uOttawa

L'Université canadienne
Canada's university

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES

uOttawa

L'Université canadienne
Canada's university

FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

Danielle Léger

AUTEUR DE LA THÈSE / AUTHOR OF THESIS

M.Sc. (Mathematics and Statistics)

GRADE / DEGREE

Department of Mathematics and Statistics

FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

A Study of Classification Techniques

TITRE DE LA THÈSE / TITLE OF THESIS

M. Alvo

DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

CO-DIRECTEUR (CO-DIRECTRICE) DE LA THÈSE / THESIS CO-SUPERVISOR

EXAMINATEURS (EXAMINATRICES) DE LA THÈSE / THESIS EXAMINERS

D. McDonald

M. Mojirsheibani

Gary W. Slater

Le Doyen de la Faculté des études supérieures et postdoctorales / Dean of the Faculty of Graduate and Postdoctoral Studies

A Study of Classification Techniques

Danielle Léger

Thesis Submitted to the Faculty of Graduate and Postdoctoral Studies
In partial fulfilment of the requirements for the degree of Master of Science in Mathematics [1]

Department of Mathematics and Statistics
Faculty of Science
University of Ottawa

[1]The M.Sc. program is a joint program with Carleton University, administered by the Ottawa-Carleton Institute of Mathematics and Statistics

Your file Votre référence
ISBN: 978-0-494-59472-8
Our file Notre référence
ISBN: 978-0-494-59472-8

Abstract

The following thesis studies both parametric and nonparametric approaches to classification. Among the various methods which exist, three are developed in detail. They are the Bayes rule, classification based on kernel density estimation and Fisher's discriminant function applied to the ranked data. Furthermore, we propose a new classification rule based on ranks. A Monte Carlo simulation study is then performed to test this new method and compare it with the other three classification techniques. The simulations indicate that the new classification rule performs well in many cases and that it is most effective when the number of dimensions are high and few observations are available. In this particular situation, the new classification rule proposed had the lowest probability of misclassification.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr. Mayer Alvo, for his guidance and support throughout the past two years. Thank you for sharing your time, ideas and advice. It is greatly appreciated and I am truly grateful for all of your help.

Dedication

To my Mom and Dad for their unconditional love, understanding and support. Thank you for your guidance, your wisdom and your constant encouragement in all my endeavors. You are wonderful parents and I love you very much.

Contents

List of Tables

Chapter 1

Introduction

Classification deals with the problem of allocating new observations to previously defined groups or classes. Such problems commonly arise in everyday living. Diagnosing a patient, deciding on the admission of a student to a university and assessing an individual's credit risk are all examples of situations where classification is utilized. Throughout the years, many classification procedures have been developed. This thesis will propose a new classification rule and compare it with existing methods.

We begin, in Chapter 2, by studying parametric classification. The theory of classification and how to evaluate a particular classification procedure are discussed. Further, we present the classic parametric classifier, the Bayes rule, and show that it is the optimal classification procedure. That is, it has the smallest probability of misclassifying a new observation among all methods.

In the following three chapters, we then turn our attention to nonparametric classification techniques. A broad range of such methods have been proposed in the literature. Some of these procedures seek to apply the Bayes rule using a nonparametric estimate of the density function of each population. For instance, Greblicki

and Pawlak (1981) utilize Fourier series estimators of density functions whereas Silverman (1986) suggests the use of kernel density estimation techniques. The latter is a more common approach and is discussed in detail in Chapter 3.

There also exist many nonparametric procedures which do not make use of the Bayes rule. Lee (2004) presents a classification technique based on neural networks. Fix and Hodges (1951) developed the nearest neighbor classification method which was later modified by Cover and Hart (1967). In this approach, a distance function (e.g. Euclidean distance) is used to determine the k nearest neighbors of the new observation that we wish to classify. The group or class which contains the majority of the k nearest neighbors will be the class assigned to this new observation. Finally, a classic nonparametric technique, presented in Johnson and Wichern (2002), is Fisher's discriminant function. The first section of Chapter 4 gives a summary of Fisher's discriminant analysis while the second shows how this leads to a classification rule. In the third section, we discuss the application of Fisher's classification rule to the ranks of the observations which was first proposed by Conover and Iman (1980).

The idea of using the ranks of the data instead of the observations themselves is common in many nonparametric classification methods. It is a popular approach since the ranks are robust to outliers. Thus, extreme values are less likely to affect the classification rule if ranks are used. In fact, just like Fisher's discriminant function was applied to ranks by Conover and Iman (1980), many of the other previously mentioned classification techniques have been modified to work with the ranked data instead of with the direct observations. For example, Ćwik and Mielniczuk (1995) explored the idea of using kernel density estimators constructed from samples of ranks and Das Gupta and Lin (1980) proposed a nearest neighbor approach based on ranks.

We equally propose a classification procedure based on ranks. This new method combines Fisher's discriminant function along with certain ideas from Mantel and Valand (1970). The details of our classification rule along with an example to illustrate the technique can be found in Chapter 5 of the present thesis.

In order to assess the quality of our new classification rule, a Monte Carlo simulation study was conducted. Simulations were performed for the four classification methods discussed in the thesis: the Bayes rule, classification based on kernel density estimation, Fisher's classification rule applied to the ranked data as well as our new classification procedure. More details including the results of these simulations are given in Chapter 6. Lastly, Chapter 7 provides a comparison of the various classification approaches, indicating the advantages and limitations of each method.

Chapter 2

Parametric Classification

Our study of classification techniques begins by looking at the parametric approach. This subject has been thoroughly developed in Chapter 11 of Johnson and Wichern (2002) and the present chapter is a summary of their work. All definitions, theorems and examples are taken from Johnson and Wichern (2002) unless otherwise stated.

Let $X_1, \ldots, X_{n_1}$ be independent, identically distributed observations from one population, π_1 and let $Y_1, \ldots, Y_{n_2}$ be independent, identically distributed observations from another population, π_2. These observations can be univariate or multivariate. In what follows, we will assume that they are $p-$dimensional observations. Let Z be a new observation arising from π_1 or π_2. The goal of classification is to correctly identify the true class or group membership of Z based on the information provided by the training sample: $X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2}$.

With parametric classification, we not only know the true class membership of each observation in the training sample but we also know the underlying density, $f_i(t)$, of each population $\pi_i, i = 1, 2$. Using this information, it is possible to obtain a rule for classifying a new observation. The classification rule can be expressed as a func-

tion $\varphi : A \rightarrow \{1, 2\}$ which takes any observation in $A = \{\text{domain of} f_1(z)\} \bigcup \{\text{domain of} f_2(z)\}$ and returns its class membership. Such a function is often called a classifier.

2.1 Finding a Good Classifier

Any function $\varphi : A \rightarrow \{1, 2\}$ can be a classifier. For example

$$\varphi(z) = \begin{cases} 1 & \text{if it rains today} \\ 2 & \text{if it does not rain today} \end{cases}$$

is a possible classifier. Evidently, this is a useless classification rule since it does not incorporate any information about the new observation Z or the training sample. Therefore, it is important to have a criterion to evaluate a classifier and determine if it is adequate and satisfactory.

The usual criterion used to evaluate a classification rule is the total probability of misclassification (abbreviated TPM). No classifier is perfect and thus will, on occasion, incorrectly classify an observation. Hence we would like to know the probability of this occurring for a given classifier. The classifier with minimal TPM would be the best or ideal classification rule. The total probability of misclassification is calculated as follows.

$$\begin{aligned} TPM &= P(\text{misclassifying the new observation } z) \\ &= P(\text{classifying } Z \text{ in } \pi_1 \text{when } Z \in \pi_2) + P(\text{classifying } Z \text{ in } \pi_2 \text{when } Z \in \pi_1) \\ &= P(\text{classifying } Z \text{ in } \pi_1 | Z \in \pi_2) \cdot P(Z \in \pi_2) + \\ &\quad P(\text{classifying } Z \text{ in } \pi_2 | Z \in \pi_1) \cdot P(Z \in \pi_1) \\ &= P(\pi_1|\pi_2) \cdot p_2 + P(\pi_2|\pi_1) \cdot p_1 \end{aligned}$$

Sometimes, it is also important to consider the cost of misclassification. For instance, when determining if an individual has cancer, it is a very grave mistake to misdiagnose the patient as not having the disease when in fact they do. Hence, the cost of such a misclassification is quite high. We can take these costs into account by considering the expected cost of misclassification (abbreviated ECM). In this case, the classifier with minimum ECM would be optimal. The ECM is given by

$$ECM = c(\pi_1|\pi_2) \cdot P(\pi_1|\pi_2) \cdot p_2 + c(\pi_2|\pi_1) \cdot P(\pi_2|\pi_1) \cdot p_1$$

where $c(\pi_i|\pi_j)$ is the cost attributed when an observation from π_j is classified in π_i $(i, j = 1, 2;\ c(\pi_i|\pi_i) = 0)$. We note that when $c(\pi_1|\pi_2) = c(\pi_2|\pi_1)$, we can take these costs to be 1 and in this case ECM=TPM.

As previously mentioned, we would like to find a classifier which will minimize the expected cost of misclassification. We begin with a general classifier of the form

$$\varphi(z) = \begin{cases} 1 & \text{if } z \in R_1 \\ 2 & \text{if } z \in R_2 \end{cases}$$

where R_1 and R_2 are regions satisfying $R_1 \bigcup R_2 = A = \{\text{domain of} f_1(z)\} \bigcup \{\text{domain of} f_2(z)\}$ and $R_1 \bigcap R_2 = \emptyset$. Hence, we also note that for $i = 1, 2$

$$\begin{aligned} 1 &= \int_A f_i(z)\, dz \\ &= \int_{R_1 \bigcup R_2} f_i(z)\, dz \\ &= \int_{R_1} f_i(z)\, dz + \int_{R_2} f_i(z)\, dz \end{aligned} \tag{2.1.1}$$

For this classifier, the ECM is given by:

$$\begin{aligned} ECM &= c(\pi_1|\pi_2) \cdot P(\pi_1|\pi_2) \cdot p_2 + c(\pi_2|\pi_1) \cdot P(\pi_2|\pi_1) \cdot p_1 \\ &= c_{12} \cdot p_2 \cdot P(Z \in R_1 | Z \in \pi_2) + c_{21} \cdot p_1 \cdot P(Z \in R_2 | Z \in \pi_1) \\ &= c_{12} \cdot p_2 \int_{R_1} f_2(z)\, dz + c_{21} \cdot\ p_1 \int_{R_2} f_1(z)\, dz \\ &= c_{12} \cdot p_2 \int_{R_1} f_2(z)\, dz + c_{21} \cdot\ p_1 \cdot \left[1 - \int_{R_1} f_1(z)\, dz\right] \quad \text{by (2.1.1)} \\ &= c_{21} \cdot p_1 + \int_{R_1} [c_{12} p_2 f_2(z) - c_{21} p_1 f_1(z)]\, dz \end{aligned}$$

Since $c_{21} \cdot p_1$ is a constant, minimizing the ECM is equivalent to minimizing

$$\int_{R_1} [c_{12} p_2 f_2(z) - c_{21} p_1 f_1(z)]\, dz.$$

Thus we are searching for a region R_1 that will minimize the expected cost of misclassification. Once this region has been identified, we will have the optimal classification rule.

Johnson and Wichern (2002) note that since c_{12}, c_{21}, p_1 and p_2 are nonnegative constants and $f_1(z)$ and $f_2(z)$ are nonnegative for all z, then the ECM will be minimized by taking R_1 to be the following region:

$$\begin{aligned} R_1 &= \{z : c_{12} p_2 f_2(z) - c_{21} p_1 f_1(z) \leqslant 0\} \\ &= \left\{z : \frac{f_1(z)}{f_2(z)} \geqslant \frac{c_{12} p_2}{c_{21} p_1}\right\}. \end{aligned}$$

Therefore, the classifier which minimizes the ECM (and TPM under the assumption of equal costs) is given by

$$\varphi(z) = \begin{cases} 1 & \text{if } \frac{f_1(z)}{f_2(z)} \geqslant \frac{c_{12}p_2}{c_{21}p_1} \\ 2 & \text{otherwise} \end{cases}.$$

2.2 The Bayes Rule

In what follows, we will assume the costs of misclassification are equal and can thus be taken as having a value of 1. In this particular case, the classifier which minimizes the ECM=TPM is

$$\varphi_*(z) = \begin{cases} 1 & \text{if } \frac{f_1(z)}{f_2(z)} \geqslant \frac{p_2}{p_1} \\ 2 & \text{otherwise} \end{cases}.$$

We will now show that the same classification rule can be obtained using a Bayesian approach. The typical Bayes classifier utilizes posterior probabilities when defining the classification rule and is given by

$$\varphi_B(z_0) = \begin{cases} 1 & \text{if } P(Z \in \pi_1 | Z = z_0) \geqslant P(Z \in \pi_2 | Z = z_0) \\ 2 & \text{otherwise} \end{cases}.$$

From Bayes theorem, we know that

$$\begin{aligned} P(Z \in \pi_i | Z = z_0) &= \frac{P(Z \in \pi_i) \cdot P(Z = z_0 | Z \in \pi_i)}{\sum_{j=1}^{2} P(Z \in \pi_j) \cdot P(Z = z_0 | Z \in \pi_j)} \\ &= \frac{p_i \cdot f_i(z_0)}{p_1 f_1(z_0) + p_2 f_2(z_0)}. \end{aligned} \tag{2.2.1}$$

Substituting equation (2.2.1) into $\varphi_B(z_0)$ gives

$$\begin{aligned}
\varphi_B(z_0) &= \begin{cases} 1 & \text{if } \frac{p_1 \cdot f_1(z_0)}{p_1 f_1(z_0)+p_2 f_2(z_0)} \geqslant \frac{p_2 \cdot f_2(z_0)}{p_1 f_1(z_0)+p_2 f_2(z_0)} \\ 2 & \text{otherwise} \end{cases} \\
&= \begin{cases} 1 & \text{if } \frac{f_1(z_0)}{f_2(z_0)} \geqslant \frac{p_2}{p_1} \\ 2 & \text{otherwise} \end{cases} \\
&= \varphi_*(z_0).
\end{aligned}$$

Therefore, we can conclude that the Bayes rule is the optimal classifier in terms of the total probability of misclassification. That is, it is the classifier which minimizes the TPM and hence has the lowest probability of misclassification among all classifiers.

2.3 Examples of Optimal Classifiers

In this section, the optimal classifier, $\varphi_B(z)$, will be calculated for two examples in which the population densities are both multivariate normal. We selected these particular problems because considerable simplifications can be made when calculating $\frac{f_1(z)}{f_2(z)}$ and hence an easier and more direct classification rule can be obtained. We would like to note that such simplifications are generally not possible. Most often, the expression $\frac{f_1(z)}{f_2(z)}$ cannot be further simplified when at least one of the populations has a density which is not normal (e.g. Cauchy, exponential, logistic, etc). In these cases, we simply use the optimal rule as stated in terms of the density functions, $f_i(z)$, $i = 1, 2$.

Example 2.3.1 Suppose $X_1, \ldots, X_{n_1} \sim N_p(\mu_1, \Sigma)$ represents the observations of the first population, π_1 and $Y_1, \ldots, Y_{n_2} \sim N_p(\mu_2, \Sigma)$ represents the observations of the second population, π_2, where $N_p(\mu, \Sigma)$ designates a $p-$dimensional multivariate normal distribution with mean μ and covariance matrix Σ. Let p_i be the probability that a new observation Z comes from $\pi_i, i = 1, 2$. Then we can calculate the optimal classification rule, $\varphi_B(z)$ for this particular case. First, we note that

$$\begin{aligned} f_1(z) &= \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \cdot \exp\left[-\frac{1}{2} \cdot (z-\mu_1)'\Sigma^{-1}(z-\mu_1)\right] \text{ and} \\ f_2(z) &= \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \cdot \exp\left[-\frac{1}{2} \cdot (z-\mu_2)'\Sigma^{-1}(z-\mu_2)\right]. \end{aligned}$$

Therefore,

$$\begin{aligned} \frac{f_1(z)}{f_2(z)} &= \frac{\exp\left[-\frac{1}{2} \cdot (z-\mu_1)'\Sigma^{-1}(z-\mu_1)\right]}{\exp\left[-\frac{1}{2} \cdot (z-\mu_2)'\Sigma^{-1}(z-\mu_2)\right]} \\ &= \exp\left[-\frac{1}{2} \cdot (z-\mu_1)'\Sigma^{-1}(z-\mu_1) + \frac{1}{2} \cdot (z-\mu_2)'\Sigma^{-1}(z-\mu_2)\right] \quad (2.3.1) \end{aligned}$$

Substituting equation (2.3.1) in $\varphi_B(z)$ and applying the natural logarithm, we obtain

$$\begin{aligned} \varphi_B(z) &= \begin{cases} 1 & \text{if } -\frac{1}{2} \cdot (z-\mu_1)'\Sigma^{-1}(z-\mu_1) + \frac{1}{2} \cdot (z-\mu_2)'\Sigma^{-1}(z-\mu_2) \geqslant \ln\left(\frac{p_2}{p_1}\right) \\ 2 & \text{otherwise} \end{cases} \\ &= \begin{cases} 1 & \text{if } (\mu_1-\mu_2)'\Sigma^{-1}z - \frac{1}{2}(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1+\mu_2) \geqslant \ln\left(\frac{p_2}{p_1}\right) \\ 2 & \text{otherwise} \end{cases} \end{aligned}$$

In the special case where $p_1 = p_2 = 1/2$, the optimal classifier is

$$\varphi_B(z) = \begin{cases} 1 & \text{if } (\mu_1-\mu_2)'\Sigma^{-1}z - \frac{1}{2}(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1+\mu_2) \geqslant 0 \\ 2 & \text{otherwise} \end{cases}.$$

Furthermore, in this particular case, it is also possible to derive an expression for the total probability of misclassification, TPM, of this optimal classification rule. Recall that the TPM is calculated as

$$\begin{aligned} TPM &= p_1 \cdot P(\pi_2|\pi_1) + p_2 \cdot P(\pi_1|\pi_2) \\ &= \frac{1}{2} \cdot P(\pi_2|\pi_1) + \frac{1}{2} \cdot P(\pi_1|\pi_2) \end{aligned}$$

Let's study each probability separately. First,

$$\begin{aligned} P(\pi_2|\pi_1) &= P(\text{classifying } Z \text{ in } \pi_2 | Z \in \pi_1) \\ &= P\left((\mu_1 - \mu_2)'\Sigma^{-1}Z - \frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2) < 0 | Z \in \pi_1\right) \\ &= P\left((\mu_1 - \mu_2)'\Sigma^{-1}Z < \frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2) | Z \in \pi_1\right) \\ &= P\left(Y < \frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2) | Z \in \pi_1\right) \end{aligned}$$

where $Y = (\mu_1 - \mu_2)'\Sigma^{-1}Z$. If $Z \in \pi_1$, then $Z \sim N_p(\mu_1, \Sigma)$. This implies that the linear combination $Y = (\mu_1 - \mu_2)'\Sigma^{-1}Z$ is normally distributed with mean $\mu_{y1} = (\mu_1 - \mu_2)'\Sigma^{-1}\mu_1$ and variance $\sigma_y^2 = (\mu_1 - \mu_2)'\Sigma^{-1}\Sigma\Sigma^{-1}(\mu_1 - \mu_2) = (\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2)$. Moreover,

$$\begin{aligned} P(\pi_2|\pi_1) &= P\left(Y < \frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2) | Z \in \pi_1\right) \\ &= P\left(\frac{Y - \mu_{y1}}{\sigma_y} < \frac{\frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2) - \mu_{y1}}{\sigma_y}\right) \\ &= P\left(\frac{Y - \mu_{y1}}{\sigma_y} < -\frac{\sigma_y}{2}\right) \\ &= \Phi\left(-\frac{\sigma_y}{2}\right) \end{aligned}$$

where $\Phi(z)$ is the cumulative distribution function of a standard normal distribution.

Similarly,

$$\begin{aligned} P(\pi_1|\pi_2) &= P(\text{classifying } Z \text{ in } \pi_1 | Z \in \pi_2) \\ &= P\left((\mu_1-\mu_2)'\Sigma^{-1}Z - \frac{1}{2}(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1+\mu_2) \geqslant 0 | Z \in \pi_2\right) \\ &= P\left((\mu_1-\mu_2)'\Sigma^{-1}Z \geqslant \frac{1}{2}(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1+\mu_2) | Z \in \pi_2\right) \\ &= P\left(Y \geqslant \frac{1}{2}(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1+\mu_2) | Z \in \pi_2\right) \end{aligned}$$

Now, when $Z \in \pi_2$, then $Z \sim N_p(\mu_2, \Sigma)$ and so the linear combination $Y = (\mu_1 - \mu_2)'\Sigma^{-1}Z$ is normally distributed with mean $\mu_{y2} = (\mu_1-\mu_2)'\Sigma^{-1}\mu_2$ and variance $\sigma_y^2 = (\mu_1-\mu_2)'\Sigma^{-1}\Sigma\Sigma^{-1}(\mu_1-\mu_2) = (\mu_1-\mu_2)'\Sigma^{-1}(\mu_1-\mu_2)$. This implies that

$$\begin{aligned} P(\pi_1|\pi_2) &= P\left(Y \geqslant \frac{1}{2}(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1+\mu_2) | Z \in \pi_2\right) \\ &= P\left(\frac{Y-\mu_{y2}}{\sigma_y} \geqslant \frac{\frac{1}{2}(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1+\mu_2) - \mu_{y2}}{\sigma_y}\right) \\ &= P\left(\frac{Y-\mu_{y2}}{\sigma_y} \geqslant \frac{\sigma_y}{2}\right) \\ &= 1 - P\left(\frac{Y-\mu_{y2}}{\sigma_y} < \frac{\sigma_y}{2}\right) \\ &= 1 - \Phi\left(\frac{\sigma_y}{2}\right) \\ &= \Phi\left(-\frac{\sigma_y}{2}\right) \end{aligned}$$

Hence, the total probability of misclassification of the optimal classifier is given by

$$\begin{aligned} TPM &= \frac{1}{2}\cdot P(\pi_2|\pi_1) + \frac{1}{2}\cdot P(\pi_1|\pi_2) \\ &= \frac{1}{2}\cdot\Phi\left(-\frac{\sigma_y}{2}\right) + \frac{1}{2}\cdot\Phi\left(-\frac{\sigma_y}{2}\right) \\ &= \Phi\left(-\frac{\sigma_y}{2}\right) \end{aligned}$$

This is the lowest probability of misclassification possible when both populations are normally distributed with equal covariance matrices.

Example 2.3.2 Suppose $X_1, \ldots, X_{n_1} \sim N_p(\mu_1, \Sigma_1)$ represents the observations of the first population, π_1 and $Y_1, \ldots, Y_{n_2} \sim N_p(\mu_2, \Sigma_2)$ represents the observations of the second population, π_2, where $N_p(\mu, \Sigma)$ designates a $p-$dimensional multivariate normal distribution with mean μ and covariance matrix Σ. Let p_i be the probability that a new observation Z comes from $\pi_i, i = 1, 2$. This is very similar to Example (2.3.1). However, in this case, the populations have different covariance matrices. We will calculate the optimal classification rule, $\varphi_B(z)$ for this problem. To begin,

$$\begin{aligned} f_1(z) &= \frac{1}{(2\pi)^{p/2}|\Sigma_1|^{1/2}} \cdot \exp\left[-\frac{1}{2} \cdot (z-\mu_1)'\Sigma_1^{-1}(z-\mu_1)\right] \text{ and} \\ f_2(z) &= \frac{1}{(2\pi)^{p/2}|\Sigma_2|^{1/2}} \cdot \exp\left[-\frac{1}{2} \cdot (z-\mu_2)'\Sigma_2^{-1}(z-\mu_2)\right] \end{aligned}$$

and so,

$$\begin{aligned} \frac{f_1(z)}{f_2(z)} &= \frac{|\Sigma_2|^{1/2} \exp\left[-\frac{1}{2} \cdot (z-\mu_1)'\Sigma_1^{-1}(z-\mu_1)\right]}{|\Sigma_1|^{1/2} \exp\left[-\frac{1}{2} \cdot (z-\mu_2)'\Sigma_2^{-1}(z-\mu_2)\right]} \\ &= \left(\frac{|\Sigma_2|}{|\Sigma_1|}\right)^{1/2} \exp\left\{-\frac{1}{2} \cdot (z-\mu_1)'\Sigma_1^{-1}(z-\mu_1) + \right. \\ &\qquad \left. \frac{1}{2}(z-\mu_2)'\Sigma_2^{-1}(z-\mu_2)\right\}. \end{aligned} \tag{2.3.2}$$

Substituting equation (2.3.2) in $\varphi_B(z)$ and applying the natural logarithm, we obtain

$$\begin{aligned} \varphi_B(z) &= \begin{cases} 1 & \text{if } \frac{1}{2}\ln\left(\frac{|\Sigma_2|}{|\Sigma_1|}\right) - \frac{1}{2} \cdot (z-\mu_1)'\Sigma_1^{-1}(z-\mu_1) + \\ & \frac{1}{2} \cdot (z-\mu_2)'\Sigma_2^{-1}(z-\mu_2) \geqslant \ln\left(\frac{p_2}{p_1}\right) \\ 2 & \text{otherwise} \end{cases} \\ &= \begin{cases} 1 & \text{if } \frac{1}{2}\ln\left(\frac{|\Sigma_2|}{|\Sigma_1|}\right) - \frac{1}{2}z'(\Sigma_1^{-1} - \Sigma_2^{-1})z + (\mu_1'\Sigma_1^{-1} - \mu_2'\Sigma_2^{-1})z + \\ & \frac{1}{2}(\mu_1'\Sigma_1^{-1}\mu_1 - \mu_2'\Sigma_2^{-1}\mu_2) \geqslant \ln\left(\frac{p_2}{p_1}\right) \\ 2 & \text{otherwise} \end{cases} \end{aligned}$$

$$= \begin{cases} 1 & \text{if } -\frac{1}{2}z'(\Sigma_1^{-1} - \Sigma_2^{-1})z + (\mu_1'\Sigma_1^{-1} - \mu_2'\Sigma_2^{-1})z + k \geqslant \ln\left(\frac{p_2}{p_1}\right) \\ 2 & \text{otherwise} \end{cases}$$

where $k = \frac{1}{2}\ln\left(\frac{|\Sigma_2|}{|\Sigma_1|}\right) + \frac{1}{2}(\mu_1'\Sigma_1^{-1}\mu_1 - \mu_2'\Sigma_2^{-1}\mu_2)$.

This is the optimal classification rule when both populations have normal distributions with unequal covariance matrices. Unfortunately, for this example, the presence of the quadratic term, $z'(\Sigma_1^{-1} - \Sigma_2^{-1})z$, prohibits us from finding a simple and direct expression for the calculation of the total probability of misclassification. This problem can also occur when either one or both population densities are not normal (e.g. Cauchy, exponential, logistic, etc) or when using a different classifier. Although no general expression can be obtained in such instances, the TPM can sometimes be calculated for specific problems.

Table 2.1 provides values of the lowest probability of misclassification for several particular examples. In this table, it is assumed that $p_1 = p_2$ and the following notation is utilized: $N_p(\mu, \Sigma)$ is a $p-$dimensional normal distribution with mean $= \mu$ and covariance matrix $= \Sigma$, $C(\mu, \sigma)$ is a Cauchy distribution with location $= \mu$ and scale $= \sigma$, $Log(\mu, \sigma)$ is a logistic distribution with location $= \mu$ and scale $= \sigma$ and $Exp(\mu, \sigma)$ is an exponential distribution with location $= \mu$ and scale $= \sigma$. Unfortunately, there are still many classification problems for which the calculations performed in Table 2.1 cannot be done. We can, however, use simulations along with cross-validation or bootstrap methods to estimate the total probability of misclassification for those cases.

Finally, it is important to note that the Bayes rule is the optimal classifier provided that both population densities are known as well as all density parameters. In

practice, the density parameters are rarely known. They can however be estimated using maximum likelihood. These estimates can then be substituted in the classification rule. For Example 2.3.1, the usual estimators of μ_1, μ_2 and Σ are:

$$\widehat{\mu}_1 = \overline{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i;\ \widehat{\mu}_2 = \overline{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i$$

$$\widehat{\Sigma} = S_p = \frac{(n_1 - 1)S_x + (n_2 - 1)S_y}{n_1 + n_2 - 2}$$

where $(n_1 - 1)S_x = \sum_{i=1}^{n_1}(X_i - \overline{X})(X_i - \overline{X})'$ and $(n_2 - 1)S_y = \sum_{i=1}^{n_2}(Y_i - \overline{Y})(Y_i - \overline{Y})'$.

Similarly, the natural estimators of the parameters in Example 2.3.2 are:

$$\widehat{\mu}_1 = \overline{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i;\ \widehat{\mu}_2 = \overline{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i$$

$$\widehat{\Sigma}_1 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_i - \overline{X})(X_i - \overline{X})'$$

$$\widehat{\Sigma}_2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(Y_i - \overline{Y})(Y_i - \overline{Y})'.$$

The classification rule that is based on the estimates of the parameters is no longer optimal. Nevertheless, as n_1 and n_2 increase, $\widehat{\mu}_1, \widehat{\mu}_2, \widehat{\Sigma}, \widehat{\Sigma}_1$ and $\widehat{\Sigma}_2$ are more accurate and thus this classifier becomes a good approximation of the optimal rule. Hence, we can say that it is asymptotically optimal.

Table 2.1: Lowest TPM for several examples

Population 1	Population 2	Lowest TPM (%)
$N_1(0,1)$	$N_1(2,1)$	15.87
$N_1(0,1)$	$N_1(2,2)$	23.17
$C(0,1)$	$C(2,1)$	25.0
$C(0,1)$	$C(2,2)$	28.71
$Exp(0,1)$	$Exp(2,1)$	6.77
$Exp(0,1)$	$Exp(2,2)$	6.77
$Log(0,1)$	$Log(2,1)$	26.89
$Log(0,1)$	$Log(2,2)$	29.3
$N_2\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right)$	$N_2\left(\begin{bmatrix}2\\2\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right)$	7.86
$N_2\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}4 & 3\\3 & 4\end{bmatrix}\right)$	$N_2\left(\begin{bmatrix}1\\-3\end{bmatrix}, \begin{bmatrix}4 & 3\\3 & 4\end{bmatrix}\right)$	7.50
$N_2\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0.3\\0.3 & 2\end{bmatrix}\right)$	$N_2\left(\begin{bmatrix}5\\-5\end{bmatrix}, \begin{bmatrix}1 & 0.3\\0.3 & 2\end{bmatrix}\right)$	0.03

Chapter 3

Kernel Density Estimation and Classification

In most classification problems, the underlying population densities are not known. Hence, we are unable to use the Bayes rule, presented in Chapter 1, since this classification procedure requires knowledge about $f_1(z)$ and $f_2(z)$. However, we can apply a nonparametric approach to solve the classification problems.

With nonparametric classification, no assumptions are made about the population densities and no information is required about them. Instead, an estimate of the density functions can be constructed from the observed data. These estimates of the population densities can then be used as substitutes for the true, unknown, density functions in the Bayes rule. The classifier in these cases is given by

$$\varphi(z) = \begin{cases} 1 & \text{if } \frac{\hat{f}_1(z)}{\hat{f}_2(z)} \geqslant \frac{p_2}{p_1} \\ 2 & \text{otherwise} \end{cases}$$

where $\hat{f}_1$ and $\hat{f}_2$ are the estimated densities of π_1 and π_2 respectively. These estimates will be more reliable as the number of observations increase. Thus, this

classifier should also be asymptotically optimal. However, its optimality will not only depend on the number of observations available but equally upon the method utilized to obtain the estimates of the densities.

There are many techniques available to estimate a density function. Greblicki and Pawlak (1981) utilize Fourier series estimators while Silverman (1986) discusses using k nearest neighbors. Nevertheless, kernel density estimation is the most commonly applied method. It has been widely researched and has been shown to perform well in many classification problems. Silverman (1986) provides a thorough study of this subject in both univariate and multivariate cases. The present chapter is a summary of this work and in what follows, all definitions, theorems and tables are taken from Silverman (1986) unless otherwise stated.

3.1 Univariate Kernel Density Estimation

Let $X_1, \ldots, X_n$ be a random sample from a given univariate population whose true density is $f(x)$. The kernel density estimate of the population density function is given by

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{3.1.1}$$

where $K(x)$ is the kernel function and h is the bandwidth or smoothing parameter.

We see from equation (3.1.1) that the properties of $\hat{f}(x)$ depend on the choice of the kernel $K(x)$. For instance, if $\int K(x)\,dx = 1$ then $\int \hat{f}(x)\,dx = 1$, which is a necessary condition for $\hat{f}$ to be a proper density function. Likewise, if $K(x) \geqslant 0$, $\forall x$ and $K(x)$ has derivatives of all order, then $\hat{f}(x)$ will possess the same characteristics. Furthermore, equation (3.1.1) indicates that the density estimate will be influenced by the smoothing parameter. Hence, the quality and accuracy of the estimator $\hat{f}$ will depend on the selection of the kernel as well as the bandwidth.

In order to select the best kernel and smoothing parameter, a criterion must be used to evaluate the estimator $\hat{f}$ in comparison to the true density, f. The most commonly used method of measuring the overall accuracy of $\hat{f}$ is to use its mean integrated square error, MISE. We seek, the kernel function and bandwidth which minimize the MISE. The mean integrated square error is defined as

$$\begin{aligned}
\text{MISE}(\hat{f}(x)) &= E\left[\int \left\{\hat{f}(x) - f(x)\right\}^2 dx\right] \\
&= \int E\left[\left\{\hat{f}(x) - f(x)\right\}^2\right] dx \\
&= \int (\text{mean square error of } \hat{f})\,dx \\
&= \int \left\{\text{var}(\hat{f}(x)) + \left[E(\hat{f}(x) - f(x))\right]^2\right\} dx \\
&= \int \text{var}(\hat{f}(x))\,dx + \int \left[E(\hat{f}(x) - f(x))\right]^2 dx \\
&= \int \text{var}(\hat{f}(x))\,dx + \int \left[\text{ bias}(\hat{f}(x))\right]^2 dx. \qquad (3.1.2)
\end{aligned}$$

Therefore, in order to compute the MISE$(\hat{f}(x))$, we need the expectation and variance of $\hat{f}$. To find these, we note that $\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)$ is a sum of independent, identically distributed variables and so we will let Y represent a variable with the

same distribution as $X_1, \ldots, X_n$. Then the variance and expected value of $\hat{f}(x)$ are given by

$$\begin{aligned} E(\hat{f}(x)) &= E\left(\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)\right) \\ &= \frac{1}{nh}\sum_{i=1}^{n} E\left(K\left(\frac{x-X_i}{h}\right)\right) \\ &= \frac{1}{nh}\sum_{i=1}^{n} E\left(K\left(\frac{x-Y}{h}\right)\right) \\ &= \frac{1}{h}E\left(K\left(\frac{x-Y}{h}\right)\right) \\ &= \int \frac{1}{h}K\left(\frac{x-y}{h}\right) f(y)\, dy \end{aligned} \tag{3.1.3}$$

and

$$\begin{aligned} \operatorname{var}(\hat{f}(x)) &= \operatorname{var}\left\{\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)\right\} \\ &= \frac{1}{nh^2}\operatorname{var}\left\{K\left(\frac{x-Y}{h}\right)\right\} \\ &= \frac{1}{nh^2}E\left\{K\left(\frac{x-Y}{h}\right)^2\right\} - \frac{1}{nh^2}\left[E\left\{K\left(\frac{x-Y}{h}\right)\right\}\right]^2 \\ &= \frac{1}{n}\int \frac{1}{h^2}\left[K\left(\frac{x-y}{h}\right)\right]^2 f(y)\, dy - \\ &\quad \frac{1}{n}\left[\int \frac{1}{h}K\left(\frac{x-y}{h}\right) f(y)\, dy\right]^2 \end{aligned} \tag{3.1.4}$$

Substituting the expressions found in (3.1.3) and (3.1.4) into equation (3.1.2), we can now obtain an exact formula for the mean integrated square error of $\hat{f}(x)$.

However, in many cases, minimizing the exact MISE in order to find the best kernel function and smoothing parameter can be difficult, computationally intensive and impractical. To rectify this, Silverman (1986) suggests using approximate values for the variance and bias. This leads to a simpler, approximate formula for the mean integrated square error. Silverman (1986) further indicates that this approximation is quite satisfactory since minimizing it will give results that coincide closely to the results that would have been obtained if the exact expression of the MISE had been used instead, even for small sample sizes.

Theorem 3.1.1 *Suppose the kernel $K(t)$ is a symmetric function that satisfies $\int K(t)\,dt = 1$, $\int tK(t)\,dt = 0$ and $\int t^2K(t)\,dt = b \neq 0$. Then, the approximate mean integrated square error is given by*

$$MISE \approx \frac{1}{4}h^4b^2 \int (f''(x))^2\,dx + \frac{1}{nh}\int (K(t))^2\,dt$$

Proof: Let's begin by calculating an approximation for the bias of $\hat{f}(x)$. We know that

$$\begin{aligned} \text{bias}(\hat{f}(x)) &= E\left(\hat{f}(x) - f(x)\right) \\ &= E\left(\hat{f}(x)\right) - f(x) \\ &= \int \frac{1}{h}K\left(\frac{x-y}{h}\right) f(y)\,dy - f(x) \qquad \text{by (3.1.3)} \end{aligned}$$

Note that $K(x)$ is a symmetric function and so $K\left(\frac{x-y}{h}\right) = K\left(\frac{y-x}{h}\right)$ and make the change of variable $y = x + ht$, then

$$\text{bias}(\hat{f}(x)) = \int K(t)f(x+ht)\,dt - f(x)$$

$$= \int K(t)f(x+ht)\,dt - f(x)\int K(t)\,dt \qquad \text{since } \int K(t)\,dt = 1$$
$$= \int K(t)\{f(x+ht) - f(x)\}\,dt$$

Using a Taylor series expansion of $f(x+ht)$ about x gives

$$f(x+ht) = f(x) + (ht)f'(x) + \frac{(h^2t^2)}{2}f''(x) + \ldots \tag{3.1.5}$$

and so

$$\begin{aligned}\text{bias}(\hat{f}(x)) &= \int K(t)\left\{(ht)f'(x) + \frac{(h^2t^2)}{2}f''(x) + \ldots\right\}dt \\ &= hf'(x)\int tK(t)\,dt + \frac{h^2}{2}f''(x)\int t^2K(t)\,dt + \text{higher order terms in } h \\ &= \frac{h^2}{2}f''(x)b + \text{higher order terms in } h\end{aligned}$$

since $\int tK(t)\,dt = 0$ and $\int t^2K(t)\,dt = b \neq 0$ by assumption. Furthermore, since h generally takes on small values, we have that

$$\text{bias}(\hat{f}(x)) \approx \frac{h^2}{2}f''(x)b \tag{3.1.6}$$

Similarly, we can derive an approximate expression for the variance of $\hat{f}(x)$. We know from equation (3.1.4) that

$$\begin{aligned}\text{var}(\hat{f}(x)) &= \frac{1}{n}\int \frac{1}{h^2}\left[K\left(\frac{x-y}{h}\right)\right]^2 f(y)\,dy - \frac{1}{n}\left[\int \frac{1}{h}K\left(\frac{x-y}{h}\right)f(y)\,dy\right]^2 \\ &= \frac{1}{n}\int \frac{1}{h^2}\left[K\left(\frac{x-y}{h}\right)\right]^2 f(y)\,dy - \frac{1}{n}\left[E(\hat{f}(x))\right]^2 \\ &= \frac{1}{n}\int \frac{1}{h^2}\left[K\left(\frac{x-y}{h}\right)\right]^2 f(y)\,dy - \frac{1}{n}\left[f(x) + \text{ bias }\right]^2\end{aligned}$$

$$\approx \frac{1}{n}\int \frac{1}{h^2}\left[K\left(\frac{x-y}{h}\right)\right]^2 f(y)\,dy - \frac{1}{n}\left[f(x)+O(h^2)\right]^2$$

since we know that the bias is approximately $O(h^2)$ from equation (3.1.6). Again, noting that $K(x)$ is a symmetric function and so $K\left(\frac{x-y}{h}\right) = K\left(\frac{y-x}{h}\right)$ and making the substitution $y = x + ht$, we then have

$$\text{var}(\hat{f}(x)) \approx \frac{1}{nh}\int (K(t))^2 f(x+ht)\,dt - \frac{1}{n}\left[f(x)+O(h^2)\right]^2$$

Assuming n is large and h is small and using the Taylor series expansion of $f(x+ht)$ around x, given in equation (3.1.5), we obtain

$$\begin{aligned}
\text{var}(\hat{f}(x)) &\approx \frac{1}{nh}\int (K(t))^2\left\{f(x)+(ht)f'(x)+\frac{h^2t^2}{2}f''(x)+\ldots\right\}dt + O\left(n^{-1}\right) \\
&\approx \frac{1}{nh}\int (K(t))^2 f(x)\,dt + O\left(n^{-1}\right) \\
&\approx \frac{f(x)}{nh}\int (K(t))^2\,dt \qquad (3.1.7)
\end{aligned}$$

Combining equations (3.1.6) and (3.1.7) and substituting them in equation (3.1.2), we find that the approximate mean integrated square error is given by

$$\begin{aligned}
\text{MISE} &= \int \text{var}(\hat{f}(x))\,dx + \int \left[\text{ bias}(\hat{f}(x))\right]^2 dx \\
&\approx \int \frac{f(x)}{nh}\left[\int (K(t))^2\,dt\right]dx + \int \left[\frac{h^2}{2}f''(x)b\right]^2 dx \\
&= \frac{1}{nh}\int (K(t))^2\,dt + \frac{h^4}{4}b^2\int (f''(x))^2\,dx
\end{aligned}$$

since $\int f(x)\,dx = 1$. ∎

Now that we have a simpler expression for the mean integrated square error, we can proceed to find the kernel function and bandwidth that will minimize it.

We begin with the smoothing parameter, h. Using simple calculus, it can easily be shown that the optimal bandwidth, h_{opt}, that is the bandwidth which minimizes the approximate mean integrated square error of Theorem 3.1.1, is given by

$$h_{opt} = b^{-2/5} \left\{ \int (K(t))^2 \, dt \right\}^{1/5} \left\{ \int (f''(x))^2 \, dx \right\}^{-1/5} n^{-1/5}$$

Hence, it can be seen that the optimal smoothing parameter will depend on the kernel, $K(t)$, and the true density $f(x)$. This can be problematic since in practice no information is known about f. Nevertheless, Silverman (1986) proposes a solution to this problem which will be discussed further in this section.

We now turn our attention to finding the optimal kernel function, $K(t)$. Substituting h_{opt} in the expression for the approximate MISE gives

$$\begin{aligned}
\text{MISE} &\approx \frac{1}{4} h_{opt}^4 b^2 \int (f''(x))^2 \, dx + \frac{1}{n h_{opt}} \int (K(t))^2 \, dt \\
&= \frac{1}{4} b^{-8/5} \left\{ \int (K(t))^2 \, dt \right\}^{4/5} \left\{ \int (f''(x))^2 \, dx \right\}^{-4/5} b^2 n^{-4/5} \left\{ \int (f''(x))^2 \, dx \right\} + \\
&\quad n^{-1} b^{2/5} \left\{ \int (K(t))^2 \, dt \right\}^{-1/5} \left\{ \int (f''(x))^2 \, dx \right\}^{1/5} n^{1/5} \left\{ \int (K(t))^2 \, dt \right\} \\
&= \frac{1}{4} b^{2/5} \left\{ \int (K(t))^2 \, dt \right\}^{4/5} \left\{ \int (f''(x))^2 \, dx \right\}^{1/5} n^{-4/5} + \\
&\quad b^{2/5} \left\{ \int (K(t))^2 \, dt \right\}^{4/5} \left\{ \int (f''(x))^2 \, dx \right\}^{1/5} n^{-4/5} \\
&= \frac{5}{4} b^{2/5} \left\{ \int (K(t))^2 \, dt \right\}^{4/5} \left\{ \int (f''(x))^2 \, dx \right\}^{1/5} n^{-4/5}
\end{aligned}$$

Since $\frac{5}{4}\left\{\int (f''(x))^2\,dx\right\}^{1/5} n^{-4/5}$ will have the same value for all kernels, then the optimal kernel will be the one which minimizes $b^{2/5}\left\{\int (K(t))^2\,dt\right\}^{4/5}$ where $b = \int t^2 K(t)\,dt$. Hodges and Lehmann (1956) showed that the optimal kernel is the Epanechnikov kernel and it is defined as

$$K_e(t) = \begin{cases} \frac{3}{4\sqrt{5}}\left(1 - \frac{1}{5}t^2\right) & \text{if } -\sqrt{5} \leqslant t \leqslant \sqrt{5} \\ 0 & \text{otherwise.} \end{cases}$$

It is important to note that there exist many other choices for the kernel function. Table 3.1 lists four other possible choices and further indicates the efficiency of each kernel in comparison to the Epanechnikov kernel. We notice from this table that the efficiencies are quite high and so we can conclude that the kernels do not differ greatly in terms of the mean integrated square error. Hence, as Silverman (1986) states, "it is perfectly legitimate, and indeed desirable, to base the choice of the kernel on other considerations, for example the degree of differentiability required or the computational effort involved". In fact, it is precisely for these reasons that we will choose to use a Gaussian kernel in the Monte Carlo simulation study of Chapter 6.

The problem of choosing the smoothing parameter remains. Recall that

$$h_{opt} = b^{-2/5}\left\{\int (K(t))^2\,dt\right\}^{1/5}\left\{\int (f''(x))^2\,dx\right\}^{-1/5} n^{-1/5}.$$

Once the kernel has been selected, b and $\int (K(t))^2\,dt$ can easily be computed. However, $\int (f''(x))^2\,dx$ cannot be evaluated without having information about the true density $f(x)$. Silverman (1986) suggests using "a standard family of distributions to assign a value to the term $\int (f''(x))^2\,dx$". A natural choice is to use a normal

Table 3.1: Some kernels and their efficiencies

Kernel	$K(t)$	Efficiency
Epanechnikov	$\begin{cases} \frac{3}{4\sqrt{5}}\left(1-\frac{1}{5}t^2\right) & \text{if } -\sqrt{5} \leqslant t \leqslant \sqrt{5} \\ 0 & \text{otherwise} \end{cases}$	1
Biweight	$\begin{cases} \frac{15}{16}(1-t^2)^2 & \text{if } \lvert t\rvert < 1 \\ 0 & \text{otherwise} \end{cases}$	$\left(\frac{3087}{3125}\right)^{1/2} \approx 0.9939$
Triangular	$\begin{cases} 1-\lvert t\rvert & \text{if } \lvert t\rvert < 1 \\ 0 & \text{otherwise} \end{cases}$	$\left(\frac{243}{250}\right)^{1/2} \approx 0.9859$
Gaussian	$\frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-(1/2)t^2}, \quad -\infty < t < \infty$	$\left(\frac{36\pi}{125}\right)^{1/2} \approx 0.9512$
Rectangular	$\begin{cases} \frac{1}{2} & \text{if } \lvert t\rvert < 1 \\ 0 & \text{otherwise} \end{cases}$	$\left(\frac{108}{125}\right)^{1/2} \approx 0.9295$

distribution with variance σ^2. In this case, if a Gaussian kernel is equally utilized, then h_{opt} is found to be

$$h_{opt} = 1.06\sigma n^{-1/5}. \tag{3.1.8}$$

Usually the population variance, σ^2, is unknown and so σ can be replaced in (3.1.8) by the estimated standard deviation of the observations $X_1, \ldots, X_n$. Sometimes a more robust estimate of the spread is needed and so Silverman (1986) suggests using

$$h^*_{opt} = 1.06\mathrm{A}n^{-1/5}$$

where A=min(standard deviation, interquartile range/1.34). The book then goes on

to say that this choice of smoothing parameter works very well when the true density of the population, $f(x)$, is unimodal.

It is therefore this smoothing parameter that will be used when conducting the simulations presented in Chapter 6. In this Monte Carlo simulation study, all simulated data were obtained from unimodal densities and so this choice of bandwidth seems reasonable and appropriate.

In cases where the population does not appear to have a unimodal density and no additional information is available about the true density $f(x)$, then h^*_{opt} is no longer the best choice of bandwidth. Instead, there exist other methods, such as least squares cross-validation or likelihood cross-validation, which lead to a better selection of the smoothing parameter. Additionally, it is possible to select the bandwidth by minimizing a criterion other than the mean integrated square error. Such techniques were not needed for the purpose of this thesis and so they are not presented. The interested reader is referred to Silverman (1986), Sheather and Jones (1991) and Scott (1992) where other methods of selecting the ideal bandwidth are discussed.

3.2 Multivariate Kernel Density Estimation

The results of the previous section can now be extended to multivariate populations. The multivariate results of this section will be given without proof. However, more information regarding rigorous theoretical details can be found in Cacoullos (1966), Epanechnikov (1969) and Silverman (1986).

Let $X_1, \ldots, X_n$ be a random sample from a $p-$dimensional population with density function $f(x)$. The multivariate kernel density estimate of the density function,

$f(x)$, constructed from the observations is given by

$$\hat{f}(x) = \frac{1}{nh^p}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

where $K(x)$ is the kernel function and h is the smoothing parameter.

As in the univariate case, we can find an expression for the approximate mean integrated square error of $\hat{f}(x)$ and minimize it in order to find the optimal bandwidth and kernel function.

Let $K(t)$ be a radially symmetric probability density function and let $\alpha = \int t_1^2 K(t)\,dt$ and $\beta = \int (K(t))^2\,dt$. Then combining multivariate Taylor series expansions and similar manipulations as those presented in Theorem 3.1.1, we have the following approximations

$$\text{bias}(\hat{f}(x)) \approx \frac{1}{2}h^2\alpha\nabla^2 f(x) \tag{3.2.1}$$

and

$$\text{var}(\hat{f}(x)) \approx n^{-1}h^{-p}\beta f(x). \tag{3.2.2}$$

Substituting (3.2.1) and (3.2.2) into (3.1.2) gives the following approximate value of the mean integrated square error

$$\text{MISE} \approx \frac{1}{4}h^4\alpha^2 \int \left\{\nabla^2 f(x)\right\}^2 dx + n^{-1}h^{-p}\beta. \tag{3.2.3}$$

Applying simple calculus to equation (3.2.3) reveals that the optimal bandwidth is

$$h_{opt}^{p+4} = p\beta\alpha^{-2}\left\{\int \left(\nabla^2 f(x)\right)^2 dx\right\}^{-1} n^{-1}. \tag{3.2.4}$$

Moreover, it can be shown that the kernel function which minimizes the approximate MISE is a multivariate version of the Epanechnikov kernel and is defined as

$$K_e(x) = \begin{cases} \frac{1}{2}v_p^{-1}(p+2)(1-x'x)^2 & \text{if } x'x < 1 \\ 0 & \text{otherwise} \end{cases}$$

where v_p is the volume of the unit $p-$dimensional sphere.

Like the univariate case, many other kernel functions are available. Table 3.2 presents three other possible multivariate kernels that can be used when estimating a density function. The first kernel in the table, $K_g(x)$, is the multivariate extension of the Gaussian kernel. Although the efficiencies of these kernels relative to the Epanechnikov kernel are not provided, Silverman (1986) does state that the three kernels of Table 3.2 "can achieve very similar mean integrated square errors" to that of the Epanechnikov kernel and hence, have a high efficiency. Thus, it is recommended that the selection of the kernel should be based on other criteria such as the differentiability requirements and computational considerations. Again, for these reasons, the multivariate Gaussian kernel will be utilized for all multivariate simulations of Chapter 6.

Once the kernel has been selected, values for α and β can readily be calculated and substituted in (3.2.4) to find h_{opt}. However, an exact value for this smoothing

Table 3.2: Three multivariate kernels

Kernel	$K(x)$
Multivariate Gaussian Kernel	$K_g(x) = (2\pi)^{-p/2} \exp\left(-\frac{1}{2}x'x\right)$
Multivariate Kernel 2	$K_2(x) = \begin{cases} 3\pi^{-1}(1-x'x)^2 & \text{if } x'x < 1 \\ 0 & \text{otherwise} \end{cases}$
Multivariate Kernel 3	$K_3(x) = \begin{cases} 4\pi^{-1}(1-x'x)^3 & \text{if } x'x < 1 \\ 0 & \text{otherwise} \end{cases}$

parameter still cannot be obtained since h_{opt} depends on the true population density, $f(x)$. To solve this problem, we proceed as in the univariate case and select a family of distributions in order to obtain a value for the term $\int \{\nabla^2 f(x)\}^2 \, dx$. A natural choice is to take $f(x)$ as the standard $p-$dimensional normal density. In this case,

$$\int \{\nabla^2 f(x)\}^2 \, dx = (2\sqrt{\pi})^{-p} \left(\frac{2p+p^2}{4}\right).$$

Then, if a multivariate Gaussian kernel is equally used, we have

$$h_{opt} = \{4/(2p+1)\}^{1/(p+4)} \, n^{-1/(p+4)}.$$

Silverman (1986) adds that if instead the true population density, $f(x)$, is multivariate normal with covariance matrix Σ, then h_{opt} can be modified to take this fact into account. Let S be a, possibly robust, estimate of the true covariance matrix Σ and let s_{ii} be the i^{th} diagonal element of S. Then a more appropriate smoothing

parameter might be σh_{opt} where

$$\sigma^2 = \frac{1}{p}\sum_{i=1}^{p} s_{ii}.$$

This is the smoothing parameter which will be used for the multivariate simulations of Chapter 6. This bandwidth was selected since all multivariate data sets utilized in the Monte Carlo simulation study were generated from various multivariate normal distributions.

It is important to note that like the optimal univariate smoothing parameter, $\sigma h_{opt} = \sigma \{4/(2p+1)\}^{1/(p+4)} n^{-1/(p+4)}$, should be used with caution. Although this approach provides a quick and easy way of obtaining an initial value of the bandwidth, it is not the optimal value for all situations. Other univariate methods, such as least-squares cross-validation and likelihood cross-validation, can be extended to the multivariate case and may yield a more ideal value for the smoothing parameter. Also, in some cases, a vector of bandwidths may even be more appropriate than a single value. For more information regarding these techniques, see Silverman (1986).

Finally, an additional concern presents itself when doing multivariate kernel density estimation. We must now consider the sample size required in order to obtain an accurate estimate of the true population density, $f(x)$. Evidently, as the number of dimensions (or variables) increases, the amount of data needed to produce an adequate estimate will also augment. The problem is that the number of observations required can increase alarmingly fast. As an example, Silverman (1986) studies the following problem. "Suppose that the true density f is unit multivariate normal, and that the kernel is normal. [Also,] suppose that it is of interest to estimate f at the point 0, and that the [bandwidth] h has been chosen to minimize the mean square error at this point. A not unreasonable aim would be to ensure that the relative mean

square error $E\{\hat{f}(0) - f(0)\}^2/f(0)^2$ is fairly small, say less than 0.1." Table 3.3 gives the approximate sample sizes, accurate to about three figures, required to meet the criteria of this problem for various dimensions.

Table 3.3: Sample sizes required for increasing dimensionality

Dimensionality	Required sample size
1	4
2	19
3	67
4	223
5	768
6	2790
7	10700
8	43700
9	187000
10	842000

It should be noted that for classification problems, we are more interested in determining if the ratio $\hat{f}_1(z)/\hat{f}_2(z)$ is greater than p_2/p_1 rather than correctly estimating the value of $f_1(z)$ or $f_2(z)$. In this case, an incorrect value of $f_1(z)$ or $f_2(z)$ may still lead to a correct classification. Hence, the sample sizes needed to correctly classify a new observation may indeed be smaller than those given in Table 3.3. However, although the sample sizes may be smaller, they are nevertheless still large and continue to increase with the dimensions.

Chapter 4

Fisher's Discriminant Function and the Rank Transform

The following nonparametric approach to classification does not use the Bayes rule in any way. Unlike the previous chapter, we are not interested in estimating the population densities, f_1 and f_2. Rather, the classification rule will be developed based on a discrimination analysis of the data.

The purpose of a discriminant function differs from that of a classifier. Discrimination analysis seeks to separate or find a separation between two groups or populations whereas classification will determine how to allocate new observations to each population. They can be used jointly to develop a classification rule. Such is the case with Fisher's discriminant function, presented in Johnson and Wichern (2002). All definitions, lemmas and theorems are taken from this book unless otherwise stated.

4.1 Fisher's Discriminant Function

Fisher's idea was to transform the multivariate observations $x_1, \ldots, x_{n_1}, y_1, \ldots, y_{n_2}$ into simpler functions of the data. To do so, he considered linear combinations which would transform the data to a set of univariate observations. This led to a new training sample $w_{11}, \ldots, w_{1n_1}, w_{21}, \ldots, w_{2n_2}$ where $w_{1i} = a'x_i$, $i = 1, \ldots, n_1$ and $w_{2j} = a'y_j$, $j = 1, \ldots, n_2$.

Using these new univariate observations, Fisher's goal was to find the maximal separation between the two populations. In his discriminant analysis, he measured the separation between the two groups as a ratio of the distance between the means of both groups to the variation within each group. That is

$$\text{separation} = \frac{(\overline{w}_1 - \overline{w}_2)^2}{s_w^2}$$

where $\overline{w}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} w_{1i}$, $\overline{w}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} w_{2j}$ and $s_w^2 = \frac{\sum_{i=1}^{n_1}(w_{1i}-\overline{w}_1)^2+\sum_{j=1}^{n_2}(w_{2j}-\overline{w}_2)^2}{n_1+n_2-2}$.

Now this separation can be written in terms of the original multivariate observations. We begin by noting that

$$\begin{aligned} \overline{w}_1 &= \tfrac{1}{n_1}\textstyle\sum_{i=1}^{n_1} w_{1i} & \overline{w}_2 &= \tfrac{1}{n_2}\textstyle\sum_{j=1}^{n_2} w_{2j} \\ &= \tfrac{1}{n_1}\textstyle\sum_{i=1}^{n_1} a'x_i & &= \tfrac{1}{n_2}\textstyle\sum_{j=1}^{n_2} a'y_j \\ &= a'\overline{x} & &= a'\overline{y} \end{aligned}$$

and

$$\begin{aligned} s_w^2 &= \frac{\sum_{i=1}^{n_1}(w_{1i}-\overline{w}_1)(w_{1i}-\overline{w}_1)' + \sum_{j=1}^{n_2}(w_{2j}-\overline{w}_2)(w_{2j}-\overline{w}_2)'}{n_1+n_2-2} \\ &= \frac{\sum_{i=1}^{n_1} a'(x_i-\overline{x})(x_i-\overline{x})'a + \sum_{j=1}^{n_2} a'(y_j-\overline{y})(y_j-\overline{y})'a}{n_1+n_2-2} \end{aligned}$$

$$= \frac{a'(n_1-1)S_1a + a'(n_2-1)S_2a}{n_1+n_2-2}$$
$$= a'S_pa$$

Therefore,

$$\text{separation} = \frac{(a'\overline{x} - a'\overline{y})^2}{a'S_pa}$$
$$= \frac{(a'd)^2}{a'S_pa}$$

where $d = \overline{x} - \overline{y}$. Thus, Fisher sought the linear combination of the data which would maximize this ratio or equivalently, maximize the separation between both populations.

In order to find this linear combination, a few maximization lemmas are needed.

Lemma 4.1.1 *Let X and Y be $p-$dimensional vectors, then*

$$(X'Y)^2 \leqslant (X'X)(Y'Y).$$

Proof: Recall that $u'v =< u, v >= \|u\| \cdot \|v\| \cos\theta$ where $< u, v >$ denotes the scalar product of u and v and $\|u\| = \sqrt{u_1^2 + \ldots + u_p^2} = \sqrt{u'u}$ denotes the Euclidean norm of $u = [u_1, \ldots, u_p]'$. Then

$$X'Y = \|X\| \cdot \|Y\| \cos\theta$$
$$= (X'X)^{1/2}(Y'Y)^{1/2}\cos\theta$$

and hence,

$$(X'Y)^2 = (X'X)(Y'Y)(\cos\theta)^2$$
$$\leqslant (X'X)(Y'Y)$$

since $(\cos\theta)^2 \leqslant 1$. ∎

Lemma 4.1.2 *Let X and $Y \in \mathbb{R}^p$ and let A be a $p \times p$ symmetric positive definite matrix, then*

$$(X'Y)^2 \leqslant (X'AX)(Y'A^{-1}Y).$$

Proof: Recall that for any symmetric matrix, A, we can find, by the spectral decomposition theorem, an orthogonal matrix, P, and a diagonal matrix, $\Lambda = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_p \end{bmatrix}$, such that $A = P\Lambda P'$ and $A^{-1} = P\Lambda^{-1}P'$.

Let $\Lambda^{1/2} = \begin{bmatrix} \lambda_1^{1/2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_p^{1/2} \end{bmatrix}$ and $\Lambda^{-1/2} = \begin{bmatrix} \lambda_1^{-1/2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_p^{-1/2} \end{bmatrix}$ and define $A^{1/2} = P\Lambda^{1/2}P'$ and $A^{-1/2} = P\Lambda^{-1/2}P'$. Then note that $A^{1/2}$ and $A^{-1/2}$ are symmetric since $\Lambda^{1/2}$ and $\Lambda^{-1/2}$ are diagonal matrices and therefore symmetric. Also, we have that

$$\begin{aligned} A^{1/2}A^{1/2} &= P\Lambda^{1/2}P'P\Lambda^{1/2}P' \\ &= P\Lambda P' \\ &= A \end{aligned} \tag{4.1.1}$$

since $\Lambda^{1/2}\Lambda^{1/2} = \Lambda$ and $PP' = P'P = I =$ because P is an orthogonal matrix. Similarly,

$$\begin{aligned} A^{-1/2}A^{-1/2} &= P\Lambda^{-1/2}P'P\Lambda^{-1/2}P' \\ &= P\Lambda^{-1}P' \\ &= A^{-1}. \end{aligned} \tag{4.1.2}$$

Moreover,

$$\begin{aligned} A^{1/2}A^{-1/2} &= P\Lambda^{1/2}P'P\Lambda^{-1/2}P' \\ &= P\Lambda^{1/2}\Lambda^{-1/2}P' \end{aligned}$$

$$\begin{aligned} &= PIP' \\ &= PP' \\ &= I. \end{aligned} \tag{4.1.3}$$

Therefore,

$$\begin{aligned} (X'Y)^2 &= (X'IY)^2 \\ &= (X'A^{1/2}A^{-1/2}Y)^2 \quad \text{by (4.1.3)} \\ &= \left[(A^{1/2}X)'(A^{-1/2}Y)\right]^2 \\ &\leqslant (A^{1/2}X)'(A^{1/2}X)\cdot(A^{-1/2}Y)'(A^{-1/2}Y) \quad \text{by Lemma 4.1.1} \\ &= (X'A^{1/2}A^{1/2}X)(Y'A^{-1/2}A^{-1/2}Y) \\ &= (X'AX)(Y'A^{-1}Y) \quad \text{by (4.1.1) and (4.1.2).} \end{aligned}$$

∎

Lemma 4.1.3 *Let X and $Y \in \mathbb{R}^p$ and let A be a $p \times p$ symmetric positive definite matrix, then*

$$\max_{X \neq 0} \frac{(X'Y)^2}{X'AX} = Y'A^{-1}Y$$

and occurs when $X = A^{-1}Y$.

Proof: From Lemma 4.1.2, we know that $(X'Y)^2 \leqslant (X'AX)(Y'A^{-1}Y)$ which implies that $\frac{(X'Y)^2}{X'AX} \leqslant Y'A^{-1}Y$. If $X = A^{-1}Y$, then the maximum is attained since

$$\begin{aligned} \frac{(X'Y)^2}{X'AX} &= \frac{((A^{-1}Y)'Y)^2}{(A^{-1}Y)'A(A^{-1}Y)} \\ &= \frac{(Y'A^{-1}Y)^2}{Y'A^{-1}AA^{-1}Y} \end{aligned}$$

$$\begin{aligned} &= \frac{(Y'A^{-1}Y)^2}{Y'A^{-1}Y} \\ &= Y'A^{-1}Y. \end{aligned}$$

∎

Thus, we can now find Fisher's discriminant function. This is the linear combination of the observations which maximizes the separation between both populations. By applying Lemma 4.1.3, we find that the ratio $\frac{(a'd)^2}{a'S_pa}$, where $d = \overline{x} - \overline{y}$, is maximized by taking $a' = d'S_p^{-1} = (\overline{x} - \overline{y})'S_p^{-1}$. Furthermore, the maximum separation between the two populations is given by $(\overline{x} - \overline{y})'S_p^{-1}(\overline{x} - \overline{y})$.

4.2 Fisher's Classification Rule

The discriminant analysis performed in the previous section can now be used to derive a classification rule. Since the transformed observations have the greatest separation, we can base our classifier on this modified data. First, note that

$$\begin{aligned} \overline{w}_1 - \overline{w}_2 &= a'(\overline{x} - \overline{y}) \\ &= (\overline{x} - \overline{y})'S_p^{-1}(\overline{x} - \overline{y}) \end{aligned}$$

and that S_p is a positive definite matrix and thus, so is its inverse, S_p^{-1}. By definition, a matrix A is positive definite if and only if $x'Ax > 0$, for all nonzero x. Consequently, $\overline{w}_1 - \overline{w}_2$ will be greater than zero unless $\overline{x} = \overline{y}$ in which case $\overline{w}_1 = \overline{w}_2$. We may then conclude that $\overline{w}_1 - \overline{w}_2 \geqslant 0$ which implies that $\overline{w}_1 \geqslant \overline{w}_2$.

We can also estimate the point of separation between the two populations as $m = \frac{1}{2}(\overline{w}_1 + \overline{w}_2)$. This is the midpoint between $\overline{w}_1$ and $\overline{w}_2$. Naturally, if the linear

combination of the new observation, $w_0 = a'z = (\overline{x} - \overline{y})'S_p^{-1}z$, is larger than m, then it belongs to π_1 since $\overline{w}_1 \geqslant \overline{w}_2$. Hence, Fisher's classification rule, also frequently referred to as Anderson's classification function, can be stated as

$$\begin{aligned}
\varphi_F(w_0) &= \begin{cases} 1 & \text{if } w_0 \geqslant \frac{1}{2}(\overline{w}_1 + \overline{w}_2) \\ 2 & \text{otherwise} \end{cases} \\
\varphi_F(z) &= \begin{cases} 1 & \text{if } (\overline{x} - \overline{y})'S_p^{-1}z \geqslant \frac{1}{2}(\overline{x} - \overline{y})S_p^{-1}(\overline{x} + \overline{y}) \\ 2 & \text{otherwise} \end{cases} \\
&= \begin{cases} 1 & \text{if } (\overline{x} - \overline{y})'S_p^{-1}z - \frac{1}{2}(\overline{x} - \overline{y})'S_p^{-1}(\overline{x} + \overline{y}) \geqslant 0 \\ 2 & \text{otherwise} \end{cases}
\end{aligned}$$

Fisher's approach is entirely nonparametric since no assumptions were made about the underlying population densities. The method, however, does implicitly suppose that both populations have equal covariance matrices since a pooled estimate of the covariance matrix, S_p, is utilized. Remarkably, under the assumptions of equal misclassification costs and equal prior probabilities, Fisher's classifier is the same as the Bayes classification rule (with the parameters replaced by their estimates) obtained when both populations have multivariate normal densities with equal covariance matrices. Therefore, under these circumstances, Fisher's classifier is asymptotically optimal. Nevertheless, there is no guarantee that it will be a reasonable classification rule when the population densities are no longer normal.

4.3 The Rank Transform

Another problem with Fisher's classifier is that the rule depends on the means $\overline{w}_1$ and $\overline{w}_2$. These values can be easily influenced by extreme observations. To minimize the effect of such outliers Conover and Iman (1980) propose applying Fisher's discriminant function to the ranks of the observations instead. Hence, we transform the data into a new set of observations. These new observations are the ranks of the original data. We then apply Fisher's classification rule using these transformed observations. Therefore, $\overline{x}$ will now be replaced by $\overline{r}_x$, the average rank of the observations of the first population. Similarly, $\overline{y}$ will be replaced by $\overline{r}_y$, the average rank of the observations of the second population and S_p will now be a pooled estimate of the variance of the ranks of the observations.

Not only is this rank transform robust to outliers but such a transformation may also help when the population densities are not normal or when the populations have unequal covariance matrices. This stems from the fact that we are no longer dealing with the population densities but rather with the distribution of the ranked data.

Finally, it is important to note that Akritas (1990) and Thompson (1991) both express that the rank transform is most effective when populations differ in location. When only a change in scale is present, the rank transform is less useful. Thankfully, in most classification problems, the populations differ in location and so the rank transform is often applicable.

Chapter 5

A New Classification Rule Based on Ranks

In the previous chapters, three commonly used classification techniques were presented and discussed. We will now propose a new nonparametric classification rule. This classification method, which is based on the ranks of the observations, is derived by modifying the test statistic of Mantel and Valand (1970) and combining it with Fisher's linear discriminant function.

5.1 Preliminary Requirements

Before detailing the new classification procedure, we begin by outlining some of the ideas of Mantel and Valand (1970) which will be useful in developing our new classifier.

Let $W_i = [W_{1i}, \ldots, W_{pi}]', i = 1, \ldots, n$ be independent $p-$dimensional observations coming from two possible populations, π_1 and π_2, where it is assumed that the true class membership of each observation is known. Also, let $R_{\alpha i}$ be the rank of the

α^{th} variable on the i^{th} observation, $\alpha = 1, \ldots, p$ and $i = 1, \ldots, n$. That is, $R_{\alpha i}$ is the rank of $W_{\alpha i}$ among $W_{\alpha 1}, \ldots, W_{\alpha n}$ and can be calculated as

$$R_{\alpha i} = \sum_{j=1}^{n} u(W_{\alpha i} - W_{\alpha j}) \tag{5.1.1}$$

where $u(x) = \begin{cases} 1 & \text{if } x \geqslant 0 \\ 0 & \text{otherwise} \end{cases}$. Then Mantel and Valand (1970) define a general measure of distance between observations i and j as

$$Y_{ij} = \sum_{\alpha=1}^{p} \eta_\alpha f_\alpha(R_{\alpha i}, R_{\alpha j})$$

where f_α is a positive symmetric function and η_α is a nonnegative weight.

Furthermore, let

$$X_{ij} = \begin{cases} 1 & \text{if } W_i \text{ and } W_j \text{ are from the same class} \\ 0 & \text{otherwise.} \end{cases}$$

Then, in order to demonstrate differences between groups, Mantel and Valand (1970) propose the following test statistic

$$\begin{aligned} Z &= \sum_{i=1}^{n}\sum_{j=1}^{n} X_{ij}Y_{ij} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} X_{ij} \sum_{\alpha=1}^{p} \eta_\alpha f_\alpha(R_{\alpha i}, R_{\alpha j}) \\ &= \sum_{\alpha=1}^{p} \eta_\alpha \sum_{i=1}^{n}\sum_{j=1}^{n} X_{ij} f_\alpha(R_{\alpha i}, R_{\alpha j}) \end{aligned}$$

Evidently, there exists many possible choices for the function $f_\alpha(R_{\alpha i}, R_{\alpha j})$. One possibility, suggested by Mantel and Valand (1970), is $f_\alpha(R_{\alpha i}, R_{\alpha j}) = |R_{\alpha i} - R_{\alpha j}|$.

However, for the new classification rule, we will use instead

$$f_\alpha(R_{\alpha i}, R_{\alpha j}) = f(R_{\alpha i}, R_{\alpha j}) = \frac{1}{2}(R_{\alpha i} - R_{\alpha j})^2. \tag{5.1.2}$$

We made this particular choice because $\sum_i \sum_j (R_{\alpha i} - R_{\alpha j})^2$ gives a measure of the variation between the ranks of the observations for the α^{th} variable. In fact, recall that the variance of the ranks of the observations of the α^{th} variable is given by

$$\begin{aligned} S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(R_{\alpha i} - \overline{R}_\alpha)^2 \\ &= \frac{1}{n-1}\left[\sum_{i=1}^{n} R_{\alpha i}^2 - 2\sum_{i=1}^{n} R_{\alpha i}\overline{R}_\alpha + \sum_{i=1}^{n}\overline{R}_\alpha^2\right] \\ &= \frac{1}{n-1}\left[\sum_{i=1}^{n} R_{\alpha i}^2 - n\overline{R}_\alpha^2\right] \end{aligned}$$

since $\sum_{i=1}^{n} R_{\alpha i} = n\overline{R}_\alpha$. Also note that

$$\begin{aligned} S_*^2 &= \frac{1}{2n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}(R_{\alpha i} - R_{\alpha j})^2 \\ &= \frac{1}{2n(n-1)}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} R_{\alpha i}^2 - 2\sum_{i=1}^{n}\sum_{j=1}^{n} R_{\alpha i}R_{\alpha j} + \sum_{i=1}^{n}\sum_{j=1}^{n} R_{\alpha j}^2\right] \\ &= \frac{1}{2n(n-1)}\left[n\sum_{i=1}^{n} R_{\alpha i}^2 - 2\sum_{i=1}^{n}\sum_{j=1}^{n} R_{\alpha i}R_{\alpha j} + n\sum_{j=1}^{n} R_{\alpha j}^2\right] \end{aligned}$$

and since $\sum_{i=1}^{n} R_{\alpha i}^2 = \sum_{j=1}^{n} R_{\alpha j}^2$ and $\sum_{i=1}^{n} R_{\alpha i} = \sum_{j=1}^{n} R_{\alpha j} = n\overline{R}_\alpha$, we have

$$S_*^2 = \frac{1}{2n(n-1)}\left[2n\sum_{i=1}^{n} R_{\alpha i}^2 - 2n^2\overline{R}_\alpha^2\right]$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n} R_{\alpha i}^2 - n\overline{R}_{\alpha}^2\right]$$
$$= S^2.$$

Hence, we can see that $\sum_{i=1}^{n}\sum_{j=1}^{n}(R_{\alpha i} - R_{\alpha j})^2$ is proportional to the variance of the ranked data. We can thus modify the test statistic of Mantel and Valand (1970) to obtain a measure of the variation of the observations. More specifically, for univariate data, if we take f as in (5.1.2) and if we let

$$X_{ij} = \begin{cases} 1 & W_i \text{ and } W_j \text{ are from the same class} \\ 0 & \text{otherwise.} \end{cases}$$

and

$$X_{ij}^* = \begin{cases} 1 & W_i \text{ and } W_j \text{ are not from the same class} \\ 0 & \text{otherwise} \end{cases}$$

then

$$\begin{aligned} B &= 2\sum_{i=1}^{n}\sum_{j=1}^{n} X_{ij}^* f(R_i, R_j) \\ &= 2\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{1}{2} X_{ij}^* (R_i - R_j)^2 \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} X_{ij}^* (R_i - R_j)^2 \end{aligned}$$

represents the variation between the two populations where as

$$\begin{aligned} W &= \sum_{i=1}^{n}\sum_{j=1}^{n} X_{ij} f(R_i, R_j) \\ &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} X_{ij} (R_i - R_j)^2 \end{aligned}$$

$$= \sum_{i<j} X_{ij}(R_i - R_j)^2$$

gives a measurement of the internal variation as it is the sum of the variation of the observations within each population.

We can then combine these two values and apply the idea of Fisher's discriminant function to get a measurement of separation between the two populations. Recall from Chapter 4 that Fisher measured the separation between two populations as a ratio of the distance between the groups to the internal variation of both groups. Likewise, we can now define our measure of separation as

$$\begin{aligned} \mathrm{D} &= \frac{\text{average variation between populations}}{\text{average internal variation of the populations}} \\ &= \frac{B/n_1 n_2}{W/(n_1(n_1-1)/2 + n_2(n_2-1)/2)} \\ &= \frac{n_1(n_1-1) + n_2(n_2-1)}{2n_1 n_2} \cdot \frac{B}{W} \\ &= \frac{n_1(n_1-1) + n_2(n_2-1)}{2n_1 n_2} \cdot \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} X_{ij}^{*}(R_i - R_j)^2}{\sum_{i<j} X_{ij}(R_i - R_j)^2} \end{aligned}$$

where n_1 and n_2 represent the number of observations among $W_1, \ldots, W_n$ that come from π_1 and π_2 respectively.

In the case of multivariate data, our measure of separation becomes

$$\mathrm{D} = \sum_{\alpha=1}^{p} \eta_\alpha D_\alpha$$

$$
\begin{aligned}
&= \sum_{\alpha=1}^{p} \frac{n_1(n_1-1)+n_2(n_2-1)}{2n_1n_2} \cdot \eta_\alpha \frac{B_\alpha}{W_\alpha} \\
&= \frac{n_1(n_1-1)+n_2(n_2-1)}{2n_1n_2} \sum_{\alpha=1}^{p} \eta_\alpha \frac{B_\alpha}{W_\alpha} \\
&= \frac{n_1(n_1-1)+n_2(n_2-1)}{2n_1n_2} \sum_{\alpha=1}^{p} \eta_\alpha \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} X_{ij}^{*}(R_{\alpha i}-R_{\alpha j})^2}{\sum_{i<j} X_{ij}(R_{\alpha i}-R_{\alpha j})^2}
\end{aligned}
$$

where $B_\alpha = \sum_{i=1}^{n}\sum_{j=1}^{n} X_{ij}^{*}(R_{\alpha i}-R_{\alpha j})^2$ measures the variation between the populations for the α^{th} variable, $W_\alpha = \sum_{i<j} X_{ij}(R_{\alpha i}-R_{\alpha j})^2$ is a measurement the internal variation of the populations for the α^{th} variable and η_α is a nonnegative weight. For simplicity, we will take $\eta_\alpha = 1$, $\forall\alpha$ in what follows.

5.2 A New Classification Rule

Using the measure of separation derived in the previous section,

$$
\mathrm{D} = \frac{n_1(n_1-1)+n_2(n_2-1)}{2n_1n_2} \sum_{\alpha=1}^{p} \frac{B_\alpha}{W_\alpha}
$$

we can now develop a new classification rule. Let $X_1,\ldots,X_{n_1}$ and $Y_1,\ldots,Y_{n_2}$ be two random samples of $p-$dimensional observations arising from two different populations, π_1 and π_2, respectively. Also, let Z be a new observation which we wish to classify. Equivalently, to be consistent with the notation of Section 5.1, we can let $W_1,\ldots,W_{n_1},W_{n_1+1}\ldots,W_{n_1+n_2},W_{n_1+n_2+1}$ represent the observations $X_1,\ldots,X_{n_1}$, $Y_1,\ldots,Y_{n_2}$, Z. Moreover, let $R_{\alpha i}$, which can be calculated using equation (5.1.1), be the rank of the observation which corresponds to the α^{th} variable of W_i, $\alpha = 1,\ldots,p$ and $i = 1,\ldots,n = n_1+n_2+1$. Then the classification rule is obtained as follows.

1. Calculate the separation between π_1 and π_2 assuming that Z, the new observation, comes from π_1. Call this separation D_1. Note that there are now $n_1 + 1$ observations from π_1 and n_2 observations that are from π_2 and so

$$\begin{aligned} \mathrm{D}_1 &= \frac{(n_1+1)n_1 + n_2(n_2-1)}{2(n_1+1)n_2} \sum_{\alpha=1}^{p} \frac{B_\alpha}{W_\alpha} \\ &= \frac{(n_1+1)n_1 + n_2(n_2-1)}{2(n_1+1)n_2} \sum_{\alpha=1}^{p} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} X_{ij}^{*}(R_{\alpha i} - R_{\alpha j})^2}{\sum_{i<j} X_{ij}(R_{\alpha i} - R_{\alpha j})^2} \end{aligned}$$

2. Calculate the separation between π_1 and π_2 assuming that Z, the new observation, comes from π_2. Call this separation D_2. Note that there are now n_1 observations from π_1 and $n_2 + 1$ observations that are from π_2 and so

$$\begin{aligned} \mathrm{D}_2 &= \frac{n_1(n_1-1) + (n_2+1)n_2}{2n_1(n_2+1)} \sum_{\alpha=1}^{p} \frac{B_\alpha}{W_\alpha} \\ &= \frac{n_1(n_1-1) + (n_2+1)n_2}{2n_1(n_2+1)} \sum_{\alpha=1}^{p} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} X_{ij}^{*}(R_{\alpha i} - R_{\alpha j})^2}{\sum_{i<j} X_{ij}(R_{\alpha i} - R_{\alpha j})^2} \end{aligned}$$

3. Like with Fisher's discriminant function, we seek a maximal separation between the populations. Therefore, classify Z as coming from π_1 if $D_1 \geqslant D_2$ and classify Z in π_2 otherwise.

   Hence the new classifier can be written as

$$\varphi(z) = \begin{cases} 1 & \text{if } D_1 - D_2 \geqslant 0 \\ 2 & \text{otherwise.} \end{cases}$$

Let's now consider a simple example to illustrate how this new classification procedure works.

Example 5.2.1 Let π_1 and π_2 be two univariate populations. The following table summarizes the data and their corresponding ranks.

Table 5.1: Data for Example 5.2.1

Population	Observations	Ranks
π_1	3, 7, 10	1, 2, 4
π_2	9, 14, 17	3, 5, 6
Z	18	7

We begin by assuming Z comes from π_1 and proceed to find D_1. For this univariate problem, $n_1 = 3$, $n_2 = 3$, $n = n_1 + n_2 + 1 = 7$ and so we have

$$\begin{aligned}
\mathrm{B}_1 &= \sum_{i=1}^{7}\sum_{j=1}^{7} X_{ij}^{*}(R_i - R_j)^2 \\
&= (1-3)^2 + (1-5)^2 + (1-6)^2 + (2-3)^2 + \ldots + (7-5)^2 + (7-6)^2 \\
&= 98
\end{aligned}$$

$$\begin{aligned}
\mathrm{W}_1 &= \sum_{i<j} X_{ij}(R_i - R_j)^2 \\
&= \left\{(1-2)^2 + (1-4)^2 + (1-7)^2 + (2-4)^2 + (2-7)^2 + (4-7)^2\right\} + \\
&\quad \left\{(3-5)^2 + (3-6)^2 + (5-6)^2\right\} \\
&= 84 + 14 \\
&= 98
\end{aligned}$$

and hence, by combining these two values we find that

$$\mathrm{D}_1 = \frac{(n_1+1)n_1 + n_2(n_2-1)}{2(n_1+1)n_2} \cdot \frac{B_1}{W_1}$$

$$= \frac{3}{4} \cdot \frac{98}{98}$$
$$= \frac{3}{4}$$

We can now repeat the above calculations assuming that Z comes from π_2 in order to find D_2. In this case, we have

$$\begin{aligned} B_1 &= \sum_{i=1}^{7}\sum_{j=1}^{7} X_{ij}^{*}(R_i - R_j)^2 \\ &= (1-3)^2 + (1-5)^2 + (1-6)^2 + (1-7)^2 + \ldots + (4-6)^2 + (4-7)^2 \\ &= 147 \end{aligned}$$

$$\begin{aligned} W_1 &= \sum_{i<j} X_{ij}(R_i - R_j)^2 \\ &= \left\{(1-2)^2 + (1-4)^2 + (2-4)^2\right\} + \\ &\quad \left\{(3-5)^2 + (3-6)^2 + (3-7)^2 + (5-6)^2 + (5-7)^2 + (6-7)^2\right\} \\ &= 14 + 35 \\ &= 49 \end{aligned}$$

and hence, by combining these two values we find that

$$\begin{aligned} D_2 &= \frac{n_1(n_1-1) + (n_2+1)n_2}{2n_1(n_2+1)} \cdot \frac{B_1}{W_1} \\ &= \frac{3}{4} \cdot \frac{147}{49} \\ &= \frac{9}{4} \end{aligned}$$

Therefore, the new observation Z is classified in π_2 since $D_1 - D_2 = -3/2 < 0$.

Although this was a simple example with few observations, some advantages of the new classification rule are clear. Unlike the three previously studied classification procedures, this new method does not estimate any density functions or density parameters. Hence, it can work well even without the presence of large sample sizes, as was the case in Example 5.2.1. Furthermore, the procedure is not computationally complex and can easily be implemented on a computer. We note, however, that because the new classification rule is based on ranks, it will be, like the rank transform, most effective when the populations differ in locations. Also, due to the use of ranks, the exact probability of misclassification is unfortunately very difficult to obtain for this classifier. Hence, a Monte Carlo simulation study will be conducted to further study the properties and efficiency of this new classification rule.

Chapter 6

Simulation Results

A Monte Carlo simulation study was conducted in order to test our new rank classification procedure's ability to correctly classify a new observation. Both univariate and multivariate simulations were performed. Among the multivariate simulations, many are based on a bivariate distribution. However, a special set of simulations were done in the case where each population has a higher number of dimensions and only a few observations.

For each simulation performed, the rank classifier was compared with the three other methods discussed in this thesis: the Bayes rule (with density parameters estimated from the training sample), the kernel density classifier and Fisher's classifier applied to the ranked data. In each case, 100 training samples were simulated and the performance of each classification rule was measured using a leave-one-out cross-validation technique, which provides an estimate of the true probability of misclassification. The final error rate provided in the tables is the average cross-validation error rate across all 100 samples.

Cross-validation is not the only method which can be used to estimate the probability of misclassification. Alternatively, when simulating the training samples, we can also generate additional observations which are not part of the training sample but are utilized to test the classification rule and thus estimate the error rate. As will be demonstrated with the univariate simulation results, this procedure gave results that were very similar to the leave-one-out cross-validation technique.

In all, over 170 different simulations were performed. The results indicate that the new rank classification procedure is most effective when the populations have many variables and few observations. In this case, the new classification rule performed better than all three other classifiers 87% of the time. Further details as well as all simulation results are provided in the following sections. Moreover, an analysis of the results and a comparison of the four classification rules can be found in Chapter 7.

6.1 Univariate Simulation Results

Three sets of univariate simulations were conducted. The first set were done under the assumption of equal sample sizes and the results can be found in Table 6.1. For the second set, presented in Table 6.2, the populations had an unequal number of observations in the training sample. In both cases, leave-one-out cross-validation was used to estimate the error rate. The simulations of Table 6.1 were then repeated using an alternate method to estimate the probability of misclassification. For these simulations, found in Table 6.3, each training sample consisted of 100 observations, where 50 observations were generated from each population. Additionally, an extra 50 observations from each population were then simulated and utilized to test the classification rule.

For all three groups of simulations, the samples of observations were obtained using various density functions. The normal, Cauchy, exponential and logistic distributions were all utilized for f_i, where f_i is the density of π_i, $i = 1, 2$. Simulations were then done under four different conditions:

1. f_1 and f_2 are of the same family of distributions and have equal scale parameters but unequal location parameters;

2. f_1 and f_2 are of the same family of distributions and have unequal location and scale parameters;

3. f_1 and f_2 are of different families of distributions and have equal scale parameters but unequal location parameters;

4. f_1 and f_2 are of different families of distributions and have unequal location and scale parameters.

The following abbreviations are utilized for the density functions:
$N(\mu, \sigma^2)$ is a normal distribution with mean $= \mu$ and variance $= \sigma^2$;
$C(\mu, \sigma)$ is a Cauchy distribution with location $= \mu$ and scale $= \sigma$;
$Log(\mu, \sigma)$ is a logistic distribution with location $= \mu$ and scale $= \sigma$;
$Exp(\mu, \sigma)$ is an exponential distribution with location $= \mu$ and scale $= \sigma$.

Table 6.1: Univariate Simulations: $n_1 = 50, n_2 = 50$

Population Density		Cross-Validation Error Rate (%)			
Population 1 π_1	Population 2 π_2	Bayes Rule	Kernel Density	Rank Transform	New Rank Method
$N(0,1)$	$N(2,1)$	16.13	16.51	16.7	16.19
$N(0,1)$	$N(2,2)$	24.89	24.06	24.44	24.98
$C(0,1)$	$C(2,1)$	26.06	26.79	25.22	48.39
$C(0,1)$	$C(2,2)$	28.61	31.39	30.98	48.85
$Exp(0,1)$	$Exp(2,1)$	7.26	9.43	12.52	9.45
$Exp(0,1)$	$Exp(2,2)$	7.35	8.27	10.44	22.3
$Log(0,1)$	$Log(2,1)$	26.19	27.66	26.92	27.05
$Log(0,1)$	$Log(2,2)$	30.43	31.08	33.62	36.91
$C(0,1)$	$Log(2,1)$	24.5	25.62	26.12	47.1
$C(0,1)$	$Log(3,2)$	24.83	26.34	25.94	43.79
$Log(0,1)$	$N(2,1)$	21.17	20.87	22.12	23.28
$Log(0,3)$	$N(3,1)$	14.75	15.83	30.34	36.77
$C(0,1)$	$N(1,1)$	29.29	30.22	33.64	48.53
$C(0,1)$	$N(3,2)$	22.0	19.76	20.36	43.39

Table 6.2: Univariate Simulations: $n_1 = 25, n_2 = 50$

Population Density		Cross-Validation Error Rate (%)			
Population 1 π_1	Population 2 π_2	Bayes Rule	Kernel Density	Rank Transform	New Rank Method
$N(0,1)$	$N(2,1)$	15.07	15.51	18.55	14.33
$N(0,1)$	$N(2,2)$	29.73	28.4	25.27	27.65
$C(0,1)$	$C(2,1)$	25.05	26.37	26.51	49.91
$C(0,1)$	$C(2,2)$	29.55	34.77	31.4	55.33
$Exp(0,1)$	$Exp(2,1)$	5.63	6.24	5.91	17.04
$Exp(0,1)$	$Exp(2,2)$	5.52	6.03	16.19	17.39
$Log(0,1)$	$Log(2,1)$	25.89	28.6	27.64	24.45
$Log(0,1)$	$Log(2,2)$	32.6	33.89	33.92	44.28
$C(0,1)$	$Log(2,1)$	25.17	26.03	27.01	28.6
$C(0,1)$	$Log(3,2)$	27.15	29.01	27.55	31.24
$Log(0,1)$	$N(2,1)$	18.52	18.29	24.23	17.19
$Log(0,3)$	$N(3,1)$	12.48	12.24	32.05	21.36
$C(0,1)$	$N(1,1)$	26.93	27.11	35.48	29.24
$C(0,1)$	$N(3,2)$	23.16	20.17	23.43	26.12

Table 6.3: Univariate Simulations: Alternate Error Rate

Population Density		Estimated Error Rate (%)			
Population 1 π_1	Population 2 π_2	Bayes Rule	Kernel Density	Rank Transform	New Rank Method
$N(0,1)$	$N(2,1)$	16.16	16.33	16.32	16.32
$N(0,1)$	$N(2,2)$	24.94	23.63	25.16	24.93
$C(0,1)$	$C(2,1)$	24.51	26.06	24.76	48.82
$C(0,1)$	$C(2,2)$	28.55	31.07	31.8	48.98
$Exp(0,1)$	$Exp(2,1)$	7.8	9.43	12.86	10.11
$Exp(0,1)$	$Exp(2,2)$	7.62	8.37	11.47	21.39
$Log(0,1)$	$Log(2,1)$	27.53	27.75	27.41	27.76
$Log(0,1)$	$Log(2,2)$	29.97	30.36	33.59	36.23
$C(0,1)$	$Log(2,1)$	23.65	24.69	26.11	47.42
$C(0,1)$	$Log(3,2)$	25.27	27.11	26.94	42.42
$Log(0,1)$	$N(2,1)$	21.15	21.1	22.52	23.31
$Log(0,3)$	$N(3,1)$	15.39	15.21	31.29	37.15
$C(0,1)$	$N(1,1)$	28.91	29.68	33.83	48.64
$C(0,1)$	$N(3,2)$	22.36	19.99	21.27	43.77

6.2 Multivariate Simulation Results

In the case of multivariate data, two sets of simulations were also conducted. Just like the univariate case, the first set of simulations, given in Table 6.4, were performed with populations that had an equal number of observations whereas the second set of simulations were done with populations of unequal sample sizes and can be found in Table 6.5. However, unlike the univariate case, for both sets of simulations, only the bivariate normal density was used for f_1 and f_2 and simulations were done under two different conditions:

1. f_1 and f_2 have unequal means but equal covariance matrices;
2. f_1 and f_2 have unequal means and unequal covariance matrices.

In order to simplify notation in Tables 6.4 and 6.5, the means utilized in the simulations are indicated below.

$$\mu_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \mu_2 = \begin{bmatrix} 2 \\ 2 \end{bmatrix}, \mu_3 = \begin{bmatrix} 1 \\ -3 \end{bmatrix}, \text{ and } \mu_4 = \begin{bmatrix} 5 \\ -5 \end{bmatrix}$$

Table 6.4: Multivariate Simulations: $n_1 = 50, n_2 = 50$

Multivariate Normal Density Parameters		Cross-Validation Error Rate (%)			
Population 1 π_1	Population 2 π_2	Bayes Rule	Kernel Density	Rank Transform	New Rank Method
$\mu_1, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\mu_2, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	7.84	8.77	8.29	8.42
$\mu_1, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\mu_3, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	5.97	6.45	6.09	6.07
$\mu_1, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\mu_4, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	0.04	0.07	0.03	0.06

Continued on next page...

Table 6.4 – continued from previous page

Multivariate Normal Density Parameters		Cross-Validation Error Rate (%)			
Population 1 π_1	Population 2 π_2	Bayes Rule	Kernel Density	Rank Transform	New Rank Method
$\mu_1, \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$	$\mu_2, \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$	2.55	2.85	2.75	3.88
$\mu_1, \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$	$\mu_3, \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$	6.6	6.81	6.24	7.43
$\mu_1, \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$	$\mu_4, \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$	0.25	0.33	0.23	0.24
$\mu_1, \Sigma = \begin{bmatrix} 4 & 3 \\ 3 & 4 \end{bmatrix}$	$\mu_2, \Sigma = \begin{bmatrix} 4 & 3 \\ 3 & 4 \end{bmatrix}$	29.83	33.3	30.15	29.34
$\mu_1, \Sigma = \begin{bmatrix} 4 & 3 \\ 3 & 4 \end{bmatrix}$	$\mu_3, \Sigma = \begin{bmatrix} 4 & 3 \\ 3 & 4 \end{bmatrix}$	7.84	9.63	9.12	18.42
$\mu_1, \Sigma = \begin{bmatrix} 4 & 3 \\ 3 & 4 \end{bmatrix}$	$\mu_4, \Sigma = \begin{bmatrix} 4 & 3 \\ 3 & 4 \end{bmatrix}$	0.0	0.07	0.1	0.39
$\mu_1, \Sigma = \begin{bmatrix} 1 & 0.3 \\ 0.3 & 2 \end{bmatrix}$	$\mu_2, \Sigma = \begin{bmatrix} 1 & 0.3 \\ 0.3 & 2 \end{bmatrix}$	14.11	13.79	13.27	14.06
$\mu_1, \Sigma = \begin{bmatrix} 1 & 0.3 \\ 0.3 & 2 \end{bmatrix}$	$\mu_3, \Sigma = \begin{bmatrix} 1 & 0.3 \\ 0.3 & 2 \end{bmatrix}$	9.63	10.81	10.44	11.14
$\mu_1, \Sigma = \begin{bmatrix} 1 & 0.3 \\ 0.3 & 2 \end{bmatrix}$	$\mu_4, \Sigma = \begin{bmatrix} 1 & 0.3 \\ 0.3 & 2 \end{bmatrix}$	0.03	0.1	0.13	0.9
$\mu_1, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\mu_2, \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$	11.69	12.24	12.97	13.64
$\mu_1, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\mu_3, \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$	9.71	10.18	10.53	10.3

Continued on next page...

Table 6.4 – continued from previous page

Multivariate Normal Density Parameters		Cross-Validation Error Rate (%)			
Population 1 π_1	Population 2 π_2	Bayes Rule	Kernel Density	Rank Transform	New Rank Method
$\mu_1, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\mu_4, \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$	0.31	0.33	0.23	0.22
$\mu_1, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\mu_2, \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$	12.74	13.69	14.45	14.66
$\mu_1, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\mu_3, \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$	11.38	12.37	12.49	12.3
$\mu_1, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\mu_4, \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$	0.33	0.42	0.44	0.41
$\mu_1, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\mu_2, \Sigma = \begin{bmatrix} 2 & -0.5 \\ -0.5 & 3 \end{bmatrix}$	11.18	11.53	13.81	14.03
$\mu_1, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\mu_3, \Sigma = \begin{bmatrix} 2 & -0.5 \\ -0.5 & 3 \end{bmatrix}$	12.78	13.29	13.08	13.1
$\mu_1, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\mu_4, \Sigma = \begin{bmatrix} 2 & -0.5 \\ -0.5 & 3 \end{bmatrix}$	0.6	0.55	0.76	0.74
$\mu_1, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\mu_2, \Sigma = \begin{bmatrix} 4 & 1 \\ 1 & 4 \end{bmatrix}$	15.89	16.56	19.49	19.71
$\mu_1, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\mu_3, \Sigma = \begin{bmatrix} 4 & 1 \\ 1 & 4 \end{bmatrix}$	9.92	11.95	14.52	13.17
$\mu_1, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\mu_4, \Sigma = \begin{bmatrix} 4 & 1 \\ 1 & 4 \end{bmatrix}$	0.53	0.81	0.81	0.85
$\mu_1, \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$	$\mu_2, \Sigma = \begin{bmatrix} 5 & -1 \\ -1 & 5 \end{bmatrix}$	11.27	12.41	15.62	18.19
$\mu_1, \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$	$\mu_3, \Sigma = \begin{bmatrix} 5 & -1 \\ -1 & 5 \end{bmatrix}$	11.8	12.94	18.35	17.23

Continued on next page...

Table 6.4 – continued from previous page

Multivariate Normal Density Parameters		Cross-Validation Error Rate (%)			
Population 1 π_1	Population 2 π_2	Bayes Rule	Kernel Density	Rank Transform	New Rank Method
$\mu_1, \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$	$\mu_4, \Sigma = \begin{bmatrix} 5 & -1 \\ -1 & 5 \end{bmatrix}$	2.4	2.49	3.94	3.69
$\mu_1, \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$	$\mu_2, \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$	6.76	7.36	7.72	8.28
$\mu_1, \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$	$\mu_3, \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$	5.65	5.82	6.09	6.45
$\mu_1, \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$	$\mu_4, \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$	0.01	0.03	0.12	0.12
$\mu_1, \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$	$\mu_2, \Sigma = \begin{bmatrix} 3 & 2 \\ 2 & 4 \end{bmatrix}$	13.9	14.89	15.78	15.68
$\mu_1, \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$	$\mu_3, \Sigma = \begin{bmatrix} 3 & 2 \\ 2 & 4 \end{bmatrix}$	9.23	9.13	14.9	14.28
$\mu_1, \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$	$\mu_4, \Sigma = \begin{bmatrix} 3 & 2 \\ 2 & 4 \end{bmatrix}$	0.13	0.23	0.97	0.93
$\mu_1, \Sigma = \begin{bmatrix} 4 & 2 \\ 2 & 3 \end{bmatrix}$	$\mu_2, \Sigma = \begin{bmatrix} 5 & -1 \\ -1 & 5 \end{bmatrix}$	22.27	26.71	27.07	26.96
$\mu_1, \Sigma = \begin{bmatrix} 4 & 2 \\ 2 & 3 \end{bmatrix}$	$\mu_3, \Sigma = \begin{bmatrix} 5 & -1 \\ -1 & 5 \end{bmatrix}$	18.82	22.47	19.63	20.86
$\mu_1, \Sigma = \begin{bmatrix} 4 & 2 \\ 2 & 3 \end{bmatrix}$	$\mu_4, \Sigma = \begin{bmatrix} 5 & -1 \\ -1 & 5 \end{bmatrix}$	2.82	3.92	3.6	3.64
$\mu_1, \Sigma = \begin{bmatrix} 4 & 2 \\ 2 & 3 \end{bmatrix}$	$\mu_2, \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$	17.91	19.4	21.43	20.59

Continued on next page...

Table 6.4 – continued from previous page

Multivariate Normal Density Parameters		Cross-Validation Error Rate (%)			
Population 1 π_1	Population 2 π_2	Bayes Rule	Kernel Density	Rank Transform	New Rank Method
$\mu_1, \Sigma = \begin{bmatrix} 4 & 2 \\ 2 & 3 \end{bmatrix}$	$\mu_3, \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$	5.11	5.87	7.17	10.96
$\mu_1, \Sigma = \begin{bmatrix} 4 & 2 \\ 2 & 3 \end{bmatrix}$	$\mu_4, \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$	0.03	0.03	0.03	0.1
$\mu_1, \Sigma = \begin{bmatrix} 4 & 2 \\ 2 & 3 \end{bmatrix}$	$\mu_2, \Sigma = \begin{bmatrix} 3 & 2 \\ 2 & 4 \end{bmatrix}$	27.91	32.4	27.7	27.64
$\mu_1, \Sigma = \begin{bmatrix} 4 & 2 \\ 2 & 3 \end{bmatrix}$	$\mu_3, \Sigma = \begin{bmatrix} 3 & 2 \\ 2 & 4 \end{bmatrix}$	11.8	13.54	12.6	17.39
$\mu_1, \Sigma = \begin{bmatrix} 4 & 2 \\ 2 & 3 \end{bmatrix}$	$\mu_4, \Sigma = \begin{bmatrix} 3 & 2 \\ 2 & 4 \end{bmatrix}$	0.33	0.36	0.22	0.59

Table 6.5: Multivariate Simulations: $n_1 = 25, n_2 = 50$

Multivariate Normal Density Parameters		Cross-Validation Error Rate (%)			
Population 1 π_1	Population 2 π_2	Bayes Rule	Kernel Density	Rank Transform	New Rank Method
$\mu_1, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\mu_2, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	7.88	8.61	9.65	10.15
$\mu_1, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\mu_3, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	5.16	6.61	8.85	9.64
$\mu_1, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\mu_4, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	0.04	0.03	2.49	3.09
$\mu_1, \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$	$\mu_2, \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$	2.11	2.8	3.36	5.65
$\mu_1, \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$	$\mu_3, \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$	6.55	7.45	10.39	11.48
$\mu_1, \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$	$\mu_4, \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$	0.25	0.49	4.93	6.01
$\mu_1, \Sigma = \begin{bmatrix} 4 & 3 \\ 3 & 4 \end{bmatrix}$	$\mu_2, \Sigma = \begin{bmatrix} 4 & 3 \\ 3 & 4 \end{bmatrix}$	30.04	33.17	31.09	28.32
$\mu_1, \Sigma = \begin{bmatrix} 4 & 3 \\ 3 & 4 \end{bmatrix}$	$\mu_3, \Sigma = \begin{bmatrix} 4 & 3 \\ 3 & 4 \end{bmatrix}$	7.55	10.63	9.45	18.56
$\mu_1, \Sigma = \begin{bmatrix} 4 & 3 \\ 3 & 4 \end{bmatrix}$	$\mu_4, \Sigma = \begin{bmatrix} 4 & 3 \\ 3 & 4 \end{bmatrix}$	0.03	0.05	0.15	1.19
$\mu_1, \Sigma = \begin{bmatrix} 1 & 0.3 \\ 0.3 & 2 \end{bmatrix}$	$\mu_2, \Sigma = \begin{bmatrix} 1 & 0.3 \\ 0.3 & 2 \end{bmatrix}$	13.39	16.17	15.21	14.61
$\mu_1, \Sigma = \begin{bmatrix} 1 & 0.3 \\ 0.3 & 2 \end{bmatrix}$	$\mu_3, \Sigma = \begin{bmatrix} 1 & 0.3 \\ 0.3 & 2 \end{bmatrix}$	10.47	11.65	11.16	13.39

Continued on next page...

Table 6.5 – continued from previous page

Multivariate Normal Density Parameters		Cross-Validation Error Rate (%)			
Population 1 π_1	Population 2 π_2	Bayes Rule	Kernel Density	Rank Transform	New Rank Method
$\mu_1, \Sigma = \begin{bmatrix} 1 & 0.3 \\ 0.3 & 2 \end{bmatrix}$	$\mu_4, \Sigma = \begin{bmatrix} 1 & 0.3 \\ 0.3 & 2 \end{bmatrix}$	0.03	0.05	1.59	2.0
$\mu_1, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\mu_2, \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$	11.12	14.23	13.93	12.67
$\mu_1, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\mu_3, \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$	9.53	10.91	11.77	12.39
$\mu_1, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\mu_4, \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$	0.23	0.32	2.12	3.39
$\mu_1, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\mu_2, \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$	11.56	14.93	15.19	13.36
$\mu_1, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\mu_3, \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$	10.93	14.67	13.57	13.8
$\mu_1, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\mu_4, \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$	0.31	0.61	2.71	3.47
$\mu_1, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\mu_2, \Sigma = \begin{bmatrix} 2 & -0.5 \\ -0.5 & 3 \end{bmatrix}$	9.85	12.68	12.52	12.32
$\mu_1, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\mu_3, \Sigma = \begin{bmatrix} 2 & -0.5 \\ -0.5 & 3 \end{bmatrix}$	11.59	15.55	14.91	13.88
$\mu_1, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\mu_4, \Sigma = \begin{bmatrix} 2 & -0.5 \\ -0.5 & 3 \end{bmatrix}$	0.57	0.87	3.51	3.47
$\mu_1, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\mu_2, \Sigma = \begin{bmatrix} 4 & 1 \\ 1 & 4 \end{bmatrix}$	13.69	19.77	20.97	14.92
$\mu_1, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\mu_3, \Sigma = \begin{bmatrix} 4 & 1 \\ 1 & 4 \end{bmatrix}$	9.05	13.63	15.28	13.29

Continued on next page...

Table 6.5 – continued from previous page

Multivariate Normal Density Parameters		Cross-Validation Error Rate (%)			
Population 1 π_1	Population 2 π_2	Bayes Rule	Kernel Density	Rank Transform	New Rank Method
$\mu_1, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\mu_4, \Sigma = \begin{bmatrix} 4 & 1 \\ 1 & 4 \end{bmatrix}$	0.65	0.73	1.79	3.4
$\mu_1, \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$	$\mu_2, \Sigma = \begin{bmatrix} 5 & -1 \\ -1 & 5 \end{bmatrix}$	9.67	15.11	16.91	12.41
$\mu_1, \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$	$\mu_3, \Sigma = \begin{bmatrix} 5 & -1 \\ -1 & 5 \end{bmatrix}$	10.43	17.0	18.84	14.87
$\mu_1, \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$	$\mu_4, \Sigma = \begin{bmatrix} 5 & -1 \\ -1 & 5 \end{bmatrix}$	1.96	2.85	5.48	6.16
$\mu_1, \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$	$\mu_2, \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$	7.64	8.05	12.35	11.41
$\mu_1, \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$	$\mu_3, \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$	5.19	5.71	6.0	9.08
$\mu_1, \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$	$\mu_4, \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$	0.07	0.05	0.39	0.48
$\mu_1, \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$	$\mu_2, \Sigma = \begin{bmatrix} 3 & 2 \\ 2 & 4 \end{bmatrix}$	11.77	18.4	18.8	12.51
$\mu_1, \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$	$\mu_3, \Sigma = \begin{bmatrix} 3 & 2 \\ 2 & 4 \end{bmatrix}$	8.73	12.04	13.35	13.84
$\mu_1, \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$	$\mu_4, \Sigma = \begin{bmatrix} 3 & 2 \\ 2 & 4 \end{bmatrix}$	0.21	0.25	0.39	1.24
$\mu_1, \Sigma = \begin{bmatrix} 4 & 2 \\ 2 & 3 \end{bmatrix}$	$\mu_2, \Sigma = \begin{bmatrix} 5 & -1 \\ -1 & 5 \end{bmatrix}$	23.2	26.79	26.16	24.64

Continued on next page...

Table 6.5 – continued from previous page

Multivariate Normal Density Parameters		Cross-Validation Error Rate (%)			
Population 1 π_1	Population 2 π_2	Bayes Rule	Kernel Density	Rank Transform	New Rank Method
$\mu_1, \Sigma = \begin{bmatrix} 4 & 2 \\ 2 & 3 \end{bmatrix}$	$\mu_3, \Sigma = \begin{bmatrix} 5 & -1 \\ -1 & 5 \end{bmatrix}$	22.04	23.88	22.12	22.61
$\mu_1, \Sigma = \begin{bmatrix} 4 & 2 \\ 2 & 3 \end{bmatrix}$	$\mu_4, \Sigma = \begin{bmatrix} 5 & -1 \\ -1 & 5 \end{bmatrix}$	3.13	3.41	6.8	7.57
$\mu_1, \Sigma = \begin{bmatrix} 4 & 2 \\ 2 & 3 \end{bmatrix}$	$\mu_2, \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$	20.2	16.47	23.65	23.44
$\mu_1, \Sigma = \begin{bmatrix} 4 & 2 \\ 2 & 3 \end{bmatrix}$	$\mu_3, \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$	4.21	5.15	9.61	10.93
$\mu_1, \Sigma = \begin{bmatrix} 4 & 2 \\ 2 & 3 \end{bmatrix}$	$\mu_4, \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$	0.0	0.03	0.65	0.72
$\mu_1, \Sigma = \begin{bmatrix} 4 & 2 \\ 2 & 3 \end{bmatrix}$	$\mu_2, \Sigma = \begin{bmatrix} 3 & 2 \\ 2 & 4 \end{bmatrix}$	28.35	31.41	28.28	26.61
$\mu_1, \Sigma = \begin{bmatrix} 4 & 2 \\ 2 & 3 \end{bmatrix}$	$\mu_3, \Sigma = \begin{bmatrix} 3 & 2 \\ 2 & 4 \end{bmatrix}$	13.0	14.41	13.67	18.29
$\mu_1, \Sigma = \begin{bmatrix} 4 & 2 \\ 2 & 3 \end{bmatrix}$	$\mu_4, \Sigma = \begin{bmatrix} 3 & 2 \\ 2 & 4 \end{bmatrix}$	0.28	1.73	0.69	0.44

6.3 High Dimension Simulations with Few Observations

A special set of multivariate simulations were performed in the case where the populations have many variables and few observations. The observations of the simulations presented in Table 6.6 were obtained using a 5-dimensional multivariate normal density. Furthermore, equal sample sizes of 10 observations were utilized for the various simulations.

Again, to simplify the notation in Table 6.6, the parameters of the multivariate normal distribution utilized in these simulations are provided below.

The means are

$$\mu_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \mu_2 = \begin{bmatrix} 2 \\ 2 \\ 2 \\ 2 \\ 2 \end{bmatrix}, \mu_3 = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix}, \mu_4 = \begin{bmatrix} 1 \\ -3 \\ 1 \\ -3 \\ 0 \end{bmatrix}, \text{ and } \mu_5 = \begin{bmatrix} 5 \\ -5 \\ 0 \\ 0 \\ 2 \end{bmatrix}.$$

The variances are

$$\Sigma_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 2 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 3 \end{bmatrix}, \Sigma_3 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 0 & 5 \end{bmatrix},$$

$$\Sigma_4 = \begin{bmatrix} 3 & 0 & 2 & 2 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 2 & 1 & 9 & -2 & 0 \\ 2 & 0 & -2 & 4 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \text{ and } \Sigma_5 = \begin{bmatrix} 4 & -1 & \frac{1}{2} & -\frac{1}{2} & 0 \\ -1 & 3 & 1 & -1 & 0 \\ \frac{1}{2} & 1 & 6 & 1 & -1 \\ -\frac{1}{2} & -1 & 1 & 4 & 0 \\ 0 & 0 & -1 & 0 & 2 \end{bmatrix}.$$

Table 6.6: Multivariate Simulations: $n_1 = 10, n_2 = 10$

Multivariate Normal Density Parameters		Cross-Validation Error Rate (%)			
Population 1 π_1	Population 2 π_2	Bayes Rule	Kernel Density	Rank Transform	New Rank Method
μ_1, Σ_1	μ_2, Σ_1	3.15	3.45	3.5	2.45
μ_1, Σ_1	μ_3, Σ_1	0.25	0.0	0.6	0.0
μ_1, Σ_1	μ_4, Σ_1	3.05	3.4	3.35	1.45
μ_1, Σ_1	μ_5, Σ_1	0.1	0.1	1.15	0.15
μ_1, Σ_2	μ_2, Σ_2	12.6	17.35	14.45	13.2
μ_1, Σ_2	μ_3, Σ_2	4.65	4.4	4.4	2.75
μ_1, Σ_2	μ_4, Σ_2	13.25	15.45	13.05	10.75
μ_1, Σ_2	μ_5, Σ_2	1.85	1.15	2.35	0.6
μ_1, Σ_3	μ_2, Σ_3	12.15	16.95	12.35	11.65
μ_1, Σ_3	μ_3, Σ_3	6.0	6.6	6.1	4.4
μ_1, Σ_3	μ_4, Σ_3	12.7	16.8	12.65	11.45
μ_1, Σ_3	μ_5, Σ_3	0.55	0.8	2.05	0.2

Continued on next page...

Table 6.6 – continued from previous page

Multivariate Normal Density Parameters		Cross-Validation Error Rate (%)			
Population 1 π_1	Population 2 π_2	Bayes Rule	Kernel Density	Rank Transform	New Rank Method
μ_1, Σ_4	μ_2, Σ_4	12.1	20.75	11.0	10.0
μ_1, Σ_4	μ_3, Σ_4	0.45	1.35	2.3	0.9
μ_1, Σ_4	μ_4, Σ_4	6.4	10.85	6.95	5.3
μ_1, Σ_4	μ_5, Σ_4	0.3	0.75	1.1	0.5
μ_1, Σ_5	μ_2, Σ_5	12.5	20.1	18.65	17.3
μ_1, Σ_5	μ_3, Σ_5	2.4	2.95	2.8	1.8
μ_1, Σ_5	μ_4, Σ_5	9.8	16.7	15.85	15.15
μ_1, Σ_5	μ_5, Σ_5	5.65	7.7	5.55	4.0
μ_1, Σ_1	μ_2, Σ_2	19.25	9.95	11.35	9.5
μ_1, Σ_1	μ_3, Σ_2	7.1	1.1	1.75	1.05
μ_1, Σ_1	μ_4, Σ_2	16.65	7.75	7.6	6.0
μ_1, Σ_1	μ_5, Σ_2	5.1	0.6	2.0	0.15
μ_1, Σ_1	μ_2, Σ_3	14.89	8.4	8.9	7.65
μ_1, Σ_1	μ_3, Σ_3	8.0	1.9	2.6	1.8
μ_1, Σ_1	μ_4, Σ_3	17.3	8.9	9.15	7.05
μ_1, Σ_1	μ_5, Σ_3	3.65	0.2	1.55	0.15
μ_1, Σ_1	μ_2, Σ_4	13.95	9.0	9.85	7.75
μ_1, Σ_1	μ_3, Σ_4	3.05	1.6	2.75	1.1

Continued on next page...

Table 6.6 – continued from previous page

Multivariate Normal Density Parameters		Cross-Validation Error Rate (%)			
Population 1 π_1	Population 2 π_2	Bayes Rule	Kernel Density	Rank Transform	New Rank Method
μ_1, Σ_1	μ_4, Σ_4	11.65	6.35	8.0	5.8
μ_1, Σ_1	μ_5, Σ_4	3.4	1.2	1.8	0.35
μ_1, Σ_1	μ_2, Σ_5	15.75	10.1	12.5	13.4
μ_1, Σ_1	μ_3, Σ_5	6.0	1.7	2.6	1.3
μ_1, Σ_1	μ_4, Σ_5	12.45	7.35	9.6	8.75
μ_1, Σ_1	μ_5, Σ_5	8.2	3.0	3.0	1.3
μ_1, Σ_2	μ_2, Σ_3	24.8	15.45	14.3	13.35
μ_1, Σ_2	μ_3, Σ_3	11.5	3.8	3.8	3.0
μ_1, Σ_2	μ_4, Σ_3	24.3	16.0	12.7	11.05
μ_1, Σ_2	μ_5, Σ_3	6.0	1.1	1.7	0.45
μ_1, Σ_2	μ_2, Σ_4	20.7	16.75	16.05	15.6
μ_1, Σ_2	μ_3, Σ_4	7.75	2.1	3.85	2.1
μ_1, Σ_2	μ_4, Σ_4	16.7	11.7	11.45	9.95
μ_1, Σ_2	μ_5, Σ_4	5.6	0.95	2.95	0.55
μ_1, Σ_2	μ_2, Σ_5	22.45	16.35	15.4	14.6
μ_1, Σ_2	μ_3, Σ_5	10.5	3.05	3.5	1.85
μ_1, Σ_2	μ_4, Σ_5	22.7	15.2	15.3	13.8

Continued on next page...

Table 6.6 – continued from previous page

Multivariate Normal Density Parameters		Cross-Validation Error Rate (%)			
Population 1 π_1	Population 2 π_2	Bayes Rule	Kernel Density	Rank Transform	New Rank Method
μ_1, Σ_2	μ_5, Σ_5	12.6	4.75	4.7	2.35
μ_1, Σ_3	μ_2, Σ_4	18.4	15.65	15.8	13.15
μ_1, Σ_3	μ_3, Σ_4	8.65	3.4	4.9	2.85
μ_1, Σ_3	μ_4, Σ_4	15.45	11.35	11.05	9.8
μ_1, Σ_3	μ_5, Σ_4	3.35	0.85	2.05	0.5
μ_1, Σ_3	μ_2, Σ_5	21.5	17.15	15.95	15.85
μ_1, Σ_3	μ_3, Σ_5	11.65	4.1	4.9	3.95
μ_1, Σ_3	μ_4, Σ_5	21.95	16.35	14.55	15.1
μ_1, Σ_3	μ_5, Σ_5	12.0	4.65	5.0	2.6
μ_1, Σ_4	μ_2, Σ_5	16.65	16.0	14.5	13.0
μ_1, Σ_4	μ_3, Σ_5	6.45	2.1	3.4	1.5
μ_1, Σ_4	μ_4, Σ_5	16.4	14.35	12.3	11.45
μ_1, Σ_4	μ_5, Σ_5	9.5	3.55	2.25	4.15

Chapter 7

An Analysis and Comparison of Classification Rules

There exist many different classification procedures and each one has its own advantages and limitations. Selecting the best classification rule is not always an easy task. Unfortunately, there does not exist one method that is best for all situations. The selection of the ideal classification rule for a given situation will depend upon the information available about the training sample, the number of variables and observations as well as the computational resources available. The present chapter attempts to analyze the simulation results presented in Chapter 6 and thus compare the new rank classification procedure proposed in this thesis with three other commonly used techniques: the Bayes rule, the kernel density classifier and Fisher's classifier applied to the ranked data.

To begin, Chapter 2 states that the Bayes rule, which uses estimates of the parameters, is the best classification method provided that n_1 and n_2 are large enough to produce good estimates. This was clearly supported by the Monte Carlo simulation study. The results in Tables 6.1 - 6.5 all indicate that the Bayes rule has the

lowest probability of misclassification. Also, as expected, Table 6.6 shows that when the number of observations are insufficient to generate adequate estimates of the parameters, the Bayes rule is no longer optimal.

The simulation results further suggest that the Bayes rule performs well for all underlying distributions. In Tables 6.1 and 6.2, the Bayes classifier had the lowest error rate when the population densities were normal, Cauchy, exponential, logistic or a mixture of these. The only problem with the optimal classification rule is that in order to be applicable, it requires knowledge of these population densities and in practice, such information is rarely known.

It is for this reason that nonparametric methods are quite popular and widely researched. Nonparametric classification techniques do not require previous knowledge about the population densities and the simulation study demonstrates that a good classifier can be obtained without this information. In fact, the simulation results show that the kernel density classifier is a competitive nonparametric alternative to the Bayes rule. Not only does it consistently have a comparable misclassification error rate but the classification approach based on kernel density estimation works equally well with any underlying population densities.

Moreover, Table 6.6 indicates that for situations of higher dimensions and fewer observations where the populations have unequal covariance matrices, the kernel density classifier often had a smaller probability of misclassification than the Bayes rule. Nevertheless, it should be noted that for the simulations presented in Table 6.6, the kernel density classifier was not the best classification method and as the dimensions continue to increase while the number of observations remain small, Silverman (1986) stipulates that the misclassification error rate will only get worse for this method.

Another problem with the kernel density classifier is that it requires the selection of the kernel and the bandwidth parameter, h. As discussed in Chapter 3, there are numerous kernel functions available as well as various techniques to select the bandwidth parameter. The performance of the classifier is strongly influenced by these factors and will thus depend upon selecting the best kernel and bandwidth for a given situation. In our simulation study, the densities were all unimodal and often normal. Hence, the selection of the bandwidth parameter, detailed in Chapter 3, was often optimal in terms of the mean integrated square error. Likewise, using the standard normal distribution as the kernel seems reasonable in such situations. However, without such prior information about the densities, there is no systematic way of knowing which kernel and bandwidth parameter will be ideal for a particular case.

Unlike the kernel density classifier, Fisher's classifier applied to the ranks of the observations necessitates no such selection of functions or parameters. It is one of the simplest classification procedures based on ranks and is one of the least computationally intensive methods to implement. The simulation results in Chapter 6 indicate that it often has a comparative error rate to the kernel density classifier. In the cases when the populations are bivariate normal with a common covariance matrix, this classification rule often had a slightly smaller error rate than the kernel density classifier. This is not surprising since in such cases, Fisher's classification rule has been shown, in Chapter 4, to be asymptotically optimal and so we would expect it to perform equally well when applied to the ranks of the observations.

We note that Ćwik and Mielniczuk (1995) arrived at the same conclusions. However, they also demonstrated that under severe departures from normality, Fisher's classifier applied to the ranked data did not perform well and that kernel density estimation techniques are a better nonparametric classification approach in such cases. Finally, according to our simulations, Fisher's classification based on ranks was not

the optimal method when there were many variables and few observations. The results in Table 6.6 are mainly due to the fact that like the Bayes rule, Fisher's classification method requires reliable estimates of the means and variances of the ranks of the observations.

Like Fisher's classifier, the new classification method based on ranks does not require the selection of any parameters or functions. Also, it is straightforward to implement and not computationally exhaustive. The Monte Carlo simulation study demonstrated that the method performs reasonably well. Although for the univariate and bivariate simulations it was not consistently the best nonparametric method, its error rate was often comparable to that of Fisher's classifier based on ranks. The simulations equally revealed that the method is most effective when the data has a large number of variables and few observations. In these cases, this new classification technique had the lowest probability of misclassification, among all four classification procedures, nearly 87% of the time.

This results mostly from the fact that this classifier does not need to estimate any population parameters or densities. Instead, it measures the average variability of the ranks when the new observation is in π_1 as well as in π_2 and compares these two values. This computation does not require a large sample size nor does it require any previous knowledge about the population densities. It does however prevent the method from being adequate when the internal variation of the population is high. This can be clearly seen by looking at all simulations in Tables 6.1 and 6.2 that involve the Cauchy distribution. In these situations, the new classification approach had a very high error rate.

Additionally, the new classification rule based on ranks assumes a difference in the location of the populations. As stated by Akritas (1990) and Thompson (1991), any procedure which deals with ranks will perform best when there is a change in location and is not as useful if only a change in scale is present. Nevertheless, the assumption of a difference in location is verified in many practical cases and so the rank classification is readily applicable to any situation in which the number of variables is large but only a few observations are available.

Finally, we end our analysis of the various classification rules by briefly discussing classification and regression trees and comparing this approach with our new classification rule. Classification trees are a more recent classification technique proposed by Breiman, Friedman, Olshen and Stone (1993) and this particular method of solving classification problems differs from those presented in this thesis.

The general idea is to construct a classification tree based on the training sample. To do so, no information about the population densities is required. The observations are initially split into two groups and each group is subsequently split into subgroups. The splitting criterion at each step is based on the various variables that have been measured on the observations and should lead to a maximal separation between the groups. This splitting process continues until a suitable stopping criterion is achieved. The class membership of each final subgroup is determined by the class membership of the majority of the observations falling into that particular final subgroup.

Breiman, Friedman, Olshen and Stone (1993) showed that classification trees are consistent. That is, their total probability of misclassification will converge to the TPM of the Bayes rule when the number of observations in the training sample is very large. Also, this type of classification can accommodate more complex data. For instance, classification trees can be used with categorical data or when the observations

have different dimensions. Under these circumstances, the four methods discussed in this thesis cannot be utilized since these classifiers require numerical data. Therefore, it is truly in these situations that classification and regression trees are most useful.

However, it should be noted that classification trees can quickly become quite complex and can be very computationally intensive. Moreover, Breiman, Friedman, Olshen and Stone (1993) state that many observations are required in order to find the best splitting criteria and hence obtain an efficient classifier. Thus, when the observations are numerical, we can see that the new classification rule based on ranks will be more effective, in some situations, than a classification tree. More precisely, even without conducting simulations, it is clear that the new classification rule will not only be less computationally demanding but will, again, have a lower probability of misclassification when the number of variables are high and only a few observations are available. This is simply because in this case, there will not be enough information to build an effective classification tree.

Chapter 8

Conclusion

In summary, this study of classification techniques demonstrated that there does not exist one classification rule that works best in all situations. Unfortunately, many classification methods are available and selecting the best one for a given classification problem can be difficult. However, the research of this thesis did reveal that when information is given about the population densities, the ideal classifier is, without a doubt, the Bayes rule. If no such information is known and sufficiently large samples of observations are available, then the simulations indicated that the classifier based on kernel density estimation is consistently the best nonparametric approach. On the other hand, when only a few observations are present in the training sample, the new classification rule based on ranks is the most effective classifier. In such cases, it had the lowest probability of misclassification, among all four methods presented in this thesis, nearly 87% of the time. Hence, for problems where there is little data available, the new classification rule is the method of choice, especially if the dimensions are high.

Although the performance of the new classification rule was satisfactory, it would be interesting to see how effective this classifier is when a different function is selected

for $f_\alpha(R_{\alpha i}, R_{\alpha j})$. For instance, we could use

$$f_\alpha(R_{\alpha i}, R_{\alpha j}) = |R_{\alpha i} - R_{\alpha j}|$$

or any other positive symmetric function. Perhaps utilizing a different function would lower the probability of misclassification of the new classification rule in other situations, such as when the variation of the observations of a population is high. If this can be achieved, it would make the new classification rule based on ranks more widely applicable. A Monte Carlo simulation study could be conducted in order to test the new classification rule when other functions are used. Regrettably, this research was beyond the scope of this thesis. It does, nevertheless, provide good ideas for future work.

Bibliography

[1] Akritas, M.G. (1990). "The Rank Transform Method in Some Two-Factor Designs" in *Journal of the American Statistical Association*, **85**, pp. 73-78.

[2] Breiman, L., J.H. Friedman, R.A. Olshen, C.J. Stone (1993). **Classification And Regression Trees**, Chapman and Hall Ltd.

[3] Cacoullos, T. (1966). "Estimation of a multivariate density" in *Annals of the Institute of Statistical Mathematics*, **18**, pp. 179-189.

[4] Conover, W.J. and R.L. Iman. (1980). "The rank transformation as a method of discrimination with some examples" in *Communications in Statistics: Theory and Methods*, **9**, pp. 3057-3069.

[5] Cover, T.M. and P.E. Hart. (1967). "Nearest neighbor pattern classification" in *IEEE Transactions on Information Theory*, **13**, pp. 21-26.

[6] Ćwik, Jan and Jan Mielniczuk. (1995). "Nonparametric rank discrimination method" in *Computational Statistics and Data Analysis*, **19**, pp. 59-74.

[7] Das Gupta, S. and H.E. Lin. (1980). "Nearest neighbor rules for statistical classification based on ranks" in *Sankhya A*, **42**, pp. 219-230.

[8] Epanechnikov, V.A. (1969). "Nonparametric estimation of a multidimensional probability density" in *Theory of Probability and its Applications*, **14**, pp. 153-158.

[9] Fix, E. and J.L. Hodges. (1951). **Nonparametric discrimination: consistency properties**, US Air Force School of Aviation Medicine, Report No. 4 Randolph Field, Texas.

[10] Greblicki, W. and M. Pawlak. (1981). "Classification using the Fourier Series Estimate of Multivariate Density Functions" in *IEEE Transactions on Systems, Man and Cybernetics*, **11**, No. 10, pp. 726-730.

[11] Hodges, J.L. and E.L. Lehmann. (1956) "The efficiency of some nonparametric competitors of the *t*-test" in *Annals of Mathematical Statistics*, **27**, pp. 324-335.

[12] Johnson, Richard A. and Dean W. Wichern. (2002). **Applied Multivariate Statistical Analysis**, Fifth Edition, Prentice Hall, Inc.

[13] Lee, Herbert K. (2004). "Bayesian nonparametrics via neural networks" in *ASA-SIAM Series on Statistics and Applied Probability*, Society for Industrial and Applied Mathematics, Philadelphia.

[14] Mantel, Nathan and R.S. Valand. (1970). "A technique of nonparametric multivariate analysis" in *Biometrics*, **26**, No. 3, pp. 547-558.

[15] Scott, D.W. (1992). **Multivariate Density Estimation: Theory, Practice, and Visualization**, Wiley.

[16] Sheather, S.J. and M.C. Jones. (1991). "A reliable data-based bandwidth selection method for kernel density estimation" in *Journal of the Royal Statistical Society series B*, **53**, pp. 683690.

[17] Silverman, B.W. (1986). **Density Estimation for Statistics and Data Analysis**, Chapman and Hall Ltd.

[18] Thompson, G.L. (1991). "A unified approach to rank tests for multivariate and repeated measures designs" in *Journal of the American Statiscal Association*, **86**, pp. 410-419.