Stockholms universitet

A Comparative Simulation Study of Logistic Regression and Linear Discriminant Analysis for Classification

Sofie Jörgensen

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2019:10
Matematisk statistik
Juni 2019

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholm University

Mathematical Statistics
Stockholm University
Bachelor Thesis **2019:10**
http://www.math.su.se

A Comparative Simulation Study of Logistic Regression and Linear Discriminant Analysis for Classification

Sofie Jörgensen*

June 2019

Abstract

The process of learning from data is central in statistical learning and the problem of selecting an appropriate method for a particular situation can be rather challenging. Two widely used linear classification methods are logistic regression and linear discriminant analysis. This thesis aims to compare these linear classifiers, partly from a theoretical perspective and partly through practical simulations, in order to study their similarities and differences in several aspects. The theoretical part outlines the concept of statistical learning and provides a detailed presentation of the methods of interest. The simulation part contains four experiments with different setups in order to evaluate the predictive power for each method. It turned out that logistic regression and linear discriminant analysis performed similarly, despite the fact that a variety of simulated data sets were used. Some notable differences were observed, which can be explained by the two methods' different ways of estimating parameters. Overall, the simulation study agreed with the theory provided in this thesis, that the two methods give similar results but that their prediction accuracy might deviate slightly from each other in some situations. This emphasizes the importance of examining the underlying structure of data before determining which method to use.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: jorgensen.sof@gmail.com. Supervisor: Ola Hössjer.

Acknowledgements

This is a bachelor's thesis of 15 ECTS in Mathematical Statistics at the Department of Mathematics at Stockholm University. First, I want to thank my supervisor Ola Hössjer for the invaluable theoretical discussions, advice and guidance in mathematical statistics with great commitment and encouragement. Further, I also want to pay attention to my fellow students for the rewarding discussions and suggestions. I also want to express my gratitude to my family and friends who have supported me throughout my education and my work on this thesis. Last but not least, I want to give a special thanks to my childhood friend Nicola Fitzgerald for proofreading this thesis.

Contents

1 Introduction

The central concept of statistics and machine learning is the process of *learning from data*. Many of today's problems, for instance in the fields of medicine, industries and finance, can advantageously be solved by learning from available information [6]. Although this concept is based on knowledge from years of research, this area is still under enormous development as the demand for solving *Big Data* problems is increasing. Problems which were previously considered to be unsolvable, can now be solved by using techniques from *statistical learning*. For this reason, the elements of statistical learning have a significant role in science and research. Roughly speaking, statistical learning can be divided into two categories; *supervised* and *unsupervised*. The main difference between these two categorizations of statistical learning is the availability of the outcome. In supervised learning, the learning process is guided by the presence of both input and output variables, in order to predict the output for new observations. In contrast, a more challenging situation occurs when only measurement of the input variables are available and this falls under the category unsupervised learning. The problems in supervised learning can in turn be divided into *regression* and *classification* depending on whether the type of the output variable is either *quantitative* or *qualitative*. All classification methods which are used for identifying and assigning an observation to a specific category or *class*, are called *classifiers*. With regard to linear classifiers, *logistic regression* and *linear discriminant analysis* are two well-known parametric methods for predicting qualitative responses and are commonly used in many fields.

Further, the question often arises about which method is more suitable to apply on a particular set of data. Since both logistic regression and linear discriminant analysis usually perform similarly, we are interested in investigating the distinctions between these two methods. Thus, the intention is to give a motivation of when one method is more preferred over the other. The approach is as follows: In this thesis, we are going to study two linear statistical classification methods, namely logistic regression and linear discriminant analysis. To facilitate the study, this thesis is divided into two parts; one theory part and one practical part. We are going to compare these two linear classifiers in order to evaluate their predictive power in different situations from a mathematical and statistical point of view.

1.1 Background

In supervised learning, logistic regression and linear discriminant analysis are two widely used methods for statistical classification problems [6, 7]. Logistic regression is a commonly used method in many fields and is frequently used in binary classification. On the other hand, another frequently used method is linear discriminant analysis, which is based on more assumptions of the underlying data compared to logistic regression. Despite this, linear discriminant analysis usually has higher prediction accuracy when all the assumptions are approximately satisfied. As long as data approximately comes from a multivariate normal distribution with a common covariance matrix, then a good choice of method is linear discriminant analysis. For a real-life data set, it is rather unlikely that all conditions in linear discriminant analysis are satisfied. Because of this, logistic regression is considered to be preferred in such situations. Generally, logistic regression is often assumed to be more flexible and robust in this context and it is not affected by outliers to the same extent as linear discriminant analysis. However, linear discriminant analysis is usually preferred over logistic regression with respect to nominal outputs. The main difference between logistic regression and linear discriminant analysis is the procedure of estimating the parameters. Overall, these two linear classifiers are closely related and usually produce similar results.

This leads us to some interesting questions that we find worth investigating. Can we expect similar results despite violated assumptions for linear discriminant analysis? Is it possible to find situations where a method proves to be significantly better in prediction accuracy than the other? Or conversely, is there no major difference between the two classifiers with respect to predictive performance and thus no distinction of which method to be selected? In more complicated cases, linear classification methods can be incapable of capturing the underlying structure of the data, which often leads to underfitting. One option is to use a more complex non-linear method such as support vector machines, tree-based methods or neural networks. Nevertheless, there are several extensions of both logistic regression and linear discriminant analysis that can improve the adaptation to data and one extension of linear discriminant analysis is quadratic discriminant analysis, which will also be considered in this thesis. However, since the focus is to compare these two linear classifiers with respect to their properties and performance in prediction accuracy, we will delimit ourselves and only consider quadratic discriminant analysis more briefly, in order to prevent the study from becoming too extensive.

1.2 Aim and Purpose

A comparison of logistic regression and linear discriminant analysis aims to give a deeper understanding of linear classification methods. The purpose is to statistically evaluate the properties for each method, in order to investigate how much influence the choice of either logistic regression or linear discriminant analysis may have on the predictive power, both theoretically and practically. In particular, the intention is to provide some guidelines and emphasize the importance of selecting the most appropriate linear classifier for the situation at hand. In more detail, we are going to study the adaptability of these classifiers on a variety of simulated data sets by modelling the number of observations, dimensions, correlation, etcetera. From this we will distinguish the similarities and differences, as well as discuss the advantages and disadvantages of each method. Moreover, the interpretation and computational complexity will be discussed. All this will be the basis for the analysis of when one method dominates and is preferred over the other.

1.3 Outline

The outline of this thesis starts with section 2, covering the central theory of statistical learning and the classification methods. The setup for the simulation is explained in section 3, followed by a presentation of the results in section 4. Section 5 contains the discussion of the content of the theory, the simulation results, the overall study and possible improvements. Lastly, the conclusions are stated in section 6.

2 Theory

The following section covers the key concepts and framework of statistical learning, as well as all relevant theory of the two parametric linear classification methods of interest, namely logistic regression and linear discriminant analysis. One natural extension of linear discriminant analysis is quadratic discriminant analysis and a general description of this method will also be given. The fundamental idea is to provide a useful theoretical toolbox in a mathematical and statistical aspect, which prepares us to be able to perform the statistical applications in practice. With this intention in mind, we are able to understand the underlying computations in the upcoming simulation study in section 3. Moreover, the elegance of mathematics is difficult to resist, at least for some of us.

2.1 Statistical Learning

This section aims to give an overview of statistical learning and unless otherwise stated we refer to [6]. The key concept of statistical learning can be described as the process of learning from data, which can be considered as a function approximation.

We start by letting $f(x)$ be a function that describes the relationship between inputs and outputs, and whose domain is the p-dimensional Euclidean space denoted by $\mathbb{R}^p$. The aim is to find a useful approximation $\hat{f}(x)$ to the function $f(x)$, that predicts the outcome given a set of observations. Before getting started with the implementation of the learning process, the set of observations is initially split into a *training set* for learning and a *test set* for evaluation. Consider training data $\mathcal{T} = \{x_i, y_i\}$, for $i = 1, ..., N$, with pairs (x_i, y_i) corresponding to points in the $p + 1$-dimensional Euclidean space. From now on, we will refer to inputs as *predictors* with the representation $X = (X_1, ..., X_p)^T$ such that $x_i = (x_{i1}, ..., x_{ip})^T$ where p is the dimension of the predictors, and outputs as *responses* denoted by Y. Further, suppose that the function $Y_i = f(x_i) + \epsilon_i$, where ϵ_i is the error with $E[\epsilon_i] = 0$, is used for the setting of the learning, based on training data $\mathcal{T}$. This leads to a learning algorithm which desirably produces approximations $\hat{f}(x_i)$ that are useful.

In general, statistical learning is often divided into *supervised* and *unsupervised*. The main difference between these two categorizations of statistical learning is the availability of the response. In short, the learning process in supervised learning is guided by the presence of both predictors as well as the matching responses. The ambition is to approximate a function $\hat{Y} = \hat{f}(X)$ with minimal errors $\hat{f}(X) - f(X)$ and a high accuracy in $Y - \hat{Y}$ for prediction. In contrast, a more challenging situation occurs when only the predictors are available. The absence of a supervisor Y means that there are no responses to learn from and this issue goes under the category unsupervised learning. Methods in statistical learning can be either parametric or non-parametric, which depends on whether assumptions are made of the form of the function f or not [7]. Problems that occur in supervised learning are usually categorized into *regression problems* and *classification problems*, depending on the outcome type. Regression refers to responses taking quantitative values in $\mathbb{R}$, while classification have nominal or ordinal responses assuming values in the set of classes $\mathcal{C}$. Regarding supervised learning for classification methods, the intention of the learning process is to build an optimal classifier with perfect prediction accuracy based on training data, in order to obtain an excellent classifier with minimal errors in predictive performance of the test data.

2.1.1 Statistical Decision Theory

As mentioned in section 2.1, the goal in statistical classification is to find a function $f(X)$ for predicting the categorical response variable Y taking values in the set of classes $\mathcal{C}$, given values of the predictor variables X, that minimizes the amount of errors. Therefore, we will provide a statistical framework for building such classifiers, referring to section 2.4 in [6] and section 2.2.3 in [7]. Let L be a $K \times K$ *loss matrix* or also called *loss function*, for $K = \text{card}(\mathcal{C})$, penalizing errors in prediction. It is common to use the *0-1 loss function* with zeros on the diagonal and ones outside the diagonal. Then the cost of classifying observations that belong to class $\mathcal{C}_k$ as $\mathcal{C}_l$ is given by

$$L(k,l) = \begin{cases} 0, & \text{if } k = l \\ 1, & \text{if } k \neq l. \end{cases} \tag{1}$$

Now, also let the estimated response variable $\hat{Y}$ take values in the set of classes $\mathcal{C}$. Using the 0-1 loss function (1), the expected prediction error can be defined as

$$\text{EPE}(\hat{f}) = \text{E}[L(Y, \hat{f}(X))],$$

with respect to the joint probability $Pr(Y, X)$. Further, conditioning on X and the definition of expectation gives

$$\begin{aligned} \text{EPE}(\hat{f}) &= E_X E_{Y|X}[L[Y, \hat{f}(X)]|X = x)] \\ &= E_X \textstyle\sum_{k=1}^{K} L[\mathcal{C}_k, \hat{f}(X)] Pr(\mathcal{C}_k|X). \end{aligned}$$

The expected prediction error is minimized pointwise by

$$\begin{aligned} \hat{f}(X) &= \underset{c \in \mathcal{C}}{\text{argmin}} \textstyle\sum_{k=1}^{K} L[\mathcal{C}_k, c] Pr(\mathcal{C}_k|X = x) \\ &= \underset{c \in \mathcal{C}}{\text{argmin}} [1 - Pr(c|X = x)] \\ &= \underset{c \in \mathcal{C}}{\text{argmax}}\, Pr(c|X = x), \end{aligned} \tag{2}$$

assuming a 0-1 function, where the minimum is obtained for $c \in \mathcal{C}$ for which the probability $Pr(c|X = x)$ is the largest. The result derived from

(2) leads to the optimal *Bayes' classifier*, when all K classes are a priori equally likely. This gives the lowest possible test error rate called *Bayes' rate*. The Bayes' classifier averagely minimizes the test error rate by assigning observations to the most probable class based on the conditional probability of the response, given values of the predictors. That is, given a test set, the predictor x_0 is classified as class $\mathcal{C}_k$ such that the conditional probability $Pr(Y = \mathcal{C}_k|X = x_0)$ is the largest. In binary classification, where $\mathcal{C} = \{0, 1\}$, Bayes' classifier assigns observations to one class if it fulfills $Pr(Y = 1|X = x_0) > 0.5$, otherwise observations are assigned to the other class. The prediction is set by the *Bayes' decision boundary* which constitutes a linear or a non-linear border between the classes. Usually, it is necessary to estimate an approximation of the optimal Bayes' classifier since the conditional probabilities in most real-world situations are unknown.

2.1.2 The Bias-Variance Trade-off

The process of finding a useful approximation $\hat{f}$ of a function f can be considered as an optimization problem, since a minimal expected test error is desirable, while maintaining both low bias and low variance. A statistical learning method with high bias and low variance underfits the data, while low bias and high variance overfits the data. This issue of avoiding underfitting and overfitting data leads us straight to the bias-variance trade-off. This section follows section 2.2.2 in [7], unless otherwise stated.

For the sake of convenience, we use regression to highlight the central parts of the bias-variance trade-off since its setting is rather convincing. The expected test *mean squared error* (MSE) can be decomposed into non-negative terms consisting of the variance of $\hat{f}(x_0)$, the squared bias of $\hat{f}(x_0)$ and the variance of the error ϵ, for some value of x_0 in the test set. Thus the expected test MSE, for predicting $Y_0 = f(x_0) + \epsilon$, can be expressed as

$$\mathrm{E}[(Y_0 - \hat{f}(x_0))^2] = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon). \tag{3}$$

An evaluation of (3) at all possible values of x_0 in the test set enables computation of the average test MSE. It is clear that the expected test MSE is larger or equals the variance of the error, because of the non-negativity of the terms in the decomposition in (3). Hence, it is obvious that a statistical learning method with both low variance and low bias is preferred.

The remaining part of this section focuses on giving a general explanation of the bias-variance trade-off among statistical methods with different properties. Methods with high complexity tend to have lower bias and higher variance compared to less flexible methods. In other words, as the complexity increases, the bias decreases simultaneously as the variance increases

[6]. Methods with high variance can be sensitive to small changes in the function-approximation $\hat{f}$. The relationship between training data and test data with respect to prediction error varies with method complexity. The difference between the estimated training error and the test error increases with higher complexity, although the test data has larger prediction errors than training data irrespective of complexity. To exemplify, assume a remarkably complicated situation where observations appear to be non-linear. If a parametric method with a linear adaption to data is used in this case, the simplicity of the method underfits data. Hence, the inflexible method will have low variance but suffer from high bias. Challenges in model selection arises when trying to find a good trade-off between a small bias and variance of $\hat{f}$. In particular, the awareness of the bias-variance trade-off makes it possible to avoid mistakes that would result in devastating consequences.

2.2 Classification Methods

This section covers the general theory of the classification methods logistic regression, linear discriminant analysis and quadratic discriminant analysis.

2.2.1 Logistic Regression

The theory provided in this section comes from chapter 5, 6 and 8 in [1], unless otherwise stated. *Logistic regression* (LR) is a widely used model in many fields for dealing with categorical response data. Specifically, the model is commonly used in binary classification problems. The main purpose of logistic regression is to study and understand the relationship between inputs and outcomes. In real-world problems, it is often of interest to fit a model to predict the outcome, given a set of observations. The fitted logistic regression model consists of predictor variables as inputs with main effects and possible interaction terms in order to predict the response variable. As the number of interaction terms of higher order increases, the more difficult it is to make a sensible interpretation of the model. For this reason, it is preferable to fit a model that is complex enough to be able to perform predictions with high accuracy, but still easy to interpret. There are many techniques to use when finding a suitable model that fits data well, but a survey of these is out of the scope of this thesis.

2.2.2 Multiple Logistic Regression

Suppose we have a binary response variable Y with two possible outcomes, usually taking values $Y = 1$ and $Y = 0$, where the two events correspond to a "success" and a "failure".

Let $\pi(x) = Pr(Y = 1|X = x) = 1 - Pr(Y = 0|X = x)$ be the probability of a success given values $X = x = (x_1, x_2, ..., x_p)^T$ of p predictor variables. To clarify, the response variable is always qualitative, but the predictor variables can be qualitative as well as quantitative. Then the conditional probability for the multiple logistic regression model is

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)} = \frac{\exp(\beta_0 + \beta^T x)}{1 + \exp(\beta_0 + \beta^T x)},$$

where β_0 is the intercept and $\beta = (\beta_1, \beta_2, ..., \beta_p)^T$ are the coefficient parameters. Equivalently, the *logit* or *log odds* transformation is the alternative formula of a linear representation which is given by

$$\text{logit}[\pi(x)] = \log\frac{\pi(x)}{1 - \pi(x)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p.$$

In other words, multiple logistic regression can be used to predict the conditional probabilities for the response variable given a set of observations. However, this requires that the intercept β_0 and the coefficients β can be estimated by some method, such as *Maximum Likelihood Estimation*. The sensitivity of the log odds of success, with respect to changes in x_j, corresponds to the value of the parameter β_j. In particular, the interpretation of $\exp(\beta_j)$ is reasonably intuitive where a one-unit increase in the corresponding x_j can be considered as a multiplicative effect on the odds of success, when the remaining levels of x_k are held fixed. Now, consider the classification problem of assigning the conditional probabilities of the response variable to either $Y = 1$ or $Y = 0$. Thus, introducing a cutoff c such that

$$Pr(Y = 1|X = x_0) > c,$$

for $0 < c < 1$, enables us to classify new observations x_0 to class 1, otherwise to class 2. One common option is to set the cutoff c to 0.5, since Bayes' classifier assigns observations to one class if $Pr(Y = 1|X = x_0) > 0.5$, as mentioned in section 2.1.1.

2.2.3 Fitting Logistic Regression Models

A necessary step in the model fitting procedure is to estimate the parameters of the logistic regression model. Hence, finding estimators corresponding to the observed data enables us to fit the logistic regression model. In general, the established method called Maximum Likelihood Estimation is used to estimate the value of the parameter by maximizing the probability of the observed values.

Let $x_i = (x_{i1}, x_{i2}, ..., x_{ip})^T$ denote the value of the p predictor variables for setting $i = 1, ..., N$. As a result of using a data set containing quantitative predictors, there is one outcome for each setting x_i. Moreover, let y_i be the response variable for the *ith* setting, where the corresponding random variables Y_i are independent and binomially distributed. For one trial, the likelihood function is given by

$$L(\beta_0, \beta) = \prod_{i=1}^{N} \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i}$$

$$= \left\{ \prod_{i=1}^{N} \exp\left[\log\left(\frac{\pi(x_i)}{1 - \pi(x_i)} \right)^{y_i} \right] \right\} \left\{ \prod_{i=1}^{N} [1 - \pi(x_i)] \right\}$$

$$= \left\{ \exp\left[\sum_{i=1}^{N} y_i \log\left(\frac{\pi(x_i)}{1 - \pi(x_i)} \right) \right] \right\} \left\{ \prod_{i=1}^{N} [1 - \pi(x_i)] \right\}$$

$$= \left\{ \exp\left[\sum_{i=1}^{N} y_i \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \right) \right] \right\} \left\{ \prod_{i=1}^{N} \left[1 + \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \right) \right]^{-1} \right\}.$$

By taking the logarithm of the likelihood function we obtain the *log-likelihood function*

$$\log\, L(\beta_0, \beta) = l(\beta_0, \beta)$$

$$= \sum_{i=1}^{N} y_i \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \right) - \sum_{i=1}^{N} \log\, \left[1 + \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \right) \right].$$

We obtain the likelihood equations by differentiating the log-likelihood function, with respect to β_0 and β respectively, and then letting the partial

derivatives equal zero. A computation of the partial derivatives gives that

$$\begin{aligned}\frac{\partial l(\beta_0,\beta)}{\partial \beta_0} &= \sum_{i=1}^{N} y_i - \sum_{i=1}^{N} \frac{\exp(\beta_0+\sum_{k=1}^{p}\beta_k x_{ik})}{1+\exp(\beta_0+\sum_{k=1}^{p}\beta_k x_{ik})},\\ \frac{\partial l(\beta_0,\beta)}{\partial \beta_j} &= \sum_{i=1}^{N} y_i x_{ij} - \sum_{i=1}^{N} x_{ij}\frac{\exp(\beta_0+\sum_{k=1}^{p}\beta_k x_{ik})}{1+\exp(\beta_0+\sum_{k=1}^{p}\beta_k x_{ik})},\end{aligned}$$

for $j = 1, ..., p$. Thus, the likelihood equations are

$$\begin{aligned}\sum_{i=1}^{N} y_i - \sum_{i=1}^{N} \hat{\pi}_i &= 0,\\ \sum_{i=1}^{N} y_i x_{ij} - \sum_{i=1}^{N} \hat{\pi}_i x_{ij} &= 0,\end{aligned}$$

for $j = 1, ..., p$, and where $\hat{\pi}_i = \exp\big(\hat{\beta}_0 + \sum_{k=1}^{p}\hat{\beta}_k x_{ik}\big)/\big[1 + \exp\big(\hat{\beta}_0 + \sum_{k=1}^{p}\hat{\beta}_k x_{ik}\big)\big]$ is referred to as the maximum likelihood estimate of the conditional probability $\pi(x_i)$. Since the likelihood equations are non-linear, they can with advantage be solved by using the Newton-Raphson algorithm [6]. An application of the iterative algorithm results in the maximum likelihood estimates $\hat{\beta}_0$ and $\hat{\beta}$ of the parameters β_0 and β. In other words, $\hat{\beta}_0$ and $\hat{\beta}$ are estimated such that the likelihood function is maximized. Finally, the estimators are obtained and a response curve is fitted.

2.2.4 Multinomial Logistic Regression

Suppose now that we are interested in a model that can handle more than two quantitative outcomes. A generalization of the logistic regression is the *Multinomial Logistic Regression*, where the restriction of a binary response variable is expanded into a multinomial response variable Y with K classes. In total, there are $\binom{K}{2}$ pairs of categories in the multinomial logistic regression, which describes the log odds for all such pairs. Let $\pi_k(x) = Pr(Y = k|X = x)$ be the probability of the event $Y = k$, for $k = 1, ..., K$, given a setting x of p predictor variables, with the constraint $\sum_k \pi_k(x) = 1$. We choose the last category to be used as a baseline. In that case, the conditional probability of the multinomial logistic regression model is

$$\begin{aligned}\pi_k(x) &= \frac{\exp(\beta_{k0}+\beta_k^T x)}{1+\sum_{l=1}^{K-1}\exp(\beta_{l0}+\beta_l^T x)},\\ \pi_K(x) &= \frac{1}{1+\sum_{l=1}^{K-1}\exp(\beta_{l0}+\beta_l^T x)},\end{aligned}$$

for $k = 1, ..., K-1$. Alternatively, the equivalent and more trivial representation is the logit transformation given by

$$\log\frac{\pi_k(x)}{\pi_K(x)} = \beta_{k0} + \beta_k^T x, \tag{4}$$

for $k = 1, ..., K-1$. The fitting procedure for multinomial logistic regression is similar to the multivariate case, but also more complex. Therefore, we are content with the derivation of the model fitting for the multivariate logistic regression in section 2.2.3.

2.2.5 Linear Discriminant Analysis

Back in 1936, R.A. Fisher [5] came up with the idea of using *linear discriminant analysis* (LDA) to classify observations into two well-defined classes, which was later expanded to classify observations into more than two classes [14]. The main idea is to find a linear combination of a set of observations for an optimal partition into different classes. All theory covered in this section follows chapter 4 in [6], unless otherwise stated. Generally, linear discriminant analysis is based on more assumptions, in contrast to the alternative linear statistical classification method based on logistic regression. Let Y denote the response variable with k classes. One assumption is that the $p > 1$ predictor variables $X \in \mathbb{R}^p$ come from a multivariate normal distribution within each class. Another assumption is that each class should have a common covariance matrix $\mathbf{\Sigma}_k = \mathbf{\Sigma}$, for all k. In other words, the linear discriminant analysis is characterized by the assumption

$$X|Y = k \sim N(\mu_k, \mathbf{\Sigma}), \tag{5}$$

with mean vector μ_k and a positive definite covariance matrix $\mathbf{\Sigma}$ [12], for all k. In consideration of the assumption (5), this yields that each class has the multivariate normal density function

$$f_k(x) = \left(\frac{1}{2\pi}\right)^{p/2} \frac{1}{\sqrt{\det\,\mathbf{\Sigma}}} \exp\left[-\frac{1}{2}(x-\mu_k)^T \mathbf{\Sigma}^{-1}(x-\mu_k)\right].$$

Futhermore, an application of Bayes' theorem gives us

$$Pr(Y = k|X = x) = \frac{Pr(X = x|Y = k)\pi_k}{\sum_{l=1}^K Pr(X = x|Y = l)\pi_l} = \frac{f_k(x)\pi_k}{\sum_{l=1}^K f_l(x)\pi_l}, \tag{6}$$

where $\pi_k = Pr(Y = k)$ with $\sum_{k=1}^{K} \pi_k = 1$. By using Bayes' theorem (6) we are able to compute estimates of $Pr(X = x_0|Y = k)$, for a new observation x_0 of class k. This in turn leads to that these estimates can be used in order to calculate an estimation of $Pr(Y = k|X = x_0)$. First we need to estimate the prior probability, the mean vector and the covariance matrix, given a set of training data where observations come from the multivariate normal distribution. Let N_k denote the number of observations of class k. Then the estimations of these parameters are given by

$$\hat{\pi}_k = \frac{N_k}{N}, \tag{7}$$

$$\hat{\mu}_k = \frac{1}{N_k} \sum_{y_i=k} x_i, \tag{8}$$

$$\hat{\Sigma} = \frac{1}{N-K} \sum_{k=1}^{K} \sum_{y_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T. \tag{9}$$

for $k = 1, ..., K$. Now we can insert the estimations (7), (8) and (9) such that

$$\widehat{Pr}(X = x|Y = k) = \left(\frac{1}{2\pi}\right)^{p/2} \frac{\exp[-\frac{1}{2}(x - \hat{\mu}_k)^T \hat{\mathbf{\Sigma}}^{-1}(x - \hat{\mu}_k)]}{\sqrt{\det\ \hat{\mathbf{\Sigma}}}}. \tag{10}$$

Finally, by using Bayes' theorem (6) and the conditional probability (10) obtained for class k, for given a set of observations, the requested estimation is given by

$$\begin{aligned} \widehat{Pr}(Y = k|X = x) &= \frac{\widehat{Pr}(X=x|Y=k)\hat{\pi}_k}{\sum_{l=1}^{K} \widehat{Pr}(X=x|Y=l)\hat{\pi}_l} \\ &= \frac{\exp[-\frac{1}{2}(x-\hat{\mu}_k)^T \hat{\mathbf{\Sigma}}^{-1}(x-\hat{\mu}_k)]\hat{\pi}_k}{\sum_{l=1}^{K} \exp[-\frac{1}{2}(x-\hat{\mu}_l)^T \hat{\mathbf{\Sigma}}^{-1}(x-\hat{\mu}_l)]\hat{\pi}_l}. \end{aligned} \tag{11}$$

This establishes the fundamental principle of linear discriminant analysis. In summary, the classifier uses the maximized estimation given a new observation x_0 and assigns it to the class k. A *decision boundary* is a $(p-1)$-dimensional hyperplane that separates each pair of classes, which is described as the set of values of x such that the log odds equals zero in the p-dimensional space. In particular, the border is a line in a two-dimensional

space. Further, a comparison between two classes k and l can be made when considering the estimated log ratio

$$\log \frac{\widehat{Pr}(Y=k|X=x)}{\widehat{Pr}(Y=l|X=x)} = \log \frac{\exp[-\frac{1}{2}(x-\hat{\mu}_k)^T\mathbf{\Sigma}^{-1}(x-\hat{\mu}_k)]\hat{\pi}_k}{\exp[-\frac{1}{2}(x-\hat{\mu}_l)^T\mathbf{\Sigma}^{-1}(x-\hat{\mu}_l)]\hat{\pi}_l}, \quad (12)$$

due to (11). Further simplification gives

$$-\tfrac{1}{2}(x-\hat{\mu}_k)^T\hat{\mathbf{\Sigma}}^{-1}(x-\hat{\mu}_k) - (-\tfrac{1}{2}(x-\hat{\mu}_l)^T\hat{\mathbf{\Sigma}}^{-1}(x-\hat{\mu}_l)) + \log\frac{\hat{\pi}_k}{\hat{\pi}_l}$$

$$= -\tfrac{1}{2}(\hat{\mu}_k+\hat{\mu}_l)^T\hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_k-\hat{\mu}_l) + x^T\hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_k-\hat{\mu}_l) + \log\frac{\hat{\pi}_k}{\hat{\pi}_l},$$

and it can be noticed that both the normalization factor and the quadratic parts in the exponent are canceled out. Consequently, when the log odds (12) equals zero, it leads to the fact that

$$x^T\hat{\mathbf{\Sigma}}^{-1}\hat{\mu}_k - \tfrac{1}{2}\hat{\mu}_k^T\hat{\mathbf{\Sigma}}^{-1}\hat{\mu}_k + \log\,\hat{\pi}_k = x^T\hat{\mathbf{\Sigma}}^{-1}\hat{\mu}_l - \tfrac{1}{2}\hat{\mu}_l^T\hat{\mathbf{\Sigma}}^{-1}\hat{\mu}_l + \log\,\hat{\pi}_l. \quad (13)$$

The equation (13) corresponds to the set $\{x : \hat{\delta}_k(x) = \hat{\delta}_l(x)\}$, which describes that the estimated decision boundary between each pair of classes k and l is linear in x. In summary, from the derivation we obtained

$$\delta_k(x) = x^T\mathbf{\Sigma}^{-1}\mu_k - \frac{1}{2}\mu_k^T\mathbf{\Sigma}^{-1}\mu_k + \log\,\pi_k,$$

which is called the *linear discriminant function*. The classifier computes the discriminant values by finding linear combinations of the predictor variables, in order to assign observations to a specific class k for which the linear discriminant value is the largest. Each of the first $K-1$ differences between the linear discriminant functions $\delta_k(x) - \delta_K(x)$ requires $p+1$ parameters. These differences define the linear discriminant classifier for K classes with $(K-1)(p+1)$ parameters.

2.2.6 Quadratic Discriminant Analysis

So far, we have only discussed linear classification methods to separate observations. An extension of linear discriminant analysis is the more flexible,

non-linear method *quadratic discriminant analysis* (QDA) which is preferable to use when the Bayes' decision boundary is quadratic. If the assumption that each class has a common covariance matrix is violated, the quadratic term in x remains in (12). This classifier assigns an observation to the class for which the quadratic discriminant function

$$\delta_k(x) = -\frac{1}{2}\log\,(\det\boldsymbol{\Sigma}_k) - \frac{1}{2}(x-\mu_k)^T\boldsymbol{\Sigma}_k^{-1}(x-\mu_k) + \log\,\pi_k$$

is the largest. The quadratic equation $\{x : \delta_k(x) = \delta_l(x)\}$ describes the decision boundary between each pair of classes k and l. To clarify, the quadratic discriminant analysis is based on the assumption

$$X|Y = k \sim N(\mu_k, \boldsymbol{\Sigma}_k),$$

with mean vector μ_k and covariance matrix $\boldsymbol{\Sigma}_k$ for class k. That is, quadratic discriminant analysis allows a unique covariance matrix for each class. For quadratic linear discriminant analysis there are $(K-1)(p(p+3)/2+1)$ parameters.

2.2.7 Relationship Between Logistic Regression and Linear Discriminant Analysis

Previous sections 2.2.1 and 2.2.5 cover the most central parts of the theory of the two statistical classification methods logistic regression and linear discriminant analysis. Both classifiers are used in order to find the relationship between a set of quantitative predictor variables and a qualitative response variable. Remember that logistic regression does not make any assumption about the predictor variables, only on the distribution of Y given x [1]. Unlike logistic regression, linear discriminant analysis also makes an assumption of X given y. In one sense, these methods are closely related, leading us to the interesting question of when one method is more preferable in a particular situation over the other method. At this point, it is obvious that logistic regression and linear discriminant analysis perform overall similar results in most cases [6]. As a reminder, the log odds of the logistic regression in the multinomial case (4) is given by

$$\log\,\frac{\pi_k(x)}{\pi_K(x)} = \log\,\frac{Pr(Y=k|X=x)}{Pr(Y=K|X=x)} = \beta_{k0} + \beta_k x, \tag{14}$$

and the log odds for linear discriminant analysis (12) can be written on the form

$$\begin{aligned}
\log \frac{\pi_k(x)}{\pi_K(x)} &= \log \frac{Pr(Y=k|X=x)}{Pr(Y=K|X=x)} \\
&= -\tfrac{1}{2}(\mu_k + \mu_K)^T \mathbf{\Sigma}^{-1}(\mu_k - \mu_K) + x^T \mathbf{\Sigma}^{-1}(\mu_k - \mu_K) + \log \frac{\pi_k}{\pi_K} \\
&= \alpha_{k0} + \alpha_k x,
\end{aligned} \tag{15}$$

where α_{k0} and α_k are functions of π_k, π_K, μ_k , μ_K and $\mathbf{\Sigma}$, for $k = 1, ..., K-1$. Both (14) and (15) are linear functions of x and have linear decision boundaries for $p > 1$. As mentioned before, the parameters β_{k0} and β_k in logistic regression can be estimated by using maximum likelihood estimation. On the other hand, linear discriminant analysis makes assumptions of a multivariate normal distribution (5) and uses the estimated mean and variance to get an estimation of the parameters α_{k0} and α_k. Also, each class needs to have a common covariance matrix. It can be realized that the difference between these linear classifiers is the method of parameter estimation.

In general, logistic regression is often assumed to be more flexible and robust, since it makes no assumptions on distribution of the predictor variables and handles outliers better, whereas linear discriminant analysis makes more assumptions of the underlying data and is strongly affected by outliers. For a real-life data set, it is rather unlikely that the conditions in linear discriminant analysis are satisfied, thus logistic regression is considered to be preferable in those situations. Despite this, linear discriminant analysis usually has higher prediction accuracy when all the assumptions are approximately satisfied. As long as data approximately comes from a multivariate normal distribution, with different means and a common covariance matrix, then a good choice of method is linear discriminant analysis. It is also suitable when Bayes' decision boundary is linear and it is also a good choice for small sample sizes, when few parameters need to be estimated. It is argued that linear discriminant analysis is the right choice, when it comes to nominal response variables [7].

2.3 Evaluation Metrics

In order to do a decent comparison of different classification methods in the simulations in section 3, some kind of metric is required for assessing the predictive power. Since different metrics have their advantages and disadvantages, we will use two types of distinct measures to evaluate the performance of the classifiers. Also, a metric for measuring distances is presented.

2.3.1 Misclassification Rate

The following section follows section 2.2.3 in [7]. The *misclassification rate* or *error rate* is an intuitive measure to assess the prediction accuracy of a classifier, where the proportion of incorrect observations is computed.

Given a training set, let y_i be the response of the actual class and let $\hat{y}_i$ be the predicted class, for observation $i = 1, ..., N$. Then *training error rate* is given by

$$\frac{1}{N}\sum_{i=1}^{N} I(y_i \neq \hat{y}_i),$$

where $I(y_i \neq \hat{y}_i)$ denotes an *indicator variable* that equals 1 if $y_i \neq \hat{y}_i$, meaning that the *ith* observation is misclassified by the classifier. Otherwise, the indicator variable equals 0 for a correct classification. This corresponds to the average of the loss function (1) in section 2.1.1 for all observations. The main interest is to calculate the *test error rate* since the observations in the test set, as explained in section 2.1, were not involved in the learning process of the classifier. A classifier is considered to perform well when a low error rate is obtained.

2.3.2 AUC

Another option to evaluate the predictive power of classifiers is to use ROC analysis [7]. In binary classification problems, the ROC curve, *receiver operating characteristics*, is widely used and graphically displays the relationship between the *true positive rate* (TPR) and the *false positive rate* (FPR) of all possible cut-off points or *thresholds* in two-dimensional space. This section continues to follow [4].

Table 1: *The four possible outcomes TP, FN, TN and FP are presented in a confusion matrix, which refers to the number of observations in a data set that ends up in each category.*

		Predicted class	
		Positive	Negative
Actual class	Positive	TP	FN
	Negative	FP	TN

Consider the actual response of an observation labeled as either positive or negative. Suppose that the actual outcome is regarded as positive. In this case, if the classifier predicts the response to be positive, it is called *true*

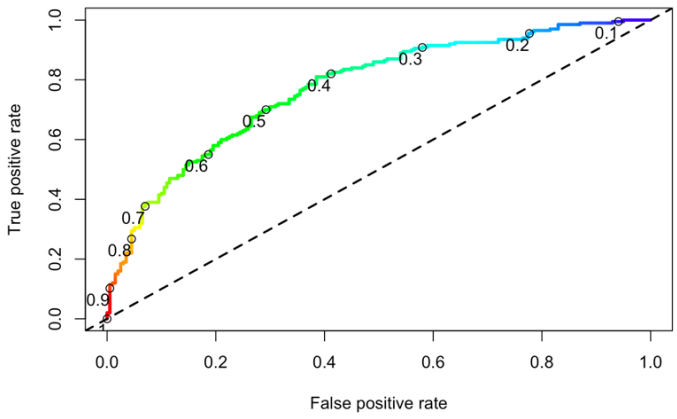


Figure 1: *An illustration of a graph in the two-dimensional space showing the corresponding ROC curve for all possible thresholds. This ROC curve gives an AUC-value of 0.77.*

positive (TP). However, if the response instead is incorrectly classified as negative, then it is called *false negative* (FN). In contrast, suppose now that the response is negative, which is also predicted by the classifier. Unsurprisingly, this situation is termed as *true negative* (TN), otherwise *false positive* (FP). In summary, there are four possible outcomes in binary classification that can be presented in a *confusion matrix*, see Table 1. There are many computations that can be made to obtain different performance metrics. The definition of true positive rate and false positive rate is

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}},$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}},$$

which also can be recognized as *sensitivity* and *1-specificity* respectively. The true positive rate and false positive rate can be changed by adjusting the threshold settings. This trade-off between these two rates is made clear when plotting the false positive rate on the x-axis against the true positive rate on the y-axis in two-dimensional space, which yields the ROC curve with terminal points $(0, 0)$ and $(1, 1)$, for all possible thresholds, see Figure 1. The point $(0, 0)$ corresponds to that none of the observations are classified as positive, so that no false negative errors as well as no true positives are observed. On the contrary, the point $(1, 1)$ is representing that all of the observations will be classified as positive, meaning that no false negative or

true negative results are obtained. A perfect classification corresponds to the point $(0, 1)$. One measure of the performance of a classifier is the *area under the ROC curve* (AUC), representing the proportion of area of the unit square. Consequently, the AUC takes values between 0 and 1, where the optimal AUC-value of 1 indicates a perfect prediction accuracy while random guessing corresponds to a ROC curve along the dashed line in 1, and hence an AUC-value of 0.5.

2.3.3 Mahalanobis Distance

The distance between two mean vectors of two classes; class k and class l, in the p-dimensional Euclidean space can be measured using the metric Mahalanobis distance [9]. The definition of the squared Mahalanobis distance is

$$\Delta^2 = (\mu_k - \mu_l)^T \mathbf{\Sigma}^{-1} (\mu_k - \mu_l),$$

where each class has a mean vector μ_k and μ_l respectively, with a common positive definite covariance matrix $\mathbf{\Sigma}$ for both classes.

3 Simulation and Modeling

In the following sections we will describe the simulation process for the statistical classification methods logistic regression and linear discriminant analysis. Quadratic discriminant analysis will also be considered in one situation. The purpose is to study the adaptability and the performance in prediction accuracy of these classifiers on a variety of simulated data sets. The simulation study is divided into four experiments, in order to distinguish these variations, which will facilitate the upcoming discussions of the results later in section 5. Each experiment consists of different simulated data sets, based on different assumptions, where some parameters are adjusted while others are held fixed.

Throughout this thesis, the statistical software package R is used for all simulations, computational issues, and graphical views. We use the implementations in R of logistic regression, linear discriminant analysis and quadratic discriminant analysis for the fitting procedure. The *glm()* function for logistic regression is included in the base package, whereas the *lda()* and *qda()* function for linear discriminant analysis and quadratic discriminant analysis respectively, are provided by the *MASS* package. Data sets generated from a multivariate normal distribution are also included in the *MASS* package.

3.1 Simulation Setup

The simulation study consisting of Experiments A-C have the following default setting: First, random samples $x = (x_1, x_2, ..., x_p)^T$ of p predictor variables of size N and M are generated from two multivariate normal distribution groups with density function

$$f_k(x) = \left(\frac{1}{2\pi}\right)^{p/2} \frac{1}{\sqrt{\det\ \mathbf{\Sigma}_k}} \exp\left[-\frac{1}{2}(x-\mu_k)^T \mathbf{\Sigma}_k^{-1}(x-\mu_k)\right],$$

having mean vector μ_k and covariance matrix $\mathbf{\Sigma}_k > 0$ for class $k = \{1, 2\}$. For all $p > 1$, the mean vector of class 1 is consistently set to zero, such that $\mu_1 = (0, ..., 0)$, while the mean vector of class 2 is set in the direction of the first principal component of the covariance matrix along which the variance is the greatest. The angle of the direction of μ_2 in relation to μ_1 is determined by the spectral decomposition of the covariance matrix

$$\mathbf{\Sigma} = QDQ^T,$$

where Q is a $p \times p$ matrix with columns containing the unit eigenvectors and D is a diagonal matrix with p eigenvalues [8]. Consequently, the direction of the first principal component is obtained by calculating the unit eigenvector with the highest corresponding eigenvalue. That unit eigenvector is scaled such that the distance between μ_1 and μ_2 for classes with a common covariance matrix has a certain value measured in the Mahalanobis distance.

A symmetric matrix with equal elements on the diagonal, or more specifically, an *autocovariance matrix* or an *autoregressive covariance matrix* (AR) are constructed according to

$$\mathbf{\Sigma}_k = \sigma_k^2 \begin{bmatrix} 1 & \rho_k & \rho_k^2 & \dots & \rho_k^{p-1} \\ \rho_k & 1 & \rho_k & \dots & \rho_k^{p-2} \\ \rho_k^2 & \rho_k & 1 & \dots & \rho_k^{p-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_k^{p-1} & \rho_k^{p-2} & \rho_k^{p-3} & \dots & 1 \end{bmatrix} = \sigma_k^2 P_k \tag{16}$$

with variance σ_k^2 and correlation $-1 < \rho_k < 1$, for class $k = \{1, 2\}$ with p predictor variables [3]. The autocovariance matrix (16) is a positive definite Toeplitz matrix [10] which satisfies the assumption (5) of a non-singular multivariate normal distribution [12]. We let the variance σ_k^2 equal 1, such that the autocovariance matrix and the *autocorrelation matrix* P_k are identical. Then (16) measures the correlation between pairs of predictor variables and the closer the elements are to the diagonal, the higher correlation between

the pairs. These correlations vary exponentially with the distance to the diagonal. Thereafter, the observations generated from the first group of size N are assigned to class 1, whereas the observations from the second group of size M are assigned to class 2, which will form a data set. The responses are balanced when the number of observations generated from the two groups are equal, that is to say if $N = M$, and the cutoff $c = 0.5$ is being used for logistic regression. Then, the total number of observations equals $2N$. The next step is to randomly split the data set into a training set and a test set. The splitting consists of 50 percent training set which is intended for the learning process for the classification methods of interest. The remaining 50 percent test set of size N is for evaluating the predictive power in terms of misclassification rate and AUC-value, both provided in section 2.3, for each classifier. This simulation procedure will be replicated 50 times. The average values of the evaluation metrics, as well as the standard deviation, are computed and the final results are presented for each classification method.

3.2 Simulation Experiments

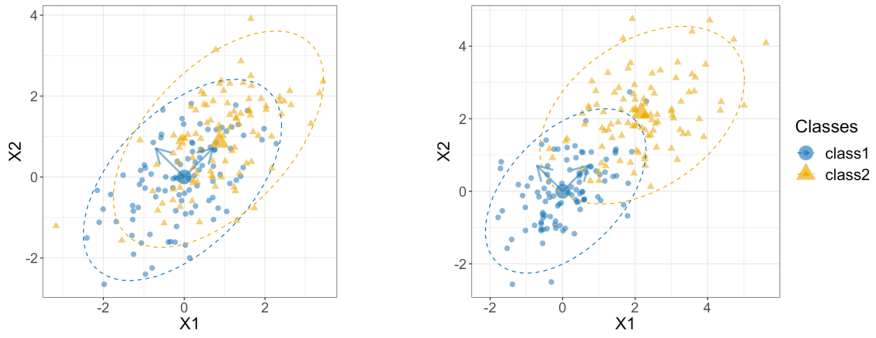
This section provides the different setups for each of the simulation experiments in more detail.

3.2.1 Experiment A - Separability between Classes

In the first experiment, the separation of two classes is going to be investigated when adjusting the Mahalanobis distance between the two mean vectors in the direction of the first principal component of the covariance matrix. The main idea of the experiment can intuitively be illustrated in the two-dimensional space as in Figures 2a and 2b. In order to evaluate the performance of separability of logistic regression and linear discriminant analysis, they are both examined on data from two multivariate normal distribution groups with different mean vectors but with common covariance matrix according to the AR-process (16) with correlation $\rho = 0.5$. In this context, the assumptions of linear discriminant analysis are met. In this experiment we fix the number of observations N to 500 and let the Mahalanobis distance take the values $\Delta = \{0.5, 1.0, 1.5, 2.0, 2.5, 3.0\}$, both with two and twenty predictor variables.

3.2.2 Experiment B - Number of Observations and Predictor Variables

The focus in the second experiment lies on varying the number of observations and predictor variables. All the assumptions remain from the setting



(a) $\Delta = 1$ (b) $\Delta = 2.5$

Figure 2: *Two classes, each of size $N = 100$, generated from a normal distribution with different mean vectors in the direction of the first principal component, and common covariance matrix with variances 1 and correlation 0.5. The blue arrows show the unit eigenvectors which represent the two principal components in the two-dimensional space. The Bayes' decision boundary is linear.*

of Experiment A and follows the default setting in section 3.1. The number of observations N varies within a range of 20 to 3000 in two dimensions and the two mean vectors differ along the direction of the first principal component. Then, the number of observations N are fixed to 500 while changing the dimensionality from 2 up to 100. The Mahalanobis distance is constantly set to 1.0, using the common covariance matrix with correlation $\rho = 0.5$ according to the AR-process (16).

3.2.3 Experiment C - Classes with Different Covariance Matrices

In this experiment we adjust the covariance matrix for one class according to an AR-process (16), while the covariance matrix for the other class remain with correlation $\rho = 0$, which corresponds to the identity matrix. Because of this setup, we let the mean vector μ_1 be the zero vector with the identity covariance matrix $\mathbf{\Sigma_1}$. The mean vector μ_2 for class 2 is set according to the same procedure as in previous experiments, but its location is based on the associated autocovariance matrix $\mathbf{\Sigma_2}$ in (16), where the correlation $\rho = \{0, 0.2, 0.3, 0.4, 0.6, 0.8\}$ is modified stepwise. We fix μ_2 as the scaled unit vector in the direction of the first principal component when the Mahalanobis distance equals 1.0 and its evaluation is based on the case when the covariance is the identity matrix. That is to say, we set $\Delta = 1.0$ based on the covariance matrix ($\rho = 0$) of class 1, regardless of the covariance matrix of class 2. This simulation will be evaluated for 20 predictor variables. As a result of the violation of the assumption of equal covariance matrices, the

Bayes' decision boundary becomes quadratic and we will study the change in adaptation for logistic regression and linear discriminant analysis, but also consider the method quadratic discriminant analysis.

3.2.4 Experiment D - The Effect of Non-Normality

The final experiment aims to study the impact of non-normality on the predictive power for logistic regression and linear discriminant analysis. In other words, the adaptability of these methods will be examined on data sets, for which the multivariate normal distribution is replaced by a more heavy-tailed one, to study how the methods are affected by outliers. For this purpose, we are going to generate spherically symmetric random samples, and then modify the distribution of the radius. Consequently, spherically symmetric observations for both classes according to the same symmetric distribution are generated around their mean vectors. Its stochastic representation follows the report [2], unless otherwise stated. Regardless of the distribution of the p-dimensional random vector Z, it can be expressed as

$$Z = R \cdot U = R \cdot \frac{Y}{||Y||},$$

with radius $R = ||Z|| > 0$ of Z. The radius is independent of the direction represented by $U = Y/||Y||$ which is uniformly distributed $\mathcal{U}_{\mathcal{S}_1}$ on the p-dimensional sphere $\mathcal{S}_1$ of unit radius in the Euclidean space, after normalization of Y. In this experiment, we let Y be a standard normal random vector. If Z is also a spherical standard normal random vector, the squared radius R^2 follows a Chi-squared distribution with p degrees of freedom, that is $R^2 = ||Z||^2 \sim \chi^2(p)$, for $Z \sim N_p(0, I_p)$.

For this setup, random samples are generated from a distribution Z with mean vector $\mu_1 = (0, ..., 0)$ for class 1, while the distribution for class 2 is offset according to the mean vector μ_2, that is $\mu_2 + Z$, and is scaled in an arbitrary direction such that the Mahalanobis distance (with an identity matrix in its definition) equals 1.0. Here it is utilized that a spherically symmetric random vector X around its mean vector μ can be decomposed as $X = \mu + Z = \mu + R \cdot U$, where $R = ||X - \mu||$ and $U = (X - \mu)/||X - \mu|| \sim \mathcal{U}_{\mathcal{S}_1}$. Figure 3 illustrates the direction with a uniform distribution on the three-dimensional sphere of unit radius.

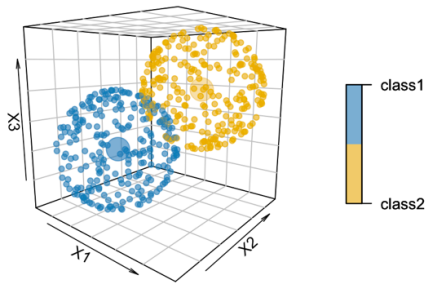


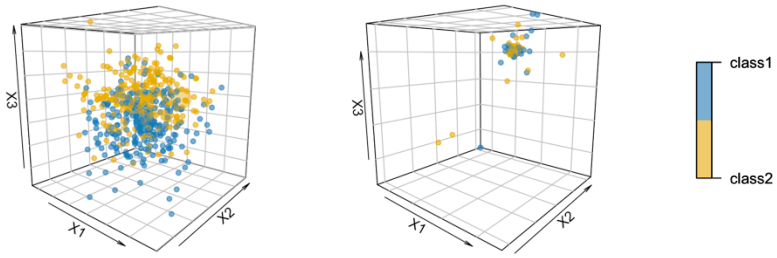
Figure 3: *An illustration of uniformly distributed observations on the three-dimensional sphere of unit distance to their mean vectors, for two classes around their respective mean vector. This corresponds to a setting where R has a one point distribution at 1.*

The *inverse transformation method* will be used to simulate different continuous random variables, in order to model the radius R [11]. First, the random variable V is generated from a standard uniform distribution. Then V is passed through the inverse cumulative distribution function F_R^{-1} to obtain the radius R with the continuous cumulative distribution function F_R. That is, for any continuous function F_R that satisfies

$$R = F_R^{-1}(V)$$

for $V \sim U(0, 1)$, then the radius R has the continuous cumulative distribution function F_R. We introduce the limitation of using the one-tailed Cauchy distribution for the radius, where the location parameter is set to zero such that the two-tailed Cauchy probability density function is symmetric around zero, while its scale parameter is set to 0.5, 1.0 and 2.0.

First, we let the radius follow the square root of a Chi-squared distribution with same degrees of freedom as the number of dimensions, which corresponds to that the observations come from a standard normal distribution. Thereafter, the distribution of the radius is interchanged to the Cauchy distribution which will generate extreme values, which is the main focus in this experiment. The number of observations N is set to 500 for both three and ten predictor variables. Figures 4a and 4b visualize the effect of changing the distribution of the radius in the three-dimensional space.



(a) Square root of Chi-square (b) Cauchy

Figure 4: *An illustration of two different distributions of the radius for observations that are spherically symmetric around their centers, for two classes with 500 observations respectively.*

4 Results

This section presents the results from Experiment A-D obtained from the simulations in section 3. The results are described and shown in boxplots in Figures 5-14. The horizontal line inside the box shows the median. A more detailed presentation of the results can be found in Tables 2-14 in the Appendix. Overall, the results show that the prediction accuracy of the linear classifiers were similar, despite the fact that a variety of simulated data sets were used. Some notable differences in performance were observed.

4.1 Experiment A - Separability between Classes

There are no clear differences in separability between the classes in the two-dimensional space and the results are more or less identical. Figures 5 and 6 show that linear discriminant analysis does not separate classes better than logistic regression or vice versa, regardless of distance between the two classes. When increasing the dimensionality to twenty, the prediction follows the same pattern but with a slightly worse accuracy, see Tables 4 and 5 in the Appendix. Same interpretation can be made out of the two metrics, but one can observe that the misclassification rate decreases more linearly, while the AUC-value increases in a more nonlinear way, as the Mahalanobis distance increases.

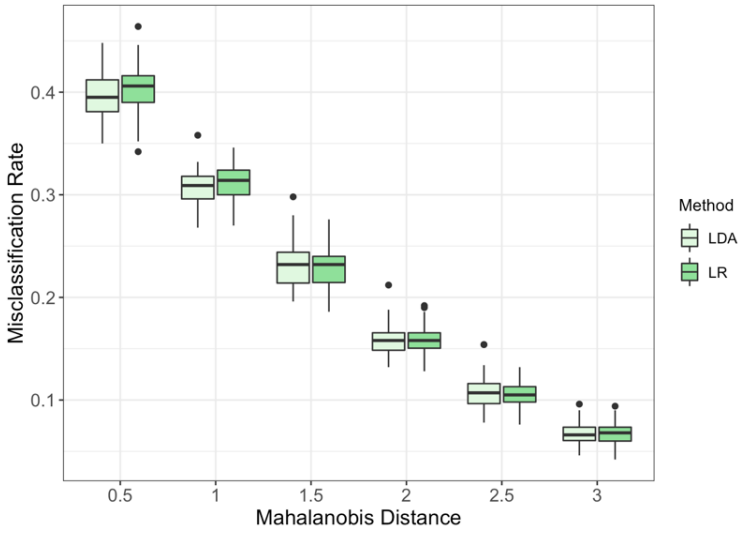


Figure 5: *Experiment A. Separability of classes in two dimensions. Metric: MR. Parameters:* $N = 500$, $p = 2$, $\rho = 0.5$.

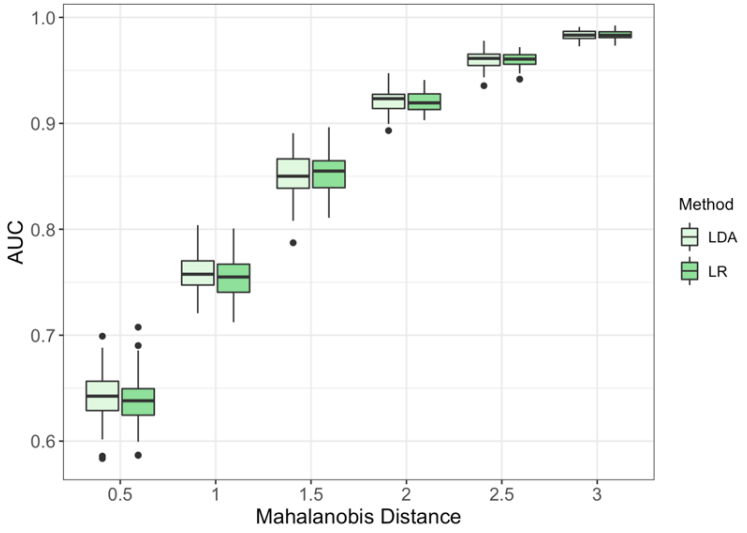


Figure 6: *Experiment A. Separability of classes in two dimensions. Metric: AUC. Parameters:* $N = 500$, $p = 2$, $\rho = 0.5$.

4.2 Experiment B - Number of Observations and Predictor Variables

The results of varying the number of observations in two dimensions can be viewed in Figures 7 and 8 which indicate that both logistic regression and linear discriminant analysis perform similarly. Noteworthy is that smaller sample sizes lead to higher uncertainty and the spread is much wider, but the pattern is quite clear; the prediction accuracy stabilizes when the number of observations increases. Further, it can be seen in Figures 9 and 10 that both logistic regression and linear discriminant analysis perform similar and get worse in prediction accuracy as the number of predictor variables increases, for a fixed number of observations. Consequently, the methods perform worse as the dimension of the data increases.

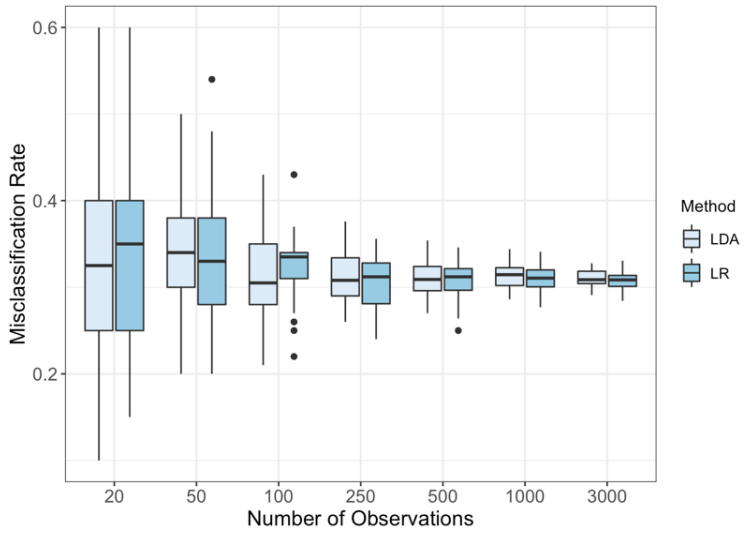


Figure 7: *Experiment B. Number of observations N in two dimensions. Metric: MR. Parameters: Mahalanobis distance* $\Delta = 1$, $p = 2$, $\rho = 0.5$.

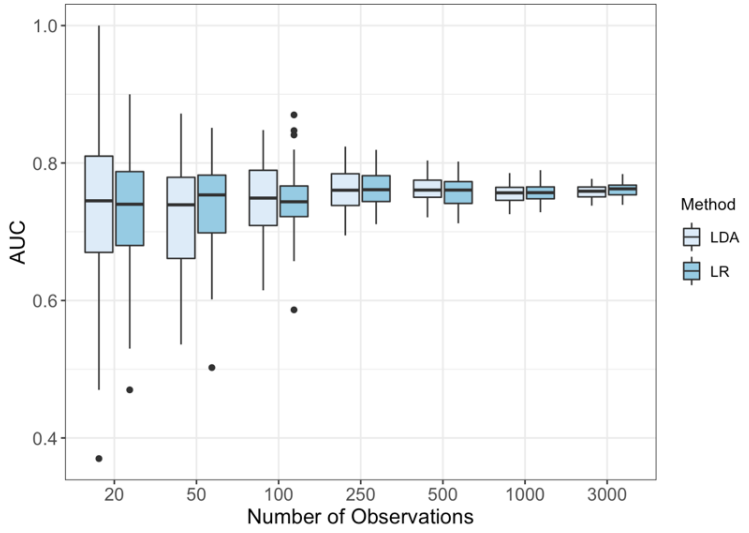


Figure 8: *Experiment B. Number of observations N in the two-dimensional space. Metric: AUC. Parameters: Mahalanobis distance $\Delta = 1$, $p = 2$, $\rho = 0.5$.*

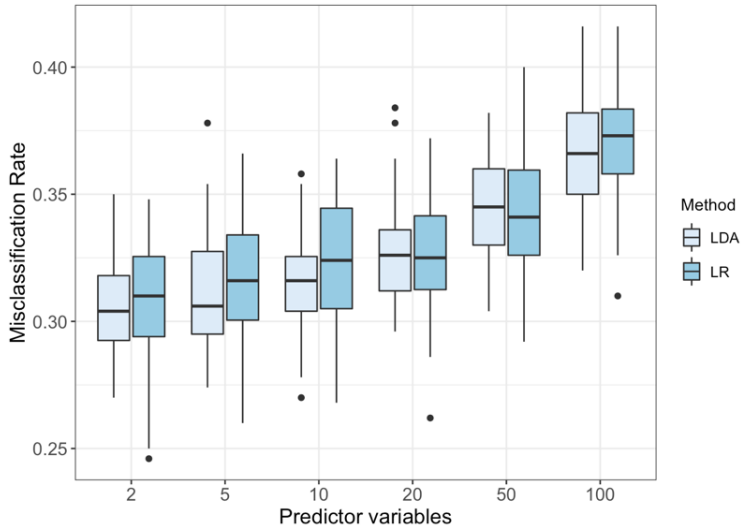


Figure 9: *Experiment B. Number of predictors. Metric: MR. Parameters: Mahalanobis distance $\Delta = 1$, $N = 500$, $\rho = 0.5$.*

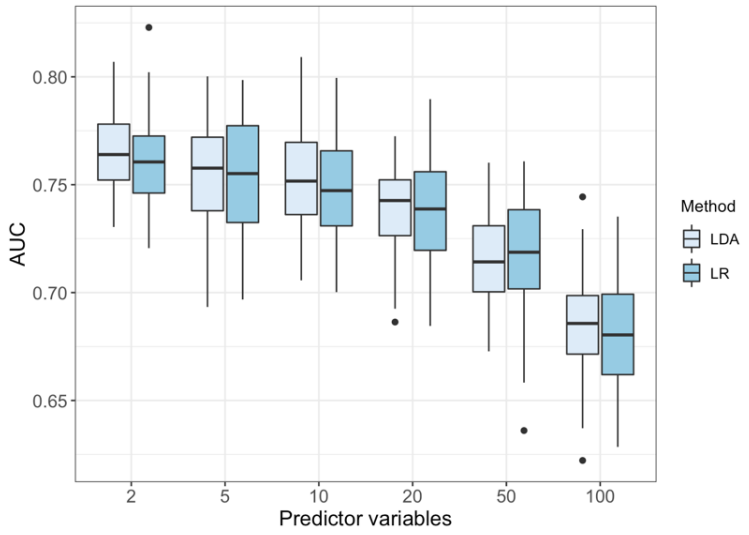


Figure 10: *Experiment B. Number of predictors. Metric: MR. Parameters: Mahalanobis distance* $\Delta = 1$, $N = 500$, $\rho = 0.5$.

4.3 Experiment C - Classes with Different Covariance Matrices

The consequences of letting the two classes have different covariance matrices are illustrated in Figures 11 and 12. The results show that quadratic discriminant analysis improves its predictive power as the correlation between the predictor variables of the second population increases, in contrast to logistic regression and linear discriminant analysis, which instead appear to be relatively unchanged. It can be observed that the linear classification methods perform slightly better in prediction for the highest tested correlation $\rho = 0.8$ compared to the case where the covariance matrices are approximately equal. Again, the two linear classifiers tend to perform similarly for the tested correlations, but quadratic discriminant analysis outperforms them both in prediction accuracy when the adjusted correlation ρ for one class is equal to or greater than 0.3.

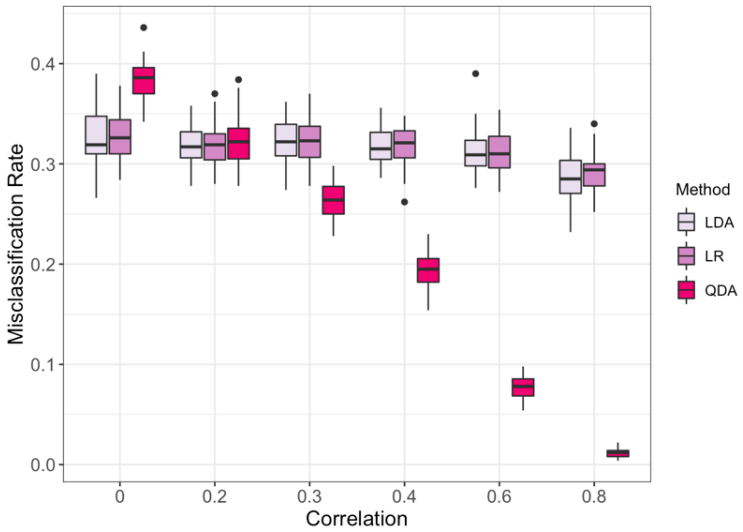


Figure 11: *Experiment C. Classes with different covariance matrices. Metric: MR. Parameters:* $N = 500$, $p = 20$. *The correlation* ρ *is adjusted along the x-axis of the graph for one class, while the other class is fixed at* $\rho = 0$.

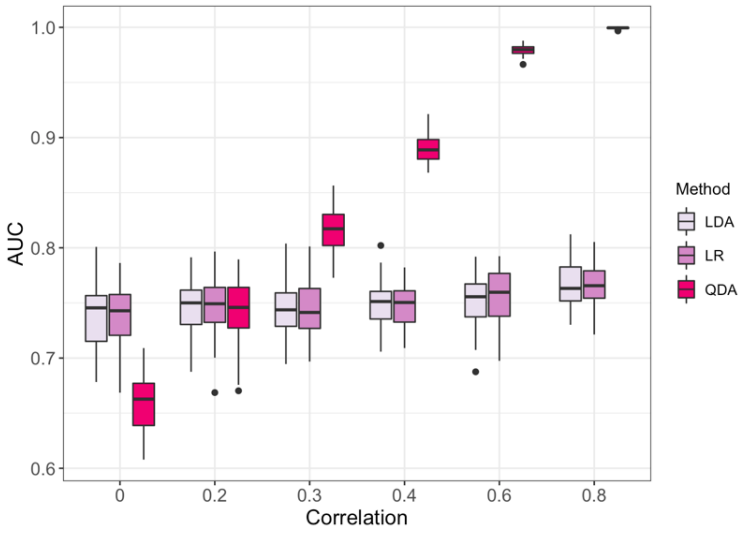


Figure 12: *Experiment C. Classes with different covariance matrices. Metric: AUC. Parameters:* $N = 500$, $p = 20$. *The correlation* ρ *is adjusted along the x-axis of the graph for one class, while the other class is fixed at* $\rho = 0$.

4.4 Experiment D - The Effect of Non-Normality

The adaptability of logistic regression and linear discriminant analysis is examined for spherically symmetric random samples where the distribution of the radius is modified as heavy-tailed. When the radius alternates to follow a one-tailed Cauchy distribution with different settings in ten dimensions, it can be observed in Figures 13 and 14 that the prediction accuracy decreases when increasing the value of the scale parameter. For linear discriminant analysis, the spread becomes wider as the value of the scale parameter increases, in contrast to logistic regression which displays another pattern. When the scale parameter is set to 0.5, the dots excluded from the whiskers confirm that the use of logistic regression results in a low prediction accuracy slightly better than guessing for some of the simulations. The pattern of the spread among logistic regression and linear discriminant analysis can also be confirmed, in terms of standard deviation, in Tables 13 and 14 in the Appendix. There are no other clear differences in performance between the linear classifiers when the results of each modification of the radius are considered separately in ten dimensions.

The same pattern occurs in the three-dimensional space, but we can notice that logistic regression has a somewhat lower average of the misclassification rate and a higher average of the AUC-value, particularly in the case where the radius follows a standard Cauchy distribution, see Tables 13 and 14. From the same tables, the standard deviation is noticeably higher in three dimensions than in ten dimensions. Similar results as in previous experiments are achieved when the radius follows the square root of a Chi-squared distribution with the number of dimensions set as degrees of freedom, which can be found in Table 12 in the Appendix.

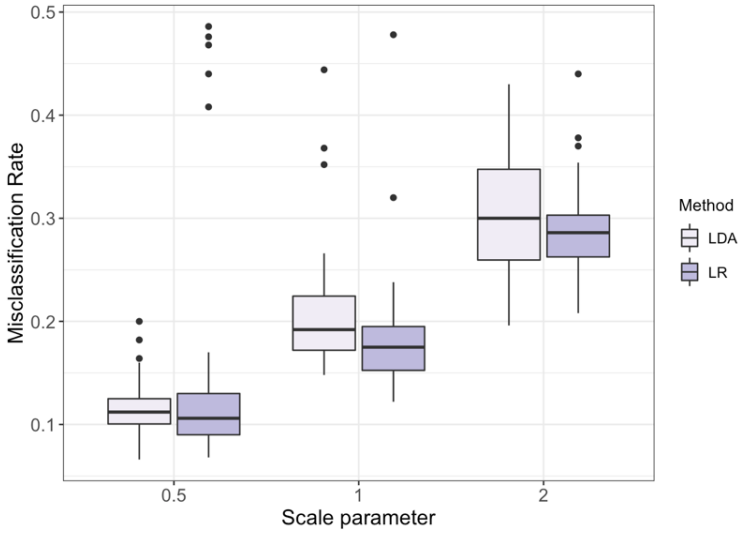


Figure 13: *Experiment D. The radius follows a one-tailed Cauchy distribution with the specified scale parameter. Metric: MR. Parameters: Mahalanobis distance* $\Delta = 1$, $N = 500$, $p = 10$.

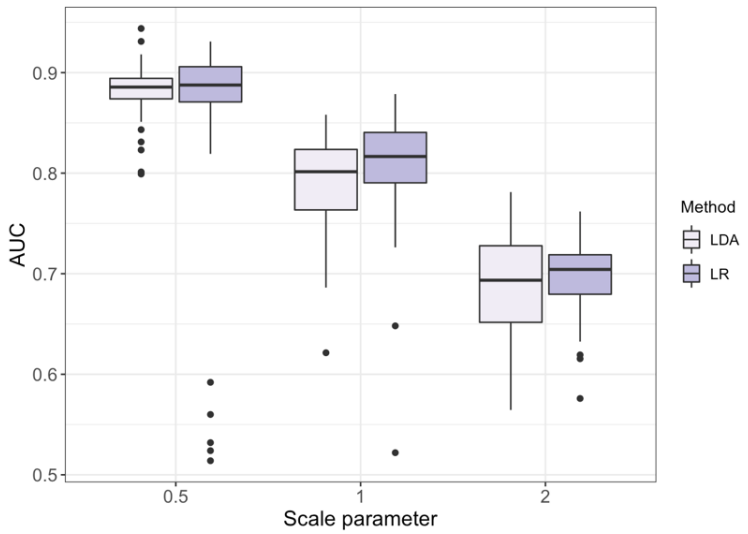


Figure 14: *Experiment D. The radius follows a one-tailed Cauchy distribution with the specified scale parameter. Metric: AUC. Parameters: Mahalanobis distance* $\Delta = 1$, $N = 500$, $p = 10$.

5 Discussion

The classification methods logistic regression and linear discriminant analysis, as well as quadratic discriminant analysis, are statistically evaluated based on the simulation plan of section 3, which contains the practical simulations with four experiments with different setups. In the following section we are going to do the final comparison between logistic regression and linear discriminant analysis based on the results presented in section 4. It turned out that the linear classification methods performed similarly in most cases. However, some notable differences in prediction accuracy were observed, which will be highlighted and possible explanations of the behaviour in each experiment will be discussed, with the underlying theory in mind. Improvements of the study will also be suggested. Lastly, these linear classification methods will be put in a broader perspective.

5.1 Predictive Power

The results of Experiment A-D in section 3.2.1 - 3.2.4 will be discussed and linked to the theory, with respect to the predictive power. To clarify, all discussions concerning the most suitable method in these experiments are based on that particular setup.

5.1.1 Experiment A - Separability between Classes

The Mahalanobis distance between the two mean vectors, which are related in the direction of the first principal component of the common covariance matrix, is varied for both two and twenty normally distributed predictor variables. Data with this setup implies that the Bayes' decision boundary divides observations linearly. Therefore, a linear classification method is probably a wise choice. This setting can be viewed as tailor-made for linear discriminant analysis since it fulfills all assumptions discussed in the theory part of section 2.2.5. For that reason, linear discriminant analysis could be expected to perform better compared to logistic regression, but this does not seem to be consistent with the results. Figures 5 and 6 show that the task of classifying observations correctly is more difficult when the two populations are close to each other, but both methods follow the same pattern and perform similarly. With reference to the theoretical relationship between logistic regression and linear discriminant analysis in section 2.2.7, both classifiers have linear decision boundaries but differ in their estimation of the parameters. This means that the methods' different parameter estimates contribute to hyperplanes that separate data with different slopes. However,

the difference is not large enough to be able to distinguish the prediction accuracy of the methods from each other.

5.1.2 Experiment B - Number of Observations and Predictor Variables

When exploring the number of observations in the two-dimensional space with the same default setting as in Experiment A, the results show the high uncertainty with few observations and that the prediction accuracy stabilizes when the number of observations increases. The reason for the convergence can be explained by the fact that normally distributed random samples approximate the distributions they are drawn from better, as the number of observations increases. In section 2.2.7, linear discriminant analysis is said to be a good choice when the sample sizes are small. However, the results do not show any remarkable indication of that linear discriminant analysis tends to work better than logistic regression for small sample sizes. On the contrary, when the number of observations are fixed and the number of predictor variables increases, both methods perform worse and the reason can be explained by overfitting. Apparently, logistic regression and linear discrimination analysis lack the ability to maintain the predictive power as the number of predictors increases. A brief comment with a suggestion of improvement for logistic regression is to use shrinking methods, for instance ridge regression or the lasso, by introducing penalty terms [6], which can prevent overfitting and improve the performance in prediction accuracy of new unseen data. Corresponding regularization techniques exist for linear discriminant analysis as well. Nowadays, it is usual to encounter high-dimensional data in real-world problems, such that the number of predictor variables is much bigger than the number of observations, that is $p >> N$. Speaking of high-dimensional data, an interesting expansion of this study could be to consider regularization techniques for solving high-dimensional problems, and investigate the distinctions in performance between the linear classifiers in such cases.

5.1.3 Experiment C - Classes with Different Covariance Matrices

This experiment investigates the adaptability for logistic regression, linear discriminant analysis and quadratic discriminant analysis on data containing two classes where the covariance matrix for one class is adjusted while holding the other fixed as the identity matrix. This construction leads to that the Bayes' decision boundary is quadratic. Consequently, classification methods with linear adaption to data face a risk of underfitting. This establishes the fact that when the decision boundary is non-linear, other methods, for in-

stance quadratic discriminant analysis in this case, should perform better in terms of predictive power. This situation can be confirmed by the results viewed in Figures 11 and 12. Nevertheless, linear discriminant analysis still performs remarkably well and does not seems to be disturbed by unequal covariance matrices, even though the assumption is not satisfied. It can be seen as an indication of the robustness of the linear classifiers, despite that they might suffer from high bias. This suggests that there may be situations where linear classifiers still are considered to be useful. Further, quadratic discriminant analysis seems to be preferable when the adjusted correlation ρ for one class is equal to or greater than 0.3. A low correlation will rapidly converge to zero since the correlation between the pairs changes exponentially according to an AR-process (16). This means that the correlations close to the diagonal are low and the other correlations can be considered negligible. Thus, the assumptions are approximately fulfilled, which is one possible reason of why the effect is relatively unnoticed with a correlation less than 0.3.

5.1.4 Experiment D - The Effect of Non-Normality

The approach of this experiment is to generate data sets containing outliers and confirms that the prediction accuracy can vary between logistic regression and linear discriminant analysis. The results of the simulations presented in Figures 13 and 14 show that both classifiers are sensitive to outliers, in terms of scattered and uncertain results. The visualizations show that the prediction accuracy gets worse when increasing the scale parameter. An explanation is that the probability density function of the Cauchy distribution gets more heavy-tailed as the value of the scale parameter increases, which in turn results in outliers even further away from the mean. The boxplot captures the inability of logistic regression to predict some of the data sets, when the scale parameter equals 0.5, corresponding to a narrow probability density function. Nevertheless, the overall results of the simulations indicates that logistic regression might not be affected to the same extent as linear discriminant analysis, which is consistent with the theory given in section 2.2.7. On several occasions, probabilities in the fitting procedure for logistic regression were close to 0 or 1 when the distribution of the radius was chosen as heavy-tailed. This problem is triggered by the outliers. However, logistic regression still produces reasonable classification results and did not seem to be noticeably affected by these numerical problems when they occurred, compared to the case when they did not occur. Therefore, the problem is ignored since the aim is to compare the performance of logistic regression and linear discriminant analysis and their ability to handle extreme values for these particular data sets.

5.2 Evaluation, Interpretation and Complexity

The main focus in this thesis is to study the predictive power for linear classification methods. First of all, there are many options when it comes to measuring the prediction accuracy and we used misclassification rate and AUC for this purpose, although the reliability of these measurements is sometimes questionable. The advantage with these evaluation metrics is that they have an understandable interpretation, but we should be aware of their disadvantages. For instance, the misclassification rate appears to be misleading in some cases since correct classifications are equally treated and the decomposition of the test error rate is not taken into account [13]. As a consequence, the prediction accuracy can be somewhat optimistic. Various functions of the four possible outcomes in the confusion matrix, see Table 1, can be of interest in order to optimize the trade-off between correct and incorrect classifications from a certain perspective. Then, the overall predictive power is not always the primary interest.

Sometimes the interpretability of classification methods can be of great importance and preferred over the predictability. An advantage among these linear classifiers is the intuitive interpretation of the relationship between the predictors and the response. Less flexible methods are simpler to interpret and can to a greater extent identify parameters that have an impact on the response, compared to more complex methods. In short, the recurring topic is the bias-variance trade-off, explained in section 2.1.2. The result from Experiment C showed that the non-linear optimal Bayes' decision boundary did not affect the linear classifiers remarkably. The argument of choosing a relatively simple method that is interpretable and has a less computational complexity is quite convincing if the linear classifier still meets the required level of acceptable performance. On the other hand, imagine a situation where one group is totally surrounded by another group. Then, linear classifiers are incapable of adapting to data since they can only divide observations linearly, and hence the precision in prediction accuracy is not better than guessing. Underfitting arises and then a more flexible method is preferable. However, complex methods have the tendency to overfit data, and this is also important to take into account. This confirms the pursuit of maintaining the balance between bias and variance and the choice of method should depend on the underlying structure of the data. The simulation part is rather restricted and only regards binary classification problems in selected situations. A natural extension of the study is to consider observations with more than two classes. Moreover, it might be of interest to examine a number of other experiments by combining different situations and vary several parameters, while isolating others. The performance of logistic regression and linear discriminant analysis can be evaluated from a different perspective and other evaluation metrics can be used.

6 Conclusions

The aim of this study is to get a deeper understanding of the linear classification methods logistic regression and linear discriminant analysis. Overall, the simulations in this thesis agreed with the provided theory; logistic regression and linear discriminant analysis perform similarly in terms of prediction accuracy, despite that a variety of simulated data sets were used. However, some notable differences between the linear classification methods were observed.

Unlike logistic regression, linear discriminant analysis is based on more assumptions of the data. The simulations show that linear discriminant analysis still has high prediction accuracy, even though the assumptions are violated. The results also show that the performance of the linear classifiers is relatively unchanged when the optimal Bayes' decision boundary is non-linear. This is an indication that linear adaptation to data does not necessarily have devastating consequences in prediction accuracy, but it may be valuable to consider other more flexible methods.

The main difference between the two linear classifiers is the procedure of estimating the parameters, which turns out to have an impact in some cases. There are no clear differences in prediction accuracy between logistic regression and linear discriminant analysis, for observations generated from a multivariate normal distribution, when it comes to the distance between two classes, sample sizes and the number of predictor variables. Further, both methods are affected by overfitting when increasing the dimensionality and appear to be sensitive to extreme observations, where logistic regression tends to handle outliers better than linear discriminant analysis.

To conclude, this study can support the fact that the underlying structure of data should preferably be examined before determining which method to use.

7 Appendix

The results of the simulation experiments in section 3 are presented in Tables 2-14. They contain information about the predictive power of the evaluated methods logistic regression (LR), linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA), for a particular setup. The simulation procedure is replicated 50 times for each setup. The performance in prediction accuracy is measured in terms of the average values (Ave.) and the standard deviation (sd) of the metrics misclassification rate (MR) and area under the ROC-curve (AUC).

7.1 Experiment A - Separability between Classes

	LR		LDA	
Δ	Ave.	sd	Ave.	sd
0.5	0.4023	0.0252	0.3979	0.0204
1.0	0.3109	0.0190	0.3076	0.0176
1.5	0.2304	0.0206	0.2326	0.0209
2.0	0.1586	0.0148	0.1596	0.0161
2.5	0.1061	0.0121	0.1078	0.0165
3.0	0.0675	0.0110	0.0664	0.0104

Table 2: *Experiment A. Separability of classes (Mahalanobis distance Δ) in two dimensions. Metric: MR. Parameters:* $N = 500$, $p = 2$, $\rho = 0.5$.

	LR		LDA	
Δ	Ave.	sd	Ave.	sd
0.5	0.6392	0.0261	0.6418	0.0237
1.0	0.7564	0.0214	0.7593	0.0189
1.5	0.8528	0.0203	0.8496	0.0211
2.0	0.9208	0.0093	0.9215	0.0107
2.5	0.9604	0.0067	0.9597	0.0083
3.0	0.9833	0.0044	0.9833	0.0042

Table 3: *Experiment A. Separability of classes (Mahalanobis distance Δ) in two dimensions. Metric: AUC. Parameters:* $N = 500$, $p = 2$, $\rho = 0.5$.

	LR		LDA	
Δ	Ave.	sd	Ave.	sd
0.5	0.4260	0.0285	0.4248	0.0254
1.0	0.3245	0.0211	0.3230	0.0226
1.5	0.2425	0.0198	0.2440	0.0202
2.0	0.1698	0.0158	0.1710	0.0166
2.5	0.1152	0.0152	0.1147	0.0153
3.0	0.0733	0.0107	0.0714	0.0114

Table 4: *Experiment A. Separability of classes (Mahalanobis distance Δ) with 20 predictors. Metric: MR. Parameters:* $N = 500$, $p = 20$, $\rho = 0.5$.

	LR		LDA	
Δ	Ave.	sd	Ave.	sd
0.5	0.6053	0.0344	0.6079	0.0289
1.0	0.7433	0.0216	0.7403	0.0243
1.5	0.8387	0.0188	0.8366	0.0184
2.0	0.9124	0.0114	0.9133	0.0122
2.5	0.9544	0.0095	0.9559	0.0085
3.0	0.9801	0.0040	0.9810	0.0042

Table 5: *Experiment A. Separability of classes (Mahalanobis distance Δ) with 20 predictors. Metric: AUC. Parameters:* $N = 500$, $p = 20$, $\rho = 0.5$.

7.2 Experiment B - Number of Observations and Predictor Variables

	LR		LDA	
N	Ave.	sd	Ave.	sd
20	0.3380	0.1043	0.3250	0.1131
50	0.3352	0.0743	0.3480	0.0722
100	0.3260	0.0351	0.3142	0.0481
250	0.3065	0.0279	0.3126	0.0289
500	0.3076	0.0214	0.3098	0.0198
1000	0.3098	0.0143	0.3133	0.0150
3000	0.3079	0.0102	0.3101	0.0090

Table 6: *Experiment B. Number of observations in two dimensions. Metric: MR. Parameters: Mahalanobis distance* $\Delta = 1$, $p = 2$, $\rho = 0.5$.

	LR		LDA	
N	Ave.	sd	Ave.	sd
20	0.7310	0.0884	0.7396	0.1253
50	0.7360	0.0709	0.7233	0.0865
100	0.7433	0.0483	0.7465	0.0554
250	0.7624	0.0264	0.7588	0.0314
500	0.7599	0.0229	0.7602	0.0196
1000	0.7580	0.0137	0.7561	0.0145
3000	0.7609	0.0101	0.7586	0.0092

Table 7: *Experiment B. Number of observations in two dimensions. Metric: AUC. Parameters: Mahalanobis distance* $\Delta = 1$, $p = 2$, $\rho = 0.5$.

	LR		LDA	
p	Ave.	sd	Ave.	sd
2	0.3070	0.0230	0.3049	0.0202
5	0.3151	0.0244	0.3141	0.0238
10	0.3225	0.0267	0.3141	0.0198
20	0.3277	0.0238	0.3263	0.0193
50	0.3436	0.0244	0.3443	0.0193
100	0.3700	0.0223	0.3668	0.0229

Table 8: *Experiment B. Number of predictors. Metric: MR. Parameters:* $N = 500$, *Mahalanobis distance* $\Delta = 1$, $\rho = 0.5$.

	LR		LDA	
p	Ave.	sd	Ave.	sd
2	0.7623	0.0221	0.7647	0.0187
5	0.7534	0.0249	0.7529	0.0256
10	0.7477	0.0250	0.7543	0.0224
20	0.7389	0.0252	0.7394	0.0191
50	0.7159	0.0273	0.7149	0.0209
100	0.6809	0.0241	0.6846	0.0252

Table 9: *Experiment B. Number of predictors. Metric: AUC. Parameters:* $N = 500$, *Mahalanobis distance* $\Delta = 1$, $\rho = 0.5$.

7.3 Experiment C - Classes with Different Covariance Matrices

	LR		LDA		QDA	
ρ	Ave.	sd	Ave.	sd	Ave.	sd
0	0.3268	0.0229	0.3268	0.0266	0.3833	0.0196
0.2	0.3199	0.0205	0.3182	0.0190	0.3216	0.0257
0.3	0.3226	0.0227	0.3224	0.0213	0.2644	0.0175
0.4	0.3184	0.0183	0.3175	0.0189	0.1936	0.0169
0.6	0.3120	0.0211	0.3112	0.0207	0.0774	0.0119
0.8	0.2921	0.0194	0.2880	0.0236	0.0118	0.0045

Table 10: *Experiment C. Classes with different covariance matrices. Metric: MR. Parameters:* $N = 500$, $p = 20$. *The correlation* ρ *given in the table is adjusted for one class, while the other class is fixed at* $\rho = 0$.

	LR		LDA		QDA	
ρ	Ave.	sd	Ave.	sd	Ave.	sd
0	0.7386	0.0267	0.7384	0.0271	0.6599	0.0232
0.2	0.7472	0.0261	0.7468	0.0240	0.7438	0.0286
0.3	0.7447	0.0263	0.7432	0.0235	0.8166	0.0189
0.4	0.7475	0.0182	0.7477	0.0207	0.8900	0.0128
0.6	0.7558	0.0241	0.7529	0.0214	0.9794	0.0043
0.8	0.7668	0.0204	0.7668	0.0218	0.9994	0.0006

Table 11: *Experiment C. Classes with different covariance matrices. Metric: AUC. Parameters:* $N = 500$, $p = 20$. *The correlation* ρ *given in the table is adjusted for one class, while the other class is fixed at* $\rho = 0$.

7.4 Experiment D - The Effect of Non-Normality

		LR		LDA	
p	Metric	Ave.	sd	Ave.	sd
3	MR	0.3072	0.0210	0.3071	0.0226
10	MR	0.3176	0.0164	0.3219	0.0192
3	AUC	0.7571	0.0189	0.7626	0.0220
10	AUC	0.7498	0.0174	0.7441	0.0196

Table 12: *Experiment D. The distance to the center of symmetry has a square root of a Chi-squared distribution. Parameters: Mahalanobis distance* $\Delta = 1$, $N = 500$.

		LR		LDA	
p	Scale	Ave.	sd	Ave.	sd
3	0.5	0.1999	0.1120	0.2334	0.1065
3	1.0	0.3208	0.1009	0.3700	0.1084
3	2.0	0.4190	0.0789	0.4495	0.0666
10	0.5	0.1422	0.1083	0.1156	0.0246
10	1.0	0.1830	0.0548	0.2054	0.0557
10	2.0	0.2896	0.0422	0.3065	0.0626

Table 13: *Experiment D. The distance to the center of symmetry has a one-tailed Cauchy distribution. Metric: MR. Parameters: Mahalanobis distance* $\Delta = 1$, $N = 500$.

		LR		LDA	
p	Scale	Ave.	sd	Ave.	sd
3	0.5	0.7952	0.1144	0.7923	0.0990
3	1.0	0.6813	0.1036	0.6398	0.1188
3	2.0	0.5749	0.0829	0.5604	0.0877
10	0.5	0.8547	0.1076	0.8810	0.0275
10	1.0	0.8056	0.0596	0.7924	0.0453
10	2.0	0.6951	0.0407	0.6902	0.0500

Table 14: *Experiment D. The distance to the center of symmetry has a one-tailed Cauchy distribution. Metric: AUC. Parameters: Mahalanobis distance* $\Delta = 1$, $N = 500$.

References

[1] AGRESTI, A. (2013). *Categorical Data Analysis*, third edition. John Wiley & Sons, Inc., Hoboken, New Jersey.

[2] BOISBUNON, A. (2012). The class of multivariate spherically symmetric distributions. Université de Rouen, Technical report (2012-005).

[3] BOX, G., JENKINS, G. AND REINSEL, G. (2008). *Time Series Analysis*, fourth edition. John Wiley & Sons, Inc., Hoboken, New Jersey. (p.25-30)

[4] FAWCETT, T. (2006). Introduction to ROC Analysis, *Pattern Recognition Letters*, 27(8), 861-874. Elsevier.

[5] FISHER, R. (1936). Use of Multiple Measurements in Taxonomic Problems, *Annals of Eugenics*, 7(2), 179–188.

[6] HASTIE, T., TIBSHIRANI, R. AND FRIEDMAN, J. (2017). *The Elements of Statistical Learning*, second edition. Springer, New York.

[7] JAMES, G., WITTEN, D., HASTIE, T. and TIBISHIRANI, R. (2017). *An Introduction to Statistical Learning*. Springer, New York.

[8] JOLLIFFE, I (2002). *Principal Component Analysis*, second edition. Springer, New York. (p.11-13)

[9] MCLACHLAN, G. (1999). Mahalanobis Distance. The University of Queensland. *Resonance* 4(6):20-26

[10] MUKHERJEE, B. and MAITI, S. (1988). On Some Properties of Positive Definite Toeplitz Matrices and Their Possible Applications. *Linear Algebra and its Applications*, 102: 211-240. Elsevier.

[11] ROSS, S. (2014). *Introduction to Probability Models*, eleventh edition. Elsevier Inc. (p.649-650)

[12] TONG, Y. (1990). *The Multivariate Normal Distribution*. Springer, New York. (p.26)

[13] WEBB, A. (2002). *Statistical Pattern Recognition*, second edition. John Wiley & Sons, Ltd.

[14] WELCH, B. (1939). Note on Discriminant Functions, *Biometrika*, 31, 218–220.