STATISTICAL CLASSIFICATION USING SELECTION AND RANKING METHODOLOGIES WITH STATISTICAL LEARNING

By

Jeong Jun Lee

Bachelor of Business Administration
Yonsei University
2001

Master of Arts in Business Economics
University of California, Santa Barbara
2003

Master of Arts in Applied Statistics
University of California, Santa Barbara
2004

A dissertation submitted in partial fulfillment
of the requirements for the

Doctor of Philosophy - Mathematical Sciences

Department of Mathematical Sciences
College of Sciences
The Graduate College

University of Nevada, Las Vegas
May 2024

UNLV | GRADUATE COLLEGE

Dissertation Approval

The Graduate College
The University of Nevada, Las Vegas

April 19, 2024

This dissertation prepared by

Jeong Jun Lee

entitled

Statistical Classification using Selection and Ranking Methodologies with Statistical Learning

is approved in partial fulfillment of the requirements for the degree of

Doctor of Philosophy - Mathematical Sciences
Department of Mathematical Sciences

Hokwon Cho, Ph.D.
Examination Committee Chair

Malwane Ananda, Ph.D.
Examination Committee Member

Kaushik Ghosh, Ph.D.
Examination Committee Member

Amei Amei, Ph.D.
Examination Committee Member

Jaewon Lim, Ph.D.
Graduate College Faculty Representative

Alyssa Crittenden, Ph.D.
Vice Provost for Graduate Education &
Dean of the Graduate College

Abstract

STATISTICAL CLASSIFICATION USING SELECTION AND RANKING METHODOLOGIES WITH STATISTICAL LEARNING

by

Jeong Jun Lee

Professor Hokwon Cho, Ph.D., Examination Committee Chair

University of Nevada, Las Vegas, USA

The subject of Statistical Classification is concerned with identifying and allocating future observations into one of the pre-categorized classes based on the characteristics of the objects. Typically, these decisions to classify and categorize the objects have been dependent on identifying a system of classification, and from there, determining attributes for sorting.

In past decades, from discriminant analysis, various methods have been developed for classification. In particular, the rise of artificial intelligence (AI), machine learning, and statistical learning theory has made it possible to consider improving the existing methods along with new developments and more comprehensive schemes in conjunction with data-driven methods.

In this dissertation, we propose innovation using multiple decision-theoretic perspectives, such as the Indifference-Zone (IZ) approach and the Subset Selection (SS) method, to improve and clarify how these classification decisions can be made.

Acknowledgments

It’s such an overwhelming moment to finally be able to write this part. This moment would never have been achieved without the constant encouragement, support, and guidance of Dr. Hokwon Cho, my advisor and mentor throughout the entire process. I would like to extend my sincere appreciation to him. He not only showed me how to become a researcher but taught me how to live as a human being. Meeting him in his office is always an opportunity that makes me think broadly. I am very grateful that he accepted me as his student and did not give up on me. My next commitment is to become a researcher like him.

I would like to thank my professors, Dr. Amei Amei, Dr. Malwane Ananda, and Dr. Kaushik Ghosh. They willingly participated in my dissertation committee and provided me with helpful comments and suggestions along with encouragement so that I could complete this long journey. I would also like to thank Dr. Jaewon Lim for taking the time to participate as a committee member.

I would like to express my deepest gratitude to my parents, Minkeun Lee and Mansoon Seo, who allowed me to begin this study and gave me endless support and faith. This work could not have been completed without your support.

My wife, Youkyung, deserves all the credit for this achievement. No matter how many times I let her down, her faith, confidence, support, and love never disappeared. Thanks to her sacrifice and understanding over a long time, I was able to finish successfully. My gratitude cannot be expressed in words. I thank you so much for being by my side.

I would also like to thank my colleagues Mark and Max for taking the time to read my dissertation and suggesting better ways of writing.

Dedication

For my wife, Youkyung.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

In many fields of study, namely financial, educational, and social problems, we are faced with assigning observations to one of two or more groups based on given information. For example, a department of a college sends an admission or rejection letter to an applicant who submitted the exam scores. The department already has data from previous years with all the test scores of the applicants and the corresponding group information of either admitted or rejected. Based on such data, the admission officer can make a rule to classify the applicants into one of two groups. Any procedure related to the above problem is called classification. Classification is a statistical procedure that identifies and allocates the observations into a known number of groups(or classes) based on a classifying rule or a classifier. In a classification problem, we have observations that are grouped and we want to classify a new observation into one of the known groups. Then, we need to construct a rule from the given grouped data set and apply the rule to the new observation. When we do so, we divide the data set into two parts; training data set and testing data set. After constructing the classifying rule using the training data, the rule is evaluated by checking the classification performance through the remaining data, the test data. Statistical learning takes place while finding and improving classifying rules through training data and testing data.

Since the data set for statistical classification consists of both the response variable as a class and the explanatory variables, statistical classification is referred to as supervised learning. An example of unsupervised learning is clustering, which identifies and groups data sets based on similarities, regardless of specific outcomes. The main objective of statistical classification has been focused on finding rules that assign individuals to the groups and finding better rules that improve the accuracy of the process. Moreover, most of the various classification methods to achieve this goal focus on reducing the dimensionality of the data set. Since Fisher's linear discriminant analysis, statistical classification has evolved towards finding better classifiers by reducing the dimensionality of the variables.

1.1 Motivation of the Problem

Following the advent and rise of artificial intelligence (AI) and learning theory, statistical classification can be viewed from the perspective of statistical decision theory. James O. Berger (1989) states that "Statistical decision theory is concerned with the making of decisions when in the presence of statistical knowledge (data) which sheds light on some of the uncertainties involved in the decision problem"(p. 217) [9]. The basic concept of statistical decision theory starts with an action and we quantify the gain or loss as we take the action. When we have two possible actions, we call it a binary decision problem, and when there are more than two possible actions, it is called a multiple decision problem. To solve the multiple decision problems, we use a method of multiple comparisons or the Selection and Ranking Methodologies. We will describe more about the Selection and Ranking Methodologies in Chapter 2. A statistical multiple decision-theoretic perspective on classification did not come out of nowhere. A. Wald, known as a founder of statistical decision theory, formulated to solve the classification problem of two groups by setting the hypotheses and using the Neyman-Pearson lemma in his 1944 paper [66]. Bechhofer et al. (1968) [6] approached and solved the selection problems as an identification problem using Wald's sequential analysis. Statistical multiple decision theory therefore has roots in statistical classification.

Looking at statistical classification and statistical multiple decision theory, this is both an identification process and a decision-making process. Thus, we want to combine the components of two procedures. From the perspective of statistical multiple decision theory, assigning an object to the group it belongs to will result in a correct decision, while assigning an object to the wrong group will result in a wrong decision. We could say that improving statistical classification is about finding some way to make the correct decision (correct classification) better by increasing the probability of the correct decision. In this dissertation, we are interested in the methods of improving statistical classification from a multiple decision-theoretic perspective and propose two methods that contribute to statistical classification.

One approach we propose is to introduce new variables into the classification problem. In various methods, statistical classification has been improved to find a better classifying function by reducing the dimensionality of the variables or using existing variables to create higher-order dimensions. Our first suggestion is not limited to finding such classifying functions. Instead of creating higher-order classifying functions, we strive to find variables that help separate the groups better, i.e., we add a new variable to the problem. If adding a new variable to the problem increases the separability, which is a measure of how far apart the groups are, in a higher dimensional space, one should include that variable in the problem rather than looking for a way to reduce dimensionality. As the cost of storing and extracting data has become more affordable and, thanks to the advancement of technology, incorporating a variable or vector of variables into a classification process has become much more pragmatic. We also propose a method to improve statistical classification by selecting the predictor variables that increase the separability among the existing variables. Decision makers already have a large number of variables and want to select those that will increase the relevance and decrease the redundancy. We apply the Selection and Ranking Methodologies for this process and will select the variables with a high correlation to the response variable and a low correlation to the predictor variables. The Indifference-Zone approach of the Selection and Ranking Methodologies with the multiple correlation coefficients will be used for this method.

In addition, one of the most frequently discussed areas in recent statistical studies is statistical learning theory. As mentioned above, learning can be categorized into two big parts: supervised learning and unsupervised learning. The most familiar problems in the supervised learning are the classification and the regression. Then, clustering belongs to unsupervised learning. In light of the statistical learning theory, the selection and ranking methodologies can be viewed as either supervised or unsupervised learning because the Indifference-Zone approach and the Subset Selection method are related to the classification and clustering, respectively.

Thus, in this dissertation, we investigate the application of selection problems to statistical classification and extend selection problems further to statistical learning theory. Eventually, we will examine ways to increase the probability of correct decision in classification problem.

1.2 Outline of the Thesis

We begin by reviewing statistical classification, the Selection and Ranking Methodologies, and statistical learning in Chapter 2. We review statistical classification by checking the various techniques. The probabilities that measure the accuracy of the classification process are explained along with the total probability of misclassification (TPM). A decision boundary to allocate a new observation based on the minimum TPM rule is provided. Then, we review the selection and ranking methodologies, which are the methodologies of selection in the field of statistical multiple decision theory. Here, two approaches to the selection methods are presented and we describe the formulations of both the indifference-zone approach and the subset selection method. The probability requirements to guarantee a correct decision for both approaches are illustrated. Also, a multivariate indifference-zone approach which is an extension of the univariate indifference-zone approach is described. Then, a review of Statistical Learning is provided.

In Chapter 3 a new method to improve statistical classification is proposed. If we could find a variable that separates the groups better and we examine whether the probability of a correct decision gets improved. If we attain a higher probability, we call the variable the preferable predictor vector. A simulated bivariate example is provided and the probability of correct classification and the accuracy from the confusion matrix are used to measure the performance of the classification procedure. In addition, the conditions of the preferable predictor vector are examined. It is shown that they are connected with the separability.

In Chapter 4 we investigate the indifference-zone approach from the point view of classification and statistical learning. We update the discerning measure of distance, δ, in the indifference-zone approach to improve the probability of a correct decision in the sense of statistical learning.

In Chapter 5 we suggest a method to improve statistical classification by selecting the predictor variables among the existing ones based on relevance and redundancy. We will use the indifference-zone approach using the correlation coefficients and the multiple correlation coefficients to select appropriate variables.

In Chapter 6 we conclude this dissertation with a summary and discussion of the future research.

Chapter 2

Review of the Literature and Existing Methods

2.1 Statistical Classification

In modern statistics, classification is of great interest to researchers and scientists who need to analyze rich data and make predictions based on it. Reflecting such needs, various classification methods have been developed since it was discussed by R.A. Fisher (1938) [24]. R.A. Fisher introduced a function that maximizes the separation of observations from two populations through a linear combination of their features. This linear combination transforms multivariate observations to univariate observations by maximizing the ratio of $(\frac{\text{squared distance between sample means}}{\text{sample variance}})$. He assumed that both populations had the same covariance matrix and the function was called linear discriminant.

The goal of statistical classification is to allocate a new observation to one of the groups based on the features (input variables) that are associated with observation. The groups are either known or sorted by the practitioner. The practitioner already has the data from experience and it includes the information about the group to which an observation belongs. When the group information is included, the practitioner does not need to sort them into groups and the group information is considered a response variable. Otherwise, there are data consisting of only input variables without response variable, and the practitioner needs to group them. There are data consisting of only input variables without response variables, and practitioners need to group them. The former situation can be represented by classification, and the latter is well-known as clustering or pattern recognition. The process of allocating the observations can be divided into two steps. In the first step, one uses a part of the data, the training data set, to build a rule to assign the observations from the experience to one of the known groups. The rule called a classifier (like Fisher's discriminant), is built based on the input variables through the learning process. Thus, this rule is the target function of the learning process. In the second step, we want to validate the optimized target function in the first step using the rest of the data, called the testing data set. This learning process is called supervised learning because the response variable (as the group information) exists in the data set. If the data set does not have any response variable like in the clustering case, it is categorized as unsupervised learning. In the second step of validation, the performance of classification (i.e. the accuracy of the rule) must be assessed and the misclassification rate is used as a measure of accuracy.

From Fisher's discriminant, the techniques of statistical classification have been expanded. Linear Discriminant Analysis (LDA) finds a projection to maximize the ratio of the between group variances to the within group variances under the assumption of the normal distributions and the same covariance matrix structures. Quadratic Discriminant Analysis (QDA) loosens the assumption of the identical covariance structures from LDA. K-Nearest Neighbor (KNN) classifies the data with distribution free model only depending on the distance measures. Classification and Regression Trees (CART) and Random Forest are popular Decision Tree Learning methods. Neural Networks or Artificial Neural Networks (ANNs) is based on the dynamics of connections between the nodes as the input and the output.

2.1.1 Existing Methods of Classification

Discriminant Analysis (Linear or Quadratic Discriminant Analysis)

Discriminant Analysis classifies the observations into two or more known categories with no overlapping parts. It is very similar to regression except that the response variable is categorical. It started with Fisher's idea to separate two groups by making the spread within groups small and the mean difference between groups big. Suppose $f_i(\mathbf{x})$ is the p-variate normal density with a mean vector of $\mu_{\mathbf{i}}$ and the covariance matrix of Σ_i, $i = 1, 2$. Then, the goal is to maximize the ratio of the between-groups variance to the within-groups variance. The result shows that $\mathbf{w}'\mathbf{x} = \mathbf{c}$ is the decision boundary, where $\mathbf{w}'\mathbf{x}$ is the dot product, and $\mathbf{c}$ is a threshold of the dot product, given by $\mathbf{c} = \mathbf{w}'\frac{\mathbf{1}}{\mathbf{2}}(\mu_{\mathbf{1}} + \mu_{\mathbf{2}})$. The $\mathbf{w}$ that maximizes the above ratio is $\mathbf{w} = (\mathbf{\Sigma_1} + \mathbf{\Sigma_2})^{\mathbf{-1}}(\mu_{\mathbf{1}} - \mu_{\mathbf{2}})$. Here, $\mathbf{w}$ is the projection vector onto which $\mathbf{x}$ is transformed and where the mean difference of transformed values is maximized. Also, $\mathbf{w}$ is orthogonal to the decision rule.

Linear Discriminant Analysis (LDA) assumes the same covariance matrix, $\Sigma_1 = \Sigma_2 = \Sigma$, and then $\mathbf{w} = \mathbf{\Sigma}^{\mathbf{-1}}(\mu_{\mathbf{1}} - \mu_{\mathbf{2}})$.

The decision rule on a new observation, $\mathbf{x_0}$, becomes:

Allocate $\mathbf{x_0}$ to group 1 if $(\mu_1 - \mu_2)'\Sigma^{-1}\mathbf{x_0} > \frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2)$, otherwise allocate to group 2.

This result is the same as the one using the minimum expected cost of misclassification (ECM) rule under the same cost and the same prior, that is, the classification rule of LDA is identical to the allocation rule from the minimum ECM rule, where

$$ECM = \sum_{\substack{all \\ misclassified \\ points}} \text{P(misclassification)} \times \text{ P(prior)} \times \text{(cost of misclassification)}.$$

If we assume non-homogeneity of the variance-covariance matrices, using the minimum ECM rule, we get the decision rule as:

Allocate $\mathbf{x_0}$ to group 1 if

$$-\frac{1}{2}\mathbf{x}'(\Sigma_1^{-1} - \Sigma_2^{-1})'\mathbf{x} + (\mu_1'\Sigma_1^{-1} - \mu_2'\Sigma_2^{-1})\mathbf{x} > \frac{1}{2}(\mu_1'\Sigma_1^{-1}\mu_1 - \mu_2'\Sigma_2^{-1}\mu_2) + \frac{1}{2}\log\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right).$$

Allocate $\mathbf{x_0}$ to group 2 otherwise.

This rule is calculated from the ratio of joint densities of $\mathbf{x}$ of group 1 and group 2, $f_1(\mathbf{x})/f_2(\mathbf{x})$, and has the form of quadratic function of $\mathbf{x}$. That is why this classification rule is called a quadratic discriminant analysis (QDA).

Support Vector Machines

Support Vector Machines (SVMs) were invented and developed by Vladimir N. Vapnik in 1963 [65]. There are three different types of classifiers depending on the situation. One of the most popular techniques as a classifier is the Maximal Margin Classifier (MMC), which uses the separating hyperplane. The other classifier is the Support Vector Classifier (SVC) and the third one is Support Vector Machine (SVM). The MMC is used when the groups are strictly linearly separable. The SVC is used when the classes are not linearly separable. In both cases, the classifier is still a linear hyperplane of the feature space. If the boundary of the groups cannot be linear, then the Support Vector Machine is appropriate. First, let's focus on the case when the decision boundary is linear.

When the classes are linearly separable, we can imagine a separating hyperplane between classes as a decision boundary. The dimension of the separating hyperplane is 1 less than the dimension of the data space. For example, if the process has two variables, the separating hyperplane will be a line, and if the model has three variables, the separating hyperplane will be a two-dimensional plane. The linear decision boundary separates the groups completely and if a new point falls on one side of the decision boundary, the point can be assigned to that group.

Then, how can we decide the decision boundary? There are plenty of hyperplanes between two groups when they are linearly separable. Among those hyperplanes, the one that creates the furthest distance to each group as a classifier or a decision boundary is the maximal margin classifier (MMC). The distance is defined as the perpendicular distance from the hyperplane to the nearest data point of the group and this distance is called the Margin. Here, the MMC depends only on the nearest data points of each group and those data points are called Support Vectors. Then, the margin is the perpendicular distance from the separating hyperplane to support vectors. In a linearly separable case, the margin is positive and called a hard margin. To find the MMC we first draw two parallel hyperplanes passing through the support vectors of each group. Then, we choose the orientation of the hyperplane to maximize the perpendicular distance between parallel hyperplanes. The hyperplane located right in the middle of two parallel hyperplanes is the MMC. In the two-group case, we have paired p-dimensional data points **x** and y = either 1 or -1 depending on the group. Then, the problem reduces to finding a separating hyperplane such that $\mathbf{w}'\mathbf{x} + c = 0$, where $\mathbf{w}$ is the normal vector to the hyperplane and c is a constant. Therefore, finding such a hyperplane is solving an optimization problem under some constraints for n data points as

$$\min \ \frac{1}{2}||w||^2 \quad \text{s.t.} \quad y_i(w'x_i + c) \geq 1, \ \ i = 1, 2, \ldots, n,$$

where $||w||$ is L_2 norm.

When it comes to the linearly inseparable groups, the points located on the wrong side of the linear decision boundary seem inevitable and it becomes impossible to get the MMC. We need to consider whether we still want to create the linear separator keeping some misclassified points that generate negative margins. A soft margin occurs in this situation. To create the linear separator, we sacrifice some points as misclassified and still maximize the margin. There is a trade-off between maximizing the margin and minimizing the number of misclassified points. The problem is modified by adding a new variable (slack variable) as the penalty for the misclassified points. Under the constraints of the slack variable, a linear decision boundary is the solution to the following optimization problem.

$$\min \ \frac{1}{2}||w||^2 + M\sum \xi_i \quad \text{s.t.} \quad y_i(w'x_i + c) \geq 1 - \xi_i \ \ , \ \ i = 1, 2, \ldots, n,$$

where ξ_i is the distance of x_i as a margin if it is on the wrong side of the decision boundary, otherwise, it's zero. Here, M is a parameter for the trade-off above. If M is large, the optimization puts more weight on avoiding misclassification, even if the margins are kept small. When M is small, the optimization focuses more on maximizing the margin rather than on misclassification. If M is large, it is close to the hard margin case. We use Lagrange Multiplier to solve this problem to fit the separating hyperplane. This decision boundary is the Support Vector Classifier (SVC).

Another way to solve the linearly inseparable problem is to use Support Vector Machine(SVM). In this method, data are mapped into a higher dimension and, when mapped, data are transformed using kernel function since it transforms the dot product of data. This procedure is called Kernel Trick. Then, we find the decision boundary in the transformed data space. This decision boundary is the hyperplane of the transformed data space and is linear in the transformed data space. In the original data space, however, the decision boundary becomes nonlinear. There are 3 widely used kernel functions.

- Polynomial Kernel : $k(x_i, x_j) = (x_i \cdot x_j + a)^b$
- Radial basis Kernel : $k(x_i, x_j) = e^{-\frac{||x_i - x_j||^2}{2\sigma^2}}$
- Sigmoidal Kernel : $k(x_i, x_j) = \tanh(ax_i \cdot x_j - b)$,

where $k(x_i, x_j)$ is a kernel function, $x_i \cdot x_j$ is a dot product, and a and b are the parameters defining kernel's behavior.

K-Nearest Neighbor

K-Nearest Neighbor is an algorithm to classify the observations based on the distance from the observation to its neighbors and was developed by Thomas M. Cover and Peter E. Hart in 1967 [17]. In two groups case, for an observation to be classified, we measure the distance of the observation to all neighbors and order them. All neighbors are already classified as one of two groups so KNN is a supervised learning. We allocate the observation to the group with a majority among the K nearest neighbors. Here, one needs to make the decision on K and the distance measure. A bias-variance tradeoff for observational prediction (or classification) may occur depending on K. When K is too small, the model is affected by outliers and it is less stable. The model has a smaller bias but shows a higher variance in prediction. On the other hand, when K is too large, the model is more stable with less variance but the bias in prediction happens. In practice, $K = \sqrt{(n)}$, n is the number of observations, is the rule of thumb. Euclidean distance, Manhattan distance, and Minkowski distance are used as distance metrics.

Decision Tree Learning

A decision tree is a non-parametric supervised learning procedure for prediction, i.e., classification or regression, and was proposed by James N. Morgan and John A. Sonquist in 1963 [49]. It has several advantages in that it is easy to understand and interpret and does not require distributional assumptions. Nodes and branches make tree-shaped decision flow. On each node, a test on a variable is run and the results of the test follow the branches which connect to the next nodes. The initial node is the root of the tree and the terminal node(leaf) indicates the class of the input data. In each node, the variable to be tested is determined by the calculation of Gini Impurity or Information Gain. Gini Impurity is a number between 0 and 1, where 0 means all data points are assigned to one class and 1 means randomly assigning the data to classes.

$$\text{Gini impurity(D)} = 1 - \sum_{c \in C} p(c)^2,$$

where D is any variable in the data set with C classes and $p(\cdot)$ is the probability that an observation belongs to class of c.

Information Gain uses the entropy of variables at the given node. Entropy ranges from 0 to 1. When all the data points fall in one class, the entropy is 0. If half of the data points are in one class and the other half are in the other class, then the entropy equals 1. On top of the tree, the first variable tested is the variable with the smallest entropy. Then, a variable with a higher Information Gain is tested on the next node. To choose this variable, Information Gain from the current variable to the next each variable needs to be calculated. Repeat this process to the final variable. Again, D is any variable in the data set, C represents the classes in the variable of D, and A is the variable with class a.

$$\text{Entropy}(D) = -\sum_{c \in C} p(c) \log_2 p(c)$$

$$\begin{aligned}\text{Information Gain}(D, a) &= \text{Entropy}(D) - \text{Entropy}(D|a) \\ &= -\sum_{c \in C} p(c) \log_2 p(c) - \sum_{c \in C} -p(c|a) \log_2 p(c|a).\end{aligned}$$

On average,

$$\text{Information Gain}(D, A) = \text{Entropy}(D) - \text{Entropy}(D|A), \text{ where } a \in A,$$

and

$$\text{Entropy}(D|A) = \sum_{a \in A} p(a) \sum_{c \in C} -p(c|a) \log_2 p(c|a).$$

The performance of a Decision Tree can be measured by accuracy, precision, or sensitivity from the confusion matrix that is explained in Table 2.1. Classification and Regression Tree (CART) and Random Forest are popular algorithms.

Neural Networks

Neural networks are a series of decision-making algorithms or non-linear statistical data modeling similar in structure to neurons in the human brain. McCulloch and Pitts (1943) [46] first introduced neural network with neurons and layers. A neural network has many different layers and each layer consists of interconnected neurons. On the one hand, the input layer receives the data and, on the other hand, there is an output layer. In between, there could be multiple hidden layers. If the data are given to the neurons of the input layer, the weighted sum of neurons of the input layer added by some constant, a bias, produces the output to the neuron in the next layer(hidden layer). It has the form of $\sum(weight \times input) + bias$. Then, a nonlinear function called an activation function evaluates the output of the previous layer which is the input to the neuron of the current layer, and determines whether the current neuron will be included for the calculation of the output to the next layer. The sigmoid function, $f(x) = (1 + e^x)^{-1}$, is one of the most widely used activation functions.

Other choices are Tanh function, $f(x) = \tanh(x)$, and ReLU function, $f(x) = \max(0, x)$. A weighted sum of the activated neurons and the activation function passes the data to the next layer again.

In this process, the data propagates forward from the input layer to the output layer. In the learning process, based on the results of the classification of the test data, the system adjusts the weights and biases of each layer backward to minimize the misclassification rate. The gradient descent method using the backpropagation algorithm is used to adjust the weights and biases. Through the learning process, this propagation forward and backpropagation iterate until the system classifies data correctly.

2.1.2 Probabilities of Correct Classification and Incorrect Classification

Suppose there are two populations, $\mathbf{G_1}$ and $\mathbf{G_2}$, representing each class. An observation $\mathbf{X}$ of the p-variate vector comes from one of these two groups. Let $f_1(\cdot)$ and $f_2(\cdot)$ denote the pdf of observations from $\mathbf{G_1}$ and $\mathbf{G_2}$, respectfully. All observations belong to one of two sets, Ω_1 or Ω_2, which consists of Ω, the sample space. When we need to allocate a new observation, we assign this new observation to population $\mathbf{G_1}$ if it belongs to Ω_1 or if it is included in the set of Ω_2, then it is classified to population $\mathbf{G_2}$. In case $\Omega_1 \cap \Omega_2$ exists and we need to allocate the observations to one of two classes based on a certain classification rule, chances are that we will allocate the observations incorrectly, that is, we will assign an observation from $\mathbf{G_1}$ to $\mathbf{G_2}$, or assign an observation from $\mathbf{G_2}$ to $\mathbf{G_1}$. We want to construct a rule, a classification function, that minimizes the chance of making such mistakes, i.e., the probability of misclassification.

The rule divides the sample space into two regions, $\mathbf{R_1}$ and $\mathbf{R_2}$ which are disjoint. If the observation belongs to $\mathbf{R_1}$, we assign the observation to $\mathbf{G_1}$. If the observation belongs to $\mathbf{R_2} = \mathbf{\Omega} - \mathbf{R_1}$, we classify the observation as $\mathbf{G_2}$. The other factor we can consider is the prior probability of occurrence. Also, the cost of misclassification may be taken into account. If allocating an object from $\mathbf{G_1}$ as $\mathbf{G_2}$ costs much more than allocating an object from $\mathbf{G_2}$ as $\mathbf{G_1}$, this allocation must be decided with greater caution. We can express the probability of misclassification or correct classification of an object from $\mathbf{G_1}$ in terms of the conditional probabilities, $P(\cdot|\mathbf{G_1})$. The probability of misclassification of an object from $\mathbf{G_1}$ can be written as the conditional probability of $P(\mathbf{G_2}|\mathbf{G_1})$ and the probability of correct classification of an object from $\mathbf{G_1}$ is the conditional probability of $P(\mathbf{G_1}|\mathbf{G_1})$. If we use the density functions,$f_1(\mathbf{x})$ or $f_2(\mathbf{x})$, and the integration over the corresponding set, $\mathbf{R_1}$ or $\mathbf{R_2}$, the probabilities of correct classification are

$$\begin{aligned}
P(\mathbf{G_1}|\mathbf{G_1}) &= P(\mathbf{X} \text{ is allocated to } \mathbf{G_1}|\mathbf{X} \text{ is from } \mathbf{G_1}) \\
&= P(\mathbf{X} \in \mathbf{R_1}|\mathbf{X} \sim \mathbf{f_1}(\mathbf{x})) = \int_{\mathbf{R_1}} \mathbf{f_1}(\mathbf{x})\mathbf{dx} \\
P(\mathbf{G_2}|\mathbf{G_2}) &= P(\mathbf{X} \text{ is allocated to } \mathbf{G_2}|\mathbf{X} \text{ is from } \mathbf{G_2}) \\
&= P(\mathbf{X} \in \mathbf{R_2}|\mathbf{X} \sim \mathbf{f_2}(\mathbf{x})) = \int_{\mathbf{R_2}} \mathbf{f_2}(\mathbf{x})\mathbf{dx}.
\end{aligned}$$

Likewise,
$P(\mathbf{G_1}|\mathbf{G_2}) = P(\mathbf{X} \in \mathbf{R_1}|\mathbf{X} \sim \mathbf{f_2}(\mathbf{x})) = \int_{\mathbf{R_1}} \mathbf{f_2}(\mathbf{x})\mathbf{dx}$ and
$P(\mathbf{G_2}|\mathbf{G_1}) = P(\mathbf{X} \in \mathbf{R_2}|\mathbf{X} \sim \mathbf{f_1}(\mathbf{x})) = \int_{\mathbf{R_2}} \mathbf{f_1}(\mathbf{x})\mathbf{dx}$ are for the case of incorrect classifications. If we consider the prior probability to draw the unconditional probabilities, we can express the probability of correct classification of an observation that is assigned to $\mathbf{G_1}$ is as follows.

$$\begin{aligned}
&P(\ \mathbf{X} \text{ is } \textbf{correctly} \text{ classified as } \mathbf{G_1}) \\
&= P(\ \mathbf{X} \text{ is from } \mathbf{G_1} \text{ and allocated to } \mathbf{G_1}) \\
&= P(\mathbf{X} \text{ is from } \mathbf{G_1}) \times P(\mathbf{X} \text{ is allocated to } \mathbf{G_1}|\mathbf{X} \text{ is from } \mathbf{G_1}) \\
&= P(\mathbf{G_1})P(\mathbf{X} \in \mathbf{R_1}|\mathbf{G_1}) \\
&= p_1 P(\mathbf{G_1}|\mathbf{G_1})
\end{aligned}$$

The unconditional probability of incorrect classification of an observation that is assigned to $\mathbf{G_1}$ is

$$\begin{aligned}
&P(\ \textbf{X} \text{ is } \textbf{incorrectly} \text{ classified as } \mathbf{G_1})\\
&= P(\ \textbf{X} \text{ is from } \mathbf{G_2} \text{ and allocated to } \mathbf{G_1})\\
&= P(\mathbf{X} \text{ is from } \mathbf{G_2}) \times P(\mathbf{X} \text{ is allocated to } \mathbf{G_1} | \mathbf{X} \text{ is from } \mathbf{G_2})\\
&= P(\mathbf{G_2})P(\mathbf{X} \in \mathbf{R_1}|\mathbf{G_2})\\
&= p_2 P(\mathbf{G_1}|\mathbf{G_2})
\end{aligned}$$

Then,

$$\begin{aligned}
&P(\ \textbf{X} \text{ is } \textbf{correctly} \text{ classified as } \mathbf{G_2}) = p_2 P(\mathbf{G_2}|\mathbf{G_2})\\
&P(\ \textbf{X} \text{ is } \textbf{incorrectly} \text{ classified as } \mathbf{G_2}) = p_1 P(\mathbf{G_2}|\mathbf{G_1}).
\end{aligned}$$

Here, $P(\mathbf{G_1}) = p_1$ and $P(\mathbf{G_2}) = p_2$ are the prior probabilities and $p_1 + p_2 = 1$.

Taking the cost of misclassification into consideration, let $c(\mathbf{G_1}|\mathbf{G_2})$ be the cost when we incorrectly assign **X** as $\mathbf{G_1}$ and $c(\mathbf{G_2}|\mathbf{G_1})$ be the cost when we incorrectly assign **X** as $\mathbf{G_2}$. There is no cost when the classification is correct ($c(\mathbf{G_1}|\mathbf{G_1}) = c(\mathbf{G_2}|\mathbf{G_2}) = 0$). Then, we can calculate the expected cost of misclassification (ECM) from 2.1.1 as

$$\begin{aligned}
ECM &= \sum_{\substack{all\ misclassified\\ points}} \text{P(misclassification)} \times \text{ P(prior)} \times \text{(cost of misclassification)}\\
&= P(\mathbf{G_2}|\mathbf{G_1}) \times p_1 \times c(\mathbf{G_2}|\mathbf{G_1}) + P(\mathbf{G_1}|\mathbf{G_2}) \times p_2 \times c(\mathbf{G_1}|\mathbf{G_2})
\end{aligned}$$

and we find a classifying rule by minimizing the ECM.

2.1.3 Total Probability of Misclassification and Apparent Error Rate

Total Probability of Misclassification

The total probability of misclassification (TPM) given any classification rule can be calculated when the distribution of the population is known and the prior probability also is known. Suppose again $\mathbf{G_1} \sim f_1(\cdot)$, $\mathbf{G_2} \sim f_2(\cdot)$, $P(\mathbf{G_1}) = p_1$, and $P(\mathbf{G_2}) = p_2$. From Figure 2.1, area A and area B are the cases where the misclassifications happen if the blue line is the classification rule.

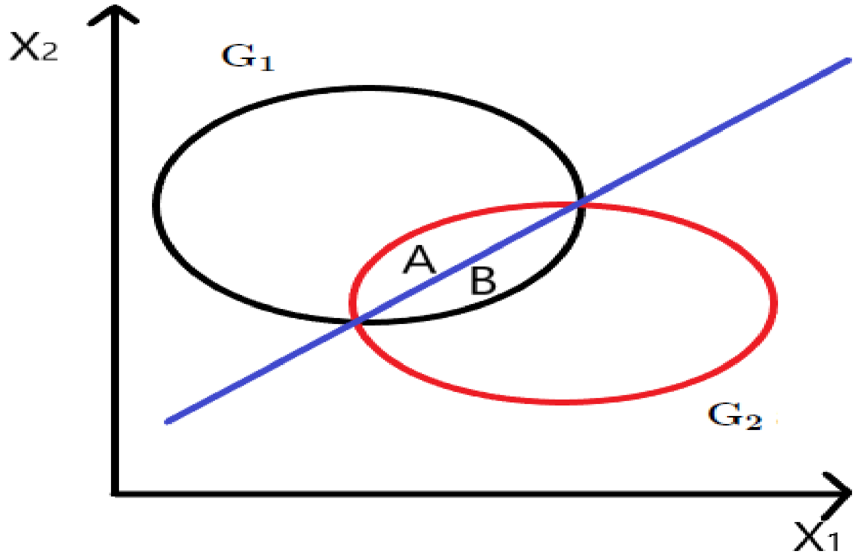


Figure 2.1: Bivariate Classification Example

Thus, the TPM is the probability for the area of A and B. Then, the TPM using the conditional probabilities for the bivariate case can be expressed as follows.

$$
\begin{aligned}
TPM &= P(\text{from } \mathbf{G_1} \text{ and misclassified }) + P(\text{ from } \mathbf{G_2} \text{ and misclassified}) \\
&= P(\text{ from } \mathbf{G_1} \text{ }) \times P(\text{ assigned to } \mathbf{G_2} \mid \text{from } \mathbf{G_1} \text{ }) \\
&+ P(\text{ from } \mathbf{G_2} \text{ }) \times P(\text{ assigned to } \mathbf{G_1} \mid \text{from } \mathbf{G_2} \text{ }) \\
&= p_1 P(\mathbf{G_2}|\mathbf{G_1}) + p_2 P(\mathbf{G_1}|\mathbf{G_2}) \\
&= p_1 \int_{\mathcal{B}} f_1(\mathbf{x}) d\mathbf{x} + p_2 \int_{\mathcal{A}} f_2(\mathbf{x}) d\mathbf{x},
\end{aligned} \tag{2.1}
$$

where $\mathbf{x}$ is the observation vector, i.e., $\mathbf{x} = (x_1, x_2)$. The penultimate equality is simply the sum of two incorrect classification probabilities from the last section.

When we find the classifying criteria by minimizing the total probability of misclassification (TPM), it is the same problem as we minimize the ECM with identical costs. Then, we have a rule to allocate a new observation, $\mathbf{x_0}$ as:

Allocate $\mathbf{x_0}$ as $\mathbf{G_1}$ if

$$\frac{f_1(\mathbf{x_0})}{f_2(\mathbf{x_0})} \geq \frac{p_2}{p_1}.$$

Otherwise, we allocate $\mathbf{x_0}$ as $\mathbf{G_2}$.

If we have equal prior probabilities for populations, i.e., $p_1 = p_2 = \frac{1}{2}$,

$$TPM = \frac{1}{2} \left(\int_{\mathcal{B}} f_1(\mathbf{x}) d\mathbf{x} + \int_{\mathcal{A}} f_2(\mathbf{x}) d\mathbf{x} \right). \tag{2.2}$$

Then, the corresponding classifying rule becomes “classify $\mathbf{x_0}$ as $\mathbf{G_1}$ if $\dfrac{f_1(\mathbf{x_0})}{f_2(\mathbf{x_0})} \geq 1$”.

Apparent Error Rate

The calculation of TPM depends on the distributions of the populations. If we don’t know the distributions of the populations, we still can measure the probability of misclassification. From the ratio of the frequencies by simply counting the number of observations classified correctly or incorrectly, we can. The below show the confusion matrix for calculating the ratio.

		Assigned to		Total
		$\mathbf{G_1}$	$\mathbf{G_2}$	
Observations from	$\mathbf{G_1}$	$N_{\mathbf{G_1}cor}$	$N_{\mathbf{G_1}mis}$	$N_{\mathbf{G_1}}$
	$\mathbf{G_2}$	$N_{\mathbf{G_2}mis}$	$N_{\mathbf{G_2}cor}$	$N_{\mathbf{G_2}}$

Table 2.1: Confusion Matrix

- $N_{\mathbf{G_1}}$ = number of observations from population 1, $\mathbf{G_1}$
- $N_{\mathbf{G_2}}$ = number of observations from population 2, $\mathbf{G_2}$
- $N_{\mathbf{G_1}cor}$ = number of observations correctly assigned to $\mathbf{G_1}$
- $N_{\mathbf{G_1}mis}$ = number of observations incorrectly assigned to $\mathbf{G_2}$
- $N_{\mathbf{G_2}cor}$ = number of observations correctly assigned to $\mathbf{G_2}$
- $N_{\mathbf{G_2}mis}$ = number of observations incorrectly assigned to $\mathbf{G_1}$
- $N_{\mathbf{G_1}mis} = N_{\mathbf{G_1}} - N_{\mathbf{G_1}cor}$ and $N_{\mathbf{G_2}mis} = N_{\mathbf{G_2}} - N_{\mathbf{G_2}cor}$

The apparent error rate(APER) can be expressed as a ratio of the number of misclassified observations to the total number of observations,

$$\frac{N_{\mathbf{G_1}mis} + N_{\mathbf{G_2}mis}}{N_{\mathbf{G_1}} + N_{\mathbf{G_2}}}.$$

If we apply the relationship of counts between correctly classified and incorrectly classified,

$$\begin{aligned} APER &= \frac{N_{\mathbf{G_1}mis} + N_{\mathbf{G_2}mis}}{N_{\mathbf{G_1}} + N_{\mathbf{G_2}}} \\ &= \frac{N_{\mathbf{G_1}} - N_{\mathbf{G_1}cor} + N_{\mathbf{G_2}} - N_{\mathbf{G_2}cor}}{N_{\mathbf{G_1}} + N_{\mathbf{G_2}}} \\ &= 1 - \frac{N_{\mathbf{G_1}cor} + N_{\mathbf{G_2}cor}}{N_{\mathbf{G_1}} + N_{\mathbf{G_2}}} \\ &= 1 - Accuarcy. \end{aligned}$$

2.1.4 Classifying Multivariate Normal Populations with Minimum TPM Rule

Let $f_1(\mathbf{x})$, $f_2(\mathbf{x})$ be p-variate normal densities with mean vectors $\mu_\mathbf{1}$ and $\mu_\mathbf{2}$ and the covariance matrices Σ_1 and Σ_2, respectively. For simplicity, let's assume the covariance matrices are identical($\Sigma_1 = \Sigma_2 = \Sigma$) and is positive definite. Then, the density of $\mathbf{X}$ from population $\mathbf{G_1}$ is

$$f_1(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{p}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mu_\mathbf{1})'\mathbf{\Sigma^{-1}}(\mathbf{x}-\mu_\mathbf{1})}, \tag{2.3}$$

where $|\Sigma|$ is the determinant of Σ. The joint density of $\mathbf{X}$ from population $\mathbf{G_2}$ is likewise.

If we use the minimum TPM rule from 2.1.3, a new observation of $\mathbf{x_0}$ is allocated to $\mathbf{G_1}$ if

$$\frac{f_1(\mathbf{x_0})}{f_2(\mathbf{x_0})} \geq \frac{p_2}{p_1},$$

where p_1 and p_2 are prior probabilities. Then, the rule can be written as follows taking logarithm on both sides of the above inequality.

$$(\mu_1 - \mu_2)'\Sigma^{-1}\mathbf{x_0} - \frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2) \geq \log\left(\frac{p_2}{p_1}\right) \tag{2.4}$$

The rule is a linear function of $\mathbf{x_0}$. Also, if we assume the same prior, $p_1 = p_2$, this rule becomes identical to the decision rule of LDA in 2.1.1.

$$(\mu_1 - \mu_2)'\Sigma^{-1}\mathbf{x_0} \geq \frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2) \tag{2.5}$$

When μ_1, μ_2, and Σ are unknown, we estimate those parameters and plug $\bar{\mathbf{x}}_1$, $\bar{\mathbf{x}}_2$, and S_{pooled} in, where $S_{pooled} = \dfrac{S_1(n_1-1)}{(n_1-1)+(n_2-1)} + \dfrac{S_2(n_2-1)}{(n_1-1)+(n_2-1)}$ since we assume the same covariance matrices. S_1 and S_2 are the sample covariance matrices. Then, we get

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'S_{pooled}^{-1}\mathbf{x_0} \geq \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'S_{pooled}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2). \tag{2.6}$$

2.2 The Selection and Ranking Methodologies

The selection and ranking methodologies (SRM) have long historical roots. The methodologies were introduced by several prominent statisticians through the publications of Bechhofer(1954) [4], Bechhofer and Sobel (1954) [8], and Gupta (1956) [31]. Since then, the results of more than two decades of research are summarized in two books; Gibbons et al (1977) [27] and Gupta and Panchapakesan (1979) [33]. The goal of the selection and ranking methodologies is to select the "best" population among many alternatives. In the problem of multiple comparisons of $k(\geq 2)$ populations, the methodologies provide more diversified approaches than the conventional approaches such as the analysis of variance (ANOVA) or investigating the least significant difference (LSD). Following those books, many research papers and books were published and the selection and ranking methodologies contributed great impacts on the development and advancement of the statistical multiple decision theory.

In classical statistics, the researchers conducted hypothesis tests to compare the multiple populations. The null hypothesis is set such that the values in which the researcher is interested are common for all populations. Then, the null hypothesis may or may not be rejected against alternatives when samples are taken to test it. Depending on the test results (rejecting or failing to reject the hypothesis) a decision can be made as to whether the values of interest are homogeneous or there exist differences among them. One of the most widely used methods of comparing problems of multiple populations was ANOVA. When the ANOVA test rejected the null hypothesis of common value of interest, the researchers were confronted with new questions such as "If the difference among populations exists, then which population is different from which other populations?" or "Which one could we select as the best(or worst) population?" After deciding to reject the null hypothesis, it was natural for those questions to follow. This is because the purpose of statistical tests for the null hypothesis of homogeneity is generally not to indicate homogeneity of population values, but to determine whether differences exist. For example, when a pharmaceutical company launches a new medical pill, the company wants to show that the new pill performs better than the existing products on the market. Moreover, when an investor is trying to decide about a stock portfolio to invest money in, she wants to find the best performing stock portfolio among several stock portfolios.

To address these problems, the multiple comparison procedures performed a pair(or more than a pair) of hypothesis tests. The multiple comparison procedures could efficiently answer the question of verifying the differences between the populations or the degree of the difference between the populations. Meanwhile, the selection and ranking methodologies were introduced and developed to answer the problem of selecting the "best" population. Since R. Bechhofer's pioneering work on the selection and ranking methodologies, it has been developed in two directions. One approach is to select a fixed number of best populations and the other approach is to select a (randomized) group of populations that contains the best population. The former uses the indifference-zone approach and was developed by R. Bechhofer and M. Sobel, while the latter is called the subset selection method and was introduced by S. Gupta and M. Sobel and developed by S. Gupta [31]. Since both approaches have the main goal of selecting several populations such that the number is either fixed or random, the selection and ranking methodologies can be considered as selection problems.

Suppose that there are *k* populations, $\mathbf{G_1}, \mathbf{G_2}, \ldots, \mathbf{G_k}$, where $\mathbf{G_i}$ has the distribution function, $F_{\theta_i}(\cdot), i = 1, 2, \ldots, k(k < \infty)$ and θ_i is an unknown real valued parameter, $\theta \in \Theta$. Also, θ_i is associated with $\mathbf{G_i}$ and the ordered θ_i is denoted by $\theta_{[1]} \leq \cdots \leq \theta_{[k]}$. We assume no prior knowledge concerning the association between $\mathbf{G_i}$ and $\theta_{[i]}$. The primary goal of the selection and ranking methodologies is to attempt to formulate the decision problem of selecting the best population out of *k* populations or selecting a non-empty subset of the *k* populations so that the subset contains the best population with a minimum of the prespecified probability level. The best population could be defined as the one with the largest(or smallest) mean or depending on the problems or the situations. These formulations are formally known as the following two main approaches; the Indifference-Zone (IZ) approach developed by Bechhofer [4] and the Subset Selection (SS) method developed by Gupta [33].

2.2.1 Indifference-Zone Approach

In the indifference-zone approach, the goal is to select the best population [4] (Bechhofer, 1954). Let $\hat{\theta}$ denote the quantity from the sample by which corresponds to θ in the population. The ordered quantities are denoted by $\hat{\theta}_{[1]} \leq \hat{\theta}_{[2]} \leq \cdots \leq \hat{\theta}_{[k]}$ and the population which is associated with $\hat{\theta}_{[k]}$ becomes the best population if we prefer the larger value of the parameter. Then, two kinds of decisions are available in this selection problem. If the selected population is the true best population, the correct selection is made. If the selected population is not the true best population, we commit a wrong decision. In this procedure, we want to maintain a certain level of probability of making a correct decision, just as we controlled for the error rate in the hypothesis test. We prefer this probability to be as high as possible, and under some conditions, it is bounded by a specific value. Then, we define and range the probability of correct selection as follows. The guaranteed probability of making a correct selection whenever the difference between the best value of θ and the second best value of θ is at least some fixed amount is at least P^*.

$$P\left(CS\right) \geq P^* \text{ whenever } \theta_{[\mathbf{k}]} - \theta_{[\mathbf{k-1}]} \geq \delta^* \text{ for prespecified } P^*\ and\ \delta^*, \tag{2.7}$$

where CS denotes “correct selection”. Then, the parameter space can be divided into a part where the difference, $\delta = \theta_{[k]} - \theta_{[k-1]}$, is less than the prespecified δ^*, and the rest of the parameter space where $\delta \geq \delta^*$. Thus, a *k*-dimensional parameter space is reduced to 2-dimensional space since selecting the best population only concerns the populations with the two largest values of parameters, $\theta_{[k]}$ and $\theta_{[k-1]}$. In the reduced parameter space, the region where $\delta < \delta^*$ is called the indifference-zone (IZ) and the region where $\delta \geq \delta^*$ is called the preference-zone (PZ). In the PZ, we have a strong preference to make a correct selection because there exists a gap between the largest and the second largest population and in the IZ we are indifferent about making a selection. Since the probability of the correct selection, P(CS), must be guaranteed only in the PZ, we are only interested in the configuration of PZ. There exist many different configurations of parameter space within PZ which make P(CS) at least P*. Among those configurations, we can find the configuration for which the P(CS) is minimum over the preference-zone and we call it the least favorable configuration(LFC). If the P(CS) at LFC can be equal to P*, then we can attain the probability requirement of (2.7) above. For the problem of the population mean where we prefer the larger mean, the LFC happens when

$$\theta_{[1]} = \theta_{[2]} = \cdots = \theta_{[k-1]} = \theta_{[k]} - \delta$$

and the minimum sample size to meet the probability requirement can be calculated given prespecified P* and δ^* later.

Selection of the Best Population; Normal Distribution with Common Known Variance

Consider *k*(≥ 2) independent normal populations with unknown means, $\mu_i, i = 1, 2, \ldots, k$ $(k < \infty)$ and a known common variance, σ^2. When we want to find the best population regarding the population mean, if a larger mean is considered better, the best population is defined as the population with the largest mean. Then, we collect samples from each population of size N. Now, X denotes the observation from the sample, then $X_{1,1}$ stands for the first observation from the first population. Also, $X_{1,N}$ is the N^{th} observation obtained from the first population. Then, $X_{1,j}$ follows a normal distribution with mean μ_1 and the variance σ^2, $j = 1, 2, \dots, N$, and each observation is not correlated to others. If we set $\bar{X}_1 = \sum_{j=1}^{N} X_{1,j}/N$, then $E(\bar{X}_1) = \mu_1$. For the second population, $X_{2,j} \sim N(\mu_2, \sigma^2), j = 1, 2, \dots, N$ and $\bar{X}_2 = \sum_{j=1}^{N} X_{2,j}/N$, then $E(\bar{X}_2) = \mu_2$. Thus, $X_{i,j} \sim N(\mu_i, \sigma^2)$, $\bar{X}_i = \sum_{j=1}^{N} X_{i,j}/N$, and $E(\bar{X}_i) = \mu_i,$ for $i = 1, 2, \dots, k,$ and $j = 1, 2, \dots, N$. We assume no correlation within the sample from each population as well as no correlation between the samples. Then, the goal is to select a population with the largest μ, denoted by $\mu_{[k]}$, where $\mu_{[k]} \geq \mu_{[k-1]} \geq \mu_{[k-2]} \geq \cdots \geq \mu_{[2]} \geq \mu_{[1]}$.

A rational way to select the population with $\mu_{[k]}$ is to find a sample with $\bar{X}_{[k]}$ and select the corresponding population as the best population. However, when the largest mean and the other means are close to each other, it is difficult to select a population as the best one. Therefore, we need a rule that guarantees the selected one is with the largest mean. A measure used in this rule is the distance between the highest mean and the second highest mean, $\mu_{[k]} - \mu_{[k-1]} = \delta$, i.e., the discerning measure of distance. At the same time, we want to preserve the probability of making a correct decision to select a population associated with the sample producing $\bar{X}_{[k]}$ as the best population as long as the distance defined above is greater than or equal to some value, δ^*. The probabilistic condition can be written as follows:

$$P(CS) \geq P^* \text{ whenever } \mu_{[k]} - \mu_{[k-1]} = \delta \geq \delta^* \text{ for prespecified } P^* \, and \, \delta^*,$$

where CS stands for “correct selection” and δ^* is a discerning threshold. To make the problem meaningful, the value of P^* has the range of $\frac{1}{k} < P^* < 1$ since the probability of $\frac{1}{k}$ can be attained by a random selection. The minimum of P(CS), P^*, happens when

$$\mu_{[1]} = \mu_{[2]} = \cdots = \mu_{[k-1]} = \mu_{[k]} - \delta^*$$

and we call this parameter configuration the *least favorable configuration* (LFC). When this probability requirement is set with a pair (P^*, δ^*), we can choose the sample size, N, using a table provided by R. Bechhofer [4]. Then, the selection procedure given (P^*, δ^*) is

1. Select a sample of size N from each population.

2. Calculate $\bar{X}_{[1]}, \bar{X}_{[2]}, \bar{X}_{[3]}, \ldots, \bar{X}_{[k-1]}, \bar{X}_{[k]}$.

3. Select the population associated with $\bar{X}_{[k]}$ as the best.

N is determined by the probability requirement, $P(CS) \geq P^*$, where

$$P(max(\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_{k-1}) < \bar{X}_{[k]}) = \int_{-\infty}^{\infty} F(y+d)^{k-1} f(y)dy = P^*,$$

$d = \sqrt{N}\frac{\delta^*}{\sigma}$, F($\cdot$) and f($\cdot$) are cdf and pdf of the standard normal random variable, and d is found from the table by R. Bechhofer (1954) [4].

Selection of t Best Populations

Suppose we want to find *t* best populations, where $1 < t < k$, under the same situation as the previous section, 2.2.1, with normal distributions and a known common variance. $X_{i,j} \sim N(\mu_i, \sigma^2)$, for $i = 1, 2, \ldots, k$, and $j = 1, 2, \ldots, N$. Then, the populations we need to select are those corresponding to $\mu_{[k]}, \mu_{[k-1]}, \mu_{[k-2]}, \ldots, \mu_{[k-t+1]}$. However, the order of the populations is not required. Consequently, the discerning measure of distance is defined as the difference between the smallest mean of the *t* best populations and the largest mean of the remaining *(k-t)* populations, i.e., $\delta = \mu_{[k-t+1]} - \mu_{[k-t]}$ from Gibbons et al (1977) [27]. The least favorable configuration from R. Bechhofer (1954) [4] can be expressed as follows:

$$\mu_{[k]} - \mu_{[k-t+1]} = 0,$$

$$\mu_{[k-t+1]} - \mu_{[k-t]} = \delta,$$

$$\mu_{[k-t]} - \mu_{[1]} = 0.$$

Given P^* and δ^* for the probability requirement, the probability of correct selection is greater than or equal to P^* whenever the discerning measure of distance, δ, is greater than or equal to δ^*.

$$P(CS) \geq P^* \text{ whenever } \mu_{[k-t+1]} - \mu_{[k-t]} = \delta \geq \delta^*.$$

If the sample of size N is taken from each population and sample means are calculated, we can order them as $\bar{X}_{[1]} \leq \bar{X}_{[2]} \leq \cdots \leq \bar{X}_{[k]}$.

Then, the *t* best populations are the populations that produce the *t* largest sample means, $\bar{X}_{[k]}$, $\bar{X}_{[k-1]}, \ldots, \bar{X}_{[k-t+1]}$ with guaranteed probability of P^*.

To determine the sample size N given P^* and δ^*, refer the table from [4]. Given P^* and δ^*, we get the sample size from the equation below.

$$P^* = tP(\max(\bar{X}_1, \ldots, \bar{X}_{k-t}) < \bar{X}_{k-t+1} < \min(\bar{X}_{k-t+2}, \ldots, \bar{X}_k))$$

$$= t \int_{-\infty}^{\infty} F(y+d)^{k-t}(1-F(y))^{t-1} f(y) dy,$$

where $d = \sqrt{N}\frac{\delta^*}{\sigma}$, F($\cdot$) and f($\cdot$) are cdf and pdf of Standard Normal random variable, and d is found from the table by R. Bechhofer [4].

2.2.2 Subset Selection Method

The subset selection method is a procedure for selecting a group of populations that includes the best population without identifying the best population. Here, a very distinctive aspect of subset selection is that the size of the selected group is determined randomly and not predetermined as in the indifference-zone approach. In this method, the correct selection occurs if the best population is included in the selected subset. If the best population is contained in the other set of populations, an error occurs. Under the same setting as above, 2.2, we calculate $\hat{\theta}_{[1]},\ \hat{\theta}_{[2]}, \ldots, \hat{\theta}_{[k]}$. Then, place the population whose associated $\hat{\theta}$ is included in the interval of $[\hat{\theta}_{[k]} - d,\ \hat{\theta}_{[k]}]$ as the selected subset. Here, d is determined by the condition that the infimum of the P(CS) over the whole parameter space is at least P^*. Unlike the indifference-zone approach, the probability is calculated over the whole parameter space. P^* in the indifference-zone approach is the probability calculated over the configuration of the preference-zone in the parameter space. However, P^* in the subset selection method is the probability calculated over the entire parameter space and it has the meaning of the minimum probability such that the selected subset contains the population with the largest mean value [27] (Gibbons et all, 1977). Also, since $\hat{\theta}_{[k]}$ is always included in the interval above, the selected subset cannot be empty.

The main difference between the indifference-zone approach and the subset selection method is that the subset selection method has no indifference zone. In addition, we do not identify the best population among the selected subset once the subset is determined, thus the subset selection method is less precise.

Suppose we have the same situation as the previous two cases in 2.2.1, where we have *k*(≥ 2) normal populations with a known common variance. Now, we want to select a subset of random size that includes the population with the largest mean. The sample from each population is collected with a fixed size of *n*. $X_{i,j} \sim N(\mu_i, \sigma^2)$, $\bar{X}_i = \sum_{j=1}^{n} X_{i,j}/n$, and $E(\bar{X}_i) = \mu_i,$ for $i = 1, 2, \ldots, k,$ and $j = 1, 2, \ldots, n.$ As we did in the indifference-zone approach, we need to determine P^* as the probability requirement in advance.

The rule of selection is to allocate the population as the selected subset if the population produces the sample mean greater than or equal to the value of $\bar{X}_{[k]} - d\frac{\sigma}{\sqrt{n}}$, where d is obtained from the table of R. Bechhofer (1954) [4] given P^* and *k*. That is, for any $i = 1, 2, \ldots, k$, the *i*th population is located in the subset if and only if $\bar{X}_i \geq \bar{X}_{[k]} - d\frac{\sigma}{\sqrt{n}}$, where $\bar{X}_i$ is the sample mean from the *i*th population. By rearranging the inequality, we can get the interval of selection procedure as $\left[\bar{X}_{[k]} - d\frac{\sigma}{\sqrt{n}}, \bar{X}_{[k]}\right]$ and thus, the subset can't be empty.

There are two different types of problems in the subset selection method. One is comparing to an unknown control population and the other is comparing to a known standard value of θ, θ_0. For these two types of the selection problems, however, the selected subset could be empty.

2.2.3 Multivariate Indifference-Zone Approach

Bivariate Normal Populations

Suppose we have populations that follow bivariate normal distributions. Each population has two variables as measurements and the collected samples are recorded as X_1 and X_2 for each variable. X_1 and X_2 are jointly normally distributed with the mean vector and covariance matrix shown below:

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}\right),$$

where μ_1 is the mean of X_1, μ_2 is the mean of X_2, σ_1 is the standard deviation of X_1, σ_2 is the standard deviation of X_2, and ρ is the correlation coefficient of (X_1, X_2).

Suppose there are 3 populations with bivariate normal distribution.

$$Pop_1 : \begin{pmatrix} X_{11} \\ X_{12} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{11} \\ \mu_{12} \end{pmatrix}, \Sigma_1\right)$$

$$Pop_2 : \begin{pmatrix} X_{21} \\ X_{22} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{21} \\ \mu_{22} \end{pmatrix}, \Sigma_2\right)$$

$$Pop_3 : \begin{pmatrix} X_{31} \\ X_{32} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{31} \\ \mu_{32} \end{pmatrix}, \Sigma_3\right),$$

where $\Sigma_1, \Sigma_2,$ and Σ_3 are the variance-covariance matrices and they are positive definite.

Let's assume marginally, $\mu_{31} > \mu_{21} > \mu_{11}$ and $\mu_{12} > \mu_{32} > \mu_{22}$. Then, we can have the following scatter plot.

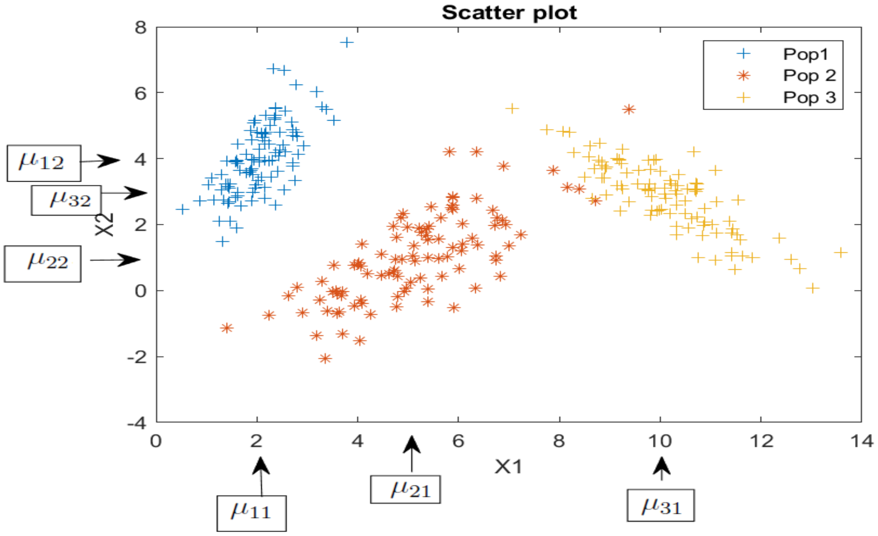


Figure 2.2: Bivariate Example with 3 Populations

If we select the best population by considering the population means marginally and prefer the population with a larger mean, Population 3 becomes the best population when we only consider the variable X_1 because $\mu_{31} > \mu_{21} > \mu_{11}$. On the other hand, if we select the best population in terms of the variable X_2, Population 1 is selected as the best population since $\mu_{12} > \mu_{32} > \mu_{22}$. Therefore, the decision about the best population depends on the choice of the marginal variable and the decisions are not identical in this example. To overcome the discrepancy, let's consider the means simultaneously and select the best population.

Selection of the Best Population Using a Linear Combination

Suppose there are k populations and each population follows a multivariate normal distribution. For the population $\mathbf{G_i}$, the mean is μ_i and the covariance is Σ_i, where μ_i is a column vector and Σ_i is a positive definite square matrix, $i = 1, 2, \ldots, k$. If the population has p variates, the dimension of μ_i is p×1 and

$$\mu_i = \begin{pmatrix} \mu_{i1} \\ \mu_{i2} \\ \vdots \\ \mu_{ip} \end{pmatrix}.$$

The dimension of Σ_i is p×p and let's assume $\Sigma_i = \Sigma$ for all populations. Then, to find the best population considering the means simultaneously, we compute θ_i, the linear combination of the means of p-variates for population $\mathbf{G_i}$, as

$$\theta_i = A\mu_i = \alpha_1 * \mu_{i1} + \alpha_2 * \mu_{i2} + \cdots + \alpha_p * \mu_{ip},$$

where $A = (\alpha_1, \alpha_2, \ldots, \alpha_p)$, $\sum_{j=1}^{p} \alpha_j = 1$ and $\alpha_j > 0$ for all *j, j* = 1, 2 , . . ., p. Here, α_j's are known and then we can order the θ_i as

$$\theta_{[1]} \leq \theta_{[2]} \leq \cdots \leq \theta_{[k]}.$$

If we define the best population as the one with the largest linear combination of means, the population associated with $\theta_{[k]}$ is selected as the best population.

If we want to select t best populations for a larger θ, the populations corresponding to $\theta_{[k]}, \theta_{[k-1]}, \theta_{[k-2]}, \ldots, \theta_{[k-t+1]}$ are selected. Then, the discerning measure of distance for this selection problem(selection of t best populations) is $\theta_{[k-t+1]} - \theta_{[k-t]}$ and let the value be δ, $\delta > 0$. The probability requirement is

$$P(CS) \geq P^* \text{ whenever } \theta_{[k-t+1]} - \theta_{[k-t]} \geq \delta^*,$$

where P^* and δ^* are predetermined. The discerning measure of distance and the probability requirement are very similar to those of the univariate population problem.

The different sets of $\alpha_j, j = 1, 2, \ldots, p$ (different weights) make different decision rules for finding the best population. On searching such set of α_j's, we can apply the statistical learning.

Selection of the Best Population Using Mahalanobis Distance

Assume again the same situation with k multivariate normal populations as in the previous section. For the population $\mathbf{G_i}$, the mean is μ_i and the covariance is Σ_i, where μ_i is a column vector and Σ_i is a positive definite square matrix, $i = 1, 2, \ldots, k$. In the linear combination method, the covariance matrix was not included in the discerning measure of distance. Since the covariance term includes additional information, it must be taken into account. To consider the means simultaneously with information from covariance, the discerning measure of distance for this formulation includes the form of $\theta_i = \mu_i' \Sigma_i^{-1} \mu_i$, the Mahalanobis distance function, and we define $\mathbf{G_i}$ is better than $\mathbf{G_j}$ if $\mu_i' \Sigma_i^{-1} \mu_i > \mu_j' \Sigma_j^{-1} \mu_j$. If the goal here is to select the t best populations, then we need to select the populations corresponding to the t largest sample $\hat{\theta}$s such as $\hat{\theta}_{[k]}, \hat{\theta}_{[k-1]}, \ldots, \hat{\theta}_{[k-t+1]}$, where $\hat{\theta}_i = \bar{X}_i' \Sigma_i^{-1} \bar{X}_i$. In this problem, we have two discerning measures of distance, δ_1 and δ_2.

$$\delta_1 = \theta_{[k-t+1]} - \theta_{[k-t]}, \quad \delta_1 \geq 0$$

$$\delta_2 = \frac{\theta_{[k-t+1]}}{\theta_{[k-t]}}, \quad \delta_2 \geq 1$$

The probability requirement is

$$P(CS) \geq P^* \quad \text{whenever} \quad \theta_{[k-t+1]} - \theta_{[k-t]} \geq \delta_1^* \quad \text{and} \quad \theta_{[k-t+1]}/\theta_{[k-t]} \geq \delta_2^*,$$

where P^*, δ_1^*, and δ_2^* are predetermined.

The preference zone (PZ) for this selection problem is the intersection of the parameter space with $\delta_1 \geq \delta_1^*$ and $\delta_2 \geq \delta_2^*$. To find out the smallest sample size n to satisfy the probability requirement given P^*, δ_1^*, and δ_2^*, we can refer to the table from Milton and Rizvi (1989) [48].

(1) **Selecting t Best Population When Σ_i Is Known**

If we take samples of size n from population i, $\mathbf{G_i}$, we can denote the sample mean vector as $\bar{X}_i$ and the sample variance-covariance matrix as S_i.

When we assume that Σ_i is known, we use $\bar{X}_i'\Sigma_i^{-1}\bar{X}_i$ and let $U_i = \bar{X}_i'\Sigma_i^{-1}\bar{X}_i$, where nU_i has the non-central chi-square distribution with p degrees of freedom and non-centrality parameter of $n(\mu_i'\Sigma_i^{-1}\mu_i)$ [1]. The least favorable configuration can be expressed as

$$\theta_{[1]} = \cdots = \theta_{[k-t]} = \delta_1(\delta_2 - 1)^{-1}$$

$$\theta_{[k-t+1]} = \cdots = \theta_{[k]} = \delta_1\delta_2(\delta_2 - 1)^{-1}.$$

The sample size n given $P*$ is calculated from

$$P^* = t\int_0^\infty F_p(x, \frac{n\delta_1}{\delta_2 - 1})^{k-t}\{1 - F_p(x, \frac{n\delta_1\delta_2}{\delta_2 - 1})\}^{t-1} f_p(x, \frac{n\delta_1\delta_2}{\delta_2 - 1})dy,$$

where $f_p(x,\theta)$ is the pdf of a noncentral chi-square random variable with p degrees of freedom and non-centrality parameter of θ and $F_p(x,\theta)$ is the cdf of that [33]. We select the populations associated with $U_{[k-t+1]}, U_{[k-t+2]}, \ldots, U_{[k]}$ as *t* best populations.

(2) **Selecting t Best Population When Σ_i Is Unknown**

If Σ_i is unknown, we let $V_i = \frac{(\bar{X}_i'S_i^{-1}\bar{X}_i)(n-p)}{np}$ and nV_i has the non-central F distribution with p, $(n-p)$ degrees of freedom and the non-centrality parameter of $n(\mu_i'\Sigma_i^{-1}\mu_i)$. We select the populations associated with $V_{[k-t+1]}, V_{[k-t+2]}, \ldots, V_{[k]}$ as *t* best populations.

2.2.4 Example of Bivariate Population with k=3

We simulated 80 data points from 3 bivariate normal populations with a mean of (2,4), (5,1), (10,3), and the common covariance matrix $\begin{bmatrix} 1.5 & 1 \\ 1 & 1.5 \end{bmatrix}$ as in 2.2.3. We select the best population for means.

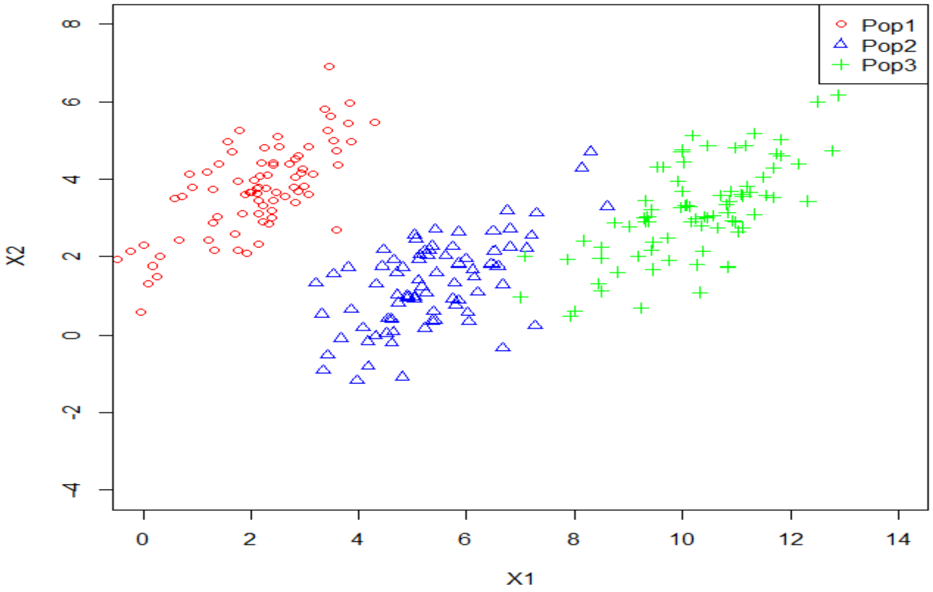


Figure 2.3: Simulated 3 Populations

We have sample data as follows.

Sample mean of Pop1 : (2.1186, 3.7382), Sample covariance of Pop1 : $\begin{bmatrix} 1.1466 & 0.8875 \\ 0.8875 & 1.3524 \end{bmatrix}$

Sample mean of Pop2 : (5.3593, 1.3112), Sample covariance of Pop2 : $\begin{bmatrix} 1.3223 & 0.8150 \\ 0.8150 & 1.3304 \end{bmatrix}$

Sample mean of Pop3 : (10.1987, 3.1717), Sample covariance of Pop3 : $\begin{bmatrix} 1.5712 & 1.0600 \\ 1.0600 & 1.4898 \end{bmatrix}$

Linear Combinations

If we set $\alpha_1 = 0.4, \alpha_2 = 0.6$, we have $\theta_i = 0.4 \times \mu_{i1} + 0.6 \times \mu_{i2}$ for i = 1,2,3. Then, $\hat{\theta_i} = 0.4 \times \bar{X_{i1}} + 0.6 \times \bar{X_{i2}}$ and

$$\hat{\theta_1} = 0.4 \times 2.1186 + 0.6 \times 3.7382 = 3.0904,$$

$$\hat{\theta_2} = 0.4 \times 5.3593 + 0.6 \times 1.3112 = 2.9304,$$

$$\hat{\theta_3} = 0.4 \times 10.1987 + 0.6 \times 3.1717 = 5.9825.$$

Thus, we select Population 3 as the best population.

Mahalanobis Distance with known variance

If we assume $P^* = 95\%$, $\delta_1 = 1$, and $\delta_2 = 2$, respectfully, the linear interpolation for $k = 3$ gives n= 43.148 from Table S.1 in [27]. We will use first the 44 data entries for calculation. The common covariance matrix is known as $\Sigma = \begin{bmatrix} 1.5 & 1 \\ 1 & 1.5 \end{bmatrix}$ and we have $\theta_i = \mu_i' \Sigma_i^{-1} \mu_i$ and $\hat{\theta_i} = \bar{x}_i' \Sigma_i^{-1} \bar{x}_i$. Then,

$$\hat{\theta_1} = (2.0906, 3.6532) \begin{pmatrix} 1.5 & 1 \\ 1 & 1.5 \end{pmatrix}^{-1} \begin{pmatrix} 2.0906 \\ 3.6532 \end{pmatrix} = 9.0401$$

$$\hat{\theta_2} = (5.4502, 1.4546) \begin{pmatrix} 1.5 & 1 \\ 1 & 1.5 \end{pmatrix}^{-1} \begin{pmatrix} 5.4502 \\ 1.4546 \end{pmatrix} = 25.5003$$

$$\hat{\theta_3} = (10.2069, 3.0655) \begin{pmatrix} 1.5 & 1 \\ 1 & 1.5 \end{pmatrix}^{-1} \begin{pmatrix} 10.2069 \\ 3.0655 \end{pmatrix} = 86.2313$$

Thus, we select Population 3 as the best one.

Mahalanobis Distance with unknown variance

Let's assume we don't know Σ. Then, we use the sample covariance matrix, S, to calculate the $\hat{\theta}_i$. $\hat{\theta}_i = \bar{x}_i' S_i^{-1} \bar{x}_i$. Given the same condition as the previous section, $P^* = 95\%$, $\delta_1 = 1$, and $\delta_2 = 2$, the sample size from Table 1 of [48] is 76.7. Then, we use the first 77 data entries for calculation.

$$\hat{\theta_1} = (2.1208, 3.7392) \begin{pmatrix} 1.1177 & 0.8388 \\ 0.8388 & 1.2974 \end{pmatrix}^{-1} \begin{pmatrix} 2.1208 \\ 3.7392 \end{pmatrix} = 10.9297$$

$$\hat{\theta_2} = (5.3416, 1.3598) \begin{pmatrix} 1.2918 & 0.8313 \\ 0.8313 & 1.2841 \end{pmatrix}^{-1} \begin{pmatrix} 5.3416 \\ 1.3598 \end{pmatrix} = 27.8495$$

$$\hat{\theta_3} = (10.15, 3.1203) \begin{pmatrix} 1.4744 & 0.9583 \\ 0.9583 & 1.4072 \end{pmatrix}^{-1} \begin{pmatrix} 10.15 \\ 3.1203 \end{pmatrix} = 85.2886$$

Thus, we select Population 3 as the best one.

2.3 Statistical Learning on Statistical Classification

2.3.1 Machine Learning and Statistical Learning

We already mentioned many times about the learning process. Statistical learning has originated from machine learning, which is a procedure that creates and develops algorithms to solve a problem using the given data set and answers it when a new data set is given based on the algorithm. From a given data set, when the training data set is entered into the machine, it returns the result. The more data entering into the machine, the better the machine develops the algorithm and returns improved results. This is where the learning takes place. In the learning process, if the data consists of input data and output data(the output data are labeled), the learning is called supervised learning. Based on the labeled output data, the machine learns and improves the algorithm. It's like you are preparing for the exam. You have a bunch of example problems with the answers. You solve the problem and check your answer with the given answer to see if it is correct or not. When you have the wrong answer, you check your procedure and fix it to get the correct answer. You do this process until you get all the correct answers and take the exam. This learning is supervised by yourself or a teacher and it's called supervised learning when you have input and output data. If there is no output data, it is called unsupervised learning. The main goal of supervised learning is to predict data, on the other hand, the goal of unsupervised learning is to find the associations or the hidden patterns from the unlabeled data. Classification is supervised learning and clustering is unsupervised learning. Then, what about the selection and ranking methodologies? We can categorize the indifference-zone approach as supervised learning and the subset selection methods as unsupervised learning.

Statistical learning theory is the field of machine learning with statistical inference involved. There is a probability distribution in the data space and we want to find a function that explains the association between the input variable and the output variable, in supervised learning. Then, we compare the predicted output from the function to the actual output and measure the difference as the loss. We find the best function that minimizes the loss, i.e., the difference between the predicted output and the actual output and it's where the learning takes place. This leads us to consider our approach to classification, a supervised learning, from the perspective of multiple decision theory. We will talk about the multiple decision theory in Chapter 4.

Chapter 3

The Statistical Classification and High Dimension

In this chapter, we propose one way to improve the performance of the classification process. From the multiple decision theoretic point of view, we will focus on the correct decision, and thus we want to increase the probability of correct classification. For two-dimensional data with two groups, a scatter plot can show areas where the two groups overlap. If we use linear discriminant analysis to find a classifier, there may be many misclassified points. To improve performance, we first focus on the classifier itself. We can improve the classifier by increasing the degree of the function to a polynomial function of degree 2 or higher. As a result, the classifier becomes more complex. Also, this will reduce the bias but increase the variance. The bias is the measure of the difference between the target value and the predicted value. The variance measures the expected difference of deviation from the actual value. The bias results from the model selection that is related to the assumptions. A higher bias model has more assumptions on the target function such as linear function. The lower bias model has fewer assumptions that can build a nonlinear function. High variance is observed by a huge inconsistency when changing the training data set. Overfitted target function leads to a high variance. Linear discrimination analysis shows low variance and decision trees, support vector machine, and k-nearest neighbor are with high variance. There exists the bias-variance tradeoff in the statistical learning procedure.

The above methods focused on improving the classifier to make fewer misclassified points and the techniques of classification have improved in this direction. Then, how about changing the way we approach this problem a bit? Instead of improving the classifier itself, let's look at this problem from a slightly different perspective. When there exist overlapping areas in given variables, can we add more variable(s) to the current problem and better separate the groups in the end? In an era of big data, finding new variables and adding them to a model is no longer that expensive. So adding a new variable and comparing performance would be an uncomplicated alternative. In this approach, we can improve classification performance by adding new variables rather than manipulating the classifier or reducing the dimensionality of the variables. The space of the problem will be expanded but we can have a hyperplane classifier that is still not too complicated to build. That is, what we propose is to increase the dimensionality of the variables by introducing a new variable that better discriminates groups in higher dimensions.

3.1 A Preferable Predictor Vector and The Probability of Correct Decision

We find the related variables using Data Mining or Machine Learning. Then, adding those variables to the classification model results in good separation of the groups. This means that the separability is improved and the probability of correct classification (or correct decision) is higher.

In this section, we want to focus on the probability of correct classification or the probability of correct decision. Suppose there are two populations (or groups) $\mathbf{G_1}$ and $\mathbf{G_2}$ and we use two variables X_1 and X_2 to classify the objects. Also, suppose there exists a linear classifying rule. Let's take a look at Figure 2.1 again. The blue line indicates the classifying rule. The correct decision happens when the object is correctly classified to the population where it comes from. Then, A and B are the areas where an incorrect decision could happen. Area A is where some objects are assigned to $\mathbf{G_1}$ even if they are from $\mathbf{G_2}$ by the given classification rule. In area B, some objects are assigned to $\mathbf{G_2}$ but they are originally from $\mathbf{G_1}$. Then, if we can find a variable (or a series of variables) that contributes separating the populations well and leads reducing such area in a higher dimensional space, we include the new vector to the problem. As a result, we have a new classification process that leads to a higher probability of a correct decision. Suppose, for example, a new variable, X_3, is added to the current object vector, (X_1, X_2), and this allows the overlapping regions under the 2-variable plot to be well separated as Figure 3.1 shows below.

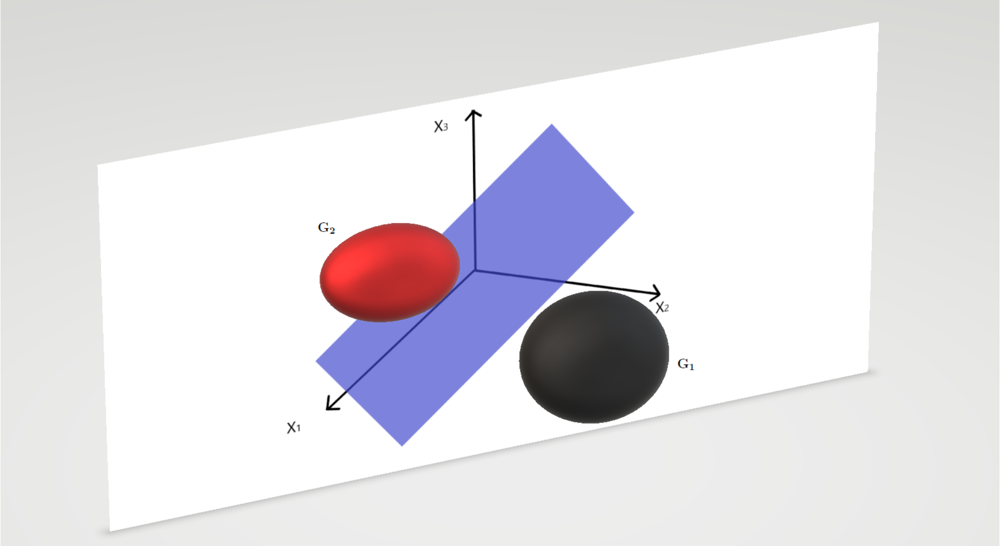


Figure 3.1: 3D Plot with X_3 Added

Consequently, by adding a new variable, two populations do not share any common area, as an extreme example, and the blue plane works as the classification rule. Here, we can notice that, as we introduce a new variable, the dimension of the classification rule also increases and the classification rule becomes a plane from a straight line. Then, we would like to compare the probability of correct decision of both cases, before and after adding X_3. If we can achieve a higher probability of correct classification by adding X_3, we need to find X_3 and include it in the classification process. We call such variables a Preferable Predictor Vector. It can be seen that observing fewer misclassified points means a higher probability of correct classification. Thus, we will compare the probability of misclassification, P(misclassification), for both cases and our goal is to have a smaller value for the case with X_3 added. Before checking the change of P(misclassification), we will take a look at the numerical result from the generated data.

3.2 Numerical Studies

We generated two sets of data using R from multivariate normal populations ($\mathbf{G_1}$ and $\mathbf{G_2}$). For simplicity, we start a multivariate normal distribution with two dimensions. There are 80 data points for each group. Let X_1 and X_2 be two variables that make the observation vector **X** = $(X_1, X_2)'$. The mean vectors are $\mu_\mathbf{1}$ and $\mu_\mathbf{2}$ and we assume the covariance matrices are diagonal, $\mathbf{\Sigma_1}$ and $\mathbf{\Sigma_2}$, for population 1 and population 2, respectively. The mean vector and covariance matrix for population 1, $\mathbf{G_1}$, are

$$\mu_1 = \begin{bmatrix} 5 \\ 2 \end{bmatrix}, \Sigma_1 = \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix}.$$

The mean vector and covariance matrix for population 2, $\mathbf{G_2}$, are

$$\mu_2 = \begin{bmatrix} 1 \\ 4 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 2 & 0 \\ 0 & 4 \end{bmatrix}.$$

Within each population, two variables (X_1 and X_2) are independent of each other. Based on the generated data, we first create a scatter plot: see Figure 3.2. Red points on the scatter plot are the data from $\mathbf{G_1}$ and blue points represent the data from $\mathbf{G_2}$. A solid black line is the example of the classifier and the misclassified observations are marked with observation numbers.

There are 19 misclassified observations altogether. There are 8 observations (9, 17, 30, 32, 41, 45, 52, 61 in red) from $\mathbf{G_1}$ but classified as $\mathbf{G_2}$ and 11 observations (17, 18, 26, 30, 55, 56, 58, 62, 72, 73, 77 in blue) from $\mathbf{G_2}$ but classified as $\mathbf{G_1}$.

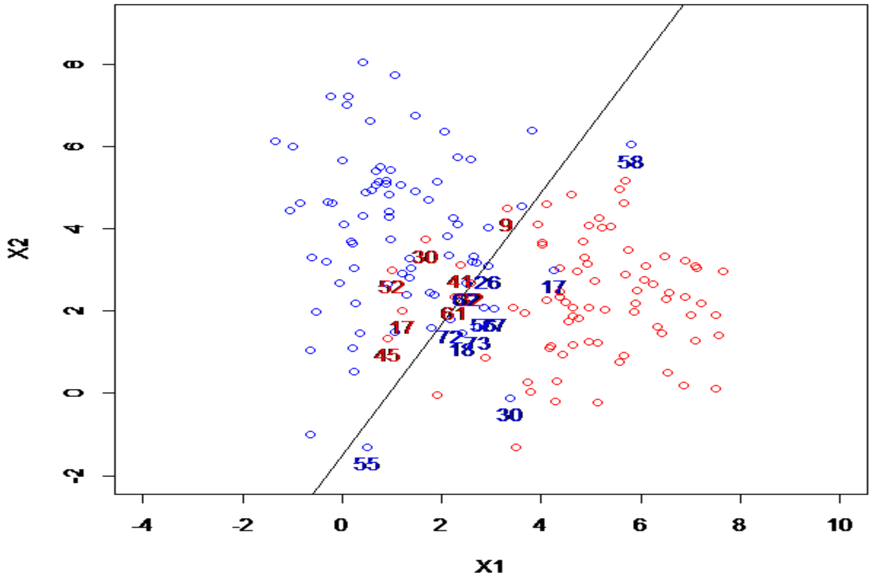


Figure 3.2: Scatter Plot of Multivariate Normal with Two Dimensions

Now, to increase the dimension of data, X_3 for each population is generated from a normal distribution. For $\mathbf{G_1}$, X_3 is generated from the population with a mean of 1 and a standard deviation of 1. For $\mathbf{G_2}$, the mean is 4 and the standard deviation is 1. Then, attach X_3 to the existing vector of **X** = (X_1, X_2) with the same order by assuming that X_3 is independent of (X_1, X_2). The table of generated data with X_3 added is provided in the **Appendix**.

The sample covariance matrices for $\mathbf{G_1}$ and $\mathbf{G_2}$ are

$$S_1 = \begin{bmatrix} 2.48763405 & 0.06649369 & -0.02741130 \\ 0.06649369 & 1.87922785 & 0.01407858 \\ -0.02741130 & 0.01407858 & 1.03841911 \end{bmatrix},$$

$$S_2 = \begin{bmatrix} 1.80773779 & -0.1651126 & -0.07293863 \\ -0.16511263 & 3.7808627 & 0.04513610 \\ -0.07293863 & 0.0451361 & 1.04058013 \end{bmatrix}.$$

From the sample covariance matrices, X_3 looks independent of X_1 and X_2.

The 3D plot of the data is shown below in Figure 3.3.

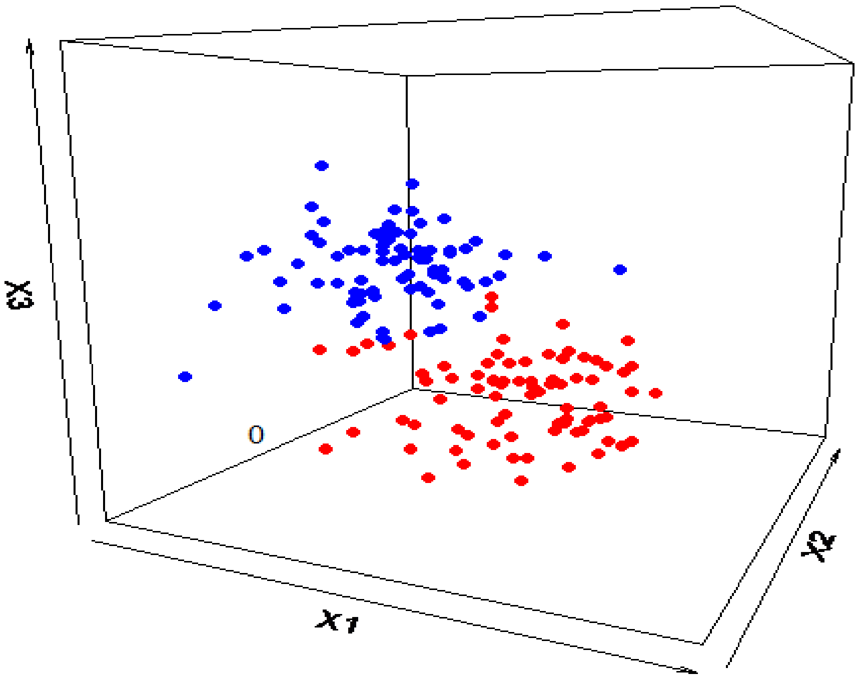


Figure 3.3: 3D Plot of Data

We observe that the data can be separated by a hyperplane classifier and the number of misclassified points are reduced.

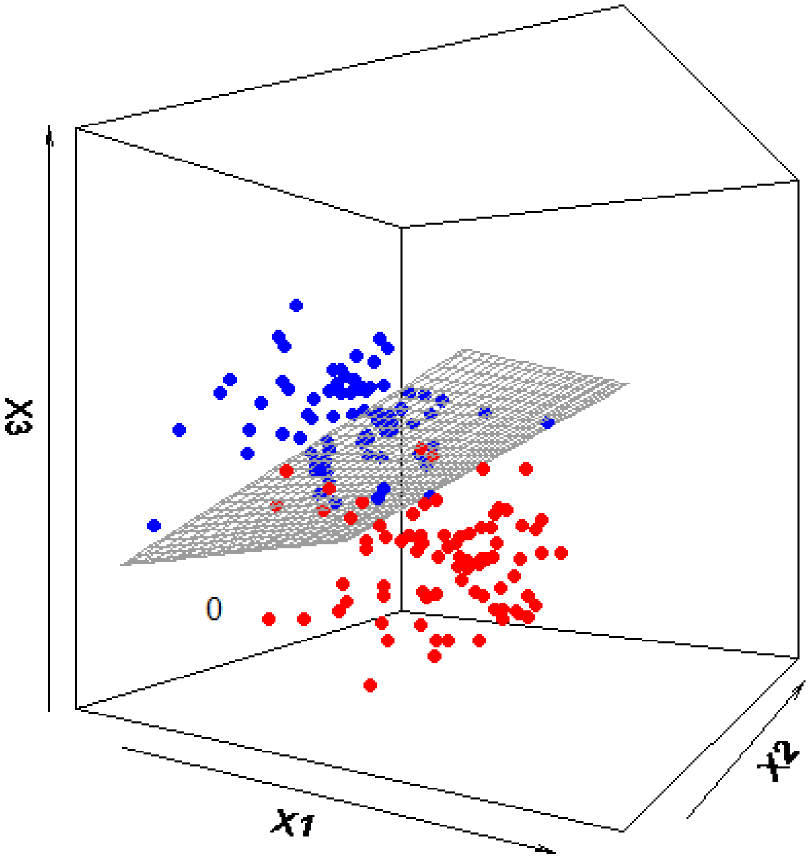


Figure 3.4: 3D Plot of Data with a Classifier

Here is a 3D plot with an arbitrary hyperplane classifier.

3D-plots of different angles are shown below.

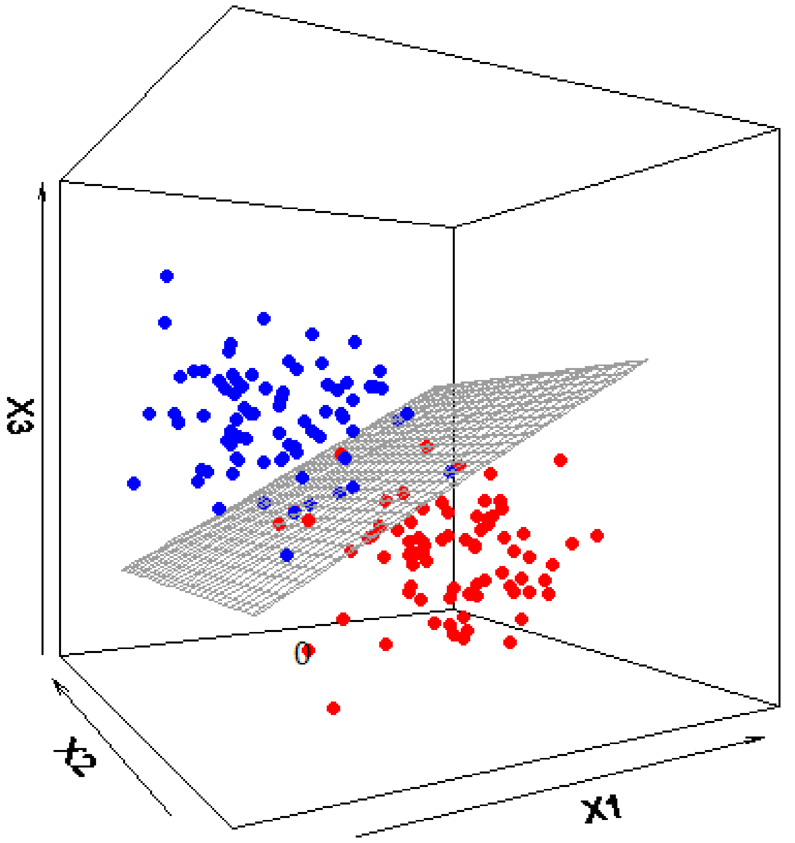


Figure 3.5: 3D Plot of Data with a Classifier from a Different Angle

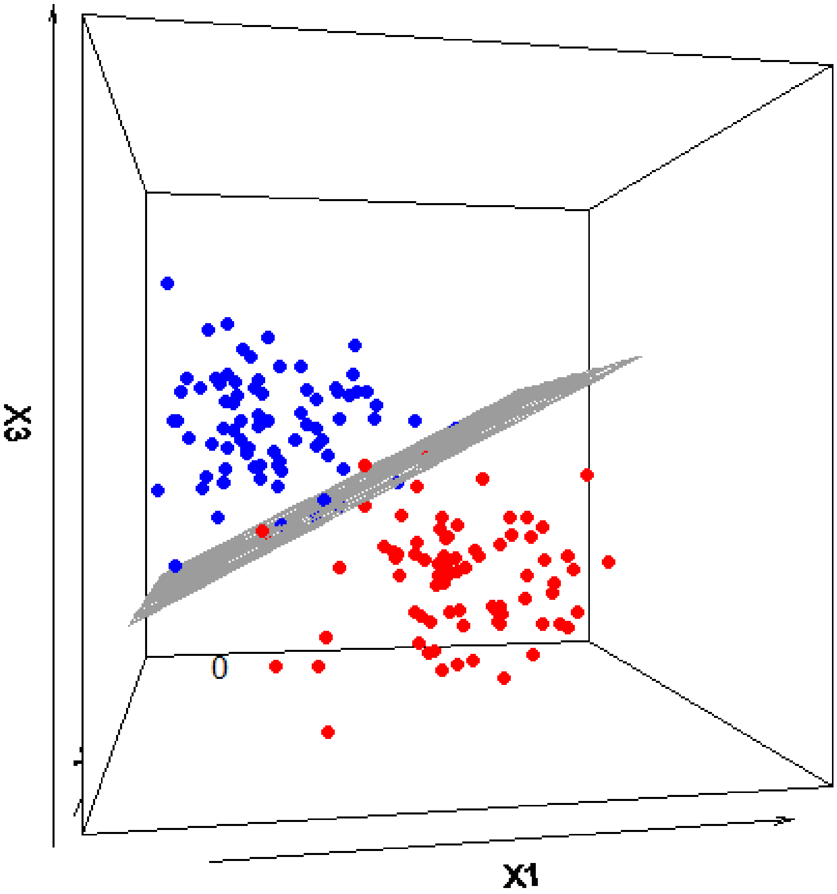


Figure 3.6: 3D Plot of Data with a Classifier with Another Angle

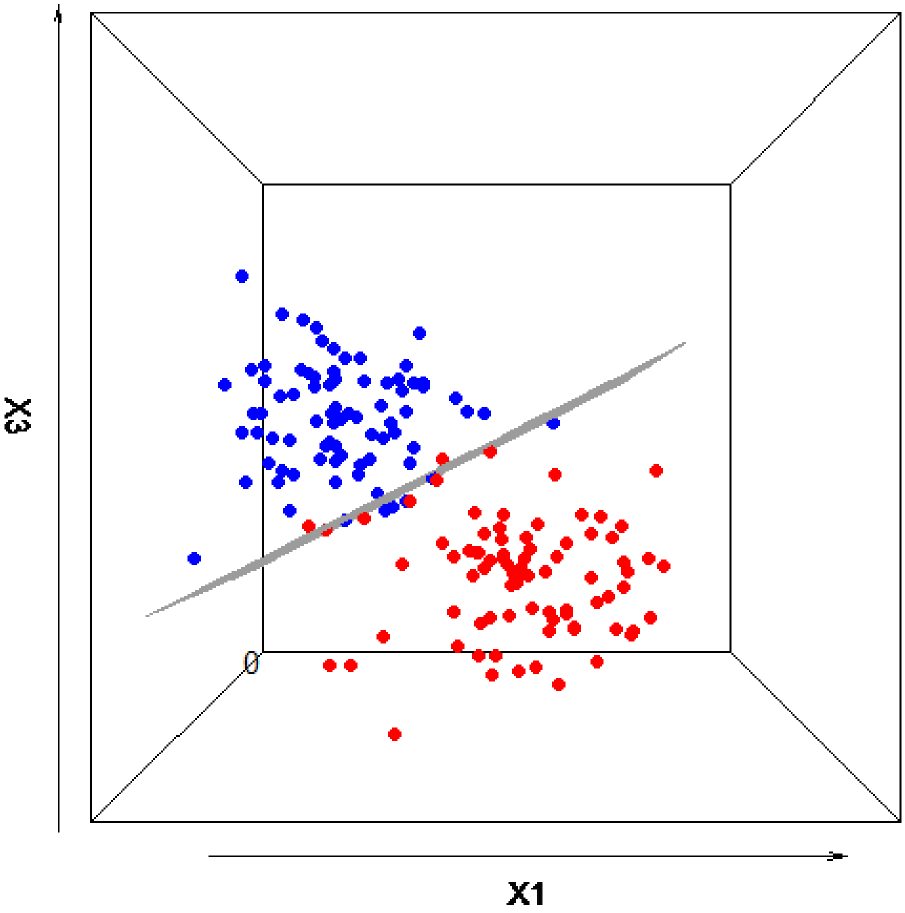


Figure 3.7: 3D Plot of Data with a Classifier with a Better Angle

We observe the much less misclassified points from several different angles.

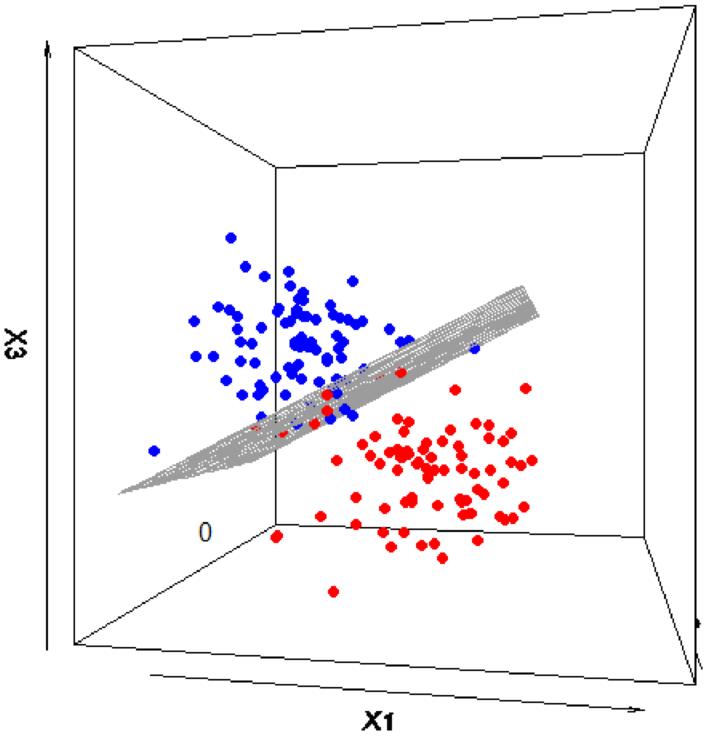


Figure 3.8: 3D Plot of Data with a Classifier with a Better Angle 2

There are only two misclassified red points and there are about 5 misclassified blue points. By introducing X_3 to the model, two populations can be separated with fewer misclassified points. Thus, we have no reason to hesitate to introduce a new variable to make a better classification. Now, we will compare the probabilities of misclassifications before and after adding X_3 using the total probability of misclassification (TPM) and the apparent error rate (APER).

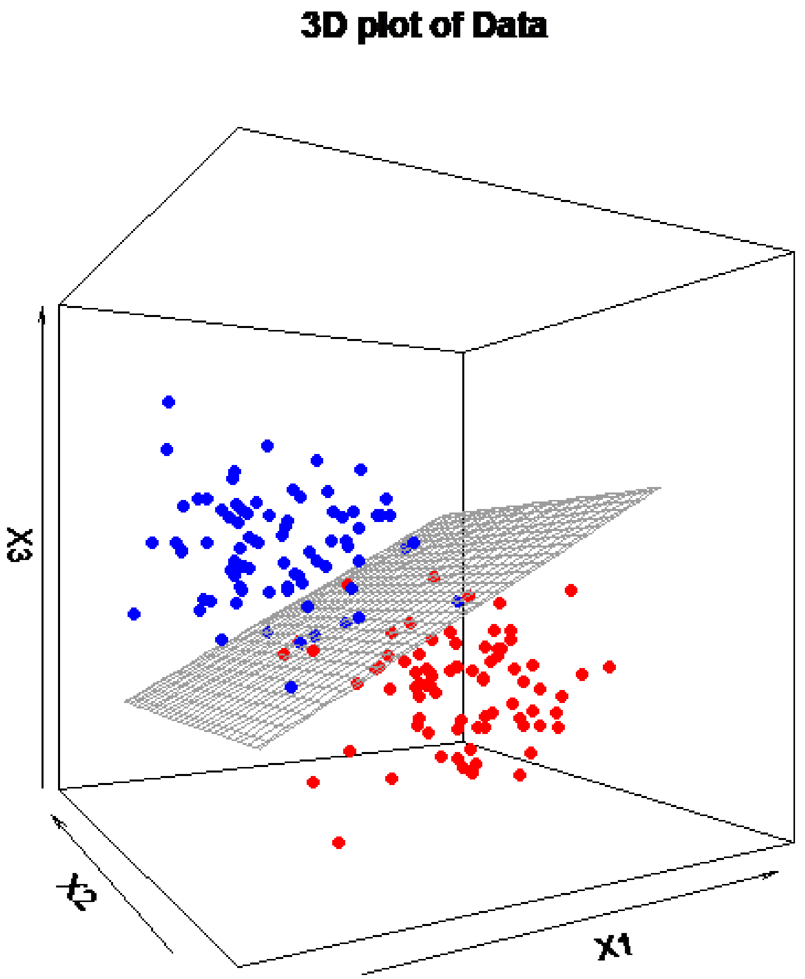


Figure 3.9: Animated 3D Plot of Data

3.3 View Point of Multiple Decision Theory

3.3.1 The Probability of Correct Decision from the Total Probability of Misclassification or the Apparent Error Rate

Multiple Decision Theory is concerned with making of decisions in the presence of statistical knowledge (data) which sheds light on some of the uncertainties involved in the decision problem. Statistical classification can be viewed as a decision-making procedure about allocating the observation to the correct group. When we classify the observation to the correct group, we make a correct decision. If we assign the observation to the group where it did not belong, we commit a wrong decision. Then, the probability of making a correct decision can be calculated from the total probability of misclassification when the distributions are provided or from the APER when we have no distributional information. The total probability of misclassification is the probability of making a wrong decision. Then, the probability of a correct decision can be calculated by $1 -$ TPM. Also, $1 -$ APER can be the probability of a correct decision.

3.4 A Preferable Predictor Vector and Calculation of TPM and APER

3.4.1 Calculation of TPM when $\Sigma_1 = \Sigma_2 = \Sigma$

We want to calculate a TPM of (2.1) to compare TPMs with or without the X_3 variable to check if adding X_3 improves classification. Under the same setup as Section 2.1.3, $\mathbf{G_1} \sim f_1(\cdot)$, $\mathbf{G_2} \sim f_2(\cdot)$, $P(\mathbf{G_1}) = p_1$, and $P(\mathbf{G_2}) = p_2$.

$$\begin{aligned} TPM &= P(\text{from } \mathbf{G_1} \text{ and misclassified }) + P(\text{ from } \mathbf{G_2} \text{ and misclassified}) \\ &= p_1 P(\mathbf{G_2}|\mathbf{G_1}) + p_2 P(\mathbf{G_1}|\mathbf{G_2}) \\ &= p_1 \int_{\mathcal{B}} f_1(\mathbf{x}) d\mathbf{x} + p_2 \int_{\mathcal{A}} f_2(\mathbf{x}) d\mathbf{x}. \end{aligned}$$

In (2.4), the classification rule for the multivariate normal distribution case was stated as a new observation of $\mathbf{x_0}$ is allocated to $\mathbf{G_1}$ if

$$(\mu_1 - \mu_2)'\Sigma^{-1}\mathbf{x_0} - \frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2) \geq \log\left(\frac{p_2}{p_1}\right), \tag{3.1}$$

where we assume the common covariance matrix under the multivariate normal distribution and p_1 and p_2 are prior probabilities. If we have equal prior probability, the right-hand side of the inequality (3.1) becomes 0. Then, we can calculate the probability of misclassifications from $\mathbf{G_1}$, $P(\mathbf{G_2}|\mathbf{G_1})$, as follows because we assign $\mathbf{x_0}$ to $\mathbf{G_2}$ if

$$(\mu_1 - \mu_2)'\Sigma^{-1}\mathbf{x_0} - \frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2) < 0. \tag{3.2}$$

$$P(\mathbf{G_2}|\mathbf{G_1}) = P((\mu_1 - \mu_2)'\Sigma^{-1}\mathbf{X} < \frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2)) \tag{3.3}$$

The left hand side of inequality, $(\mu_1 - \mu_2)'\Sigma^{-1}\mathbf{X}$ is a linear combination of p random variables, $l'\mathbf{X}$, so let's denote it as W, $W = l'\mathbf{X} = (\mu_1 - \mu_2)'\Sigma^{-1}\mathbf{X}$. Then,

$$P(\mathbf{G_2}|\mathbf{G_1}) = P(W < \frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2)), \tag{3.4}$$

where W= $(\mu_1 - \mu_2)'\Sigma^{-1}\mathbf{X} = l'\mathbf{X}$ and $\mathbf{X}\sim$ MVN(μ_1, Σ).

Since $\mathbf{X}\sim$ MVN(μ_1, Σ), we calculate and denote the mean and the variance of W as follows.

$$E(W) = (\mu_1 - \mu_2)'\Sigma^{-1}\mu_1 = l'\mu_1.$$

$$Var(W) = Var(l'\mathbf{X}) = l'\Sigma l = (\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2) = \sigma_W^2 = M^2.$$

Both are scalars and W follows a normal distribution. Thus,

$$\begin{aligned}
P(\mathbf{G_2}|\mathbf{G_1}) &= P\left(\frac{W-l'\mu_1}{\sigma_W} < \frac{\frac{1}{2}(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1+\mu_2)-l'\mu_1}{\sigma_W}\right)\\
&= P\left(\frac{W-l'\mu_1}{\sigma_W} < \frac{\frac{1}{2}(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1+\mu_2)-(\mu_1-\mu_2)'\Sigma^{-1}\mu_1}{\sigma_W}\right)\\
&= P\left(Z < \frac{\frac{1}{2}(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1)+\frac{1}{2}(\mu_1-\mu_2)'\Sigma^{-1}(\mu_2)-(\mu_1-\mu_2)'\Sigma^{-1}\mu_1}{\sigma_W}\right)\\
&= P\left(Z < \frac{-\frac{1}{2}(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1)+\frac{1}{2}(\mu_1-\mu_2)'\Sigma^{-1}(\mu_2)}{\sigma_W}\right) \qquad (3.5)\\
&= P\left(Z < \frac{-\frac{1}{2}(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1-\mu_2)}{\sigma_W}\right)\\
&= P\left(Z < \frac{-\frac{1}{2}M^2}{\sigma_W}\right)\\
&= \Phi(-\frac{M}{2}),
\end{aligned}$$

where $\Phi(\cdot)$ is the CDF of N(0,1) random variable and M=$\sqrt{(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1-\mu_2)}$.
In the same way, $P(\mathbf{G_1}|\mathbf{G_2}) = \Phi(-\frac{M}{2})$. Thus, the corresponding TPM, when the prior probabilities are equal, $p_1 = p_2 = 1/2$, is

$$TPM = \frac{1}{2}\Phi(-\frac{M}{2}) + \frac{1}{2}\Phi(-\frac{M}{2}) = \Phi(-\frac{M}{2}), \qquad (3.6)$$

where M =$\sqrt{(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1-\mu_2)}$. Thus, we need a bigger M to have less TPM.

If we use the example above with an assumption of a common Σ,

$$\mu_1 = \begin{bmatrix} 5 \\ 2 \end{bmatrix}, \mu_2 = \begin{bmatrix} 1 \\ 4 \end{bmatrix}, \text{ and assume that } \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 4 \end{bmatrix}.$$

If X_3 which is assumed to be independent of X_1 and X_2 is added,

$$\mu_{31} = \begin{bmatrix} 5 \\ 2 \\ 1 \end{bmatrix}, \mu_{32} = \begin{bmatrix} 1 \\ 4 \\ 4 \end{bmatrix}, \Sigma_3 = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Then, $M_1^2 = (\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2) = 9$ and $M_3^2 = (\mu_{31} - \mu_{32})'\Sigma_3^{-1}(\mu_{31} - \mu_{32}) = 18$.
TPM for $(X_1, X_2)'$ is 0.0668 and TPM for $(X_1, X_2, X_3)'$ is 0.01695. Thus, adding X_3 improves the probability of misclassification by almost 74.6%.

If we let the probability of correct classification be (1-TPM), we can derive the following rate to measure the improvement of the probability of correct classification.
Suppose we have $TPM_1 > TPM_2$.
The rate of improvement on TPM(%) = $\dfrac{TPM_1 - TPM_2}{TPM_1}$ x 100.
The rate in terms of the probability of correct classification(%)

$$=\frac{(1 - TPM_2) - (1 - TPM_1)}{(1 - TPM_1)} \text{ x } 100$$

$$=\frac{TPM_1 - TPM_2}{(1 - TPM_1)} \text{ x } 100.$$

There is a 5.34% improvement in probability of correct classification.

Then, we change the variance of the new variable from 0.1 to 20 to check whether the new TPM is still less than the TPM before adding X_3 and the plot is shown below. When the variance is 20, the TPM is 0.0621 and it is at least improved after adding X_3.

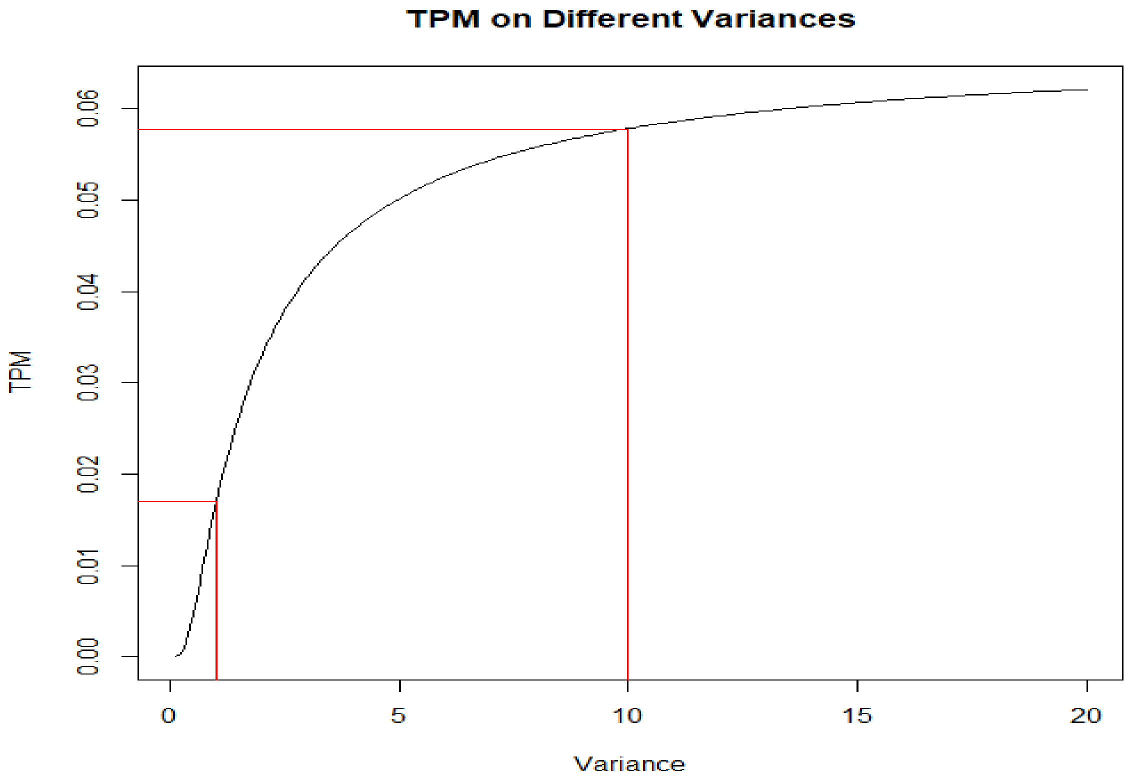


Figure 3.10: Changing the Variance from 0.1 to 20: At 10, TPM = 0.0578. At 20, TPM = 0.0621

3.4.2 Calculation of TPM when $\Sigma_1 \neq \Sigma_2$

From the density function (2.3) with Σ_1 and Σ_2 for two multivariate normal populations, the classification rule for a new observation, $\mathbf{x_0}$, is that $\mathbf{x_0}$ is allocated to $\mathbf{G_1}$ if

$$\frac{f_1(\mathbf{x_0})}{f_2(\mathbf{x_0})} \geq \frac{p_2}{p_1}, \text{ i.e.,}$$

$$-\frac{1}{2}\mathbf{x_0}'(\Sigma_1^{-1} - \Sigma_2^{-1})\mathbf{x_0} + (\mu_1'\Sigma_1^{-1} - \mu_2'\Sigma_2^{-1})\mathbf{x_0} - \beta \geq \log\left(\frac{p_2}{p_1}\right), \tag{3.7}$$

where $\beta = \frac{1}{2}\log\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) + \frac{1}{2}\left(\mu_1'\Sigma_1^{-1}\mu_1 - \mu_2'\Sigma_2^{-1}\mu_2\right)$ by the minimum TPM rule. If the prior probabilities are the same, the right-hand side of the inequality (3.7) is 0 again. Then, the probability of misclassifications from $\mathbf{G_1}$, $P(\mathbf{G_2}|\mathbf{G_1})$, can be calculated as follows since we allocate $\mathbf{x_0}$ to $\mathbf{G_2}$

if

$$-\frac{1}{2}\mathbf{x_0}'(\Sigma_1^{-1}-\Sigma_2^{-1})\mathbf{x_0}+(\mu_1'\Sigma_1^{-1}-\mu_2'\Sigma_2^{-1})\mathbf{x_0}-\beta<0. \tag{3.8}$$

Thus,

$$P(\mathbf{G_2}|\mathbf{G_1})=P(\mathbf{X}'(\Sigma_1^{-1}-\Sigma_2^{-1})\mathbf{X}-2(\mu_1'\Sigma_1^{-1}-\mu_2'\Sigma_2^{-1})\mathbf{X}>-2\beta), \tag{3.9}$$

where $\mathbf{X}\sim MVN(\mu_1,\Sigma_1)$ and β as above in (3.7).
In the same way,

$$P(\mathbf{G_1}|\mathbf{G_2})=P(\mathbf{X}'(\Sigma_1^{-1}-\Sigma_2^{-1})\mathbf{X}-2(\mu_1'\Sigma_1^{-1}-\mu_2'\Sigma_2^{-1})\mathbf{X}\leq-2\beta), \tag{3.10}$$

where $\mathbf{X}\sim MVN(\mu_2,\Sigma_2)$ and β as above in (3.7).
The probability here involves the quadratic functions of X. The first term is the quadratic form of a multivariate normal random variable and the second term is the linear combination of normal random variables, i.e., the random variable in this probability is the sum of noncentral chi-squared random variables and normal random variables and it follows the generalized chi-squared distribution. We will use the generalized chi-squared distribution to calculate the TPM for the case of $(X_1,X_2)'$ and the case of $(X_1,X_2,X_3)'$ by using the numerical result above.

Calculation of $P(\mathbf{G_2}|\mathbf{G_1})$

From (3.9),

$$\begin{aligned}P(\mathbf{G_2}|\mathbf{G_1})&=P(\mathbf{X}'(\Sigma_1^{-1}-\Sigma_2^{-1})\mathbf{X}-2(\mu_1'\Sigma_1^{-1}-\mu_2'\Sigma_2^{-1})\mathbf{X}>-2\beta)\\&=P((\mathbf{X}-h)'(\Sigma_1^{-1}-\Sigma_2^{-1})(\mathbf{X}-h)-h'(\Sigma_1^{-1}-\Sigma_2^{-1})h>-2\beta)\\&=P((\mathbf{X}-h)'(\Sigma_1^{-1}-\Sigma_2^{-1})(\mathbf{X}-h)>h'(\Sigma_1^{-1}-\Sigma_2^{-1})h-2\beta),\end{aligned} \tag{3.11}$$

where $h=-\frac{1}{2}(\Sigma_1^{-1}-\Sigma_2^{-1})^{-1}(-2(\mu_1'\Sigma_1^{-1}-\mu_2'\Sigma_2^{-1}))'=(\Sigma_1^{-1}-\Sigma_2^{-1})^{-1}(\mu_1'\Sigma_1^{-1}-\mu_2'\Sigma_2^{-1})'$ and let $C=h'(\Sigma_1^{-1}-\Sigma_2^{-1})h-2\beta$. The second equality holds because we make a complete square form in terms of **X**. Then,

$$P(\mathbf{G_2}|\mathbf{G_1}) = P((\mathbf{X}-h)'(\Sigma_1^{-1}-\Sigma_2^{-1})(\mathbf{X}-h) > C).$$

Here, we denote $(\mathbf{X}-h)$ as Q and $(\Sigma_1^{-1}-\Sigma_2^{-1})$ as A. Then,

$$P(\mathbf{G_2}|\mathbf{G_1}) = P(Q'AQ > C), \text{ where } Q \sim MVN(\mu_1 - h, \Sigma_1)$$

because X is from $\mathbf{G_1}$.

$Q'AQ$ is the quadratic form of **X** and can be written as the linear combination of the noncentral chi-squared variables. Let $\mathbf{Y} = \Sigma_1^{-\frac{1}{2}}(\mathbf{X}-h) = \Sigma_1^{-\frac{1}{2}}Q$. Then, $E(\mathbf{Y}) = \Sigma_1^{-\frac{1}{2}}(\mu_1 - h)$. Also, let $\mathbf{Z} = \mathbf{Y} - \Sigma_1^{-\frac{1}{2}}(\mu_1 - h) = \mathbf{Y} - E(\mathbf{Y})$. Then, $E(\mathbf{Z}) = \mathbf{0}$ and $Var(\mathbf{Z}) = Var(\mathbf{Y}) = I$.

Then,

$$\mathbf{Z} = \mathbf{Y} - \Sigma_1^{-\frac{1}{2}}(\mu_1 - h) = \Sigma_1^{-\frac{1}{2}}(\mathbf{X}-h) - \Sigma_1^{-\frac{1}{2}}(\mu_1 - h) \text{ and}$$

$$(\mathbf{X}-h) = \Sigma_1^{\frac{1}{2}}(\mathbf{Z} + \Sigma_1^{-\frac{1}{2}}(\mu_1 - h)).$$

Thus,

$$\begin{aligned} Q'AQ &= (\mathbf{X}-h)'(\Sigma_1^{-1}-\Sigma_2^{-1})(\mathbf{X}-h) \\ &= (\mathbf{Z} + \Sigma_1^{-\frac{1}{2}}(\mu_1 - h))'\Sigma_1^{\frac{1}{2}}A\Sigma_1^{\frac{1}{2}}(\mathbf{Z} + \Sigma_1^{-\frac{1}{2}}(\mu_1 - h)), \end{aligned} \tag{3.12}$$

By spectral theorem, $\Sigma_1^{\frac{1}{2}}A\Sigma_1^{\frac{1}{2}} = P'\Lambda P$, where P is an orthogonal matrix and Λ is a diagonal matrix of the eigenvalues, λ_i. Then,

$$\begin{aligned} Q'AQ &= (Z + \Sigma_1^{-\frac{1}{2}}(\mu_1 - h))'P'\Lambda P(Z + \Sigma_1^{-\frac{1}{2}}(\mu_1 - h)) \\ &= (PZ + P\Sigma_1^{-\frac{1}{2}}(\mu_1 - h))'\Lambda(PZ + P\Sigma_1^{-\frac{1}{2}}(\mu_1 - h)). \end{aligned} \tag{3.13}$$

Let $U = PZ$ and $b = P\Sigma_1^{-\frac{1}{2}}(\mu_1 - h)$. Then, $U \sim MVN(0, I_n)$ because P is the orthogonal matrix, i.e., $PP' = P'P = I$. E(PZ)=PE(Z) = 0

Then,

$$Q'AQ = (U+b)'\Lambda(U+b) = \sum_{j=1}^{n} \lambda_i (U_j + b_j)^2.$$

$Q'AQ$ is a linear combination of noncentral chi-squared variables, and thus

$$P(\mathbf{G_2}|\mathbf{G_1}) = P(Q'AQ > C) = P((U+b)'\Lambda(U+b) > C). \tag{3.14}$$

Calculation of $P(\mathbf{G_1}|\mathbf{G_2})$

In a similar way, $P(\mathbf{G_1}|\mathbf{G_2}) = P(Q'AQ \leq C)$, where $Q \sim MVN(\mu_2 - h, \Sigma_2)$ because X is from $\mathbf{G_2}$. Then,

$$P(\mathbf{G_1}|\mathbf{G_2}) = P((U+b)'\Lambda(U+b) \leq C)$$

but b and Λ here are different from those in (3.14).

Here, $\mathbf{Y} = \Sigma_2^{-\frac{1}{2}}(\mathbf{X} - h)$. Then, $E(\mathbf{Y}) = \Sigma_2^{-\frac{1}{2}}(\mu_2 - h)$ and $\mathbf{Z} = \mathbf{Y} - \Sigma_2^{-\frac{1}{2}}(\mu_2 - h) = \mathbf{Y} - E(\mathbf{Y})$. From Spectral Theorem, $\Sigma_2^{\frac{1}{2}} A \Sigma_2^{\frac{1}{2}} = P'\Lambda P$. Thus, Λ and P are from this decomposition and $U = PZ$ and $b = P\Sigma_2^{-\frac{1}{2}}(\mu_2 - h)$.

Suppose we have a similar example in Section 3.2 with different covariance matrices

$$\mu_1 = \begin{bmatrix} 5 \\ 2 \end{bmatrix}, \Sigma_1 = \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix}, \mu_2 = \begin{bmatrix} 1 \\ 4 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 2 & 0 \\ 0 & 4 \end{bmatrix}.$$

Also, if X_3 which is assumed to be independent of X_1 and X_2 is added, we have

$$\mu_{31} = \begin{bmatrix} 5 \\ 2 \\ 1 \end{bmatrix}, \Sigma_{31} = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \mu_{32} = \begin{bmatrix} 1 \\ 4 \\ 4 \end{bmatrix}, \Sigma_{32} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 0.5 \end{bmatrix}.$$

Here, we set the new variance for population 2 as 0.5 (not 1) to avoid a singular problem in the calculation. For $(X_1, X_2)'$ case, $P(\mathbf{G_2}|\mathbf{G_1})$ = 0.08611 and $P(\mathbf{G_1}|\mathbf{G_2})$ = 0.0737, then TPM = 0.5(0.0861 + 0.0737) = 0.0799. When X_3 is added, $P(\mathbf{G_2}|\mathbf{G_1})$ = 0.01367 and $P(\mathbf{G_1}|\mathbf{G_2})$ =

0.01018, then TPM = 0.5(0.01367 + 0.01018) = 0.011925. Thus, by adding X_3, the total probability of misclassification(TPM) significantly decreases. It improves the probability of misclassification by 85.1%. The rate in terms of the probability of correct classification is $(0.0799 - 0.0119)/(1 - 0.0799) = 0.0739$ and there is a 7.39% improvement. Also, we change the variance of population 2 to check how the TPM changes as the variance differs. The plot is shown below.

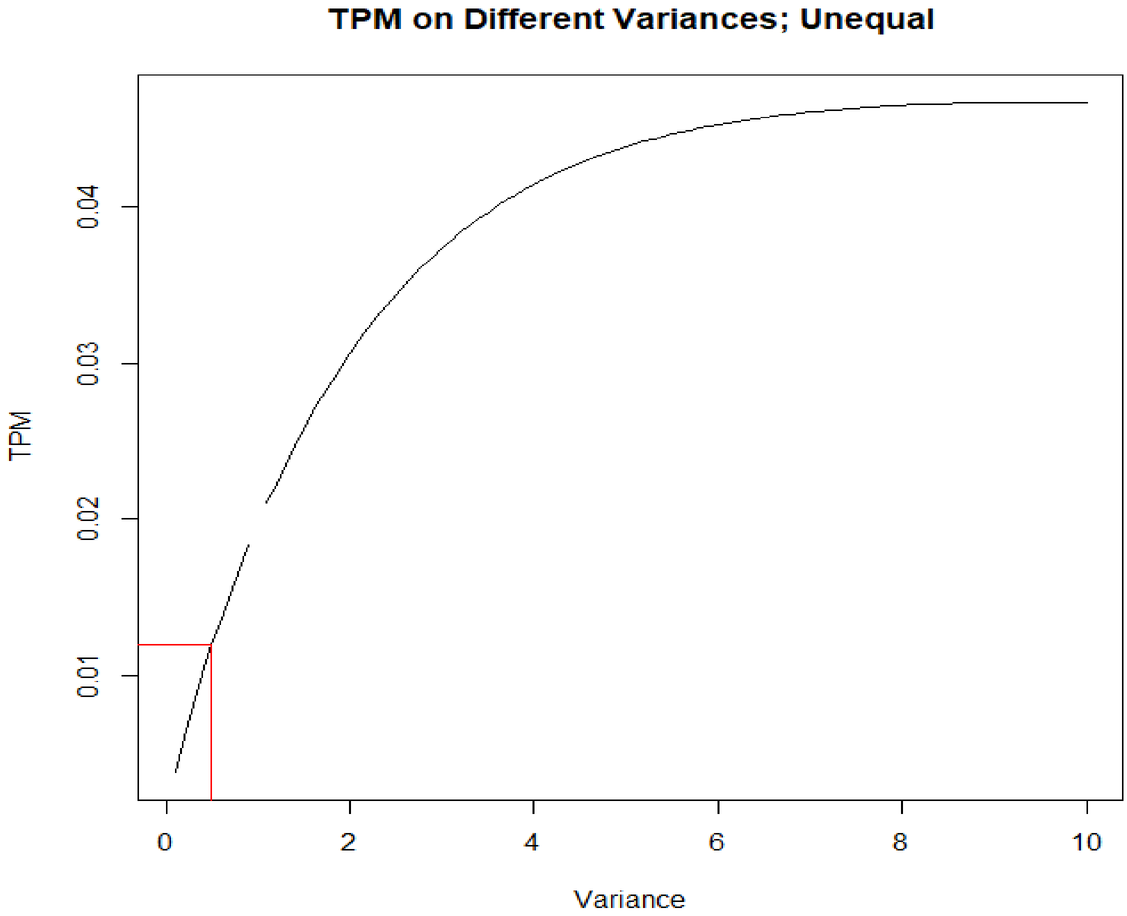


Figure 3.11: Changing the Variance from 0.1 to 10: At 10, TPM = 0.0467.

3.4.3 Calculation of APER

The APERs are calculated from the numerical result in Section 3.2.

X_1 and X_2		Assigned to		
		$\mathbf{G_1}$	$\mathbf{G_2}$	
Observations from	$\mathbf{G_1}$	72	8	80
	$\mathbf{G_2}$	11	69	80

Table 3.1: Confusion Matrix Before Adding A Preferable Predictor Vector

The apparent error rate (APER) = $\frac{8+11}{80+80} = \frac{19}{160} = 0.11875$

X_1, X_2 and X_3		Assigned to		
		$\mathbf{G_1}$	$\mathbf{G_2}$	
Observations from	$\mathbf{G_1}$	78	2	80
	$\mathbf{G_2}$	5	75	80

Table 3.2: Confusion Matrix After Adding a Preferable Predictor Vector

The apparent error rate (APER) = $\frac{2+5}{80+80} = \frac{7}{160} = 0.0438$

There is a 63.12$\%$ drop in APER by adding X_3. The Accuracy has increased by 8.51$\%$ from 0.88125 to 0.95625. The probability of a correct decision has been increased by adding a preferable predictor vector.

Below are the confusion matrices when LDA and QDA are applied before and after adding the preferable predictor vector from data in Section 3.2. lda and qda functions in R are used to make the confusion matrices.

		Assigned to		
		Pop1	Pop2	
Observations from	Pop1	71	9	80
	Pop2	7	73	80

Table 3.3: Confusion Matrix of 2D LDA

		Assigned to		
		Pop1	Pop2	
Observations from	Pop1	76	4	80
	Pop2	1	79	80

Table 3.4: Confusion Matrix of LDA with Preferable Predictor Vector

		Assigned to		
		Pop1	Pop2	
Observations from	Pop1	71	9	80
	Pop2	8	72	80

Table 3.5: Confusion Matrix of 2D QDA

		Assigned to		
		Pop1	Pop2	
Observations from	Pop1	75	5	80
	Pop2	0	80	80

Table 3.6: Confusion Matrix of QDA with Preferable Predictor Vector

The table below shows APERs calculated from four confusion matrices above.

	2D	3D
LDA	0.1	0.031
QDA	0.106	0.031

Table 3.7: Table of APERs

In both cases of LDA and QDA, by adding a preferable predictor vector, there are almost 70% drops in APER. The accuracy increases by 7.67% and 8.39% each for LDA and QDA, respectively. We can notice that classification performance is improved. However, there is almost no improvement when the classification method was changed from LDA to QDA for this example.

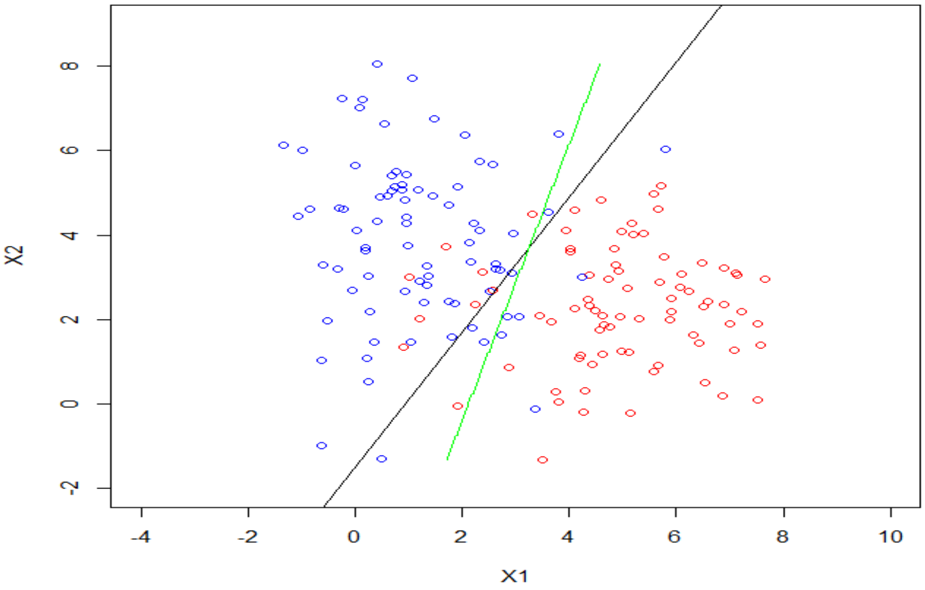


Figure 3.12: Figure 3.2 Added by LDA Classifier (Green Line)

3.4.4 Summary

We used the TPM and the APER when we showed the improvement in correct classification by adding a new variable, a preferable predictor vector. When we calculate the TPM under the assumptions of distribution with homogeneous variance, the classification rule becomes a linear combination of predictor variables. So, it uses the values on a projected line and it means the calculation is always reduced to 1-dimension. We also calculated TPMs under the assumption of a non-homogeneous variance structure. When we use APER without the distribution assumptions, however, we created the hyperplane as a classifying rule and counted the incorrectly classified observations. So, there is a difference in the dimension of the classifying rule between the two methods. We can use SVM to find the hyperplane classifier for APER calculation. We also used LDA and QDA, under the assumption of the distribution, to create the confusion matrix and calculate APERs. In all cases, the preferable predictor vector helps to increase the probability of correct classification.

3.5 Conditions of the Preferable Predictor Vector.

We can suggest a few ways to search for the preferable predictor vector. First, we use the neighboring data set. From the existing variables, we extract some features and characteristics, then look for a variable that possesses the related information. Second, we do data mining from the big data. By setting some conditions on the data set, we can collect the variables that seem to be useful for our classification. Once we are ready with potential preferable predictor vectors, try each vector in turn and label one as a preferable predictor vector if the vector improves the probability of misclassification or as an inferior vector otherwise. In the following section, we will investigate the conditions of the preferable predictor vector when minimizing the total probability of misclassification (TPM) is used as the rule for classification.

3.5.1 Preferable Predictor Variable chosen by Statistical Learning

Suppose we have two populations as the picture shows below.

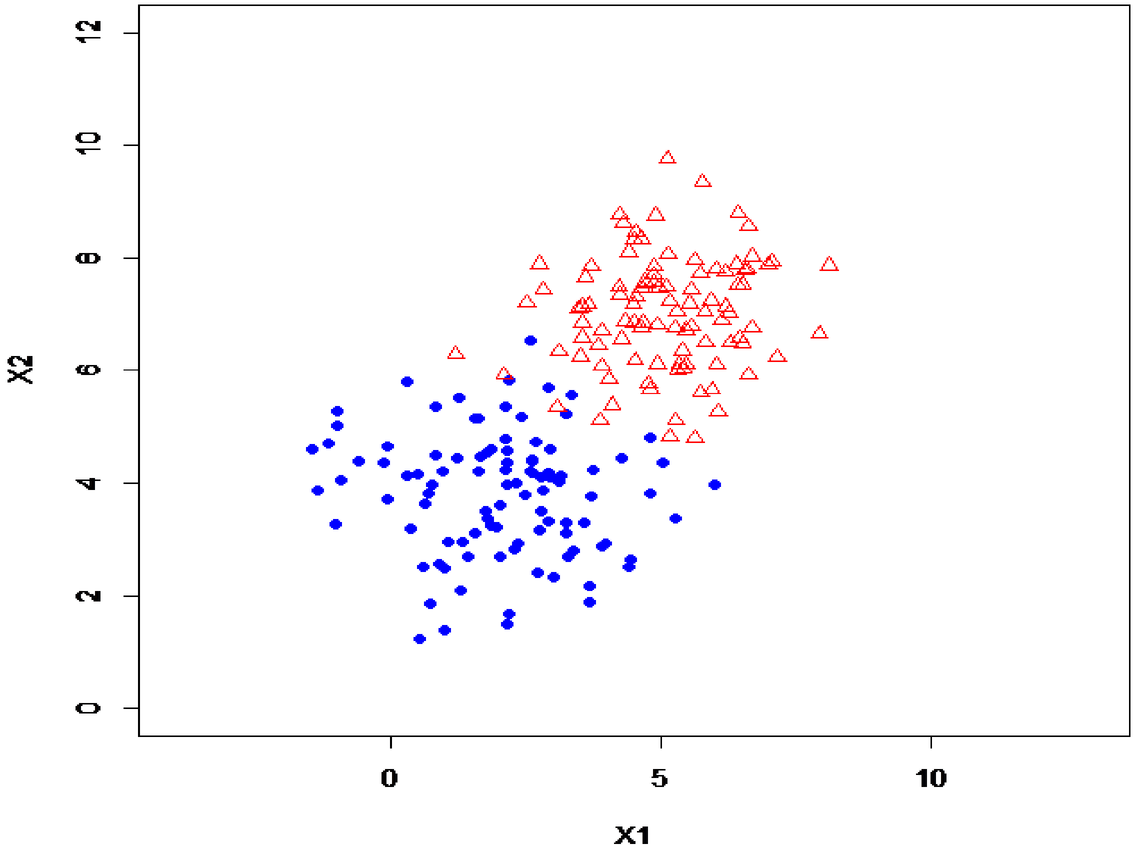


Figure 3.13: Two Populations with Overlapped Area in 2D

We assume that two populations have the same covariance structure. If there exists X_3 that makes the following graph, the separation becomes easier and we want to find such a variable(variables).

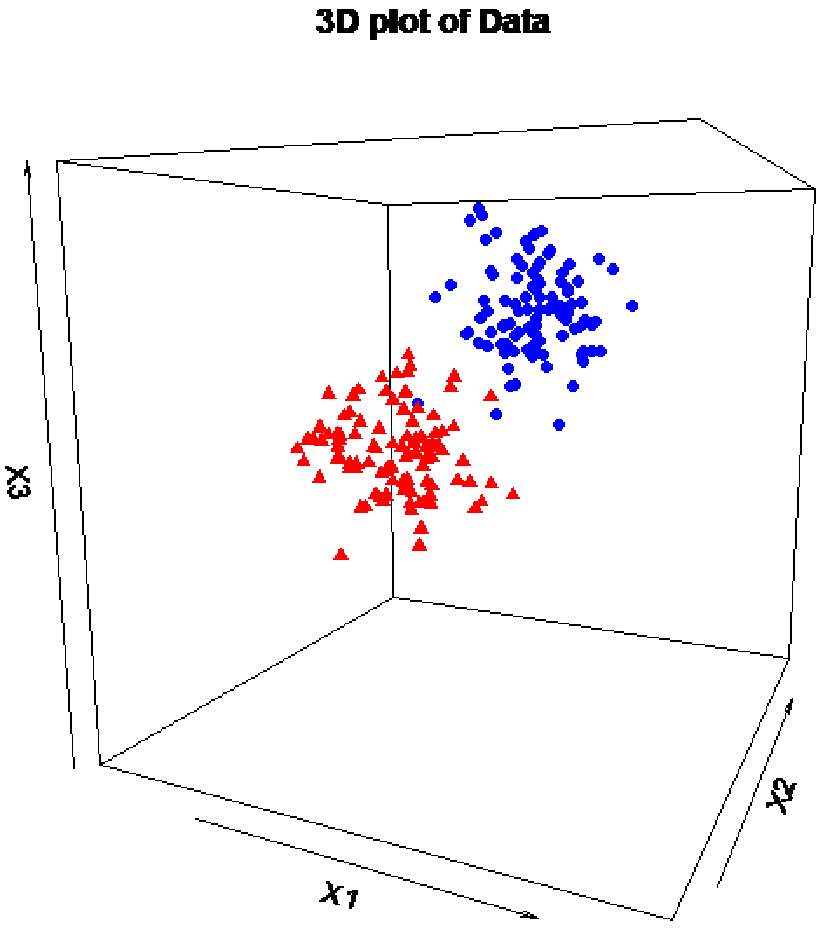


Figure 3.14: Two Populations Separated by X_3

To get the above plot, we need to add an X_3 variable to the model and we need to consider two parts when we choose X_3. The variance of the X_3 variable and the difference in means of X_3 between two populations. For example, if there is no variation within the X_3 variable, i.e., X_3 is a constant, then it is not too difficult to separate two populations with a small difference in means of the X_3 variable between two populations. The picture below shows the case with no variation in the X_3 variable and the mean difference of 3 between the two populations.

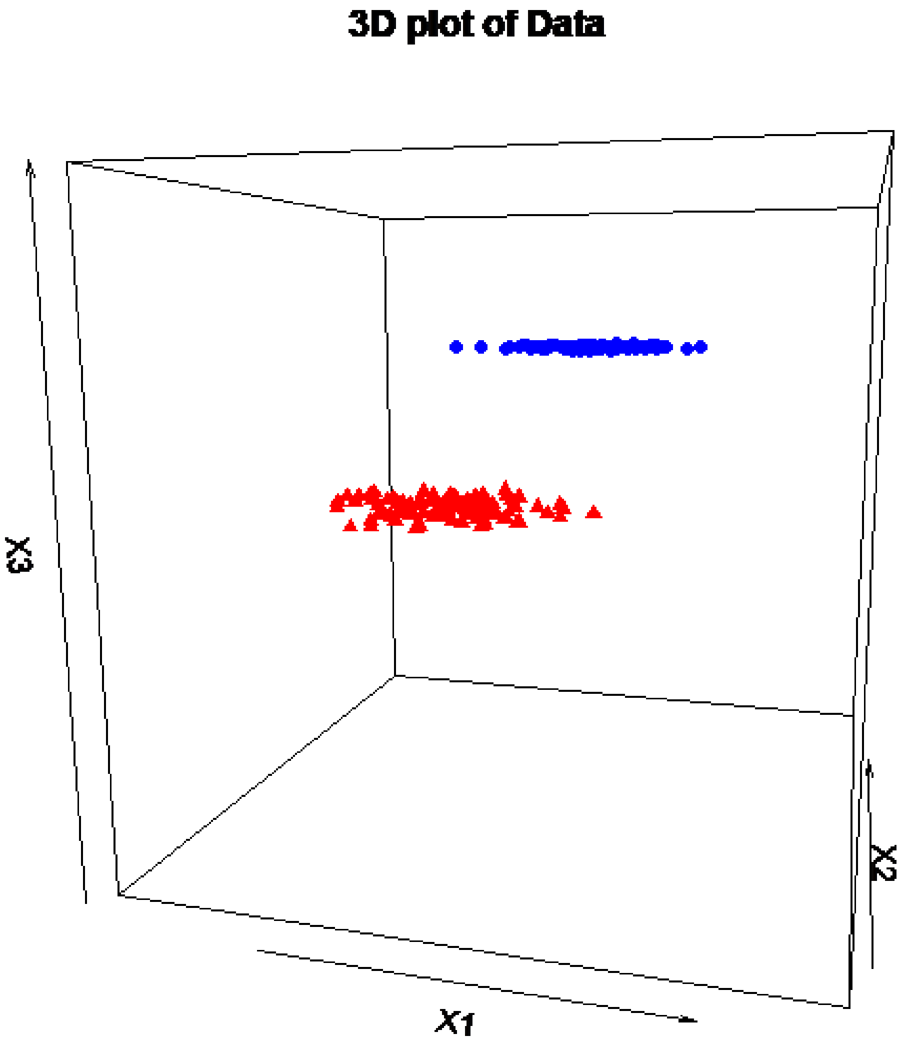


Figure 3.15: Variance of X_3 is Zero

Even when the difference of means is 1, it is still easy to find the separating hyperplane. The following picture shows the cases with the difference of 1 through 4.

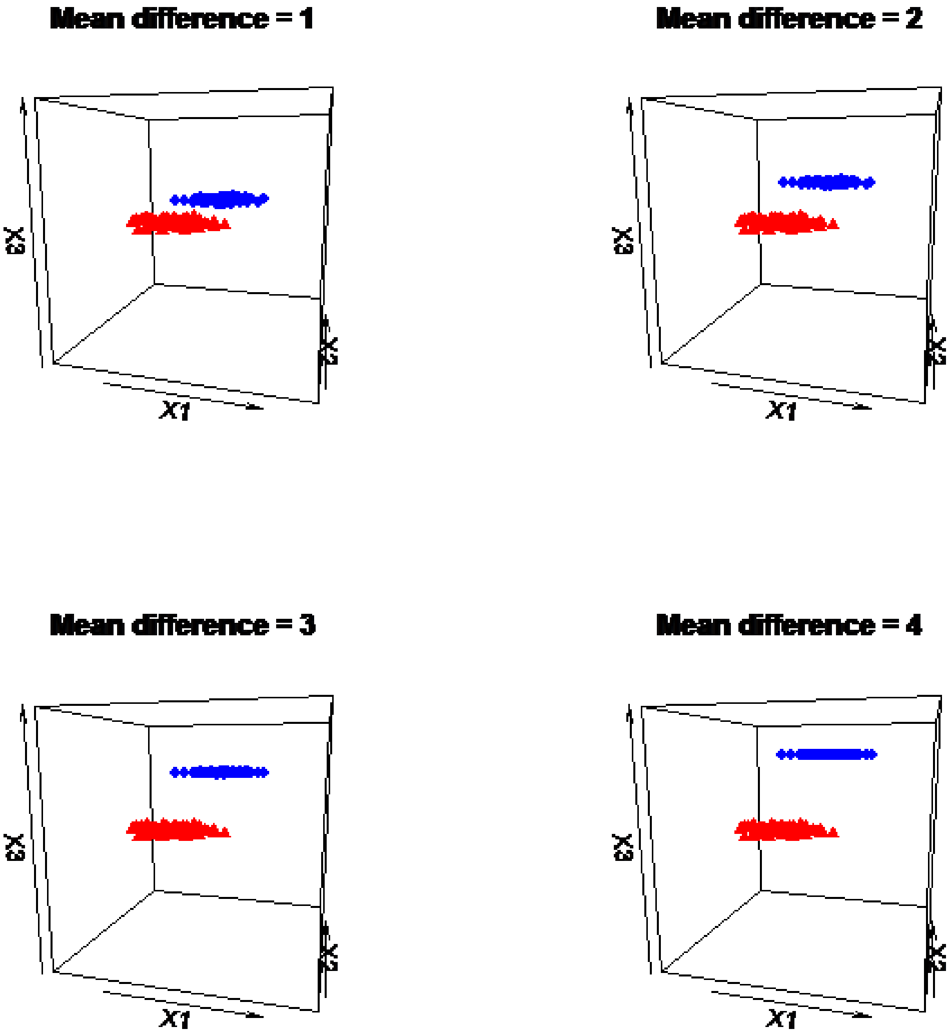


Figure 3.16: Two Populations Separated by Various Mean Differences

Thus, if the X_3 variable has no variation at all, we can add X_3 to the model and find the separating hyperplane with a slight difference in means of X_3 between two populations. If we add X_3 with a variance of 1 and mean of 4 and 7 for each population, respectively, we have the following picture.

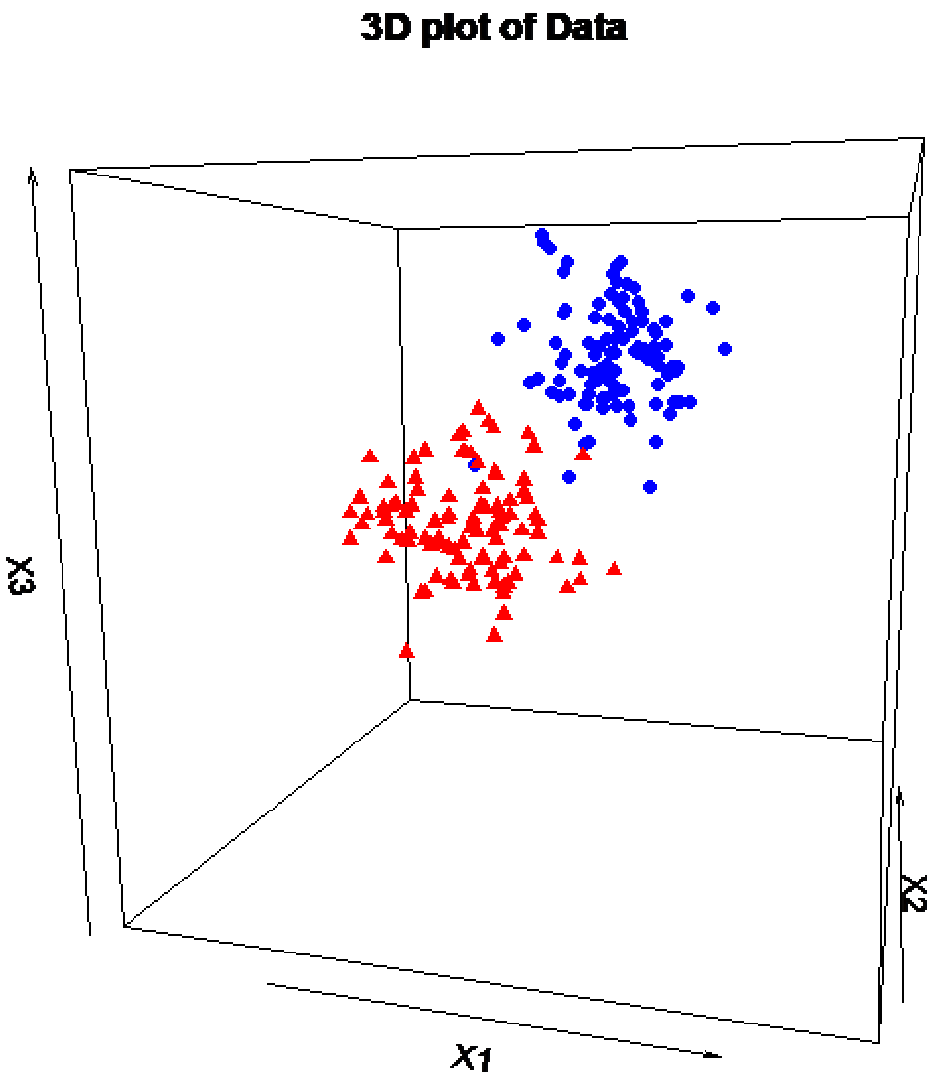


Figure 3.17: Two Populations with Variance = 1 and Mean Difference =3

Plots of different variances given the same mean difference is shown below.

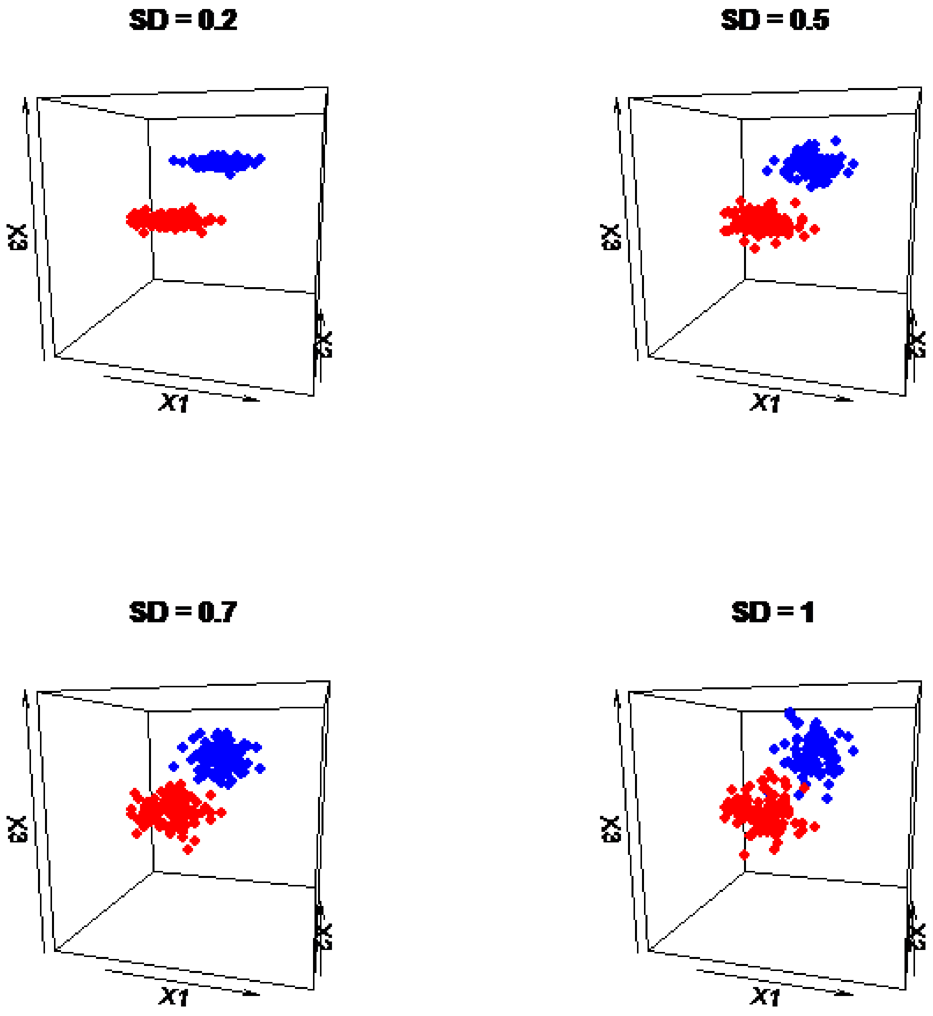


Figure 3.18: Various Variances with Mean Difference = 3

Therefore, finding an appropriate variable X_3 involves the variance of the X_3 and the mean difference in X_3 between the two populations. We can find such variable X_3 by statistical learning. We want to investigate further on the conditions in the next section. We also need to check the correlation between X_3 and X_1 or X_2, the new variable, and the existing variable(s).

3.5.2 Conditions of the Preferable Predictor Vector With Calculation of TPM

Now, we need to compare the TPM of the bivariate normal case to the TPM of a multivariate normal with X_3 added. The TPM of bivariate normal is $\Phi(-\frac{M}{2})$ from (3.6), where M=$\sqrt{(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2)}$ when we assume the common covariance matrix. Then, we denote the TPM with X_3 as $\Phi(-\frac{M_3}{2})$ and need to have

$$\Phi(-\frac{M_3}{2}) \leq \Phi(-\frac{M}{2})$$

or $M_3 \geq M$ to attain a higher probability of correct classification by adding X_3.

Suppose $\mathbf{X} = (X_1, X_2)$ and $\mathbf{T} = (X_1, X_2, X_3)$. Let μ_1 be the mean of $\mathbf{X}$ for $\mathbf{G_1}$ and μ_2 be the mean of $\mathbf{X}$ for $\mathbf{G_2}$ with cov($\mathbf{X}$) = Σ. The mean of $\mathbf{T}$ for $\mathbf{G_1}$ is μ_{13} and the mean of $\mathbf{T}$ for $\mathbf{G_2}$ is μ_{23} with cov$(\mathbf{T}) = \Sigma_3$. Both covariance matrices are positive definite. Then, $M^2 = \sigma_W^2 = l'\Sigma l$, where W =$(\mu_1 - \mu_2)'\Sigma^{-1}\mathbf{X}$ and $l' = (\mu_1 - \mu_2)'\Sigma^{-1}$ as in (3.3). In a similar way, $M_3^2 = \sigma_{W_3}^2 = l_3'\Sigma_3 l_3$, where $W_3 = (\mu_{13} - \mu_{23})'\Sigma_3^{-1}\mathbf{T}$ and $l_3' = (\mu_{13} - \mu_{23})'\Sigma_3^{-1}$. Thus, $M_3^2 \geq M^2$ can be rewritten as

$$(\mu_{13} - \mu_{23})'\Sigma_3^{-1}(\mu_{13} - \mu_{23}) \geq (\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2).$$

To calculate M^2 and M_3^2, we set the following;

$$(\mu_1 - \mu_2) = \begin{bmatrix} u \\ v \end{bmatrix}, (\mu_{13} - \mu_{23}) = \begin{bmatrix} u \\ v \\ w \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix},$$

$$\Sigma_3 = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix}.$$

Here, $u, v, w, \sigma_{12}, \sigma_{13}, \sigma_{23}, \sigma_{21}, \sigma_{31},$ and σ_{32} are constants. Also, $\sigma_{11}, \sigma_{22},$ and σ_{33} are positive constants. Now,

$$M^2 = [u, v] \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}^{-1} \begin{bmatrix} u \\ v \end{bmatrix}$$

and

$$M_3^2 = [u, v, w] \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix}^{-1} \begin{bmatrix} u \\ v \\ w \end{bmatrix}.$$

Then, using the symmetric matrix of Σ and Σ_3,

$$M^2 = [u, v] \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}^{-1} \begin{bmatrix} u \\ v \end{bmatrix} = \frac{1}{(\sigma_{11}\sigma_{22} - \sigma_{12}^2)} \left([u, v] \begin{bmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{12} & \sigma_{11} \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} \right)$$

and

$$\begin{aligned} M_3^2 &= [u, v, w] \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} \end{bmatrix}^{-1} \begin{bmatrix} u \\ v \\ w \end{bmatrix} \\ &= \frac{1}{|\Sigma_3|} \left([u, v, w] \begin{bmatrix} \sigma_{22}\sigma_{33} - \sigma_{23}^2 & \sigma_{13}\sigma_{23} - \sigma_{12}\sigma_{33} & \sigma_{12}\sigma_{23} - \sigma_{22}\sigma_{13} \\ \sigma_{13}\sigma_{23} - \sigma_{12}\sigma_{33} & \sigma_{11}\sigma_{33} - \sigma_{13}^2 & \sigma_{12}\sigma_{13} - \sigma_{11}\sigma_{23} \\ \sigma_{12}\sigma_{23} - \sigma_{22}\sigma_{13} & \sigma_{12}\sigma_{13} - \sigma_{11}\sigma_{23} & \sigma_{11}\sigma_{22} - \sigma_{12}^2 \end{bmatrix} \begin{bmatrix} u \\ v \\ w \end{bmatrix} \right), \end{aligned}$$

where $|\Sigma_3| = \sigma_{11}\sigma_{22}\sigma_{33} - \sigma_{11}\sigma_{23}^2 - \sigma_{12}^2\sigma_{33} + 2\sigma_{12}\sigma_{13}\sigma_{23} - \sigma_{22}\sigma_{13}^2$, the determinant of Σ_3.

Then,

$$M^2 = \frac{1}{(\sigma_{11}\sigma_{22} - \sigma_{12}^2)} (u^2\sigma_{22} - 2uv\sigma_{12} + v^2\sigma_{11}), \tag{3.15}$$

$$\begin{aligned}M_3^2 &= w\big(\frac{u(\sigma_{12}\sigma_{23}-\sigma_{22}\sigma_{13})}{|\Sigma_3|}+\frac{v(\sigma_{12}\sigma_{13}-\sigma_{11}\sigma_{23})}{|\Sigma_3|}+\frac{w(\sigma_{11}\sigma_{22}-\sigma_{12}^2)}{|\Sigma_3|}\big)\\ &+v\big(\frac{u(\sigma_{13}\sigma_{23}-\sigma_{12}\sigma_{33})+v(\sigma_{11}\sigma_{33}-\sigma_{13}^2)+w(\sigma_{12}\sigma_{13}-\sigma_{11}\sigma_{23})}{|\Sigma_3|}\big)\\ &+u\big(\frac{u(\sigma_{22}\sigma_{33}-\sigma_{23}^2)+v(\sigma_{13}\sigma_{23}-\sigma_{12}\sigma_{33})+w(\sigma_{12}\sigma_{23}-\sigma_{22}\sigma_{13})}{|\Sigma_3|},\end{aligned} \tag{3.16}$$

where $|\Sigma_3| = \sigma_{11}\sigma_{22}\sigma_{33} - \sigma_{11}\sigma_{23}^2 - \sigma_{12}^2\sigma_{33} + 2\sigma_{12}\sigma_{13}\sigma_{23} - \sigma_{22}\sigma_{13}^2$.

To compare M^2 to M_3^2, we suppose 4 different cases according to the dependence of X_3 to X_1 and X_2.

(1) X_3 is not correlated with both X_1 and X_2, i.e., $\sigma_{13} = \sigma_{23} = 0$.

(i) X_1 and X_2 are not correlated, i.e., $\sigma_{12} = 0$. Then, the off-diagonal elements of the covariance matrix are all zero.

Thus, Σ and Σ_3 become diagonal matrices and M^2, M_3^2 are simplified as

$$M^2 = \frac{\sigma_{11}v^2 + \sigma_{22}u^2}{\sigma_{11}\sigma_{22}} = \frac{v^2}{\sigma_{22}} + \frac{u^2}{\sigma_{11}} \tag{3.17}$$

$$M_3^2 = \frac{w^2\sigma_{11}\sigma_{22} + v^2\sigma_{11}\sigma_{33} + u^2\sigma_{22}\sigma_{33}}{\sigma_{11}\sigma_{22}\sigma_{33}} = \frac{w^2}{\sigma_{33}} + \frac{v^2}{\sigma_{22}} + \frac{u^2}{\sigma_{11}} = M^2 + \frac{w^2}{\sigma_{33}}. \tag{3.18}$$

Since σ_{33} is positive and w^2 is nonnegative, $M^2 \leq M_3^2$. As w is larger and σ_{33} is smaller, M_3^2 gets greater. Then, we need X_3 with a small variance and a big difference in means between $\mathbf{G_1}$ and $\mathbf{G_2}$. Also, we want to make that ratio larger.

(ii) X_1 and X_2 are correlated, i.e., $\sigma_{12} \neq 0$.

From (3.15) and (3.16),

$$M^2 = \frac{1}{(\sigma_{11}\sigma_{22} - \sigma_{12}^2)}(u^2\sigma_{22} - 2uv\sigma_{12} + v^2\sigma_{11})$$

$$
\begin{aligned}
M_3^2 &= \frac{w^2(\sigma_{11}\sigma_{22} - \sigma_{12}^2) + v(u(-\sigma_{12}\sigma_{33}) + v(\sigma_{11}\sigma_{33})) + u(u\sigma_{22}\sigma_{33} + v(-\sigma_{12}\sigma_{33}))}{\sigma_{11}\sigma_{22}\sigma_{33} - \sigma_{12}^2\sigma_{33}} \\
&= \frac{1}{\sigma_{33}(\sigma_{11}\sigma_{22} - \sigma_{12}^2)}\big(w^2(\sigma_{11}\sigma_{22} - \sigma_{12}^2) + v^2\sigma_{11}\sigma_{33} + u^2\sigma_{22}\sigma_{33} - 2uv\sigma_{12}\sigma_{33}\big) \\
&= \frac{w^2}{\sigma_{33}} + \frac{v^2\sigma_{11} + u^2\sigma_{22} + 2uv\sigma_{12}}{\sigma_{11}\sigma_{22} - \sigma_{12}^2} \\
&= \frac{w^2}{\sigma_{33}} + M^2
\end{aligned}
\tag{3.19}
$$

Since $\frac{w^2}{\sigma_{33}} \geq 0$, $M_3^2 \geq M^2$. Thus, when X_3 is not correlated to X_1 and X_2, all we need for the new variable is to make the ratio of the mean difference to the variance large.

(2) X_3 is uncorrelated with X_1 but correlated with X_2, i.e., $\sigma_{13} = 0$ and $\sigma_{23} \neq 0$.

(i) X_1 and X_2 are not correlated, i.e., $\sigma_{12} = 0$.

Then,

$$
\Sigma = \begin{bmatrix} \sigma_{11} & 0 \\ 0 & \sigma_{22} \end{bmatrix}, \Sigma_3 = \begin{bmatrix} \sigma_{11} & 0 & 0 \\ 0 & \sigma_{22} & \sigma_{23} \\ 0 & \sigma_{23} & \sigma_{33} \end{bmatrix}.
$$

Then, from (3.15),

$$
M^2 = \frac{\sigma_{11}v^2 + \sigma_{22}u^2}{\sigma_{11}\sigma_{22}}
$$

and

$$
\begin{aligned}
M_3^2 &= [u, v, w] \begin{bmatrix} \sigma_{11} & 0 & 0 \\ 0 & \sigma_{22} & \sigma_{23} \\ 0 & \sigma_{23} & \sigma_{33} \end{bmatrix}^{-1} \begin{bmatrix} u \\ v \\ w \end{bmatrix} \\
&= \frac{1}{\sigma_{11}\sigma_{22}\sigma_{33} - \sigma_{11}\sigma_{23}^2} \left([u, v, w] \begin{bmatrix} \sigma_{22}\sigma_{33} - \sigma_{23}^2 & 0 & 0 \\ 0 & \sigma_{11}\sigma_{33} & -\sigma_{11}\sigma_{23} \\ 0 & -\sigma_{11}\sigma_{23} & \sigma_{11}\sigma_{22} \end{bmatrix} \begin{bmatrix} u \\ v \\ w \end{bmatrix} \right) \\
&= \frac{u^2(\sigma_{22}\sigma_{33} - \sigma_{23}^2) + v^2(\sigma_{11}\sigma_{33}) + w^2(\sigma_{11}\sigma_{22}) - 2vw(\sigma_{11}\sigma_{23})}{\sigma_{11}\sigma_{22}\sigma_{33} - \sigma_{11}\sigma_{23}^2} \\
&= \frac{v^2(\sigma_{11}\sigma_{33}) + u^2(\sigma_{22}\sigma_{33})}{\sigma_{11}\sigma_{22}\sigma_{33} - \sigma_{11}\sigma_{23}^2} + \frac{w^2(\sigma_{11}\sigma_{22}) - u^2(\sigma_{23}^2) - 2vw(\sigma_{11}\sigma_{23})}{\sigma_{11}\sigma_{22}\sigma_{33} - \sigma_{11}\sigma_{23}^2} \\
&= \frac{\sigma_{11}v^2 + \sigma_{22}u^2}{\sigma_{11}\sigma_{22} - \frac{\sigma_{11}}{\sigma_{33}}\sigma_{23}^2} + \frac{w^2(\sigma_{11}\sigma_{22}) - u^2(\sigma_{23}^2) - 2vw(\sigma_{11}\sigma_{23})}{\sigma_{11}\sigma_{22}\sigma_{33} - \sigma_{11}\sigma_{23}^2}.
\end{aligned}
$$

Here, the first term is greater than M^2 because the denominator is less than that of M^2 with the same numerator. Thus, we need to have a positive value on the second term to ensure that $M_3^2 \geq M^2$. Then, we need the same signs of the numerator and the denominator in the second fraction. Since the denominator of the second term is the determinant of the matrix, it's always positive due to the positive definite matrix. So, we only consider both positive values, i.e., $\sigma_{11}\sigma_{22}\sigma_{33} - \sigma_{11}\sigma_{23}^2 > 0$ and $w^2(\sigma_{11}\sigma_{22}) - u^2(\sigma_{23}^2) - 2vw(\sigma_{11}\sigma_{23}) > 0$. Also, since $\sigma_{11} > 0$, the first inequality becomes $\sigma_{22}\sigma_{33} - \sigma_{23}^2 > 0$. Here, we can get the range of σ_{23} as

$$-\sqrt{\sigma_{22}\sigma_{33}} < \sigma_{23} < \sqrt{\sigma_{22}\sigma_{33}}$$

given σ_{33}. This doesn't help much about the range of σ_{23} because it indicates the range of correlation coefficient between -1 and 1 when dividing the inequality by $\sqrt{\sigma_{22}\sigma_{33}}$. We can get the range of σ_{33} given σ_{23} instead. $\sigma_{33} > \frac{\sigma_{23}^2}{\sigma_{22}}$. Then, from the second inequality, we can get the range of w, the mean difference of X_3. Given σ_{23},

$$w > \sigma_{23}\Big(\frac{\sigma_{11}v + \sqrt{(\sigma_{11}v)^2 + \sigma_{11}\sigma_{22}u^2}}{\sigma_{11}\sigma_{22}}\Big) \text{ or } w < \sigma_{23}\Big(\frac{\sigma_{11}v - \sqrt{(\sigma_{11}v)^2 + \sigma_{11}\sigma_{22}u^2}}{\sigma_{11}\sigma_{22}}\Big)$$

by solving the quadratic equation for w.

We found the range of the variance of X_3 and the corresponding mean difference of X_3 given the covariance of (X_2,X_3).

(ii) X_1 and X_2 are correlated, i.e., $\sigma_{12} \neq 0$.

Then,

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}, \Sigma_3 = \begin{bmatrix} \sigma_{11} & \sigma_{12} & 0 \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ 0 & \sigma_{23} & \sigma_{33} \end{bmatrix},$$

$$M^2 = \frac{\sigma_{11}v^2 + \sigma_{22}u^2 - 2uv(\sigma_{12})}{\sigma_{11}\sigma_{22} - \sigma_{12}^2},$$

and

$$\begin{aligned} M_3^2 &= [u, v, w] \begin{bmatrix} \sigma_{11} & \sigma_{12} & 0 \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ 0 & \sigma_{23} & \sigma_{33} \end{bmatrix}^{-1} \begin{bmatrix} u \\ v \\ w \end{bmatrix} \\ &= \frac{1}{\text{Det}}\left([u, v, w] \begin{bmatrix} \sigma_{22}\sigma_{33} - \sigma_{23}^2 & -\sigma_{12}\sigma_{33} & \sigma_{12}\sigma_{23} \\ -\sigma_{12}\sigma_{33} & \sigma_{11}\sigma_{33} & -\sigma_{11}\sigma_{23} \\ \sigma_{12}\sigma_{23} & -\sigma_{11}\sigma_{23} & \sigma_{11}\sigma_{22} - \sigma_{12}^2 \end{bmatrix} \begin{bmatrix} u \\ v \\ w \end{bmatrix}\right) \\ &= \frac{1}{\text{Det}}\Big(u^2(\sigma_{22}\sigma_{33} - \sigma_{23}^2) + v^2(\sigma_{11}\sigma_{33}) + w^2(\sigma_{11}\sigma_{22} - \sigma_{12}^2) \\ &\qquad - 2uv(\sigma_{12}\sigma_{33}) + 2uw(\sigma_{12}\sigma_{23}) - 2vw(\sigma_{11}\sigma_{23})\Big) \\ &= \frac{\sigma_{33}(\sigma_{11}v^2 + \sigma_{22}u^2 - 2uv(\sigma_{12}))}{\text{Det}} \\ &\qquad + \frac{-u^2\sigma_{23}^2 + w^2(\sigma_{11}\sigma_{22} - \sigma_{12}^2) + 2uw\sigma_{12}\sigma_{23} - 2vw\sigma_{11}\sigma_{23}}{\text{Det}}, \end{aligned} \tag{3.20}$$

where Det = $\sigma_{11}\sigma_{22}\sigma_{33} - \sigma_{12}^2\sigma_{33} - \sigma_{11}\sigma_{23}^2$. The first term is greater than M^2 because the same numerator with a smaller denominator. Then, we want to make the second fraction positive again. Since the matrix is positive definite, the denominator that is the determinant of M_3^2 is positive, i.e., $\sigma_{33}(\sigma_{11}\sigma_{22} - \sigma_{12}^2) - \sigma_{11}\sigma_{23}^2 > 0$. Thus, we need to have $-u^2\sigma_{23}^2 + w^2(\sigma_{11}\sigma_{22} - \sigma_{12}^2) + 2uw\sigma_{12}\sigma_{23} - 2vw\sigma_{11}\sigma_{23} > 0$. Then, we have the range of σ_{33} from the first inequality given σ_{23} as

$$\sigma_{33} > \frac{\sigma_{11}\sigma_{23}^2}{\sigma_{11}\sigma_{22} - \sigma_{12}^2}.$$

If we solve the second inequality for w given σ_{23} with $\sigma_{11}\sigma_{22} - \sigma_{12}^2 > 0$, w has the range of

$$w > \sigma_{23}\left(\frac{-(\sigma_{12}u - \sigma_{11}v) + \sqrt{(\sigma_{12}u - \sigma_{11}v)^2 + (\sigma_{11}\sigma_{22} - \sigma_{12}^2)u^2}}{\sigma_{11}\sigma_{22} - \sigma_{12}^2}\right)$$

or

$$w < \sigma_{23}\left(\frac{-(\sigma_{12}u - \sigma_{11}v) - \sqrt{(\sigma_{12}u - \sigma_{11}v)^2 + (\sigma_{11}\sigma_{22} - \sigma_{12}^2)u^2}}{\sigma_{11}\sigma_{22} - \sigma_{12}^2}\right).$$

We can rewrite the range of w as

$$w > \sigma_{23}\left(\frac{-B + \sqrt{B^2 + Du^2}}{D}\right)$$

or

$$w < \sigma_{23}\left(\frac{-B - \sqrt{B^2 + Du^2}}{D}\right),$$

where B = $\sigma_{12}u - \sigma_{11}v$ and D = $\sigma_{11}\sigma_{22} - \sigma_{12}^2$.

If both the numerator and denominator are negative, it is not appropriate because the matrix is positive definite.

(3) X_3 is correlated with X_1 and uncorrelated with X_2, i.e., $\sigma_{13} \neq 0$ and $\sigma_{23} = 0$. The result is similar to (2).

(i) X_1 and X_2 are not correlated, i.e., $\sigma_{12} = 0$.

Then,

$$\Sigma = \begin{bmatrix} \sigma_{11} & 0 \\ 0 & \sigma_{22} \end{bmatrix}, \Sigma_3 = \begin{bmatrix} \sigma_{11} & 0 & \sigma_{13} \\ 0 & \sigma_{22} & 0 \\ \sigma_{13} & 0 & \sigma_{33} \end{bmatrix}.$$

Then, from (3.15),

$$M^2 = \frac{\sigma_{11}v^2 + \sigma_{22}u^2}{\sigma_{11}\sigma_{22}}$$

and

$$\begin{aligned}
M_3^2 &= [u, v, w] \begin{bmatrix} \sigma_{11} & 0 & \sigma_{13} \\ 0 & \sigma_{22} & 0 \\ \sigma_{13} & 0 & \sigma_{33} \end{bmatrix}^{-1} \begin{bmatrix} u \\ v \\ w \end{bmatrix} \\
&= \frac{1}{\sigma_{11}\sigma_{22}\sigma_{33} - \sigma_{22}\sigma_{13}^2} \left([u, v, w] \begin{bmatrix} \sigma_{22}\sigma_{33} & 0 & -\sigma_{22}\sigma_{13} \\ 0 & \sigma_{11}\sigma_{33} - \sigma_{13}^2 & 0 \\ -\sigma_{22}\sigma_{13} & 0 & \sigma_{11}\sigma_{22} \end{bmatrix} \begin{bmatrix} u \\ v \\ w \end{bmatrix} \right) \\
&= \frac{u^2(\sigma_{22}\sigma_{33}) + v^2(\sigma_{11}\sigma_{33} - \sigma_{13}^2) + w^2(\sigma_{11}\sigma_{22}) - 2uw(\sigma_{22}\sigma_{13})}{\sigma_{11}\sigma_{22}\sigma_{33} - \sigma_{22}\sigma_{13}^2} \\
&= \frac{v^2(\sigma_{11}\sigma_{33}) + u^2(\sigma_{22}\sigma_{33})}{\sigma_{11}\sigma_{22}\sigma_{33} - \sigma_{22}\sigma_{13}^2} + \frac{w^2(\sigma_{11}\sigma_{22}) - v^2(\sigma_{13}^2) - 2uw(\sigma_{22}\sigma_{13})}{\sigma_{11}\sigma_{22}\sigma_{33} - \sigma_{22}\sigma_{13}^2} \\
&= \frac{\sigma_{11}v^2 + \sigma_{22}u^2}{\sigma_{11}\sigma_{22} - \frac{\sigma_{22}}{\sigma_{33}}\sigma_{13}^2} + \frac{w^2(\sigma_{11}\sigma_{22}) - 2u(\sigma_{22}\sigma_{13})w - v^2(\sigma_{13}^2)}{\sigma_{11}\sigma_{22}\sigma_{33} - \sigma_{22}\sigma_{13}^2}.
\end{aligned}$$

As in (2) (i), the first term is greater than M^2 and we need to have both positive values of the second fraction, i.e., $\sigma_{11}\sigma_{22}\sigma_{33} - \sigma_{22}\sigma_{13}^2 > 0$ and $w^2(\sigma_{11}\sigma_{22}) - 2u(\sigma_{22}\sigma_{13})w - v^2(\sigma_{13}^2) > 0$. Since $\sigma_{22} > 0$, the first inequality becomes $\sigma_{11}\sigma_{33} - \sigma_{13}^2 > 0$, which is always true. Now, we get the range of σ_{33} given σ_{13} as

$$\sigma_{33} > \frac{\sigma_{13}^2}{\sigma_{11}}.$$

Then, from the second inequality, we can get the range of w, the mean difference of X_3. Given σ_{13},

$$w > \sigma_{13}(\frac{\sigma_{22}u + \sqrt{(\sigma_{22}u)^2 + \sigma_{11}\sigma_{22}v^2}}{\sigma_{11}\sigma_{22}}) \text{ or } w < \sigma_{13}(\frac{\sigma_{22}u - \sqrt{(\sigma_{22}u)^2 + \sigma_{11}\sigma_{22}v^2}}{\sigma_{11}\sigma_{22}})$$

by solving the quadratic equation for w. Both negative signs case is inappropriate because the denominator can't be negative. We found the range of the variance of X_3 and the corresponding mean difference of X_3 given the covariance of (X_1,X_3).

(ii) X_1 and X_2 are correlated, i.e., $\sigma_{12} \neq 0$.

Then,

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}, \Sigma_3 = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & 0 \\ \sigma_{13} & 0 & \sigma_{33} \end{bmatrix},$$

$$M^2 = \frac{\sigma_{11}v^2 + \sigma_{22}u^2 - 2uv(\sigma_{12})}{\sigma_{11}\sigma_{22} - \sigma_{12}^2},$$

and

$$
\begin{aligned}
M_3^2 &= [u,v,w]\begin{bmatrix}\sigma_{11} & \sigma_{12} & \sigma_{13}\\ \sigma_{12} & \sigma_{22} & 0\\ \sigma_{13} & 0 & \sigma_{33}\end{bmatrix}^{-1}\begin{bmatrix}u\\ v\\ w\end{bmatrix}\\
&= \frac{1}{\text{Det}}\left([u,v,w]\begin{bmatrix}\sigma_{22}\sigma_{33} & -\sigma_{12}\sigma_{33} & -\sigma_{22}\sigma_{13}\\ -\sigma_{12}\sigma_{33} & \sigma_{11}\sigma_{33}-\sigma_{13}^2 & \sigma_{12}\sigma_{13}\\ -\sigma_{22}\sigma_{13} & \sigma_{12}\sigma_{13} & \sigma_{11}\sigma_{22}-\sigma_{12}^2\end{bmatrix}\begin{bmatrix}u\\ v\\ w\end{bmatrix}\right)\\
&= \frac{1}{\text{Det}}\Big(u^2(\sigma_{22}\sigma_{33}) + v^2(\sigma_{11}\sigma_{33}-\sigma_{13}^2) + w^2(\sigma_{11}\sigma_{22}-\sigma_{12}^2)\\
&\qquad - 2uv(\sigma_{12}\sigma_{33}) - 2uw(\sigma_{22}\sigma_{13}) + 2vw(\sigma_{12}\sigma_{13})\Big)\\
&= \frac{\sigma_{33}(\sigma_{11}v^2+\sigma_{22}u^2-2uv(\sigma_{12}))}{\text{Det}}\\
&\qquad + \frac{-v^2\sigma_{13}^2 + w^2(\sigma_{11}\sigma_{22}-\sigma_{12}^2) - 2uw\sigma_{22}\sigma_{13} + 2vw\sigma_{12}\sigma_{13}}{\text{Det}},
\end{aligned}
\tag{3.21}
$$

where Det = $\sigma_{11}\sigma_{22}\sigma_{33}-\sigma_{12}^2\sigma_{33}-\sigma_{22}\sigma_{13}^2$. The first term is greater than M^2 because of the less denominator with the same numerator. Then, we want to make the second fraction positive again. If both the numerator and the denominator are positive, $\sigma_{33}(\sigma_{11}\sigma_{22}-\sigma_{12}^2)-\sigma_{22}\sigma_{13}^2 > 0$ and $-v^2\sigma_{13}^2 + w^2(\sigma_{11}\sigma_{22}-\sigma_{12}^2) - 2uw\sigma_{22}\sigma_{13} + 2vw\sigma_{12}\sigma_{13} > 0$. Then, we have the range of σ_{33} from the first inequality given σ_{13} as

$$\sigma_{33} > \frac{\sigma_{22}\sigma_{13}^2}{\sigma_{11}\sigma_{22}-\sigma_{12}^2}.$$

If we solve the second inequality for w given σ_{13} with $\sigma_{11}\sigma_{22}-\sigma_{12}^2 > 0$, w has the range of

$$w > \sigma_{13}\left(\frac{-(\sigma_{12}v-\sigma_{22}u)+\sqrt{(\sigma_{12}v-\sigma_{22}u)^2+(\sigma_{11}\sigma_{22}-\sigma_{12}^2)v^2}}{\sigma_{11}\sigma_{22}-\sigma_{12}^2}\right)$$

or

$$w < \sigma_{13}\left(\frac{-(\sigma_{12}v - \sigma_{22}u) - \sqrt{(\sigma_{12}v - \sigma_{22}u)^2 + (\sigma_{11}\sigma_{22} - \sigma_{12}^2)v^2}}{\sigma_{11}\sigma_{22} - \sigma_{12}^2}\right)$$

We can rewrite this as

$$w > \sigma_{13}\left(\frac{-C + \sqrt{C^2 + Dv^2}}{D}\right)$$

or

$$w < \sigma_{13}\left(\frac{-C - \sqrt{C^2 + Dv^2}}{D}\right),$$

where C = $\sigma_{12}v - \sigma_{22}u$ and D = $\sigma_{11}\sigma_{22} - \sigma_{12}^2$.

(4) X_3 is correlation with X_1 and X_2, i.e., $\sigma_{13} \neq 0$ and $\sigma_{23} \neq 0$. From (3.15) and (3.16),

$$M^2 = \frac{1}{(\sigma_{11}\sigma_{22} - \sigma_{12}^2)}(u^2\sigma_{22} - 2uv\sigma_{12} + v^2\sigma_{11}), \tag{3.22}$$

$$\begin{aligned} M_3^2 = {} & w\big(\frac{u(\sigma_{12}\sigma_{23} - \sigma_{22}\sigma_{13})}{|\Sigma_3|} + \frac{v(\sigma_{12}\sigma_{13} - \sigma_{11}\sigma_{23})}{|\Sigma_3|} + \frac{w(\sigma_{11}\sigma_{22} - \sigma_{12}^2)}{|\Sigma_3|}\big) \\ & + v\big(\frac{u(\sigma_{13}\sigma_{23} - \sigma_{12}\sigma_{33}) + v(\sigma_{11}\sigma_{33} - \sigma_{13}^2) + w(\sigma_{12}\sigma_{13} - \sigma_{11}\sigma_{23})}{|\Sigma_3|}\big) \\ & + u\big(\frac{u(\sigma_{22}\sigma_{33} - \sigma_{23}^2) + v(\sigma_{13}\sigma_{23} - \sigma_{12}\sigma_{33}) + w(\sigma_{12}\sigma_{23} - \sigma_{22}\sigma_{13})}{|\Sigma_3|}, \end{aligned} \tag{3.23}$$

where $|\Sigma_3| = \sigma_{11}\sigma_{22}\sigma_{33} - \sigma_{11}\sigma_{23}^2 - \sigma_{12}^2\sigma_{33} + 2\sigma_{12}\sigma_{13}\sigma_{23} - \sigma_{22}\sigma_{13}^2$.

Then,

$$\begin{aligned} M_3^2 = {} & \frac{u^2\sigma_{22}\sigma_{33} - 2uv\sigma_{12}\sigma_{33} + v^2\sigma_{11}\sigma_{33}}{|\Sigma_3|} \\ & + \frac{-u^2\sigma_{23}^2 - v^2\sigma_{13}^2 + w^2(\sigma_{11}\sigma_{22} - \sigma_{12}^2)}{|\Sigma_3|} \\ & + \frac{2uw(\sigma_{12}\sigma_{23} - \sigma_{22}\sigma_{13}) + 2vw(\sigma_{12}\sigma_{13} - \sigma_{11}\sigma_{23}) + 2uv\sigma_{13}\sigma_{23}}{|\Sigma_3|}. \end{aligned} \tag{3.24}$$

We can rewrite the determinant of Σ_3 as

$$\begin{aligned}|\Sigma_3| &= \sigma_{11}\sigma_{22}\sigma_{33} - \sigma_{11}\sigma_{23}^2 - \sigma_{12}^2\sigma_{33} + 2\sigma_{12}\sigma_{13}\sigma_{23} - \sigma_{22}\sigma_{13}^2 \\ &= \sigma_{33}\big(\sigma_{11}\sigma_{22} - \sigma_{12}^2 - \frac{\sigma_{11}\sigma_{23}^2 + \sigma_{22}\sigma_{13}^2 - 2\sigma_{12}\sigma_{23}\sigma_{13}}{\sigma_{33}}\big)\end{aligned} \tag{3.25}$$

Since the determinant is positive, we get the range of σ_{33} as

$$\sigma_{33} > \frac{\sigma_{11}\sigma_{23}^2 + \sigma_{22}\sigma_{13}^2 - 2\sigma_{12}\sigma_{23}\sigma_{13}}{\sigma_{11}\sigma_{22} - \sigma_{12}^2}.$$

Here, the numerator of the fraction part of the last term in (3.25),

$$\begin{aligned}0 &\le (\sqrt{\sigma_{11}}\sigma_{23} - \sqrt{\sigma_{22}}\sigma_{13})^2 \\ &= \sigma_{11}\sigma_{23}^2 + \sigma_{22}\sigma_{13}^2 - 2\sqrt{\sigma_{11}\sigma_{22}}\sigma_{23}\sigma_{13} \\ &< \sigma_{11}\sigma_{23}^2 + \sigma_{22}\sigma_{13}^2 - 2\sigma_{12}\sigma_{23}\sigma_{13}.\end{aligned} \tag{3.26}$$

The last inequality holds because $\sigma_{11}\sigma_{22} - \sigma_{12}^2 > 0$ or $-\sqrt{\sigma_{11}\sigma_{22}} < \sigma_{12} < \sqrt{\sigma_{11}\sigma_{22}}$. Thus, the first term of (3.24) is bigger than M^2 because the denominator is smaller since both $\sigma_{33} > 0$ and $\sigma_{11}\sigma_{23}^2 + \sigma_{22}\sigma_{13}^2 - 2\sigma_{12}\sigma_{23}\sigma_{13} > 0$ from (3.26).

Then, we get the range of w from the second and third terms of (3.24) given the denominator of those are positive. We solve the inequality for w.

$$\begin{gathered}(\sigma_{11}\sigma_{22} - \sigma_{12}^2)w^2 + 2w\big(u(\sigma_{12}\sigma_{23} - \sigma_{22}\sigma_{13}) + v(\sigma_{12}\sigma_{13} - \sigma_{11}\sigma_{23})\big) \\ + 2uv\sigma_{13}\sigma_{23} - u^2\sigma_{23}^2 - v^2\sigma_{13}^2 > 0\end{gathered} \tag{3.27}$$

or

$$(\sigma_{11}\sigma_{22} - \sigma_{12}^2)w^2 + 2w\big(u(\sigma_{12}\sigma_{23} - \sigma_{22}\sigma_{13}) + v(\sigma_{12}\sigma_{13} - \sigma_{11}\sigma_{23})\big) - (u\sigma_{23} - v\sigma_{13})^2 > 0.$$

Then,

$$w > \frac{-F + \sqrt{F^2 + DG}}{D}$$

$$w < \frac{-F - \sqrt{F^2 + DG}}{D},$$

where $D = (\sigma_{11}\sigma_{22} - \sigma_{12}^2)$, $F = u(\sigma_{12}\sigma_{23} - \sigma_{22}\sigma_{13}) + v(\sigma_{12}\sigma_{13} - \sigma_{11}\sigma_{23})$, and $G = (u\sigma_{23} - v\sigma_{13})^2$.

3.5.3 Summary

In this section, we investigated the conditions of the preferable predictor vector, especially for 2-dimensional problems. We first suggested using statistical learning method to search for the variable. Then, we use the bivariate normal case with TPM. In both ways, it reduced to the values of the mean differences between two populations and the variance of the new variable or the covariance with the existing variables. Below is the summary of the conditions from the TPM calculation. In summary, let $D = (\sigma_{11}\sigma_{22} - \sigma_{12}^2)$,

(1) $\sigma_{13} = \sigma_{23} = 0$

(i) $\sigma_{12} = 0$: Make $\frac{w^2}{\sigma_{33}}$ big.

(ii) $\sigma_{12} \neq 0$: Make $\frac{w^2}{\sigma_{33}}$ big.

(2) $\sigma_{13} = 0, \sigma_{23} \neq 0$

(i) $\sigma_{12} = 0$

$$\sigma_{33} > \frac{\sigma_{23}^2}{\sigma_{22}},$$

$$w > \sigma_{23}\big(\frac{\sigma_{11}v + \sqrt{(\sigma_{11}v)^2 + \sigma_{11}\sigma_{22}u^2}}{\sigma_{11}\sigma_{22}}\big) \text{ or } w < \sigma_{23}\big(\frac{\sigma_{11}v - \sqrt{(\sigma_{11}v)^2 + \sigma_{11}\sigma_{22}u^2}}{\sigma_{11}\sigma_{22}}\big)$$

(ii) $\sigma_{12} \neq 0$

$$\sigma_{33} > \frac{\sigma_{11}\sigma_{23}^2}{D},$$

$$w > \sigma_{23}\left(\frac{-(\sigma_{12}u - \sigma_{11}v) + \sqrt{(\sigma_{12}u - \sigma_{11}v)^2 + Du^2}}{D}\right)$$

or

$$w < \sigma_{23}\left(\frac{-(\sigma_{12}u - \sigma_{11}v) - \sqrt{(\sigma_{12}u - \sigma_{11}v)^2 + Du^2}}{D}\right).$$

(3) $\sigma_{13} \neq 0, \sigma_{23} = 0$

(i) $\sigma_{12} = 0$

$$\sigma_{33} > \frac{\sigma_{13}^2}{\sigma_{11}}$$

$$w > \sigma_{13}\Big(\frac{\sigma_{22}u + \sqrt{(\sigma_{22}u)^2 + \sigma_{11}\sigma_{22}v^2}}{\sigma_{11}\sigma_{22}}\Big) \text{ or } w < \sigma_{13}\Big(\frac{\sigma_{22}u - \sqrt{(\sigma_{22}u)^2 + \sigma_{11}\sigma_{22}v^2}}{\sigma_{11}\sigma_{22}}\Big)$$

(ii) $\sigma_{12} \neq 0$

$$\sigma_{33} > \frac{\sigma_{22}\sigma_{13}^2}{D},$$

$$w > \sigma_{13}\left(\frac{-(\sigma_{12}v - \sigma_{22}u) + \sqrt{(\sigma_{12}v - \sigma_{22}u)^2 + Dv^2}}{D}\right)$$

or

$$w < \sigma_{13}\left(\frac{-(\sigma_{12}v - \sigma_{22}u) - \sqrt{(\sigma_{12}v - \sigma_{22}u)^2 + Dv^2}}{D}\right)$$

(4) $\sigma_{13} \neq 0$ and $\sigma_{23} \neq 0$.

$$\sigma_{33} > \frac{\sigma_{11}\sigma_{23}^2 + \sigma_{22}\sigma_{13}^2 - 2\sigma_{12}\sigma_{23}\sigma_{13}}{D},$$

$$w > \frac{-F + \sqrt{F^2 + DG}}{D}$$

or

$$w < \frac{-F - \sqrt{F^2 + DG}}{D},$$

where $F = u(\sigma_{12}\sigma_{23} - \sigma_{22}\sigma_{13}) + v(\sigma_{12}\sigma_{13} - \sigma_{11}\sigma_{23})$ and $G = (u\sigma_{23} - v\sigma_{13})^2$.

When the new variable is independent of the two existing variables, we only need to make the ratio $\frac{w^2}{\sigma_{33}}$ larger. Thus, to increase the probability of a correct decision, we have to select a new variable that is independent of the existing variables with a small variance and a big difference in means. If the new variable is independent of only one of the existing variables, adding a new variable does not always guarantee in the probability of a correct decision, but we can get the range of the variance of X_3, σ_{33}, along with the range of the mean difference of X_3, w, given covariances. If we can fix the covariances either 0 or not zero, all we need to do is to maximize $\dfrac{w^2}{\sigma_{33}}$. Then, in any given range of σ_{33}, we take the smallest value for it and take w as big as possible.

Example 1

We check the condition of X_3 with examples. Suppose there are two bivariate normal populations. Population 1 has the mean vector of $(5, 2)'$ and Population 2 has the mean vector of $(1, 4)'$. The covariance matrix is common but we only set the variances as 2 and 4, respectively. Then, the covariance matrix is

$$\begin{bmatrix} 2 & \sigma_{12} \\ \sigma_{12} & 4 \end{bmatrix}.$$

To check the condition of X_3, the mean difference becomes $(4, -2, w)'$ and the covariance matrix is set as

$$\begin{bmatrix} 2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & 4 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} \end{bmatrix}.$$

Then, let's check the example for each case above. We will generate 80 data points for each population given the mean vector and the covariance matrix. We use lda from R to find the classifying rule and create the confusion matrices.

Let's calculate the ranges of w and σ_{33}.

(1) $\sigma_{13} = \sigma_{23} = 0$

(i) $\sigma_{12} = 0$.

(ii) $\sigma_{12} \neq 0$. In both cases, we just need to make the ratio, $\dfrac{w^2}{\sigma_{33}}$, big.

(2) $\sigma_{13} = 0, \sigma_{23} \neq 0$ and let $\sigma_{23} = 2$.

(i) $\sigma_{12} = 0$. We have $\sigma_{33} > 1$ and $w > 2$.

(ii) $\sigma_{12} \neq 0$ and let $\sigma_{12} = 1$. We have $\sigma_{33} > 1.1428$ and $w > 1.5047$.

(3) $\sigma_{13} \neq 0, \sigma_{23} = 0$ and let $\sigma_{13} = 1.5$.

(i) $\sigma_{12} = 0$. We have $\sigma_{33} > 1.125$ and $w > 4.835$.

(ii) $\sigma_{12} \neq 0$ and let $\sigma_{12} = 1$. We have $\sigma_{33} > 1.2857$ and $w > 7.877$.

(4) No zero entry in covariance matrix.

Let $\Sigma = \begin{bmatrix} 2 & 0 & 1.5 \\ 0 & 4 & 2 \\ 1.5 & 2 & 2.5 \end{bmatrix}$, $\Sigma = \begin{bmatrix} 2 & 1 & 1.5 \\ 1 & 4 & 2 \\ 1.5 & 2 & 2 \end{bmatrix}$ for $\sigma_{12} = 0$ and $\sigma_{12} \neq 0$, respectively. Since we have $\sigma_{33} > 2.125$ and $w > 2.1514$ when $\sigma_{12} = 0$. Also, we have $\sigma_{33} > 1.5714$ and $w > 2.873$ when $\sigma_{12} \neq 0$.

Thus, to create the confusion matrices, let's first fix σ_{33} then, change w to increase the ratio, $\frac{w^2}{\sigma_{33}}$.

(1) $\sigma_{13} = \sigma_{23} = 0$ and $\sigma_{33} = 2$. w takes 3 and 6 by (4,1) and (7,1), respectively.

(i) $\sigma_{12} = 0$. Then, the confusion matrices before and after the preferable predictor vector are

Before		To		
		G_1	G_2	
From	G_1	77	3	80
	G_2	5	75	80

w=3		To		
		G_1	G_2	
From	G_1	79	1	80
	G_2	1	79	80

w=6		To		
		G_1	G_2	
From	G_1	80	0	80
	G_2	1	79	80

Table 3.8: Confusion Matrices before/after the Preferable Predictor Vector

when $\sigma_{12} = \sigma_{13} = \sigma_{23} = 0$

(ii) $\sigma_{12} \neq 0$, then make $\sigma_{12} = 1$.

Before		To		
		G_1	G_2	
From	G_1	78	2	80
	G_2	3	77	80

w=3		To		
		G_1	G_2	
From	G_1	76	4	80
	G_2	1	79	80

w=6		To		
		G_1	G_2	
From	G_1	80	0	80
	G_2	0	80	80

Table 3.9: Confusion Matrices before/after the Preferable Predictor Vector

when $\sigma_{12} = 1$ and $\sigma_{13} = \sigma_{23} = 0$

(2) $\sigma_{13} = 0, \sigma_{23} \neq 0$, then, make $\sigma_{23} = 2$.

(i) $\sigma_{12} = 0$. Then, $\sigma_{33} = 2$ and w takes 3 and 6.

Before		To		
		G_1	G_2	
From	G_1	77	3	80
	G_2	5	75	80

w=3		To		
		G_1	G_2	
From	G_1	79	1	80
	G_2	0	80	80

w=6		To		
		G_1	G_2	
From	G_1	80	0	80
	G_2	0	80	80

Table 3.10: Confusion Matrices before/after the Preferable Predictor Vector

when $\sigma_{12} = \sigma_{13} = 0$ and $\sigma_{23} = 2$

(ii) $\sigma_{12} \neq 0$ and let $\sigma_{12} = 1$. Then, $\sigma_{33} = 2$ and w takes 3 and 6.

Before		To		
		G_1	G_2	
From	G_1	78	2	80
	G_2	3	77	80

w=3		To		
		G_1	G_2	
From	G_1	80	0	80
	G_2	0	80	80

w=6		To		
		G_1	G_2	
From	G_1	80	0	80
	G_2	0	80	80

Table 3.11: Confusion Matrices before/after the Preferable Predictor Vector

when $\sigma_{12} = 1, \sigma_{13} = 0,$ and $\sigma_{23} = 2$

(3) $\sigma_{13} \neq 0, \sigma_{23} = 0$, then make $\sigma_{13} = 1.5$.

(i) $\sigma_{12} = 0$. Also, $\sigma_{33} = 2$ and w takes 5.

		Assigned to		
		Pop1	Pop2	
Observations from	Pop1	77	3	80
	Pop2	5	75	80

		Assigned to		
		Pop1	Pop2	
Observations from	Pop1	79	1	80
	Pop2	1	79	80

Table 3.12: Confusion Matrices before/after the Preferable Predictor Vector

when $\sigma_{12} = 0, \sigma_{13} = 1.5,$ and $\sigma_{23} = 0$

(ii) $\sigma_{12} \neq 0$, then $\sigma_{12} = 1$. Let $\sigma_{33} = 2$ and w takes 8.

		Assigned to		
		Pop1	Pop2	
Observations from	Pop1	78	2	80
	Pop2	3	77	80

		Assigned to		
		Pop1	Pop2	
Observations from	Pop1	80	0	80
	Pop2	0	80	80

Table 3.13: Confusion Matrices before/after the Preferable Predictor Vector

when $\sigma_{12} = 1, \sigma_{13} = 1.5,$ and $\sigma_{23} = 0$

(4) No zero entry in covariance matrix.

Let $\Sigma = \begin{bmatrix} 2 & 0 & 1.5 \\ 0 & 4 & 2 \\ 1.5 & 2 & 2.5 \end{bmatrix}$ or $\Sigma = \begin{bmatrix} 2 & 1 & 1.5 \\ 1 & 4 & 2 \\ 1.5 & 2 & 2 \end{bmatrix}$. Take w= 3 for both cases. Then, confusion matrices are

		Assigned to		
		Pop1	Pop2	
Observations from	Pop1	78	2	80
	Pop2	3	77	80

		Assigned to		
		Pop1	Pop2	
Observations from	Pop1	78	2	80
	Pop2	1	79	80

Table 3.14: Confusion Matrices with the PPV when $\sigma_{13} = 1.5$ and $\sigma_{23} = 2$

Summary

When the conditions for the preferable predictor vector are met, the probability of correct decision is at least improved based on the calculation of APER for all examples. However, the location of the means of X_3 affects the rate of improvement and it needs to be further investigated. In this example, the mean of X_3 of the second group was fixed as 1 and calculated the mean of the first group depending on the given w. When we change the fixed mean to a different value and change the corresponding other mean, we observe a different number of misclassified points. It must be related to the dispersion of the group and the location of the mean of new variable.

Chapter 4

Multiple Decision Procedures and Statistical Classification

4.0.1 The Indifference-Zone Approach as a Statistical Classification

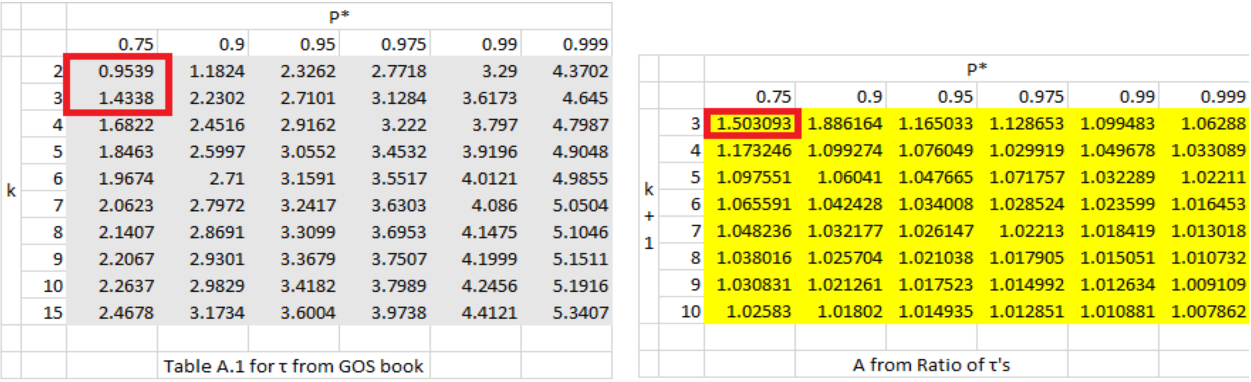
One way to explain the classification process is that it is one of the identifying processes and the selection and ranking methodologies are identifying processes, too. The indifference-zone approach results in two groups of populations, one with the best population and the other without the best population. When we have the result of the indifference-zone approach, we can look it from the statistical classification viewpoint and propose to update the procedure by applying the statistical learning.

4.1 The Indifference-Zone Approach with Statistical Learning

When we consider the indifference-zone approach with statistical learning , we can focus on the discerning measure of distance, δ. We can think of it in two ways. First, we update the discerning measure of distance after adding a new population. Second, we update the discerning measure of distance by increasing the required probability of correct selection.

Updating δ When A New Population Is Added

Let's take a look at the first way to update the discerning measure of distance. In the indifference-zone approach, we select the best population by the samples whose size is calculated given δ^* and P^* with known variance. Suppose we take t best populations from k populations. From the Table A.1 in GOS book, the value τ was calculated, where $\tau = \sqrt{n}\frac{\delta^*}{\sigma}$. We notice that there is a positive relationship between δ and the probability of a correct decision if you keep the other values (sample size and variance) the same. After selecting the best population, if we add one more population, the number of populations goes from K to $K+1$ in the table, so we need a new discerning measure of distance that should be larger than before. Then, a new discerning measure of distance, $\delta_{k+1} = A\delta^*$, is required and, at least to keep the same probability of correct selection level, we need to have a new larger δ between the best and the second best population since the number of population is increased. A's($A > 1$) are calculated by taking ratio of $\frac{\tau(k+1)}{\tau(k)}$ under the same P^* level and a table for A is provided below.



		P*					
		0.75	0.9	0.95	0.975	0.99	0.999
k	2	0.9539	1.1824	2.3262	2.7718	3.29	4.3702
	3	1.4338	2.2302	2.7101	3.1284	3.6173	4.645
	4	1.6822	2.4516	2.9162	3.222	3.797	4.7987
	5	1.8463	2.5997	3.0552	3.4532	3.9196	4.9048
	6	1.9674	2.71	3.1591	3.5517	4.0121	4.9855
	7	2.0623	2.7972	3.2417	3.6303	4.086	5.0504
	8	2.1407	2.8691	3.3099	3.6953	4.1475	5.1046
	9	2.2067	2.9301	3.3679	3.7507	4.1999	5.1511
	10	2.2637	2.9829	3.4182	3.7989	4.2456	5.1916
	15	2.4678	3.1734	3.6004	3.9738	4.4121	5.3407

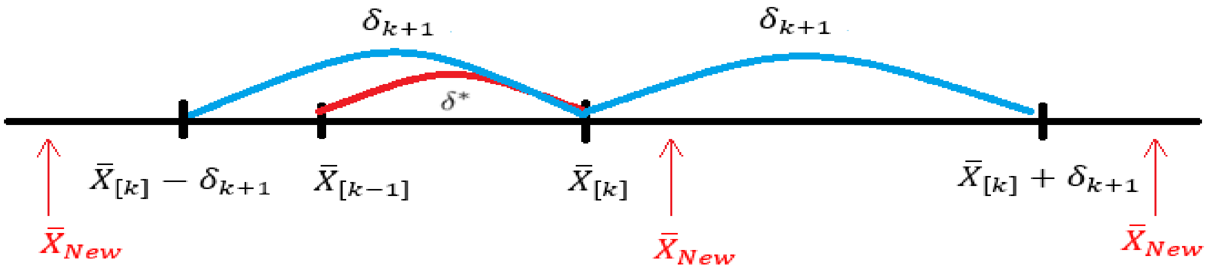
Table A.1 for τ from GOS book

		P*					
		0.75	0.9	0.95	0.975	0.99	0.999
k + 1	3	1.503093	1.886164	1.165033	1.128653	1.099483	1.06288
	4	1.173246	1.099274	1.076049	1.029919	1.049678	1.033089
	5	1.097551	1.06041	1.047665	1.071757	1.032289	1.02211
	6	1.065591	1.042428	1.034008	1.028524	1.023599	1.016453
	7	1.048236	1.032177	1.026147	1.02213	1.018419	1.013018
	8	1.038016	1.025704	1.021038	1.017905	1.015051	1.010732
	9	1.030831	1.021261	1.017523	1.014992	1.012634	1.009109
	10	1.02583	1.01802	1.014935	1.012851	1.010881	1.007862

A from Ratio of τ's

Figure 4.1: Table A.1 from GOS book(left) and a Table of A Calculated(right)

When the new sample mean is greater than the new δ_{k+1} from the old largest sample mean, where $\delta_{k+1} = \bar{X}_{[k+1]} - \bar{X}_{[k]}$, we select the new population as the best population with at least the same probability of correct selection.



$\bar{X}_{[k]} + \delta_{k+1} < \bar{X}_{[New]} \quad \Rightarrow \bar{X}_{[New]}$ is selected as the best population.

$\bar{X}_{[k-1]} < \bar{X}_{[New]} \leq \bar{X}_{[k]} + \delta_{k+1} \Rightarrow$ Indifferent between $\bar{X}_{[k]}$ and $\bar{X}_{[New]}$.

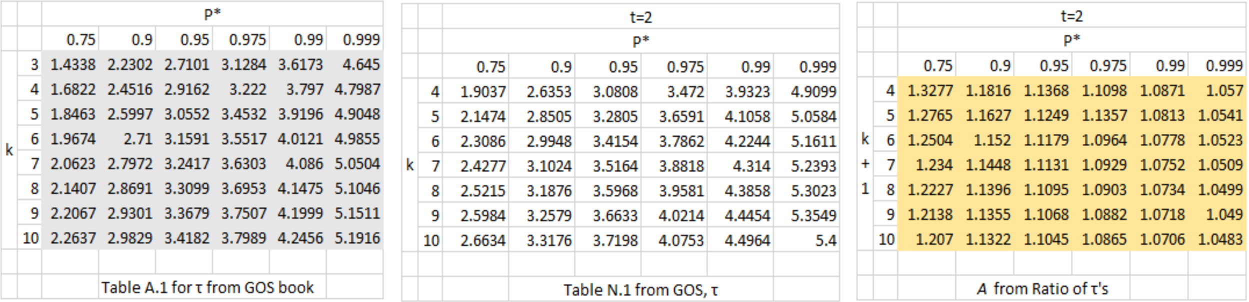
$\bar{X}_{[New]} \leq \bar{X}_{[k-1]} \quad \Rightarrow$ Indifferent between $\bar{X}_{[k]}$ and $\bar{X}_{[k-1]}$.

Figure 4.2: Location of $\bar{X}_{New}$ and the Selection Rule

If we want to select 2 best populations after you add one population, the number of populations becomes $k+1$ from k and the discerning measure of distance becomes δ_{k+1},

$$\begin{aligned} \delta_{k+1} &= \bar{X}_{[k]} - \bar{X}_{[k-1]} \quad \text{given} \\ \delta_k &= \bar{X}_{[k]} - \bar{X}_{[k-1]}. \end{aligned} \tag{4.1}$$

When we use the same way as above, $\delta_{k+1} = A\delta_k$, we have the following table for A below.



		p*					
		0.75	0.9	0.95	0.975	0.99	0.999
k	3	1.4338	2.2302	2.7101	3.1284	3.6173	4.645
	4	1.6822	2.4516	2.9162	3.222	3.797	4.7987
	5	1.8463	2.5997	3.0552	3.4532	3.9196	4.9048
	6	1.9674	2.71	3.1591	3.5517	4.0121	4.9855
	7	2.0623	2.7972	3.2417	3.6303	4.086	5.0504
	8	2.1407	2.8691	3.3099	3.6953	4.1475	5.1046
	9	2.2067	2.9301	3.3679	3.7507	4.1999	5.1511
	10	2.2637	2.9829	3.4182	3.7989	4.2456	5.1916

Table A.1 for τ from GOS book

		t=2					
		p*					
		0.75	0.9	0.95	0.975	0.99	0.999
k	4	1.9037	2.6353	3.0808	3.472	3.9323	4.9099
	5	2.1474	2.8505	3.2805	3.6591	4.1058	5.0584
	6	2.3086	2.9948	3.4154	3.7862	4.2244	5.1611
	7	2.4277	3.1024	3.5164	3.8818	4.314	5.2393
	8	2.5215	3.1876	3.5968	3.9581	4.3858	5.3023
	9	2.5984	3.2579	3.6633	4.0214	4.4454	5.3549
	10	2.6634	3.3176	3.7198	4.0753	4.4964	5.4

Table N.1 from GOS, τ

		t=2					
		p*					
		0.75	0.9	0.95	0.975	0.99	0.999
k+1	4	1.3277	1.1816	1.1368	1.1098	1.0871	1.057
	5	1.2765	1.1627	1.1249	1.1357	1.0813	1.0541
	6	1.2504	1.152	1.1179	1.0964	1.0778	1.0523
	7	1.234	1.1448	1.1131	1.0929	1.0752	1.0509
	8	1.2227	1.1396	1.1095	1.0903	1.0734	1.0499
	9	1.2138	1.1355	1.1068	1.0882	1.0718	1.049
	10	1.207	1.1322	1.1045	1.0865	1.0706	1.0483

A from Ratio of τ's

Figure 4.3: *A* for Selecting Two Best Populations

From the relationship between δ_k and δ_{k+1} above, they share the same location for the discerning measure of distance, $\bar{X}_{[k]} - \bar{X}_{[k-1]}$. For example, $\tau(3)$ from Table A.1, 1.4338, and $\tau(4)$ Table N.1, 1.9037, include the same discerning measure of distance under the same column. However, $\tau(k)$

from Table A.1 is always smaller than $\tau(k+1)$ from Table N.1 given the same P^*. Thus, if we let $\delta_{k+1} = A\delta_k$ and $A > 1$, A can be calculated from $\dfrac{\tau(k+1)\ in\ TableN.1}{\tau(k)\ in\ TableA.1}$ as the right table from Figure 4.3. So, when we need to select 2 best populations after adding one more population with the same level of probability of correct selection, we need $\delta_{k+1} > \delta_k$, which is unattainable even when the new sample mean is greater than $\bar{X}_{[k]}$ because $\delta_{k+1} = \delta_k$ already in (4.1). Then, other values being equal, the probability of correct selection should decrease, $P^*(k+1) < P^*(k)$ if we want to select the two best population after adding one new population. We can't achieve the higher probability of correct selection for this case.

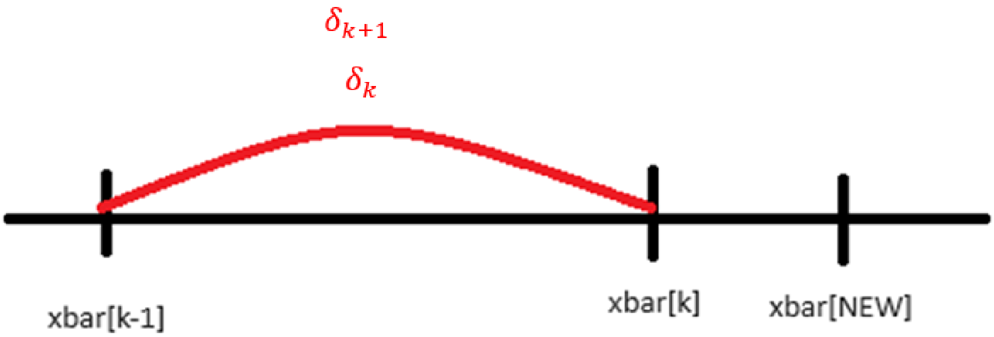
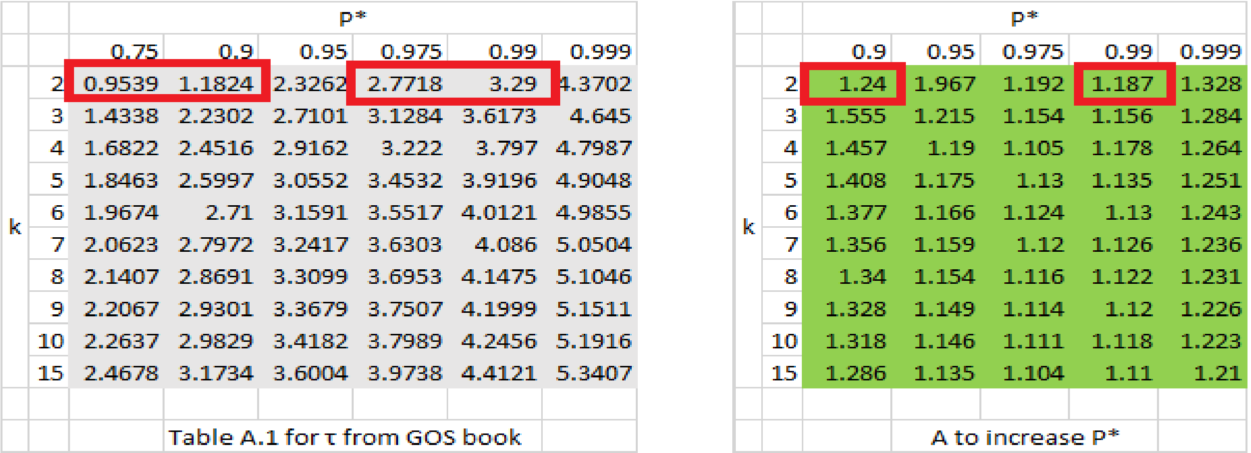


Figure 4.4: $\delta_{k+1} > \delta_k$ is Required.

Updating δ When Increasing P*

The next approach is to increase the probability of correct selection, P^*. We keep the same number of populations and have the same variances. From Table A.1 of the GOS book, it's moving one column to the next or the other on the right. If we increase P^* to the next level provided, i.e., from 0.75 to 0.9 or from 0.975 to 0.99, we can create a table of A for $\delta(new) = A\delta(old)$ and can be calculated from the ratio of $\dfrac{\tau(P^*_{new})}{\tau(P^*_{old})}$, where $\tau(P^*_{new}) > \tau(P^*_{old})$.



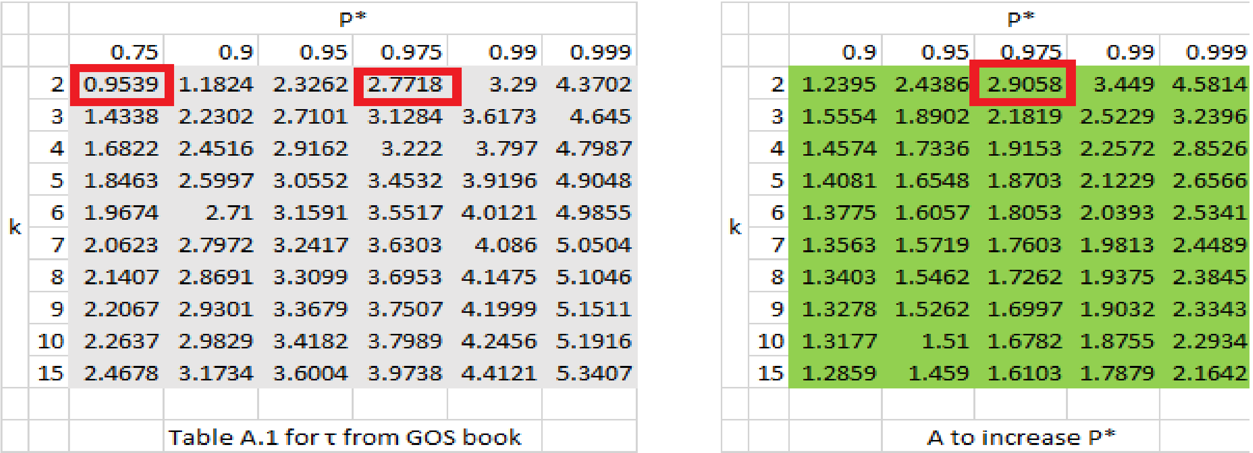
	P*					
k	0.75	0.9	0.95	0.975	0.99	0.999
2	0.9539	1.1824	2.3262	2.7718	3.29	4.3702
3	1.4338	2.2302	2.7101	3.1284	3.6173	4.645
4	1.6822	2.4516	2.9162	3.222	3.797	4.7987
5	1.8463	2.5997	3.0552	3.4532	3.9196	4.9048
6	1.9674	2.71	3.1591	3.5517	4.0121	4.9855
7	2.0623	2.7972	3.2417	3.6303	4.086	5.0504
8	2.1407	2.8691	3.3099	3.6953	4.1475	5.1046
9	2.2067	2.9301	3.3679	3.7507	4.1999	5.1511
10	2.2637	2.9829	3.4182	3.7989	4.2456	5.1916
15	2.4678	3.1734	3.6004	3.9738	4.4121	5.3407

Table A.1 for τ from GOS book

	P*				
k	0.9	0.95	0.975	0.99	0.999
2	1.24	1.967	1.192	1.187	1.328
3	1.555	1.215	1.154	1.156	1.284
4	1.457	1.19	1.105	1.178	1.264
5	1.408	1.175	1.13	1.135	1.251
6	1.377	1.166	1.124	1.13	1.243
7	1.356	1.159	1.12	1.126	1.236
8	1.34	1.154	1.116	1.122	1.231
9	1.328	1.149	1.114	1.12	1.226
10	1.318	1.146	1.111	1.118	1.223
15	1.286	1.135	1.104	1.11	1.21

A to increase P*

Figure 4.5: Increasing P^* to the Next Level and Updating δ^*



	P*					
k	0.75	0.9	0.95	0.975	0.99	0.999
2	0.9539	1.1824	2.3262	2.7718	3.29	4.3702
3	1.4338	2.2302	2.7101	3.1284	3.6173	4.645
4	1.6822	2.4516	2.9162	3.222	3.797	4.7987
5	1.8463	2.5997	3.0552	3.4532	3.9196	4.9048
6	1.9674	2.71	3.1591	3.5517	4.0121	4.9855
7	2.0623	2.7972	3.2417	3.6303	4.086	5.0504
8	2.1407	2.8691	3.3099	3.6953	4.1475	5.1046
9	2.2067	2.9301	3.3679	3.7507	4.1999	5.1511
10	2.2637	2.9829	3.4182	3.7989	4.2456	5.1916
15	2.4678	3.1734	3.6004	3.9738	4.4121	5.3407

Table A.1 for τ from GOS book

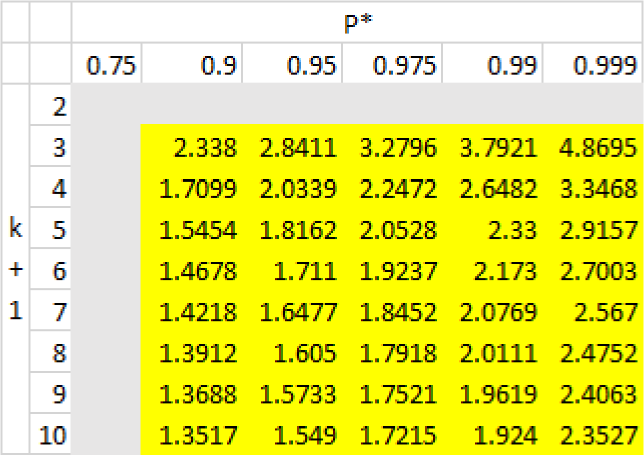
	P*				
k	0.9	0.95	0.975	0.99	0.999
2	1.2395	2.4386	2.9058	3.449	4.5814
3	1.5554	1.8902	2.1819	2.5229	3.2396
4	1.4574	1.7336	1.9153	2.2572	2.8526
5	1.4081	1.6548	1.8703	2.1229	2.6566
6	1.3775	1.6057	1.8053	2.0393	2.5341
7	1.3563	1.5719	1.7603	1.9813	2.4489
8	1.3403	1.5462	1.7262	1.9375	2.3845
9	1.3278	1.5262	1.6997	1.9032	2.3343
10	1.3177	1.51	1.6782	1.8755	2.2934
15	1.2859	1.459	1.6103	1.7879	2.1642

A to increase P*

Figure 4.6: Increasing P^* from 0.75 to a New P^* (a different column) and Updating δ^*

We can improve the selection procedure with a higher probability of CS by updating the discerning measure of distance.

Now, we combine the previous two methods by adding one more population after increasing the probability of correct selection. We make the product of A's of Figure 4.6 and Figure 4.1. Figure 4.7 below shows the multiplier, A, when we increase the probability (from P=0.75 to a new column) and a new population is added. We select one best population when k→k+1 with a new probability of correct selection.



		p*					
		0.75	0.9	0.95	0.975	0.99	0.999
	2						
	3		2.338	2.8411	3.2796	3.7921	4.8695
	4		1.7099	2.0339	2.2472	2.6482	3.3468
k	5		1.5454	1.8162	2.0528	2.33	2.9157
+	6		1.4678	1.711	1.9237	2.173	2.7003
1	7		1.4218	1.6477	1.8452	2.0769	2.567
	8		1.3912	1.605	1.7918	2.0111	2.4752
	9		1.3688	1.5733	1.7521	1.9619	2.4063
	10		1.3517	1.549	1.7215	1.924	2.3527

Figure 4.7: Increase the Probability and Add one more Population

(P^*= 0.75 to a new column and k → k+1)

In this section we update δ, the discerning measure of distance of the indifference-zone approach, based on the situation, either adding one population or increasing the probability of a correct selection. To have at least the same level of probability of correct selection, we need to have a larger value of δ than before. After the selection has been done, if we add one more population, we only can select one best population with the same level of probability. If the two best populations are selected, then the probability level must decrease.

Chapter 5

The Statistical Classification Utilizing the Indifference-Zone Approach

5.1 Variable Selection with the Indifference-Zone approach

In recent studies in classification, researchers have huge data to deal with, especially, many variables as predictors. Including all the variables when modeling classification makes the model complex, takes time in computation, makes it difficult to understand (or interpret) after the model is set, and sometimes decreases the accuracy of the model. Thus, reducing the dimensionality of predictor variables is one of the most studied fields in classification and statistical learning. This process is meaningless if reducing the dimensionality loses the important information the variables have. Thus, we want to keep the information as equal as possible while reducing the number of variables.

There are two approaches to dimension reduction methods; variable (feature) extraction and variable (feature) selection. Variable extraction uses the projection of the variables into a new variable space with lower dimensionality. Variable selection selects a subset of the variables that optimize the relevance and redundancy. Principle Component Analysis (PCA) and LDA are examples of variable extraction. Variable selection includes the techniques of Information Gain, Fisher Score, and Lasso. We focus on variable selection because variable selection retains the original variables and is easier to interpret than variable extraction.

Also, since variable selection is the learning using the training data, there are supervised variable selection and unsupervised variable selection depending on the labeled target variable. Supervised variable selection, then, can be categorized into filter methods, wrapper methods, and embedded methods. Filter methods happen before the classification process using some properties such as distance, correlation, or information. Filter methods choose the best subset of variables by evaluating them based on certain criteria. So, variables are evaluated individually. Wrapper methods consist of a series of steps; select a subset of variable, continue to finish the classification process, evaluate the performance of classification, and select a new subset to iterate the procedure. Since filter methods do not proceed to classification, they are computationally faster and less expensive than wrapper methods. Wrapper methods predict more accurately than filter methods. Embedded methods combine the advantages of filter methods and wrapper methods by including interactions of variables with the classification process with reasonable computational costs. Filter methods use Information Gain, Fisher Score, Chi-square test, and correlation coefficient as variable selection techniques. Wrapper methods' algorithms are forward variable selection, backward variable selection, and recursive variable selection. In embedded methods, there are techniques such as Lasso Regularization, Bridge regularization, and Random Forest Importance.

Among filter methods, the correlation coefficient evaluates the relevance between the variable and the target variable at the same time it measures the correlation among predictor variables. So it can evaluate the relevance and the redundancy together unlike the other filter methods. If a variable is highly correlated with the target variable and uncorrelated with other predictor variables, it should be a good variable to be selected. For this selection method, we incorporate the selection and ranking methodologies, especially, the indifference-zone approach using the correlation coefficient and multiple correlation coefficients.

5.1.1 IZ approach using Correlation Coefficients

When there are k pairs of variables, $(Y, X_1), (Y, X_2), \ldots, (Y, X_k)$, and each pair has the bivariate normal distributions with correlation coefficients, $\rho_1, \rho_2, \ldots, \rho_k$, we can select the pair with the largest ρ, $\rho_{[k]}$ with a probability condition from the indifference-zone approach methods. Also, we can select t pairs of variables from the largest correlation coefficients as the t most highly correlated pairs. We can order the correlation coefficients from the smallest to the largest as $\rho_{[1]} \leq \rho_{[2]} \leq \cdots \leq \rho_{[k]}$. If we apply this setup to a classification problem, Y is the response variable and X_is are the predictor variables, $i = 1, 2, \ldots, k$. This selection procedure based on the indifference-zone methods yields the same results as the variable selection process for predictors in classification. If we assume that we have a multivariate normal distribution for **X** = $(X_1, X_2, \ldots, X_k)'$ and the response is the linear combination of these, then the pair between the response variable and one of the predictor variables follows the bivariate normal distribution. Since the higher correlation between the response variable and the predictor variable means a higher relevance between them, we prefer the variable with a higher correlation coefficient. For a selection problem of the largest ρ, the probability requirement is

$$P(CS) \geq P^* \text{ whenever } \rho_{[k]} - \rho_{[k-1]} = \delta \geq \delta^*,$$

where P^* and δ^* are prespecified. In the same way, for the problem of selecting t largest ρ variables, the probability requirement is

$$P(CS) \geq P^* \text{ whenever } \rho_{[k-t+1]} - \rho_{[k-t]} = \delta \geq \delta^*,$$

where P^* and δ^* are prespecified. Thus, by using the method of selecting t largest correlation coefficients among k populations, we can select the variables to include the classification process.

Suppose there are 4 variables, (X_1, X_2, X_3, X_4), from a multivariate normal distribution with mean $(2, 4, 3, 7)'$ and covariance matrix

$\begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}$. We assume the variables are independent. Let's set the response variable as a sum of these four variables, Y =$(X_1 + X_2 + X_3 + X_4)$. Then, we calculate the correlation coefficients between the response variable and four variables, cor(Y,X_i), i = 1,2,3,4. They are 0.378, 0.567, 0.945, 0.756, respectively. Thus, the first variable shows the lowest relevance and we can get rid of this variable from our predictors for classification. When we generate 100 observations and calculate the sample correlations, they are 0.345, 0.394, 0.682,0.458, respectively. It still shows the least relevance between the first variable and the response variable. If we make two classes for the response variable, we make one class if the response is greater than 15 and the other class otherwise. From the same generated data, the correlations are 0.2785, 0.4274, 0.3230, 0.4241. The first variable has the lowest correlation with the response again. Then, in the indifference-zone approach, we can set the discerning measure of distance and the level of probability of a correct decision and find the number of samples needed from Table I of [4]. Then, we can select a fixed number of variables to use in the classification process.

5.1.2 IZ Approach Using Multiple Correlation Coefficients

The correlation between the response variable and the predictor variable represents the relevance. We selected variables with a high correlation with the response variable in the previous section. If there is a high correlation between two predictors, one of the variables overlaps with the other and the second variable does not add much information to the model. These variables are said to exhibit redundancy. Also, the presence of redundant variables overfits the classification model and lengthens computations. Thus, we want to select variables with high relevance and low redundancy. If we observe high redundancy, we do not include it in the subset of selected predictor variables for classification. In other words, the multicollinearity represents the dependency among the variables and can be measured by the variance inflation factor(VIF). The larger the VIF, the higher the dependency. Since VIF and Multiple Correlation Coefficient(MCC) are related by the following equation, $VIF = \frac{1}{1 - MCC^2}$, we use MCC for variable selection process based on the indifference-zone approach. We remove the variables with MCC closer to 1. For MCC, the subset selection method is also available by Gupta and Panchapakesan (1969).

The multiple correlation coefficient measures the relationship between one variable and the others. Suppose there are k variables, $X_1, X_1, \ldots, X_k$. Then, the multiple correlation coefficient, ρ_i, measures the relationship between X_i and $(X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_k)$, $i = 1, 2, \ldots, k$.

$$\rho_i = \sqrt{1 - \frac{|R(X)|}{R_i(X)}} = \sqrt{-\frac{|R(X)| - R_i(X)}{R_i(X)}},$$

where $R(X) = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1k} \\ \rho_{21} & 1 & \rho_{23} & \cdots & \rho_{2k} \\ \rho_{31} & \rho_{32} & 1 & \cdots & \rho_{3k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{k1} & \rho_{k2} & \rho_{k3} & \cdots & 1 \end{bmatrix}$ is the matrix with all correlations, $|\cdot|$ is the determinant of the matrix and $R_i(X)$ is the minor of $R(X)$ with i^{th} row and column are removed.

Suppose there are 4 variables from a multivariate normal distribution with mean vector of $(2, 4, 3, 7)'$ and the covariance matrix, $\begin{bmatrix} 2 & 0 & 1 & 2 \\ 0 & 2 & 0 & 0 \\ 1 & 0 & 4 & 1 \\ 2 & 0 & 1 & 4 \end{bmatrix}$. We generated 100 data for each variable and calculated MCC for each variable.

$\rho_1 = 0.75895$,

$\rho_2 = 0.12468$,

$\rho_3 = 0.32115$,

$\rho_4 = 0.74338$.

The first variable has the highest MCC and the second variable shows the lowest. From the correlation matrix below, we can verify that these numbers make sense. First of all, the second variable is independent of all other variables so it has a very low MCC, 0.1246. The first variable is highly correlated with the fourth variable and still correlated with the third variable. Thus, it has the highest MCC, 0.75895.

$$R(X) = \begin{bmatrix} 1.000 & -0.110 & 0.321 & 0.742 \\ -0.110 & 1.000 & -0.031 & -0.121 \\ 0.321 & -0.031 & 1.000 & 0.225 \\ 0.742 & -0.121 & 0.225 & 1.000 \end{bmatrix}.$$

Example 2

We generate 1000 samples from two populations with a common covariance matrix for 10000 times. The mean vectors and the covariance matrix are

$$\mu_1 = \begin{bmatrix} 3 \\ 4 \\ 3 \\ 5 \end{bmatrix}, \quad \mu_2 = \begin{bmatrix} 3 \\ 1 \\ 3 \\ 5 \end{bmatrix}, \text{ and } \quad \Sigma = \begin{bmatrix} 2 & 1 & 1 & 2 \\ 1 & 2 & 0 & 0 \\ 1 & 0 & 4 & 1 \\ 2 & 0 & 1 & 4 \end{bmatrix}.$$

We use LDA to classify and record the misclassified points by removing one variable each. Also, MCC from the last sample is recorded.

	Total Misclassified Points
Full Model	1,192,012
Without X_1	2,881,374
Without X_2	9,717,441
Without X_3	1,333,033
Without X_4	2,090,249

	MCC
X_1	0.7983
X_2	0.5061
X_3	0.3763
X_4	0.7495

Table 5.1: Results from Simulation: Misclassified Points (left) and MCC from the Last Simulation (right)

From MCC, X_1 needs to be removed from classification. However, when we check the cumulative misclassified points with one of the variables removed, X_3 is the one that can be removed. Removing X_1 increases the misclassified points almost three times. If we check the sample correlation matrix among the input variables along with the response variable from the last sample, we can notice why X_3 shows the least changes but X_2 shows the most changes.

$$R = \begin{bmatrix} 1.000 & 0.3234 & 0.3449 & 0.6965 & 0.0430 \\ 0.3234 & 1.000 & -0.0285 & -0.0361 & -0.7138 \\ 0.3449 & -0.0285 & 1.000 & 0.2630 & 0.0274 \\ 0.6965 & -0.0361 & 0.2630 & 1.000 & 0.0352 \\ 0.0430 & -0.7138 & \mathbf{0.0274} & 0.0352 & 1.000 \end{bmatrix}.$$

X_3 has the smallest correlation with the response variable and X_2 shows the largest correlation with the response variable. Thus, we can remove X_3 from the classification but need to keep X_2. Variable selection appears to be more influenced by correlation than MCC.

5.1.3 Summary

Using the indifference-zone approach, we select several variables that show high relevance with the response variable and low redundancy among the input variables. The correlation coefficients for the relevance and the multiple correlation coefficients for the redundancy are used in the indifference-zone approach.

Chapter 6

Conclusion

6.1 Concluding Remarks

In this dissertation, we focused on improving statistical classification using the multiple decision-theoretic perspective along with the probability of a correct decision. First, we showed that the probability of correct decision got improved as we proposed a method of increasing dimensions by introducing a preferable predictor vector to the classification problem when the populations are not linearly separable in the current vector space. Adding a preferable predictor vector to the 2-dimensional classification problem resulted in improving the separability of the populations and causing a higher probability of correct decision in the case of either with Normal population assumption or without the distribution assumptions. To show the improvement in the probability of correct decision, the total probability of misclassification (TPM) was calculated in the case of a normal distribution, and the apparent error rate (APER) was used in the case where the distribution of the population was not considered. In both cases, we showed that the calculated TPM and APER got smaller after adding a preferable predictor vector to the problem.

We investigated the conditions on the preferable predictor vector especially when we add one variable to the 2-dimensional multivariate normal distribution case. The conditions depend on the value that is called the separability, the ratio between the mean difference and the variance of the new variable. If the new variable makes the mean distance larger and the variance of itself smaller, the performance of classification improves significantly. Also, we find the range of mean difference of the new variable between the groups and the variance of the new variable depending on the various situations of covariance. This helps to find a new variable by data mining or machine learning.

Also, we updated the discerning measure of distance, δ, which works as the classifier in the indifference-zone approach. The indifference-zone approach can be viewed as a classification process and we update the classifier as we intend to improve the correct selection by updating δ. We found a new decision rule (allocation rule) after we update δ. When you add a new population once the indifference-zone approach classifies the populations into two groups, the mean of the new sample needs to be greater than the value of updated δ added to the current largest sample mean to be selected as the best population.

Lastly, we apply the indifference-zone approach to variable selection method. The variable selection is one field of techniques that improves the performance of classification. We use the correlation coefficient to select the variables in terms of relevance and the multiple correlation coefficient (MCC) concerning the redundancy. In the former case, the correlations between the response variable and predictor variables are used and the variables with high correlation are selected. The MCC was calculated among the predictor variables and the variables with high MCC are eliminated from the classification procedure in the latter case. The indifference-zone approach was incorporated to variable selection process as a part of filter methods which is the supervised learning process. The indifference-zone approach also can be viewed as supervised learning and this would be the first step to improve classification with the help of multiple decision-theoretic approaches.

6.2 Future Research

We can apply the selection and ranking methodologies to clustering by using the subset selection method and extend to unsupervised learning. We investigated the indifference-zone approach with classification as a supervised learning. The subset selection method chooses a random number of subsets of populations based on the distance measure from the best population, and it seems very similar to the approach of clustering that identifies the groups based on similarity.

When we searched for the preferable predictor vector, we mainly assumed no correlation in the populations of a two-dimensional space. It can be extended to a model with correlated variables in higher dimensions. The cases when the existing variables are correlated, i.e., the presence of multicollinearity, or the new variable is also correlated to the existing variables need to be studied. Also, if a correlation exists, the difference in the means of the new variable and the position of the means according to the covariance structure must be considered.

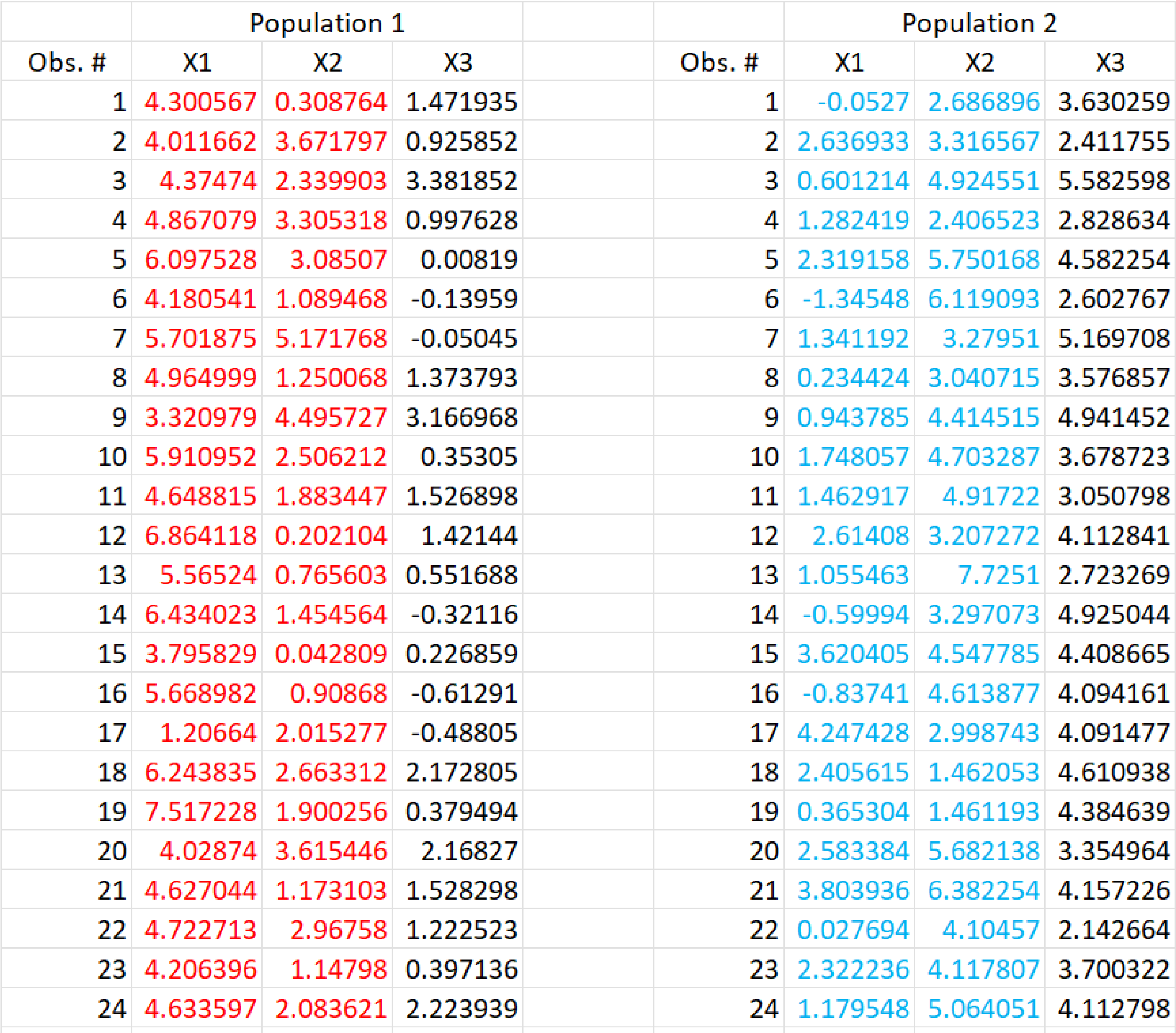
Additionally, we can find the conditions of new variables under the assumption of different covariance matrices, non-homogeneous variance structure, which results in using the quadratic discriminant analysis (QDA).

In a variable selection, the subset selection method can be used. We used the indifference-zone approach in selecting variables as a filter method. Then, the subset selection method using the correlation coefficients or the multiple correlation coefficients can be suggested.

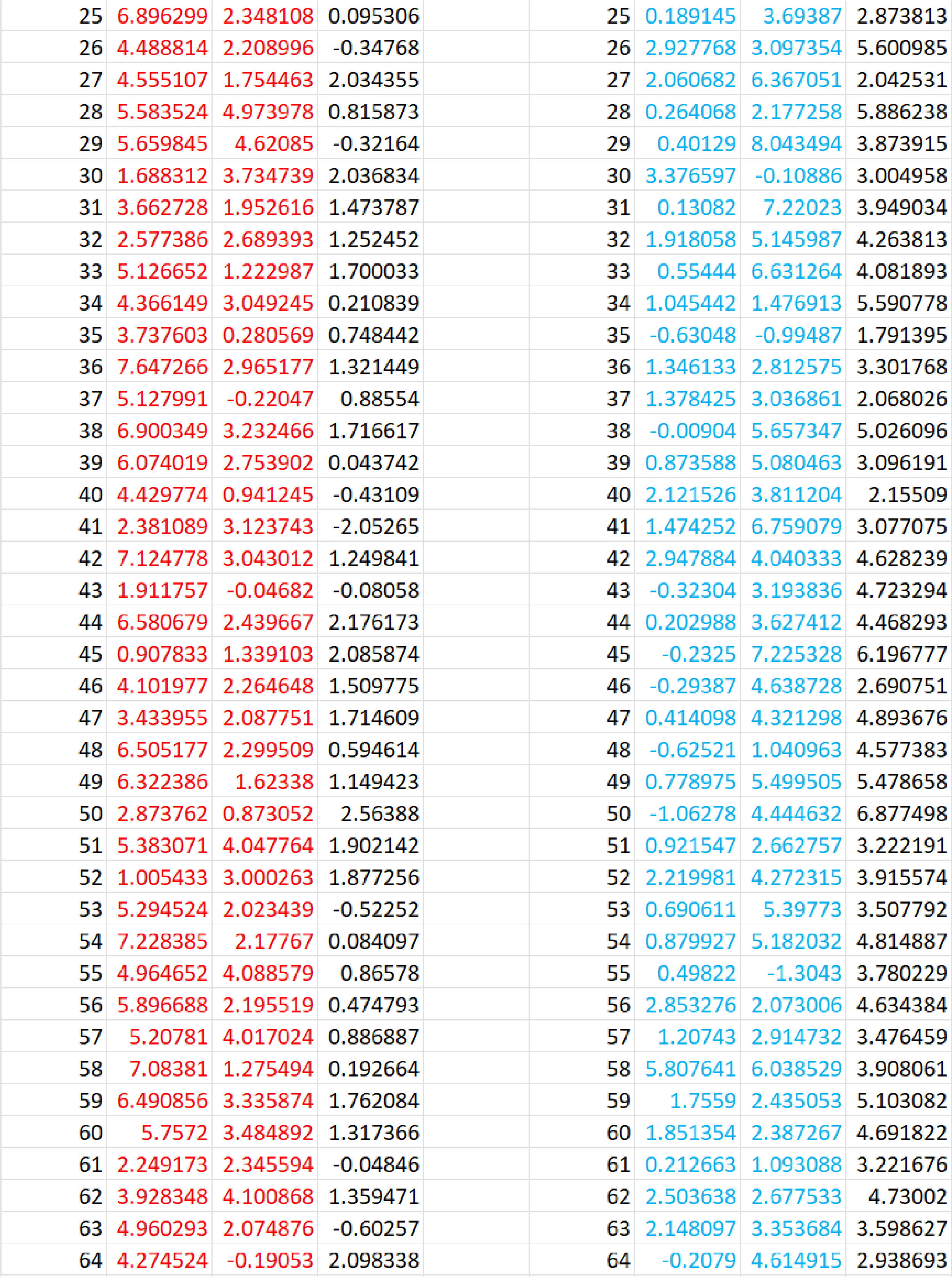
Finally, it is open to a different distribution than the normal distribution in the indifference-zone approach such as multinomial distribution to select the best population and apply to classification.

Appendix A

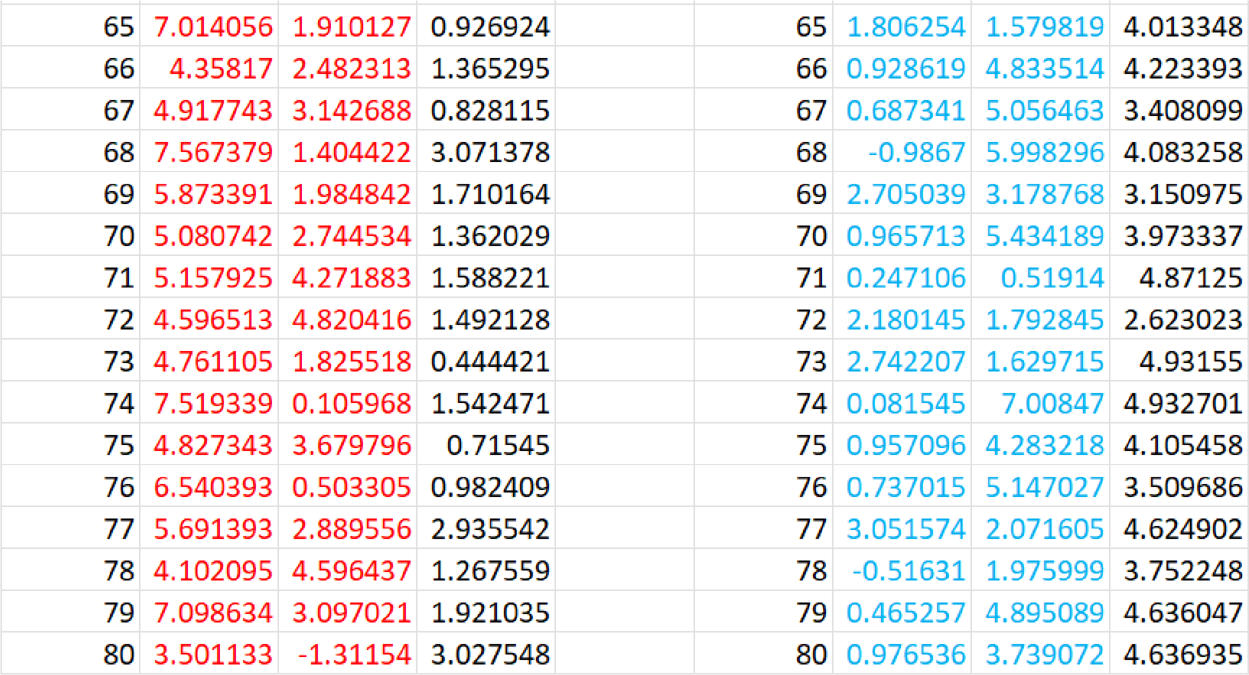
Table of Simulated Data



	Population 1					Population 2		
Obs. #	X1	X2	X3		Obs. #	X1	X2	X3
1	4.300567	0.308764	1.471935		1	-0.0527	2.686896	3.630259
2	4.011662	3.671797	0.925852		2	2.636933	3.316567	2.411755
3	4.37474	2.339903	3.381852		3	0.601214	4.924551	5.582598
4	4.867079	3.305318	0.997628		4	1.282419	2.406523	2.828634
5	6.097528	3.08507	0.00819		5	2.319158	5.750168	4.582254
6	4.180541	1.089468	-0.13959		6	-1.34548	6.119093	2.602767
7	5.701875	5.171768	-0.05045		7	1.341192	3.27951	5.169708
8	4.964999	1.250068	1.373793		8	0.234424	3.040715	3.576857
9	3.320979	4.495727	3.166968		9	0.943785	4.414515	4.941452
10	5.910952	2.506212	0.35305		10	1.748057	4.703287	3.678723
11	4.648815	1.883447	1.526898		11	1.462917	4.91722	3.050798
12	6.864118	0.202104	1.42144		12	2.61408	3.207272	4.112841
13	5.56524	0.765603	0.551688		13	1.055463	7.7251	2.723269
14	6.434023	1.454564	-0.32116		14	-0.59994	3.297073	4.925044
15	3.795829	0.042809	0.226859		15	3.620405	4.547785	4.408665
16	5.668982	0.90868	-0.61291		16	-0.83741	4.613877	4.094161
17	1.20664	2.015277	-0.48805		17	4.247428	2.998743	4.091477
18	6.243835	2.663312	2.172805		18	2.405615	1.462053	4.610938
19	7.517228	1.900256	0.379494		19	0.365304	1.461193	4.384639
20	4.02874	3.615446	2.16827		20	2.583384	5.682138	3.354964
21	4.627044	1.173103	1.528298		21	3.803936	6.382254	4.157226
22	4.722713	2.96758	1.222523		22	0.027694	4.10457	2.142664
23	4.206396	1.14798	0.397136		23	2.322236	4.117807	3.700322
24	4.633597	2.083621	2.223939		24	1.179548	5.064051	4.112798



25	6.896299	2.348108	0.095306	25	0.189145	3.69387	2.873813
26	4.488814	2.208996	-0.34768	26	2.927768	3.097354	5.600985
27	4.555107	1.754463	2.034355	27	2.060682	6.367051	2.042531
28	5.583524	4.973978	0.815873	28	0.264068	2.177258	5.886238
29	5.659845	4.62085	-0.32164	29	0.40129	8.043494	3.873915
30	1.688312	3.734739	2.036834	30	3.376597	-0.10886	3.004958
31	3.662728	1.952616	1.473787	31	0.13082	7.22023	3.949034
32	2.577386	2.689393	1.252452	32	1.918058	5.145987	4.263813
33	5.126652	1.222987	1.700033	33	0.55444	6.631264	4.081893
34	4.366149	3.049245	0.210839	34	1.045442	1.476913	5.590778
35	3.737603	0.280569	0.748442	35	-0.63048	-0.99487	1.791395
36	7.647266	2.965177	1.321449	36	1.346133	2.812575	3.301768
37	5.127991	-0.22047	0.88554	37	1.378425	3.036861	2.068026
38	6.900349	3.232466	1.716617	38	-0.00904	5.657347	5.026096
39	6.074019	2.753902	0.043742	39	0.873588	5.080463	3.096191
40	4.429774	0.941245	-0.43109	40	2.121526	3.811204	2.15509
41	2.381089	3.123743	-2.05265	41	1.474252	6.759079	3.077075
42	7.124778	3.043012	1.249841	42	2.947884	4.040333	4.628239
43	1.911757	-0.04682	-0.08058	43	-0.32304	3.193836	4.723294
44	6.580679	2.439667	2.176173	44	0.202988	3.627412	4.468293
45	0.907833	1.339103	2.085874	45	-0.2325	7.225328	6.196777
46	4.101977	2.264648	1.509775	46	-0.29387	4.638728	2.690751
47	3.433955	2.087751	1.714609	47	0.414098	4.321298	4.893676
48	6.505177	2.299509	0.594614	48	-0.62521	1.040963	4.577383
49	6.322386	1.62338	1.149423	49	0.778975	5.499505	5.478658
50	2.873762	0.873052	2.56388	50	-1.06278	4.444632	6.877498
51	5.383071	4.047764	1.902142	51	0.921547	2.662757	3.222191
52	1.005433	3.000263	1.877256	52	2.219981	4.272315	3.915574
53	5.294524	2.023439	-0.52252	53	0.690611	5.39773	3.507792
54	7.228385	2.17767	0.084097	54	0.879927	5.182032	4.814887
55	4.964652	4.088579	0.86578	55	0.49822	-1.3043	3.780229
56	5.896688	2.195519	0.474793	56	2.853276	2.073006	4.634384
57	5.20781	4.017024	0.886887	57	1.20743	2.914732	3.476459
58	7.08381	1.275494	0.192664	58	5.807641	6.038529	3.908061
59	6.490856	3.335874	1.762084	59	1.7559	2.435053	5.103082
60	5.7572	3.484892	1.317366	60	1.851354	2.387267	4.691822
61	2.249173	2.345594	-0.04846	61	0.212663	1.093088	3.221676
62	3.928348	4.100868	1.359471	62	2.503638	2.677533	4.73002
63	4.960293	2.074876	-0.60257	63	2.148097	3.353684	3.598627
64	4.274524	-0.19053	2.098338	64	-0.2079	4.614915	2.938693



65	7.014056	1.910127	0.926924		65	1.806254	1.579819	4.013348
66	4.35817	2.482313	1.365295		66	0.928619	4.833514	4.223393
67	4.917743	3.142688	0.828115		67	0.687341	5.056463	3.408099
68	7.567379	1.404422	3.071378		68	-0.9867	5.998296	4.083258
69	5.873391	1.984842	1.710164		69	2.705039	3.178768	3.150975
70	5.080742	2.744534	1.362029		70	0.965713	5.434189	3.973337
71	5.157925	4.271883	1.588221		71	0.247106	0.51914	4.87125
72	4.596513	4.820416	1.492128		72	2.180145	1.792845	2.623023
73	4.761105	1.825518	0.444421		73	2.742207	1.629715	4.93155
74	7.519339	0.105968	1.542471		74	0.081545	7.00847	4.932701
75	4.827343	3.679796	0.71545		75	0.957096	4.283218	4.105458
76	6.540393	0.503305	0.982409		76	0.737015	5.147027	3.509686
77	5.691393	2.889556	2.935542		77	3.051574	2.071605	4.624902
78	4.102095	4.596437	1.267559		78	-0.51631	1.975999	3.752248
79	7.098634	3.097021	1.921035		79	0.465257	4.895089	4.636047
80	3.501133	-1.31154	3.027548		80	0.976536	3.739072	4.636935

Table A.1: Simulated Data in 3.2

Bibliography

[1] Alam, K. and Rizvi, M. H. (1966). Selection from multivariate normal populations. *Annals of the Institute of Statistical Mathematics*, 18(1):307–318.

[2] Alam, K., Rizvi, M. H., and Solomon, H. (1976). Selection of largest multiple correlation coefficients: Exact sample size case. *The Annals of Statistics*, 4(3):614–620.

[3] Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. Wiley series in probability and statistics. Wiley-Interscience, Hoboken, N.J., 3rd edition.

[4] Bechhofer, R. E. (1954). A single-sample multiple decision procedure for ranking means of normal populations with known variances. *The Annals of Mathematical Statistics*, 25(1):16–39.

[5] Bechhofer, R. E., Dunnett, C. W., and Sobel, M. (1954). A two-sample multiple decision procedure for ranking means of normal populations with a common unknown variance. *Biometrika*, 41(1-2):170–176.

[6] Bechhofer, R. E., Kiefer, J., and Sobel, M. (1968). *Sequential identification and ranking procedures: with special reference to Koopman-Darmois populations*. Chicago, University of Chicago Press.

[7] Bechhofer, R. E., Santner, T. J., and Goldsman, D. M. (1995). *Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons*. New York : Wiley.

[8] Bechhofer, R. E. and Sobel, M. (1954). A single-sample multiple decision procedure for ranking variances of normal populations. *The Annals of Mathematical Statistics*, 25(2):273–289.

[9] Berger, J. O. (1980). *Statistical decision theory, foundations, concepts, and methods*. Springer series in statistics. Springer-Verlag, New York.

[10] Bishop, C. M. (1995). *Neural networks for pattern recognition*. Clarendon Press ; Oxford University Press, Oxford [England] : New York.

[11] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Information science and statistics. Springer, New York.

[12] Brown, L. D. (2000). An essay on statistical decision theory. *Journal of the American Statistical Association*, 95(452):1277–1281.

[13] Bunge, J. A. and Judson, D. H. (2004). Data mining. In *Encyclopedia of Social Measurement*, volume 1, pages 617–624.

[14] Cannings, T. I. and Samworth, R. J. (2017). Random-projection ensemble classification. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 79(4):959–1035.

[15] Chernoff, H. and Yahav, J. (1977). A subset selection problem employing a new criterion. In *Statistical Decision Theory and Related Topics*, pages 93–119. Elsevier Inc, United States.

[16] Cho, H. A. (2009). Statistical identification in multinomial models with sequential sampling. *American Journal of Mathematical and Management Sciences*, 29(1-2):139–156.

[17] Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27.

[18] Duda, R. O. (2001). *Pattern Classification*. Wiley, New York, 2nd edition.

[19] Dudewicz, E. J. (1980). Ranking (ordering) and selection: An overview of how to select the best. *Technometrics*, 22(1):113–119.

[20] Dunnett, C. W. (1960). On selecting the largest of k normal population means. *Journal of the Royal Statistical Society. Series B, Methodological*, 22(1):1–40.

[21] Eatwell, J., Milage, M., and Newman, P. (1989). *Game Theory*. The New Palgrave. The Macmillan Press Ltd., London.

[22] Efron, B. (1975). Biased versus unbiased estimation. *Advances in mathematics (New York. 1965)*, 16(3):259–277.

[23] Faraway, J. J. (2006). *Extending the linear model with R generalized linear, mixed effects and nonparametric regression models*. Texts in statistical science. Chapman & Hall/CRC, Boca Raton.

[24] Fisher, R. A. (1938). The statistical utilization of multiple measurements. *Annals of Eugenics*, 8(4):376–386.

[25] Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association*, 84(405):165–175.

[26] Geisser, S. (1967). Estimation associated with linear discriminants. *The Annals of mathematical statistics*, 38(3):807–817.

[27] Gibbons, J. D., Olkin, I., and Sobel, M. (1977). *Selecting and Ordering Populations: A New Statistical Methodology*. New York : Wiley.

[28] Golberg, M. A. and Cho, H. A. (2010). *Introduction to Regression Analysis*. WIT Press ; Computational Mechanics Inc., Southhampton, UK : Billerica, MA.

[29] Goldstein, M. (1978). *Discrete discriminant analysis*. Wiley series in probability and mathematical statistics. Wiley, New York.

[30] Gordon, A. D. (1999). *Classification*. Monographs on statistics and applied probability (Series) ; 82. Chapman & Hall/CRC, Boca Raton, 2nd edition.

[31] Gupta, S. S. (1956). *On a Decision Rule for a Problem in Ranking Means*. University of North Carolina at Chapel Hill.

[32] Gupta, S. S. (1977). *Selection and Ranking Procedures: A Brief Introduction*.

[33] Gupta, S. S. and Panchapakesan, S. (1979). *Multiple Decision Procedures : Theory and Methodology of Selecting and Ranking Populations*. New York : Wiley.

[34] Gupta, S. S., Panchapakesan, S., and Balakrishnan, N. (1997). *Advances in statistical decision theory and applications*. Statistics for industry and technology. Birkhäuser, Boston ; Basel ; Berlin, 1st ed. 1997.. edition.

[35] Gupta, S. S. and Sobel, M. (1957). On a statistic which arises in selection and ranking problems. *The Annals of Mathematical Statistics*, 28(4):957–967.

[36] Hand, D. J. D. J. (1981). *Discrimination and classification*. Wiley series in probability and mathematical statistics. Applied probability and statistics. J. Wiley, Chichester [Chichestershire] ; New York.

[37] Hand, D. J. D. J. (1997). *Construction and assessment of classification rules*. Wiley series in probability and statistics. Probability and statistics. Wiley, Chichester, West Sussex, England ; New York.

[38] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in statistics. Springer, New York.

[39] Hsu, J. C. (1996). *Multiple Comparisons : Theory and Methods*. Chapman & Hall, London, 1st edition.

[40] Härdle, W. (2003). *Applied multivariate statistical analysis*. Springer, Berlin ; New York.

[41] James, M. (1985). *Classification algorithms*. Wiley, New York.

[42] Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and physical sciences*, 186(1007):453–461.

[43] Johnson, R. A. and Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*. Pearson Prentice Hall, Upper Saddle River, N.J., 6th edition.

[44] Mahamunulu, D. M. (1967). Some fixed-sample ranking and selection problems. *The Annals of mathematical statistics*, 38(4):1079–1091.

[45] McCullagh, P. P. (1998). *Generalized linear models*. Monographs on statistics and applied probability (Series) ; 37. Chapman & Hall/CRC, Boca Raton, Fla., 2nd ed. edition.

[46] McCulloch, W. S. and Pitts, W. (1990). A logical calculus of the ideas immanent in nervous activity. *Bulletin of mathematical biology*, 52(1):99–115.

[47] Miller, R. G. (1966). *Simultaneous statistical inference*. McGraw-Hill, New York.

[48] Milton, R. C. and Rizvi, M. H. (1989). On computation of integrals for selection from multivariate normal populations on the basis of distances. *Annals of the Institute of Statistical Mathematics*, 41(2):401–408.

[49] Morgan, J. N. and Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, 58(302):415–434.

[50] Mosteller, F. and Tukey, J. W. (1950). Significance levels for a k-sample slippage test. *The Annals of mathematical statistics*, 21(1):120–123.

[51] Nilsson, N. J. (1965). *Learning machines ; foundations of trainable pattern-classifying systems*. McGraw-Hill series in systems science. McGraw-Hill, New York.

[52] Olkin, I., Sobel, M., and Tong, Y. L. (1976). *Estimating the True Probability of Correct Selection for Location and Scale Parameter Families*.

[53] Panchapakesan, S. (2005). Restricted subset selection procedures for normal means: A brief review with a fresh look at the classical formulations of bechhofer and gupta. *Communications in statistics. Theory and methods*, 34(6):1265–1273.

[54] Paulson, E. (1949). A multiple decision procedure for certain problems in the analysis of variance. *The Annals of mathematical statistics*, 20(1):95–98.

[55] Peter O. Anderson, T. A. B. and Dudewicz, E. J. (1977). Indifference-zone ranking and selection: confidence intervals for true achieved p(cd). *Communications in Statistics - Theory and Methods*, 6(11):1121–1132.

[56] Rizvi, M. H. and Solomon, H. (1973). Selection of largest multiple correlation coefficients: Asymptotic case. *Journal of the American Statistical Association*, 68(341):184–188.

[57] Ross Quinlan, J. and Rivest, R. L. (1989). Inferring decision trees using the minimum description length principle. *Information and computation*, 80(3):227–248.

[58] Sobel, M. and Huyett, M. J. (1957). Selecting the best one of several binomial populations. *Bell System Technical Journal*, 36(2):537–576.

[59] Sobel, M. and Tong, Y. L. (1971). Optimal allocation of observations for partitioning a set of normal populations in comparison with a control. *Biometrika*, 58(1):177–181.

[60] Sugiyama, M. (2015). *Introduction to Statistical Machine Learning*. Elsevier Science & Technology, San Diego, 1 edition.

[61] Tang, J., Alelyani, S., and Liu, H. (2014). Feature selection for classification: A review. In *Data Classification: Algorithms and Applications*, pages 37–64.

[62] Tong, Y. L. (1969). On partitioning a set of normal populations by their locations with respect to a control. *The Annals of mathematical statistics*, 40(4):1300–1324.

[63] Vapnik, V. N. (1998). *Statistical Learning Theory*. New York : Wiley.

[64] Vapnik, V. N. (2000). *The Nature of Statistical Learning Theory*. New York : Springer, 2nd edition.

[65] Vapnik, V. N. and Lerner, A. Y. (1963). Recognition of patterns with help of generalized portraits. *Avtomatika i Telemekhanika*, 24(6):774–780.

[66] Wald, A. (1944). On a statistical problem arising in the classification of an individual into one of two groups. *The Annals of Mathematical Statistics*, 15(2):145–162.

[67] Wilcox, R. R. (1980). Some exact sample sizes for comparing the squared multiple correlation coefficient to a standard. *Educational and psychological measurement*, 40(1):119–124.

Curriculum Vitae

Graduate College
University of Nevada, Las Vegas

Jeong Jun Lee
hyperjjl@gmail.com

Degrees:

Bachelor in Business Administration, Feb 2001
Yonsei University, Korea

Master of Art in Business Economics, Jun 2003
University of California, Santa Barbara

Master of Art in Applied Statistics, Sep, 2004
University of California, Santa Barbara

Dissertation Title:

Statistical Classification Using Selection and Ranking Methodologies with Statistical Learning

Dissertation Examination Committee:

Chairperson, Hokwon Cho, Ph.D.
Committee Member, Amei Amei, Ph.D.
Committee Member, Malwane Ananda, Ph.D.
Committee Member, Kaushik Ghosh, Ph.D.
Graduate Faculty Representative, Jaewon Lim, Ph.D.

Presentation:

"Classification Using Statistical Learning with Multiple Decision Theoretic Perspective", ASA Nevada Chapter Fall Symposium, Las Vegas, NV, Oct 2022

Work Experience:

Statistical Consultant, Statistical Consulting Laboratory, UCSB, Sep 2007-Mar, 2008
Head Teaching Assistant, Department of Statistics and Applied Probability, UCSB, Jan 2009-Jun 2011