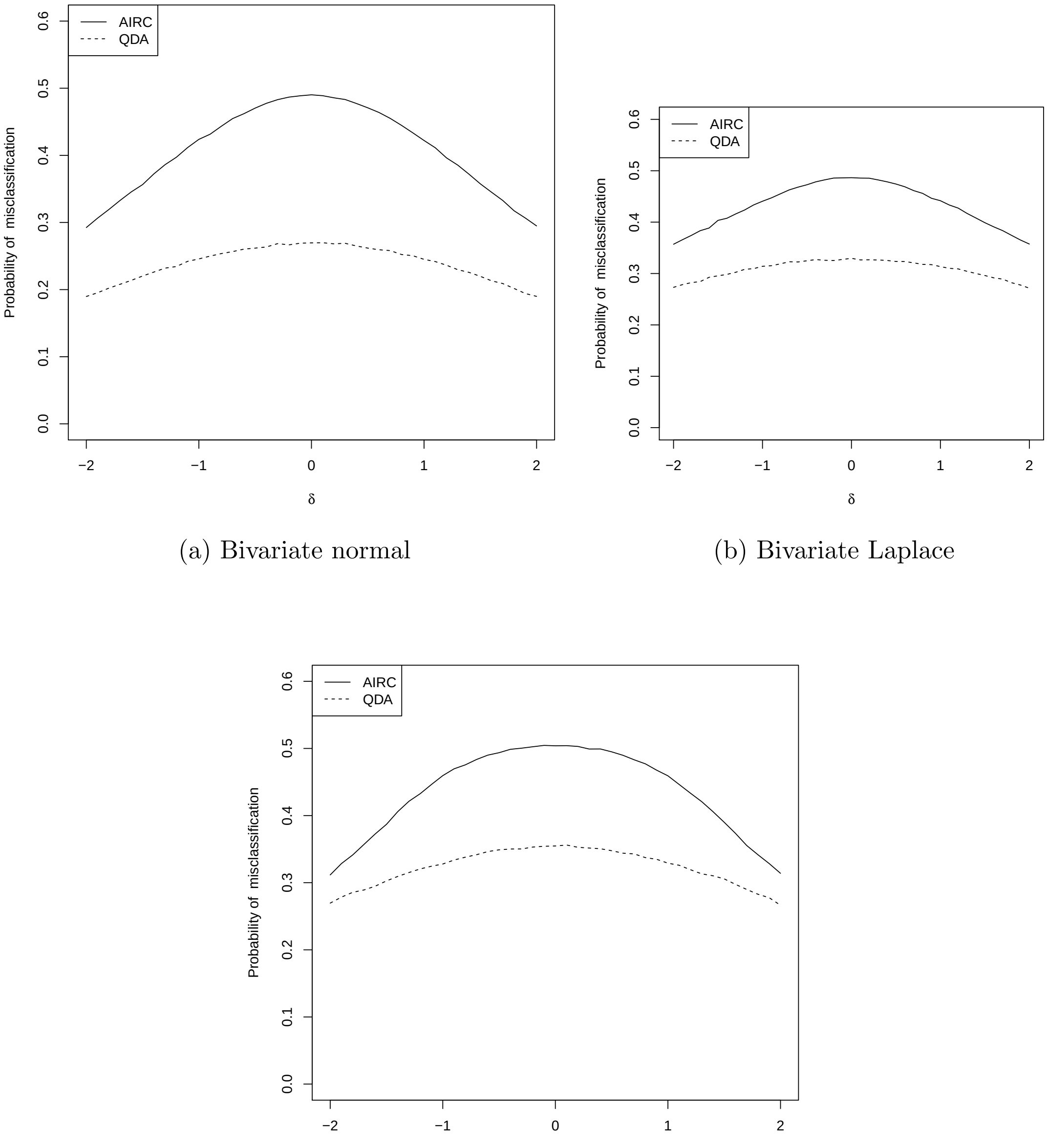
		Classifiers						
Distribution	δ	Bayes	LDA	SVM	O-D	P-D	RC	AIRC
Bivariate normal	1	0.3085	0.3148	0.3157	0.3181	0.3213	0.3129	0.3168
	2	0.1587	0.1612	0.1602	0.1649	0.1660	0.1615	0.1623
Bivariate Laplace	1	0.3576	0.3770	0.3814	0.3831	0.3729	0.3693	0.3727
	2	0.2415	0.2464	0.2573	0.2503	0.2508	0.2475	0.2506
Bivariate t	1	0.3339	0.3746	0.3505	0.3707	0.3400	0.3418	0.3475
	2	0.2019	0.2220	0.2137	0.2185	0.2053	0.2060	0.2113

by $\{\mathbf{X}(\alpha)\}^{-1}$ for all $i \notin \alpha$. Unlike minimal rank classifier whose performance is enhanced by spherical symmetry of the distribution of the training data, AIRC performs well for elliptically symmetric distributions.

Comparing AIRC with some of other classifiers for location shift case with $\mathbf{\Sigma} = \mathbf{I}_2$, the result is presented in Table 2.2. These classifiers are support vector machine (SVM), maximum depth classifier based on Oja depth(O-D) and projection depth(P-D), LDA and RC. Table 2.2 shows that AIRC competes favourably with other classifiers for location shift problem for the three multivariate distributions.

Example : Scale Shift

Now consider $\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\mu}_2 = \begin{pmatrix} \delta \\ 0 \end{pmatrix}$ and $\mathbf{\Sigma}_1 = \mathbf{I}_2, \mathbf{\Sigma}_2 = \sigma^2\mathbf{I}_2$. Minimal affine invariant rank classifier performs very poor like minimal rank classifier when there is scale shift irrespective of whether there is location shift or not. The mislassification rates are higher in AIRC than QDA when normality is assumed for different values of non-central parameter δ and $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$. This can be seen in Figure 2.8(a). Similarly for non-normal distributions, AIRC has higher misclassification rates than QDA when $\sigma = 2$ as shown in Figure 2.8(b)-(c). This poor performance of AIRC, as well as RC, when there is scale shift can be attributed to the lack of scale parameter in the formulation of spatial rank



(a) Bivariate normal

(b) Bivariate Laplace

(c) Bivariate t

Figure 2.8: Performance of AIRC for location-scale shift when $\sigma = 2.0$.

function. To overcome this problem, we shall develop a classifier based on rank region in the next chapter.

Chapter 3

Classifier Based on Volume of Central Rank Regions

3.1 Central Rank Regions and its Associated Volume

In this chapter, we propose a classification method based on volumes of central rank regions. It is imperative to briefly discuss central rank region and its associated volume. The Definition 3.1.1 below gives a mathematical expression for central rank region.

Definition 3.1.1 *Suppose* $\mathbf{X} \in \mathbb{R}^d$ *is a random vector with distribution function* F*. The central rank region is defined as*

$$C_F(r) = \{\mathbf{x} : \|\mathrm{rank}_F(\mathbf{x})\| \leqslant r\}, \quad 0 < r < 1 \tag{3.1.1}$$

where $\mathrm{rank}_F(\mathbf{x})$ *is the spatial rank of* $\mathbf{x}$ *with respect to* F*.*

Central rank region, $C_F(r)$ is equivariant under location shift, orthogonal and homogenous scale transformations (Liu, Parelius and Singh, 1999). It is equivariant under affine transformation if and only if $\mathrm{rank}_F(\mathbf{x})$ is invariant under affine transformation. Serfling (2002, 2004), Guha (2012) and Guha and Chakraborty (2013) also defined central rank regions but with respect to probability mass. Now we can define volume of multivariate central rank regions.

Definition 3.1.2 *Suppose* $0 < r < 1$ *and* $C_F(r)$ *is the central rank region of* $\mathbf{X} \in \mathbb{R}^d$ *having distribution function* F*. The volume of central rank region, denoted by* $V_F(r)$*, is defined as*

$$V_F(r) = \textit{Volume of } C_F(r). \tag{3.1.2}$$

Some authors called $V_F(r)$, the volume functional. It is well known that $V_F(r)$ characterises the spread of the distribution F in terms of central rank region $C_F(r)$ as r increases. $V_F(r)$ measures, for small r, the overall spread of the data around the spatial median while it measures overall spread of the distribution as r increases (Guha and Chakraborty, 2013). Volume of central rank region is equivariant under orthogonal rotation and its d-root is equivariant under homogenous scale transformations. Central regions are ordered and increase with respect to r that describes their boundaries. That is, if $r_1 < r_2$, then $C_F(r_1) \subseteq C_F(r_2)$ (Serfling, 2002 and 2004). Consequently, the central rank regions and its associated volume functional can equivalently be indexed by the probability weight of the central region. The notion of central region and its corresponding volume is well discussed in Liu, Parelius and Singh (1999), Serfling (2002), Wang and Serfling (2004) using data depth. Guha and Chakraborty (2013) studied central region and its volume based on spatial rank.

3.2 Classifier Based on Volume of Rank Regions

The possibility of solving classification problem by computing volume of an observation with respect to each of the competing populations is raised here. An observation is assigned to the class for which it attains minimum volume. Suppose an observation belongs to a particular class of observations, the class will have the least rank outlyingness among the competing classes and thereby has the least volume. Spatial rank in Definition 3.1.1

may be as defined in Definition 2.1.1 or 2.4.3, depending on whether the distribution of the data cloud is spherically symmetric or elliptically symmetric respectively. Since we are interested in classification method that can handle both, we define $\text{rank}_F(\mathbf{x})$ in Definition 3.1.1 as in equation (2.4.3). It is shown in Lemma 2.4.1 that $\text{rank}_F(\mathbf{x})$ is invariant under affine transformation, hence $C_F(r)$ and $V_F(r)$ are equivariant under general affine transformations.

Obviously, $||\text{rank}_F(\mathbf{x})|| \in [0,1]$ and $||\text{rank}_F(\mathbf{x})|| \to 1$ as $\|\mathbf{x}\| \to \infty$. Following Definition 3.1.1, set $\{\mathbf{x} : ||\text{rank}_F(\mathbf{x})|| = r\}$ is nonempty for all $0 < r < 1$. Following Definition 3.1.2, $V_F(r)$ is finite, a function of r and strictly increasing for $r < 1$.

Definition 3.2.1 *Suppose F and G are absolutely continuous with respect to Lebesgue measure in $\mathbb{R}^d$, $C_F(r)$ and $C_G(r)$ are the central rank region of $\mathbf{X} \in \mathbb{R}^d$ having distribution functions F and G respectively. Define*

$$V_F(r) = \textit{Volume of } C_F(r) \ \textit{ and } \ V_G(r) = \textit{Volume of } C_G(r).$$

Also define

$$r_F(\mathbf{x}) = \|\text{rank}_F(\mathbf{x})\| \ \textit{ and } \ r_G(\mathbf{x}) = \|\text{rank}_G(\mathbf{x})\|,$$

where $\text{rank}_F(\mathbf{x})$ *and* $\text{rank}_G(\mathbf{x})$ *are as defined in equation (2.4.3). Then the classification rule based on volume of central rank region is to assign* $\mathbf{x}$ *to F if*

$$V_F\big(r_F(\mathbf{x})\big) \leqslant V_G\big(r_G(\mathbf{x})\big)$$

and to G if

$$V_F\big(r_F(\mathbf{x})\big) > V_G\big(r_G(\mathbf{x})\big).$$

The probability of misclassification associated with this classification method is

$$\Delta = p_1 P\big(V_F\big(r_F(\mathbf{x})\big) > V_G\big(r_G(\mathbf{x})\big) \mid \mathbf{x} \in \pi_1\big) + p_2 P\big(V_F\big(r_F(\mathbf{x})\big) \leqslant V_G\big(r_G(\mathbf{x})\big) \mid \mathbf{x} \in \pi_2\big).$$

Suppose there are $J(> 2)$ populations, then assign $\mathbf{x}$ to population π_k with distribution F_k, $1 \leqslant k \leqslant J$ if

$$V_{F_k}\big(r_{F_k}(\mathbf{x})\big) = \min_j V_{F_j}\big(r_{F_j}(\mathbf{x})\big),$$

where $F_1, F_2, \ldots, F_J$ are absolutely continuous distributions corresponding to J populations. The corresponding probability of misclassification is

$$\Delta = \sum_{j=1}^{J} p_j P\{V_{F_k}\big(r_{F_k}(\mathbf{x})\big) \leqslant V_{F_j}\big(r_{F_j}(\mathbf{x})\big) \mid \mathbf{x} \in F_j\},$$

where $p_1, p_2, \ldots, p_J$ are prior probability of populations $\pi_1, \pi_2, \ldots, \pi_J$ respectively. For the rest of this thesis, we shall call the classification rule based on volume of central rank region, rank region classifier (RRC) and investigate the properties of this classifier by some theorems below.

When samples drawn from populations are only available, we compute the empirical version of volume functional, $V_{F_n}\big(r_{F_n}(\mathbf{x})\big)$, carry out the classification and compute the probability of misclassification based on the samples. Theorem 3.2.1 below shows that $V_{F_n}\big(r_{F_n}(\mathbf{x})\big)$ converges to its population version with probability one.

Theorem 3.2.1 *Suppose Theorem 2.2.2 hold for elliptically symmetric distribution F, then for sufficiently large n and $r_F(\mathbf{x}) \in [0, 1]$,*

$$\big|V_{F_n}\big(r_{F_n}(\mathbf{x})\big) - V_F\big(r_F(\mathbf{x})\big)\big| \overset{a.s.}{\to} 0$$

Proof: Define

$$\begin{aligned}
V_{F_n}\big(r_{F_n}(\mathbf{x})\big) &= \text{volume}\{\mathbf{y} : r_{F_n}(\mathbf{y}) \leqslant r_{F_n}(\mathbf{x})\} \\
V_F\big(r_{F_n}(\mathbf{x})\big) &= \text{volume}\{\mathbf{y} : r_F(\mathbf{y}) \leqslant r_{F_n}(\mathbf{x})\} \\
V_F\big(r_F(\mathbf{x})\big) &= \text{volume}\{\mathbf{y} : r_F(\mathbf{y}) \leqslant r_F(\mathbf{x})\}
\end{aligned}$$

We note from Theorem 2.1.1 that $r_F(\mathbf{x})$ is continuous and bounded. Also, $V_F\big(r_F(\mathbf{x})\big)$ is continuous on $r_F(\mathbf{x}) \in (0,1)$.

$$\begin{aligned}
\big|V_{F_n}\big(r_{F_n}(\mathbf{x})\big) - V_F\big(r_F(\mathbf{x})\big)\big| \leqslant &\big|V_{F_n}\big(r_{F_n}(\mathbf{x})\big) - V_F\big(r_{F_n}(\mathbf{x})\big)\big| \\
&+ \ \big|V_F\big(r_{F_n}(\mathbf{x})\big) - V_F\big(r_F(\mathbf{x})\big)\big| = S_1 + S_2
\end{aligned}$$

By Theorem 2.2.2 and continuous mapping theorem,

$$S_2 = \big|V_F\big(r_{F_n}(\mathbf{x})\big) - V_F\big(r_F(\mathbf{x})\big)\big| \to 0$$

almost surely.

$$\begin{aligned}
S_1 =&\big|V_{F_n}\big(r_{F_n}(\mathbf{x})\big) - V_F\big(r_{F_n}(\mathbf{x})\big)\big| \\
&= \big|\text{volume}\{\mathbf{y} : r_{F_n}(\mathbf{y}) \leqslant r_{F_n}(\mathbf{x})\} - \text{volume}\{\mathbf{y} : r_F(\mathbf{y}) \leqslant r_{F_n}(\mathbf{x})\}\big| \\
&= \text{volume}\{\mathbf{y} : \min\big(r_{F_n}(\mathbf{y}), r_F(\mathbf{y})\big) \leqslant r_{F_n}(\mathbf{x}) \leqslant \max\big(r_{F_n}(\mathbf{y}), r_F(\mathbf{y})\big)\} \qquad (3.2.1)
\end{aligned}$$

Taking limit of equation (3.2.1) as $n \to \infty$,

$$\lim_{n\to\infty} \big|V_{F_n}\big(r_{F_n}(\mathbf{x})\big) - V_F\big(r_{F_n}(\mathbf{x})\big)\big| \to 0$$

almost surely. The proof is complete. □

Now, we will like to show that RRC is a Bayes rule for location shift under the same conditions as in Theorem 2.4.1. This is given formally in Theorem 3.2.2 below:

Theorem 3.2.2 *Let f_1 and f_2 be the probability density functions of populations, π_1 and π_2 having elliptically symmetric distributions F and G respectively from the same family of multivariate distributions such that $G(\mathbf{x}) = F(\mathbf{x} - \boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \mathbb{R}^d$ is a location shift. Suppose $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}$, $p_1 = p_2$ and Theorem 3.2.1 hold, then the Bayes rule is equivalent to*

$$\begin{cases} V_F\big(r_F(\mathbf{x})\big) \leqslant V_G\big(r_G(\mathbf{x})\big) & \Rightarrow \quad \textit{assign } \mathbf{x} \textit{ to population } \pi_1 \\ V_F\big(r_F(\mathbf{x})\big) > V_G\big(r_G(\mathbf{x})\big) & \Rightarrow \quad \textit{assign } \mathbf{x} \textit{ to population } \pi_2 \end{cases}$$

Proof: This proof follows from the proof of Theorem 2.4.1 for elliptically symmetric F and G. So,

$$\|\text{rank}_F(\mathbf{x})\| = h_1(\|\mathbf{\Sigma}^{-\frac{1}{2}}(\mathbf{x} - \boldsymbol{\theta}_F)\|)$$

and

$$\|\text{rank}_G(\mathbf{x})\| = h_2(\|\mathbf{\Sigma}^{-\frac{1}{2}}(\mathbf{x} - \boldsymbol{\theta}_G)\|).$$

for some increasing functions h_1 and h_2. $V_F\big(r_F(\mathbf{x})\big)$ and $V_G(r_G(\mathbf{x}))$ depend only on $||\text{rank}_F(\mathbf{x})||$ and $||\text{rank}_G(\mathbf{x})||$ respectively. Write $V_F\big(r_F(\mathbf{x})\big) = U_F(||\text{rank}_F(\mathbf{x})||)$, where U_F is an increasing function of $||\text{rank}_F(\mathbf{x})||$. Then

$$\|\mathbf{\Sigma}^{-\frac{1}{2}}(\mathbf{x} - \boldsymbol{\theta}_F)\| = \psi_F^{-1}(V_F(r_F(\mathbf{x})))$$

for some increasing function $\psi_F = U_F \circ h_1$. Similarly,

$$\|\mathbf{\Sigma}^{-\frac{1}{2}}(\mathbf{x} - \boldsymbol{\theta}_G)\| = \psi_G^{-1}(V_G(r_G(\mathbf{x})))$$

for some increasing function $\psi_G = U_G \circ h_2$. Then we can write

$$f_1(\mathbf{x}) \geqslant f_2(\mathbf{x}) \text{ implies } g_1\big(\psi_F^{-1}(V_F(r_F(\mathbf{x})))\big) \geqslant g_2\big(\psi_G^{-1}(V_G(r_G(\mathbf{x})))\big)$$

where g_1 and g_2 are non-negative real-valued decreasing functions. For $G(\mathbf{x}) = F(\mathbf{x} - \boldsymbol{\theta})$, $\psi_F = \psi_G = \psi$ (say) and $g_1 = g_2 = g$ (say). This implies $f_1(\mathbf{x}) = g\big(\psi^{-1}(V_F(r_F(\mathbf{x})))\big)$ and $f_2(\mathbf{x}) = g\big(\psi^{-1}(V_G(r_G(\mathbf{x})))\big)$. Also, $g \circ \psi^{-1}$ is a decreasing function since ψ is an increasing function and g is a decreasing function. Hence

$$V_F\big(r_F(\mathbf{x})\big) \leqslant V_G\big(r_G(\mathbf{x})\big) \text{ whenever } f_1(\mathbf{x}) \geqslant f_2(\mathbf{x}).$$

The proof is complete. □

3.3 Numerical Example: Simulation

Here, we carry out simulation study to investigate the performance of rank region classifier (RRC) for location-scale shift problem using simulation information in Section 2.3.3. Suppose

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\mu}_2 = \begin{pmatrix} \delta \\ 0 \end{pmatrix}, \boldsymbol{\Sigma}_1 = \mathbf{I}_2, \boldsymbol{\Sigma}_2 = \sigma^2 \mathbf{I}, n_1 = n_2 = m = 100.$$

For various values of σ and δ, the experiment is repeated 1000 times and average misclassification errors associated with RRC are determined and compared with misclassification errors associated with some other classifiers.

Table 3.1 presents the comparison of classifiers' performance based on their average misclassification errors. The simulation study is based on information in Section 2.3 for location-scale shift. Suppose $\mathbf{X}$ and $\mathbf{Y}$ are bivariate normally distributed samples such that $\mathbf{X} \sim N_d(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathbf{Y} \sim N_d(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, with sizes n_1 and n_2 respectively, where

Table 3.1: Misclassification rates of classifiers when $\boldsymbol{\Sigma}_1 = \mathbf{I}_2$, $\boldsymbol{\Sigma}_2 = 4\mathbf{I}_2$ and $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$.

			Misclassification error				
Distribution	Classifier	δ	$\sigma = 0.2$	$\sigma = 0.5$	$\sigma = 1.0$	$\sigma = 2.0$	$\sigma = 5.0$
Bivariate normal	RRC	1	0.1577	0.2307	**0.3086**	0.3599	0.3832
	QDA		**0.0556**	**0.1889**	0.3136	**0.2448**	**0.0813**
	AIRC		0.2979	0.2927	0.3156	0.4222	0.4827
	RC		0.3012	0.2929	0.3123	0.4222	0.4820
	O-D		0.5000	0.4983	0.3181	0.5000	0.5000
	P-D		0.2966	0.2962	0.3213	0.4206	0.4816
	SVM		0.0609	0.1969	0.3255	0.2547	0.0854
	RRC	2	0.0759	0.0863	**0.1522**	0.2300	0.2943
	QDA		**0.0184**	**0.0751**	0.1614	**0.1891**	**0.0784**
	AIRC		0.0765	0.1026	0.1630	0.2949	0.4553
	RC		0.0773	0.1035	0.1616	0.2922	0.4532
	O-D		0.5000	0.3442	0.1631	0.4984	0.5000
	P-D		0.0812	0.1078	0.1660	0.2945	0.4512
	SVM		0.0198	0.0792	0.1685	0.1954	0.0816
Bivariate Laplace	RRC	1	0.2153	0.3047	**0.3667**	0.4056	0.4165
	QDA		**0.1045**	0.2714	0.3766	**0.3109**	**0.1436**
	AIRC		0.3629	0.3554	0.3705	0.4417	0.4839
	RC		0.3603	0.3541	0.3706	0.4001	0.4519
	O-D		0.5000	0.4998	0.3831	0.5000	0.5000
	P-D		0.3643	0.3572	0.3729	0.4418	0.4823
	SVM		0.1077	**0.2606**	0.3806	0.3176	0.1475
	RRC	2	0.1236	0.1703	**0.2416**	0.3049	0.3534
	QDA		**0.0564**	0.1613	0.2471	0.2709	**0.1379**
	AIRC		0.2082	0.2145	0.2487	0.3573	0.4629
	RC		0.2111	0.2233	0.2458	0.2961	0.3773
	O-D		0.5000	0.4800	0.2512	0.4999	0.5000
	P-D		0.2055	0.2116	0.2508	0.3572	0.4634
	SVM		0.0584	**0.1587**	0.2560	**0.2617**	0.1419
Bivariate t	RRC	1	0.1977	0.2732	0.3401	0.3782	0.3795
	QDA		0.1123	0.2685	0.3716	0.3295	0.1611
	AIRC		0.3092	0.3244	0.3677	0.4592	0.4858
	RC		0.3231	0.3201	0.3431	0.4295	0.4827
	P-D		0.3224	0.3182	**0.3400**	0.4257	0.4801
	SVM		**0.1048**	**0.2357**	0.3508	**0.3045**	**0.1479**
	RRC	2	0.1071	0.1366	**0.2001**	0.2655	0.3141
	QDA		**0.0510**	0.1373	0.2225	0.2692	0.1547
	AIRC		0.1906	0.2017	0.2296	0.3116	0.4680
	RC		0.1469	0.1597	0.2067	0.3188	0.4548
	P-D		0.1494	0.1609	0.2053	0.3215	0.4540
	SVM		0.0514	**0.1274**	0.2142	**0.2361**	**0.1420**

$\mathbf{\Sigma}_1, \mathbf{\Sigma}_2, n_1, n_2, \boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are as defined above. It is seen that RRC competes favourably with SVM and QDA for some values of δ and σ. For $\sigma < 1$, it is seen that RRC competes favourably with other six classifiers, which are AIRC, RC, SVM, QDA, O-D and P-D. When $\sigma = 1$, misclassification errors associated with the seven classifiers are equivalent, as shown in Table 3.1 below. This means that at $\sigma = 1$, the distributions of competing classes will have homogenous scale, for which all these classifiers are optimal if class prior probabilities are equal. This is also the case for bivariate Laplace distributed samples and bivariate t distributed samples with 3 degrees of freedom. When $\sigma > 1$, RRC competes favourably with others as well. It is known that QDA is Bayes procedure under normality, so it has the least misclassification rates, even when normality assumption is violated because of its robustness to deviation from normality assumption. It is seen that misclassification error associated with RRC increases with the increase in σ. O-D, P-D, RC and AIRC have highest misclassification rates for $\sigma < 1$ and $\sigma > 1$. Also, their associated misclassification errors increase with the increase in σ.

3.4 Improved Classifier Based on Volume of Central Rank Regions

Balanda and MacGillivray (1990) introduced scale-scale plot for comparing univariate distributions. Guha and Chakraborty (2013) used scale-scale plot as an efficient visual tool to validate distributional assumptions for multivariate data. Suppose $F = G$ in $\mathbb{R}^d$, the plot of $V_F\big(r_F(\mathbf{x})\big)$ against $V_G\big(r_G(\mathbf{x})\big)$ becomes concentrated along 45 degree through the origin and exhibits a noticeable departure from 45 degree line if they differ. Based on this, we can assign $\mathbf{x}$ to F if $V_F\big(r_F(\mathbf{x})\big) \leqslant V_G\big(r_G(\mathbf{x})\big)$, and vice versa. This is analogous to classification method based on DD plot. Alternatively, suppose F and G are elliptically symmetric distributions and differ in location and scale, Guha and Chakraborty (2013) have shown that $V_G(r) = kV_F(r)$ for some $k > 0$, $0 \leqslant r < 1$ (See Theorem 3.1 of the paper) and the slope of $V_G(r) = kV_F(r)$ is determined as ratio of the determinants of the scale matrices associated with F and G. Then for a known and defined $k > 0$, the scale-scale plot is a straight line.

Suppose $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^d$ have distributions F and G respectively, which are elliptically symmetric. If $\mathbf{Y} = \mathbf{AX} + \boldsymbol{\theta}$ for some $d \times d$ non-singular matrix $\mathbf{A}$ and d-dimensional vector $\boldsymbol{\theta}$, $\boldsymbol{\Sigma}_G = \mathbf{A}\boldsymbol{\Sigma}_F\mathbf{A}^T$, $(\mathbf{Y} - \boldsymbol{\theta}_G)^T\boldsymbol{\Sigma}_G^{-1}(\mathbf{Y} - \boldsymbol{\theta}_G) = (\mathbf{X} - \boldsymbol{\theta}_F)^T\boldsymbol{\Sigma}_F^{-1}(\mathbf{X} - \boldsymbol{\theta}_F)$ and by Lemma 2.4.1, $||\text{rank}_F(\mathbf{x})|| = ||\text{rank}_G(\mathbf{Ax} + \boldsymbol{\theta})||$. It then follows from Theorem 2.2 of Guha and Chakraborty (2013),

$$\begin{aligned} V_G\big(r_G(\mathbf{Ax} + \boldsymbol{\theta})\big) &= \frac{\pi^{d/2}|\boldsymbol{\Sigma}_G|^{1/2}\big(r_G(\mathbf{Ax} + \boldsymbol{\theta})\big)^d}{\mathit{\Gamma}(\frac{d}{2} + 1)} \\ &= \frac{|\boldsymbol{\Sigma}_G|^{1/2}}{|\boldsymbol{\Sigma}_F|^{1/2}} \frac{\pi^{d/2}|\boldsymbol{\Sigma}_F|^{1/2}\big(r_F(\mathbf{x})\big)^d}{\mathit{\Gamma}(\frac{d}{2} + 1)} = \frac{|\boldsymbol{\Sigma}_G|^{1/2}}{|\boldsymbol{\Sigma}_F|^{1/2}} V_F\big(r_F(\mathbf{x})\big) \end{aligned} \tag{3.4.1}$$

where $\boldsymbol{\Sigma}_X$ and $\boldsymbol{\Sigma}_Y$ are positive definite. From numerical example in Section 3.3 above, $|\boldsymbol{\Sigma}_2|^{1/2} = \sigma^2$, $|\boldsymbol{\Sigma}_1|^{1/2} = 1$ and so $k = \sigma^2$. The separating hyperplane between F and G is the line that passes through $V_G\big(r_G(\mathbf{x})\big) = \sigma^2 V_F\big(r_F(\mathbf{x})\big)$.

The regions of classification of $\mathbf{x}$ to F and to G denoted by R_F and R_G respectively, are defined as

$$\begin{aligned} R_F &: V_F\big(r_F(\mathbf{x})\big)/V_G\big(r_G(\mathbf{x})\big) \leqslant 1/k \\ R_G &: V_F\big(r_F(\mathbf{x})\big)/V_G\big(r_G(\mathbf{x})\big) > 1/k. \end{aligned} \tag{3.4.2}$$

We denote this classifier by RRC-1. The Bayes equivalency of RRC-1 is examined by the theorem below.

Theorem 3.4.1 *Let f_1 and f_2 be the probability density functions of populations, π_1 and π_2 having elliptically symmetric distributions F and G repectively from the same family of multivariate distributions such that $G(\mathbf{x}) = F(\mathbf{x} - \boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \mathbb{R}^d$ is a location shift.*

Suppose $p_1 = p_2$ and Theorem 3.2.1 hold, then the Bayes rule is equivalent to

$$\begin{cases} |\mathbf{\Sigma}_F|^{-1/2}V_F\big(r_F(\mathbf{x})\big) \leqslant |\mathbf{\Sigma}_G|^{-1/2}V_G\big(r_G(\mathbf{x})\big) & \Rightarrow \textit{ assign } \mathbf{x} \textit{ to population } \pi_1 \\ |\mathbf{\Sigma}_F|^{-1/2}V_F\big(r_F(\mathbf{x})\big) > |\mathbf{\Sigma}_G|^{-1/2}V_G\big(r_G(\mathbf{x})\big) & \Rightarrow \textit{ assign } \mathbf{x} \textit{ to population } \pi_2 \end{cases}$$

Proof: It follows from the proof of Theorem 3.2.2 that

$$f_1(\mathbf{x}) \geqslant f_2(\mathbf{x}) \text{ implies } g_1\big(\psi_F^{-1}(V_F(r_F(\mathbf{x})))\big) \geqslant g_2\big(\psi_G^{-1}(V_G(r_G(\mathbf{x})))\big).$$

for some decreasing functions g_1 and g_2. By Lemma 2.4.1, $\text{rank}_F(\mathbf{x}) = \text{rank}_{F_0}\big(\mathbf{\Sigma}_F^{-\frac{1}{2}}(\mathbf{x} - \boldsymbol{\theta}_F)\big)$ and by equation (3.4.1), $V_F(r_F(\mathbf{x})) = |\mathbf{\Sigma}_F|^{-1/2}V_{F_0}(r_F(\mathbf{x}))$, where F_0 is the distribution of $\mathbf{\Sigma}_F^{-\frac{1}{2}}(\mathbf{x} - \boldsymbol{\theta}_F)$, which is spherically symmetric about $\mathbf{0}$ if $\mathbf{x} \in \pi_1$. Similarly, $\text{rank}_G(\mathbf{x}) = \text{rank}_{G_0}\big(\mathbf{\Sigma}_G^{-\frac{1}{2}}(\mathbf{x} - \boldsymbol{\theta}_G)\big)$ and $V_G(r_G(\mathbf{x})) = |\mathbf{\Sigma}_G|^{-1/2}V_{G_0}(r_G(\mathbf{x}))$, where G_0 is the distribution of $\mathbf{\Sigma}_G^{-\frac{1}{2}}(\mathbf{x} - \boldsymbol{\theta}_G)$. Then

$$f_1(\mathbf{x}) \geqslant f_2(\mathbf{x}) \text{ implies } g_1\big(\psi_F^{-1}(|\mathbf{\Sigma}_F|^{-1/2}V_{F_0}(r_F(\mathbf{x})))\big) \geqslant g_2\big(\psi_G^{-1}(|\mathbf{\Sigma}_G|^{-1/2}V_{G_0}(r_G(\mathbf{x})))\big).$$

For $G(\mathbf{x}) = F(\mathbf{x} - \boldsymbol{\theta})$,

$$f_1(\mathbf{x}) = g\big(\psi^{-1}(|\mathbf{\Sigma}_F|^{-1/2}.V_{F_0}(r_F(\mathbf{x})))\big)$$

and

$$f_2(\mathbf{x}) = g\big(\psi^{-1}(|\mathbf{\Sigma}_G|^{-1/2}.V_{G_0}(r_G(\mathbf{x})))\big).$$

Also, $g \circ \psi^{-1}$ is a decreasing function since ψ is an increasing function and g is a decreasing function. Then $kV_{F_0}(r_F(\mathbf{x})) \leqslant V_{G_0}(r_G(\mathbf{x}))$. By Theorem 3.2.2 for $G(\mathbf{x}) = F(\mathbf{x} - \boldsymbol{\theta})$ and $\mathbf{\Sigma}_F = \mathbf{\Sigma}_G = \mathbf{\Sigma}$, $f_1(\mathbf{x}) > f_2(\mathbf{x})$ implies $V_F(r_F(\mathbf{x}))/V_G(r_G(\mathbf{x})) < 1$, which is equivalent to

$V_{F_0}(r_F(\mathbf{x}))/V_{G_0}(r_G(\mathbf{x})) < 1$ under this setting. Hence

$$kV_F(r_F(\mathbf{x})) \leqslant V_G(r_G(\mathbf{x})),$$

where $k = |\mathbf{\Sigma}_G|^{1/2}/|\mathbf{\Sigma}_F|^{1/2}$. The proof is complete. □

Estimates of $\mathbf{\Sigma}_F$ and $\mathbf{\Sigma}_G$ based on the moment of F and G respectively can be used but in order to get robust estimates from the training samples, minimum covariance determinant (MCD) estimates of $\mathbf{\Sigma}_F$ and $\mathbf{\Sigma}_G$ are used (see, for example, Rousseeuw and Leroy, 1987 for detail).

3.5 Numerical Example: Real data

We analyse seven benchmark data set to illustrate the performances of our methods (RC, AIRC, RRC and RRC-1). These datasets include iris data, Pima Indians diabetes (PID) data, banknote data, biomedical data, yeast data, cloud data and seed data. Iris data set (Fisher, 1936) contains three classes of Iris plants data. The classes are Iris setosa, Iris versicolor and Iris virginica. A training sample of size 30 and validation sample of size 20 are chosen for each of the three groups with four features (sepal length, sepal width, petal length and petal width, all measured in cm). Pima Indian diabetes (PID) data set, owned by the National Institute of Diabetes and Digestive and Kidney Diseases, consists of two groups ("tested negative" and "tested positive"). Training samples and validation samples of sizes 100 are chosen for each of the two groups. Banknote authentication dataset (Lohweg, 2013), denoted by banknote, consists of two classes which are "genuine"(size = 762) and "forged"(size = 610) with four input features. The features are variance of wavelet transformed image, skewness of wavelet transformed image, kurtosis of wavelet transformed image and entropy of image. A training sample of size 100 and a validation sample of size 100 from each of the two classes are chosen. Biomedical data

(Cox, Johnson and Kafadar, 1982) consists of two classes with eight features. The classes are "Carrier"(size = 67) and "Normal"(size = 127) after deleting the 15 observations with missing values. Four features (ml, m2, m3 and m4), a training sample of size 50 and a validation sample of size 17 from each of the two classes are chosen. Yeast data (Nakai, 1991) consist of ten categories, of which we choose four classes, which are CYT(size = 463), NUC(size = 329), MIT(size = 244) and Others(size = 348). Class "Others" consists of ME3, ME2, ME1, EXC, VAC, POX and ERL, each of which has small sample size. A training sample of size 100 and a validation sample of size 100 from each of the four classes are chosen for four variables (mcg, gvh, alm and mit). Cloud data (Miller et al., 1979) consist of period rainfalls in inches collected in a cloud-seeding experiment in Tasmania between mid-1964 and January 1971. It consists of two classes, which are "seeded" and "unseeded" with class size 54 each and seven features. Four features (TE, TW, NC and SC), a training sample of size 40 and a validation sample of size 14 from each of the four classes are chosen. Seed data (Charytanowicz et al., 2012) consists of 3 classes with 70 observations each. A training sample of size 50 and a validation sample of size 20 from each of the two classes are used. We remove one of the two linearly dependent features from the data and use the remaining six input features. Pima Indian diabetes data, banknote data, yeast data and seed data are taken from UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets.html) while biomedical data and Cloud data are taken from StatLib Datasets Archive (http://lib.stat.cmu.edu/datasets/) except iris data, which is inbuilt in R. The summary of the training and validation samples is presented in Table 3.2, where d is the dimension of the data and k is number of groups considered.

We use MCD estimates of covariance via R package *robustbase* with $\alpha = 0.70$. For depth classifier, the experiment is repeated 100 times and average probability of correct classification is computed. For each of the datasets, we assume equal prior probabilities for competing classes. Table 3.2 presents the result of analysis of data. Iris data is known

Table 3.2: Information about real data set.

Dataset	training sample	validation sample	d	k
iris	$n_1 = n_2 = n_3 = 30$	$m_1 = m_2 = m_3 = 20$	4	3
Pima Indian diabetes (PID)	$n_1 = n_2 = 100$	$m_1 = m_2 = 100$	4	2
Banknote	$n_1 = n_2 = 100$	$m_1 = m_2 = 100$	4	2
Biomedical	$n_1 = n_2 = 50$	$m_1 = m_2 = 17$	4	2
Yeast	$n_1 = n_2 = n_3 = n_4 = 100$	$m_1 = m_2 = m_3 = m_4 = 100$	4	4
Cloud	$n_1 = n_2 = 40$	$m_1 = m_2 = 14$	4	2
Seed	$n_1 = n_2 = n_3 = 50$	$m_1 = m_2 = m_3 = 20$	6	3

to be normally distributed for which QDA is the optimal. AIRC and RRC-1 have the same misclassification error as QDA while RC has the same misclassification error as LDA. RRC and maximum depth classifier based on Oja depth has highest misclassification error. For biomedical data, QDA, AIRC and RRC-1 have least misclassification error while P-D, RC and RRC perform like LDA with relatively high misclassification error (= 0.2059). For Pima Indian diabetes data, LDA appears to perform best (with error = 0.26) while QDA and RRC-1 (with error = 0.28) perform very close to LDA. For cloud data, AIRC, RRC and RRC-1 outperform others though misclassification errors associated with each of the competing classifiers are generally high. AIRC and RRC-1 compete favourably with all other classifiers while RRC-1 outperforms other classifiers for banknote authentication data, followed by LDA, QDA, AIRC and maximum depth classifier based on Oja depth, then RRC and maximum depth classifier based on projection depth while RC has the highest error. For seed data, both RRC-1 AIRC and LDA outperforms others while RRC-1 has the lowest misclassification error. Misclassification error could not be computed for maximum depth classifier based on Oja depth for seed data because of dimensionality.

From the analysis of these data sets, AIRC and RRC-1 seem to be better than RC and RRC in terms of misclassification error. Performance of all these classifiers on the simulation and benchmark data sets is fairly competitive, compare with some parametric and nonparametric classifiers. In most of the data sets, RRC-1 generates smaller error

Table 3.3: Performance of classifiers based on real data.

	Comparison of classifiers based on misclassification errors							
Dataset	LDA	QDA	O-D	P-D	RC	AIRC	RRC	RRC-1
iris	0.0333	**0.0167**	0.1667	0.0222	0.0333	**0.0167**	0.1667	**0.0167**
PID	**0.2600**	0.2800	0.5000	0.4410	0.3600	0.2850	0.4500	0.2800
Banknote	0.0200	0.0200	0.0200	0.0446	0.2050	0.0200	0.0350	**0.0100**
Biomedical	0.2059	**0.1471**	0.5000	0.2239	0.2059	**0.1471**	0.2059	**0.1471**
Yeast	0.4950	**0.3600**	0.6450	0.5272	0.5500	0.3875	0.3850	0.3775
Cloud	0.5714	0.4286	0.5000	0.4443	0.4643	**0.3571**	**0.3571**	**0.3571**
Seed	0.1167	0.1333	-	0.2387	0.2500	0.1167	0.3167	**0.1000**

rates than AIRC. In terms of computational simplicity, it has a clear edge over maximum depth classifiers when dimension is greater than 2, especially when any of half space depth, simplicial depth, simplicial volume depth and Oja depth is used.

3.6 Classification for More Than Two Classes

Another property of classifier based on volume of central rank region is that it readily lends itself to multiclass extension. Suppose there are $J(>2)$ populations, then assign $\mathbf{x}$ to population π_k with distribution F_k, $1 \leqslant k \leqslant J$ if

$$|\boldsymbol{\Sigma}_k|^{-1/2} V_{F_k}\big(r_{F_k}(\mathbf{x})\big) = \min_j |\boldsymbol{\Sigma}_j|^{-1/2} V_{F_j}\big(r_{F_j}(\mathbf{x})\big),$$

where $F_1, F_2, \ldots, F_J$ are absolutely continuous distributions corresponding to J populations and $\boldsymbol{\Sigma}_j$ is covariance matrix of jth population and $r_{F_j}(\mathbf{x}) = \|\text{rank}_{F_j}(\mathbf{x})\|$. Its associated total probability of misclassification is

$$\Delta = \sum_{j=1}^{J} p_j P\{|\boldsymbol{\Sigma}_k|^{-1/2} V_{F_k}\big(r_{F_k}(\mathbf{x})\big) \leqslant |\boldsymbol{\Sigma}_j|^{-1/2} V_{F_j}\big(r_{F_j}(\mathbf{x})\big) \ \mid\ \mathbf{x} \in F_j\},$$

where $p_1, p_2, \ldots, p_J$ are prior probability of populations $\pi_1, \pi_2, \ldots, \pi_J$ respectively.

Chapter 4

Classifier Based on Distribution of Multivariate Rank

In the previous chapter, we have defined spatial rank function and volume of central rank region, and then proposed classification methods based on outlyingness of spatial rank and volume of central rank region. Here in this chapter we will be proposing another classification method based on distribution of outlyingness of spatial rank and its variants. In the next section we will define the distribution of outlyingness of spatial rank and propose the classification method for spherically symmetric family of distributions.

4.1 Definitions

Suppose π_1 and π_2 are two populations having distributions F and G respectively with equal prior probabilities p_1 and p_2, where F and G are absolutely continuous with respect to Lebesgue measure in $\mathbb{R}^d$. For $\mathbf{x} \in \mathbb{R}^d$, define outlyingness of $\text{rank}_F(\mathbf{x})$ as

$$r_F(\mathbf{x}) = ||\text{rank}_F(\mathbf{x})||$$

and outlyingness of $\mathrm{rank}_G(\mathbf{x})$ as

$$r_G(\mathbf{x}) = ||\mathrm{rank}_G(\mathbf{x})||.$$

The probability distribution of $r_F(\mathbf{X})$, denoted by $F_R(r)$, is defined as

$$F_R(r) = P\Big(r_F(\mathbf{X}) \leqslant r\Big) \tag{4.1.1}$$

and the probability distribution of $r_G(\mathbf{Y})$, denoted by $G_R(r)$, is defined as

$$G_R(r) = P\Big(r_G(\mathbf{Y}) \leqslant r\Big). \tag{4.1.2}$$

Following equations (4.1.1) and (4.1.2), $F_R(r)$ and $G_R(r)$ depend solely on r. Also, $F_R(r)$ and $G_R(r)$ are increasing functions of r.

The classification rule based on probability distribution of outlyingness of spatial rank is to assign an observation $\mathbf{x}$ into population π_1 if

$$F_R\big(r_F(\mathbf{x})\big) \leqslant G_R\big(r_G(\mathbf{x})\big) \tag{4.1.3}$$

otherwise, assign $\mathbf{x}$ to population π_2. We shall call this classification method minimal rank distribution classifier, denoted by RDC. The probability of misclassification corresponding to two populations, π_1 and π_2, denoted by Δ, is

$$\Delta = p_1 P\big(F_R\big(r_F(\mathbf{x})\big) > G_R\big(r_G(\mathbf{x})\big) \ \mid\ \mathbf{x} \in \pi_1\big) + p_2 P\big(F_R\big(r_F(\mathbf{x})\big) \leqslant G_R\big(r_G(\mathbf{x})\big) \ \mid\ \mathbf{x} \in \pi_2\big).$$

This classification method is completely data driven, easy to compute and can be extended to higher dimension. It easily lends itself to multiclass extension. Suppose there are $J(> 2)$ populations, then assign $\mathbf{x}$ to population π_k with distribution F_k and prior probability p_k, $1 \leqslant k \leqslant J$ if

$$F_R\big(r_{F_k}(\mathbf{x})\big) = \min_j F_R\big(r_{F_j}(\mathbf{x})\big),$$

where $F_1, F_2, \ldots, F_J$ are absolutely continuous distributions corresponding to J populations. The total probability of misclassification corresponding to J competing populations is

$$\Delta = \sum_{\substack{j=1 \\ j \neq k}}^{J} p_j P\{F_R\big(r_{F_k}(\mathbf{x})\big) \leqslant F_R\big(r_{F_j}(\mathbf{x})\big) \ \mid\ \mathbf{x} \in F_j\},$$

where $p_1, p_2, \ldots, p_J$ are prior probabilities corresponding to populations $\pi_1, \pi_2, \ldots, \pi_J$ respectively.

When only training samples $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m$ from π_1 and $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n$ from π_2 are available, then we replace the population version of the rank functions rank_F and rank_G by their empirical versions rank_{F_m} and rank_{G_n} respectively to construct the empirical classification rule. Define the outlyingness of spatial rank of $\mathbf{x}$ based on F_m and G_n as

$$\begin{aligned} r_{F_m}(\mathbf{x}) &= ||\text{rank}_{F_m}(\mathbf{x})|| \ \text{ based on } \ \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m \sim F_m, \\ r_{G_n}(\mathbf{x}) &= ||\text{rank}_{G_n}(\mathbf{x})|| \ \text{ based on } \ \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n \sim G_n, \end{aligned}$$

respectively. Then,

$$\begin{aligned} \widehat{F}_R\big(r_{F_m}(\mathbf{x})\big) &= \frac{1}{m}\sum_{i=1}^{m} I\{r_{F_m}(\mathbf{X}_i) \leqslant r_{F_m}(\mathbf{x})\} \\ \widehat{G}_R\big(r_{G_n}(\mathbf{x})\big) &= \frac{1}{n}\sum_{i=1}^{n} I\{r_{G_n}(\mathbf{Y}_i) \leqslant r_{G_n}(\mathbf{x})\} \end{aligned}$$

where I is the indicator function. It is shown in Theorem 2.2.2 that $||\text{rank}_{F_m}(\mathbf{x})||$ and $||\text{rank}_{G_n}(\mathbf{x})||$ converge to respective rank functions $||\text{rank}_F(\mathbf{x})||$ and $||\text{rank}_G(\mathbf{x})||$ almost surely. The classification rule based on empirical distribution functions of $r_{F_m}(\mathbf{x})$ and $r_{G_n}(\mathbf{x})$ is to assign $\mathbf{x}$ to π_1 if

$$\widehat{F}_R\big(r_{F_m}(\mathbf{x})\big) < \widehat{G}_R\big(r_{G_n}(\mathbf{x})\big),$$

otherwise to π_2. The misclassification error associated with classifier based on empirical distribution functions, denoted by $\widehat{\Delta}_N$, is

$$\widehat{\Delta}_N = \frac{p_1}{m}\sum_{i=1}^{m} I\{\widehat{F}_R\big(r_{F_m}(\mathbf{x}_i)\big) > \widehat{G}_R\big(r_{G_n}(\mathbf{x}_i)\big)|\mathbf{x}_i \in \pi_1\}$$
$$+ \frac{p_2}{n}\sum_{i=1}^{n} I\{\widehat{F}_R\big(r_{F_m}(\mathbf{y}_i)\big) \leqslant \widehat{G}_R\big(r_{G_n}(\mathbf{y}_i)\big)|\mathbf{y}_i \in \pi_2\}.$$

4.2 Properties of Minimal Rank Distribution Classifier

Suppose the probability measure μ of the data in $\mathbb{R}^d$ has absolutely continuous distribution function, F possessing a probability density function $f(\mathbf{x})$, which is positive for all $\mathbf{x}$ in the support of distribution F and $r_F(\mathbf{x}) \in [0, 1]$ is continuous in $\mathbf{x}$. The classification method based on minimal rank distribution classifier is Bayes rule for spherically symmetric families of distributions that differ in location. This is formally stated in Theorem 4.2.1 below.

Theorem 4.2.1 *Let f_1 and f_2 be the probability density functions of populations, π_1 and π_2 having spherically symmetric distributions F and G in $\mathbb{R}^d$ about $\boldsymbol{\theta}_F$ and $\boldsymbol{\theta}_G$ respectively from the same family of multivariate distributions such that $G(\mathbf{x}) = F(\mathbf{x} - \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a location shift in $\mathbb{R}^d$. If $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$ and $p_1 = p_2 = \frac{1}{2}$, then the Bayes rule is*

$$\textit{assign } \mathbf{x} \textit{ to population } \pi_1 \textit{ if } F_R\big(r_F(\mathbf{x})\big) \leqslant G_R\big(r_G(\mathbf{x})\big)$$

and

$$\textit{assign } \mathbf{x} \textit{ to population } \pi_2 \textit{ if } F_R\big(r_F(\mathbf{x})\big) > G_R\big(r_G(\mathbf{x})\big).$$

Proof: It follows from the proof of Theorem 2.2.1 that for some increasing functions h_1 and h_2, $\|\text{rank}_F(\mathbf{x})\| = h_1(\|\mathbf{x} - \theta_F\|)$ and $\|\text{rank}_G(\mathbf{x})\| = h_2(\|\mathbf{x} - \boldsymbol{\theta}_G\|)$. It follows from equation (4.1.1) that $F_R\big(r_F(\mathbf{x})\big)$ depends only on $r_F(\mathbf{x})$. So, define $F_R\big(r_F(\mathbf{x})\big) = \psi_1(\|\text{rank}_F(\mathbf{x})\|) = \psi_1(h_1(\|\mathbf{x} - \theta_F\|))$ and similarly, $G_R\big(r_G(\mathbf{x})\big) = \psi_2(\|\text{rank}_G(\mathbf{x})\|) = \psi_2(h_2(\|\mathbf{x} - \boldsymbol{\theta}_G\|))$ for some increasing functions ψ_1 and ψ_2, then

$$\|\mathbf{x} - \theta_F\| = \varpi_1^{-1}\big(F_R\big(r_F(\mathbf{x})\big)\big) \text{ and } \|\mathbf{x} - \theta_G\| = \varpi_2^{-1}\big(G_R\big(r_G(\mathbf{x})\big)\big),$$

where $\varpi_1 = \psi_1 \circ h_1$ and $\varpi_2 = \psi_2 \circ h_2$ are increasing functions. Therefore, we can write $f_1(\mathbf{x}) = g_1\big(\varpi_1^{-1}(F_R(r_F(\mathbf{x}))\big)$ and $f_2(\mathbf{x}) = g_2\big(\varpi_2^{-1}(G_R(r_G(\mathbf{x}))\big)$. Now,

$$f_1(\mathbf{x}) \geqslant f_2(\mathbf{x}) \text{ implies } g_1\big(\varpi_1^{-1}(F_R(r_F(\mathbf{x})))\big) \geqslant g_2\big(\varpi_2^{-1}(G_R(r_G(\mathbf{x})))\big).$$

$g_1 \circ \varpi_1^{-1}$ and $g_2 \circ \varpi_2^{-1}$ are decreasing functions since ϖ_1 and ϖ_2 are increasing functions and g_1 and g_2 are decreasing functions. Given that $G(\mathbf{x}) = F(\mathbf{x} - \boldsymbol{\theta})$, then $g_1 = g_2$ and $\varpi_1 = \varpi_2$ and as a result, we get $F_R\big(r_F(\mathbf{x})\big) \leqslant G_R\big(r_G(\mathbf{x})\big)$ whenever $f_1(\mathbf{x}) \geqslant f_2(\mathbf{x})$. The proof is complete. □

Theorem 4.2.2 *Let $r_F(\mathbf{X}_1), r_F(\mathbf{X}_2), \ldots, r_F(\mathbf{X}_n)$ be independent and identically distributed real valued random variables with probability distribution function $F_R(r) = P\big(r_F(\mathbf{X}_1) \leqslant r\big)$. Define the standard empirical distribution of $r_F(\mathbf{X})$ as*

$$\widehat{F}_R(r) = \frac{1}{n}\sum_{i=1}^{n} I\{r_F(\mathbf{X}_i) \leqslant r\}.$$

Then

$$\sup_{r\in(0,1)} |F_R(r) - \widehat{F}_R(r)| = 0$$

with probability one as $n \longrightarrow \infty$.

Proof: This theorem is an extension of Glivenko-Cantelli theorem. In this theorem, outlyingness of spatial rank of each observation in the training sample is treated as independent and identically distributed real-valued random variable since $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ are independent and identically distributed real-valued random variables. Glivenko-Cantelli theorem is proved in Durrett (2010) (See Theorem 2.4.7 and its proof in pp. 76). □

Theorem 4.2.3 *Let F be a d-variate distribution function, which is absolutely continuous. Suppose Theorem 2.2.2 hold, then for sufficiently large n*

$$\mid \widehat{F}_R\big(r_{F_n}(\mathbf{x})\big) - F_R\big(r_F(\mathbf{x})\big) \mid \ \overset{a.s.}{\to}\ 0.$$

Proof:

$$\begin{aligned}\big|\widehat{F}_R\big(r_{F_n}(\mathbf{x})\big) - F_R\big(r_F(\mathbf{x})\big)\big| &\leqslant \big|\widehat{F}_R\big(r_{F_n}(\mathbf{x})\big) - \widehat{F}_R\big(r_F(\mathbf{x})\big)\big| \\ &\quad + \ \big|\widehat{F}_R\big(r_F(\mathbf{x})\big) - F_R\big(r_F(\mathbf{x})\big)\big| = S_1 + S_2\end{aligned}$$

The almost sure convergence of the sequence S_1 is proved in Guha (2012) (See Lemma 3.1.2 of Guha, 2012). This is equivalent to saying

$$\widehat{F}_R\big(r_{F_n}(\mathbf{x})\big) - \widehat{F}_R\big(r_F(\mathbf{x})\big) = o(e^{-n}) \ \text{ as } \ n \to \infty \ \ w.p. \ 1.$$

By Theorem 4.2.2,

$$S_2 = \big|\widehat{F}_R\big(r_F(\mathbf{x})\big) - F_R\big(r_F(\mathbf{x})\big)\big| \longrightarrow 0$$

with probability one as $n \to \infty$. The proof is complete. □

Theorem 4.2.4 *Suppose f_1 and f_2 are the probability density functions of populations, π_1 and π_2 having spherically symmetric distributions F and G in $\mathbb{R}^d$ about $\boldsymbol{\theta}_F$ and $\boldsymbol{\theta}_G$ respectively with equal prior probabilities $p_1 = p_2 = \frac{1}{2}$, $F_R\big(r_F(\mathbf{x})\big)$ and $G_R\big(r_G(\mathbf{x})\big)$ are continuous and satisfy*

$$\widehat{F}_R\big(r_{F_m}(\mathbf{x})\big) \overset{a.s.}{\to} F_R\big(r_F(\mathbf{x})\big) \text{ and } \widehat{G}_R\big(r_{G_n}(\mathbf{x})\big) \overset{a.s.}{\to} G_R\big(r_G(\mathbf{x})\big)$$

respectively, as $\min(m,n) \to \infty$, then the total probability of misclassification for the classification rule based on the training sample,

$$\widehat{\Delta}_N \longrightarrow \Delta_B \text{ a.s. for } \min(m,n) \to \infty$$

where Δ_B is the Bayes risk.

Proof: Suppose $\mathbf{x}_i \in \pi_1$ and $\mathbf{y}_i \in \pi_2$. Define

$$\widehat{\Delta} = \frac{p_1}{m}\sum_{i=1}^{m} I\{F_R\big(r_F(\mathbf{x}_i)\big) > G_R\big(r_G(\mathbf{x}_i)\big)\} + \frac{p_2}{n}\sum_{i=1}^{n} I\{F_R\big(r_F(\mathbf{y}_i)\big) \leqslant G_R\big(r_G(\mathbf{y}_i)\big)\}$$

It follows that

$$|\widehat{\Delta}_N - \Delta| \leqslant |\widehat{\Delta}_N - \widehat{\Delta}| + |\widehat{\Delta} - \Delta|.$$

Define

$$|\widehat{\Delta}_N - \widehat{\Delta}| \leqslant S_{10} + S_{20}.$$

where

$$S_{10} = \left|\frac{p_1}{m}\left[\sum_{i=1}^{m} I\{\widehat{F}_R\big(r_{F_m}(\mathbf{x}_i)\big) > \widehat{G}_R\big(r_{G_n}(\mathbf{x}_i)\big)\} - \sum_{i=1}^{m} I\{F_R\big(r_F(\mathbf{x}_i)\big) > G_R\big(r_G(\mathbf{x}_i)\big)\}\right]\right|$$

and

$$S_{20} = \left|\frac{p_2}{n}\left[\sum_{i=1}^{n} I\{\widehat{F}_R\big(r_{F_m}(\mathbf{y}_i)\big) \leqslant \widehat{G}_R\big(r_{G_n}(\mathbf{y}_i)\big)\} - \sum_{i=1}^{n} I\{F_R\big(r_F(\mathbf{y}_i)\big) \leqslant G_R\big(r_G(\mathbf{y}_i)\big)\}\right]\right|$$

By Theorem 2.2.2,

$$\sup_z |r_{F_m}(\mathbf{z}) - r_F(\mathbf{z})| \overset{a.s.}{\to} 0 \text{ and } \sup_z |r_{G_n}(\mathbf{z}) - r_G(\mathbf{z})| \overset{a.s.}{\to} 0$$

are satisfied. Following this and applying Theorem 4.2.3, $S_{10} \to 0$ as $m \to \infty$ and $S_{20} \to 0$ as $n \to \infty$. Hence $|\widehat{\Delta}_N - \widehat{\Delta}| \to 0$ almost surely.

Similarly, define

$$|\widehat{\Delta} - \Delta| \leqslant T_{10} + T_{20}$$

where

$$T_{10} = \left|\frac{p_1}{m}\sum_{i=1}^{m} I\{F_R\big(r_F(\mathbf{x}_i)\big) > F_R\big(r_G(\mathbf{x}_i)\big)\} - p_1 P\Big(F_R\big(r_F(\mathbf{x})\big) > G_R\big(r_G(\mathbf{x})\big)\Big)\right|$$

and

$$T_{20} = \left|\frac{p_2}{n}\sum_{i=1}^{n} I\{F_R\big(r_F(\mathbf{y}_i)\big) \leqslant G_R\big(r_G(\mathbf{y}_i)\big)\} - p_2 P\Big(F_R\big(r_F(\mathbf{y})\big) \leqslant G_R\big(r_G(\mathbf{y})\big)\Big)\right|$$

By strong law of large number,

$$\frac{1}{m}\sum_{i=1}^{m} I\{F_R\big(r_F(\mathbf{x}_i)\big) > F_R\big(r_G(\mathbf{x}_i)\big)\} \overset{a.s.}{\to} P\Big(F_R\big(r_F(\mathbf{x})\big) > G_R\big(r_G(\mathbf{x})\big)\Big)$$

for all continuous points of $\mathbf{x}$ in F and hence $T_1 \to 0$ as $m \to \infty$. Similarly,

$$\frac{1}{n}\sum_{i=1}^{n} I\{F_R\big(r_F(\mathbf{y}_i)\big) \leqslant G_R\big(r_G(\mathbf{y}_i)\big)\} \overset{a.s.}{\to} P\Big(F_R\big(r_F(\mathbf{x})\big) \leqslant G_R\big(r_G(\mathbf{x})\big)\Big)$$

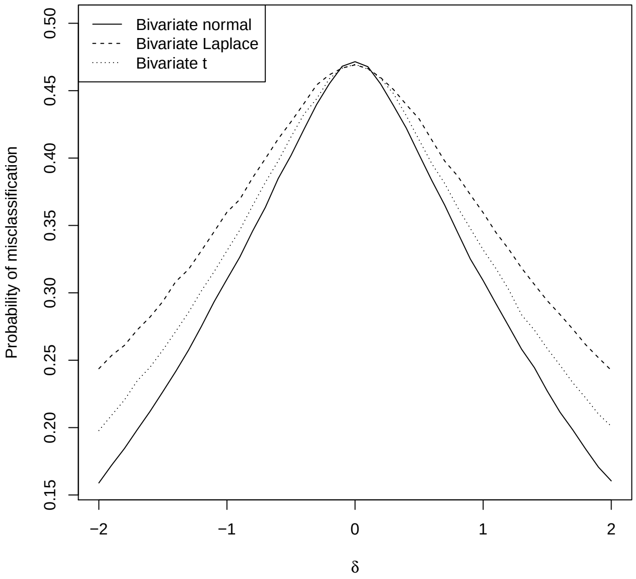


Figure 4.1: Misclassification rates for 3 different families of distributions using RDC for location shift only ($\mathbf{\Sigma} = \mathbf{I}_2, \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$).

for all continuous points of $\mathbf{x}$ in G and hence $T_2 \to 0$ as $n \to \infty$. Hence $|\widehat{\Delta} - \Delta| \to 0$ almost surely.

Following these arguments,

$$\widehat{\Delta}_N \longrightarrow \Delta \text{ with probability one as } \min(m, n) \to \infty.$$

From Theorem 4.2.1, RDC is Bayes rule under the conditions of Theorem 4.2.4. Hence

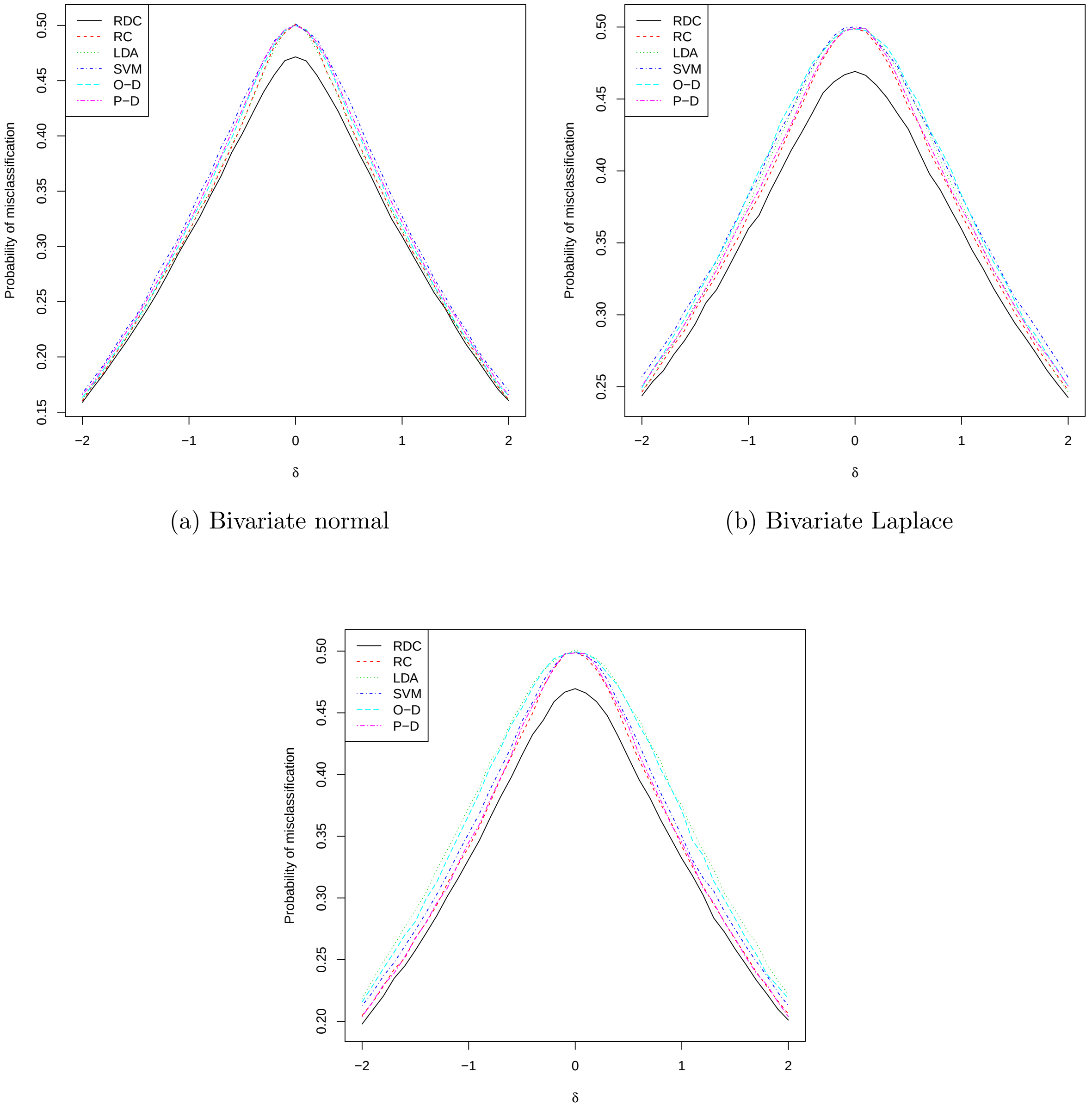
$$\widehat{\Delta}_N \longrightarrow \Delta = \Delta_B$$

with probability one. □

4.2.1 Numerical Example: Simulation

We want to investigate the performance of RDC using numerical example in Section 2.3. Suppose $\mathbf{X}$ and $\mathbf{Y}$ are bivariate normally distributed samples with $\mathbf{X} \sim N_d(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $\mathbf{Y} \sim N_d(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, with sizes n_1 and n_2 respectively. We classify m observations from each of the distributions and compute the probability of misclassification associated with RDC. Suppose $n_1 = n_2 = m = 100$, $\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and $\boldsymbol{\mu}_2$ is chosen such $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \delta^2$. For $\delta \in [-2, 2]$ and $\boldsymbol{\Sigma} = I_2$, we make plots of estimates of associated misclassification rates. The results are thereafter compared with the result from existing methods. We repeat the simulation process for bivariate t distributed samples with 3 degree of freedom and bivariate Laplace distributed samples.

Figure 4.1 shows the plot of empirical misclassification rates against the non-centrality parameter $\delta = \sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}$ for three bivariate spherically symmetric distributions using RDC. Misclassification rates are least in bivariate normally distributed samples and highest in bivariate Laplace distributed samples among the three distributions given that the competing classes have equal scale, as seen in Figure 4.1. This implies that the probability distribution of the populations from which samples are taken have implication on the misclassification rates, just as the case with RC (See Figure 2.1a). The misclassification rates at $\delta = 0$ is half for the three families of distributions because distributions of $\mathbf{X}$ and $\mathbf{Y}$ are the same at this value. As δ goes away from 0, the distinction between the two classes become clearer and misclassification error decreases as $|\delta|$ increases for each of the three distributions. Figure 4.2 compares RDC with some existing methods (Fisher's LDA, support vector machine(SVM), maximum depth classifier based on Oja depth (O-D), maximum depth classifier based on Projection depth (P-D) and minimal rank classifier (RC)). The figure shows that RDC performs well and competes favourably with other classifiers. In Figure 4.2(a), it is shown that RDC compares favorably with



(a) Bivariate normal

(b) Bivariate Laplace

(c) Bivariate t

Figure 4.2: Misclassification rates for spherically symmetric distributions with $\mathbf{\Sigma} = \mathbf{I}_2$ and $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$.

LDA, RC and other classifiers for bivariate normal samples. The misclassification rates of these classifiers are almost equivalent for each value of δ. This is similar for bivariate Laplace samples and for bivariate t samples as shown in Figure 4.2(b)-(c) respectively. RDC appears to have least misclassification errors among the classifiers for the three families of distributions.

To demonstrate robustness of minimal rank distribution classifier against deviation from the property of spherical symmetry, we use the information in above numerical example and assume $\mathbf{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. Table 4.1 shows that misclassification rates associated with RDC are not in any specific order of ρ. This means that RDC, like minimal rank classifier in Chapter 2, is not robust against the existence of correlation among variables in the population from which the sample is drawn. This is as a result of non-invariance property of spatial rank under affine transformation discussed in Section 2.4.1. In order to overcome this limitation, we suggest replacing affine non-invariant spatial rank by its affine invariant version.

4.3 Affine Invariant Version of Minimum Rank Distribution Based Classifier

Suppose $\mathbf{X} \in \mathbb{R}^d$ has a d-dimensional distribution F and $\mathbf{Y} \in \mathbb{R}^d$ has a d-dimensional distribution G, where F and G are elliptically symmetric and absolutely continuous with respect to Lebesgue measure in $\mathbb{R}^d$. The affine invariant spatial rank of $\mathbf{x} \in \mathbb{R}^d$ with respect to F is

$$\text{rank}^*_F(\mathbf{x}) = E_F\left(\frac{\{\mathbf{X}(\alpha)\}^{-1}[\mathbf{x}-\mathbf{X}]}{||\{\mathbf{X}(\alpha)\}^{-1}[\mathbf{x}-\mathbf{X}]||}\right)$$

and the affine invariant spatial rank of $\mathbf{x} \in \mathbb{R}^d$ with respect to G is

$$\text{rank}^*_G(\mathbf{x}) = E_G\left(\frac{\{\mathbf{Y}(\beta)\}^{-1}[\mathbf{x}-\mathbf{Y}]}{||\{\mathbf{Y}(\beta)\}^{-1}[\mathbf{x}-\mathbf{Y}]||}\right)$$

Table 4.1: Performance of RDC when $\mathbf{\Sigma} \neq \lambda \mathbf{I}_2$, $\lambda \in \mathbb{R}$.

Distribution	δ	LDA	Minimal rank distribution classifier(RDC)			
			$\rho = 0.00$	$\rho = 0.50$	$\rho = 0.75$	$\rho = 0.90$
Bivariate normal	0.0	0.5000	0.4702	0.4696	0.4704	0.4704
	0.5	0.4059	0.4031	0.4040	0.4031	0.4023
	1.0	0.3117	0.3107	0.3153	0.3155	0.3131
	1.5	0.2290	0.2283	0.2352	0.2345	0.2313
	2.0	0.1602	0.1583	0.1677	0.1674	0.1626
Bivariate Laplace	0.0	0.5000	0.4703	0.4703	0.4698	0.4699
	0.5	0.4361	0.4278	0.4265	0.4259	0.4259
	1.0	0.3577	0.3589	0.3587	0.3584	0.3568
	1.5	0.2960	0.2571	0.2602	0.2598	0.2570
	2.0	0.2434	0.2004	0.2043	0.2038	0.2008
Bivariate t	0.0	0.5000	0.4699	0.4701	0.4703	0.4702
	0.5	0.4216	0.4137	0.4121	0.4117	0.4111
	1.0	0.3347	0.3320	0.3333	0.3332	0.3311
	1.5	0.2618	0.2571	0.2602	0.2598	0.2570
	2.0	0.2018	0.2004	0.2043	0.2038	0.2008

where $\mathbf{X}(\alpha)$ is a transformation matrix, whose columns are $\mathbf{X}_{i_1} - \mathbf{X}_{i_0}, \mathbf{X}_{i_2} - \mathbf{X}_{i_0}, \ldots, \mathbf{X}_{i_d} - \mathbf{X}_{i_0}$ and $\mathbf{Y}(\beta)$ is a transformation matrix, whose columns are $\mathbf{Y}_{i_1} - \mathbf{Y}_{i_0}, \mathbf{Y}_{i_2} - \mathbf{Y}_{i_0}, \ldots, \mathbf{Y}_{i_d} - \mathbf{Y}_{i_0}$. We refer readers to Section 2.4 of this thesis for detail.

Define outlyingness of $\mathrm{rank}^*_F(\mathbf{X})$ and $\mathrm{rank}^*_G(\mathbf{X})$ as

$$r^*_F(\mathbf{x}) = ||\mathrm{rank}^*_F(\mathbf{x})|| \quad \text{and} \quad r^*_G(\mathbf{x}) = ||\mathrm{rank}^*_G(\mathbf{x})||$$

and the distribution functions of $r^*_F(\mathbf{x})$ and $r^*_G(\mathbf{x})$ as

$$F_R\big(r^*_F(\mathbf{x})\big) = P\big(r^*_F(\mathbf{X}) \leqslant r^*_F(\mathbf{x})\big) \quad \text{and} \quad G_R\big(r^*_G(\mathbf{x})\big) = P\big(r^*_G(\mathbf{Y}) \leqslant r^*_G(\mathbf{x})\big)$$

respectively. The classification rule based on distribution function of $r^*_F(\mathbf{x})$ and $r^*_G(\mathbf{x})$ is to assign $\mathbf{x}$ to π_1 if

$$F_R\big(r^*_F(\mathbf{x})\big) \leqslant G_R\big(r^*_G(\mathbf{x})\big)$$

otherwise, assign $\mathbf{x}$ to π_2. We shall call this classification method, minimal affine invariant rank distribution classifier (AIRDC).

Given any two populations π_1, with $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{n_1} \in \pi_1$ and π_2, with $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_{n_2} \in \pi_2$. For the training sample in the population π_1, let $\mathbf{X}_{i_0}, \ldots, \mathbf{X}_{i_d}$ be $d+1$ observations and $\alpha = \{i_0, i_1, \ldots, i_d\}$ denotes the set of $d+1$ indices. The affine invariant spatial rank function with respect to the training sample $\mathbf{X}_1, \ldots, \mathbf{X}_n \in \pi_2$ as

$$\text{rank}^*_F(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \frac{\{\mathbf{X}(\alpha)\}^{-1}(\mathbf{x} - \mathbf{X}_i)}{\|\{\mathbf{X}(\alpha)\}^{-1}(\mathbf{x}\mathbf{X}_i)\|}$$

and the affine invariant spatial rank function with respect to the training sample $\mathbf{Y}_1, \ldots, \mathbf{Y}_m \in \pi_2$ as

$$\text{rank}^*_G(\mathbf{x}) = \frac{1}{m} \sum_{i=1}^{m} \frac{\{\mathbf{Y}(\beta)\}^{-1}(\mathbf{x} - \mathbf{Y}_i)}{\|\{\mathbf{Y}(\beta)\}^{-1}(\mathbf{x} - \mathbf{Y}_i)\|}$$

where β is a set of $d+1$ indices $\{j_0, j_1, \ldots, j_d\}$ and $\mathbf{Y}(\beta)$ is the $d \times d$ matrix formed with the columns $\mathbf{Y}_{j_1} - \mathbf{Y}_{j_0}, \ldots, \mathbf{Y}_{j_d} - \mathbf{Y}_{j_0}$. Define the outlyingness of $\text{rank}^*_{F_m}(\mathbf{x})$ and $\text{rank}^*_{G_n}(\mathbf{x})$ as

$$\begin{aligned} r^*_{F_m}(\mathbf{x}) &= ||\text{rank}^*_{F_m}(\mathbf{x})|| \quad \text{based on} \quad \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m \ \text{ and} \\ r^*_{G_n}(\mathbf{x}) &= ||\text{rank}^*_{G_n}(\mathbf{x})|| \quad \text{based on} \quad \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n \end{aligned}$$

respectively, and the empirical distribution functions of $r^*_{F_n}(\mathbf{X})$ and $r^*_{G_m}(\mathbf{Y})$ as

$$\widehat{F}_R\big(r^*_{F_m}(\mathbf{x})\big) = \frac{1}{m} \sum_{i=1}^{m} I\{r^*_{F_m}(\mathbf{X}_i) \leqslant r^*_{F_m}(\mathbf{x})\} \ \text{ and } \ \widehat{G}_R\big(r^*_{G_n}(\mathbf{x})\big) = \frac{1}{n} \sum_{i=1}^{n} I\{r^*_{G_n}(\mathbf{Y}_i) \leqslant r^*_{G_n}(\mathbf{x})\}$$

respectively. The empirical classification rule based on empirical distribution functions of $r^*_{F_m}(\mathbf{x})$ and $r^*_{G_n}(\mathbf{x})$ is to assign $\mathbf{x}$ to π_1 if

$$\widehat{F}_R\big(r^*_{F_m}(\mathbf{x})\big) \leqslant \widehat{G}_R\big(r^*_{G_n}(\mathbf{x})\big)$$

Table 4.2: Comparison of AIRDC with some other classifiers based on average misclassification errors when $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{I}_2$ and $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$.

		Classifiers							
Distribution	δ	LDA	SVM	O-D	P-D	RC	AIRC	RDC	AIRDC
Bivariate	1	0.3148	0.3157	0.3181	0.3213	0.3129	0.3168	0.3104	**0.3096**
normal	2	0.1612	0.1602	0.1649	0.1660	0.1615	0.1623	0.1583	**0.1582**
Bivariate	1	0.3770	0.3814	0.3831	0.3729	0.3693	0.3727	**0.3585**	0.3598
Laplace	2	0.2464	0.2573	0.2503	0.2508	0.2475	0.2506	**0.2411**	0.2442
Bivariate t	1	0.3746	0.3505	0.3707	0.3400	0.3418	0.3475	0.3329	**0.3320**
	2	0.2220	0.2137	0.2185	0.2053	0.2060	0.2113	0.1983	**0.1970**

otherwise, assign $\mathbf{x}$ to π_2.

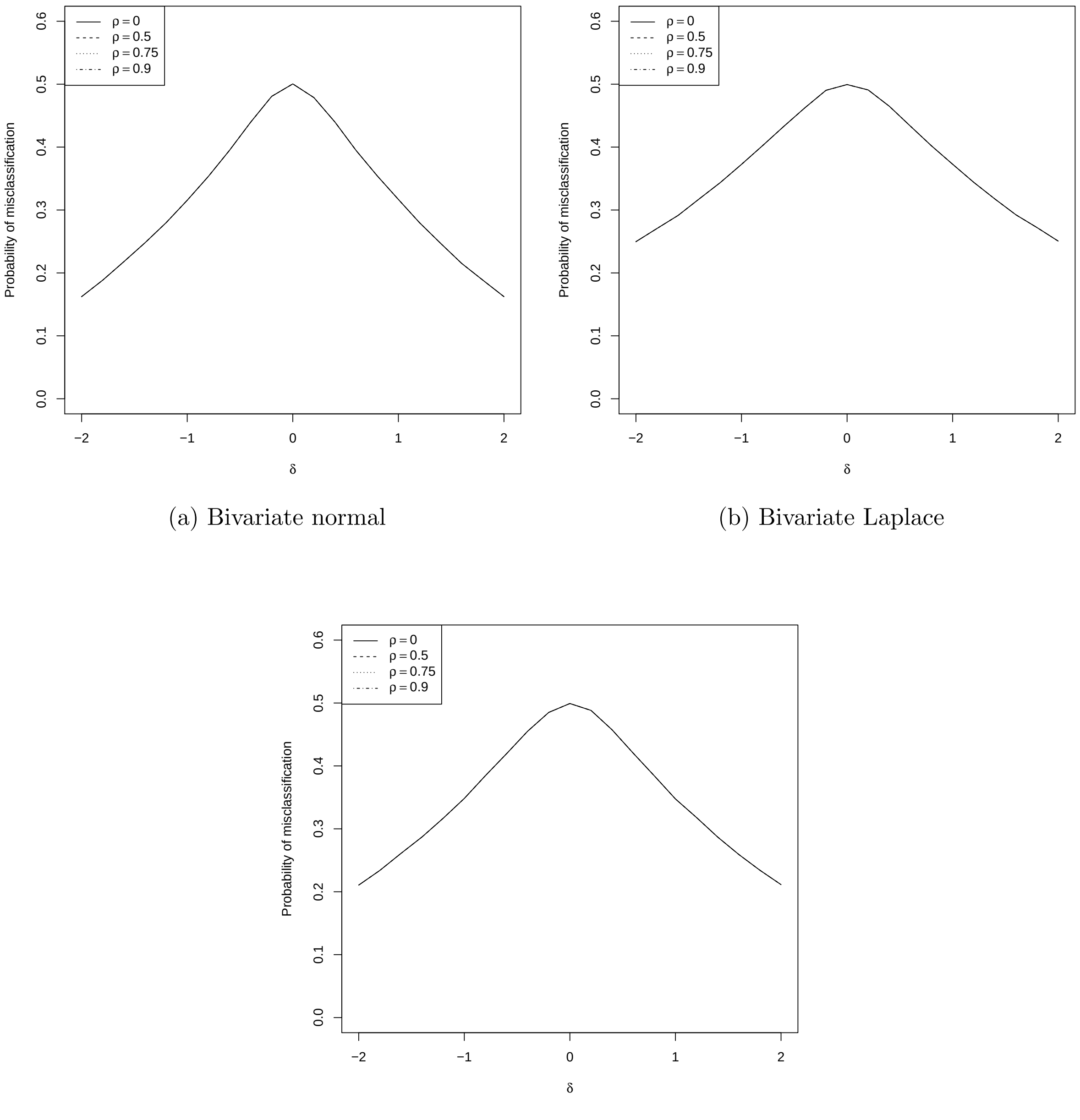
Theorem 4.3.1 *Let f_1 and f_2 be the probability density functions of populations, π_1 and π_2 having elliptically symmetric distributions F and G respectively from the same family of multivariate distributions such that $G(\mathbf{x}) = F(\mathbf{x} - \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a location parameter in $\mathbb{R}^d$. Suppose $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$, $p_1 = p_2$, then the Bayes rule is equivalent to*

$$\begin{cases} F_R\big(r^*_F(\mathbf{x})\big) \leqslant G_R\big(r^*_G(\mathbf{x})\big) & \Rightarrow \quad \textit{assign } \mathbf{x} \textit{ to population } \pi_1 \\ F_R\big(r^*_F(\mathbf{x})\big) > G_R\big(r^*_G(\mathbf{x})\big) & \Rightarrow \quad \textit{assign } \mathbf{x} \textit{ to population } \pi_2 \end{cases}$$

Proof: The proof is straight forward from the proofs of Theorem 2.4.1 and Theorem 4.2.1. □

4.3.1 Numerical Example : Simulation II

Here, we carry out simulation study on AIRDC for homogenous scale case and heterogenous scale case using simulation information in Section 2.3. Suppose π_1 has a distribution $N(\boldsymbol{\mu}_1, \mathbf{\Sigma}_1)$ and π_2 has a distribution $N(\boldsymbol{\mu}_2, \mathbf{\Sigma}_2)$, where $\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$. For homogenous scale case, $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. $\boldsymbol{\mu}_2$ is chosen in such a way that $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \delta^2$. The misclassifcation rates associated with AIRDC remain the same for various vales of ρ.



(a) Bivariate normal

(b) Bivariate Laplace

(c) Bivariate t

Figure 4.3: Robustness of AIRDC against deviation from spherical symmetry.

Table 4.3: Comparison of classifiers with theoretical Bayes rule.

			Average misclassification errors				
Distribution	δ	Bayes Risk	LDA	RC	AIRC	RDC	AIRDC
Bivariate normal	0.0	0.5000	0.5007	0.5014	0.5017	0.4718	0.4712
	0.5	0.4013	0.4049	0.4110	0.4176	0.3997	0.4046
	1.0	0.3085	0.3114	0.3148	0.3156	0.3104	0.3114
	1.5	0.2266	0.2294	0.2299	0.2309	0.2283	0.2280
	2.0	0.1587	0.1593	0.1612	0.1630	0.1583	0.1593
Bivariate Laplace	0.0	0.5000	0.4987	0.4987	0.5005	0.4704	0.4705
	0.5	0.4347	0.4558	0.4444	0.4519	0.4269	0.4313
	1.0	0.3576	0.3770	0.3693	0.3705	0.3585	0.3606
	1.5	0.2947	0.3048	0.3012	0.3049	0.2945	0.2953
	2.0	0.2415	0.2464	0.2475	0.2487	0.2411	0.2440
Bivariate t	0.0	0.5000	0.5009	0.4994	0.4980	0.4685	0.4721
	0.5	0.4231	0.4565	0.4309	0.4385	0.4137	0.4209
	1.0	0.3339	0.3746	0.3418	0.3484	0.3329	0.3350
	1.5	0.2612	0.2901	0.2663	0.2721	0.2598	0.2634
	2.0	0.2019	0.2220	0.2060	0.2106	0.1983	0.2047

This can be seen in Figure 4.3. We compare AIRDC with other classifiers, which include LDA, SVM, Maximum depth classifiers based on projection depth (P-D) and Oja depth (O-D), RC, RDC and AIRC. Consider $\rho = 0.0$, Table 4.2 gives the comparison of classifiers for $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{I}_2$ and $\delta = 1, 2$. It is seen that RDC and AIRDC compete favourably with other classifiers. For non-normal samples, RDC and AIRDC have noticeable lower misclassification errors.

We want to compare RC, AIRC, RDC and AIRDC with theoretical Bayes rule given that the competing classes normally distributed, bivariate t distributed (with 3 degrees of freedom) and bivariate Laplace distributed, based on the information in numerical example in Subsection 1.2.3 in order to confirm numerically Theorem 2.2.1, Theorem 2.4.1, Theorem 4.2.1 and Theorem 4.3.1 for RC, AIRC, RDC and AIRDC respectively. This is given in the Table 4.3. Numerical results confirm the theoretical results that RC, AIRC, RDC and AIRDC are Bayes rule for location shift problem.

Suppose $\mathbf{\Sigma}_1 = \mathbf{I}_2$ and $\mathbf{\Sigma}_2 = \sigma^2\mathbf{I}_2$, misclassification rates associated with AIRDC

Table 4.4: Comparison of classifiers in terms of average misclassification errors when $\mathbf{\Sigma}_1 = \mathbf{I}_2, \mathbf{\Sigma}_2 = \sigma^2\mathbf{I}_2$ and $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$.

			Average misclassification errors				
Distribution	Classifier	δ	$\sigma = 0.2$	$\sigma = 0.5$	$\sigma = 1.0$	$\sigma = 2.0$	$\sigma = 5.0$
Bivariate normal	AIRDC	1	0.2943	0.2892	0.3105	0.4179	0.4785
	RDC		0.2892	0.2921	**0.3101**	0.3578	0.4345
	QDA		**0.0556**	**0.1889**	0.3136	**0.2448**	**0.0813**
	AIRDC	2	0.0695	0.0987	**0.1589**	0.2895	0.4496
	RDC		0.0943	0.1258	**0.1589**	0.2102	0.3167
	QDA		**0.0184**	**0.0751**	0.1614	**0.1891**	**0.0784**
Bivariate Laplace	AIRDC	1	0.3594	0.3522	0.3614	0.4362	0.4789
	RDC		0.3616	0.3521	**0.3569**	0.4377	0.4793
	QDA		**0.1045**	**0.2714**	0.3766	**0.3109**	**0.1436**
	AIRDC	2	0.2019	0.2038	0.2425	0.3523	0.4586
	RDC		0.2028	0.2073	**0.2423**	0.3537	0.4586
	QDA		**0.0564**	**0.1613**	0.2471	**0.2709**	**0.1379**
Bivariate t	AIRDC	1	0.3187	0.3153	0.3376	0.4198	0.4782
	RDC		0.3197	0.3152	**0.3320**	0.4236	0.4777
	QDA		**0.1123**	**0.2685**	0.3716	**0.3295**	**0.1611**
	AIRDC	2	0.1408	0.1577	0.2031	0.3144	0.4509
	RDC		0.1449	0.1560	**0.2010**	0.3143	0.4514
	QDA		**0.0510**	**0.1373**	0.2225	**0.2692**	**0.1547**

increase with increase in σ. Comparing AIRDC with QDA, performance of AIRC is poor when there is scale shift, as shown in Table 4.4. This can be attributed to the lack of scale parameter in the formulation of spatial rank function.

4.3.2 Alternative approach

While defining spatial depth, Ghosh and Chaudhuri (2005b) suggested that spatial rank could be made affine invariant by simply premultiply $\mathbf{x}$ and $\mathbf{X}$ by inverse of covariance matrix of $\mathbf{X}$. That is, suppose $\mathbf{X}$ has a distribution F, which is elliptically symmetric about location parameter $\boldsymbol{\theta}$ and has covariance matrix $\mathbf{\Sigma}$, define

$$\text{rank}_F^{\sim}(\mathbf{x}) = E_F\left(\frac{\mathbf{\Sigma}^{-1/2}(\mathbf{x}-\mathbf{X})}{||\mathbf{\Sigma}^{-1/2}(\mathbf{x}-\mathbf{X})||}\right).$$

$\text{rank}_{\widetilde{F}}(\mathbf{x})$ is invariant under general affine transformation and can be used to build a classification rule that accommodates correlation among variables of the competing populations by replacing $\text{rank}_F(\mathbf{x})$ and $\text{rank}_G(\mathbf{x})$ in Section 4.1 by $\text{rank}_{\widetilde{F}}(\mathbf{x})$ and $\text{rank}_{\widetilde{G}}(\mathbf{x})$ respectively.

Suppose $\mathbf{X} \in \pi_1$ has a distribution function, F on $\mathbb{R}^d$ with prior probability p_1 and $\mathbf{Y} \in \pi_2$ has a distribution G on $\mathbb{R}^d$ with prior probability p_2. Given $\mathbf{x} \in \mathbb{R}^d$, define

$$\begin{aligned}
&r_{\widetilde{F}}(\mathbf{x}) = ||\text{rank}_{\widetilde{F}}(\mathbf{x})||, \;\; r_{\widetilde{F}}(\mathbf{x}) = ||\text{rank}_{\widetilde{G}}(\mathbf{x})||, \\
&r_{\widetilde{F}}(\mathbf{X}) = ||\text{rank}_{\widetilde{F}}(\mathbf{X})|| \;\text{ and }\; r_{\widetilde{G}}(\mathbf{Y}) = ||\text{rank}_{\widetilde{G}}(\mathbf{Y})||.
\end{aligned}$$

Define $F_R\big(r_{\widetilde{F}}(\mathbf{x})\big)$ and $G_R\big(r_{\widetilde{G}}(\mathbf{x})\big)$, the distribution functions of $r_{\widetilde{F}}(\mathbf{X})$ and $r_{\widetilde{G}}(\mathbf{Y})$ as

$$\begin{aligned}
F_R\big(r_{\widetilde{F}}(\mathbf{x})\big) &= P\Bigg(r_{\widetilde{F}}(\mathbf{X}) \leqslant r_{\widetilde{F}}(\mathbf{x})\Bigg) \;\text{ and} \\
G_R\big(r_{\widetilde{G}}(\mathbf{x})\big) &= P\Bigg(r_{\widetilde{G}}(\mathbf{Y}) \leqslant r_{\widetilde{G}}(\mathbf{x})\Bigg)
\end{aligned}$$

respectively. The classification rule based on distribution function of outlyingness of the transformed spatial ranks, $r_{\widetilde{F}}(\mathbf{x})$ and $r_{\widetilde{G}}(\mathbf{x})$ is to assign observation, $\mathbf{x}$ into population π_1 if

$$F_R\big(r_{\widetilde{F}}(\mathbf{x})\big) \leqslant G_R\big(r_{\widetilde{G}}(\mathbf{x})\big)$$

otherwise, assign $\mathbf{x}$ to population π_2. We denote this approach by RDA-A. We suggest use of minimum covariance determinant (MCD) estimate of covariance matrix in order to get robust estimates from the training sample. Here, it should be noted that RDA-A is not fully nonparametric but it is robust against deviation of class distribution from spherical symmetry.

Table 4.5: Performance of classifiers based on real data.

	Comparison of classifiers based on misclassification errors							
Dataset	LDA	QDA	RC	RDC	AIRC	RDA-A	RDA-A0	AIRDC
iris	0.0333	**0.0167**	0.0333	0.0333	**0.0167**	**0.0167**	**0.0167**	**0.0167**
Biomedical	0.2059	0.1471	0.2059	0.1766	0.1471	0.1471	**0.1176**	0.1471
PID	**0.2600**	0.2800	0.3600	0.3500	0.2850	**0.2600**	0.2650	0.2800
Cloud	0.5714	0.4286	0.4643	0.4643	0.3571	**0.2857**	0.3929	0.3571
Banknote	0.0200	0.0200	0.2050	0.1950	0.0200	0.1300	**0.0150**	**0.0150**
Seed	**0.1167**	0.1333	0.2500	0.2500	**0.1167**	0.1833	**0.1167**	**0.1167**
Haberman	0.3150	**0.2583**	0.4300	0.4000	0.3900	0.3483	0.3917	0.3917
Yeast	0.4950	0.3600	0.5500	0.5300	0.3875	0.4475	0.3750	**0.3525**

4.3.3 Numerical Example: Real data

Here, we analyse eight benchmark data set, seven of which are discussed in Section 3.5, to illustrate the performances of our methods (RDC, AIRDC and RDA-A). These datasets include iris data, Pima Indians diabetes (PID) data, banknote data, biomedical data, yeast data, cloud data, seed data and Haberman data. Haberman's survival data (see Haberman, 1976) is an unbalanced data and consists of two classes. We choose training samples of sizes (150 and 50) and validation samples of sizes (50 and 30) respectively. For clarity in our data analysis, we denote RDA-A with MCD estimate of covariance by RDA-A and RDA-A with moment estimate of covariance by RDA-A0. For computing MCD estimate of covariance via R package *robustbase*, we set $\alpha = 0.90$ for small training sample sizes (iris data, seed data and biomedical data) and $\alpha = 0.70$ for large training sample sizes.

Table 4.5 presents the result of analysis of real data. For iris data, RDA-A, RDA-A0, AIRDC and AIRC have the same misclassification error as QDA while RC and RDC has the same misclassification error as LDA. For biomedical data, RDA-A0 has the least misclassification error. RDA-A and AIRDC perform well like QDA. For Pima Indian diabetes data, RDA-A, RDA-A0 and LDA appear to perform best while QDA and AIRDC perform well. For cloud data, RDA-A outperforms others while AIRC, AIRDC and RDA-A0 outperform QDA and LDA. AIRC and RDA-A compete favourably with all other classifiers for banknote authentication data, while misclassification error is least in RDA-A0 but high in RDA-A, RC and RDC. For seed data, both AIRDC, AIRC and LDA outperform others. For Haberman data, QDA has the least misclassification error while RDA-A and LDA perform well. AIRDC, QDA, RDA-A0 and AIRC perform best among other classifiers with yeast data. In general, AIRC, AIRDC, RDA-A and RDA-A0 perform well and compete favourably with QDA while RC and RDC compete favourably with LDA.

Chapter 5

Classification of Functional Data

Median is known to be a popular choice of centre of data cloud, irrespective of its dimension. Median can also be defined as the deepest point in the data cloud with respect to a statistical depth function (Liu, Parelius and Singh, 1999). One of the main motivations for considering the median is its robustness against outlying observations. According to the traditional measures of robustness like breakdown point, median is more robust than mean. For example, spatial median has 50% breakdown point (Kemperman, 1987 and Lopuhaa and Rousseeuw, 1991). Several versions of median in finite dimensional space have been extensively studied in the literature; for example, depth oriented medians (e.g. half-space median, simplicial median, Oja median in Liu, Parelius and Singh, 1999; Zuo and Serfling, 2000a, among others), spatial median (Chaudhuri, 1996; Vardi and Zhang, 2000), among others. These different versions of multivariate median has been extended into infinite dimension. Many of these medians for finite dimensional probability measures do not extend in any natural and meaningful way into infinite dimensional spaces (See Chakraborty and Chaudhuri, 2014 for detail). Also, many of these medians for functional data are not computationally simple. These give attraction to the use of spatial median. On the other hand, spatial median has been extended into Banach spaces, see Kemperman (1987) for detail.

In this chapter, we propose classification method for functional data based on distance to some functional medians and study the properties of these classifiers. This method is completely data driven and easy to compute. It is imperative to review some existing centroid based classifiers for functional data. In Section 5.1, we review some existing literature on classification methods for functional data. In Section 5.2, we propose classification procedure based on distance to spatial median for functional data and shall establish some theoretical properties of this method. Section 5.3 contains generalisation of classifier based on distance to spatial median into L_p distance based procedure for various values of p. Numerical results based on simulation and real data, optimal choice of p and other relevant discussions are contained in succeeding subsections.

5.1 Some Classification Methods For Functional Data

5.1.1 Different forms of discriminant analysis for functional Data

In a two-class problem with same covariance matrix, LDA in Section 1.2 is equivalent to

$$\delta(\mathbf{x}) = I\{(\mathbf{x}-\boldsymbol{\mu}_a)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_b \geqslant 0\}, \tag{5.1.1}$$

where I is an indicator function, $\boldsymbol{\mu}_a = \frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)$ and $\boldsymbol{\mu}_b = \frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$. The classification rule is to assign an observation $\mathbf{x}$ to population π_1 if $\delta_F(\mathbf{x}) = 1$ and to population π_2 if $\delta_F(\mathbf{x}) = 0$ with the associated misclassification error

$$\Delta = 1 - \Phi\big((\boldsymbol{\mu}_b^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_b)^{\frac{1}{2}}\big)$$

if normality is assumed, where Φ is the cumulative distribution function of normal distribution. Minimising misclassification error is equivalent to maximising $\boldsymbol{\mu}_b^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_b$. However, LDA becomes a serious challenge to use when data are curves or functions. The reason is that the sample covariance matrix is singular and cannot be inverted. Hence it hinders the applicability of LDA. There are two common solutions to this problem. The first one is called regularisation method. This include regularising or penalising covariance matrix $\mathbf{\Sigma}$ (see Di Pillo, 1976; Friedman, 1989; Hastie, Buja and Tibshirani, 1995; Guo, Hastie and Tibshirani, 2007). The second solution is filtering method. This involves choosing a finite dimensional basis and finding the best projection of each curve onto this basis. We refer reader to James and Hastie (2001) for detail.

Hastie, Buja and Tibshirani (1995) proposed penalized discriminant analysis (PDA), which involves replacing $\mathbf{\Sigma}$ in equation (5.1.1) by $\mathbf{\Sigma}_W = \mathbf{\Sigma} + \lambda\mathbf{\Omega}$, where $\mathbf{\Omega}$ is a symmetric, nonnegative definite, roughness-type penalty matrix and λ is a smoothing parameter, which was later assumed to be absorbed in $\mathbf{\Omega}$. The LDA then follows as usual. The performance of PDA degrades when many irrelevant variables exist in the data. Witten and Tibshirani (2011) suggested recasting of Fisher's discriminant problem as a biconvex problem by applying convex penalties given that $\mathbf{\Omega}$ is a diagonal matrix.

Dudoit, Fridlyand and Speed (2002) proposed independence rule. Independence rule assumes no correlation among features. It is Bayes rule under normality given that input features are not correlated (Bickel and Levina, 2004). It involves replacing $\mathbf{\Sigma}$ in equation (5.1.1) by $\mathbf{D}$, the diagonal of pooled covariance matrix of the competing classes. Then the usual LDA is carried out on the test data. Fan, Fan and Wu (2011) gave an expression for the probability of misclassification associated with independence rule as

$$\Delta_I = \Phi\left(\frac{1}{2}\frac{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\mathbf{D}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{[(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\mathbf{D}^{-1}\mathbf{\Sigma}\mathbf{D}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]^{1/2}}\right).$$

Bickel and Levina (2004) argued that independence rule is not much lower in performance compare to Bayes rule in terms of proportion of correct classification. Fan, Feng and Tong (2012) argued that it may perform very poor when using all the features in the curves because of accumulation of noise in estimating population centroids in high dimensional feature space and shown that optimal risk using independence rule increases as correlation among features increases.

Fan, Feng and Tong (2012) proposed regularised optimal affine discriminant (ROAD) and its variants. In their proposal, $\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_b$ in equation (5.1.1) is replaced with a vector $\mathbf{w} \in \mathbb{R}^d$. The optimal choice of $\mathbf{w}$, denoted by $\mathbf{w}_c$, is

$$\mathbf{w}_c = \min_{\|\mathbf{w}\|_1 \leqslant c, \mathbf{w}^T \mu_b = 1} \mathbf{w}^T \mathbf{\Sigma} \mathbf{w}$$

where c is a small positive number such that

$$c = \frac{1}{\max_{1 \leqslant i \leqslant d} |\mu_{b,i}|},$$

$\mu_{b,i}$ is the ith component of $\boldsymbol{\mu}_b$. The classification rule is to assign $\mathbf{x}$ to π_1 if $\delta_{\mathbf{w}_c}(\mathbf{x}) = I\{\mathbf{w}_c^T(\mathbf{x} - \boldsymbol{\mu}_a) \geqslant 0\} = 1$ and to π_2 if otherwise. ROAD is robust and performs well when all variables are independent. It has two variants. The first variant of ROAD is diagonal regularised optimal affine discriminant (DROAD). This involves setting $\mathbf{w}_c$ as $\mathbf{w}_c = \min_{\|\mathbf{w}\|_1 \leqslant c, \mathbf{w}^T \mu_b = 1} \mathbf{w}^T \mathrm{diag}(\mathbf{\Sigma})\mathbf{w}$. The second variant is to perform pre-screening of all features before carrying out ROAD. This is called S-ROAD.

James and Hastie (2001) proposed functional linear discriminant analysis (FLDA) for irregularly sampled curves. The classification rule is to classify Y to class i if

$$\|\widehat{\boldsymbol{\alpha}}_{\mathbf{Y}} - \boldsymbol{\alpha}_i\| - \log_e p_i$$

is minimum, where

$$\mathbf{Y}_{ij} = \mathbf{S}_{ij}(\boldsymbol{\lambda}_0 + \mathbf{\Lambda}\boldsymbol{\alpha}_i + \boldsymbol{\gamma}_{ij}) + \boldsymbol{\varepsilon}_{ij}; \;\; \boldsymbol{\varepsilon}_{ij} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}), \;\; \boldsymbol{\gamma}_{ij} \sim N(\mathbf{0}, \mathit{\Gamma}),$$

$$\widehat{\boldsymbol{\alpha}}_{\mathbf{Y}} = (\mathbf{S}_{\mathbf{Y}}\boldsymbol{\Lambda})^{-1}(\mathbf{Y} - \mathbf{S}_{\mathbf{Y}}\boldsymbol{\lambda}_0), \quad E(\mathbf{Y}) = \boldsymbol{\mu}_{\mathbf{Y}} = \mathbf{S}_{\mathbf{Y}}^T(\boldsymbol{\lambda}_0 + \boldsymbol{\Lambda}\boldsymbol{\alpha}_i)$$

p_i is the prior probability of the class i, and $\boldsymbol{\Sigma}_{\mathbf{Y}} = \sigma^2\boldsymbol{I} + \mathbf{S}_{\mathbf{Y}}\mathit{\Gamma}\mathbf{S}_{\mathbf{Y}}^T$, $\mathbf{S}_{\mathbf{Y}}$ is the spline basis matrix for $\mathbf{Y}$ evaluated over a fine lattice of $1 \leqslant j \leqslant n$ points. This technique performs well only when fragments of the curves are observed.

Further discussions on functional discriminant analysis include study by Preda, Saporta and Leveder (2007) and Shin (2008).

5.1.2 Maximum functional depth classifier

Data depth provides criterion for ordering sample of curves from centre-outward. It helps to build nonparametric tools for functional data analysis and thereby motivates nonparametric robust statistical methodologies. Data depths for functional data include integrated depth or Fraiman-Munic depth (Fraiman and Munic, 2001), h-mode depth (Fraiman and Meloche, 1999), band depth and modified band depth (Lóopez-Pintado and Romo, 2006), random Tukey depth (Cuesta-Albertos and Nieto-Reyes, 2008), among others. Cuevas, Febrero and Fraiman (2007) proposed maximum depth classifier for functional data. It is an extension of maximum depth classifier in multivariate setting into functional data. The classification rule based on maximum functional depth classifier is to assign an observation to group with highest depth value. That is, for $J(\geqslant 2)$ groups, assign x to kth group if

$$D_k(x, X_k) = \max_{1 \leqslant j \leqslant J} D_j(x, X_j),$$

where $D_k(x, X_k)$ is the depth value of x with respect to kth group, X_k. The functional depths used in their proposal are $h-$modal depth, Fraiman-Muniz depth, random projection depth and double random projection depth. More recently, Claeskens et al. (2014)

defined multivariate functional depth function as

$$MFD(\mathbf{x}, F) = \int_{\mathcal{I}} D(\mathbf{x}, F) w(t) dt,$$

where w is the weight function defined on $\mathcal{I}$ and integrates to 1, D is the statistical depth function on $\mathbb{R}^d$, F is the distribution of continuous stochastic process $\mathbf{X}$ on $\mathbb{R}^d$ that generates continuous paths in $\mathcal{C}(\mathcal{I})^d$ and $\mathbf{x} \in \mathcal{C}(\mathcal{I})^d$. Examples of multivariate functional depth are double random projection depth and derivatives proposed in Cuevas, Febrero and Fraiman (2007) and multivariate functional halfspace depth in Claeskens et al. (2014). Both random projection depth and derivatives and multivariate functional halfspace depth suffer computational difficulty as the set of observed time points increase.

5.1.3 Bayesian approach

Naive Bayes rule

In high dimension, the curse of dimensionality and accumulation of noise limit the use of Bayes rule (see Fan, Fan and Wu, 2011 for detail). Thanks to Naive Bayes classifier, which helps to overcome this by making conditional independence assumption. Naive Bayes rule, proposed in Bickel and Levina (2004), is Bayes rule under the assumption of independence of features. Dudoit, Fridlyand and Speed (2002) argued that if correlation between features in functional data, especially for genes in microarray data, is ignored, independence rule may perform better. Bickel and Levina (2004) studied theoretical properties of naive Bayes rule and Fisher's LDA. Ackermann and Strimmer (2009) and Fan, Feng and Tong (2012) have shown that correlation among features is an essential characteristic and is not always negligible, especially in micro-array data and clinical outcomes. So, use of naive Bayes rule for such data may lead to suboptimal procedure. It may also lead to loss of critical information.

Nonparametric method for curve discrimination (NPCD)

Devroye, Gyӧrfi and Lugosi (1996) proposed kernel rules. Hall, Poskitt and Presnell (2001) and Ferraty and Vieu (2003) upgraded kernel rule into full fledged NPCD. NPCD involves estimating posterior probability of each of the competing classes given an observation x using consistent kernel estimator, then assign a new observation to the class with highest estimated posterior probability. Suppose $(X_1, Y_1), \ldots, (X_n, Y_n)$ is a collection of independently and identically distributed curve X_i and the class membership Y_i. The kernel estimator of posterior probability of jth class given x, as given in Ferraty and Vieu (2003), is

$$\widehat{P}_{j,h}(x) = \frac{\sum_{i=1}^{n_j} K(h^{-1}d(X_i, x))I_{[Y_i=j]}}{\sum_{i=1}^{n} K(h^{-1}d(X_i, x))},$$

for $j = 1, 2, \ldots, J$, where I is indicator function with value 1 if $Y_i = j$ and 0 if otherwise, K is the kernel, h is the bandwidth, d is the semi-metric and $n = \sum_{j=1}^{J} n_j$. The classification rule based on NPCD is to assign x to the class with the highest $\widehat{P}_{j,h}(x)$. On the choice of semi-metric d for NPCD, Hall, Poskitt and Presnell (2001) used functional principal component analysis based on Karhunen-Loѐve expansion for dimension reduction while Ferraty and Vieu (2003) suggested functional principal component analysis (FPCA) and successive derivatives, and then estimated the L_2 norm based on the resulting multivariate data. Ferraty and Vieu (2003) have shown in their study through simulation and analysis of real dataset that NPCD competes favourably with penalized discriminant analysis and partial least square regression method. This comes from the possibility of various semi-metric choices.

Other methods include classifiers based on functional mixed model (Zhu, Brown and Morris, 2012), which involves fitting functional mixed model to the training data with class as one of the fixed effect predictors and then perform classification of the test data using posterior predictive probabilities of class membership.

5.1.4 Classification based on Support Vector Machine

Support vector machine for functional data

Support vector machine (SVM) is a popular method for classifying both multivariate and functional data. This is first proposed in Vapnik (1982) and upgraded in Cortes and Vapnik (1995). Use of support vector machine for classifying functional data is an extension of its multivariate set-up. Suppose (x_i, y_i) is a pair of random variable in which y_i, class membership takes values in $\{-1, 1\}$ and $x_i \in \mathcal{X}$, where $\mathcal{X}$ is a set of training data points in functional space, $i = 1, 2, \ldots, n$. SVM aims at predicting the value of y_i given observed value for x_i. SVM separates two different classes of data by a hyperplane $\{x :< w, x > + b = 0\}$. The corresponding classification rule is

$$y_i(x) = \text{sign}(< w, x_i > + b)$$

where w is to be estimated and b is a constant scalar. In order to obtain the best separating hyperplane, $\|w\|$ is minimised subject to the decision rule. That is,

$$\begin{aligned} &\min_{w,b,\|w\|=1} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i \\ \text{subject to } &y_i(< w, x_i > + b) \geqslant 1 - \xi_i \;\; i = 1, 2, \ldots, n \\ &\xi_i \geqslant 0 \end{aligned}$$

Rossi and Villa (2006) proposed use of kernels with support vector machines in order to provide consistent classification in both finite dimensional spaces and infinite dimensional spaces. This involves replacing $< w, x_i >$ by kernel function $K(w, x_i)$. From the geometric perspective, SVM is a large margin classifier. Specifically, for separable data, SVM separates two classes by maximising the margin between them. For non-separable data, the soft-margin SVM chooses a separating hyperplane that splits two classes as cleanly as possible, while still maximising the distance to the support vectors, a subset of the training samples on the separating hyperplane. A desirable property of SVM is that its solution depends only on support vectors. However, since all the input variables are used for constructing the classifier, SVM cannot select important variables and its performance will degrade when many irrelevant variables exist (see Li and Yu, 2008; Hastie, Tibshirani and Friedman, 2001). According to Li and Yu (2008), decision rule of SVM suffers from presence of redundant variables.

Functional segment discriminant analysis (FSDA)

Li and Yu (2008) proposed FSDA. This method combines classical LDA as a data reduction tool with support vector machine as classifier. In their proposal, $F-$statistic is used to select the first m features with largest $F-$statistic values and then apply LDA on the selected curve segments. The resulting sequence of linear discriminant variables are then used as the extracted features on which support vector machine is performed.

5.1.5 Nearest neighbour rule

The k-nearest neighbour rule (k-NN) is a nonparametric method for classifying finite and infinite dimensional test observations based on closest training observations in the data cloud. It is proposed in Cover and Hart (1967). This involves assigning an unclassified sample point to the class that is commonest amongst its k nearest neighbours, where k is a positive integer. Suppose $x_{(i)}, i = 1, 2, \ldots, k$ are k nearest neighbours to x, the distance between x and $x_{(i)}$ is

$$d_{(i)} = \|x_{(i)} - x\|$$

where $\|.\|$ is Euclidean distance. This classification rule is to assign x to the class that is commonest amongst its k nearest neighbours. If $k = 1$, then an observation is assigned to the class of its nearest neighbour. This procedure is simply majority vote of neighbours.

This rule is independent of the underlying joint distribution of the sample points. Kim et al. (2011) suggested use of cosine method or correlation methods as an alternative for calculating the distance between x and its k-nearest neighbour. Cover and Hart (1967) showed that the single nearest neighbour rule (1-NN) is admissible and for any number of categories, the total probability of misclassification using nearest neighbour rule is bounded above by twice the total probability of misclassification using Bayes rule. The best choice of k depends on the data. Generally, larger values of k reduce the effect of noise on the classification but make boundaries between classes less distinct. One major drawback of k-NN is that training sample points from more frequent class tend to dominate the prediction of test sample points when the class distribution is skewed. According to Coomans and Massart (1982), this is because they tend to be common among the k nearest neighbours due to their large number. Similarly, the performance of k-NN algorithm can be severely degraded by the presence of noisy or irrelevant features, or if the feature scales are not consistent with their importance.

5.1.6 Classifier based on distance to centroids

Nearest centroid classifier

Hastie, Tibshirani and Friedman (2001) proposed nearest centroid classifier for classifying gene expression, a very high dimensional data. This method computes a standardized centroid for each class. Standaradized centroid is the class mean divided by the within class standard deviation for each class component. Nearest centroid classifier assigns an individual curve x to the class $\mathcal{X}$ of curves with shortest distance from its standardized centroid (say μ) to the observation. That is, $x \in \mathcal{X}$ if

$$\|\mu - x\| \text{ is minimum.}$$

This method performs well if competing classes only differ in location and features are uncorrelated. Centroid based classifiers are known for some intuitive features, such as computational simplicity, convergence of sample mean to population mean, among others. However, sample means are sensitive to the presence of outliers in the data cloud, since outliers are difficult to detect in high dimension and they can affect the analysis in many different ways (L$\acute{o}$pez-Pintado and Romo, 2006). Tibshirani et al. (2002) suggested shrinking the class centroids towards the overall centroid after standardizing by within-class standard deviation for each component, then assign x to the class of curves with shortest distance from x to its shrunken centroid. This will effectively eliminate many non-contributing genes and leave us with a small subset of genes for scientific interpretation and further analysis. Note that the class centroids of each gene are shrunken individually. This is based on the assumption that genes are independent of each other, which however, for most of the time is not totally valid (Guo, Hastie and Tibshirani, 2007). Chan and Hall (2009) presented a scale-adjusted version of centroid classifier for very high dimensional data when the principal difference between competing classes are in location. Hall and Pham (2010) argued that scale adjustment removes the tendency of scale to confound difference in means and discussed its optimal properties. Alonso, Casado and Romo (2012) proposed a weighted distance approach, which assigns weight to the distance between new curve, functional data and their derivatives.

Near perfect classification method

Delaigle and Hall (2012a) proposed near perfect classification method for functional data. It involves constructing truncated version of nearest centroid classifier for competing classes with equal covariance. The classification rule is to classify an observation x to class 1 if

$$D^2(x, \overline{X}_1) < D^2(x, \overline{X}_0)$$

otherwise to class 0, where $D(x, y) = | < x, \psi > - < y, \psi > |$, $\overline{X}_k$ is the mean of the kth class of curves, $k = 0, 1$, $< a, b >$ is inner product of a and b. The authors proposed two choices of ψ, which are

1. $$\psi^{(r)} = \sum_{j=1}^{r} \theta_j^{-1} \mu_j \phi_j$$

   where $\mu_j = \int \mu\phi_j$, the projection of μ on the respective eigenfunction. $\mu = \frac{1}{n_1} \sum_{i=1}^{n_1} X_{1i}(t) - \frac{1}{n_0} \sum_{i=1}^{n_0} X_{0i}(t)$, θ_j and ϕ_j are jth eigenvalue and its corresponding eigenfunction of covariance kernel respectively, n_0 and n_1 are sizes of the competing classes, and r is chosen by leave-one-out cross validation estimator of error rate.

2. The second is based on regression using asymptotic partial least square approach of Preda, Saporta and Leveder (2007). The authors assume that covariance of each competing class is both positive definite and uniformly bounded. In order to achieve optimal classification, it is assumed that $\sum_{j \leqslant 1} \theta_j^{-2} \mu_j^2 = \infty$. Similarly, to achieve perfect classification, it is assumed that $\sum_{j \leqslant 1} \theta_j^{-1} \mu_j^2 = \infty$.

Classifiers based on distance to trimmed mean and its variants

L*ó*pez-Pintado and Romo (2006) extended the concept of α-trimmed mean in $\mathbb{R}^d$ to functional set-up, which is the mean of $100(1-\alpha)\%$ deepest observations in the training sample and proposed classification method based on distance to the trimmed mean, weighted average distance and trimmed weighted average distance.

1. α-trimmed mean is defined as the average of $n - \lfloor n\alpha \rfloor$ deepest curves from the sample where $\lfloor n\alpha \rfloor$ is the integer part of $n\alpha$. Let $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ be the centre-outward ordered sample, where $x_{(1)}$ is the deepest observation and $x_{(n)}$ is the least

deepest one, then α-trimmed mean is

$$m^{\alpha} = \frac{\sum_{i=1}^{n-\lfloor n\alpha \rfloor} x_{(i)}}{n - \lfloor n\alpha \rfloor}.$$

The associated classification rule is to assign new curves to the group with the shortest distance between group trimmed mean and the curve. That is, for $J(\geqslant 2)$, assign x to kth group if

$$DTM(x, m_k^{\alpha}) = \min_{1 \leqslant j \leqslant J} \|x - m_j^{\alpha}\|,$$

where $\|.\|$ is the Euclidean norm.

2. Weighted average distance is the weighted average of distances to each element in the group. The weight of each observation is determined by its depth value within the group. Suppose higher weight is allocated to deeper observation and for $1 \leqslant j \leqslant J$, $A_j = \{x_1, x_2, \ldots, x_{n_j}\}$. The weighted average distance of x to A_j is

$$WAD(x, A_j) = \frac{\sum_{i=1}^{n_j} d(x, x_i) S(x_i)}{\sum_{i=1}^{n_j} S(x_i)},$$

where $S(x_i)$ is the weight of x_i in A_j, $d(x, x_i)$ is metric between x and x_i, and n_j is the size of the group A_j. The associated classification rule is to assign a new curve to the group with the shortest weighted average distance.

3. Trimmed weighted average distance involves computing weighted average distance based on fixed m deepest observations for each group, where m is less than or equal to the minimum group size $(m \leqslant n_1 \leqslant n_2 \leqslant \ldots \leqslant n_J)$. Mathematically, trimmed

weighted average distance is

$$WAD(x, A_j) = \frac{\sum_{i=1}^{m} d(x, x_{(i)})S(x_{(i)})}{\sum_{i=1}^{m} S(x_{(i)})}.$$

The associated classification rule is to assign a new curve to the group with the minimum trimmed weighted average distance.

To determine $100(1-\alpha)\%$ deepest observations and weights for group observations, band depth determined by three different curves, band depth determined by four different curves and the generalized band depth are used. Cuesta-Albertos and Nieto-Reyes (2010) suggested use of random Tukey depth while Sguera, Galeano and Lillo (2014) suggested use of spatial depth and kernelized spatial depth in place of depths used in L*ó*pez-Pintado and Romo (2006) for trimming of sample means.

In theory, if the distribution of the sample is sufficiently heavy-tailed, then the expected value of the sample mean will not be well defined. This may limit the use of mean-based classifiers (Hall, Titterington and Xue, 2009).

5.1.7 Median based classifier

Recently, Hall, Titterington and Xue (2009) proposed median based classifiers for high dimensional data, which depends on distance between componentwise L_1 median and an observation. It is analogous to mean based classifiers. In a two class problem, the classification rule is to assign a p-variate observation z to class $\mathcal{X}$ if

$$\sum_{k=1}^{p} \Big(|\text{med}\mathcal{Y}_k - z_k| - |\text{med}\mathcal{X}_k - z_k| \Big) > 0, \tag{5.1.2}$$

otherwise to $\mathcal{Y}$, where med$\mathcal{X}_k$ and med$\mathcal{Y}_k$ denote kth feature of the L_1 median of $\mathcal{X}$ and $\mathcal{Y}$ respectively. Also, the authors proposed a truncation-based classifier by defining a uniformly bounded function ψ for which $\psi(u) = -\psi(-u)$ for all u, and $\psi(u) > 0$ for

$u > 0$. The associated rule is to assign z to $\mathcal{X}$ if

$$\sum_{k=1}^{p} \psi\Bigg(|\mathrm{med}\mathcal{Y}_k - z_k| - |\mathrm{med}\mathcal{X}_k - z_k|\Bigg) > 0,$$

and to $\mathcal{Y}$ if otherwise. The truncation-based classifier is relatively insensitive to gradations in the sizes of the differences between medians. Hennig and Viroli (2013) proposed classification method based on distance to the within class θth componentwise L_1 quantiles. This is the modification of componentwise L_1 median based classifier for high dimensional data in Hall, Titterington and Xue (2009). In their proposal, L_1 median is replaced by θth componentwise L_1 quantile, whose choice depends on the value of θ that minimises misclassification error. When $\theta = 0.5$, the resulting classifier is simply median based classifier. The advantage of L_1 distance based classifiers is in their performance when data are skewed. That is, L_1 distance based classifiers are asymptotically optimal when data components are independent and double-exponential, and sample sizes of competing classes diverge as the dimension increases (Hennig and Viroli, 2013 and Hall, Titterington and Xue, 2009).

5.1.8 Feature selection

In literature, dimension reduction is performed by projecting functional data onto a finite number of functions $\psi_1, \psi_2, \ldots, \psi_d$. Then, standard multivariate classifiers are applied to d dimensional projection $(\int_I X\psi_1, \int_I X\psi_2, \ldots, \int_I X\psi_d)$, where ψ_i is chosen from the data (e.g. principal component basis) or chosen arbitrarily (e.g. spline basis). Tian, James and Wilcox (2010) proposed multivariate adaptive stochastic search method for dimension reduction, which involves projecting a high dimensional data into a lower dimensional space and then apply a conventional classification method on the resulting data.

Delaigle and Hall (2012b) proposed componentwise feature selection for classification and clustering of functional data. This method adaptively selects set of d points that contribute most to classification and apply conventional finite dimensional classifiers (LDA, QDA, a nonparametric Bayes rule, a nonparametric regression-based classifier and a classifier based on logistic regression) on resulting d-dimensional vectors $\{X(t_1), X(t_2), \ldots, X(t_d)\}$. To choose d points, let I_r denote the set of all r-vectors $t_{(r)} = (t_1, t_2, \ldots, t_r)^T$ with $t_1 < t_2 < \ldots < t_r$ and $t_1, t_2, \ldots, t_r \in I_r$. Define a cross validation estimator of error rate as

$$\widehat{err}_r(t_{(r)}) = \frac{1}{n}\sum_{i=1}^{n} I\{J(X_i, D_{-i}|t_{(r)}) \neq I_i\},$$

where $D_{-i} = D \setminus \{(X_i, I_i)\}$ denotes the dataset with ith data pair removed, I_i is class label of each X_i, $J(X_i, D_{-i})$ denotes population index, either 0 or 1, to which each x is assigned after the dimension has been reduced to $t_{(r)} = (t_1, t_2, \ldots, t_r)^T$. The most important r dimensional points $t_{(r)}$ is set as one that minimises $\widehat{err}_r(t_{(r)})$. Define

$$T_r = \inf_{t_{(r)} \in I_r} \widehat{err}_r(t_{(r)}),$$

Delaigle and Hall (2012b) suggested an estimate of d as $\widehat{d} = \inf\{r : (1-\rho)T_r \leqslant T_{r+1}\}$, where ρ is chosen to be 0.1.

Componentwise two sample t-test are often used for selecting important feature in classification problem (Tibshirani et al., 2002 and Fan and Fan, 2008). Fan and Fan (2008) proposed feature annealed independence rule (FAIR), which selects the statistically most significant m features based on componentwise two sample t-test and apply independence rule on the selected features. Biau, Bunea and Wegkamp (2005) selected finite features from infinite dimension by considering only the first d coefficients of Fourier series expansion of each element and then perform k-NN on the reduced data in $\mathbb{R}^d$. The choice of d and k are determined using simple data splitting device. In functional segment discriminant analysis, F-statistic is used to select first m features with largest F-statistic values on which LDA is applied for data reduction.

5.2 Classifiers Based on Distance to Spatial Median

5.2.1 Spatial median

Suppose $X \in \mathcal{C}(\mathcal{I})$ is a random function observed at finite points $t \in \mathcal{I}$, where $\mathcal{C}(\mathcal{I})$ is a space of continuous functions defined on $\mathcal{I}$ and $\mathcal{I}$ is a closed interval of $\mathbb{R}$. The spatial median of X, denoted by M, is defined as

$$M = \arg_f \min E \bigg\{ \int_{\mathcal{I}} |X(t) - f(t)|^2 dt \bigg\}^{1/2}.$$

Kemperman (1987) extended the notion of spatial median into Banach space and has shown that the spatial median is unique for a strictly convex Hilbert space if the distribution of X is nonatomic and not entirely supported on a line. Therefore it is the only point in the Hilbert space which satisfies $E_F\{(x - X)/\|x - X\|\} = 0$ (See Theorem 2.17 of Kemperman, 1987 and Fact 2.1 of Chakraborty and Chaudhuri, 2013), where $\|x - X\|$ is the Euclidean distance of x from X, defined as $\|a\| = \bigg\{ \int_{\mathcal{I}} |a(t)|^2 dt \bigg\}^{1/2}$.

5.2.2 Minimal distance to spatial median classifier

Suppose we observe independent and identically distributed random functions defined on a compact interval $\mathcal{I}$. Let $X_{ij}(t), i = 1, 2, \ldots, n_j, j = 1, 2, \ldots, J$ be the ith observed functional observation from jth class, with prior probability p_j, where j is the class label and $t \in \mathcal{I}$. We assume the functions are drawn from populations that differ only in mean functions and their covariance kernels are both positive definite and uniformly bounded. Define a L_2 metric $\mathcal{D}$ as

$$\mathcal{D}(z, a) = \bigg\{ \int_{\mathcal{I}} |z(t) - a(t)|^2 dt \bigg\}^{1/2},$$

for $t \in \mathcal{I}$. It is obvious that $\mathcal{D}(z, a) > 0$ if $z \neq a$ and $\mathcal{D}(z, a) = 0$ iff $z = a$.

Here, we propose a classification method based on L_2 distance of individual observation to the spatial median of each of the competing classes. The classification procedure is to assign an observation, z into the class with the least L_2 distance between z and spatial median of each of the competing classes. In a two-class classification problem, let $\mathcal{X}$ and $\mathcal{Y}$ be two classes of observations taken values from Hilbert space. Suppose $X_1, X_2, \ldots, X_{n_1} \in \mathcal{X}$ with size n_1 and prior probability p_1, and $Y_1, Y_2, \ldots, Y_{n_2} \in \mathcal{Y}$ with size n_2 and prior probability p_2. Suppose $p_1 = p_2$, M_X and M_Y are spatial medians of the data cloud, $\mathcal{X}$ and $\mathcal{Y}$ respectively. The classification rule is to assign z into class $\mathcal{X}$ if

$$\mathcal{D}(z, M_X) \leqslant \mathcal{D}(z, M_Y), \tag{5.2.1}$$

otherwise to class $\mathcal{Y}$, where

$$\mathcal{D}(z, M_X) = \left\{ \int_{\mathcal{I}} |z(t) - M_X(t)|^2 dt \right\}^{1/2} \text{ and } \mathcal{D}(z, M_Y) = \left\{ \int_{\mathcal{I}} |z(t) - M_Y(t)|^2 dt \right\}^{1/2}.$$

Assuming $P\big(\mathcal{D}(z, M_X) = \mathcal{D}(z, M_Y) \mid z\big) = 0$, this classifier is unique except for the set of points with probability zero and the separating hyperplane between $\mathcal{X}$ and $\mathcal{Y}$ is the line that passes through $\mathcal{D}^2(z, M_X) = \mathcal{D}^2(z, M_Y)$. That is, suppose $\mathcal{X}$ and $\mathcal{Y}$ are linearly separable, it is easy to show that the separating hyperplane between $\mathcal{X}$ and $\mathcal{Y}$ is

$$\int_{\mathcal{I}} z(t)[M_X(t) - M_Y(t)]dt = \frac{1}{2} \int_{\mathcal{I}} [M_X^2(t) - M_Y^2(t)]dt.$$

The classification rule in equation (5.2.1) is equivalent to assigning z to $\mathcal{X}$ if

$$\int_{\mathcal{I}} z(t)[M_X(t) - M_Y(t)]dt \geqslant \frac{1}{2} \int_{\mathcal{I}} [M_X^2(t) - M_Y^2(t)]dt$$

and z to $\mathcal{Y}$ if

$$\int_{\mathcal{I}} z(t)[M_X(t) - M_Y(t)]dt < \frac{1}{2}\int_{\mathcal{I}} [M_X^2(t) - M_Y^2(t)]dt.$$

This can be viewed as setting a threshold for classification. That is, assign z to $\mathcal{X}$ if $< z, M_X - M_Y > \geqslant \frac{1}{2} < M_X + M_Y, M_X - M_Y >$, where $< a, b >$ denotes inner product of a and b. We shall call this classification method minimal L_2 distance to spatial median classifier, denoted by DL2M. The probability of misclassification associated with DL2M in a two class problem with equal prior probabilities is

$$\Delta = \frac{1}{2}P(\mathcal{D}(z, M_X) > \mathcal{D}(z, M_Y) \mid z \in \mathcal{X}) + \frac{1}{2}P(\mathcal{D}(z, M_X) \leqslant \mathcal{D}(z, M_Y) \mid z \in \mathcal{Y}).$$

It was mentioned in Hall, Titterington and Xue (2009) that the theoretical median of the sample is not necessarily equal to the median of the population from which the data were drawn, whereas the expected value of a sample mean always equals the population mean. This means that theoretical properties of median-based classifiers can be quite different from those of their mean-based counterparts. The almost sure convergence of empirical spatial median to its population version for observations that take values in a strictly convex separable Hilbert space, given that the probability distribution of the observations is nonatomic and not entirely supported on a line in $\mathcal{X}$ was proved in Chakraborty and Chaudhuri (2013).

5.2.3 Theoretical Properties

Suppose $\mathcal{X}$ and $\mathcal{Y}$ are two competing populations of functional data from distributions F and G respectively with equal prior probabilities and differ only in mean function. Assuming the first moments, $\mu_X = E(X)$ and $\mu_Y = E(Y)$ of the distribution of $\mathcal{X}$ and $\mathcal{Y}$ respectively, exist and uniformly bounded. Suppose M_X and M_Y are spatial medians of $\mathcal{X}$ and $\mathcal{Y}$ respectively. Define $m = \int_{\mathcal{I}} \left(M_X^2(t) - M_Y^2(t)\right)dt$ and $\gamma = \int_{\mathcal{I}} z(t)\big(M_X(t) - M_Y(t)\big)dt$, then

$$E_X(\gamma) = E_X\bigg(\int_{\mathcal{I}} z(t)\big(M_X(t) - M_Y(t)\big)dt\bigg) = \int_{\mathcal{I}} \big(M_X(t) - M_Y(t)\big)E_X\big(z(t)\big)dt \text{ and}$$
$$E_Y(\gamma) = \int_{\mathcal{I}} \big(M_X(t) - M_Y(t)\big)E_Y\big(z(t)\big)dt.$$

Suppose

$$\mathrm{var}(\gamma) = \sigma_\gamma^2.$$

By Mercer's theorem (Mercer, 1909; Kac and Siegert, 1947),

$$\mathrm{cov}(X(t), X(s)) = \sum_j \theta_j\phi_j(t)\phi_j(s),$$

where θ_j and ϕ_j are jth eigenvalue of $\mathrm{cov}(X(t), X(s))$ and its corresponding eigenfunction respectively. $\mathrm{cov}(X(t), X(s))$ is positive definite if $\theta_j > 0$ for all j and uniformly bounded if $\sum_j \theta_j < \infty$. We want to show that γ is Gaussian if X is Gaussian but with restriction to finite dimensional settings. This is given in the lemma below.

Lemma 5.2.1 *Suppose X is finitely observed functional data, γ is Gaussian if X is Gaussian.*

Proof: Suppose X is Gaussian distributed with mean μ and covariance kernel K. Define $Z = X - \mu$, where Z is a zero-mean Gaussian with covariance kernel K. Karhunen-Loève expansion of Z gives $Z(t) = \sum_{j=1}^{\infty} \theta_j^{1/2} Z_j\phi_j(t)$ (Deheuvels and Martynov, 2008), where $Z_j, j = 1, 2, \ldots$ are independent normally distributed random variables, θ_j and $\phi_j(t)$ are jth eigenvalue and its corresponding eigenfunction of $K(t, s)$ and are continuous in $\mathcal{I}$, the convergence is in L^2 sense and uniform in t. Then

$$\gamma = \int_{\mathcal{I}} X(t)[M_X(t) - M_Y(t)]dt = \int_{\mathcal{I}} X(t)u(t)dt,$$

which is the same as

$$\gamma - E(\gamma) = \int_{\mathcal{I}} Z(t)[M_X(t) - M_Y(t)]dt = \int_{\mathcal{I}} Z(t)u(t)dt,$$

where $u(t) = M_X(t) - M_Y(t)$. Suppose X is observed at d finite points in $\mathcal{I}$ and d is allowed to diverge, then

$$\begin{aligned}\gamma - E(\gamma) &= \int_{\mathcal{I}} \big[\sum_{j=1}^{d} \theta_j^{1/2} Z_j \phi_j(t)\big] u(t)dt = \int_{\mathcal{I}} \sum_{j=1}^{d} Z_j \big[\theta_j^{1/2} \phi_j(t)u(t)\big]dt \\ &= \sum_{j=1}^{d} Z_j \int_{\mathcal{I}} \big[\theta_j^{1/2} \phi_j(t)u(t)\big]dt = \sum_{j=1}^{d} Z_j \xi_j\end{aligned}$$

where $\xi_j = \int_{\mathcal{I}} \big[\theta_j^{1/2} \phi_j(t)u(t)\big]dt$. Both $u(t)$, θ_j and $\phi_j(t)$ are deterministic and $\phi_j(t)$ are orthogonal functions in time domain. $W = \sum_{j=1}^{d} Z_j \xi_j$ is a finite linear combination of independent Gaussian distributed random variables Z_j and thereby Gaussian. Hence, $\gamma = W + \int_{\mathcal{I}} \mu(t)u(t)dt$ is Gaussian. □

Theorem 5.2.1 *Let $\mathcal{X}$ and $\mathcal{Y}$ be any two classes of functional data having the same covariance kernel K. Suppose the following assumptions hold:*

1. *$\mathcal{X}$ and $\mathcal{Y}$ take values in $L^2[a, b]$.*

2. *μ_X and μ_Y, the means of $\mathcal{X}$ and $\mathcal{Y}$ respectively, exist and uniformly bounded in strong sense.*

3. *K is strictly positive definite and uniformly bounded.*

Assuming that prior probabilities $P(x \in \mathcal{X}) = p_1$ and $P(x \in \mathcal{Y}) = p_2$,

1. *If the distributions of classes $\mathcal{X}$ and $\mathcal{Y}$ are Gaussian, the probability of misclassification is*

$$\Delta = p_1\Phi(-k_1) + p_2\Phi(-k_2),$$

where Φ is the distribution function of the standard normal distribution, and k_1 and k_2 are real valued.

2. *If the distributions of classes $\mathcal{X}$ and $\mathcal{Y}$ are not Gaussian, the probability of misclassification is*

$$\Delta = p_1 P(R_X < -k_1) + p_2\big[1 - P(R_Y < k_2)\big],$$

where $R_X = \big(\gamma - E_X(\gamma)\big)/\sigma_\gamma$ and $R_Y = \big(\gamma - E_Y(\gamma)\big)/\sigma_\gamma$ are zero mean and unit variance random variables.

Proof:

$$\begin{aligned}\mathcal{D}^2(z, M_X) - \mathcal{D}^2(z, M_Y) &= \int_{\mathcal{I}} |z(t) - M_X(t)|^2 dt - \int_{\mathcal{I}} |z(t) - M_Y(t)|^2 dt \\ &= \int_{\mathcal{I}} [M_X^2(t) - M_Y^2(t)]dt - 2\int_{\mathcal{I}} z(t)[M_X(t) - M_Y(t)]dt = m - 2\gamma\end{aligned}$$

where $m = \int_{\mathcal{I}} [M_X^2(t) - M_Y^2(t)]dt$ and $\gamma = \int_{\mathcal{I}} z(t)[M_X(t) - M_Y(t)]dt$. Since M_X and M_Y do not depend on z, it follows that

$$\begin{aligned}E_X(\gamma) &= E_X\bigg\{\int_{\mathcal{I}} z(t)[M_X(t) - M_Y(t)]dt\bigg\} = \int_{\mathcal{I}} u(t)E[z(t)]dt \\ &= \int_{\mathcal{I}} u(t)\mu_X(t)dt = < \mu_X, u >,\end{aligned}$$

where $u(t) = M_X(t) - M_Y(t)$. $E_X(\gamma) = < \mu_X, u >$ is the expectation of γ given that z is distributed as $\mathcal{X}$. Similarly,

$$E_Y(\gamma) = E_Y\bigg\{\int_{\mathcal{I}} z(t)[M_X(t) - M_Y(t)]dt\bigg\} = \int_{\mathcal{I}} u(t)\mu_Y(t)dt \ = < \mu_Y, u >$$

is the expectation of γ given that z is distributed as $\mathcal{Y}$. Since $\mathcal{X}$ and $\mathcal{Y}$ have the same variance, $\mathrm{var}(z(t)) = K(s,t)$. It then follows that

$$\mathrm{var}(\gamma) = \mathrm{var}\bigg(\int_{\mathcal{I}} u(t)z(t)dt\bigg) = \int_{\mathcal{I}}\int_{\mathcal{I}} u(t)K(s,t)u(s)dsdt = \sigma_\gamma^2.$$

If $z \in \mathcal{X}$, $\mathcal{D}(z, M_X) \leqslant \mathcal{D}(z, M_Y)$ and $m - 2\gamma \leqslant 0$. Then

$$\begin{aligned} P(\mathcal{J}(z,\mathcal{D}) = 0 \mid z \in \mathcal{X}) &= P(\mathcal{D}(z, M_X) - \mathcal{D}(z, M_Y) > 0 \mid z \in \mathcal{X}) \\ &= P(\mathcal{D}^2(z, M_X) - \mathcal{D}^2(z, M_Y) > 0 \mid z \in \mathcal{X}) \\ &= P(m - 2\gamma > 0 \mid z \in \mathcal{X}) = P(\gamma < m/2 \mid z \in \mathcal{X}) \\ &= P(R_X < -k_1) \end{aligned}$$

where $R_X = \frac{\gamma - E_X(\gamma)}{\sqrt{var(\gamma)}}$ has univariate distribution with mean 0 and variance 1, and $k_1 = \frac{-m/2 + E_X(\gamma)}{\sqrt{var(\gamma)}}$. Similarly, if $z \in \mathcal{Y}$, $\mathcal{D}(z, M_X) > \mathcal{D}(z, M_Y)$ and $m - 2\gamma > 0$. Then

$$\begin{aligned} P(\mathcal{J}(z,\mathcal{D}) = 1 \mid z \in \mathcal{Y}) &= P(\mathcal{D}(z, M_X) - \mathcal{D}(z, M_Y) \leqslant 0 \mid z \in \mathcal{Y}) \\ &= P(\mathcal{D}^2(z, M_X) - \mathcal{D}^2(z, M_Y) \leqslant 0 \mid z \in \mathcal{Y}) \\ &= P(m - 2\gamma \leqslant 0 \mid z \in \mathcal{Y}) = P(\gamma \geqslant m/2 \mid z \in \mathcal{Y}) \\ &= P(R_Y \geqslant k_2) = 1 - P(R_Y < k_2) \end{aligned}$$

where $R_Y = \frac{\gamma - E_Y(\gamma)}{\sqrt{var(\gamma)}}$ has univariate distribution with mean 0 and variance 1, and $k_2 = \frac{m/2 - E_Y(\gamma)}{\sqrt{var(\gamma)}}$. The probability of misclassification of z into either $\mathcal{X}$ or $\mathcal{Y}$ is

$$\Delta = p_1 P(R_X < -k_1) + p_2\big[1 - P(R_Y < k_2)\big].$$

If $\mathcal{X}$ and $\mathcal{Y}$ are Gaussian distributed, then $P(R_X < -k_1) = \Phi(-k_1)$, $1 - P(R_Y < k_2) = \Phi(-k_2)$ and

$$p_1\Phi(-k_1) + p_2\Phi(-k_2).$$

The proof is complete. □

Suppose $E(X) = 0$, it follows from the above proof that $k_1 = -m/2\sigma_\gamma^2$ and $k_2 = m/2\sigma_\gamma^2$ if $E(Y) = 0$. Note that m can be viewed as the difference between $D^2(M_X, 0)$ and $D^2(M_Y, 0)$. The probability of misclassification goes to 0 as the difference between $D^2(M_X, 0)$ and $D^2(M_Y, 0)$ goes to infinity.

5.2.4 Numerical examples - Simulation

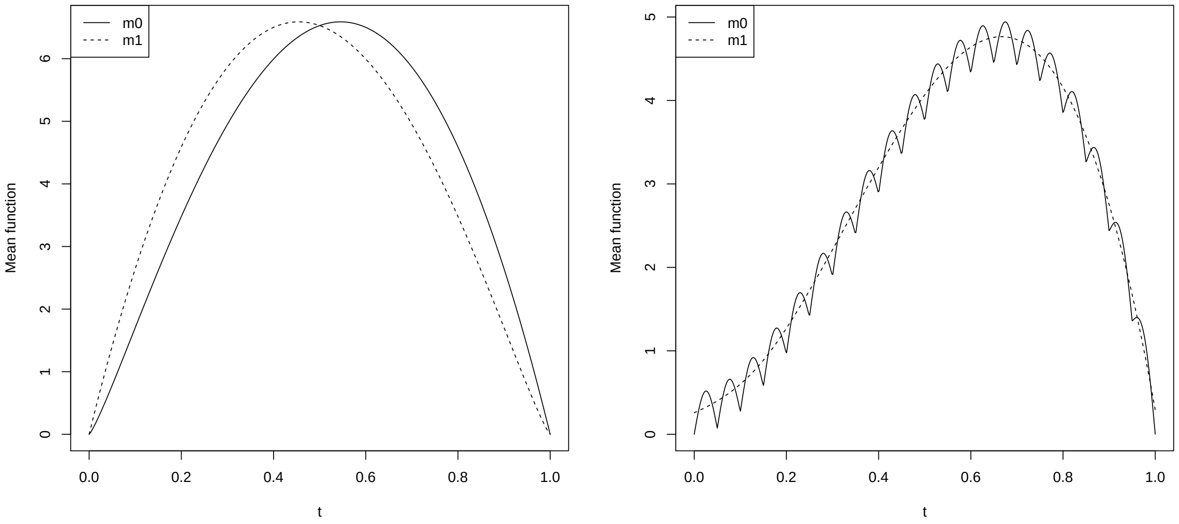
Three models have been simulated in order to generate the functional samples:

1. Model 1: The population P_0 consists of trajectories of the process $X(t) = m_0(t) + e(t)$, where $m_0(t) = 30(1-t)t^{1.2}$ and $e(t)$ is a Gaussian process with mean 0 and $\mathrm{cov}(X(s), X(t)) = 0.2\exp(-|s-t|/0.3)$. The process corresponding to P_1 differs from $X(t)$ only in the mean function and is given by $Y(t) = m_1(t) + e(t)$, where $m_1(t) = 30(1-t)^{1.2}t$.

2. Model 2: The population P_0 consists of trajectories of the process $X(t) = m_0(t) + e(t)$, where $m_0(t) = 30(1-t)t^2 + 0.5|\sin(20\pi t)|$ and $e(t)$ is a Gaussian process with mean 0 and $\mathrm{cov}(X(s), X(t)) = 0.2\exp(-|s-t|/0.3)$. Population P_1 is made of spline approximations (with 8 knots) of trajectories from the previous process.

3. Model 3: Consider model 1 above with $m_0(t) = 30(1-t)t^{1.2}$ and $m_1(t) = \delta m_0(t)$, where $\delta \in [0, 5]$.

The first two models above are adapted from Cuevas, Febrero and Fraiman (2007). We choose 500 distinct points for $t \in [0, 1]$. We compare the performance of DL2M with



(a) Model 1

(b) Model 2

Figure 5.1: Plot of means of competing samples of functional data.

some classifiers. Tables 5.1 and 5.2 present the performance of the classifiers for model 1 and model 2 respectively in term of average classification accuracy and quantiles of pobabilities of correct classification for the simulation procedures above. Figure 5.1 give the plot of class means, m_0 and m_1 of the competing populations, P_0 and P_1 respectively for model 1 and model 2. Figure 5.1(a) shows that populations P_0 and P_1 in model 1 consist of smooth functions. In model 2, population P_1 consists of smooth function of members of population P_0, as shown in Figure 5.1(b).

The possibility of using componentwise median as class centroid for median based classification method was raised in Hall, Titterington and Xue (2009) in high dimension setting. We shall extend this possibility for functional data and see its performance in the simulation study for models above. In Chapter 2, we raise the possibility of classifying an observation to the class for which it attains minimal rank in $\mathbb{R}^d$ and presented some upgraded versions of the approach. Similarly in Chapter 4, we proposed classifying an observation based on the distribution of the outlyingness of its spatial rank in $\mathbb{R}^d$ and its variants, and argued that both methods can be extended into infinite dimension. We apply this minimal rank classifier and minimal rank distribution classifier to functional data and classify observations from the above models to the class for which each achieve minimal rank and minimal rank distribution function respectively. DL2M is compared with minimum rank classifier (RC), minimal rank distribution classifier (RDC), classifier based on L_1 distance to L_1 median, independence rule of Dudoit, Fridlyand and Speed (2002), centroid classifier (Hastie, Tibshirani and Friedman, 2001) and maximum functional depth classifier (Cuevas, Febrero and Fraiman, 2007) using the above models. For the maximum functional depth classifier, four functional depths are considered. The depth functions are h-mode depth (HMD), Fraiman-Munic depth (FMD), random projection depth (RPD) and random Tukey depth(RTD). To compute these functional depth, we use R package *fda.usc* with 10% trimming, and assign observations to class with maximum depth value. Denote centroid classifier by C.C, independence rule by ind and classifier based on minimum distance to L_1 median by DL1M. We choose the sizes of both training samples and validation samples of P_0 and P_1 to be 100 and repeat the simulation 1000 times.

Classifiers based on L_2 distance to spatial median and L_1 distance to L_1 median compete favourably with other classifiers. Among the depth based classifiers, it is seen that maximum depth classifier based on h-mode depth achieves highest average probability of correct classification for models 1 and 2. All the classifiers perform well as shown in Table 5.1 and Table 5.2 for model 1 and model 2 respectively. Generally, classification procedures based on L_2 distance to spatial median and L_1 distance to the L_1 median can be seen as competitive with depth based methods, centroid classifier, 1NN and independence rule for location problem. In the next section, we shall generalise the L_1 median and L_2 median to L_p median for various values of p and examine the performance of its associated

Table 5.1: Performance of some classifiers for model 1.

	Probability of correct classification										
				Functional Depth Classifiers							
	C.C	1NN	Ind	HMD	FMD	RPD	RTD	RC	RDC	DL1M	DL2M
Minimum	0.935	0.925	0.940	0.925	0.860	0.915	0.915	0.920	0.905	0.935	0.930
25% quantile	0.969	0.965	0.970	0.970	0.920	0.955	0.955	0.970	0.955	0.965	0.970
Mean	0.974	0.971	0.975	**0.976**	0.930	0.965	0.963	**0.976**	0.962	0.972	**0.976**
Median	0.975	0.970	0.975	0.975	0.930	0.965	0.965	0.975	0.965	0.975	0.975
75% quantile	0.980	0.980	0.985	0.985	0.940	0.975	0.970	0.985	0.970	0.980	0.985
Maximum	1.000	1.000	1.000	1.000	0.980	1.000	0.995	1.000	1.000	1.000	1.000

Table 5.2: Performance of some classifiers for model 2.

	Probability of correct classification										
				Functional Depth Classifiers							
	C.C	1NN	Ind	HMD	FMD	RPD	RTD	RC	RDC	DL1M	DL2M
Minimum	0.535	0.750	0.565	0.565	0.460	0.570	0.555	0.585	0.560	0.555	0.560
25% quantile	0.730	0.825	0.765	0.805	0.640	0.660	0.630	0.770	0.745	0.710	0.765
Mean	0.786	0.842	0.845	**0.855**	0.683	0.684	0.657	0.826	0.789	0.778	0.838
Median	0.785	0.840	0.850	0.865	0.680	0.685	0.655	0.830	0.790	0.775	0.840
75% quantile	0.840	0.860	0.935	0.910	0.730	0.710	0.680	0.885	0.835	0.845	0.925
Maximum	0.975	0.930	1.000	1.000	0.870	0.785	0.765	1.000	0.955	0.990	1.000

classifiers.

5.3 Minimal Distance to L_p Median Classifier

Define a L_p metric $\mathcal{D}_p$ as

$$\mathcal{D}_p(z,a) = \left\{ \int_{\mathcal{I}} |z(t) - a(t)|^p dt \right\}^{1/p},$$

for $t \in \mathcal{I}$ and $p \in \mathbb{R}$. It is obvious that $\mathcal{D}(z,a) > 0$ if $z \neq a$ and $\mathcal{D}(z,a) = 0$ iff $z = a$. In functional analysis, L^p space is know for its completeness property when $p \geqslant 1$. The L_p metric above satisfies triangle inequality for $p \geqslant 1$ as shown in Rudin (1991). For $0 < p < 1$, L^p space is complete if its associated L_p metric satisfies the triangle inequality. For $0 < p < 1$, Rudin (1991) suggested

$$\mathcal{D}_p(z,a) = \left\{ \int_{\mathcal{I}} |z(t) - a(t)|^p dt \right\}.$$

Suppose $X_1, X_2, \ldots, X_n \in \mathcal{X}$, the L_p median of X, denoted by M_p, is defined as

$$M_p = \arg_f \min E \left\{ \int |X(t) - f(t)|^p dt \right\}^{1/p}.$$

When $p = 1$, M_1 is called co-ordinatewise median while it is spatial median or L_2 median when $p = 2$. The applicability of this L_1 median for functional data lies in assuming that functions are sampled at common distinct points.

Suppose $\mathcal{X}$ and $\mathcal{Y}$ are two classes of observations, having prior probabilities P_1 and P_2 and sizes n_1 and n_2 respectively. Let $X_1, X_2, \ldots, X_{n_1} \in \mathcal{X}$ and $Y_1, Y_2, \ldots, Y_{n_2} \in \mathcal{Y}$ take values from Banach space. Suppose $p_1 = p_2$, M_{X_p} and M_{Y_p} are L_p median of the data cloud $\mathcal{X}$ and $\mathcal{Y}$ respectively. The classification rule is to assign observation z into class $\mathcal{X}$

if

$$\mathcal{D}_p(z, M_{X_p}) \leqslant \mathcal{D}_p(z, M_{Y_p}), \tag{5.3.1}$$

otherwise to class $\mathcal{Y}$, where

$$\mathcal{D}_p(z, M_{X_p}) = \left\{ \int_{\mathcal{I}} |z(t) - M_{X_p}(t)|^p dt \right\}^{1/p} \text{ and } \mathcal{D}_p(z, M_{Y_p}) = \left\{ \int_{\mathcal{I}} |z(t) - M_{Y_p}(t)|^p dt \right\}^{1/p}.$$

We use L_p median with corresponding L_p distance because it is L_p median that minimises its corresponding L_p distance. The possibility of using L_1 median lies in fixing t for each class members. In this case, function space does not necessarily need to be a Hilbert space. Then for $p = 1$, the difference between $\mathcal{D}_1(z, M_{X_1})$ and $\mathcal{D}_1(z, M_{Y_1})$ can be viewed as $\sum_{k=1}^{d}[|z_k - M_{X_{1k}}| - |z_k - M_{Y_{1k}}|]$ in $\mathbb{R}^d$ for $d > 1$, where d can be finite or infinite which is the case of componentwise L_1 median classifier in Hall, Titterington and Xue (2009). Suppose $p_1 \neq p_2$, the classification rule based on DL2M will be to assign an observation, z into class $\mathcal{X}$ if

$$\mathcal{D}_2^2(z, M_X) - \mathcal{D}_2^2(z, M_Y) \leqslant \log_e \left(\frac{p_2}{p_1} \right), \tag{5.3.2}$$

otherwise to class $\mathcal{Y}$. This is equivalent to assigning z into class $\mathcal{X}$ if

$$\int_{\mathcal{I}} z(t)[M_X(t) - M_Y(t)]dt \leqslant C,$$

otherwise to $\mathcal{Y}$, where $C = \frac{1}{2} \int_{\mathcal{I}} [M_X(t)^2 - M_Y(t)^2]dt + \log_e \left(\frac{p_2}{p_1} \right)$.

It has been proved in Hall, Titterington and Xue (2009) that the probability of misclassification based on L_1 distance to componentwise L_1 median asymptotically goes to zero as $d \to \infty$, competing class sizes diverge and suitable conditions for componentwise L_1 median hold. This result is stated formally in Theorem 5.3.1 below.

Theorem 5.3.1 (Hall, Titterington and Xue, 2009) *Assume that the following as-*

sumptions hold for $p = 1$:

1. *$\mathcal{X}$ and $\mathcal{Y}$ take values in $L^1[a, b]$ sampled at d distinct points*
2. *μ_X and μ_Y, the means of $\mathcal{X}$ and $\mathcal{Y}$ respectively, exist and uniformly bounded in strong sense*
3. *K is strictly positive definite and uniformly bounded*
4. *components of the difference between medians of $\mathcal{X}$ and $\mathcal{Y}$ is nonzero*
5. *standard α-mixing condition hold*
6. *sample sizes n_1 and n_2 of $\mathcal{X}$ and $\mathcal{Y}$ diverge as $d \to \infty$,*

then with probability converging to 1 as d increases,

$$\Delta = p_1 P\big(\mathcal{J}(z, \mathcal{D}) = 0 \mid z \in \mathcal{X}) + p_2 P(\mathcal{J}(z, \mathcal{D}) = 1 \mid z \in \mathcal{Y}) \to 0.$$

Similar to intuitive features of RC, RRC and RDC and their variants in Chapter 2 - 4, classifier based on minimal L_p distance to L_p median enjoys easy lending to multiclass extension. Suppose there are J classes, then assign z to class $\mathcal{X}_k$, $1 \leqslant k \leqslant J$ if

$$\mathcal{D}_p(z, M_{k_p}) = \min_j \mathcal{D}_p(z, M_{j_p}),$$

where M_{j_p} is the L_p median of the jth class, $j = 1, 2, ..., J$. For $J(\geqslant 2)$ populations with prior probabilities $p_1, \cdots, p_J$, the associated probability of misclassification is

$$\Delta = \sum_{j=1}^{J} p_j P\bigg(\mathcal{D}_p(z, M_{j_p}) \text{ is not the minimum } \mid z \in \mathcal{X}_j\bigg).$$

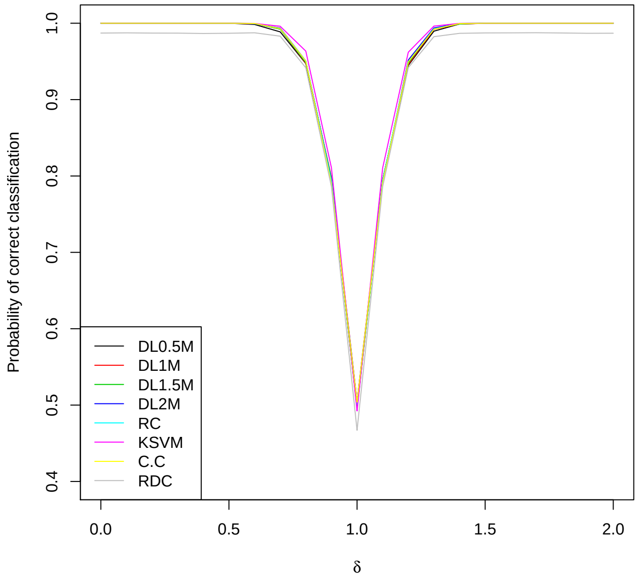


Figure 5.2: Performance of classifiers for model 3.

Table 5.3: Performance of classifier based on L_p distance to L_p median for model 1.

	Probability of correct classification for different values of p							
	p=0.5	p=1	p=1.2	p=1.5	p=2	p=2.5	p=10	p=100
Minimum	0.915	0.935	0.930	0.935	0.930	0.925	0.920	0.925
25% quantile	0.950	0.965	0.965	0.970	0.970	0.970	0.965	0.960
Mean	0.961	0.972	0.974	0.975	**0.976**	**0.976**	0.972	0.969
Median	0.960	0.975	0.975	0.975	0.975	0.975	0.975	0.970
75% quantile	0.970	0.980	0.985	0.985	0.985	0.985	0.980	0.980
Maximum	0.990	1.000	1.000	1.000	1.000	1.000	1.000	1.000
S.E.	0.0004	0.0004	0.0004	0.0003	0.0003	0.0003	0.0004	0.0004

Table 5.4: Performance classifier based on L_p distance to L_p median for model 2.

	Probability of correct classification for different values of p							
	p=0.5	p=1	p=1.2	p=1.5	p=2	p=2.5	p=10	p=100
Minimum	0.555	0.555	0.530	0.535	0.560	0.565	0.565	0.490
25% quantile	0.704	0.710	0.710	0.710	0.765	0.715	0.715	0.710
Mean	0.775	0.778	0.782	0.780	**0.838**	0.781	0.778	0.780
Median	0.775	0.775	0.775	0.780	0.840	0.783	0.775	0.780
75% quantile	0.841	0.845	0.845	0.845	0.925	0.850	0.845	0.845
Maximum	0.985	0.990	1.000	0.990	1.000	0.985	1.000	0.990
S.E.	0.0028	0.0029	0.0029	0.0029	0.0033	0.0028	0.0028	0.0029

5.4 Numerical Examples

In this section, we shall investigate the performance of classification method based on minimal L_p distance to L_p median for various values of p based on simulation information in Section 5.2.4 and analysis of real data. We shall denote this classifier by DLpM for various values of p.

5.4.1 Numerical example - simulation

Consider models 1 and 2 in subsection 5.2.4. Tables 5.3 and 5.4 give the mean and quantiles of proportion of correctly classified test data for model 1 and 2 respectively for different values of p. For model 1, the average probability of correct classification are equivalent and close to 1 for various values of p. This is similar for model 2, the average probability of correct classification are equivalent except for $p = 2$, where there is a noticeable higher value of average proportion of correctly classified test data. For separable data in L^p space, the choice of p in the metric is of less importance because of the equivalence of L_p norm for $p \geqslant 1$. This means that when observations from distinct classes take values from L^p space, irrespective of L_p median used, classifier based on its corresponding distance function will perform well for location problem. This is illustrated in Table 5.3 for model 1 and Figure 5.2 for model 3.

5.4.2 Example: real data

We applied our method to six real data examples. The real datasets are LSVT voice rehabilitation data, Phoneme data, lung cancer data, internet advertisement data, mass-spectrometry data and growth data. The LSVT voice rehabilitation data (Tsanas et al., 2014) consists of two classes of observations, which are acceptable(size = 42) and unacceptable(size = 84). A training sample of size 30 and a validation sample of size 12 are selected from each of the two classes. Phoneme frequency data, denoted by phoneme data, arose from a collaboration between Andreas Buja, Werner Stuetzle and Martin Maechler, and was used as an illustration in Hastie, Buja and Tibshirani (1995). Phoneme data was formed by selecting five phonemes based on discretized log-periodograms of digitized speech. It consist of five classes of observations, which are aa(size = 695), ao(size = 1022), dcl(size = 757), iy(size = 1163) and sh(size = 872). A training sample of size 200 and a validation sample of size 100 for each of the classes are chosen. Lung cancer data (Hong and Yang, 1991) is a sparse data with three classes of sizes 9, 13 and 10. For this data, we select training samples of sizes 5, 7, 6 and validation samples of sizes 4, 6, 4 respectively. Two out of 57 features have missing values and are removed. Internet advertisement data (Kushmerick, 1998), denoted by internet ads, is a set of possible advertisements on internet pages. It consists of two classes, ad and nonad with class sizes 459 and 2820 respectively, and 1558 features. A training sample of size 200 and a validation sample of size 100 from each of the two classes are chosen, and 1554 features are used from this dataset. Mass-spectrometry data (Mahé and Veyrieras, 2013), denoted by micromass data, consists of two classes, pure spectra and mixed spectra with class sizes 571 and 360 respectively. We choose a training sample of size 200 and a validation sample of size 150 from each of the two classes. Growth data (Ramsay and Silverman, 2005) consists of the heights (in centimeters) of 54 girls and 39 boys measured at a set of thirty one ages from one to eighteen years old. We choose a training sample of size 25 and a validation sample of size 14 from each of the two classes. Summary of these data are given in Table 5.5 below. Phoneme data can be found in a R package *fds* while others are taken from UCI Machine Learning Repository. For depth classifiers, the experiment is repeated 100 times and average probability of correct classification is computed. For each of the datasets, we assume equal prior probabilities for competing classes.

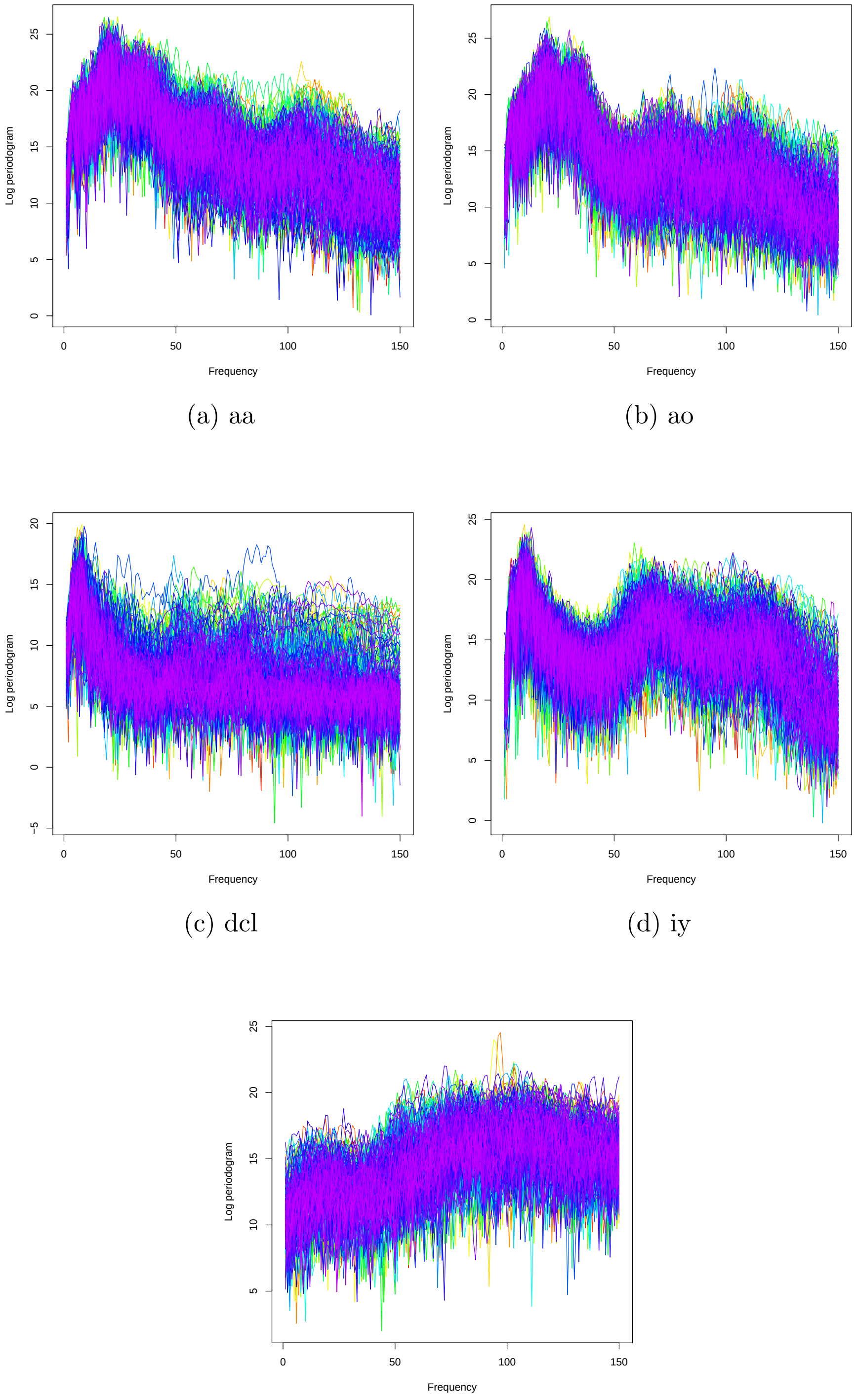
Table 5.6 presents the comparison of classifiers based on the probabilities of correct classification. For growth data, all the classifiers perform well while maximum functional depth classifier based on random projection depth perform best among others. DLpM performs poorly as well as RC for micromass and internet ads data except for $p = 2$. DL2M performs well in all cases and has highest proportion of correctly classified data except for growth data. All the functional depth classifiers compete well with other classifiers except for lung cancer data, internet advertisement data and mass-spectrometry data. Minimal rank classifier competes favourably with depth based procedures.

5.4.3 Optimal choice of p

In this subsection, we want to estimate the value of p for which classification rule based on minimal L_p distance of test data to its corresponding L_p median is optimal. In practice, p is not always unique. In low finite dimensional setting, Dutta and Ghosh (2012b) proposed classifier based on maximal L_p depth. In their proposal, p is estimated by maximising the joint likelihood function of the sample or its natural logarithm. In functional and infinite dimensional setting, estimation of p by minimising joint likelihood function of the sample is possible using kernel estimator of probability density function but has high computational time. We choose p by cross validation error.

Let us represent classification rule based on DLpM by $\mathcal{J}(z, \mathcal{D}_p)$, where $\mathcal{J}$ is

$$\mathcal{J}(z, \mathcal{D}_p) = \begin{cases} 1, & \text{if } \mathcal{D}_p(z, M_{X_p}) \leqslant \mathcal{D}_p(z, M_{Y_p}) \\ 0, & \text{if otherwise} \end{cases}$$



(a) aa

(b) ao

(c) dcl

(d) iy

(e) sh

Figure 5.3: Phoneme data

Table 5.5: Description of some real dataset.

Dataset	No of classes	Training sample size	Val. sample size	No of features
LSVT voice rehabilitation	2	30, 30	12, 12	310
Phoneme data	5	200 each	100 each	256
Lung cancer data	3	5,7,6	4,6,4	55
Internet ads	2	200 each	100 each	1554
Micromass data	2	200 each	150 each	1300
Growth data	2	25 each	14 each	31

Table 5.6: Comparison of classifiers based on probability of correct classification for some real datasets.

	Probability of correct classification									
			Maximum functional depth classifiers				Distance to L_p medians			
Dataset	C.C	RC	FMD	HMD	RPD	RTD	p = 0.5	p = 1	p = 1.5	p = 2
LSVT voice	**0.8333**	0.5833	0.5000	0.5833	0.5613	0.5579	0.6667	0.6667	0.6667	0.6667
Phoneme	0.8480	0.8380	0.8280	0.8740	0.8290	0.8213	0.8340	0.8640	0.8720	**0.8760**
Lung cancer	**0.5714**	**0.5714**	0.2857	0.4286	0.4279	0.4036	0.4286	0.5000	**0.5714**	**0.5714**
Internet ads	0.7740	0.5550	0.5000	0.5000	0.5750	0.5564	0.5000	0.5000	0.5800	**0.7950**
Micromass	0.6367	0.5000	0.8214	0.5000	0.5003	0.5003	0.5000	0.5000	0.6033	**0.6533**
Growth	0.8929	0.8571	0.8929	0.8448	**1.0000**	0.8682	0.8929	0.8929	0.8929	0.8929

Define a cross validation estimator of error rate as

$$\widehat{err}(p) = \frac{1}{n}\sum_{i=1}^{n} I\{J(x_i, \mathcal{D}_p) \neq I_i\},$$

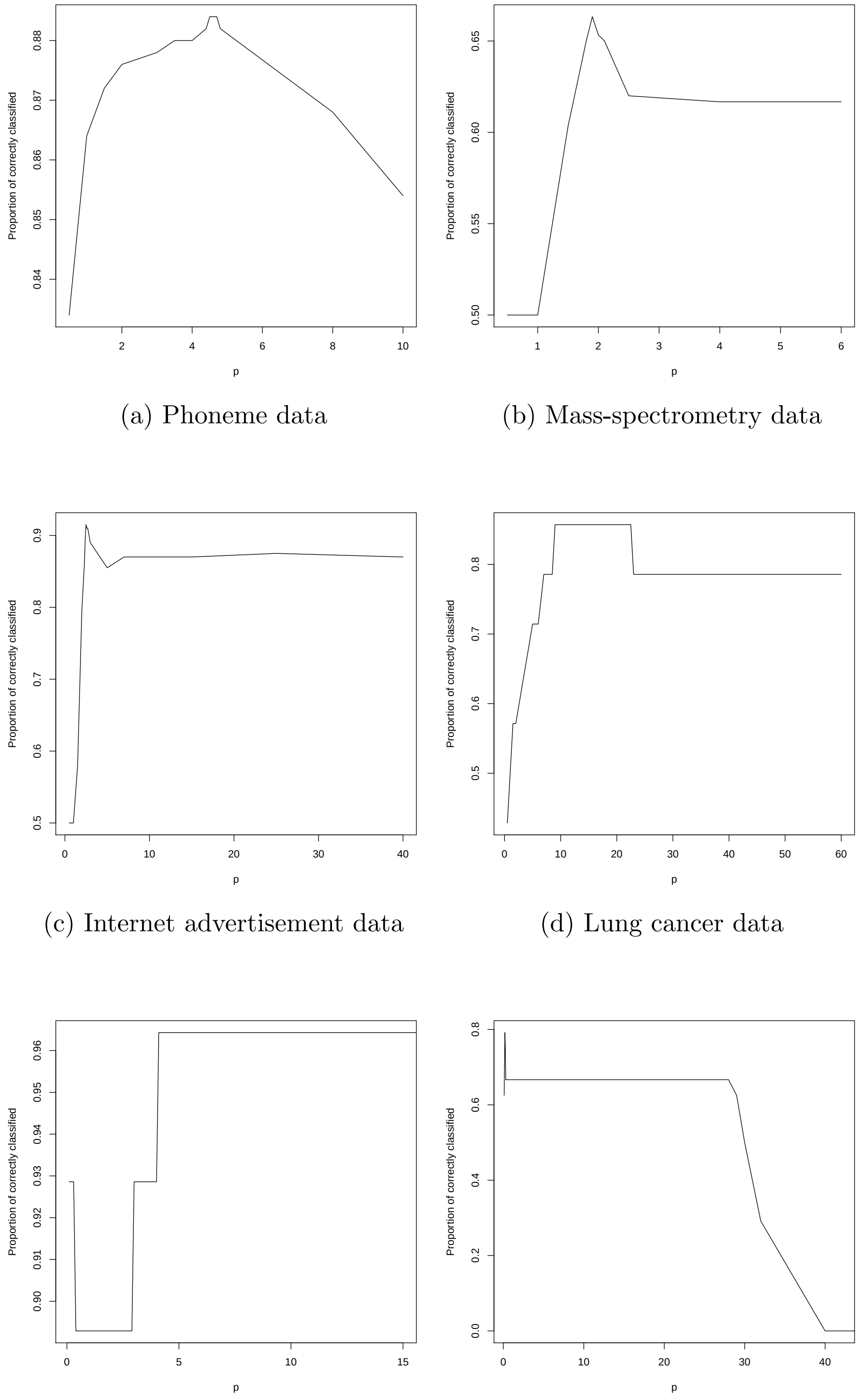
where $\mathcal{D}_p$ denotes the L_p distance function of ith data from each of the competing classes, $\mathcal{I}_i$ denotes the class label of ith observation x_i. The optimal value of p, denoted by p_o, is

$$p_o = inf_p \widehat{err}(p). \tag{5.4.1}$$

Using the real datasets discussed in Subsection 5.4.2, p_o is 1.9 with probability of correct classification being 0.6633 for the micromass data. The optimal value of p for internet advertisement data, $p_o = 2.5$ gives highest probability of correct classification which is 0.915. For the lung cancer data, optimal value of p is 8.7 with probability of correct classification being 0.8571429. For LSVT voice rehabilitation data, $p_o = 0.2$ with the probability value 0.7917. The highest probability of correct classification obtained for Phoneme data is 0.884 for $p \in [4.5, 4.7]$, and so $p_o = 4.5$. Similarly for lung cancer data and growth data, highest probability of correct classification obtained based on DLpM are 0.8571 and 0.8929 for $p \in [8.7, 22.6]$ and $p \in [4.1, \infty)$ respectively (see Figure 5.4) and hence, p_o for respective data are 8.7 and 4.1. We summarise this numerical results using some plots. Figure 5.4 present the plot of proportion of correctly classified test data against various values of p. Table 5.7 presents the optimal value of p and its corresponding proportion of correctly classified test data for the six datasets.

5.5 Samples with Different Scale

The performance of distance based methods are generally poor when competing classes have different scale or the principal difference among the competing classes is in scale. In order to overcome this, we thought of dividing each component by its standard deviation



(a) Phoneme data

(b) Mass-spectrometry data

(c) Internet advertisement data

(d) Lung cancer data

(e) growth data

(f) LSVT voice rehabilitation data

Figure 5.4: Plot of proportions of correctly classified data against various values of p.

Table 5.7: Optimal value of p and corresponding probability of correct classification for some real dataset.

Dataset	optimal p	Probability of correct classification
LSVT voice	0.2	0.7917
Phoneme	4.5	0.8840
Lung cancer	8.7	0.8571
Internet ads	2.5	0.9150
Micromass	1.9	0.6633
Growth	4.1	0.9643

and then apply DLpM. We denote this by DLpM-S. This approach may fail for some functions where the mean difference is close to zero.

To illustrate this, consider model 3 in Subsection 5.2.4. Suppose populations P_0 and P_1 consist of trajectories of the processes $X(t) = m_0(t) + e(t)$ and $Y(t) = \delta m_0(t) + e_1(t)$ respectively, where $m_0(t) = 30(1 - t)t^{1.2}$, $e(t)$ is a Gaussian process with mean 0 and $\text{cov}(X(s), X(t)) = 0.2 \exp(-|s - t|/0.3)$, $e_1(t) = 2e(t)$, $t \in [0, 1]$ and $\delta \in [0, 2]$. Figure 5.5 gives the performance of some classifiers. DL2M-S is compared with KSVM, support vector machine using kernel trick (Rossi and Villa, 2006) and some other classifiers. It is seen here that within the neighbourhood of $\delta = 1$, the principal difference between the competing classes is not in location and so, the probability of correct classification tends to 0.5. The reason is that dividing each feature by its standard deviation for each class in the neighbourhood of $\delta = 1$ makes the resulting observations in each class become alike with probability of correctly classifying an observation to either of the competing class being 0.5. As δ moves away from 1, the difference between $\mathcal{D}_2(z, M_{X_2})$ and $\mathcal{D}_2(z, M_{Y_2})$ goes away from zero with probability tending to one. Hence pre-scaling of data does not improve the chance of correctly classifying data in this case.

Furthermore, we suggest standardizing L_p median by dividing each component of the L_p median by its corresponding median absolute deviation and then perform DL2M. Using this approach on the example above, the performance of DL2M is not improved in the

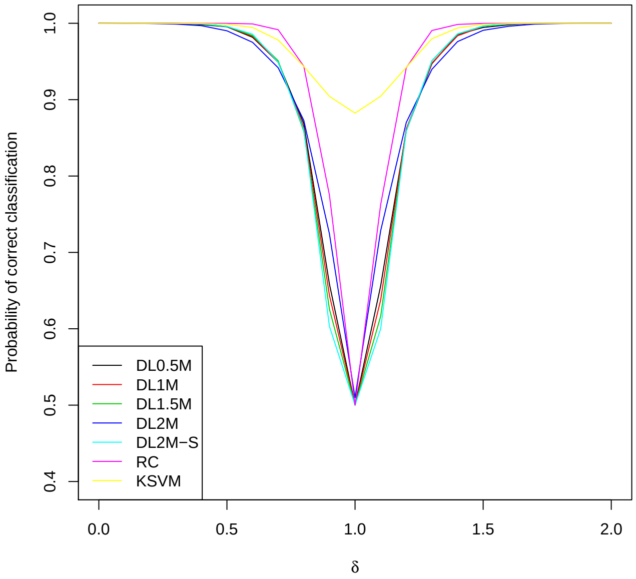


Figure 5.5: Performance of classifiers for population distributions with different covariance kernels.

neighbourhood of $\delta = 1$.

Chapter 6

Concluding Remarks and Future Work

Classification aims at obtaining rules that describe the separation between groups of observations and allocate each new observation to one of the known groups. A good classification procedure is the one that classifies observations from unknown populations correctly. Two major approaches to classification, identified in this study, are parametric and nonparametric. Parametric approach requires making assumptions about the distribution of the population while nonparametric approach does not. Parametric approaches include linear and quadratic discriminant analysis, which assume multivariate normal distribution for the data. The limitations of parametric approaches include lack of robustness against outliers. This thesis focuses on nonparametric approach. The motivation for nonparametric classification methods includes robustness against outliers, distribution-free property, easy lending to multiclass extension.

In Chapter one, we have reviewed parametric approach to discriminant analysis and investigated the optimal performance of linear and quadratic discriminant functions under normality condition based on simulation and provide solutions of some theoretical examples. We derive expressions for Bayes error for multivariate normal distributions, multivariate Laplace distributions and multivariate t distributions with the same degree of freedom, under location shift. The theoretical probabilities of misclassification were compared with empirical error rates, based on simulation, when competing populations differ in location and scale using LDA and QDA respectively. The sample estimates of probability of misclassification associated with LDA and QDA are good approximation for their respective population versions.

In Chapter two, we propose nonparametric methods for classifying d-dimensional observations based on multivariate rank. They are minimal rank classifier(RC) and affine invariant version of minimal rank classifier(AIRC). We show that these classifiers are optimal Bayes rule under suitable conditions. The performance of these methods are examined by using simulation and their results are compared with the results from existing methods. The variations in total probability of misclassification of d-dimensional observations associated with a pair of multivariate distributed random samples for the cases where location vectors and dispersion matrices are homogenous and heterogeneous are studied. Minimal rank classifier performs well when competing class distributions are spherically symmetric with equal covariance matrices. When distributions of competing classes are elliptically symmetric, the error rates associated with minimal rank classifier are not in any specific order of the value of correlation existing among variables. This is due to non-invariance of spatial rank under general affine transformation and because of this, we construct AIRC using transformation and re-transformation technique. When the competing distributions have different covariance matrices, RC and AIRC perform poorly compared to QDA because scale term is not involved in their formulation. To overcome this limitation, we construct a classifier based on volume of central rank region.

In chapter three, we propose rank region classifier (RRC) and its variant. This method assigns observations to the class for which it attains minimum volume of rank region. Affine invariant version of spatial rank is used to compute the volume of rank region to make the volume equivariant under general affine transformation of the data. RRC performs well when the principal difference among the distributions of competing populations is in location parameter. To improve this classifier, we set a threshold for assigning an observation to a population based on the ratio of volumes of rank regions of the competing populations. When the principal difference among the distributions of competing populations is in location parameter, the improved version of rank region classifier (RRC-1) reduces to usual rank region classifier. When covariance matrices ($\mathbf{\Sigma}_j, j = 1, 2, \ldots, J$) of J competing populations are the same (say $\mathbf{\Sigma}$), rank region classifier reduces to minimal affine invariant rank classifier. Also, when $\mathbf{\Sigma}$ is a scalar multiple of $\mathbf{I}_d$, the minimal affine invariant rank classifier reduces to minimal rank classifier. .

The performance of these methods (RC, AIRC, RRC and RRC-1) are examined by using simulation and real data set, and their results are compared with the results from existing methods. The methods perform competitively under necessary conditions. These classifiers can be practically implemented for large dimension, unlike depth based classifier. It may worth mentioning here that our simulation work based on volume of central rank regions, in Chapter three, was restricted by the heavy computation cost. The R programs which we used for our computation are quite time consuming. For computing misclassification errors associated with rank region classifier for two competing distributions with training samples of size 100 each and test sample of size 100 each based on 1000 iterations for five different covariance matrices took more than fifteen days in a machine with a dual-core 3.00GHz CPU with 4GB RAM. Using C programming for the same training sample size and test sample size based on 1000 iterations for five different covariance matrices, it took almost four days in the same machine. For computing the volume of central rank region in C programming, we use qhull of Barber, Dobkin and Huhdanpaa (1996). For computing the volume of central rank region in r programming, we use R-package *geometry*.

We note that RRC has a high computational time compared to LDA, QDA, RC and AIRC. Comparing the computational time of RRC with some depth based classifier like simplicial depth and half-space depth, RRC is still much better because depth based classifiers can not just do it. The reason for high computational time is that affine invariant rank is first computed before computing volume of central rank region, on which the classifier is constructed. For high dimensional data, the estimated central rank region is computationally unstable (Guha and Chakraborty, 2013) due to the curse of dimensionality. For computation of Oja depth and projection depth, we use R-packages *depth* and *fda.usc* respectively with 25% trimming for projection depth when sample size is large. For the SVM we use radial basis kernel as implemented in the R-Package *kernlab* and employ 5-fold cross-validation.

In chapter four, we propose minimal rank distribution classifier (RDC) and its affine invariant versions (AIRDC and RDA-A). Minimal rank distribution classifier assigns observations to class with least distribution function of outlyingness of spatial rank. Due to the lack of robustness of minimal rank distribution classifier against deviation of distributions of competing populations from spherical symmetry, we propose two affine invariant versions of minimal rank distribution classifier. One based on transformation and retransformation technique of Chakraborty (2001) and one based on pre-multiplying the data with the inverse of estimate of $\mathbf{\Sigma}^{\frac{1}{2}}$. Both RDC and its invariants are Bayes rule under some certain conditions. Analysis of real data show that the choice of covariance estimator has a standing implication on RDA-A. When using MCD estimator of covariance matrix for data with small size, we suggest that value of α should be close to 1 to ensure relatively low misclassification error. For competing class of data with large sample size, the choice of $\alpha \in [0.5, 1)$ does not significantly affect the performance of RDA-A. When $\mathbf{\Sigma}$ is a scalar multiple of $\mathbf{I}_d$, AIRDC reduces to RDC.

When data are functions, many multivariate techniques fail to perform well. In Chapter five, we propose classification method based on L_2 distance to spatial median. The L_2 distance classifier assigns each observation to the population for which the observation attains minimum L_2 distance to the population's spatial median. Robustness is one of the interesting features of statistical methods based on functional outliers. This is because functional outliers can affect statistical analysis in many different ways and are not always easy to identify (L*ó*pez-Pintado and Romo, 2006). Spatial median is robust against outliers and easy to compute. The classifier based on distance to spatial median enjoys easy lending to multiclass extension. When the distributions of the competing populations are Gaussian, we derive an expression for the probability of misclassification for two-class problem. This method is generalised into classification approach based on the L_p distance to its corresponding L_p median for some values of p. Throughout this chapter, the same value of p is assumed for all competing groups of functional data. The performance of this generalised L_p distance classifier is examined through simulation and real data analysis. To obtain optimal classifier, we define the optimal L_p distance classifier based on the p_o, optimal value of p, where p_o is determined by cross validation.

6.1 Further Work and Possible Extensions

When dimension of observations is greater than sample size ($d > n$), estimate of covariance matrix $\mathbf{\Sigma}$ degenerates and becomes singular. This makes estimating $v(\alpha)$, that leads to the choice of $\mathbf{X}(\alpha)$ that removes the effect of correlations among features in each of the competing classes, practically impossible. As a result, it limits the applicability of AIRC and AIRDC, as well as RRC for elliptically symmetric distributions. In future we will like to work on overcoming challenges of high dimensionality in the use of affine invariant spatial rank via transformation and re-transformation techniques. Also, the execution of RDA-A becomes practically infeasible. Hastie, Buja and Tibshirani (1995) suggested penalizing covariance functional. Dudoit, Fridlyand and Speed (2002) suggested assuming independence of components while Guo, Hastie and Tibshirani (2007) suggested regularizing covariance functional. Our future work may focus on executing RDA-A when dimension of observations is greater than sample size, without penalizing or regularizing the degenerated covariance matrix.

For functional data, minimal L_2 distance to L_2 median classifiers perform well for either univariate functional data or multivariate functional data when the principal difference among the competing classes of observations are not in covariance kernels. Several methods have been proposed under this setting but these methods fail when the principal difference are in covariance kernels. In future, we will like to work on how to overcome this problem using nonparametric approach. The possibility of incorporating different covariance kernels in functional classification problem and solving the problem non-parametrically may be extended into L_p distance classifiers for some p.

Bibliography

[1] Ackermann, M. and Strimmer, K. (2009). A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*. Volume 90, article 47.

[2] Alonso A. M., Casado, D. and Romo, J. (2012). Supervised classification for functional data: A weighted distance approach. *Computational Statistics and Data Analysis*. Volume 56, Issue 7, pp. 2334 - 2346.

[3] Anderson, T. W. (1972). Asymptotic evaluation of the probabilities of misclassification by linear discriminant functions. Stanford University, Department of Statistics. Technical Report No 10.

[4] Anderson, T. W. (1984). An introduction to multivariate statistical analysis. John Wiley & Sons, Inc., New York.

[5] Anderson, T. W. and Bahadur, R. R. (1962). Classification into two multivariate normal distributions with different covariance matrices. *The Annals of Mathematical Statistics*. Volume 33, Number 2. pp. 420 - 431.

[6] Balanda, K. P. and MacGillivray, H. L. (1990). Kurtosis and spread. *The Canadian Journal of Statistics*, Volume 18, pp. 17 - 30.

[7] Barber, C. B., Dobkin, D. P. and Huhdanpaa, H. (1996). The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software.* Volume 22, Issue 4, pp. 469 - 483.

[8] Biau G., Bunea, F. and Wegkamp, M. (2005). Functional classification in Hilbert spaces. *IEEE Transactions on Information Theory.* Volume 51, Issue 6, pp. 2163 - 2172.

[9] Bickel, P. and Levina, E. (2004). Some theory for Fisher's linear discriminant function, naive Bayes, and some alternatives when there are many more variables than observations. *Bernoulli.* Volume 10, Number 6, pp. 989 - 1010.

[10] Boser E., Guyon, M. and Vapnik, V. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the Fifth ACM Workshop on Computational Learning Theory. Pittsburgh, PA.* pp. 144 - 152.

[11] Chakraborty, A. and Chaudhuri, P. (2014). The deepest point for distributions in infinite dimensional spaces. *Statistical Methodology.* Volume 20, pp. 27 - 39.

[12] Chakraborty, B. (2001). On affine equivariant multivariate quantiles. *Annals of the Institute of Statistical Mathematics.* Volume 53, Number 2, pp. 380 - 403.

[13] Chakraborty, B. and Chaudhuri, P. (1996). On a transformation and re-transformation technique for constructing affine equivariant multivariate median. *Proceedings of the American Mathematical Society.* Volume 124, Number 8, pp. 2539 - 2546.

[14] Chakraborty, B. and Chaudhuri, P. (1998). On an adaptive transformation and re-transformation estimate of multivariate location. *Journal of the Royal Statistical Society: Series B.* Volume 60, part 1, pp. 145 - 157.

[15] Chakraborty B., and Chaudhuri, P. and Oja, H. (1998). Operating transformation and re-transformation on spatial median and angle test. *Statistica Sinica.* Volume 8, pp. 767 - 784.

[16] Chan, Y. and Hall, P. (2009). Scale adjustments for classifiers in high-dimensional, low sample size settings. *Biometrika.* Volume 96, Issue 2, pp. 469 - 478.

[17] Chang, P. C. and Afifi, A. A. (1974). Classification based on dichotomous and continuous variables. *Journal of American Statistical Association.* Volume 69, Number 346, pp. 336 - 339.

[18] Charytanowicz, M., Niewczas, J., Kulczycki, P., Kowalski, P.A., Lukasik, S. and Zak, S. (2012). UCI Machine Learning Repository [http://mlr.cs.umass.edu/ml/datasets/seeds]. Irvine, CA: University of California, School of Information and Computer Science.

[19] Chaudhuri, P. (1996). On a geometric notion of multivariate data. *Journal of American Statistical Association.* Volume 91, Number 434, pp. 862 - 872.

[20] Claeskens G., Hubert M., Slaets, L. and Vakili, K. (2014). Multivariate functional halfspace depth. *Journal of the American Statistical Association.* Volume 109, Number. 505, pp. 411 - 423.

[21] Coomans, D. and Massart, D. L. (1982). Alternative k-nearest neighbour rules in supervised pattern recognition : Part 1. k-Nearest neighbour classification by using alternative voting rules. *Analytica Chimica Acta.* Volume 136, pp. 15 - 27.

[22] Cortes, C. and Vapnik, V. N. (1995). Support-vector networks. *Machine Learning.* Volume 20, Issue 3, pp 273 - 297.

[23] Cover, T. M. (1968). Rates of convergence for nearest neighbor procedures. *Proc. Hawaii Int'l Conf. Systems Sciences.* Western Periodicals, Honolulu. pp. 413 - 415.

[24] Cover, T. M. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory.* Volume 13, Issue 1, pp. 21 - 27.

[25] Cox, L. H., Johnson, M. M. and Kafadar, K. (1982). Exposition of statistical graphics technology. *ASA Proceedings of the Statistical Computation Section.* Pp 55-56.

[26] Cuesta-Albertos, J. A. and Nieto-Reyes, A. (2008). The random Tukey depth. *Computational Statistics and Data Analysis.* Volume 52, pp. 4979 - 4988.

[27] Cuesta-Albertos, J. A. and Nieto-Reyes, A. (2010). Functional classification and the random Tukey depth. Practical issues. *Advances in Intelligent and Soft Computing - Combining Soft Computing and Statistical Methods in Data Analysis.* Volume 77, pp. 123 - 130.

[28] Cuevas, A. (2014). A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference.* Volume 147, pp. 1 - 23.

[29] Cuevas, A. and Fraiman, R. (2009). On depth measures and dual statistics. A methodology for dealing with general data. *Journal of Multivariate Analysis.* Volume 100, pp. 753 - 766.

[30] Cuevas A., Febrero, M. and Fraiman, R. (2007). Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics.* Volume 22, Issue 3, pp. 481 - 496.

[31] Cui X., Lin, L. and Yang, G. R. (2008). An extended projection data depth and its applications to discrimination. *Communications in Statistics-Theory and Methods.* Volume 37, Issue 14, pp. 2276 - 2290.

[32] Das Gupta, S. (1972). Probability inequalities and error in classification. University of Minnesota, School of Statistics, Technical report, No 190.

[33] Deheuvels, P. and Martynov, G. V. (2008). A KarhunenLoeve decomposition of a Gaussian process generated by independent pairs of exponential random variables. *Journal of functional analysis.* Volume 255, pp. 2363 - 2394.

[34] Delaigle, A. and Hall, P. (2012a). Achieving near perfect classification for functional data. *Journal of the Royal Statistical Society: Series B.* Volume74, part 2, pp. 267 - 286.

[35] Delaigle, A. and Hall, P. (2012b). Componentwise classification and clustering of functional data. *Biometrika.* Volume 99, Issue 2, pp. 299 - 313.

[36] Devroye, L., Györfi, L. and Lugosi, G. (1996). A probabilistic theory of pattern recognition. Springer-Verlag New York, Inc. First Edition.

[37] Di Pillo, P. J. (1976). The application of bias to discriminant analysis. *Communications in Statistics - Theory and Methods.* Volume 5, Issue 9, pp. 843 - 854.

[38] Donoho, D. L. (1982). Breakdown Properties of Multivariate Location Estimators. Ph.D qualifying paper, Harvard University.

[39] Dudoit S., Fridlyand, J. and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of American Statistical Association.* Volume 97, Number 457, pp. 77 - 87.

[40] Dunn, O. J. (1971). Some expected values for probability of correct classification in discriminant analysis. *Technometrics.* Volume 13, Number 2, pp. 345 - 353.

[41] Durrett, R. (2010). Probability: Theory and examples. Cambridge University Press, USA. Fourth Edition

[42] Dutta, S. and Ghosh, A. K. (2012a). On robust classification using projection depth. *Annals of the Institute of Statistical Mathematics.* Volume 64, Issue 3, pp. 657 - 676.

[43] Dutta, S. and Ghosh, A. K. (2012b). On classification based on L_p depth with an adaptive choice of p. *Technical Report No. R5/2011, Statistics and Mathematics Unit. Indian Statistical Institute, Kolkata, India.* Volume 737, pp. 657 - 676.

[44] Epifanio, L. I. (2008). Shape descriptors for classification of functional data. *Technometrics.* Volume 50, Issue 3, pp. 284 - 294.

[45] Fan, J. and Fan, Y. (2008). High dimensional classification using features annealed independence rules. *The Annals of Statistics*, Volume 36, Number 6, pp. 2605 - 2637.

[46] Fan J., Fan, Y. and Wu, Y. (2011). High-dimensional classification. In high-dimensional data analysis (Cai, T.T. and Shen, X., eds.), World Scientific, New Jersey. Pp. 3 - 37.

[47] Fan J., Feng, Y. and Tong, X. (2012). A road to classification in high dimensional space: the regularized optimal affine discriminant. *Journal of the Royal Statistical Society: Series B.* Volume 74, part 4, pp. 745 - 771.

[48] Ferraty, F. and Vieu, P. (2003). Curves discrimination: A nonparametric functional approach. *Computational Statistics & Data Analysis.* Volume 44, pp. 161 - 173.

[49] Ferraty, F. and Vieu, P. (2006). Nonparametric functional data analysis: Theory and Practice. Springer. USA.

[50] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, Volume 7, Part II, 179 - 188.

[51] Fraiman, R., Meloche, J. (1999). Multivariate L - estimation. *Test.* Volume 8, Number 2, pp. 1 - 62.

[52] Fraiman, R. and Muniz, G. (2001). Trimmed means for functional data. *Test.* Volume 10, Number 2, pp. 419 - 440.

[53] Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association.* Volume 84, Number. 405, pp. 165-175.

[54] Fromont, M. and Tuleau, C. (2006). Functional classification with margin conditions. *Learning Theory.* Volume 4005, pp. 94 - 108.

[55] Ghosh, A. K. and Chaudhuri, P. (2005a). On data depth and distribution-free discriminant analysis using separating surfaces. *Bernoulli.* Volume 11, Number 1, pp. 1 - 27.

[56] Ghosh, A. K. and Chaudhuri, P. (2005b). On maximum depth and related classifiers. *Scandinavian Journal of Statistics*, Volume 32, Number 2, pp. 327 - 350.

[57] Gilbert, E. S. (1969). The effect of unequal variance-covariance matrices on Fisher's linear discriminant function, *Biometrics*, Volume 25, Number 3, pp. 505 - 515.

[58] Glendinning, R. H. and Herbert, R. A. (2003). Shape classification using smooth principal components. *Pattern Recognition Letters.* Volume 24, Issue 12, pp. 2021 - 2030.

[59] Guha, P. (2012). On scale-scale curves for multivariate data based on rank regions. Ph.D. thesis, University of Birmingham.

[60] Guha, P. and Chakraborty, B. (2013). On scale-scale plot for comparing multivariate distributions. Accepted.

[61] Guo, Y., Hastie, T. and Tibshirani, R. (2007). Regularized discriminant analysis and its application in micro-arrays. *Biostatistics.* Volume 8, Issue 1, pp. 1 - 18.

[62] Guyon I., Weston J., Barnhill, S. and Vapnik, V. (2002). Gene selection for cancer classification using SVM. *Machine Learning.* Volume 46, Issues 1 - 3, pp. 389 - 422.

[63] Haberman, S. J. (1976). Generalized residuals for log-linear models. *Proceedings of the 9th International Biometrics Conference. Boston.* Pp. 104 - 122.

[64] Hall, P., Poskitt, D. and Presnell, B. (2001). A functional data-analytic approach to signal discrimination. *Technometrics.* Volume 43, Number 1, pp. 1 - 9.

[65] Hall P. and Pham, T. (2010). Optimal properties of centroid-based classifiers for very high-dimensional data. *The Annals of Statistics.* Volume 38, Number 2, pp. 1071 - 1093.

[66] Hall P., Titterington, D.M. and Xue, J. (2009). Median based classifiers for high dimensional data. *Journal of the American Statistical Association.* Volume 104, Number 488, pp. 1597 - 1608.

[67] Hastie T., Buja, A. and Tibshirani, R. (1995). Penalized discriminant analysis. *The Annals of Statistics.* Volume 23, Number 1, pp. 73 - 102.

[68] Hastie T., Tibshirani, R. and Friedman, J. (2001). The elements of statistical learning. Springer, NY.

[69] Hennig, C. and Viroli, C. (2013). Quantile-based classifiers. *Statistical Methodology.*

[70] Hills, M. (1967). Discrimination and allocation with discrete data. *Journal of the Royal Statistical Society: Series C.* Volume 16, Number 3, pp. 237 - 250.

[71] Hong, Z. Q. and Yang, J. Y. (1991). UCI Machine Learning Repository [mlr.cs.umass.edu/ml/datasets/Lung+Cancer]. Irvine, CA: University of California, School of Information and Computer Science.

[72] Hubert, M. and Van Driessen, K. (2004). Fast and robust discriminant analysis. *Computational Statistics and Data Analysis*. Volume 45, Issue 2, pp. 301 - 320.

[73] James, G. M. and Hastie, T. J. (2001). Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society: Series B*. Volume 63, part 3, pp. 533 - 550.

[74] Johnson, R. A. and Wichern, D. W. (2007). Applied multivariate statistical analysis. Sixth edition, Pearson Prentice Hall inc. New Jersey.

[75] Jörnsten, R. (2004). Clustering and classification based on the L_1 data depth. *Journal of Multivariate Analysis*. Volume 90, Issue 1, pp. 67 - 89.

[76] Kac, M. and Siegert, A. J. F. (1947). An explicit representation of a stationary Gaussian process. *The Annals of Mathematical Statistics*. Volume 18, Number 3, pp. 438 - 442.

[77] Kemperman, J. H. B. (1987). The median of a finite measure on a Banach space. In Statistical data analysis based on the L_1-norm and related methods (Y. Dodge (Ed.)) Amsterdam: North-Holland. Pp. 217 - 230.

[78] Kiefer, J. C. (1961). On large deviations of the empirical D. F. of vector chance variables and a law of iterated logarithm. *Pacific Journal of Mathematics*. Volume 11, Number 2, pp. 649 - 660.

[79] Kim, K. S., Choi, H. H., Moon, C. S. and Mun, C. W. (2011). Comparison of k-nearest neighbor, quadratic discriminant and linear discriminant analysis in classification of electromyogram signals based on the wrist-motion directions. *Current Applied Physics*. Volume 11, pp. 740 - 745.

[80] Koltchinskii, V. (1997). M-estimator, convexity and quantiles. *The Annals of Statistics.* Volume 25, Number 2, pp. 435 - 477.

[81] Kotz S., Kozubowski T. and Podgorski K. (2001). The Laplace distribution and generalizations: A Revisit with Applications to Communications, Economics, Engineering, and Finance.

[82] Krzanowski, W. J. (1977). The performance of fisher's linear discriminant function under non-optimal conditions. *Technometrics.* Volume 19, Number 2, pp. 191 - 200.

[83] Kushmerick, N. (1998). UCI Machine Learning Repository [http://mlr.cs.umass.edu/ml/datasets/Internet+Advertisements]. Irvine, CA: University of California, School of Information and Computer Science.

[84] Lange T., Mosler, K. and Mozharovskyi, P. (2014). Fast nonparametric classification based on data depth. *Statistical Papers.* Volume 55, Issue 1, pp 49 - 69.

[85] Leng, X. and Müller, H. G. (2006). Classification using functional data analysis for temporal gene expression data. *Bioinformatics*, Volume 22, Number 1, pp. 68 - 76.

[86] Li, B. and Yu, Q. (2008). Classification of functional data: A segmentation approach. *Computational Statistics and Data Analysis*, Volume 52, Issue 10, pp. 4790 - 4800.

[87] Li, J., Cuesta-Albertos, J. A. and Liu, R. Y. (2012). DD-Classifier: Nonparametric classification procedure based on DD-plot. *Journal of American Statistical Association.* Volume 107, Issue 498, pp. 737 - 753.

[88] Liu, R. Y. (1990). On a notion of data depth based on random simplices. *The Annals of Statistics.* Volume 18, Number 1, pp. 405 - 414.

[89] Liu, R. and Singh, K. (1993). A quality index based on data depth and multivariate rank tests. *Journal of American Statistical Association.* Volume 88, Number 421, pp. 252 - 260.

[90] Liu, R. Y., Parelius, J. M. and Singh, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *The Annals of Statistics.* Volume 27, Number 3, pp. 783 - 858.

[91] Lohweg, V. (2013). UCI Machine Learning Repository [http://mlr.cs.umass.edu/ml/datasets/banknote+authentication#]. Irvine, CA: University of California, School of Information and Computer Science.

[92] López-Pintado, S. and Romo, J (2006). Depth based classification of functional data. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science. Data Depth: Robust Multivariate Analysis, Computational Geometry and Appliations. American Mathematical Society.* Volume 72, pp. 103 - 120.

[93] Lopuhaa, H. P. and Rousseeuw, P. J. (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *The Annals of Statistics.* Volume 19, Number 1, pp. 229 - 248

[94] Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Science of India* . Volume 2, Number 1, pp. 49 - 55.

[95] Mahé, P. and Veyrieras, J. (2013). UCI Machine Learning Repository [http://mlr.cs.umass.edu/ml/datasets/MicroMass]. Irvine, CA: University of California, School of Information and Computer Science.

[96] Memon, A. Z. and Okamoto, M. (1971). Asymptotic expansion of distribution of Z-statistic in discriminant analysis. *Journal of Multivariate Analysis.* Volume 1, Issue 3, pp. 294 - 307.

[97] Mercer, J. (1909). Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society A*. Volume 209, pp. 415 - 446.

[98] Miller, A.J., Shaw, D.E., Veitch, L.G. and Smith, E.J. (1979). Analyzing the results of a cloud-seeding experiment in Tasmania. *Communications in Statistics - Theory & Methods*. Volume 8, Issue 10, pp. 1017 - 1047.

[99] Möttönen, J. and Oja, H. (1995). Multivariate spatial sign and rank methods. *Journal of Nonparametric Statistics*. Volume 5, Issue 2, pp. 201 - 213.

[100] Nakai, K. (1991). UCI Machine Learning Repository [http://mlr.cs.umass.edu/ml/datasets/Yeast]. Irvine, CA: University of California, School of Information and Computer Science.

[101] Oja, H. (1983). Descriptive statistics for multivariate distributions. *Statistics and Probability Letters*. Volume 1, Issue 6, pp. 327 - 332.

[102] Oja, H. (1999). Affine invariant multivariate sign and rank tests and corresponding estimates: A review. *Scandinavian Journal of Statistics*. Volume 26, Number 3, pp. 319 - 343.

[103] Preda C., Saporta, G. and Leveder, C. (2007). PLS classification of functional data. *Computational Statistics*. Volume 22, Issue 2, pp. 223 - 235.

[104] Ramsay, J. O. and Silverman, B.W. (2005). Functional data analysis. Second edition, New York, Springer.

[105] Rossi, F. and Villa, N., (2006). Support vector machine for functional data classification. *Neurocomputing*. Volume 69, Issues 7 - 9, pp. 730 - 742.

[106] Rousseeuw, P. J. and Hubert, M. (1999). Regression depth. *Journal of the American Statistical Association.* Volume 94, Number 446, pp. 388-402.

[107] Rousseeuw, P. J. and Leroy, A. M. (1987). Robust regression and outlier detection. Wiley.

[108] Rousseeuw, P. J. and Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics.* Volume 41, Number 3, pp. 212 - 223.

[109] Rudin, W. (1991). Functional analysis. McGraw-Hill Inc.

[110] Serfling, R. (2002). A depth function and a scale curve based on spatial quantiles. *Statistical Data Analysis Based On the L1-Norm and Related Methods.(ed. Y. Dodge). Birkhaeuser.* Pp. 25 - 38.

[111] Serfling, R. (2004). Nonparametric multivariate descriptive measures based on spatial quantiles. *Journal of Statistical Planning and Inference.* Volume 123, Issue 2, pp. 259 - 278.

[112] Serfling, R. (2006a). Multivariate symmetry and asymmetry, *In Encyclopedia of Statistical Sciences*, (S. Kotz, N. Balakrishnan, C. B. Read and B. Vidakovic, eds.), Vol. 8, pp. 5338 - 5345. Second Edition, Wiley.

[113] Serfling, R. (2006b) Depth functions in nonparametric multivariate inference. In Data Depth: Robust Multivariate Analysis, *Computational Geometry and Applications* (R. Y. Liu, R. Serfling, D. L. Souvaine, eds.), American Mathematical Society. DIMACS Series in Discrete Mathematics and Theoretical Computer Science, Volume 72, pp. 1-16.

[114] Sguera C., Galeano, P. and Lillo, R. (2014). Spatial depth based classification for functional data. *Test.*

[115] Shin, H. (2008) . An extension of Fisher's discriminant analysis for stochastic processes. *Journal of Multivariate Analysis.* Volume 99, Issue 6, pp. 1191 - 1216.

[116] Silverman, B. W. (1986) Density estimation. London: Chapman and Hall.

[117] Sitgreaves, R. (1961). Some results on the distribution of the W-classification Statistic. In: Studies in Item Analysis and Prediction (H. Solomon, ed.). Stanford University Press, Stanford, California (pp. 241 - 251).

[118] Tian T. S., James, G. M. and Wilcox, R. R. (2010). A multivariate adaptive stochastic search method for dimension reduction in classification. *The Annals of Applied Statistics.* Volume 4, Number 1, pp. 340 - 365.

[119] Tibshirani R., Hastie T., Narasimhan, B. and Chu, G. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences, USA.* Volume 99, Number 10, pp. 6567 - 6572.

[120] Tsanas A., Little M. A., Fox, C. and Ramig, L. O. (2014). Objective automatic assessment of rehabilitative speech treatment in Parkinsons disease. *IEEE Transactions on Neural Systems and Rehabilitation Engineering.* Volume 22, Number 1, pp. 181 - 190.

[121] Tukey, J. (1975). Mathematics and picturing data. *Proceedings of the 1975 International Congress of Mathematics.* Volume 2, pp. 523 - 531.

[122] Vapnik, V. N. (1982). Estimation of dependences based on empirical data. Addendum 1, New York: Springer-Verlag.

[123] Vapnik V. N. (1998). Statistical learning theory. John Wiley and Sons, New York.

[124] Vardi, Y. and Zhang, C. H. (2000). The multivariate L_1-median and associated data depth. *Proceedings of the National Academy of Sciences, USA*. Volume 97, Number 4, pp. 1423 - 1426.

[125] Vilar, J. A. and Pértega, S (2004). Discriminant and cluster analysis for Gaussian stationary processes: local linear fitting approach. Journal of Nonparametric Statistics. Volume 16, Issue 3-4, pp. 443 - 462.

[126] Wald, A. (1944). On a statistical problem arising in the classification of an individual into one of two groups. *The Annals of Mathematical Statistics.* Volume 15, Number 2, pp. 145 - 162.

[127] Wang, J. and Serfling, R. (2004). On scale curves for nonparametric description of dispersion. *Proceedings of May 2003 DIMACS Workshop on Data Depth: Robust Multivariate Analysis, Computational Geometry, and Applications.*

[128] Welch, B. L. (1939). Note on discriminant functions. *Biometrika.* Volume 31, Number 1/2, pp. 218 - 220.

[129] Witten, D. M. and Tibshirani, R. (2011). Penalized classification using Fisher's linear discriminant. *Journal of the Royal Statistical Society: Series B.* Volume 73, Part 5, pp. 753 - 772.

[130] Zhu H., Brown, P. J. and Morris, J. S. (2012). Robust classification of functional and quantitative image data using functional mixed models. *Biometrics.* Volume 68, Number 4, pp. 1260-1268.

[131] Zuo, Y. and Serfling, R. (2000a). General notions of statistical depth function. *The Annals of Statistics.* Volume 28, Number 2, pp. 461 - 482.

[132] Zuo, Y. and Serfling, R. (2000b). Structural properties and convergence results for contours of sample statistical depth functions. *The Annals of Statistics*. Volume 28, Number 2, pp. 483 - 499.