FACULTY OF SCIENCE AND TECHNOLOGY

BACHELOR'S THESIS

Study programme / specialisation: Mathematics and Physics	The spring semester, 2023 Open
Author: Stian Hammerseth Wighus	
Supervisor at UiS: Jan Terje Kvaløy	
Thesis title: Multivariate statistics	
Credits (ECTS): 20	
Keywords: Multivariate statistics Non-parametric kernel estimation Classification	Pages: 39 + appendix: 0 Stavanger, 15.05.2023

Multivariate Statistics

Stian Hammerseth Wighus

Spring 2023

Acknowledgements

I would like to give my sincere gratitude to my supervisor, Jan Terje Kvaløy, who has given me valuable guidance and a generous amount of excellent feedback. Furthermore, I would like to thank him for his patience to answer any question I could come up with, and for being willing to explain any concept in detail to give me a better understanding. I am grateful to have such a dedicated and knowledgeable supervisor.

Abstract

This bachelor thesis provides an introduction to multivariate statistics, which is the analysis of data with multiple variables using statistical methods. The thesis focuses on the generalization of the normal distribution to random vectors, properties of the multivariate normal distribution, and non-parametric kernel estimation methods for estimating densities. Additionally, the thesis presents methods for separating populations and classifying new observations within these populations using multivariate statistics. The applications of these methods is demonstrated using a real data set, and the accuracy of the classification is evaluated.

Contents

Chapter 1

Introduction

In the field of statistical analysis, measuring multiple variables is a common setting, and multivariate methods are the primary tools used for analysis. These methods enable the examination of relationships between different variables, allowing for a more comprehensive understanding of the data. The purpose of this bachelor thesis is to provide an introduction to the subject of multivariate statistics.

Specifically, we will explore the generalization of the normal distribution to random vectors and analyze some of the properties of the multivariate normal distribution. Additionally, we will discuss non-parametric kernel estimation methods for estimating densities. Using multivariate statistics, we will construct methods for separating populations and defining rules for classifying new observations within these populations. To demonstrate the effectiveness of these methods, we will apply them to a real data set and evaluate the accuracy of the classification. By delving into multivariate statistics and exploring these various techniques, this thesis aims to provide readers with an introductory understanding of this important field of statistical analysis. In chapter 2 we will build the theory we need to handle multivariate observations, in chapter 3 we will be concerned with applying the methods constructed in the previous chapter on a real data set, and finally a brief summary in chapter 4.

Chapter 2

Theory

Notation

First, we will have to familiarize ourselves with the notation. We will use capital letters, say X, for random variables. When we make an observation, we will use lowercase, for example x, to denote them. When making n observations of p different variables we will use the first index for the observation number and the second for the variable, so for the ith observation of the kth variable we will write x_{ik}, where $i = 1, 2, \ldots, n$ and $k = 1, 2, \ldots, p$. Bold will be used for vectors and matrices, like the mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$, which we will use throughout this text. And finally, for vectors and matrices, we will denote the transposed as $\mathbf{X}'$.

2.1 Univariate statistics

2.1.1 Random variables

During this text, we will be working with random variables. So it will be purposeful to familiarize ourselves with some basic concepts related to random variables. For a discrete random variable X with corresponding probability mass function $f(x)$ we have the following criteria for all possible outcomes x:

1. $f(x) \geq 0$.
2. $\sum_x f(x) = 1$.
3. $P(X = x) = f(x)$.

For a continuous random variable where possible outcomes are among the real numbers, the probability density function $f(x)$ must satisfy:

1. $f(x) \geq 0 \quad \forall\, x \in \mathbb{R}$.
2. $\int_{-\infty}^{\infty} f(x)\, dx = 1$.
3. $P(a < X < b) = \int_a^b f(x)\, dx$.

The expected value

The expected value, or mean value, is the "center of mass" of the distribution and for random variables of countably many outcomes can be thought of as the average of a long-run experiment. The expected value is defined as the first uncentered moment of f,

$$\mu = E(X) = \begin{cases} \sum_x x f(x) & \text{If } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x f(x)\, dx & \text{If } X \text{ is continuous} \end{cases}$$

Variance and covariance

The variance of a variable is the expected squared error from the mean, which is the second centered moment of f.

$$\operatorname{Var}(X) = E((x - \mu)^2) = \begin{cases} \sum_x (x - \mu)^2 f(x) & \text{If } X \text{ is discrete} \\ \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\, dx & \text{If } X \text{ is continuous} \end{cases}$$

For this text's purposes, we will use the notation σ_{kk} for the variance of variable k. This notation choice simplifies an object we will define in due time. Then with our notation $\operatorname{Var}(X_k) = \sigma_{kk}$

For two random variables X and Y with joint probability distribution $f(x, y)$, the covariance is a measure of the linear association between them. Statistically independent variables will have zero covariance.

$$\operatorname{Cov}(X, Y) = \sigma_{XY} = \begin{cases} \sum_x \sum_y (x - \mu_X)(y - \mu_Y) f(x, y) \\ \iint_{\mathbb{R}^2} (x - \mu_X)(y - \mu_Y) f(x, y)\, dA \end{cases}$$

Correlation

We will now define Pearson's product-moment correlation coefficient. This coefficient is a measure of the linear association between two variables and does not

depend on the units of measurement. The correlation coefficient for the ith and kth variables is defined as

$$\rho_{ik} = \frac{\sigma_{ik}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{kk}}}$$

This coefficient lies on the closed interval $[-1, 1]$ where a larger absolute value corresponds to a stronger positive or negative linear association, with zero implying a lack of correlation.

2.1.2 Descriptive statistics

For large datasets it is impractical to try to assess the underlying information from each individual data point, thus some descriptive statistics are used to summarise the information that we're interested in.

The arithmetic average

A commonly used statistic is the arithmetic average, or the sample mean, which is often used as an estimator for the expected value of the underlying distribution. The arithmetic average is the sum of all measurements of a variable divided by the number of measurements. So the arithmetic average of the kth variable is

$$\bar{x}_k = \frac{1}{n}\sum_{j=1}^{n} x_{jk}$$

The sample variance and covariance

Another descriptive statistic central to our understanding of multivariate statistics is the sample variance and covariance. For the kth variable the sample variance is

$$s_k^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2$$

Covariance is used as a measure for the linear association between the measurements of two variables. For the ith and kth variables the sample covariance is

$$s_{ik} = \frac{1}{n-1}\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)$$

Sample correlation coefficient

Similarly to the correlation between two random variables, we can define the sample correlation between two variables from a set of pairwise measurements by using the sample variances and covariance. So, for the sample correlation between the ith and kth variables we define

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}} = \frac{\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)}{\sqrt{\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)^2}\sqrt{\sum_{j=1}^{n}(x_{jk} - \bar{x}_k)^2}}$$

2.2 Multivariate statistics

One might wonder why you would need multivariate statistics and when you would need it. And the simple answer is that multivariate statistics comes up whenever we measure multiple variables for the same unit. So when a doctor not only measures your blood pressure, but also your height, weight, blood protein levels, blood sugar and so on, you better hope that the people analyzing your test know some multivariate statistics so that they can detect outliers, or find out whether the measures would classify you as in risk of some disease or not.

2.2.1 Multivariate objects of random variables

In the following sections, we will familiarise ourselves with some objects central to our study of multivariate statistics.

Random vectors

A multivariate random variable or **random vector** is a vector of random variables, so $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$ is a $p \times 1$ random vector where each of the elements in $\mathbf{X}$ is its own random variable.

Mean vector

Following from the random vector $\mathbf{X}$ we define the mean vector $\boldsymbol{\mu}$, which is the vector of expected values for the elements in $\mathbf{X}$. So

$$E(\mathbf{X}) = \begin{pmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_p) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} = \boldsymbol{\mu}$$

For a random vector that can be expressed as a linear combination of $\mathbf{X}$ such as $\mathbf{Y} = \mathbf{C}\mathbf{X} + \mathbf{b}$, where $\mathbf{C}$ is matrix and $\mathbf{b}$ is a vector, the mean vector $\boldsymbol{\mu}_\mathbf{Y}$ becomes

$$\begin{aligned} E(\mathbf{Y}) &= E(\mathbf{C}\mathbf{X} + \mathbf{b}) \\ &= \mathbf{C}\, E(\mathbf{X}) + \mathbf{b} \\ &= \mathbf{C}\boldsymbol{\mu}_\mathbf{X} + \mathbf{b} \end{aligned}$$

Variance-Covariance matrix

The variance-covariance matrix, or simply the covariance matrix, is a symmetric positive-definite $p \times p$ matrix composed of variances and covariances for a random vector.

$$\boldsymbol{\Sigma} = \operatorname{Cov}(\mathbf{X}) = E\left((\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})'\right)$$

$$= E\left(\begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \\ \vdots \\ X_p - \mu_p \end{bmatrix} \begin{bmatrix} X_1 - \mu_1, & X_2 - \mu_2, & \ldots, & X_p - \mu_p \end{bmatrix} \right)$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1p} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} & \cdots & \sigma_{2p} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} & \cdots & \sigma_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{1p} & \sigma_{2p} & \sigma_{3p} & \cdots & \sigma_{pp} \end{pmatrix}$$

The inverse of the covariance matrix $\boldsymbol{\Sigma}^{-1}$ is often called the precision matrix. The covariance for the linear combination $\mathbf{Y} = \mathbf{C}\mathbf{X} + \mathbf{b}$, which might be denoted as $\boldsymbol{\Sigma}_\mathbf{Y}$, can be expressed as

$$\begin{aligned} \operatorname{Cov}(\mathbf{Y}) &= \operatorname{Cov}(\mathbf{C}\mathbf{X} + \mathbf{b}) \\ &= \mathbf{C} \operatorname{Cov}(\mathbf{X})\, \mathbf{C}' \\ &= \mathbf{C}\boldsymbol{\Sigma}_\mathbf{X}\mathbf{C}' \end{aligned}$$

2.2.2 Sampling statistics

Random samples

When making a random sample, which corresponds to making n independent identically distributed multivariate observations, we will construct the data matrix

$\mathbf{X}$. Then $\mathbf{X}$ is a $n \times p$ matrix, meaning each of our p variables is a column in a matrix with rows corresponding to one of the n observations for each of the variables. Such that

$$\underset{(n\times p)}{\boldsymbol{X}} = \begin{pmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1p} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2p} \\ x_{31} & x_{32} & x_{33} & \cdots & x_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \cdots & x_{np} \end{pmatrix}$$

Unfortunately, this gives the notation $\mathbf{X}$ two different meanings, either as a random vector or as a random sample, usually this won't be a problem and the usage will be clear from the context. Through the random sample we will estimate the mean vector and covariance matrix.

Sample mean vector

The sample mean vector $\bar{\mathbf{x}}$ is the generalization of the arithmetic average and using a bit of matrix multiplication we can calculate this directly from the random sample

$$\bar{\mathbf{x}} = \frac{1}{n}\mathbf{X}'\mathbf{1} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_n \end{pmatrix}$$

Where $\mathbf{1}$ is a $n \times 1$ vector of ones.

Sample variance-covariance matrix

For the sample variance-covariance matrix the procedure is similar. We will make use of the $n \times p$ matrix of deviations $\mathbf{X} - \frac{1}{n}\mathbf{1}\mathbf{1}'\mathbf{X}$, so

$$\begin{aligned} \mathbf{S} &= \frac{1}{n-1}\left(\mathbf{X} - \frac{1}{n}\mathbf{1}\mathbf{1}'\mathbf{X}\right)'\left(\mathbf{X} - \frac{1}{n}\mathbf{1}\mathbf{1}'\mathbf{X}\right) \\ &= \frac{1}{n-1}\mathbf{X}'\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)\mathbf{X} \\ &= \begin{pmatrix} s_{11} & s_{12} & s_{13} & \cdots & s_{1p} \\ s_{12} & s_{22} & s_{23} & \cdots & s_{2p} \\ s_{13} & s_{23} & s_{33} & \cdots & s_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s_{1p} & s_{2p} & s_{3p} & \cdots & s_{pp} \end{pmatrix} \end{aligned}$$

2.3 Multivariate normal distribution

The (univariate) normal distribution is possibly the most famous probability distribution. It applies to a wide range of situations, both as a population model and as an approximation to the sampling distributions of many different statistics through the central limit theorem. Thus it should not be surprising that there exists a multivariate normal distribution. Many methods in multivariate statistics are based on the assumption that the data is generated from a multivariate normal distribution.

One advantage of the normal distribution is that it's a mathematically manageable function that behaves "nicely", this leads to many properties and results that follow directly from algebraic manipulations, some of which are included here. First, we will have a look at the probability density function for the multivariate normal distribution. For the familiar univariate normal density with mean μ and variance σ^2 we have

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

and is often denoted as $N(\mu, \sigma^2)$, an extension of this notation will also be used for the multivariate case. Notice that in the exponent of the univariate normal density function, we have the term

$$\left(\frac{x \, - \, \mu}{\sigma}\right)^2 \, = \, (x \, - \, \mu)(\sigma^2)^{-1}(x \, - \, \mu)$$

which measures the distance from the mean in standard deviation units. This can be generalized to a $p \times 1$ vector $\mathbf{x}$ of observations on several variables as

$$(\mathbf{x} \, - \, \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} \, - \, \boldsymbol{\mu}) \tag{2.1}$$

and this shall substitute the univariate distance in the exponential. The next step to generalize the univariate density function is to normalize the function, for a univariate normal distribution the normalization constant is $(2\pi)^{-1/2}(\sigma^2)^{-1/2}$, and in the multivariate case, we have

$$f(\mathbf{x}) \, = \, Ce^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

Then we need to determine the constant C such that the integral over the p-dimensional space is unity

$$1 \, = \, C \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})} \, dx_p \cdots dx_1$$

To evaluate the integral one can use the fact that $\mathbf{\Sigma}$ is positive definite to make the change of variables $\mathbf{x} - \boldsymbol{\mu} = \mathbf{\Sigma}^{1/2}\mathbf{z}$, where $\mathbf{\Sigma}^{1/2}\mathbf{\Sigma}^{-1}\mathbf{\Sigma}^{1/2} = \mathbf{I}$. The Jacobian for this change of variables is $|\mathbf{\Sigma}|^{-1/2}$. This gives us

$$\begin{aligned}
1 &= C|\mathbf{\Sigma}|^{1/2} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-\frac{1}{2}\mathbf{z}'\mathbf{z}}\, dz_p \cdots dz_1 \\
&= C|\mathbf{\Sigma}|^{1/2} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{i=1}^{p} e^{-\frac{z_i^2}{2}}\, dz_p \cdots dz_1 \\
&= C|\mathbf{\Sigma}|^{1/2} \left(\int_{-\infty}^{\infty} e^{-\frac{z^2}{2}}\, dz \right)^p \\
&= C|\mathbf{\Sigma}|^{1/2}(2\pi)^{p/2} \\
\implies C &= (2\pi)^{-p/2}|\mathbf{\Sigma}|^{-1/2}
\end{aligned}$$

And now we have the p-dimensional normal density for the random vector $\mathbf{X}' = [X_1,\, X_2, \cdots, X_p]$

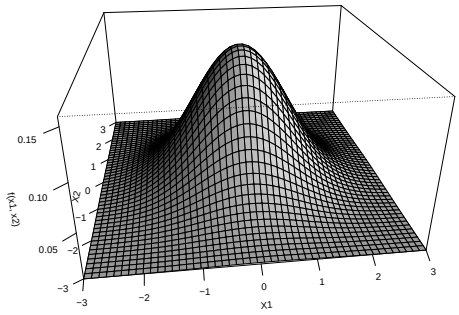
$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

Where $-\infty < x_i < \infty$ for $i = 1, 2, \ldots, p$. Analogous to the univariate normal density we will denote the p-dimensional normal density as $N_p(\boldsymbol{\mu},\, \mathbf{\Sigma})$.

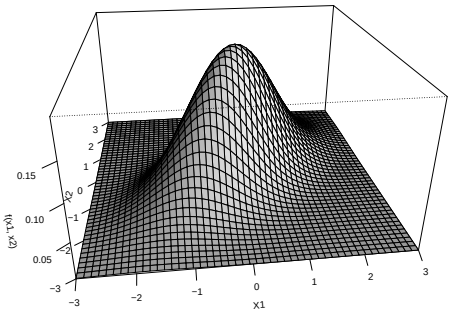
Plotting of the multivariate normal density is mainly limited to the cases $p = 1$ and $p = 2$, we will demonstrate two common methods for visualization. The first method is by perspective plots, where the height of the surface above the (X_1, X_2)-plane corresponds to the joint probability density, and the volume under the surface and above a region in the plane corresponds to the probability of a observation form this distribution to return a point in the region.

The second method is by a heat map. Unlike perspective plots where we attempt to draw the 3-dimensional surface, we instead represent the density by colour, where regions of higher density is usually illustrated with a increase in hue or intensity.

The methods described in the section above are illustrated in Figure 2.1 and 2.2. For the plots, the bivariate standard normal density $N_2(\mathbf{0}, \mathbf{I})$ and a bivariate normal density with $\boldsymbol{\mu} = \mathbf{0}$ and $\mathbf{\Sigma} = \begin{bmatrix} 1 & 0.6 \\ 0.6 & 1 \end{bmatrix}$ were used. We can then notice how the density concentrates along a line, if we instead had negative correlation we would have seen the same concentration, but along the line $x_1 = -x_2$ instead.

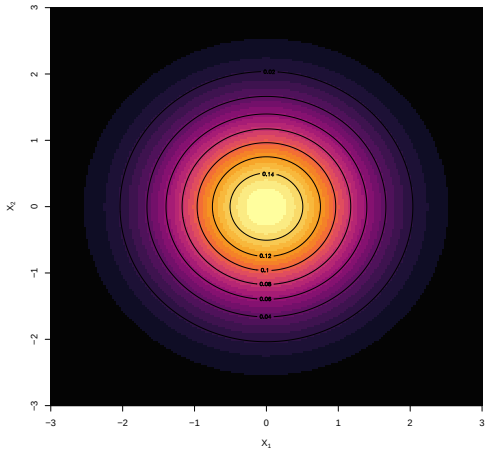


(a) Bivariate standard normal density $N_2(\mathbf{0}, \mathbf{I})$

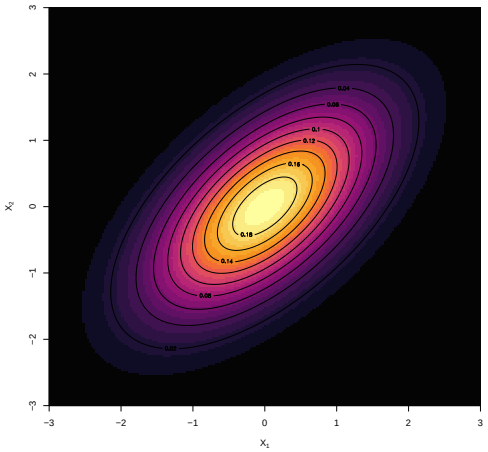


(b) Bivariate normal density with $\rho_{12} = 0.6$

Figure 2.1: Perspective plots



(a) Bivariate standard normal density $N_2(\mathbf{0}, \mathbf{I})$



(b) Bivariate normal density with $\rho_{12} = 0.6$

Figure 2.2: Heat maps

2.3.1 Properties

There are many beneficial properties of the multivariate normal distribution, we will look at some of them in the following section.

Statistical Independence between variables

Statistical independence is an important notion throughout probability theory. Two random vectors are said to be independent if

$$f(\mathbf{x}_1,\, \mathbf{x}_2) \;=\; f_1(\mathbf{x}_1) f_2(\mathbf{x}_2)\,,$$

where f is the joint probability distribution of the random vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ and f_1 and f_2 are the marginal distributions of $\mathbf{x}_1$ and $\mathbf{x}_2$ respectably. From the quadratic form in equation 2.1 we can observe that when $\mathbf{\Sigma}^{-1}$ is a diagonal matrix all the cross-terms disappear and we are left with

$$(\mathbf{x} \,-\, \boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x} \,-\, \boldsymbol{\mu}) \;=\; \left(\frac{x_1 \,-\, \mu_1}{\sigma_1}\right)^2 \,+\, \cdots \,+\, \left(\frac{x_p \,-\, \mu_p}{\sigma_p}\right)^2 .$$

Along with $|\mathbf{\Sigma}|^{1/2} = \sigma_1 \cdots \sigma_p$. Because we now have no cross-terms in the exponential, only a sum of the squared statistical distances, and all of the standard deviations as a product in the normalization constant we can write the multivariate normal density as a product of univariate normal densities, which means that the elements of $\mathbf{X}$ are mutually independent. This is also true even when some elements in $\mathbf{X}$ are dependent, and this can be shown by partitioning the random vector $\mathbf{X}$ as $\mathbf{X}' = [\mathbf{X}^{(1)\prime}, \mathbf{X}^{(2)\prime}]$ where

$$\mathbf{X}^{(1)} \;=\; \begin{pmatrix} X_1 \\ \vdots \\ X_q \end{pmatrix}, \qquad \mathbf{X}^{(2)} \;=\; \begin{pmatrix} X_{q+1} \\ \vdots \\ X_p \end{pmatrix} .$$

This partition also affects the mean vector and covariance matrix as follows

$$\boldsymbol{\mu} \;=\; \begin{pmatrix} E\left(\mathbf{x}^{(1)}\right) \\ E\left(\mathbf{x}^{(2)}\right) \end{pmatrix} \;=\; \begin{pmatrix} \boldsymbol{\mu}^{(1)} \\ \boldsymbol{\mu}^{(2)} \end{pmatrix}$$

$$\mathbf{\Sigma} \;=\; \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix} ,$$

where $\mathbf{\Sigma}_{ij} \;=\; E\left((\mathbf{x}^{(i)} - \boldsymbol{\mu}^{(i)})(\mathbf{x}^{(j)} - \boldsymbol{\mu}^{(j)})'\right)$ and $\mathbf{\Sigma}_{ij} \;=\; \mathbf{\Sigma}'_{ji}$. In the case where $\mathbf{\Sigma}_{12} = \mathbf{\Sigma}'_{21} = \mathbf{0}$, a matrix of only zeroes, the quadratic form from earlier becomes

$$\begin{aligned}
(\mathbf{x}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) &= \left[(\mathbf{x}^{(1)} - \boldsymbol{\mu}^{(1)})', \quad (\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)})'\right] \begin{pmatrix} \mathbf{\Sigma}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{22}^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{x}^{(1)} - \boldsymbol{\mu}^{(1)} \\ \mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)} \end{pmatrix} \\
&= \left[(\mathbf{x}^{(1)} - \boldsymbol{\mu}^{(1)})', \quad (\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)})'\right] \begin{pmatrix} \mathbf{\Sigma}_{11}^{-1}(\mathbf{x}^{(1)} - \boldsymbol{\mu}^{(1)}) \\ \mathbf{\Sigma}_{22}^{-1}(\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)}) \end{pmatrix} \\
&= (\mathbf{x}^{(1)} - \boldsymbol{\mu}^{(1)})'\mathbf{\Sigma}_{11}^{-1}(\mathbf{x}^{(1)} - \boldsymbol{\mu}^{(1)}) \;+\; (\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)})'\mathbf{\Sigma}_{22}^{-1}(\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)})
\end{aligned}$$

When combined with the fact $|\mathbf{\Sigma}| = |\mathbf{\Sigma}_{11}||\mathbf{\Sigma}_{22}|$ we can see that the p-dimensional density $f(\mathbf{x}^{(1)}, \mathbf{x}^{(2)})$ can be written as the product of two multivariate densities $f_1(\mathbf{x}^{(1)})$ and $f_2(\mathbf{x}^{(2)})$ of dimensions q and $(p-q)$ respectively.

Contours of constant density

Another important property of the multivariate normal density is the contours of constant probability. From the $p = 2$ cases which are plotted in Figure 2.1 and 2.2 we can see that contours of constant density are ellipsoids defined by $\{\mathbf{x} : (\mathbf{x}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = c^2\}$, where these ellipsoids are centered at $\boldsymbol{\mu}$ and have axes $\pm c\sqrt{\lambda_i}\mathbf{e}_i$ with $(\lambda_i, \mathbf{e}_i)$ as a normalized eigenvalue-eigenvector pair.

For a better understanding of the constant c^2 we will look at a result concerning the distribution of the quadratic form $(\mathbf{x}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$. First, we will use spectral decomposition to write $\mathbf{\Sigma}^{-1}$ as $\sum_{i=1}^{p} \frac{1}{\lambda_i}\mathbf{e}_i\mathbf{e}_i'$. Thus the quadratic form then becomes

$$\begin{aligned}(\mathbf{x}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) &= \sum_{i=1}^{p}\frac{1}{\lambda_i}(\mathbf{x}-\boldsymbol{\mu})'\mathbf{e}_i\mathbf{e}_i'(\mathbf{x}-\boldsymbol{\mu}) = \sum_{i=1}^{p}\frac{1}{\lambda_i}(\mathbf{e}_i'(\mathbf{x}-\boldsymbol{\mu}))^2 \\ &= \sum_{i=1}^{p}\left(\frac{1}{\sqrt{\lambda_i}}\mathbf{e}_i'(\mathbf{x}-\boldsymbol{\mu})\right)^2 = \sum_{i=1}^{p}\mathrm{Z}_i^2 = \mathbf{Z}'\mathbf{Z}\end{aligned}$$

Where it can then be shown[3] that the linear combination $\mathbf{Z}$ is distributed as $N_p(\mathbf{0}, \mathbf{I})$, thus $Z_1, Z_2, \ldots, Z_p$ are independent standard normal variables and $(\mathbf{x}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$ has a χ_p^2 distribution.

Following from the previous result the multivariate normal distribution assigns probability $1-\alpha$ to the solid ellipsoid $\{\mathbf{x} : (\mathbf{x}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \leq \chi_p^2(\alpha)\}$, where $\chi_p^2(\alpha)$ is the upper $(100\alpha)th$ percentile of the χ_p^2 distribution.

Other properties

Other properties of the normal distribution are, among others, that linear combinations of multivariate variables are multivariate normal and that the conditional distributions of components of a random vector are multivariate normal. We will not go further into this. Proofs and examples can be found in [3]

2.3.2 Maximum likelihood estimation

Unfortunately, when working with a random sample we rarely have access to $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$, therefore we will have to estimate them. Before we can find the maximum likelihood estimators for $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ we will have to define the likelihood function for our sample.

When we have made a random sample $\mathbf{X}$ with observations $\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_n$ from a multivariate normal population

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \mathbf{x}_3' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1p} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2p} \\ x_{31} & x_{32} & x_{33} & \cdots & x_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \cdots & x_{np} \end{pmatrix}$$

We can define the likelihood function $L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ as the product of the density for each observation, making it a function of the unknown parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Then the maximum likelihood estimator is the value of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ that maximizes $L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. For the univariate normal distribution with mean μ and variance σ^2 the likelihood function is

$$\begin{aligned} L(\mu, \sigma) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right) \\ &= (2\pi)^{-n/2}\sigma^{-n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right) \end{aligned}$$

It is common to take the logarithm of the likelihood function, due to the fact that the log-likelihood function will have the same estimator as its maxima, and the log is usually easier to work with.

$$l(\mu, \sigma) = \ln(L) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Setting the partial derivatives equal to zero gives us the maximum likelihood estimators

$$\frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0 \quad \Longrightarrow \quad \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i - \mu)^2 = 0 \quad \Longrightarrow \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

For the multivariate normal distribution we obtain the likelihood function

$$\begin{aligned} L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \prod_{i=1}^{n} \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})} \\ &= \frac{1}{(2\pi)^{np/2}|\boldsymbol{\Sigma}|^{n/2}} e^{-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})} \end{aligned} \tag{2.2}$$

Further we will only present the maximum likelihood estimators for the multivariate normal distribution, for a detailed look at the procedure of maximizing 2.2 one can see [1] or [3].

For the multivariate normal distribution the maximum likelihood estimators for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}} \qquad\qquad \hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})' = \frac{n-1}{n}\mathbf{S}$$

One of the first things we want to check is whether our estimators are unbiased and consistent. For $\hat{\boldsymbol{\mu}}$ we notice that $E(\hat{\boldsymbol{\mu}}) = \boldsymbol{\mu}$ and $\mathrm{Var}(\hat{\boldsymbol{\mu}}) = \frac{1}{n}\boldsymbol{\Sigma}$, meaning that it is a unbiased and consistent estimator for $\boldsymbol{\mu}$. As for the covariance $\hat{\boldsymbol{\Sigma}}$ is consistent, but unbiased.

2.3.3 Simulation

Simulation is a common method in computational statistics. It is often used when modeling complex situations or when calculating the exact answer is too difficult or expensive and an approximate answer is good enough.

For simulation of a general multivariate normal distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ it is sufficient to simulate a multivariate standard normal distribution and then transform, because if $\mathbf{Z} \sim N_p(\mathbf{0}, \mathbf{1})$ then $\mathbf{X} = \mathbf{CZ} + \boldsymbol{\mu}$ is multivariate normal. This is because $\mathbf{X}$ is a linear combination of multivariate normal variables, and if we choose $\mathbf{C}$ such that $\mathbf{CC}' = \boldsymbol{\Sigma}$ the expected value becomes

$$E(\mathbf{X}) = E(\mathbf{CZ} + \boldsymbol{\mu}) = \mathbf{C}\,E(\mathbf{Z}) + \boldsymbol{\mu} = \mathbf{C0} + \boldsymbol{\mu} = \boldsymbol{\mu}$$

and the variance-covariance matrix becomes

$$\mathrm{Cov}(\mathbf{X}) = \mathrm{Cov}(\mathbf{CZ} + \boldsymbol{\mu}) = \mathbf{C}\,\mathrm{Cov}(\mathbf{Z})\,\mathbf{C}' = \mathbf{CIC}' = \boldsymbol{\Sigma}$$

leading to the conclusion that $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ as desired.

So for n simulations of the general multivariate normal distribution we first simulate a $n \times p$ matrix $\mathbf{Z}$ of standard normal variables. Then find a matrix $\mathbf{C}$ such that $\mathbf{CC}' = \boldsymbol{\Sigma}$ and apply the transformation $\mathbf{X} = \mathbf{ZC} + \mathbf{1}\boldsymbol{\mu}'$.

A common approach for simulating from the standard normal distribution is the Box-Muller transformation, where we first simulate two variables $U, V \sim Unif(0,1)$ and then compute

$$X = \sqrt{-2\ln(U)}\cos(2\pi V)$$
$$Y = \sqrt{-2\ln(U)}\sin(2\pi V)$$

It can be shown that X and Y now are two independent $N(0,1)$ variables[2].

2.4 Non-parametric kernel estimation

One of the most fundamental concepts in statistics is the probability density. In some cases the functional form of the probability density is known when analysing a random sample, but this is not always the case. One common method for estimating the density function without making any distributional assumptions is by non-parametric kernel estimation, which we will look at in the following section.

2.4.1 The kernel

The main idea of kernel estimation is to use a weighing function $K(x)$, referred to as a kernel, to weight the distance from each data point to a point of interest and then add these kernels together to get the final estimate for the density $\hat{f}(x)$. To do this there are some properties we would like our kernels to have, the first being that it satisfies

$$\int_{-\infty}^{\infty} K(x)\,dx = 1\,.$$

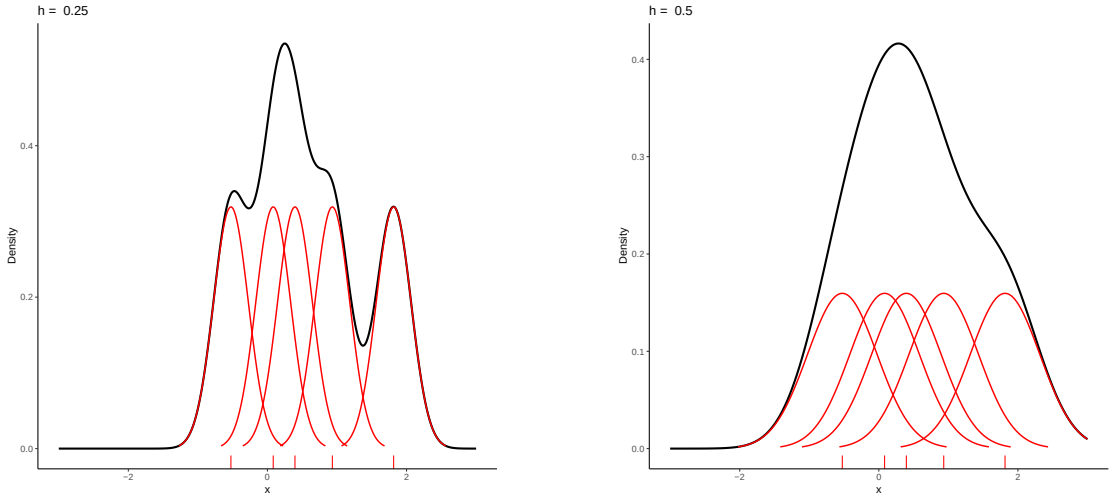
As well as $K(x) \geq 0$, this makes it easy to normalize our estimate for the density $f(x)$ by dividing by n after adding each kernel together. These conditions on $K(x)$ automatically gives us a valid probability density when considering their sum, as well as $\hat{f}(x)$ inheriting all continuity and differentiability properties of K. This is one of the reasons the normal density is a standard choice of kernel function.

2.4.2 Bandwidth

Before we write down the formula for $\hat{f}(x)$ we would like to like how to explain control the amount of smoothing for our estimate. This is done by scaling the distance from the data point and the point of interest by $\frac{1}{h}$, where $h > 0$ is our smoothing parameter, usually called bandwidth. Then our estimate for the density $f(x)$ from a random sample with observations $x_1,\ x_2,\ \cdots,\ x_n$ is

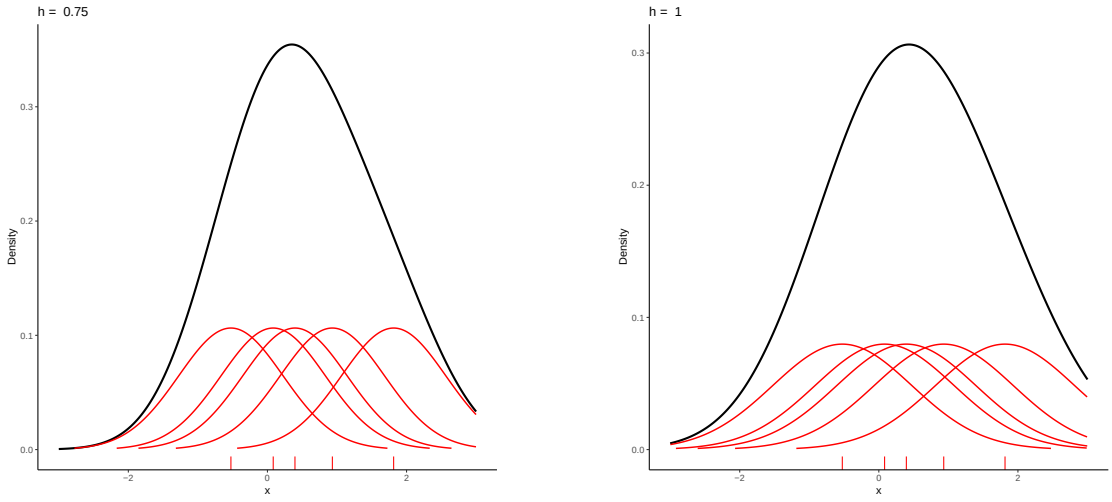
$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)\,.$$

We can note that when using a standard normal kernel, changing the parameter h is equivalent to changing the standard deviation of the normal kernel. For the following examples, we have simulated five data points from a standard normal distribution and used normal kernels with different bandwidths h. As we can see in Figure 2.4, where the individual kernels are drawn in red and the estimate for $\hat{f}$ is drawn in black, the choice of h changes our estimate of $\hat{f}(x)$. For more complex



(a) h = 0.25

(b) h = 0.5



(c) h = 0.75

(d) h = 1

Figure 2.3: Density estimates of $f_Y(x)$ with varying bandwidth from a simulated sample of 5 data points. Black: KDE $\hat{f}_Y$. Red: Individual kernels

Figure 2.4: Density estimates with varying bandwidth

densities the bandwidth can determine if some crucial bits of information can be detected or not. For the following example, we have simulated a mixture of three normal distributions, making a multimodal distribution, $f_Y(x) = 0.4f_{X_1}(x) + 0.2f_{X_2}(x) + 0.4f_{X_3}(x)$, where $X_1 \sim N(-3, 1), X_2 \sim N(0, 1)$ and $X_3 \sim N(3, 1)$. Notice how for large h it is not possible to detect the presence of X_2 and for small h the noise of the data can make it difficult to conclude anything. This illustrates that both smoothing too much and smoothing too little can give wrong results.

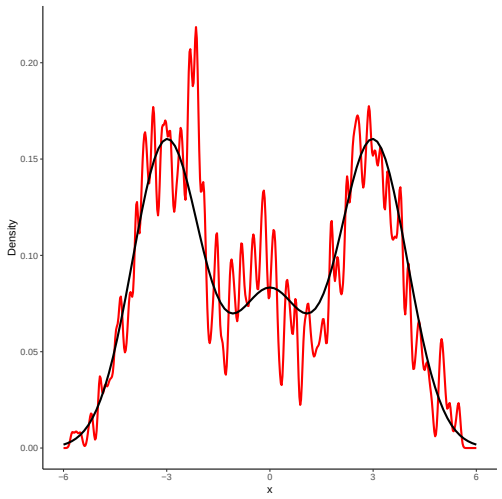
There are multiple ways to define a criterion for how optimal the choice of bandwidth is, and many methods are used to calculate a suitable bandwidth. One of the most popular methods is "Silverman's rule of thumb"[4].

For multivariate data it is still possible to perform this procedure, but the difference is that instead of using a univariate kernel we now use a multivariate kernel, usually a symmetric normal density. Similarly to how we generalized the univariate normal density, the smoothing parameter h now becomes a matrix $\mathbf{H}$ and for the multivariate normal kernel this matrix plays the part of the covariance. For a more detailed look into density estimation we refer to Silverman's book about the subject[4].

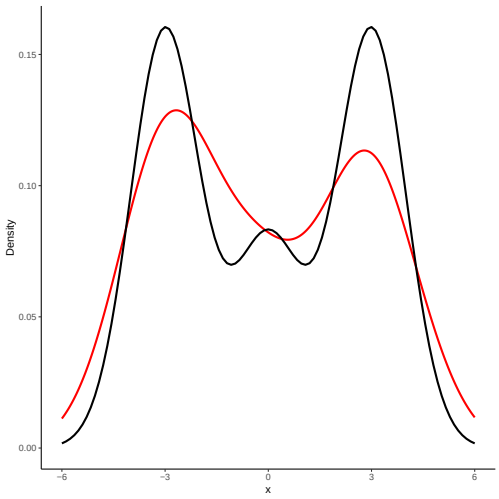
2.5 Discriminant analysis and classification

In the following section we will look at a couple of methods for discrimination and classification. Discrimination is the process of describing the features of observations that differentiate between populations and using these features to separate these populations as much as possible. This makes discrimination an excellent tool for data exploration and also for dimensionality reduction, by looking at which features separate our populations as effectively as possible. After separating our populations the next goal could be to sort new observations into labeled classes corresponding to the separated populations, this is what classification procedures try to accomplish. Thus, classification procedures are less about exploration and rather determine rules for optimally assigning new observations.

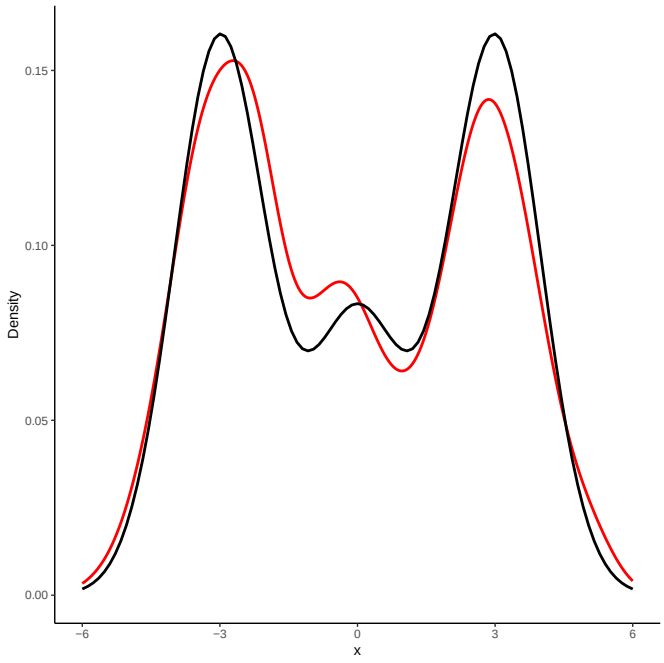
Before we start the separation we must be aware that it is usually impossible to define a perfect classification rule and always assigns an observation to the correct class. This is due to possible overlap between the densities for our populations. Knowing this we shall determine some features a "optimal" classification rule should have.



(a) h = 0.05



(b) h = 1



(c) h = 0.5

Figure 2.5: Density estimates of $f_Y(x)$ with varying bandwidth from a simulated sample of 1000 data points. Black: True density of f_Y. Red: KDE $\hat{f}_Y$

2.5.1 Expected cost of misclassification

First some notation. For the true populations, we will use ω, so for an object coming from the ith population we can write $\mathbf{x} \in \omega_i$ or simply ω_i, with corresponding

probability density function $f_i(\mathbf{x})$. We will then try to determine subsets of our sample space Ω, where observations in the region R_i correspond to us assigning our observation to be in population ω_i. We must assign the observation $\mathbf{x}$ to a unique population so the regions must be mutually exclusive and exhaustive, meaning for g populations

$$\bigcap_{i=1}^{g} R_i = \varnothing \qquad\qquad \bigcup_{i=1}^{g} R_i = \Omega$$

And now we can define the probability of misclassification, which we will denote as $P(i \mid k)$, classifying the object as belonging to ω_i when it is really from ω_k. So for the case when $g = 2$

$$P(2 \mid 1) = P(\mathbf{X} \in R_2 \mid \omega_1) = \int_{R_2} f_1(\mathbf{x})\, d\mathbf{x}$$

$$P(1 \mid 2) = P(\mathbf{X} \in R_1 \mid \omega_2) = \int_{R_1} f_2(\mathbf{x})\, d\mathbf{x}$$

Any classification rule we would consider using should make few misclassifications, meaning that the probability of a misclassification should be small. Our rule should take into account the prior probability of an observation coming from a certain population, we will denote this as p_i for the prior probability of coming from population ω_i. This makes it possible to simply state the probabilities of misclassification as the probability of misclassifying the observation, as seen above, times the probability of the observation coming from the true population.

$$P(\text{observation is misclassified as } \omega_1) = P(\mathbf{X} \in R_1 \mid \omega_2)P(\omega_2) = P(1 \mid 2)\, p_2$$

$$P(\text{observation is misclassified as } \omega_2) = P(\mathbf{X} \in R_2 \mid \omega_1)P(\omega_1) = P(2 \mid 1)\, p_1$$

We could now optimize the sum of these probabilities and find the classification rule that minimizes the total probability of misclassification, but this is not always the optimal solution. In cases where the direction of misclassification can be crucial we can define a cost of misclassification, we will denote this as $c(i \mid k)$ for the cost of incorrectly classifying a object from ω_k as coming from ω_i.

From this we define the expected cost of misclassification for a observation from ω_k

$$\text{ECM}(k) = \sum_{\substack{i=1 \\ i \neq k}}^{g} c(i \mid k)\; P(i \mid k)\;.$$

And now we have a good measure to optimize for, the total expected cost of misclassification (TECM) which is the sum of the expected cost of misclassification times the corresponding prior probability for each population

$$\begin{aligned}\text{TECM} &= p_1\text{ECM}(1) + p_2\text{ECM}(2) + \cdots + p_g\text{ECM}(g) \\ &= \sum_{k=1}^{g} p_k \left(\sum_{\substack{i=1 \\ i\neq k}}^{g} c(i \mid k)\ P(i \mid k) \right)\end{aligned}$$

And the regions that minimize the ECM are defined by allocating $\mathbf{x}$ to the population ω_k for which

$$\sum_{\substack{i=1 \\ i\neq k}}^{g} c(i \mid k)\ f_k(\mathbf{x})\ p_k$$

is the smallest. Further we will only look at the case of $g = 2$ for simplicity. Then the TECM is

$$\text{TECM} = c(2 \mid 1)p_1 \int_{R_2} f_1(\mathbf{x})\, d\mathbf{x} + c(1 \mid 2)p_2 \int_{R_1} f_2(\mathbf{x})\, d\mathbf{x}$$

when writing $P(1|2)$ and $P(2|1)$ as integrals. Because $\Omega = R_1 \cup R_2$ we can rewrite one of the integrals as

$$\int_{R_2} f_1(\mathbf{x})\, d\mathbf{x} = 1 - \int_{R_1} f_1(\mathbf{x})\, d\mathbf{x}\,.$$

This allows us to write

$$\begin{aligned}\text{TECM} &= c(2 \mid 1)p_1 \left[1 - \int_{R_1} f_1(\mathbf{x})\, d\mathbf{x} \right] + c(1 \mid 2)p_2 \int_{R_1} f_2(\mathbf{x})\, d\mathbf{x} \\ &= \int_{R_1} \Big[\, c(1 \mid 2)p_2 f_2(\mathbf{x}) - c(2 \mid 1)p_1 f_1(\mathbf{x}) \,\Big]\, d\mathbf{x} + c(2 \mid 1)p_1,\end{aligned}$$

and because all of these variables are non-negative for all $\mathbf{x}$, the TECM is minimized if R_1 includes those values $\mathbf{x}$ for which

$$c(1 \mid 2)p_2 f_2(\mathbf{x}) - c(2 \mid 1)p_1 f_1(\mathbf{x}) \leq 0$$

and excludes those $\mathbf{x}$ for which this quantity is positive. This can also be written with the ratio of densities as follows

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{c(1 \mid 2)}{c(2 \mid 1)}\ \frac{p_2}{p_1}\,.$$

2.5.2 Linear discriminant analysis

Next we assume that $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ are multivariate normal densities, with mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ respectively, we shall first consider the case of equal covariance $\boldsymbol{\Sigma}$. Then the normalization factors cancel and the regions R_1 and R_2 the minimize the TECM are

$$R_1: \quad \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)+\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right] \geq \frac{c(1\mid 2)}{c(2\mid 1)}\,\frac{p_2}{p_1}$$

$$R_2: \quad \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)+\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right] < \frac{c(1\mid 2)}{c(2\mid 1)}\,\frac{p_2}{p_1}\,.$$

Now that we have determined the regions R_1 and R_2 we can define our classification rule, where we have a new observation $\mathbf{x}_0$ and wish to allocate it to either ω_1 or ω_2. The following formulation is simplified from the above expressions by taking the logarithm, which we can do because all the terms are non-negative. When we expand the products in the exponential all the quadratic terms $\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x}$ cancel and we receive a linear function in $\mathbf{x}$. We allocate $\mathbf{x}_0$ to ω_1 if

$$(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\mathbf{x}_0-\frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)\geq \ln\left(\frac{c(1\mid 2)}{c(2\mid 1)}\,\frac{p_2}{p_1}\right)\,, \tag{2.3}$$

and we allocate $\mathbf{x}_0$ to ω_2 otherwise. This rule can also be interpreted as

$$y\geq \ln\left(\frac{c(1\mid 2)}{c(2\mid 1)}\,\frac{p_2}{p_1}\right)+m\,, \tag{2.4}$$

where $y=(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\mathbf{x}=\mathbf{a}'\mathbf{x}$ and $m=\frac{1}{2}(\mathbf{a}'\boldsymbol{\mu}_1+\mathbf{a}'\boldsymbol{\mu}_2)$.

When the expression in the logarithm is unitary we are left with comparing the scalar variable y against m, this is known as "Fisher's linear discriminant". This is equivalent to projecting the points onto a line and checking if it falls to the right or left of the midpoint m.

2.5.3 Quadratic discriminant analysis

As seen above, when our populations are normally distributed and have equal covariance we get a nice linear function separating our populations, this is not

the case when considering unequal covariance. Following the same procedure as before, assuming the populations are normally distributed with different covariance matrices, by finding the ratio of densities and simplifying we find the classification regions

$$\begin{aligned} R_1: \quad & -\frac{1}{2}\mathbf{x}'(\boldsymbol{\Sigma}_1^{-1}-\boldsymbol{\Sigma}_2^{-1})\mathbf{x}+(\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1}-\boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2^{-1})\mathbf{x}-k \\ & \geq \ln\left(\frac{c(1\mid 2)}{c(2\mid 1)}\,\frac{p_2}{p_1}\right) \\ R_2: \quad & -\frac{1}{2}\mathbf{x}'(\boldsymbol{\Sigma}_1^{-1}-\boldsymbol{\Sigma}_2^{-1})\mathbf{x}+(\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1}-\boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2^{-1})\mathbf{x}-k \\ & < \ln\left(\frac{c(1\mid 2)}{c(2\mid 1)}\,\frac{p_2}{p_1}\right) \end{aligned}$$

where

$$k=\frac{1}{2}\ln\left(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}\right)+\frac{1}{2}(\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2).$$

Notice how the first term is a quadratic form, thus the regions are defined by quadratic functions of $\mathbf{x}$. This term disappears when $\boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}_2$ and the rest of the equation reduces to 2.3.

In practice, the covariances and mean vectors are usually unknown, so for all of the discussed classification rules the estimated minimum TECM rule can be found by substituting the unknown statistics with their sample variants.

Chapter 3

Analysis

Now we shall demonstrate some of the methods we have looked at and see how effective even the linear classification rule can be in practice.

3.1 Data

The data used in this analysis is from a voluntary online questionnaire in the Faculty of Science and Technology course STA100 at the University of Stavanger. The questionnaire included questions about the participants height, shoe size, gender and more, but these are the variables we will be interested inn for demonstration. The height is measured in centimeters, shoe size in the EU continental system and the gender as one of the character strings "Male" or "Female". The results from the questionnaire for the years 2019 to 2023 have been included totaling in 1541 data points where 406 of which had the gender set as "Female" and 1135 as "Male".

3.1.1 Some changes to suspect data points

There were some data points in the original data set which has been either removed or changed, including a female with a shoe size of 385 which has been changed to 38.5. Two males with shoe size 4, both of these have been removed. And lastly a female with a height of 195cm and 48 in EU shoe size, this is such a extreme outlier that it is assumed that it is a male that has either set the wrong gender by accident or done so as a joke, therefore the gender was changed to "Male". A scatter plot of the data is illustrated in Figure 3.1 with some noise added to counteract the discretization.

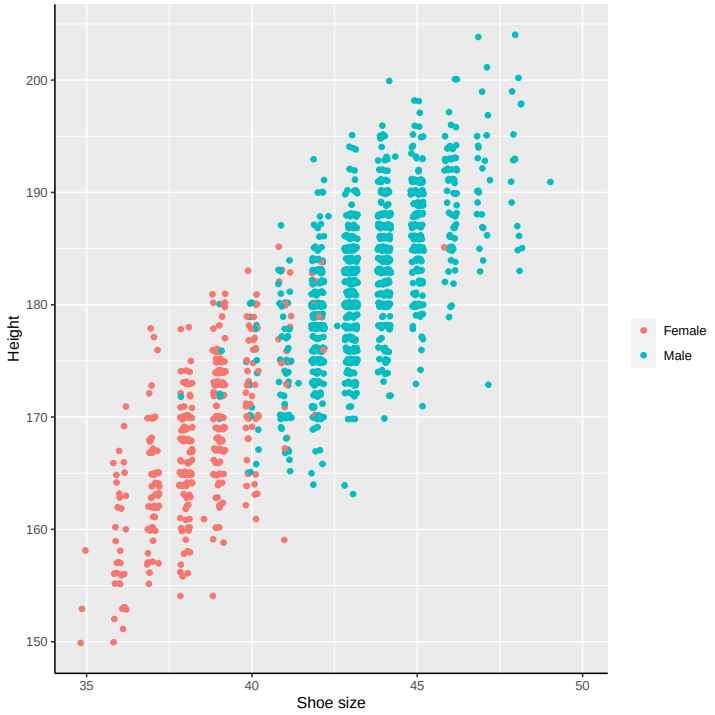


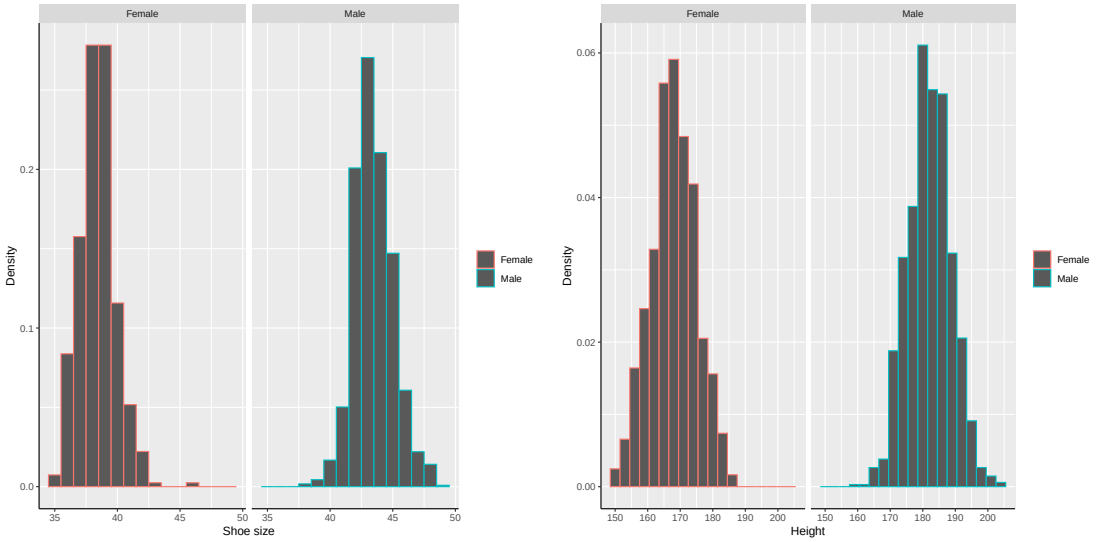
Figure 3.1: Scatter plot of the data with some noise added to the position

3.1.2 Assessing normality

It is generally that height and shoe size are approximately normally distributed, but we will still make a short attempt to justify this assumption. For a more detailed discussion about assessing the assumption of normality see [3]. We will first start with histograms for the shoe sizes and heights, separated by gender as we will assume that each gender is its own population.

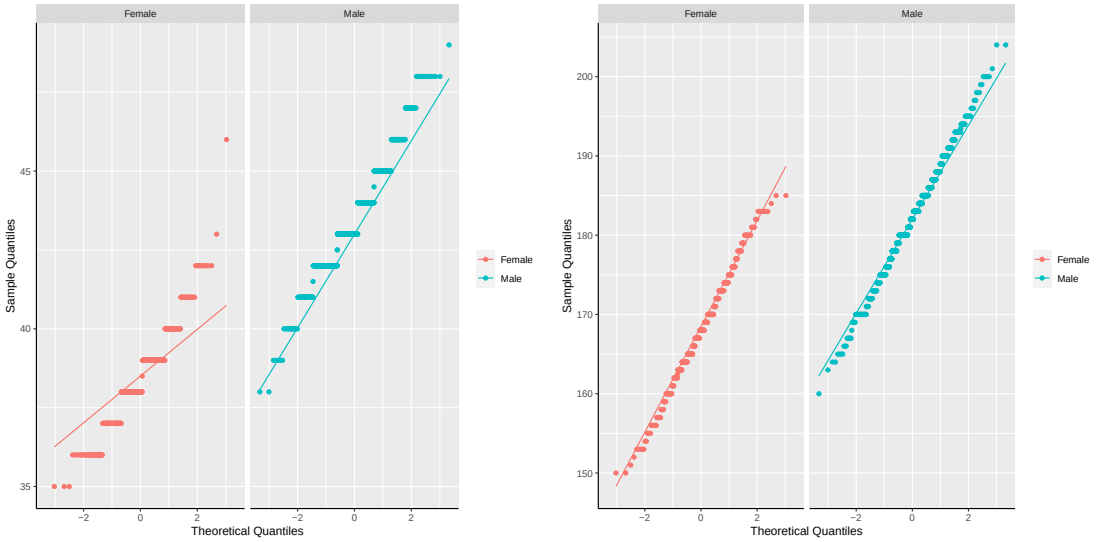
Along with these histograms we can also make Q-Q plots for each variable, here we will see the discretization that happens from most of the inputs being rounded to the nearest integer or half integer, which is completely reasonable as it is understandable that the participants don't specify their height or shoe size with multiple decimals of precision.

There seems to be deviation from normality in the female shoe sizes, and it looks like this is due to having heavier tails than a normal distribution, fortunately this does not really pose a large problem for our purposes. From this we will conclude that the shoe size and height for the males are multivariate normal and for the females we will approximate it as such, knowing it deviates slightly from normality for the shoe sizes.



(a) Histogram of shoe sizes (b) Histogram of heights

Figure 3.2: Histograms



(a) Quantile plot of shoe sizes (b) Quantile plot of heights

Figure 3.3: Quantile plots

3.2 Estimation

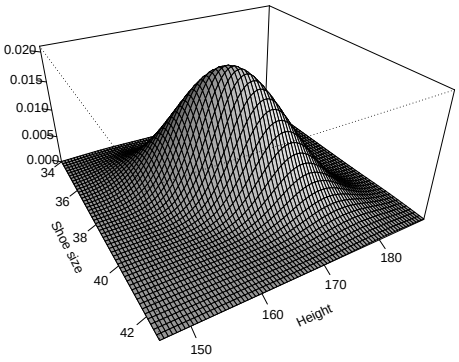
Next we will assume that the shoe size and height for females are distributed as $N_2(\boldsymbol{\mu}_F, \boldsymbol{\Sigma}_F)$ and the males as $N_2(\boldsymbol{\mu}_M, \boldsymbol{\Sigma}_M)$ and estimate the various parameters.

This is is done by using the formulas presented in section 2.2.2.

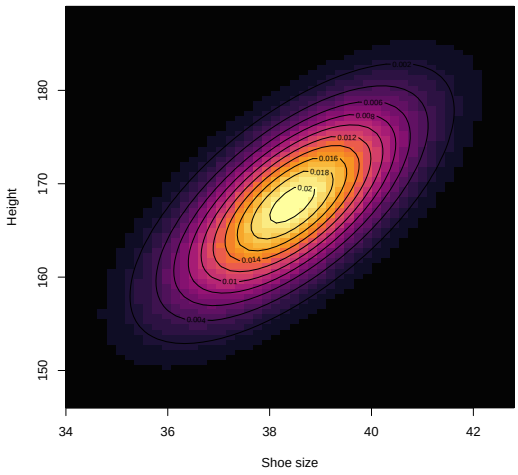
$$\hat{\boldsymbol{\mu}}_F = \bar{\mathbf{x}}_F = \begin{bmatrix} 38.44 \\ 167.9 \end{bmatrix} \qquad \hat{\boldsymbol{\Sigma}}_F = \mathbf{S}_F = \begin{bmatrix} 2.153 & 6.740 \\ 6.740 & 47.86 \end{bmatrix}$$

$$\hat{\boldsymbol{\mu}}_M = \bar{\mathbf{x}}_M = \begin{bmatrix} 43.47 \\ 182.1 \end{bmatrix} \qquad \hat{\boldsymbol{\Sigma}}_M = \mathbf{S}_M = \begin{bmatrix} 2.525 & 6.819 \\ 6.819 & 44.28 \end{bmatrix}$$

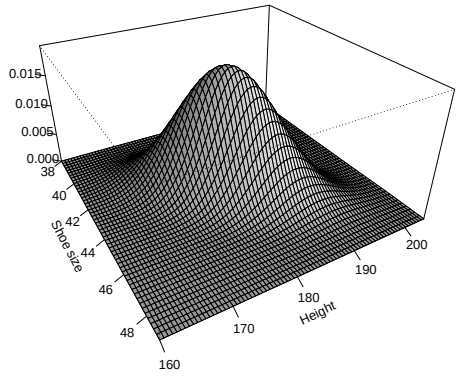
This gives us the two densities, illustrated with perspective plots and heat maps in Figure 3.4.



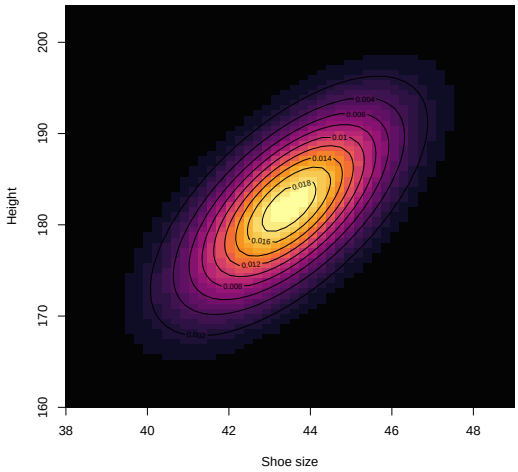
(a) Perspective plot of $N_2(\bar{\mathbf{x}}_F, \mathbf{S}_F)$



(b) Heat map of $N_2(\bar{\mathbf{x}}_F, \mathbf{S}_F)$



(c) Perspective plot of $N_2(\bar{\mathbf{x}}_M, \mathbf{S}_M)$



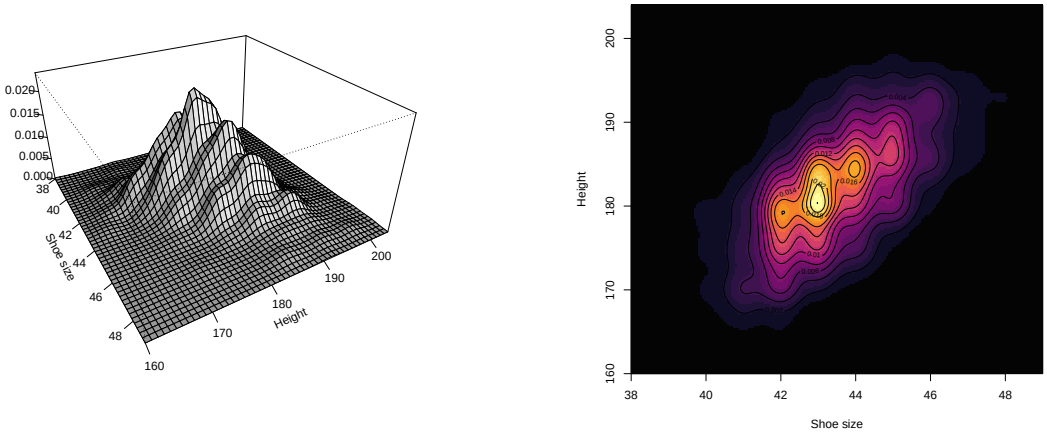
(d) Heat map of $N_2(\bar{\mathbf{x}}_M, \mathbf{S}_M)$

Figure 3.4: Various plots of the estimated normal densities

This is of course not the only way to estimate the densities, we can compare this result to the one we get from non-parametric kernel estimation. For the plots in Figure 3.5 we will only use the male data to demonstrate. If the discretization is a key feature that we wish to include in our estimate then we must choose a

bandwidth that manages to detect this, one example is the bandwidth

$$\mathbf{H} = \begin{bmatrix} 1.55 & 0 \\ 0 & 6.2 \end{bmatrix}.$$



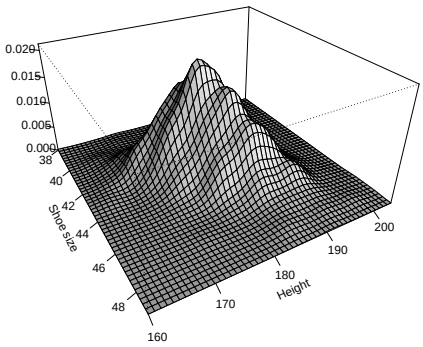
(a) Perspective plot of $\hat{f}_M(\mathbf{x})$ (b) Heat map of $\hat{f}_M(\mathbf{x})$

Figure 3.5: KDE $\hat{f}_M(\mathbf{x})$ for male data with $\mathbf{H} = \begin{bmatrix} 1.55 & 0 \\ 0 & 6.2 \end{bmatrix}$

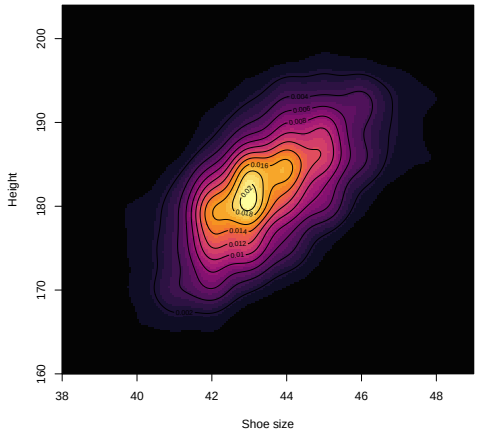
But now we want to compare with our normal estimate $N_2(\bar{\mathbf{x}}_M, \mathbf{S}_M)$ in Figure 3.4c and 3.4d, so we choose the matrix $\mathbf{H} = \begin{bmatrix} 1.8 & 0 \\ 0 & 7 \end{bmatrix}$ as our smoothing parameter and see in Figure 3.6 that the density we get from kernel estimation is similar to the one we get from assuming normality and estimating the parameters $\boldsymbol{\mu}_M$ and $\boldsymbol{\Sigma}_M$.

Another point of interest would be test how the rounding to integers effects the normal distribution, so here we have simulated 1135 data points from $N_2(\bar{\mathbf{x}}_M, \mathbf{S}_M)$ and then rounded the values to the closest integer before kernel estimation with $\mathbf{H} = \begin{bmatrix} 1.55 & 0 \\ 0 & 6.2 \end{bmatrix}$.

When comparing Figure 3.7 with Figure 3.5 we can see how similar these two plots are, this gives us more confidence in the assumption that the shoe size and height of males are multivariate normal. The true distribution of shoe sizes is probably normal, and due to the discretization and the small range of values it

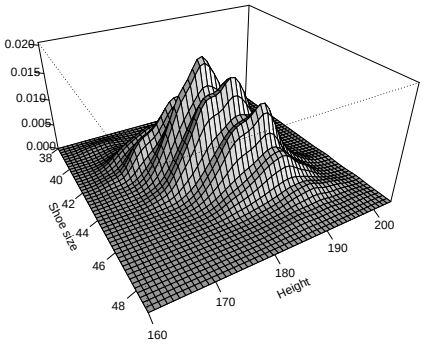


(a) Perspective plot of $\hat{f}_M(\mathbf{x})$

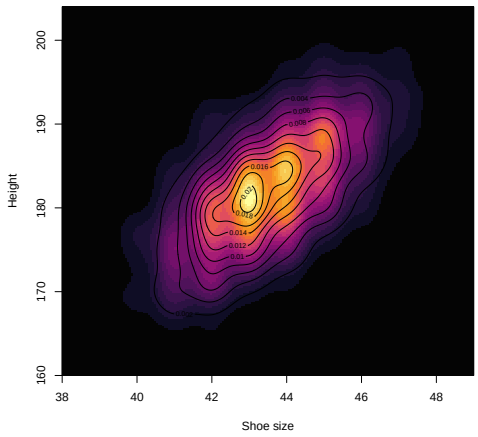


(b) Heat map of $\hat{f}_M(\mathbf{x})$

Figure 3.6: KDE $\hat{f}_M(\mathbf{x})$ for male data with $\mathbf{H} = \begin{bmatrix} 1.8 & 0 \\ 0 & 7 \end{bmatrix}$



(a) Perspective plot of the simulated density



(b) Heat map of the simulated density

Figure 3.7: KDE of simulated $N_2(\bar{\mathbf{x}}_M, \mathbf{S}_M)$ with rounding and $\mathbf{H} = \begin{bmatrix} 1.55 & 0 \\ 0 & 6.2 \end{bmatrix}$

can make it difficult to confirm this, but as we saw in Figure 3.7 we were able to recreate the result by simulating values from a normal distribution.

3.3 Classification

Now we want to find a classification rule where we can specify the shoe size and height of a individual and determine their gender. This will be done with the linear classification rule as outlined in section 2.5.

For the linear rule we assume equal covariance for each of our multivariate normal populations. As we saw in the previous section, the assumption of normality seems reasonable, and the difference between our estimates for $\mathbf{\Sigma}_F$ and $\mathbf{\Sigma}_M$ is small, so its also reasonable to assume equal covariance. Thus we will find the pooled variance $\mathbf{S}_{\text{pooled}}$

$$\mathbf{S}_{\text{pooled}} = \frac{(n_F - 1)\mathbf{S}_F + (n_M - 1)\mathbf{S}_M}{n_F + n_M - 2} = \begin{bmatrix} 2.427 & 6.798 \\ 6.798 & 45.23 \end{bmatrix}$$

$$\implies \mathbf{S}_{\text{pooled}}^{-1} = \begin{bmatrix} 0.7118 & -0.1070 \\ -0.1070 & 0.03819 \end{bmatrix}$$

and use this as our estimate for the covariance.

The costs of misclassification will be assumed to be equal. The prior probabilities are unknown for our populations, which means that these must be estimated. We will the take proportion of females in the data sett as our estimate for the prior probability for the female population p_F, and equivalently for the prior probability for the male population p_M.

$$\hat{p}_F = \frac{406}{1541} \approx 0.263 \qquad\qquad \hat{p}_M = \frac{1135}{1541} \approx 0.737$$

Using the formulation in equation 2.4, we find a vector $\hat{\mathbf{a}}_1$

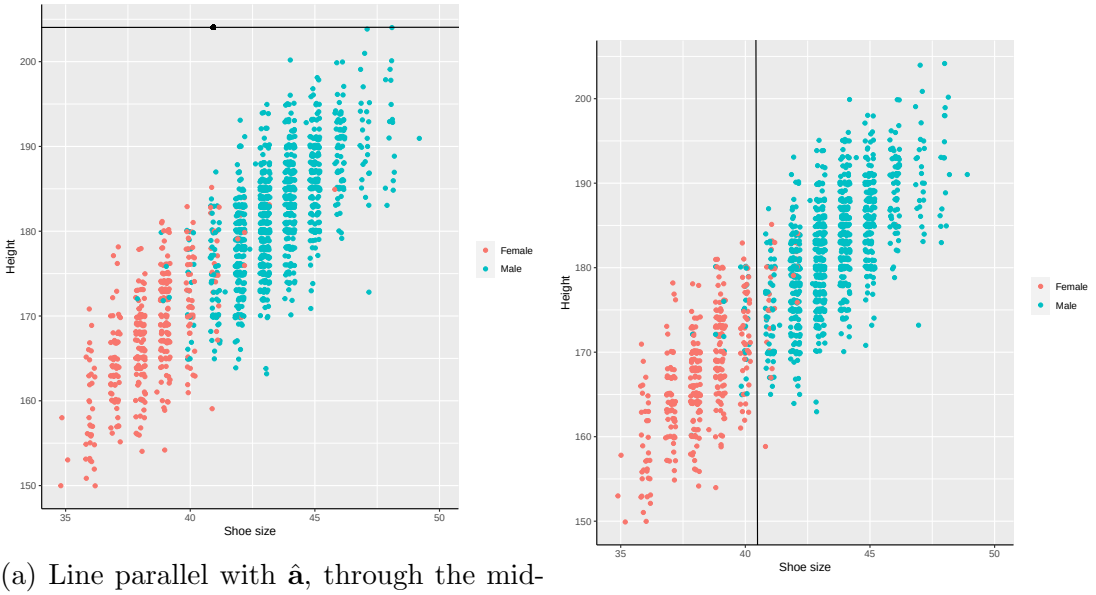
$$\hat{\mathbf{a}}_1' = (\bar{\mathbf{x}}_F - \bar{\mathbf{x}}_M)\mathbf{S}_{\text{pooled}}^{-1} = \begin{bmatrix} -2.07 & -0.00224 \end{bmatrix}$$

which is unique up to a multiplicative constant, so we will normalize and change the sign

$$\hat{\mathbf{a}} = \frac{-1}{|\hat{\mathbf{a}}_1|}\hat{\mathbf{a}_1} = \begin{bmatrix} 1.00 \\ 0.00108 \end{bmatrix}.$$

Next is determining $\hat{m}$, this is calculated directly now that we have found $\hat{\mathbf{a}}$.

$$\hat{m} = \frac{1}{2}\left(\hat{\mathbf{a}}'\bar{\mathbf{x}}_F + \hat{\mathbf{a}}'\bar{\mathbf{x}}_M\right) = 41.1$$



(a) Line parallel with $\hat{\mathbf{a}}$, through the midpoint $\hat{m}$

(b) Line separating R_F and R_M

Figure 3.8: Plots of the line parallel to $\hat{\mathbf{a}}$ and the line separating our regions

Thus we have found the line to where we project the data points and classify the points falling to the left of the midpoint $\hat{m}$ as female and the points to the right as male, this line is illustrated in Figure 3.8a and when considering this line as one of axes we will maximize the separation between the populations. This can also be illustrated with a line separating our two regions as in Figure 3.8b where the formula for our line will be $\hat{\mathbf{a}}'\mathbf{x} = \hat{m} - \frac{1}{|\hat{\mathbf{a}}_1|} \ln\left(\frac{0.737}{0.263}\right)$ in this case our separation line becomes $x_1 + 0.00108x_2 = 40.6$.

We might notice that for the discrimination the shoe size is the better variable to use, so for a practical setting it would not be a problem to only measure the shoe size to classify the individual, we can also see this in the vector $\hat{\mathbf{a}}$.

The final classification rule is expressed in terms of the scalar $\hat{y}_0 = \hat{\mathbf{a}}'\mathbf{x}_0$ as follows. Allocate the observation $\mathbf{x}_0$ to the female population if

$$\hat{y}_0 \leq \hat{m} - \frac{1}{|\hat{\mathbf{a}}_1|} \ln\left(\frac{0.737}{0.263}\right) = 40.6\,,$$

or equivalently

$$\hat{\mathbf{a}}'\mathbf{x} = x_1 + 0.00108x_2 \leq 40.6\,,$$

otherwise allocate $\mathbf{x}_0$ to the male population.

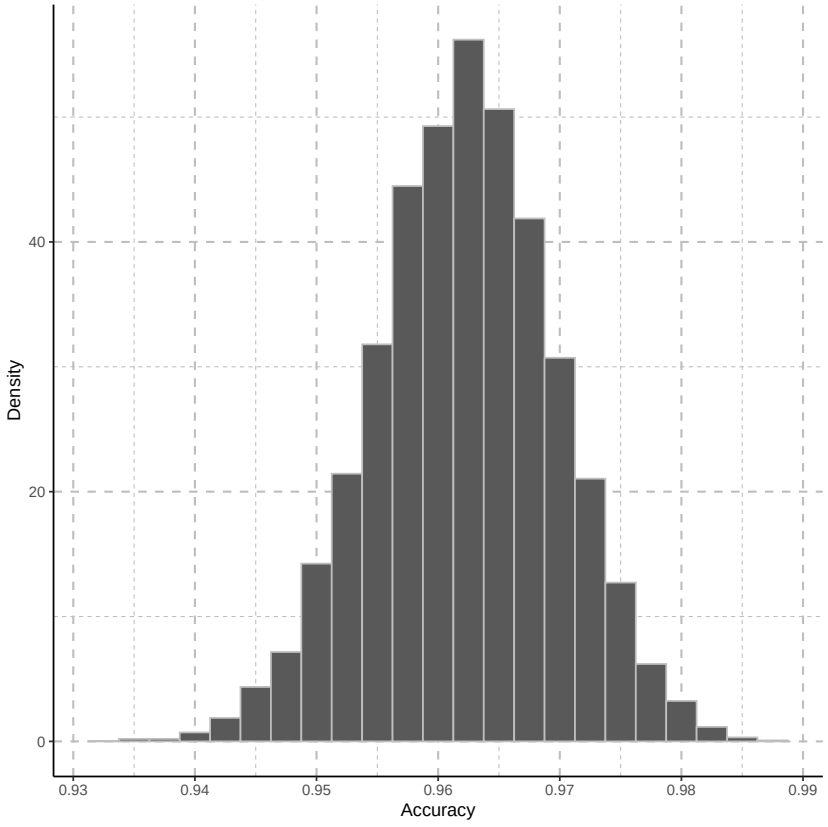


Figure 3.9: Histogram of estimated accuracy

3.3.1 Evaluating the classification rule

One common method used to evaluate the classification rule is to split the data set into two parts, one that is used for making the model and one that is used to test it. We have done so, splitting 70% into a training set and 30% for testing, and repeating this 10 000 times to make a 95% confidence interval for the accuracy of the classification rule. Through this procedure the point estimate for the accuracy is 0.962 with a 95% confidence interval of $[0.947, 0.977]$, illustrated in the histogram below.

There are more methods to evaluate the accuracy of classification rules, as well as other measures to evaluate on, such as the apparent error rate.

Other methods could be used to classify our two populations, such as Fishers linear discriminant or the quadratic classification rule. For Fishers method the only change is that the prior probabilities are assumed to be equal, for our case,

this will shift our separation line further to the right, this could be used if our populations were not the female and male students at this specific faculty where the ratio of female to male is skewed, but rather some equivalent population with a ratio of female and male students that is closer to one. As for the quadratic rule, the result does not drastically change from the linear rule and this is due to the small difference in our covariance matrices and having well-separated populations.

Chapter 4

Summary

Throughout this work, we have seen how univariate statistics generalizes to vectors of random variables with a focus on the multivariate normal distribution. For the multivariate normal distribution, we have looked in section 2.3 at some properties, maximum likelihood estimation, and lastly simulation of multivariate normal variables in section. Further, we introduced, in section 2.4, non-parametric kernel estimation and demonstrated the effect of different values of the smoothing parameter. In section 2.5 we built methods for discrimination and then the classification of new observations by defining rules such as the quadratic classification rule. In the final chapter, we demonstrated many of these methods on a real data set and discussed various challenges and the validity of our assumptions.

Bibliography

[1] Anderson, T. W. (2003). *An introduction to multivariate statistical analysis* (3rd ed.). Wiley series in probability and statistics. Hoboken, N.J: Wiley.

[2] Box, G. E. P. and M. E. Muller (1958, June). A note on the generation of random normal deviates. *The Annals of Mathematical Statistics 29*(2).

[3] Johnson, R. A. and D. W. Wichern (2014). *Applied multivariate statistical analysis* (6th ed.). Harlow: Pearson.

[4] Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Monographs on statistics and applied probability. London: Chapman and Hall.