Feature-based models for forensic likelihood ratio calculation

Supporting research for the ENFSI-LR project

F.S. Kool

Master of Science Thesis

TU Delft Delft University of Technology

Challenge the future

Feature-based models for forensic likelihood ratio calculation

Supporting research for the ENFSI-LR project

by

Fréderique Suzanne Kool

Delft, November 2016

A thesis submitted to the
Delft Institute of Applied Mathematics
in partial fulfillment of the requirements

for the degree

Master of Science
in
Applied Mathematics

Supervisor:	Prof.dr.ir. G. Jongbloed	
Thesis committee:	Prof.dr.ir. G. Jongbloed,	TU Delft
	Dr. D. Kurowicka,	TU Delft
	Dr. A. Bolck,	NFI

Delft University of Technology
Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS)

TU Delft Delft University of Technology

Nederlands Forensisch Instituut
Ministerie van Veiligheid en Justitie

Abstract

The likelihood ratio is a generally accepted measure for the strength of evidence in forensic comparison problems. These problems concern comparisons where it is investigated whether at least two items come from the same source or not, e.g. whether the DNA on the crime scene comes from the suspect or not. The use of likelihood ratios by forensic experts in practical forensic casework demands for a unified system to compute likelihood ratios. Therefore, the EU funded the "ENFSI-LR" project that aims to construct software which helps forensic experts to calculate likelihood ratios based on validated scripts and harmonized models. In this thesis some problems concerning the ENFSI-LR project are addressed. Solutions to these problems are useful for unification, validation and (future) development of the software. Throughout this thesis, the emphasis is on continuous two-level feature-based models. In the literature, these underlying models have led to two likelihood ratio formulas. In this thesis it is proved that these two formulas are exactly the same. This thesis also explores several parameter estimation methods for the two-level model. Standard estimation methods are compared with estimation methods which have not been used in forensic statistics until now: a generalized weighted mean or maximum likelihood estimation. As an extension of existing feature-based models, a model is introduced that combines discrete- and continuous evidence into one likelihood ratio.

Contents

1

Introduction

> ***Warning: dangerous xtc tablets at Amsterdam Dance Event (ADE)***
> *Wednesday the ADE started in Amsterdam. The Trimbos Insitute warns visitors for a deadly xtc tablet that contains a very high dosage of PMMA. This dangerous tablet is pink and has a Superman logo on both sides.*
> – Volkskrant, 21 October 2016

Production and trafficking of xtc tablets is a significant problem in Europe, because of several reasons such as the health risks as described in the news article above. Forensic science is a useful tool in the fight against this problem. For instance, suppose that the police seizes two consignments of pink Superman xtc tablets at the ADE. The question of interest in court is whether the seized consignments are from the same source or not. The task for forensic drug experts is to evaluate the evidence (e.g. both consignments contain pink tablets with a Superman logo and a high dosage of PMMA) and to provide the judge with the strength of this evidence. In forensic science the likelihood ratio is a generally accepted measure to evaluate the strength of evidence.

This specific example of the comparison of seized consignments of xtc tablets falls within a broader class of forensic comparison problems. In these comparison problems forensic scientists always compare whether at least two items (e.g. consignments of xtc tablets) come from the same source or not. Other examples include the comparison of a shoe print at the crime scene with a shoe of the suspect or the comparison of a bloodstain on the crime scene with the blood of the suspect. Although forensic experts agree about the use of the likelihood ratio approach in such forensic comparison problems, the calculation of the likelihood ratio is not (yet) unified in forensic science.

Various forensic statisticians in different countries have developed statistical models for the computation of likelihood ratios. Some of them have written (non-validated) scripts that enables forensic experts to compute likelihood ratios in their casework. The use of likelihood ratios in practice demands for a unified system to compute likelihood ratios. To investigate and harmonize the statistical models and existing software, the European Union funded a two year project for the European network of forensic science institutes (ENFSI), called the " ENFSI-LR" project. The aim of this project is to construct a Graphical User Interface (GUI) around software that helps forensic experts to calculate likelihood ratios based on validated scripts and harmonized models.

In this thesis some problems concerning the ENFSI-LR project are addressed, such as unifying different likelihood ratio approaches and exploring different methods for parameter estimation. Before these problems are addressed, the underlying models within the likelihood ratio approach are explained. Furthermore, some suggestions and extensions for future development of the project will be given. As a running example

throughout this thesis, xtc tablet comparison problems will be used. However, it is important to note that the theory can be applied to a great diversity of fields of expertise within forensic science.

Thesis outline

In order to obtain a full understanding of problems in forensic statistics, Chapter 2 will serve as an introduction to forensic evidence evaluation. The chapter starts off with forensic problems that can be approached with a likelihood ratio. The particular example of xtc tablet comparison will be used throughout this thesis. The chapter is concluded by describing how the likelihood ratio can be used to evaluate evidence.

The purpose of Chapter 3 is to describe existing discrete- and continuous models that can be used to calculate the likelihood ratio. In this thesis the focus will be on the continuous evidence models described in Section 3.2. In these continuous evidence models the data is modelled as a two-level model. In forensic literature modelling these two levels are known as within-source variation and between-source variation. In forensic statistics there exist two types of two-level models. In this thesis these models will be referred to as Gaussian two-level models and non-Gaussian two-level models.

In Chapter 4 the Gaussian two-level model is described and two explicit likelihood ratio formulas are derived. One of the goals of this chapter is to show that these two formulas are exactly the same. This is important in the ENFSI-LR project, for one of the objectives of the project is to agree upon likelihood ratio formulas. Furthermore, the equivalence of the two likelihood ratio formulas is important for the validation of the implemented likelihood ratio in the software.

The likelihood ratio formula in Chapter 4 depends on unknown parameters of the two-level Gaussian model. In Chapter 5 different methods to estimate these parameters are explored. The exploration of parameter estimation for Gaussian two-level models is important for the ENFSI-LR project, since the software must contain "simple" plug-in estimators as default choice. Other estimators will be implemented as optional choices. Currently, forensic statisticians are discussing which plug-in estimator should be used for the mean parameter. These estimators are compared in this chapter. Furthermore, an alternative to these estimators is given. As an optional choice for the estimators, the EM-algorithm is suggested in this thesis as an iterative method to find the maximum likelihood estimates. The chapter ends with a comparison of these methods using a simulation study.

In practice, the assumption of a Gaussian two-level model is often not valid. Therefore in forensic statistics the non-Gaussian two-level model, which uses a non-parametric estimation technique, is often applied. Because of its importance in forensic statistics and consequently in the developed ENFSI-LR software, this model is described in Chapter 6. Furthermore, some difficulties in this model are introduced which can be investigated in the future.

Chapter 7 introduces an extension of the models described in Section 3. Up to now forensic experts could only report two separate likelihood ratios, one based on discrete evidence and the other based on continuous evidence. The objective of this chapter is to describe a model that can be used to combine the discrete and continuous evidence into one likelihood ratio. In addition, in Chapter 7 the described methods in this thesis and the extended model of Chapter 7 are applied to real xtc data.

2

Forensic evidence evaluation

This chapter will give an introduction to forensic evidence evaluation. Section 2.1 describes a forensic comparison problem that will form the basis of the theory throughout this thesis. Section 2.2 describes the use of probability within such problems. This section introduces the likelihood ratio as fundamental ingredient in forensic statistics.

2.1 A forensic comparison problem

The last decades, the impact of forensic science on investigation and evidence evaluation in criminal cases became more apparent because of increasing media attention. For example, Figure 2.1(a) illustrates an investigation problem that aims to find a similarity between DNA-profiles. Figure 2.1(b) shows an example of evidence evaluation in a criminal case where glass pieces found on the suspect are compared to a car window at the crime scene.

Suspect arrested after DNA-match at NFI in criminal case of sexual abuse.

(a) A 29-year old man is arrested for sexual abuse of a 60-year old woman in her house. During technical research at the crime scene, DNA was secured and transmitted to the Netherlands Forensic Institute for further examination. The NFI found a match between the DNA-profile of the suspect and the DNA which was found at the crime scene. (Blik op nieuws (2016)).

Beating felonies using a data base containing glass piece information

Devastation after a ram raiding in Rijsenhout in 2013 (ANP).

(b) In three criminal cases the police has tracked down suspects after an analysis of the glass pieces found at the crime scene. One of the cases is a double liquidation in the Staatsliederenbuurt in Amsterdam. The balaclava from one of the suspects contained a glass piece from the getaway car which has been shot at the crime scene. (NOS (2016)).

Figure 2.1: Translated summaries of two Dutch news items that illustrate problems in forensic science.

Next to the comparison of DNA-profiles and glass pieces, several other examples of

applications in forensic science exist, e.g. evidence based on shoe print comparison or fingerprint comparison. Despite the great diversity of fields of expertise, most criminal cases will share some similar aspects (Sjerps (2004)). For example, criminal investigation often concerns the comparison of at least two items. Figure 2.1(a) shows for instance the comparison of DNA of a suspect with DNA found on a crime scene. Figure 2.1(b) shows the comparison of a piece of glass found on the suspect with a window at the crime scene. In this thesis we emphasize this form of criminal investigation. For the police and court the question of interest is whether these items come from the same source. Hence, in such criminal cases it is common to consider two hypotheses. The *prosecutor's hypothesis* (H_p) proposes that the items come from the same source. The *hypothesis of the defense* (H_d) suggests that the items come from different sources. In this thesis emphasis is put on such hypotheses, which are called hypotheses on "source level". This can only indicate whether the items (glass pieces) have the same source (window) or not. Another possibility would be to consider hypotheses on "activity level": whether the suspect smashed the window or not. We could also consider the "crime level", whether the suspect committed the crime or not. To consider such hypotheses often more (unknown) information is needed. For example, there should be information about whether the number of pieces found on the suspect is reasonable when he smashed the window. If enough information is available, currently Bayesian networks are used to solve these problems (Aitken and Taroni (2004))

Considering the prosecutors hypothesis and the hypothesis of the defense, forensic experts are asked to evaluate the evidence, i.e. to compare items, given the two hypothesis. Hence, they are expected to answer the following question

"Under which hypothesis (H_p or H_d) is the evidence the most likely?"

To evaluate the evidence, the strength of observed similarities and differences between the items should be determined. In the late twentieth century a probabilistic framework to evaluate the strength of evidence was developed. This framework serves as a basis for this thesis. This will be further explained in Section 2.2.1.

In this thesis, the focus will be on criminal cases where two or more items will be compared under hypotheses as stated above. As a running example throughout this thesis, drug comparison will be used. However, it is important to note that the theory can be applied to many other examples as well. The next section will describe the forensic problem of xtc tablet comparison. After a sketch of this problem is given, the framework of forensic evidence evaluation will be described.

2.1.1 XTC tablets comparison

To fight against drug production and trafficking in Europe, comparison on samples of drug seizures is an important forensic tool (Koper et al. (2007)). First the problem of the production of xtc tablets will be discussed. Subsequently it will be described how comparison of drug samples can be used as a forensic tool.

A remaining problem

Europe is an important market for drugs, supported by both domestic production and drugs trafficked from other regions. For example, 3,4-methylenedioxymethamphetamine (MDMA) is one of the most widely spread synthetic drugs since the 1990s. MDMA tablets, also known as *xtc* or *ecstasy*, have always been popular MDMA products on the market. However, the last decade investigations performed by the European Monitoring Centre for Drugs and Drugs Addiction (EMCDA) have shown a decline in MDMA

production. In the 1990s and 2000s, xtc tablets had a low MDMA content. Moreover, a majority of the xtc tablets on European markets contained no MDMA at all (EMCDDA (2016b)).

However, a re-emerge of the production of xtc tablets which contain higher doses of MDMA started in 2011-2012. In Figure 2.2 it can be seen that at present over half of all xtc tablets contains over 140 mg of MDMA compared to just 3% in 2009 (EMCDDA (2016b)). Moreover, "super pills" with a MDMA purity between 270-340 mg are found on the market. In addition, there are variations in the dosage in similar looking tablets. As a consequence, in 2014 the EMCDDA and Europol have issued an alert warning of health risks linked to the consumption of tablets that contain a very high MDMA purity (EMCDDA (2016a)).

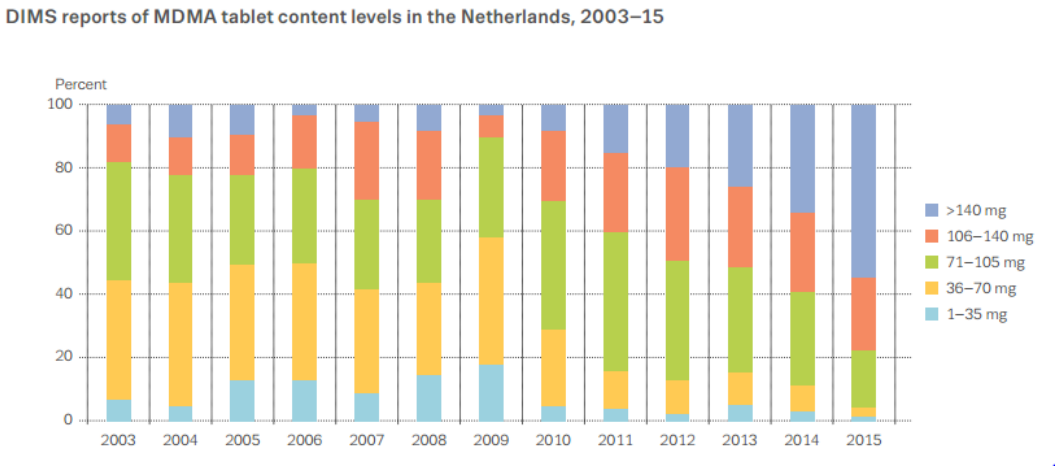


Figure 2.2: Drugs information and monitoring system (DIMS) reports of MDMA content levels in the Netherlands (EMCDDA (2016b)).

Production of MDMA in Europe appears to be concentrated around the Netherlands and Belgium, providing the largest production and producing higher purity products than elsewhere (EMCDDA (2016b)). Therefore, it may not come as a surprise that the highest mass loads of MDMA is found in the wastewater of Belgian and Dutch cities. Due to the increased production and higher purity of the tablets, wastewater MDMA loads are higher compared to the loads in 2011. The increase of dumping dangerous waste products from MDMA production processes has been reported by law enforcement agencies and is considered to be an environmental concern in the Netherlands and Belgium (EMCDDA (2016b)).

Due to both environmental concern and health risks, production and trafficking of xtc tablets remains a significant problem in Europe. To fight against this problem, comparison on samples of drug seizures can be used as a forensic tool.

Forensic comparison of drug seizures

In 2012, the Netherlands reported seizing 2.4 million xtc tablets (EMCDDA (2016a)). In most cases, the origin of the confiscated consignments with xtc filled bags is unknown. However, suspicion of links between different consignments may exist. For example, if tablets are found in the same type of bags this can indicate that the tablets originate from the same source. The court is interested in whether the tablets of two

consignments C_1 and C_2 come from the same source or not, i.e.

$$\begin{cases} H_p: & \text{Tablets of the consignments } C_1 \text{ and } C_2 \text{ come from the same source.} \\ H_d: & \text{Tablets of consignments } C_1 \text{ and } C_2 \text{ come from different sources.} \end{cases}$$

To investigate whether consignments have a common source, similarities and differences in the characteristics of the tablets are being examined.

For xtc tablets these characteristics can be distinguished in *pre-tabletting characteristics* and *post-tabletting characteristics*. This distinction is due to the two-stage production process of xtc tablets (Koper et al. (2007), Weyerman et al. (2008)). The first stage (pre-tabletting) is a synthesis process that creates a mixed powder that contains for instance the active substance MDMA. Next to this chemical composition the powder consist of impurities which arise during the synthesis process. The magnitude of the impurities can differ depending on the raw materials that are used, for instance. The chemical composition, dosage and impurities of the synthetic drug are thus considered to be pre-tabletting characteristics.[1] The impurities are more distinctive in the synthesis process than the chemical composition and dosage, because most dealers aim at more or less the same composition (e.g. 50% MDMA) while the impurities arise randomly. Consequently, often 15 selected impurities are used by forensic experts as pre-tabletting characteristics for comparison purposes. In the second stage the mixed powder is compressed using a tabletting machine. The post-tabletting characteristics are thus considered to be features such as diameter, weight, thickness, logo, color or shape. In the second stage, the so called "production batch" is created. This batch is ready for transfer, sale and usage.

Because the stages are often carried out at different locations, we can consider the stages as two different sources: the source that produces the mixed powder (stage 1) and the source that forms the production batch (stage 2). Since there are two types of sources in this case, it is important to decide on what source we are focusing in the hypothesis H_p. If this decision is not made, confusing situations could emerge. This problem is described below. If we consider that consignments originate from the same source (hypothesis H_p), three possibilities can be distinguished:

1. The consignments have a different first source (stage 1, synthesis) and the same second source (stage 2, production batch), see Figure 2.3(a).

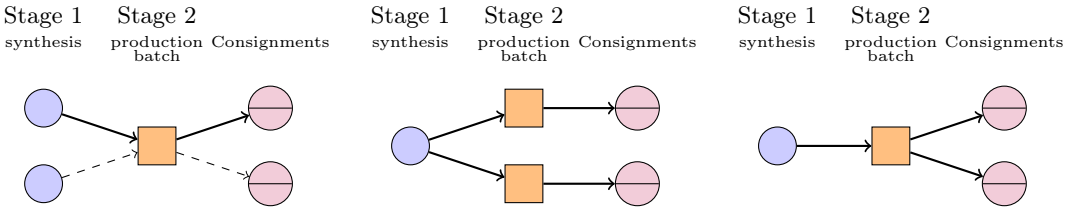
2. The consignments have the same first source (stage 1, synthesis) and a different second source (stage 2, production batch), see Figure 2.3(b).

3. The consignments have the same first source (stage 1, synthesis) and the same second source (stage 2, production batch), see Figure 2.3(c).

Consider the first possibility, see Figure 2.3(a). It is generally assumed that if tablets originate from the same source, they will have corresponding characteristics (Milliet et al. (2009)). Because the two consignments originate from the same production batch we can thus assume that they have the same post-tabletting characteristics. But, since the two consignments come from a different synthesis process, the pre-tabletting characteristics (e.g. (average) purity) can be different. If a dissimilarity in purity is measured, this would thus indicate a difference in the first source but cannot give exclusion about whether the two consignments originate from the same production

[1] Research performed by drug experts shows that often the impurities and dosage in tablets can be assumed to be distributed homogeneously. However, the impurities and dosage between tablets in one batch can vary. Therefore, we use the average impurities/dosage of a consignment.

batch. For that reason it is important to establish which source is being studied and hence which characteristics must be considered. This means that if we want to examine whether consignments have the same second source (production batch), only the post-tabletting characteristics should be studied. In Figure 2.3(b) this problem is given the other way around. Because the consignments come from a different production batch, the post-tabletting characteristics can be different. However, a dissimilarity in the post-tabletting characteristics does not imply that the consignments come from a different synthesis process. Thus, to examine whether consignments have the same first source (synthesis process), only the pre-tabletting characteristics should be studied. In Figure 2.3(c) this problem does not exist, because both the first- and second source are the same.



(a) Two consignments have a different synthesis (stage 1) but are from the same production batch (stage 2).

(b) Two consignments have the same synthesis (stage 1) but are from a different production batch (stage 2).

(c) Two consignments have the same synthesis (stage 1) and are from the same production batch (stage 2).

Figure 2.3: Examples of consignments that have the same source.

In this thesis we choose to consider the question whether the tablets of two consignments C_1 and C_2 originate from the same production batch or not. For that reason, the following prosecutor's hypothesis (H_p) and a hypothesis of the defense (H_d) can be considered (Bolck et al. (2009)):

$$\begin{cases} H_p: & \text{Tablets of the consignments } C_1 \text{ and } C_2 \text{ come from the same production batch.} \\ H_d: & \text{Tablets of consignments } C_1 \text{ and } C_2 \text{ come from different production batches.} \end{cases}$$

Hence, to investigate whether the consignments come from the same production batch, similarities and differences in the post-tabletting characteristics are studied. The pre-tabletting characteristics are thus not considered in this problem. Since post-tabletting characteristics are formed within one source, it is reliable to assume that there is a certain dependency between these characteristics. For instance, tablets with a Ferrari logo are often red.

Due to the increased marketing of xtc tablets, there is a sharp increase in the number of new tablets designs (174 new designs in 2014). This increase concerns the use of logos, shapes, bright and fluorescent colors and larger sizes and weights of tablets. For example, xtc tablets are produced specifically for individual events. Typically, these events are music events such as Amsterdam Dance Event. Examples of some of the physical features are given in Figure 2.4 and Table 2.1.

The fast increase of new designs give rise to an additional difficulty for the forensic experts. The problem is that when a new design is observed for the first time in two consignments this will automatically be considered as a rare event. Consequently this might indicate strong evidence that the consignments come from the same batch. However, such a conclusion could be a mistake in case this new design is a very popular design that is used by several manufacturers. The forensic experts do not possess this

kind of information when observing the new design for the first time, so when observing a new design they have to be cautious in drawing conclusion about rare events.

Figure 2.4: Tablet made for the Amsterdam Dance Event (ADE) electronic music festival (EMCDDA (2016b)).

Logo	Colour	Weight [mg]
Marlboro	Grey	190
Star	White	240
Euro	White	264
Peace	Pink	271
Ferrari	White	306
Mitsubishi	Beige-white	341
Dromedary	Yellow	237
Twins	Beige	305

Table 2.1: The logo, colour and weight of a sample of XTC tablets (Milliet et al. (2009)).

A methodology that can be used by forensic experts to investigate whether found similarities are probable when items have a common production batch or not is described in Section 2.2.1.

2.2 Towards a numerical framework

Until the late twentieth century, it was standard practice for forensic experts to provide conclusions as for example (Bolck et al. (2012)):

- "The xtc tablets come from the same production batch."
- "The xtc tablets do not come from the same production batch."
- "Whether the xtc tablets come from the same production batch is undecided."

To come to one of these conclusions, no general concept of evidence evaluation was established. Every field of expertise used their own methods to determine the strength of evidence (Sjerps (2004)). For example, to compare xtc tablets visual examination was frequently used.[2]

Within the likelihood ratio framework (or Bayesian framework) a numerical expression for the strength of evidence can often be computed. Using this concept, forensic experts do not draw any conclusions about whether tablets originate from the same batch or not. Instead it became more usual to give probabilistic conclusions as "The matching logos are *slightly* more probable if the consignments come from the same batch than if they come from different batches". In this way, forensic experts only provide the strength of their studied evidence. Consequently, it is up to the court only to conclude whether tablets are from the same batch.

In Section 2.2.1 it is described how the strength of evidence is expressed as a numerical expression by the likelihood ratio. The translation of the numerical expression to a verbal probabilistic conclusion is described in Section 2.2.2. The rise of the use of

[2]The interested reader is for example referred to Koper et al. (2007), Weyermann et al. (2008) or Milliet et al. (2009).

the likelihood ratio in forensic casework demands for a unified system to compute these ratios. Hence, forensic statisticians work on a European project with that purpose nowadays. Section 2.2.3 briefly describes this project and how this project relates to this thesis.

2.2.1 Likelihood ratio as strength of evidence

In the early twentieth century, forensic scientists began to develop the use of probabilistic theories in the field of forensic science (Aitken and Taroni (2004)). However, the first systematic work in the field was given in Fairley and Finkelstein (1970). They suggested a new approach to assess the strength of evidence based on Bayes' theorem. Nowadays this approach is known as the likelihood ratio approach (or Bayesian approach) and is widely accepted by forensic scientists.

To describe the likelihood ratio approach we consider a lawsuit about whether the suspect is a drug (xtc) dealer or not. Suppose the police has seized two consignments C_1 and C_2 of xtc tablets and there exists some links to the suspect. The question of interest for forensic experts is whether these consignments come from the same batch or not. Consequently, the prosecutors hypothesis H_p states that the consignments C_1 and C_2 come from the same batch, as described in Section 2.1.1. The hypothesis of the defense, H_d, claims that the consignments come from different production batches.

Every case that comes to court has certain non-scientific evidence for the jury to evaluate. This could include factors such as motive, eyewitness evidence, alibi and so on (Evett (1998)). An eyewitness in this case can be the following example: a farmer who rented his barn to the suspect saw the suspect carrying two heavy bags into his expensive car. Let I denote all background information, for example such non-scientific evidence.

During the lawsuit the prosecutor urges the judge to consider the probability that the consignments come from the same production batch given the non-scientific evidence, prior to any further (scientific) evidence. This probability is denoted by $P(H_p \mid I)$. Since it is not meaningful to consider this probability without considering an alternative, the odds in favor of H_p given the non-scientific evidence can be considered,

$$\frac{P(H_p \mid I)}{P(H_d \mid I)}. \tag{2.1}$$

If this ratio is greater than one, this means that given the non-scientific evidence, H_p is more probable than H_d.

However, in most cases scientific evidence E is available as well. This could include measurements on the tablets, such as the diameter and weight of the xtc tablets in the seized consignments.[3] As a consequence, the considered probabilities should be conditioned on both the non-scientific evidence I and the scientific evidence E. In theory, the probabilities $P(H_p \mid I)$ and $P(H_d \mid I)$ should be updated in the light of the new information E. The new ratio to be considered is

$$\frac{P(H_p \mid E, I)}{P(H_d \mid E, I)}. \tag{2.2}$$

In this sense, it seems reasonable to call the ratio in equation (2.1) the *prior odds* or the judge's "prior beliefs". After updating the prior probabilities in the light of the new information E, equation (2.2) is called the *posterior odds*. However, neither the judge

[3]Further specification of the evidence E is given in Chapter 3.

or a forensic expert can determine the probability of H_p (or H_d) directly. Hence, direct calculation of both equation (2.1) and (2.2) will be impossible. This is were Bayes' theorem comes in.

Theorem 2.2.1 (Bayes' theorem)**.** *Let A and B be events where $P(A) > 0$ and $P(B) > 0$. Then*

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)}.$$

The proof of Bayes' rule follows from the definition of conditional probabilities (Rice (2007)).

Applying this theorem to the numerator and denominator in the ratio given in equation (2.2) gives

$$\begin{aligned} \frac{P(H_p \mid E, I)}{P(H_d \mid E, I)} &= \frac{P(E, I \mid H_p)P(H_p)}{P(E, I \mid H_d)P(H_d)} \\ &= \frac{P(E \mid H_p, I)P(I \mid H_p)P(H_p)}{P(E \mid H_d, I)P(I \mid H_d)P(H_d)} \end{aligned}$$

where the latter equation is found by the definition of conditional probabilities. Then, again using the definition of conditional probabilities gives the fundamental equation for the likelihood ratio approach:

$$\frac{P(H_p \mid E, I)}{P(H_d \mid E, I)} = \frac{P(E \mid H_p, I)}{P(E \mid H_d, I)} \cdot \frac{P(H_p \mid I)}{P(H_d \mid I)}. \tag{2.3}$$

In the literature, the explicit mention of the background information I is often omitted from the latter equation for the ease of notation. In words, the formula is expressed as follows:

$$\text{Posterior odds} = \text{likelihood ratio} \times \text{prior odds}.$$

To determine the likelihood ratio the following two questions need to be answered: "what is the probability of the evidence given that the consignments C_1 and C_2 come from the same production batch?" and "what is the probability of the evidence given that the consignments C_1 and C_2 come from different production batches?". Since in many cases these questions can be answered, the likelihood ratio can be determined and thus the likelihood ratio may be thought of as the value of evidence. Equation (2.3) then demonstrates that the likelihood ratio assists the court in updating the prior odds to the posterior odds.

The task of the forensic experts is to provide the judge with the likelihood ratio. In an optimal situation, the judge will make an estimation of the prior odds, based on all background information. Subsequently, he will use the likelihood ratio to convert the prior odds into the posterior odds.

The process of updating the prior odds to a posterior odds using the likelihood ratio can in theory be an iterative process. This is because the evidence can exists of different pieces and for each piece a likelihood ratio can be calculated. To use all these likelihood ratios in the iterative process, the likelihood ratios of the different pieces of evidence must be computed under the same hypotheses H_p and H_d. For example, suppose the comparison problem as illustrated in Figure 2.3(c). A forensic expert has measured pre- and post-tabletting characteristics, but often considers these as two separate pieces of evidence. Since we consider the situation in Figure 2.3(c) the hypotheses of both pieces are the same, i.e. the consignments come from the same

source or not. Hence, first the forensic experts can give a likelihood ratio based on the pre-tabletting characteristics such that the judge can update his prior odds to a posterior odds. Subsequently, the forensic expert can give another likelihood ratio based on the post-tabletting characteristics such that the judge uses the posterior odds as a prior odds and update this to a new posterior odds.

Using the likelihood ratio approach, the forensic expert can only give a probabilistic conclusion about the strength of evidence in his field of expertise. For example: "These matching logos are 10 times more probable if consignments come from the same batch than if they come from different batches". Conclusions about the posterior odds, which state whether it is likely that the tablets come from the same production batch given the evidence, are only made by the judge. A schematic representation of this framework is given in Figure 2.5.

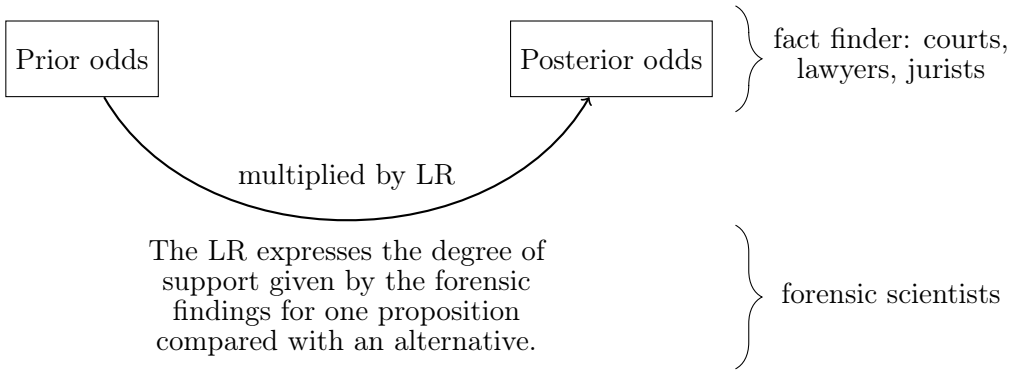


Figure 2.5: Schematic representation of the Bayesian framework (Champod (2013)). This process can be used as an iterative process if there are multiple pieces of evidence that provide multiple likelihood ratios under the same hypotheses. Then, the posterior odds can be used as the prior odds for a new likelihood ratio update.

Currently, in forensic statistics the likelihood ratio is a generally accepted measure to evaluate the strength of evidence. In addition, many experts prefer this approach over a traditional approach used for assessing evidence, e.g. using visual inspection and similarity measures to conclude that tablets have a strong link or not. In caseworks involving DNA or glass given in Figure 2.1, the likelihood ratio is a common measure to use. Since 2009 the approach is used for the comparison of xtc samples as well (Bolck et al. (2009)). However, the (numerical) likelihood ratio approach is not yet applicable in many fields of expertise and many kinds of casework. This is due to the absence of databases, to difficulties in interpretation and to the lack of information needed to consider certain hypotheses (e.g. activity level). In addition there are limitations caused by comparison problems that rely on many variables (e.g. handwriting or voice recognition) and by its demands for data.

2.2.2 Verbal likelihood ratios

In the previous section the likelihood ratio approach is described as a method to evaluate the evidence. In practice, there are two possibilities to determine this value of evidence:

- The most ideal situation is when the likelihood ratio can be computed numerically.

- As noticed before, numerical calculation of the likelihood ratio is not always possible. In that case it is common practice that the forensic expert gives a subjective (verbal) estimate of the likelihood ratio based on knowledge, expertise and experience (Bolck et al. (2012)).

In either way, reporting the value of evidence by the forensic expert is an important part in forensic casework. Suppose that the computed value of the likelihood ratio is equal to 378. In that case, the conclusion of the forensic expert would be:

> "These matching logos and colors are 378 times more probable if consignments come from the same batch than if they come from different batches."

However, such a conclusion can be hard to interpret. Especially for communicating evidence values in the courtroom it would be useful to translate the numerical expression "378 times more probable" to a verbal counterpart, such as "appreciably more probable".

Values of likelihood ratio	**Verbal equivalent**
1-2	The forensic findings provide *no assistance* in addressing the issue.
2-10	The forensic findings are *slightly more probable* given one proposition relative to the other.
10-100	The forensic findings are *more probable* given one proposition relative to the other.
100-10.000	The forensic findings are *much more probable* given one proposition relative to the other.
10.000-1.000.000	The forensic findings are *far more probable* given one proposition relative to the other.
>1.000.000	The forensic findings are *exceedingly more probable* given one proposition relative to the other.

Table 2.2: The current unified framework to relate verbal and numerical likelihood ratios (NFI (2014)).

If a numerical likelihood ratio is not computable, the forensic expert will only provide such a verbal conclusion based on experience. For example:

> "These matching logos and colors are slightly more probable if the consignments come from the same batch than if they come from different batches."

Thus, for the assessment of the conclusions in both situations it is important to allow the interpretation of different kinds of evidence in one common framework. Therefore a unified scale that relates verbal and numerical likelihood ratios needs to be introduced. In 2014 the Netherlands Forensic Institute has suggested such a unified framework, see Table 2.2. In forensic reports often only the verbal likelihood ratio is given. Note that this verbal likelihood ratio is not necessarily based on the numerical value. More discussion about this problem can for example be found in Evett et al. (2000) or Nordgaard (2012).

2.2.3 ENFSI LR software

In the beginning of Chapter 2 an overview of comparison problems in forensic casework is given. We have seen how the likelihood ratio can be used as a measure for the strength

of evidence in such cases. Although forensic statisticians agree about the framework of the likelihood ratio approach, calculation of the likelihood ratio is not (yet) unified. Various forensic statisticians in different countries have developed statistical models for the computation of likelihood ratios. Some of them have written (non-validated) scripts such that (non-statistical) forensic experts are able to compute likelihood ratios in their casework. The rise of the use of likelihood ratios in practice demands for a unified system to compute likelihood ratios. To investigate and harmonize the statistical models and existing software, the European network of forensic science institutes (ENFSI) leads the two year "ENFSI-LR" project. The aim of this project is to construct a userfriendly graphical interface around software that helps forensic experts to calculate likelihood ratios based on validated scripts and harmonized models. This userfriendly graphical interface is called SAILR.

This thesis addresses some problems concerning the ENFSI-LR project. In Section 4.2 the equality of two specific likelihood ratio expressions is proved. In Chapter 5 various possibilities for parameter estimation are investigated. Furthermore, some extensions and suggestions are given for future development of the project. In Section 4.1 some suggestions are given to check the model assumptions. In Chapter 7 the extension of existing models is described.

3

Likelihood ratios for evidence evaluation

In the previous chapter an introduction to forensic evidence evaluation is given. In Section 2.1 we have seen forensic comparison problems and in Section 2.2.1 we have discussed the importance of the likelihood ratio in such cases. The purpose of this chapter is to describe models that can be used to find a numerical value for the likelihood ratio given in equation (2.3).

The models are explained based on the drug comparison problem given in Section 2.1.1. Recall that when two consignments C_1 and C_2 are found, it is desired to know whether they come from the same production batch (H_p) or not (H_d). If consignments of xtc tablets are compared, we have discussed that we will focus on post-tabletting characteristics. These characteristics are physical characteristics and can be distinguished in either continuous features (weight, thickness) or discrete features (color, logo).

In forensic statistics various types of models exist to compute the likelihood ratio in cases like this. These models are applicable to either discrete- or continuous data (characteristics). The first section describes a model in case only discrete data are available and shows how this model can be used to find an expression for the likelihood ratio in terms of a probability mass function. Section 3.2 has the same structure, but is only applicable for continuous evidence. In Chapter 7 these models are combined, such that a likelihood ratio model is found that is applicable for a combination of discrete- and continuous evidence.

3.1 Discrete evidence

Suppose that discrete features of xtc tablets are measured by forensic experts. These features are post-tabletting characteristics such as logo or color (see Section 2.1.1). This section starts off with a likelihood ratio model applicable if measurements on such discrete characteristics are made. In Section 3.1.2 the described model is used to find an expression for the likelihood ratio.

3.1.1 Model

Suppose that p discrete features ($p > 1$) of xtc tablets are measured by forensic experts. The evidence (scientist's results) E is divided into two parts,

$$E = (\mathbf{Y}_1, \mathbf{Y}_2).$$

The discrete random vector $\mathbf{Y}_1$ represents p characteristics of the tablets from consignment C_1,

$$\mathbf{Y}_1 = (Y_{11}, \ldots, Y_{1p}).$$

This vector can be referred to as the *control data*. The control data will be compared to the *recovered data* $\mathbf{Y}_2$, that is the discrete random vector which represents p characteristics of the tablets from consignment C_2. Thus, the random variables $\{\mathbf{Y}_l\} = (Y_{lk}, l \in \{1,2\}, k \in \{1,\ldots,p\})$ represent discrete characteristics (e.g. the logo and color) of the tablets from consignments C_1 and C_2. In Section 2.1.1 we have seen that it is reasonable to assume that there exist certain dependencies between the post-tabletting characteristics of tablets from one consignment. We thus assume that there exist certain dependencies within the random vectors $\mathbf{Y}_1$ and $\mathbf{Y}_2$.

In the particular example of xtc tablets, the control- and recovered data consist of categorical variables. A categorical variable is a discrete random variable whose sample space is the set of s individually identified items (categories). The sample space can be taken as a finite sequence of integers that should be fixed beforehand. The integers are used as labels and the choice of the sequence is thus not important. For the evidence E we then have,

$$Y_{lk} \in \{1,\ldots,s_k\} \qquad \text{for} \qquad l \in \{1,2\}, \quad k \in \{1,\ldots,p\} \quad \text{and} \quad s_k \in \mathbb{N}, \tag{3.1}$$

where s_k is the number of levels of the variable, e.g. all possible logos such as Ferrari, shark or star. Here, we assume that all possible categories[1] (e.g. logos) are known and both the control- and recovered data can take the same values.

For discrete characteristics it is assumed that for each characteristic one measurement is sufficient to represent the evidence. In many applications this is a reliable assumption, since for discrete characteristics the measurement often equals the true value of the characteristic. Furthermore, it can be assumed that the xtc tablets within one consignment have the same discrete characteristics. So, if one tablet in consignment C_1 has a Ferrari logo, we assume that all of the tablets in consignment C_1 will have a Ferrari logo. For continuous random variables this is in particular not true, as we will see in Section 3.2.

Suppose that the multivariate probability mass function of the evidence E belongs to a set of probability mass functions $\mathcal{G}$, that is $g_{\mathbf{Y}_1,\mathbf{Y}_2} \in \mathcal{G}$ where

$$\mathcal{G} = \Big\{ g_{\mathbf{Y}_1,\mathbf{Y}_2} : \mathbb{N}^{2p} \to [0,1] \mid \qquad \sum_{\mathbf{v}} g_{\mathbf{Y}_1,\mathbf{Y}_2}(\mathbf{u},\mathbf{v}) = \sum_{\mathbf{v}} g_{\mathbf{Y}_1,\mathbf{Y}_2}(\mathbf{v},\mathbf{u}) \qquad \forall \mathbf{u} \Big\}. \tag{3.2}$$

This means that we consider probability mass functions $g_{\mathbf{Y}_1,\mathbf{Y}_2}$ such that the probability mass functions of $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are the same, i.e. $g_{\mathbf{Y}_1} = g_{\mathbf{Y}_2} = g$. This is a valid assumption, because we can assume that tablets in consignment C_1 and C_2 will have the same probability of a certain feature (e.g. Ferrari logo). For now, it will be assumed that the marginal probability mass functions are known. Later on, estimation of the probability mass functions is discussed.

3.1.2 An expression for the likelihood ratio

Suppose that a forensic expert has measured control- and recovered data $\mathbf{y}_1$ and $\mathbf{y}_2$ from two consignments C_1 and C_2. Recall from Section 2.2.1 that the goal of the

[1] In practice it is often not feasible to know all possible categories. For example, 174 new designs were discovered in 2014 (see Section 2.1.1). Thus "all possible categories" refers to categories that have been observed before.

forensic expert is to determine the likelihood ratio,

$$\mathrm{LR}(\mathbf{y}_1, \mathbf{y}_2) = \frac{P(E = (\mathbf{y}_1, \mathbf{y}_2) \mid H_p, I)}{P(E = (\mathbf{y}_1, \mathbf{y}_2) \mid H_d, I)}. \tag{3.3}$$

This is the ratio of the probability of the evidence given two opposite hypotheses,

$$\begin{cases} H_p: & \text{Tablets of the consignments } C_1 \text{ and } C_2 \text{ come from the same production batch.} \\ H_d: & \text{Tablets of consignments } C_1 \text{ and } C_2 \text{ come from different production batches.} \end{cases}$$

To find a useful expression for the likelihood ratio, the two hypotheses H_p and H_d need to be formulated more mathematically. First the formulas will be given and these will be explained thereafter.

$$\begin{cases} H_p: & g_{\mathbf{Y}_1, \mathbf{Y}_2 \mid I} \in \mathcal{G}_p \\ H_d: & g_{\mathbf{Y}_1, \mathbf{Y}_2 \mid I} \in \mathcal{G}_d, \end{cases} \tag{3.4}$$

where

$$\begin{aligned} \mathcal{G}_p &= \left\{ g_{\mathbf{Y}_1, \mathbf{Y}_2 \mid I} : \mathbb{N}^{2p} \to [0,1] \;\middle|\; g_{\mathbf{Y}_1, \mathbf{Y}_2 \mid I}(\mathbf{y}_1, \mathbf{y}_2 \mid I) = g(\mathbf{y}_1 \mid I) \mathbb{1}_{\{\mathbf{y}_1 = \mathbf{y}_2\}} \right\} \\ \mathcal{G}_d &= \left\{ g_{\mathbf{Y}_1, \mathbf{Y}_2 \mid I} : \mathbb{N}^{2p} \to [0,1] \;\middle|\; g_{\mathbf{Y}_1, \mathbf{Y}_2 \mid I}(\mathbf{y}_1, \mathbf{y}_2 \mid I) = g(\mathbf{y}_1 \mid I) g(\mathbf{y}_2 \mid I) \right\}. \end{aligned}$$

To understand this expression, we first consider the numerator of the likelihood ratio, i.e. H_p is true. In this case the consignments C_1 and C_2 come from the same batch. In Section 2.1.1 we have seen that it can be assumed that tablets from the same batch have the same (discrete) characteristics, i.e.

$$g_{\mathbf{Y}_1, \mathbf{Y}_2 \mid I}(\mathbf{y}_1, \mathbf{y}_2 \mid I) = P(\mathbf{Y}_1 = \mathbf{y}_1, \mathbf{Y}_1 = \mathbf{y}_2 \mid I).$$

And hence,

$$g_{\mathbf{Y}_1, \mathbf{Y}_2 \mid I}(\mathbf{y}_1, \mathbf{y}_2 \mid I) = g(\mathbf{y}_1 \mid I) \mathbb{1}_{\{\mathbf{y}_1 = \mathbf{y}_2\}},$$

where g is the marginal probability mass function of $\mathbf{Y}_1$ as defined in Section 3.1.1. To derive the denominator of the likelihood ratio, we assume that the hypothesis H_d is true. If the consignments C_1 and C_2 come from different batches, this does not necessarily imply that tablets in the two consignments have different features. For example, they could have the same logos. The only thing we do know in this case is that a feature of the tablets of consignment C_1 does not affect the probability of that feature of the tablets from the other consignment. And thus each characteristic in the control data is independent of the corresponding characteristic in the recovered data, that is $Y_{1k} \perp Y_{2k}$ for every characteristic k. In this case, this also implies that $Y_{1k} \perp Y_{2k'}$ for all $k \neq k'$ with $k' = 1, \ldots, p$. Hence,

$$g_{\mathbf{Y}_1, \mathbf{Y}_2 \mid I}(\mathbf{y}_1, \mathbf{y}_2 \mid I) = g(\mathbf{y}_1 \mid I) g(\mathbf{y}_2 \mid I).$$

Using (3.4), the likelihood ratio in equation (3.3) can be written as

$$\mathrm{LR}(\mathbf{y}_1, \mathbf{y}_1) = \frac{g_p(\mathbf{y}_1, \mathbf{y}_1 \mid I)}{g_d(\mathbf{y}_1, \mathbf{y}_1 \mid I)} = \frac{1}{g(\mathbf{y}_1 \mid I)}. \tag{3.5}$$

If the control- and recovered data are not the same ($\mathbf{y}_1 \neq \mathbf{y}_2$), the likelihood ratio will be zero because of the assumption that tablets from the same batch should have the same discrete characteristics.

In Section 3.1.1 we have assumed that there exist certain dependencies between the features of tablets from one consignment. However, in some applications it is reasonable to assume that the characteristics are independent of each other, that is $Y_{1k} \perp Y_{1k'}$ and $Y_{2k} \perp Y_{2k'}$ for all $k, k' \in \{1, \ldots, p\}$. In that case the likelihood ratio in equation (3.5) is the product of the likelihood ratios of the separate discrete characteristics.

For the strength of evidence, i.e. the value of the likelihood ratio, not only the fact that the measurements are the same is important. The value itself also plays an important role. Suppose that both consignments contain tablets with the logo shark. If sharks do not occur frequently, the strength of evidence will be higher since the probability of a shark will be smaller.

So far, the probability functions are assumed to be known. However, in practice the probability $g(\mathbf{y}_1 \mid I)$ has to be estimated in order to calculate a likelihood ratio. The most straightforward way is to estimate this probability based on the frequency of the features in the relevant background data. By relevant background data we mean the database should contain random sampled xtc consignments of the relevant population. The relevant population is possibly determined by the background information I. For example, suppose that our total database contains samples of xtc consignments from the "total population" Europe. But, because of the background information I there is reason to assume that both consignments are produced in the Netherlands.[2] Then, theoretically the relevant background data should contain features of (random) samples of xtc consignments that are produced in the Netherlands, i.e. a subset of the total database. However, in practice such a relevant database is often not available for drug comparison. In some areas (e.g. DNA) there already exist databases for some ethnic groups (possible relevant populations). Other areas work towards more specific resources too, but it takes a lot of time and input to accomplish that goal.

In the literature, see for example Aitken and Taroni (2004) or Bolck et al. (2012), the likelihood ratio is given by[3]

$$\frac{1}{P(\mathbf{Y} = \mathbf{y} \mid H_d, I)},$$

where often I is omitted from the expression. Although there is a small difference in notation, in practice this likelihood will give exactly the same value as the one given in equation (3.5). Since it is assumed that the probability of the recovered data is independent of whether the two consignments come from the same source or not, we know that

$$P(\mathbf{Y} = \mathbf{y} \mid H_d) = P(\mathbf{Y} = \mathbf{y}).$$

Conditioning on H_d is thus just a choice of notation.

Recall that for drug comparisons it is a valid assumption that tablets from the same source will have corresponding characteristics. Consequently, we have seen that the likelihood ratio is equal to zero if the control- and recovered data are not the same. In this thesis we will work under this assumption and thus with the likelihood ratio results as discussed previously. However, in other forensic fields this assumption (items from the same source have the same discrete characteristics) is not necessarily true. To give a more complete illustration of reality, we will briefly discuss this situation. If tablets from the same source will not necessarily have corresponding characteristics (e.g. tablets

[2] In this case the background information can be for example the place where the consignments were seized. Another possibility could be that the tablets have logos from a dance event in the Netherlands. Then, the feature 'logo' is the discrete evidence, but it influences the background information as well.

[3] In that literature, the control data is denoted by $\mathbf{X}$ and the recovered data by $\mathbf{Y}$.

from the same source have different logos), the situation under H_p changes. Hence, if different control- and recovered data are found we could find a positive likelihood ratio:

$$\mathrm{LR}(\mathbf{y}_1, \mathbf{y}_2) = \frac{P(E = (\mathbf{y}_1, \mathbf{y}_2) \mid H_p, I)}{P(\mathbf{Y}_1 = \mathbf{y}_1 \mid I)P(\mathbf{Y}_2 = \mathbf{y}_2 \mid I)} = \frac{P(\mathbf{Y}_2 = \mathbf{y}_2 \mid \mathbf{Y}_1 = \mathbf{y}_1, H_p, I)}{P(\mathbf{Y}_2 = \mathbf{y}_2 \mid I)}.$$

In addition to $g(\mathbf{y}_2 \mid I)$, the probability in the numerator has to be estimated as well (Bolck et al. (2012)).

3.2 Continuous evidence

Suppose that continuous features of xtc tablets are measured by forensic experts. These features are post-tabletting characteristics such as weight, thickness or diameter[4] (see Section 2.1.1). This section starts off with a likelihood ratio model applicable when measurements on such continuous characteristics are available. In Section 3.2.2 the described model is used to find an expression for the likelihood ratio.

3.2.1 Model

Suppose that p continuous features ($p > 1$) of xtc tablets are measured by forensic experts. In Section 3.1 it is explained that when discrete characteristics are measured, it is sufficient to determine the characteristics on a single tablet in a consignment. It is reasonable to assume that if one tablet has the logo Ferrari, all the tablets within that consignment will have the same logo. For continuous characteristics this is often not true. For example, suppose that the weight of the tablets is measured. Due to different types of errors, the measured weights will vary around a certain value. These errors can include measurement errors or errors due to an inhomogeneous production process. Hence, it might be that tablets on top of the seized consignment have lower weights than other tablets within that consignment. Therefore it is important that drug experts take a random sample from tablets in a consignment, instead of a single tablet to measure.

Let n_1 be the number of tablets that can be measured in consignment C_1. The control data is given by

$$\mathbf{Y}_1 = (\mathbf{Y}_{11}, \ldots, \mathbf{Y}_{1n_1})$$

where

$$\mathbf{Y}_{1j} = (Y_{1j1}, \ldots, Y_{1jp}) \qquad \text{for} \qquad j = 1, \ldots, n_1.$$

The vectors $\mathbf{Y}_{11}, \ldots, \mathbf{Y}_{1n_1}$ thus represent p measured features of n_1 different tablets in consignment C_1. Similarly, let n_2 be the number of tablets measured in consignment C_2 such that the recovered data is given by $\mathbf{Y}_2 = (\mathbf{Y}_{21}, \ldots, \mathbf{Y}_{2n_2})$. Although it is desirable that the forensic experts take a random sample from tablets in a consignment, in practice sometimes only one tablet can be recovered.

In forensic statistics, two-level models are often used to model the data. Using such a multilevel model is appropriate because the data is organized at more than one level: the n_l tablets (first level) are nested in either the control- or recovered consignment (second level). In the forensic literature the assumption of variation between the n_l

[4]Currently it is being discussed whether diameter should be treated as a continuous variable or a discrete variable. Forensic experts have strong suspicions that tabletting machines can only produce tablet of 4, 5 or 6 millimeter. If this is indeed true, the measurements can be categorized. In this thesis we will consider the diameter as a continuous variable.

tablets within the same consignment is known as *within-source variation*. The variation between the consignments is known as *between-source variation*.

To model the within-source variation, it is reasonable to assume the presence of noise within both the control- and recovered data due to errors for each feature as described above. Let $\boldsymbol{\varepsilon}_{lj} = (\varepsilon_{lj1}, \ldots, \varepsilon_{ljp}), l \in \{1, 2\}, j \in \{1, \ldots, n_l\}$ be random vectors of p random errors for the control- and recovered data respectively. Assume that the noise vectors are independent identically distributed normal random vectors,

$$\boldsymbol{\varepsilon}_{lj} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}) \quad \text{for} \quad l \in \{1, 2\}, \quad j \in \{1, \ldots, n_l\},$$

with

$$\boldsymbol{\varepsilon}_{1j} \perp \boldsymbol{\varepsilon}_{2j} \quad \forall j, \quad \text{and} \quad \boldsymbol{\varepsilon}_{lj} \perp \boldsymbol{\varepsilon}_{lj'} \quad \forall j \neq j'.$$

For now, we will assume that $\boldsymbol{\Sigma}$ is known. In Chapter 5 methods to estimate the covariance matrix $\boldsymbol{\Sigma}$ are discussed.

We assume that the control- and recovered data vary around their group means $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, i.e. let

$$\boldsymbol{\theta}_l = (\theta_{l1}, \ldots, \theta_{lp}) \qquad \text{for} \qquad l \in \{1, 2\}$$

then

$$\mathbf{Y}_{lj} \;=\; \boldsymbol{\theta}_l + \boldsymbol{\varepsilon}_{lj} \qquad \text{for} \qquad l \in \{1, 2\}, \quad j \in \{1, \ldots, n_l\},$$

To model the between-source variation we assume that $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are random group means who share their distribution,

$$\boldsymbol{\theta}_l \sim h \text{ for } l \in \{1, 2\}.$$

At this moment, we will assume that the between-source density h is known. In Chapter 4 and 6 we will make this probability function more explicit.

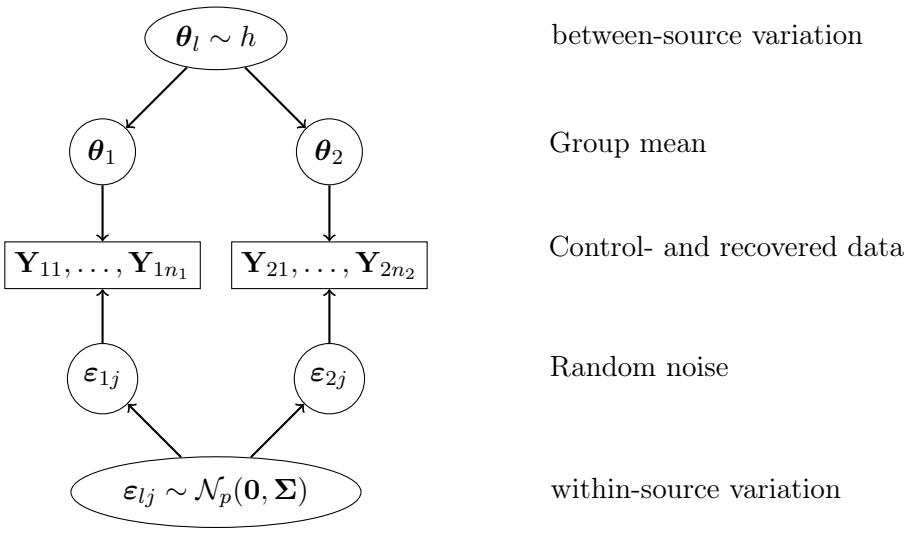


Figure 3.1: Schematic representation of the two-level model for the continuous control- and recovered data, $l \in \{1, 2\}$, $j \in \{1, \ldots, n_l\}$.

We can thus see the data as generated in a two-level process, where first the groups means $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are drawn from the between-source density h. Subsequently, each measurement is

$$\mathbf{Y}_{lj} \mid \boldsymbol{\theta}_l \;\overset{iid}{\sim}\; \mathcal{N}_p(\boldsymbol{\theta}_l, \boldsymbol{\Sigma}) \qquad \text{for} \qquad l \in \{1, 2\}, \quad j \in \{1, \ldots, n_l\},$$

that is the group mean plus some random noise such that the (conditional) variation within-source is normally distributed. Figure 3.1 represents this model schematically.

In this problem, it is generally accepted to reduce the original data to the means of the measurements (Bolck and Alberink (2011), Lindley (1977)). Then, the evidence E is given by

$$E = (\overline{\mathbf{Y}}_1, \overline{\mathbf{Y}}_2),$$

where

$$\overline{\mathbf{Y}}_l = \frac{1}{n_l}\sum_{j=1}^{n_l} \mathbf{Y}_{lj} \qquad \text{for} \qquad l \in \{1,2\}$$

such that

$$\overline{\mathbf{Y}}_l \mid \boldsymbol{\theta}_l \sim \mathcal{N}_p(\boldsymbol{\theta}_l, n_l^{-1}\boldsymbol{\Sigma}) \qquad \text{for} \qquad l \in \{1,2\}. \tag{3.6}$$

3.2.2 An expression for the likelihood ratio

In Section 3.1.2 we have seen that an expression for the likelihood ratio in equation (3.3) could be determined for discrete control- and recovered data. In this section we assume to have continuous data from the model described in Section 3.2.1. The evidence is given by the means of the observations, $E = (\overline{\mathbf{Y}}_1, \overline{\mathbf{Y}}_2)$ and suppose $E \sim F$, with F the distribution function. Since we are considering continuous random vectors the likelihood ratio from definition (3.3) does not apply directly. In fact, both the numerator and denominator in equation (3.3) are zero for continuous random vectors. Hence, we will consider the following approximation:

$$\mathrm{LR}(\overline{\mathbf{y}}_1, \overline{\mathbf{y}}_2) = \frac{P(\overline{\mathbf{Y}}_1 \in [\overline{\mathbf{y}}_1 \pm \delta \mathbf{1}_p]\,, \overline{\mathbf{Y}}_2 \in [\overline{\mathbf{y}}_2 \pm \delta \mathbf{1}_p] \mid H_p, I)}{P(\overline{\mathbf{Y}}_1 \in [\overline{\mathbf{y}}_1 \pm \delta \mathbf{1}_p]\,, \overline{\mathbf{Y}}_2 \in [\overline{\mathbf{y}}_2 \pm \delta \mathbf{1}_p] \mid H_d, I)} \tag{3.7}$$

for some small and positive constant δ and $\mathbf{1}_p$ a vector of ones. This likelihood ratio can be rewritten to its continuous version, that is in terms of the probability density function f of the evidence. In order to do that, we will use the lemma that is given below.

Lemma 3.2.1. *Let $\overline{Y}_1$ and $\overline{Y}_2$ be continuous random variables with joint probability density f and suppose its partial derivatives $f_{\overline{y}_1}$ and $f_{\overline{y}_2}$ are continuous in the neighbourhood of $(\overline{y}_1, \overline{y}_2)$. Then, the probability $P(\overline{\boldsymbol{Y}}_1 \in [\overline{\boldsymbol{y}}_1 \pm \delta \boldsymbol{1}_p]\,, \overline{\boldsymbol{Y}}_2 \in [\overline{\boldsymbol{y}}_2 \pm \delta \boldsymbol{1}_p]\,)$ can be approximated by $(2\delta)^{2p} f(\overline{\boldsymbol{y}}_1, \overline{\boldsymbol{y}}_2)$ if $\delta \to 0$.*

Proof. For the ease of notation we will assume $p = 1$. However, for $p > 1$ the same steps can be applied. By definition we know that

$$P(\overline{Y}_1 \in [\overline{y}_1 \pm \delta]\,, \overline{Y}_2 \in [\overline{y}_2 \pm \delta]\,) = \int_{\overline{y}_2-\delta}^{\overline{y}_2+\delta}\int_{\overline{y}_1-\delta}^{\overline{y}_1+\delta} f(y_1^*, y_2^*)dy_1^*dy_2^*. \tag{3.8}$$

Furthermore, we know that

$$\int_{\overline{y}_2-\delta}^{\overline{y}_2+\delta}\int_{\overline{y}_1-\delta}^{\overline{y}_1+\delta} f(\overline{y}_1, \overline{y}_2)dy_1^*dy_2^* = \int_{\overline{y}_2-\delta}^{\overline{y}_2+\delta} 2\delta f(\overline{y}_1, \overline{y}_2)dy_2^* = 4\delta^2 f(\overline{y}_1, \overline{y}_2). \tag{3.9}$$

By combining equation (3.8) and equation (3.9) we have that

$$\left|P\big(\overline{Y}_1\in[\overline{y}_1\pm\delta],\overline{Y}_2\in[\overline{y}_2\pm\delta]\big)-4\delta^2 f(\overline{y}_1,\overline{y}_2)\right|=\left|\int\limits_{\overline{y}_2-\delta}^{\overline{y}_2+\delta}\int\limits_{\overline{y}_1-\delta}^{\overline{y}_1+\delta}\big(f(y_1^*,y_2^*)-f(\overline{y}_1,\overline{y}_2)\big)dy_1^*dy_2^*\right|. \tag{3.10}$$

By the generalization of the mean value theorem (Adams and Essec (2010)) we have

$$\begin{aligned}\left|\int\limits_{\overline{y}_2-\delta}^{\overline{y}_2+\delta}\int\limits_{\overline{y}_1-\delta}^{\overline{y}_1+\delta}\big(f(y_1^*,y_2^*)-f(\overline{y}_1,\overline{y}_2)\big)dy_1^*dy_2^*\right| &= \left|\int\limits_{\overline{y}_2-\delta}^{\overline{y}_2+\delta}\int\limits_{\overline{y}_1-\delta}^{\overline{y}_1+\delta}\big((y_1^*-\overline{y}_1)f_{\overline{y}_1}(\boldsymbol{\zeta})+(y_2^*-\overline{y}_2)f_{\overline{y}_2}(\boldsymbol{\zeta})\big)dy_1^*dy_2^*\right| \\ &\leq \left|\int\limits_{\overline{y}_2-\delta}^{\overline{y}_2+\delta}\int\limits_{\overline{y}_1-\delta}^{\overline{y}_1+\delta}(y_1^*-\overline{y}_1)f_{\overline{y}_1}(\boldsymbol{\zeta})dy_1^*dy_2^*\right|+\left|\int\limits_{\overline{y}_2-\delta}^{\overline{y}_2+\delta}\int\limits_{\overline{y}_1-\delta}^{\overline{y}_1+\delta}(y_2^*-\overline{y}_2)f_{\overline{y}_2}(\boldsymbol{\zeta})dy_1^*dy_2^*\right|\end{aligned}$$

where the last inequality is true by the triangle inequality and $\boldsymbol{\zeta}$ lies in the open line segment between (y_1^*, y_2^*) and $(\overline{y}_1, \overline{y}_2)$. Then it can be used that,

$$\begin{aligned}\left|\int\limits_{\overline{y}_2-\delta}^{\overline{y}_2+\delta}\int\limits_{\overline{y}_1-\delta}^{\overline{y}_1+\delta}(y_1^*-\overline{y}_1)\,f_{\overline{y}_1}(\boldsymbol{\zeta})dy_1^*dy_2^*\right| &\leq \int\limits_{\overline{y}_2-\delta}^{\overline{y}_2+\delta}\int\limits_{\overline{y}_1-\delta}^{\overline{y}_1+\delta}\left|(y_1^*-\overline{y}_1)\,f_{\overline{y}_1}(\boldsymbol{\zeta})\right|dy_1^*dy_2^* \\ &\leq \left|\left|f_{\overline{y}_1}\right|\right|_\infty\cdot 2\delta^3 \end{aligned}\tag{3.11}$$

where the latter inequality is true, because we have assumed that the partial derivatives of f are continuous and because we can rewrite the following integral

$$\begin{aligned}\int\limits_{\overline{y}_2-\delta}^{\overline{y}_2+\delta}\int\limits_{\overline{y}_1-\delta}^{\overline{y}_1+\delta}|y_1^*-\overline{y}_1|\,dy_1^*dy_2^* &= \int\limits_{\overline{y}_2-\delta}^{\overline{y}_2+\delta}\left\{\int\limits_{\overline{y}_1-\delta}^{\overline{y}_1}-(y_1^*-\overline{y}_1)dy_1^*+\int\limits_{\overline{y}_1}^{\overline{y}_1+\delta}(y_1^*-\overline{y}_1)dy_1^*\right\}dy_2^* \\ &= \int\limits_{\overline{y}_2-\delta}^{\overline{y}_2+\delta}\delta^2 dy_2^* \\ &= 2\delta^3.\end{aligned}$$

By combining equation (3.11) with equation (3.10) it then follows that

$$\begin{aligned}\left|P(\overline{Y}_1\in[\overline{y}_1\pm\delta]\,,\overline{Y}_2\in[\overline{y}_2\pm\delta]\,)-4\delta^2 f(\overline{y}_1,\overline{y}_2)\right| &\leq \Big(\left|\left|f_{\overline{y}_1}\right|\right|_\infty+\left|\left|f_{\overline{y}_2}\right|\right|_\infty\Big)2\delta^3 \\ &:= \kappa 2\delta^3.\end{aligned}$$

Hence,

$$\delta^{-2}\left|P(\overline{Y}_1\in[\overline{y}_1\pm\delta]\,,\overline{Y}_2\in[\overline{y}_2\pm\delta]\,)-4\delta^2 f(\overline{y}_1,\overline{y}_2)\right|\to 0 \qquad \text{if } \delta\to 0. \tag{3.12}$$

□

In the definition of the likelihood ratio in equation (3.3), the numerator and denominator are zero for continuous random vectors. By using the approximation in equation (3.7) and the result from Lemma 3.2.1 we can now give meaning to the likelihood ratio in equation (3.3) for continuous random vectors:

$$\mathrm{LR}(\overline{\mathbf{y}}_1,\overline{\mathbf{y}}_2) = \lim_{\delta\to 0}\frac{(2\delta)^{2p}f(\overline{\mathbf{y}}_1,\overline{\mathbf{y}}_2\mid H_p,I)}{(2\delta)^{2p}f(\overline{\mathbf{y}}_1,\overline{\mathbf{y}}_2\mid H_d,I)}=\frac{f(\overline{\mathbf{y}}_1,\overline{\mathbf{y}}_2\mid H_p,I)}{f(\overline{\mathbf{y}}_1,\overline{\mathbf{y}}_2\mid H_d,I)}. \tag{3.13}$$

Note that the conditioning on the hypotheses and the background information I is omitted in Lemma 3.2.1, but this does not affect the result. Below, a useful expression for the likelihood ratio in equation (3.13) will be derived.

First assume that H_p is true, i.e. the consignments C_1 and C_2 come from the same batch. In Section 3.1 we have used that when H_p is true, the discrete features of consignments C_1 and C_2 have the same values. For continuous features, it cannot be assumed that they have the same value, but it can be assumed that the true means $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ of the measurements in both consignments are the same. Under this assumption, the two-level model for the control- and recovered data can be seen as the process that draws one mean vector $\boldsymbol{\theta} \sim h$ such that all measurements are derived from that mean plus some random noise. So, under the hypothesis H_p

$$f_{\overline{\mathbf{Y}}_1,\overline{\mathbf{Y}}_2|I}(\overline{\mathbf{y}}_1,\overline{\mathbf{y}}_2 \mid I) = \int_{\boldsymbol{\theta}} f_{\overline{\mathbf{Y}}_1,\overline{\mathbf{Y}}_2|\boldsymbol{\theta},I}(\overline{\mathbf{y}}_1,\overline{\mathbf{y}}_2 \mid \boldsymbol{\theta},I)h(\boldsymbol{\theta} \mid I)d\boldsymbol{\theta}.$$

Because $\overline{\mathbf{Y}}_1$ and $\overline{\mathbf{Y}}_2$ have the same underlying mean $\boldsymbol{\theta}$, they are not independent. However, conditional on this mean, the value of $\overline{\mathbf{Y}}_2$ does not contain additional information about $\overline{\mathbf{Y}}_1$, i.e. the means of the measurements are conditional independent. And thus,

$$f_{\overline{\mathbf{Y}}_1,\overline{\mathbf{Y}}_2|I}(\overline{\mathbf{y}}_1,\overline{\mathbf{y}}_2 \mid I) = \int_{\boldsymbol{\theta}} f_{\overline{\mathbf{Y}}_1|\boldsymbol{\theta},I}(\overline{\mathbf{y}}_1 \mid \boldsymbol{\theta},I)f_{\overline{\mathbf{Y}}_2|\boldsymbol{\theta},I}(\overline{\mathbf{y}}_2 \mid \boldsymbol{\theta},I)h(\boldsymbol{\theta} \mid I)d\boldsymbol{\theta}.$$

Now suppose that H_d is true, i.e. the consignments C_1 and C_2 come from different batches. If H_d is true, this does not necessarily imply that the two consignments have different true means. The only thing we do know in this case is that the means of the characteristics in consignment C_1 do not affect the probability of the means of the characteristics in consignment C_2. And thus, $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are independent. Under this assumption, the two-level model can be seen as the process that draws two means $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \sim h$ independently. Subsequently the control- and recovered measurements are derived from their mean plus some random noise. Thus, under hypothesis H_d

$$f_{\overline{\mathbf{Y}}_1,\overline{\mathbf{Y}}_2|I}(\overline{\mathbf{y}}_1,\overline{\mathbf{y}}_2 \mid I) = \int_{\boldsymbol{\theta}_1}\int_{\boldsymbol{\theta}_2} f_{\overline{\mathbf{Y}}_1,\overline{\mathbf{Y}}_2|\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,I}(\overline{\mathbf{y}}_1,\overline{\mathbf{y}}_2 \mid \boldsymbol{\theta}_1,\boldsymbol{\theta}_2,I)h(\boldsymbol{\theta}_1 \mid I)h(\boldsymbol{\theta}_2 \mid I)d\boldsymbol{\theta}_1 d\boldsymbol{\theta}_2.$$

Because the means of the measurements $\overline{\mathbf{Y}}_1$ and $\overline{\mathbf{Y}}_2$ are independent given their mean vectors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, we have that

$$\begin{aligned} f_{\overline{\mathbf{Y}}_1,\overline{\mathbf{Y}}_2|I}(\overline{\mathbf{y}}_1,\overline{\mathbf{y}}_2 \mid I) &= \int_{\boldsymbol{\theta}_1}\int_{\boldsymbol{\theta}_2} f_{\overline{\mathbf{Y}}_1|\boldsymbol{\theta}_1,I}(\overline{\mathbf{y}}_1 \mid \boldsymbol{\theta}_1,I)f_{\overline{\mathbf{Y}}_2|\boldsymbol{\theta}_2,I}(\overline{\mathbf{y}}_2 \mid \boldsymbol{\theta}_2,I) \\ &\times\ h(\boldsymbol{\theta}_1 \mid I)h(\boldsymbol{\theta}_2 \mid I)d\boldsymbol{\theta}_1 d\boldsymbol{\theta}_2. \end{aligned}$$

The latter expression is exactly the product of the marginal distributions of $\overline{\mathbf{Y}}_1$ and $\overline{\mathbf{Y}}_2$,

$$\begin{aligned} f_{\overline{\mathbf{Y}}_1,\overline{\mathbf{Y}}_2|I}(\overline{\mathbf{y}}_1,\overline{\mathbf{y}}_2 \mid I) &= \int_{\boldsymbol{\theta}_1} f_{\overline{\mathbf{Y}}_1|\boldsymbol{\theta}_1,I}(\overline{\mathbf{y}}_1 \mid \boldsymbol{\theta}_1,I)h(\boldsymbol{\theta}_1 \mid I)d\boldsymbol{\theta}_1 \\ &\times \int_{\boldsymbol{\theta}_2} f_{\overline{\mathbf{Y}}_2|\boldsymbol{\theta}_2,I}(\overline{\mathbf{y}}_2 \mid \boldsymbol{\theta}_2,I)h(\boldsymbol{\theta}_2 \mid I)d\boldsymbol{\theta}_2. \end{aligned}$$

And thus, under H_d we know that the means of the measurements are independent, i.e. $\overline{\mathbf{Y}}_1 \perp \overline{\mathbf{Y}}_2$. The latter derivations of the joint density function f under the hypotheses H_p and H_d can be summarized as

$$\begin{cases} H_p: & f_{\overline{\mathbf{Y}}_1,\overline{\mathbf{Y}}_2|I} \in \mathcal{F}_p \\ H_d: & f_{\overline{\mathbf{Y}}_1,\overline{\mathbf{Y}}_2|I} \in \mathcal{F}_d, \end{cases} \tag{3.14}$$

where

$$\begin{aligned}
\mathcal{F}_p &= \Big\{ f_{\overline{\mathbf{Y}}_1,\overline{\mathbf{Y}}_2|I} : \mathbb{R}^{2p} \to [0,1] \mid \quad f_{\overline{\mathbf{Y}}_1,\overline{\mathbf{Y}}_2|I}(\overline{\mathbf{y}}_1,\overline{\mathbf{y}}_2 \mid I) = \int_{\boldsymbol{\theta}} f_{\overline{\mathbf{Y}}_1|\boldsymbol{\theta},I}(\overline{\mathbf{y}}_1 \mid \boldsymbol{\theta}, I) \\
&\times \quad f_{\overline{\mathbf{Y}}_2|\boldsymbol{\theta},I}(\overline{\mathbf{y}}_2 \mid \boldsymbol{\theta}, I) h(\boldsymbol{\theta} \mid I) d\boldsymbol{\theta} \Big\}, \\
\mathcal{F}_d &= \Big\{ f_{\overline{\mathbf{Y}}_1,\overline{\mathbf{Y}}_2|I} : \mathbb{R}^{2p} \to [0,1] \mid \qquad f_{\overline{\mathbf{Y}}_1,\overline{\mathbf{Y}}_2|I}(\overline{\mathbf{y}}_1,\overline{\mathbf{y}}_2 \mid I) = f_{\overline{\mathbf{Y}}_1|I}(\overline{\mathbf{y}}_1 \mid I) f_{\overline{\mathbf{Y}}_2|I}(\overline{\mathbf{y}}_2 \mid I) \Big\}.
\end{aligned}$$

The expression in equation (3.14) can be used to write the likelihood ratio in equation (3.13) as

$$\begin{aligned}
\mathrm{LR}(\overline{\mathbf{y}}_1,\overline{\mathbf{y}}_2) &= \frac{\int_{\boldsymbol{\theta}} f_{\overline{\mathbf{Y}}_1|\boldsymbol{\theta},I}(\overline{\mathbf{y}}_1 \mid \boldsymbol{\theta}, I) f_{\overline{\mathbf{Y}}_2|\boldsymbol{\theta},I}(\overline{\mathbf{y}}_2 \mid \boldsymbol{\theta}, I) h(\boldsymbol{\theta} \mid I) d\boldsymbol{\theta}}{\int_{\boldsymbol{\theta}_1} f_{\overline{\mathbf{Y}}_1|\boldsymbol{\theta}_1,I}(\overline{\mathbf{y}}_1 \mid \boldsymbol{\theta}_1, I) h(\boldsymbol{\theta}_1 \mid I) d\boldsymbol{\theta}_1} \\
&\times \frac{1}{\int_{\boldsymbol{\theta}_2} f_{\overline{\mathbf{Y}}_2|\boldsymbol{\theta}_2,I}(\overline{\mathbf{y}}_2 \mid \boldsymbol{\theta}_2, I) h(\boldsymbol{\theta}_2 \mid I) d\boldsymbol{\theta}_2}. \qquad (3.15)
\end{aligned}$$

In the forensic literature two other derivations for the likelihood ratio are given (see Aitken and Taroni (2004) or Bolck et al. (2012)). The derivation in Aitken and Taroni (2004) leads to exactly the same likelihood ratio as given in equation (3.15). Bolck et al. (2012) uses a different expression to describe the same likelihood ratio as given in equation (3.15). In Chapter 4 it will be shown that the expression used in Bolck et al. (2012) can be rewritten into the likelihood ratio expression used in equation (3.15) and Aitken and Taroni (2004). This is important in light of unification within the ENFSI-LR project.

From Section 3.2.1 we know that the conditional densities $f_{\overline{\mathbf{Y}}_1|\boldsymbol{\theta}_1}$ and $f_{\overline{\mathbf{Y}}_2|\boldsymbol{\theta}_2}$ are normal densities. However, to compute the likelihood ratio the between-source density h is important. Currently, in forensic statistics two possibilities for the between-source density h are distinguished. Either a normal distribution is assumed or a kernel density estimator is used to estimate the density function. Both possibilities are further discussed in Chapter 4 and Chapter 6. Based on the choice of the between-source distribution, these sections will give explicit formulas to compute the likelihood ratio as well.

4

Likelihood ratios in Gaussian two-level models

The purpose of this chapter is to make the likelihood ratio that is given in Section 3.2 more explicit. This chapter thus focuses on continuous evidence that is modeled using a two-level model. This means that the control- and recovered data are modelled such that they vary around their random group means $\boldsymbol{\theta}_l$, $l \in \{1, 2\}$, which are drawn from a between-source density h. Up to now we have assumed that the between-source density h was known. In practice, either a normal density is assumed or, if the normality assumption is not satisfied, a nonparametric density estimate is assumed to model the between-source variation. In this chapter a normal between-source density is assumed such that likelihood ratios in Gaussian two-level models can be derived. In forensic literature, these models are also called "two-level normal normal models", because both the within-source distribution and the between-source distribution are assumed to be normal. The assumption of a nonparametric density estimate for the between-source density is discussed in Chapter 6.

In this chapter two problems of interest for the ENFSI-LR project are discussed. In addition, it will give a good overall view on actual computation of likelihood ratios. Section 4.1 describes the normality assumption and suggests some first ideas to assess the assumption of normality. A description of formal methods to assess normality is valuable for further development of the ENFSI-LR project, because it is of interest to implement such methods in the software. Formal tests could help (non-statistical) forensic experts to decide whether the model described in this chapter or the one in Chapter 6 should be used for likelihood ratio calculation.

One of the objectives of the ENFSI-LR project is to agree upon a likelihood ratio formula. This is especially of interest for the validation of the implemented likelihood ratio in the software. Section 4.2 describes two approaches that lead to likelihood ratios under the normality assumption. This section shows that these different approaches lead to the exact same likelihood ratio. This validates the implemented likelihood ratio.

4.1 A Gaussian between-source distribution

This section will focus on the normality assumption for the between-source density. Section 4.1.1 describes the assumption of normality for the true means. According to a literature study, Section 4.1.2 will discuss some first ideas to assess the assumption of normality. Since in future development formal tests will be implemented in the ENFSI-LR software, these tests can be further explored.

4.1.1 The assumption of normality

In Section 3.2 we have seen that the continuous evidence E is given by $E = (\overline{\mathbf{Y}}_1, \overline{\mathbf{Y}}_2)$, where $\overline{\mathbf{Y}}_1$ and $\overline{\mathbf{Y}}_2$ are the means of the control- and recovered data. The vectors $\overline{\mathbf{Y}}_1$ and $\overline{\mathbf{Y}}_2$ are modelled as a two-level model, i.e. the mean vector is assumed to have a so called between-source distribution h,

$$\begin{aligned} \overline{\mathbf{Y}}_l &= \boldsymbol{\theta}_l + \overline{\boldsymbol{\varepsilon}_l} \qquad \text{for} \qquad l \in \{1,2\}, \\ \boldsymbol{\theta}_l &\sim h, \end{aligned}$$

where $\overline{\boldsymbol{\varepsilon}_l}$ is the mean of the n_l random error vectors in group l. So far we have assumed that the between-source density h is known. In practice, however, this is not the case. In the beginning of the development of two-level models in forensic statistics, it was common to assume a (multivariate) normal distribution for $\boldsymbol{\theta}_l$. For xtc comparison, this assumption originated from the idea which states that produced batches of xtc tablets were on average the same. For example, it was supposed that the means of the weights of the tablets in different batches have a certain overall mean and deviate from that mean depending on who produced the batches. Under this assumption the control- and recovered data are in fact modelled as a random effects model (Searle (1992)), i.e.

$$Y_{lj} = \mu + \alpha_l + \varepsilon_{lj}, \qquad \text{for} \qquad l \in \{1,2\}, \quad j \in \{1,\ldots,n_l\}$$

with μ the overall mean and

$$\alpha_l \sim N(0, \tau^2)$$

the random group effect, i.e. the effect of the lth consignment depending on who produced the batches. This is thus the same as assuming a normal distribution for the means of the weights of the batches:

$$\theta_l \sim \mathcal{N}(\mu, \tau^2).$$

The described assumption of normality can be extended to a multivariate normal distribution for $\boldsymbol{\theta}_l$ with mean vector $\boldsymbol{\mu}$, containing the overall means of the batch means for each feature, and $\mathbf{T}$ the covariance matrix,

$$\boldsymbol{\theta}_l \sim \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{T}).$$

Hence, all feature means are supposed to be normally distributed. Furthermore, we assume that

$$\boldsymbol{\theta}_l \perp \boldsymbol{\varepsilon}_{l'j} \quad \forall l \text{ and } l'.$$

Recall from Section 3.2.1 that $\overline{\boldsymbol{\varepsilon}_l} \sim \mathcal{N}_p(\mathbf{0}, n_l^{-1}\boldsymbol{\Sigma})$ because the error vectors within group l are independent. Therefore, $\overline{\mathbf{Y}}_1$ and $\overline{\mathbf{Y}}_2$ are the sums of two independent normal vectors $\boldsymbol{\theta}_l$ and $\overline{\boldsymbol{\varepsilon}_l}$. And thus

$$\overline{\mathbf{Y}}_l \sim \mathcal{N}_p(\boldsymbol{\mu}, n_l^{-1}\boldsymbol{\Sigma} + \mathbf{T}). \tag{4.1}$$

Although it is used to be reasonable to assume that the means of the features of different batches were normally distributed, nowadays this is less realistic. For instance, in Section 2.1.1 we have seen the rise of "super pills" on the market. Hence, we can imagine that for example the means of the weights do not only have a peak on μ, but they also have a peak on a higher weight, say $\mu + c$ for c a constant. The normality assumption is thus not valid in such situation. Consequently, a multivariate normal distribution for $\boldsymbol{\theta}_l$ will be even more complicated, since each feature means should have a normal distribution. The following section will discuss some ideas to assess whether a multivariate normal distribution is a valid assumption for $\boldsymbol{\theta}_l$.

4.1.2 Evaluating the assumption of normality

In Section 4.1.1 we have discussed the assumption of normality for the true means $\boldsymbol{\theta}_l$, $l \in \{1, 2\}$. Based on experience, a forensic expert assigns a normal distribution to for example the means of the weights. However, especially for higher dimensions ($p > 1$) such an assumption is harder to assess. Therefore, this section covers some initial exploration to methods to assess the (multivariate) normality assumption for the true means $\boldsymbol{\theta}_l$, $l \in \{1, 2\}$.

To evaluate the normality assumption, available background (empirical) data will be used. Let

$$\overline{\mathbf{Z}}_i = (\overline{Z}_{i1}, \ldots, \overline{Z}_{ip}) \text{ for } i = 1, \ldots, m.$$

be the batch means taken over m batches of p features that are contained in the background data. Thus, $\overline{\mathbf{Z}}_1, \ldots, \overline{\mathbf{Z}}_m$ are the mean vectors of the m consignments that are contained in the database. A more detailed description of the background data will be given in Section 5.1. Since the mean vectors $\boldsymbol{\theta}_l$ are assumed to be equally distributed for both consignments C_1 ($l = 1$) and C_2 ($l = 2$), the sample $\overline{\mathbf{Z}}_1, \ldots, \overline{\mathbf{Z}}_m$ can be used as realizations for $\boldsymbol{\theta}_l$. Furthermore, let the observations $\overline{\mathbf{Z}}_1, \ldots, \overline{\mathbf{Z}}_m$ be independent and identically distributed from some distribution H. Then we want to answer the following question:

> Are the observations $\overline{\mathbf{Z}}_1, \ldots, \overline{\mathbf{Z}}_m$ samples from a normal distribution, i.e. is H a (multivariate) normal distribution?

To answer this question we will use properties of the multivariate normal distribution. Since a normal random vector has normal marginals, we can first assess whether the marginals of $\overline{\mathbf{Z}}_i, i \in \{1, \ldots, m\}$ are normal distributed. However, individual normal random variables are not necessarily jointly normal distributed. Hence, we should test the multivariate structure on normality as well.

Normality of the marginal distributions

To assess whether the observations $\overline{Z}_{1k}, \ldots, \overline{Z}_{mk}$ come from a normal distribution for each $k \in \{1, \ldots, p\}$, either probability plots can be used or goodness of fit tests. The latter test is a more formal test. Both types will be briefly described below. In Section 7.3 these methods will be applied to real xtc data.

- QQ-plots provide a visual way to assess a certain distributional assumption. Although this is not a formal method, it is a quick tool to check the assumption. This visual test is especially interesting if the number of features p is not too high. To assess normality, the normal (probability) plot can be used as a special case of the QQ-plot. The idea of a normal plot is to compare the order statistics (sample quantiles) to quantiles from a standard normal distribution (theoretic quantiles) (Rice (2007)).

  Let $\overline{Z}_{1k}, \ldots, \overline{Z}_{mk}$ be data from some distribution H. The empirical distribution of the data $\hat{H}_m$ can be estimated by $\hat{H}_m(\overline{Z}_{[i]k}) = \frac{i}{m}$ such that $\hat{H}_m(\overline{Z}_{[i]k}) \approx H(\overline{Z}_{[i]k})$, where $\overline{Z}_{[i]k}$ is the i-th order statistic. If H is indeed a normal distribution with

mean μ and variance τ then,

$$\begin{aligned} \overline{Z}_{[i]k} &\approx H^{-1}\Big(\hat{H}_m\Big(\overline{Z}_{[i]k}\Big)\Big) \\ &= H^{-1}\left(\frac{i}{m}\right) \\ &= \mu + \tau\Phi^{-1}\left(\frac{i}{m}\right) \end{aligned} \tag{4.2}$$

where Φ^{-1} is the standard normal quantile function. Thus, if the data comes from a normal distribution we expect the points $\left(\overline{Z}_{[i]k}, \Phi^{-1}\left(\frac{i-0.5}{m}\right)\right)$ to be in a straight line. Here, the term $\frac{i}{m}$ is replaced by $\frac{i-0.5}{m}$ to ensure that $\Phi^{-1}(1)$ is not evaluated because this can be infinite.

- A more formal way to assess the normality of the observations $\overline{Z}_{1k}, \ldots, \overline{Z}_{mk}$ is to use composite goodness of fit tests. In such tests we want to test whether the distribution of the sample belong to a certain class of distribution functions. In this problem, we thus consider the testing problem

$$\begin{cases} H_0: & H \in \mathcal{H} \\ H_1: & H \notin \mathcal{H}, \end{cases}$$

where $\mathcal{H}$ is the class of normal distribution functions. An example of a well-known goodness of fit test for normality is the Shapiro-Wilk test. The test-statistic is

$$W = \frac{(\sum_{i=1}^m a_i \overline{Z}_{[i]k})^2}{\sum_{i=1}^m (\overline{Z}_{ik} - \overline{Z}_m)^2},$$

where $\overline{Z}_m$ is the sample mean and the coefficients $(a_1, \ldots, a_m)$ are based on the expected values and covariances of order statistics of independent standard normal random variables (Shapiro and Wilk (1965)). The idea behind this test statistic is that under H_0 both the numerator and denominator are estimators for $(n-1)\tau^2$ (Mardia (1980)). Thus, a test for normality is to compare the statistic W with 1. The hypothesis H_0 is rejected for small values of W. The Shapiro-Wilk test is only one example of a suitable goodness of fit test for this problem. Other tests are for example based on skewness and kurtosis. For more information about goodness of fit tests, see for example D'Agostino and Stephens (1986) or Mardia (1980).

If at least one of the marginal distributions is clearly not normal distributed, we thus know that the multivariate structure is not normal either. However, one should bear in mind that all measures of goodness of fit tests suffer from the same problem: for small samples only very aberrant behaviour will be identified as a lack of fit and thus the difference is not always detected. On the other hand, for large samples a difference (relevant or not) is always detected.

Normality of the multivariate structure

Although the presence of non-normality is often reflected in the marginal distributions (Johnson and Wichern (2007)), the multivariate structure should be assessed as well. The assessment of the multivariate structure can be done by extensions from univariate tests as described before. In Section 7.3 these methods will be applied to real xtc data.

- For a visual check to assess multivariate normality again a QQ-plot can be used. In contrary to the normal plot for the univariate case, now a chi-squared plot should be used (Johnson and Wichern (2007)). In this plot the squared generalized distances d_i^2 are used as sample quantiles where,

$$d_i^2 = (\overline{\mathbf{Z}}_i - \boldsymbol{\mu})\mathbf{T}^{-1}(\overline{\mathbf{Z}}_i - \boldsymbol{\mu}), \quad i = 1, \ldots, m. \tag{4.3}$$

  In Johnson and Wichern (2007, result 4.7) it is proved that if $\overline{\mathbf{Z}}_i$ is distributed as a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{T}$, the squared generalized distances are chi-squared distributed with p degrees of freedom, where p is the number of features. This is true, because d_i^2 can be written as the sum of independent standard normal random variables, which is exactly the definition of a chi-squared random variable. Hence, the theoretic quantiles are based on the chi-squared distribution. Again the steps of equation (4.2) can be applied, but since the chi-squared distribution is not a location-scale distribution and has only the parameter p which is known, we have

$$d_{(i)}^2 \approx q_p\left(\frac{i}{m}\right),$$

  where q_p is the quantile function of the chi-squared distribution with p degrees of freedom. Thus, if the data $d_{(i)}^2$ comes from a chi-squared distribution, we expect the points $\left(d_{(i)}^2, q_p\left(\frac{i-0.5}{m}\right)\right)$ to be on a straight line through the origin with slope one. Hence, this indicates that the observations $\overline{\mathbf{Z}}_1, \ldots, \overline{\mathbf{Z}}_m$ come from a multivariate distribution. In Johnson and Wichern (2007) it is noted that this method should only be applied if the condition $m - p > 25$ is satisfied. Furthermore, the sample mean and sample covariance can be used for $\boldsymbol{\mu}$ and $\mathbf{T}$ in the computation of d_i^2 in equation (4.3).

- For the goodness of fit tests for the multivariate normal distribution some univariate generalizations exist, such as the extension of the Shapiro-Wilk test (Malkovich and Afiffi (1973)). But also various strict multivariate procedures exist. Good overviews of these tests are e.g. in Gnanadesikan (1977) or Mardia (1980). Another test is suggested in Doornik and Hansen (2008). The proposed test has the best size and power in comparison to tests that were supposed to be the most effective in earlier experiments. Further details about these tests will not be discussed here, but these multivariate tests will be valuable in addition to the chi-square plot to assess the normality of the multivariate structure.

4.2 Derivation of the likelihood ratio

Section 4.1 covered the assumption of a Gaussian between-source distribution. In the forensic literature two different expressions for the likelihood ratio are given under this assumption of normality. Section 4.2.1 and Section 4.2.2 describe these two approaches that are used to obtain the likelihood ratio given in equation (3.15) in more detail. Finally, in Section 4.2.3 it will be shown that these different approaches lead to the same likelihood ratio. This result is therefore useful for the ENFSI-LR project. Although it is not explicitly defined in Chapter 3, from now on we will assume that the vectors $\overline{\mathbf{Y}}_l, \boldsymbol{\theta}_l$ and $\boldsymbol{\varepsilon}_{lj}$ are column vectors $\forall l \in \{1, 2\}, j \in \{1, \ldots, n_l\}$.

4.2.1 Lindley's approach

In Aitken and Lucy (2004) and Zadora et al. (2014) a likelihood ratio is given under the assumption of a normal between-group distribution. The formula that is given there, is based on the approach in Lindley (1977). The approach in Lindley (1977) is applied on the univariate model, i.e. for $p = 1$. The idea of Lindley can be extended to a multivariate model, $p > 1$, such that it leads to the formulas given in Aitken and Lucy (2004) and Zadora et al. (2014). In this section Lindley's approach for the multivariate structure will be discussed in more detail. To do this, recall the likelihood ratio given in equation (3.15). The computation of the numerator and the denominator will be given separately.

Computation of the numerator

The numerator of the likelihood ratio in equation (3.15) is given by[1]

$$\int_{\boldsymbol{\theta}} f_{\overline{\mathbf{Y}}_1|\boldsymbol{\theta}}(\overline{\mathbf{y}}_1 \mid \boldsymbol{\theta}) f_{\overline{\mathbf{Y}}_2|\boldsymbol{\theta}}(\overline{\mathbf{y}}_2 \mid \boldsymbol{\theta}) h(\boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{4.4}$$

Since we know from Section 3.2 that the conditional random vector $\overline{\mathbf{Y}}_l \mid \boldsymbol{\theta}$ is normally distributed and we have assumed that $\boldsymbol{\theta}$ is normally distributed as well, the numerator could be computed by direct integration. However, Lindley uses the fact that equation (4.4) is the unconditional joint density of $\overline{\mathbf{Y}}_1$ and $\overline{\mathbf{Y}}_2$. More precisely, he claims that this joint density is normal. We cannot immediately claim that equation (4.4) is a multivariate normal density function, because unconditional on $\boldsymbol{\theta}$ the vectors $\overline{\mathbf{Y}}_1$ and $\overline{\mathbf{Y}}_2$ are not independent. Therefore, it will be shown that the numerator is indeed a multivariate normal density below.

Lemma 4.2.1. *The integral in equation* (4.4) *is a multivariate normal density function with mean vector* $\overline{\boldsymbol{\mu}} = \begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu} \end{pmatrix}$ *and covariance matrix* $\overline{\boldsymbol{\Sigma}} = \begin{pmatrix} \boldsymbol{T} + n_1^{-1}\boldsymbol{\Sigma} & \boldsymbol{T} \\ \boldsymbol{T} & \boldsymbol{T} + n_2^{-1}\boldsymbol{\Sigma} \end{pmatrix}$.

Proof. Since $\overline{\mathbf{Y}}_1$ and $\overline{\mathbf{Y}}_2$ both have the underlying true mean vector $\boldsymbol{\theta}$, we know from Section 3.2.1 that we can write

$$\overline{\mathbf{Y}}_{lj} = \boldsymbol{\theta} + \overline{\boldsymbol{\varepsilon}}_l,$$

where

$$\overline{\boldsymbol{\varepsilon}}_l = \frac{1}{n_l} \sum_{j=1}^{n_l} \boldsymbol{\varepsilon}_{lj} \sim \mathcal{N}_p\left(\mathbf{0}, n_l^{-1}\boldsymbol{\Sigma}\right)$$

and

$$\boldsymbol{\theta} \sim \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{T}).$$

Recall from Section 3.2.1 and Section 4.1.1 that

$$\overline{\boldsymbol{\varepsilon}}_1 \perp \overline{\boldsymbol{\varepsilon}}_2 \quad \text{and} \quad \boldsymbol{\theta} \perp \overline{\boldsymbol{\varepsilon}}_l \quad \text{for} \quad l \in \{1, 2\}.$$

By definition a random vector is multivariate normal if and only if every linear combination of its elements is univariate normal (Rao (1973), p.518). Then, it follows that

$$\begin{pmatrix} \boldsymbol{\theta} \\ \overline{\boldsymbol{\varepsilon}}_1 \\ \overline{\boldsymbol{\varepsilon}}_2 \end{pmatrix} \sim \mathcal{N}_{3p}\left(\begin{pmatrix} \boldsymbol{\mu} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{T} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & n_1^{-1}\boldsymbol{\Sigma} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & n_2^{-1}\boldsymbol{\Sigma} \end{pmatrix} \right).$$

[1] For the ease of notation, the conditioning on the background information I is omitted here. In Chapter 5 it will be explained how the background information should be used in the likelihood ratio.

Now we can use the following property of the multivariate normal distribution (Rao (1973), p.519)

$$\begin{pmatrix}\overline{\mathbf{Y}}_1\\ \overline{\mathbf{Y}}_2\end{pmatrix} = A\begin{pmatrix}\boldsymbol{\theta}\\ \overline{\boldsymbol{\varepsilon}}_1\\ \overline{\boldsymbol{\varepsilon}}_2\end{pmatrix} \sim \mathcal{N}_{2p}\left(A\begin{pmatrix}\boldsymbol{\mu}\\ \mathbf{0}\\ \mathbf{0}\end{pmatrix}, A\begin{pmatrix}\mathbf{T} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & n_1^{-1}\boldsymbol{\Sigma} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & n_2^{-1}\boldsymbol{\Sigma}\end{pmatrix}A'\right)$$

with

$$A = \begin{pmatrix}\mathbf{I}_p & \mathbf{I}_p & \mathbf{0}\\ \mathbf{I}_p & \mathbf{0} & \mathbf{I}_p\end{pmatrix}.$$

Here, $\mathbf{I}_p$ is the $p \times p$ identity matrix. Since

$$A\begin{pmatrix}\boldsymbol{\mu}\\ \mathbf{0}\\ \mathbf{0}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\mu}\\ \boldsymbol{\mu}\end{pmatrix} := \overline{\boldsymbol{\mu}}$$

and

$$\begin{aligned} A\begin{pmatrix}\mathbf{T} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & n_1^{-1}\boldsymbol{\Sigma} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & n_2^{-1}\boldsymbol{\Sigma}\end{pmatrix}A' &= \begin{pmatrix}\mathbf{I}_p & \mathbf{I}_p & \mathbf{0}\\ \mathbf{I}_p & \mathbf{0} & \mathbf{I}_p\end{pmatrix}\begin{pmatrix}\mathbf{T} & \mathbf{T}\\ n_1^{-1}\boldsymbol{\Sigma} & \mathbf{0}\\ \mathbf{0} & n_2^{-1}\boldsymbol{\Sigma}\end{pmatrix}\\ &= \begin{pmatrix}\mathbf{T}+n_1^{-1}\boldsymbol{\Sigma} & \mathbf{T}\\ \mathbf{T} & \mathbf{T}+n_2^{-1}\boldsymbol{\Sigma}\end{pmatrix}\\ &:= \overline{\boldsymbol{\Sigma}}\end{aligned}$$

we obtain the required result. □

Despite the fact that we now have shown that the numerator is multivariate normal, Lindley adds an extra variable transformation to obtain a more computational favorable result. Lindley chooses to take independent linear combinations of the vectors $\overline{\mathbf{Y}}_1$ and $\overline{\mathbf{Y}}_2$. By using these variable transformations, the numerator can be expressed as the product of the two corresponding multivariate normal densities of dimension p. Let

$$\begin{aligned}\mathbf{U} &= g_1\left(\overline{\mathbf{Y}}_1, \overline{\mathbf{Y}}_2\right) = \overline{\mathbf{Y}}_1 - \overline{\mathbf{Y}}_2,\\ \mathbf{V} &= g_2\left(\overline{\mathbf{Y}}_1, \overline{\mathbf{Y}}_2\right) = \frac{n_1\overline{\mathbf{Y}}_1 + n_2\overline{\mathbf{Y}}_2}{n_1+n_2}.\end{aligned}$$

Then (Rao (1973), p.519)

$$\begin{pmatrix}\mathbf{U}\\ \mathbf{V}\end{pmatrix} = B\begin{pmatrix}\overline{\mathbf{Y}}_1\\ \overline{\mathbf{Y}}_2\end{pmatrix} \sim \mathcal{N}_{2p}(B\overline{\boldsymbol{\mu}}, B\overline{\boldsymbol{\Sigma}}B') \tag{4.5}$$

with

$$B = \begin{pmatrix}1 & -1\\ \frac{n_1}{n_1+n_2} & \frac{n_2}{n_1+n_2}\end{pmatrix} \otimes \mathbf{I}_p,$$

where $\otimes$ is the Kronecker-product. However since

$$B\overline{\boldsymbol{\Sigma}}B' = \begin{pmatrix}n_1^{-1}\boldsymbol{\Sigma} + n_2^{-1}\boldsymbol{\Sigma} & \mathbf{0}\\ \mathbf{0} & (n_1+n_2)^{-1}\boldsymbol{\Sigma} + \mathbf{T}\end{pmatrix}$$

the covariance between $\mathbf{U}$ and $\mathbf{V}$ is zero. And thus the random vectors $\mathbf{U}$ and $\mathbf{V}$ are independently distributed (Rao (1973), p.520). By using this independence result, the

following transformation for the unconditional joint density of $\overline{\mathbf{Y}}_1$ and $\overline{\mathbf{Y}}_2$ is helpful (Rice (2007))

$$f_{\overline{\mathbf{Y}}_1,\overline{\mathbf{Y}}_2}(\overline{\mathbf{y}}_1,\overline{\mathbf{y}}_2) = f_{\mathbf{U}}\left(g_1\left(\overline{\mathbf{y}}_1,\overline{\mathbf{y}}_2\right)\right) f_{\mathbf{V}}\left(g_2\left(\overline{\mathbf{y}}_1,\overline{\mathbf{y}}_2\right)\right) \left|J\left(\overline{\mathbf{y}}_1,\overline{\mathbf{y}}_2\right)\right|, \tag{4.6}$$

where $J\left(\overline{\mathbf{y}}_1,\overline{\mathbf{y}}_2\right)$ is the Jacobian of the transformation,

$$J\left(\overline{\mathbf{y}}_1,\overline{\mathbf{y}}_2\right) = \begin{vmatrix} \partial g_1/\partial\overline{\mathbf{y}}_1 & \partial g_1/\partial\overline{\mathbf{y}}_2 \\ \partial g_2/\partial\overline{\mathbf{y}}_1 & \partial g_2/\partial\overline{\mathbf{y}}_2 \end{vmatrix} = \begin{vmatrix} 1 & -1 \\ \frac{n_1}{n_1+n_2} & \frac{n_2}{n_1+n_2} \end{vmatrix} = 1.$$

By using the identity in equation (4.6), the numerator in equation (4.4) can be written as:

$$\int_{\boldsymbol{\theta}} f_{\overline{\mathbf{Y}}_1|\boldsymbol{\theta}}(\overline{\mathbf{y}}_1 \mid \boldsymbol{\theta}) f_{\overline{\mathbf{Y}}_2|\boldsymbol{\theta}}(\overline{\mathbf{y}}_2 \mid \boldsymbol{\theta}) h(\boldsymbol{\theta}) d\boldsymbol{\theta} \;=\; f_{\mathbf{U}}\left(g_1\left(\overline{\mathbf{y}}_1,\overline{\mathbf{y}}_2\right)\right) f_{\mathbf{V}}\left(g_2\left(\overline{\mathbf{y}}_1,\overline{\mathbf{y}}_2\right)\right), \tag{4.7}$$

where the random vector $\mathbf{U}$ has a multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $(n_1^{-1}+n_2^{-1})\boldsymbol{\Sigma}$ and the random vector $\mathbf{V}$ has a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $(n_1+n_2)^{-1}\boldsymbol{\Sigma}+\mathbf{T}$ according to (4.5).

Computation of the denominator

The denominator of the likelihood ratio in equation (3.15) is given by

$$\int_{\boldsymbol{\theta}_1} f_{\overline{\mathbf{Y}}_1|\boldsymbol{\theta}_1}(\overline{\mathbf{y}}_1 \mid \boldsymbol{\theta}_1) h(\boldsymbol{\theta}_1) d\boldsymbol{\theta}_1 \int_{\boldsymbol{\theta}_2} f_{\overline{\mathbf{Y}}_2|\boldsymbol{\theta}_2}(\overline{\mathbf{y}}_2 \mid \boldsymbol{\theta}_2) h(\boldsymbol{\theta}_2) d\boldsymbol{\theta}_2, \tag{4.8}$$

i.e. the product of the unconditional densities of $\overline{\mathbf{Y}}_1$ and $\overline{\mathbf{Y}}_2$. From Section 4.1.1 it is known that $\overline{\mathbf{Y}}_1$ and $\overline{\mathbf{Y}}_2$ are normally distributed with mean vector $\boldsymbol{\mu}$ and covariance matrices $\mathbf{T}+n_1^{-1}\boldsymbol{\Sigma}$ and $\mathbf{T}+n_2^{-1}\boldsymbol{\Sigma}$, respectively. The denominator is thus the product of these two densities. If we combine this result with equation (4.7), the likelihood ratio in equation (3.15) is equal to

$$\begin{aligned} \mathrm{LR}(\overline{\mathbf{y}}_1,\overline{\mathbf{y}}_2) \;=\; & \frac{\left|(n_1^{-1}+n_2^{-1})\boldsymbol{\Sigma}\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\overline{\mathbf{y}}_1-\overline{\mathbf{y}}_2)'\left((n_1^{-1}+n_2^{-1})\boldsymbol{\Sigma}\right)^{-1}(\overline{\mathbf{y}}_1-\overline{\mathbf{y}}_2)\right\}}{\left|\mathbf{T}+n_1^{-1}\boldsymbol{\Sigma}\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\overline{\mathbf{y}}_1-\boldsymbol{\mu})'(\mathbf{T}+n_1^{-1}\boldsymbol{\Sigma})^{-1}(\overline{\mathbf{y}}_1-\boldsymbol{\mu})\right\}} \\ \times \; & \frac{\exp\left\{-\frac{1}{2}\left(\frac{n_1\overline{\mathbf{y}}_1+n_2\overline{\mathbf{y}}_2}{n_1+n_2}-\boldsymbol{\mu}\right)'\left(\mathbf{T}+\frac{\boldsymbol{\Sigma}}{n_1+n_2}\right)^{-1}\left(\frac{n_1\overline{\mathbf{y}}_1+n_2\overline{\mathbf{y}}_2}{n_1+n_2}-\boldsymbol{\mu}\right)\right\}}{\left|\mathbf{T}+n_2^{-1}\boldsymbol{\Sigma}\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\overline{\mathbf{y}}_2-\boldsymbol{\mu})'(\mathbf{T}+n_2^{-1}\boldsymbol{\Sigma})^{-1}(\overline{\mathbf{y}}_2-\boldsymbol{\mu})\right\}} \\ \times \; & \left|\left(\mathbf{T}+\frac{\boldsymbol{\Sigma}}{n_1+n_2}\right)\right|^{-\frac{1}{2}}. \end{aligned} \tag{4.9}$$

4.2.2 A Bayesian approach

The two-level model that is described in Section 3.2 can also be seen as a Bayesian statistical model, because we have a parametric statistical model $f(\overline{\mathbf{y}}_l|\boldsymbol{\theta}_l)$, $l \in \{1,2\}$, and a prior on the parameter $\boldsymbol{\theta}_l$, i.e. $h(\boldsymbol{\theta}_l)$ (Hoff (2009), chapter 8). This Bayesian paradigm is used in Bolck and Alberink (2011) to find an explicit formula for the

likelihood ratio. Instead of the likelihood ratio given in equation (3.15), they start off with the following likelihood ratio[2]:

$$\mathrm{LR}(\overline{\mathbf{y}}_1, \overline{\mathbf{y}}_2) \quad = \quad \frac{\int_{\boldsymbol{\theta}} f_{\overline{\mathbf{Y}}_2|\boldsymbol{\theta}}(\overline{\mathbf{y}}_2|\boldsymbol{\theta}) h(\boldsymbol{\theta}|\overline{\mathbf{y}}_1) d\boldsymbol{\theta}}{\int_{\boldsymbol{\theta}_2} f_{\overline{\mathbf{Y}}_2|\boldsymbol{\theta}_2}(\overline{\mathbf{y}}_2|\boldsymbol{\theta}_2) h(\boldsymbol{\theta}_2) d\boldsymbol{\theta}_2}. \tag{4.10}$$

To find the latter likelihood ratio a different method is used compared to the method described in Section 3.2. However, the formula in equation (4.10) is the same as the likelihood ratio given in equation (3.15). To see this, we use the definition of the posterior density $h(\boldsymbol{\theta} \mid \overline{\mathbf{y}}_1)$ (Robert (2007)):

$$h(\boldsymbol{\theta} \mid \overline{\mathbf{y}}_1) = \frac{f(\overline{\mathbf{y}}_1 \mid \boldsymbol{\theta}) h(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} f(\overline{\mathbf{y}}_1 \mid \boldsymbol{\theta}) h(\boldsymbol{\theta}) d\boldsymbol{\theta}}.$$

Substituting the latter equation into the likelihood ratio in (3.15) results in equation (4.10), hence the two expressions are the same.

The denominator of equation (4.10) is the unconditional density of $\overline{\mathbf{Y}}_2$. Recall from equation (4.1) that this random vector is normally distributed with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{T} + n_2^{-1}\boldsymbol{\Sigma}$ and thus the denominator is known. To compute the numerator of the likelihood ratio it is necessary to know the posterior distribution. Below it will be shown that the posterior is a normal distribution.

Lemma 4.2.2. *The posterior $h(\boldsymbol{\theta} \mid \overline{\boldsymbol{y}}_1)$ is a multivariate normal distribution with mean vector $\boldsymbol{\mu}_n = \boldsymbol{T}(\boldsymbol{T} + n_1^{-1}\boldsymbol{\Sigma})^{-1}\overline{\boldsymbol{y}}_1 + n_1^{-1}\boldsymbol{\Sigma}(\boldsymbol{T} + n_1^{-1}\boldsymbol{\Sigma})^{-1}\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{T}_n = \boldsymbol{T} - \boldsymbol{T}(\boldsymbol{T} + n_1^{-1}\boldsymbol{\Sigma})^{-1}\boldsymbol{T}$.*

Proof. The posterior distribution is proportional to the likelihood times the prior distribution,

$$h(\boldsymbol{\theta} \mid \overline{\mathbf{y}}_1) \propto f(\overline{\mathbf{y}}_1 \mid \boldsymbol{\theta}) h(\boldsymbol{\theta}).$$

Since we have assumed that $\boldsymbol{\theta}$ is normally distributed with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{T}$, it follows that

$$\begin{aligned} h(\boldsymbol{\theta}|\overline{\mathbf{y}}_1) &\propto \exp\Big(-\frac{1}{2}(\overline{\mathbf{y}}_1 - \boldsymbol{\theta})'(n_1^{-1}\boldsymbol{\Sigma})^{-1}(\overline{\mathbf{y}}_1 - \boldsymbol{\theta})\Big) \exp\Big(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu})'\mathbf{T}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu})\Big) \\ &\propto \exp\Big(-\frac{1}{2}\Big(-\boldsymbol{\theta}'((n_1^{-1}\boldsymbol{\Sigma})^{-1}\overline{\mathbf{y}}_1 + \mathbf{T}^{-1}\boldsymbol{\mu}) - (\overline{\mathbf{y}}_1'(n_1^{-1}\boldsymbol{\Sigma}_x)^{-1} + \boldsymbol{\mu}'\mathbf{T}^{-1})\boldsymbol{\theta} \\ &\quad + \boldsymbol{\theta}'((n_1^{-1}\boldsymbol{\Sigma})^{-1} + \mathbf{T}^{-1})\boldsymbol{\theta}\Big)\Big) \\ &= \exp\Big(-\frac{1}{2}\Big(-\boldsymbol{\theta}'\mathbf{T}_n^{-1}\boldsymbol{\mu}_n - \boldsymbol{\mu}_n'\mathbf{T}_n^{-1}\boldsymbol{\theta} + \boldsymbol{\theta}'\mathbf{T}_n^{-1}\boldsymbol{\theta}\Big)\Big) \end{aligned} \tag{4.11}$$

Hence,

$$\boldsymbol{\theta}|\overline{\mathbf{y}}_1 \sim \mathcal{N}_p(\boldsymbol{\mu}_n, \mathbf{T}_n).$$

To see that $\boldsymbol{\mu}_n$ and $\mathbf{T}_n$ are the same as defined in the lemma, the following two matrix identities for invertible matrices $\mathbf{A}$ and $\mathbf{B}$ will be used:

$$\begin{aligned} (\mathrm{M}_1) \quad &: \quad (\mathbf{A}^{-1} + \mathbf{B}^{-1})^{-1} = \mathbf{A} - \mathbf{A}(\mathbf{A} + \mathbf{B})^{-1}\mathbf{A} = \mathbf{B} - \mathbf{B}(\mathbf{A} + \mathbf{B})^{-1}\mathbf{B} \\ (\mathrm{M}_2) \quad &: \quad (\mathbf{A}^{-1} + \mathbf{B}^{-1})^{-1} = \mathbf{A}(\mathbf{A} + \mathbf{B})^{-1}\mathbf{B} = \mathbf{B}(\mathbf{A} + \mathbf{B})^{-1}\mathbf{A} \end{aligned}$$

[2]Note that the notation of the formula given in Bolck and Alberink (2011) is adapted to the notation used in this thesis. In Bolck and Alberink (2011) the mean vector $\boldsymbol{\theta}_2$ is replaced by $\boldsymbol{\theta}$ in combination with conditioning on the hypothesis H_d in the denominator.

A proof of these identities is given in Appendix A.1. From equation (4.11) we know that,

$$\boldsymbol{\mu}_n = \mathbf{T}_n((n_1^{-1}\boldsymbol{\Sigma})^{-1}\overline{\mathbf{y}}_1 + \mathbf{T}^{-1}\boldsymbol{\mu})$$

Apply (M_2) on $\mathbf{T}_n$ results in

$$\begin{aligned}\boldsymbol{\mu}_n &= \mathbf{T}(n_1^{-1}\boldsymbol{\Sigma} + \mathbf{T})^{-1}n_1^{-1}\boldsymbol{\Sigma}(n_1^{-1}\boldsymbol{\Sigma})^{-1}\overline{\mathbf{y}}_1 + n_1^{-1}\boldsymbol{\Sigma}(n_1^{-1}\boldsymbol{\Sigma} + \mathbf{T})^{-1}\mathbf{T}\mathbf{T}^{-1}\boldsymbol{\mu} \\ &= \mathbf{T}(\mathbf{T} + n_1^{-1}\boldsymbol{\Sigma})^{-1}\overline{\mathbf{y}}_1 + n_1^{-1}\boldsymbol{\Sigma}(\mathbf{T} + n_1^{-1}\boldsymbol{\Sigma})^{-1}\boldsymbol{\mu}.\end{aligned}$$

To find $\mathbf{T}_n$ we apply (M_1) on $\mathbf{T}_n$ and hence

$$\mathbf{T}_n = \mathbf{T} - \mathbf{T}(\mathbf{T} + n_1^{-1}\boldsymbol{\Sigma})^{-1}\mathbf{T}.$$

□

Now the posterior distribution is known, the numerator of the likelihood ratio in equation (4.10) can be computed with direct integration. These steps are shown in Appendix A.2. Using this result the following formula is found (Bolck and Alberink (2011)):

$$\mathrm{LR}(\overline{\mathbf{y}}_1, \overline{\mathbf{y}}_2) = \frac{|\mathbf{U}_n|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\overline{\mathbf{y}}_2 - \boldsymbol{\mu}_n)'\mathbf{U}_n^{-1}(\overline{\mathbf{y}}_2 - \boldsymbol{\mu}_n)\right\}}{|\mathbf{U}_0|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}((\overline{\mathbf{y}}_2 - \boldsymbol{\mu})'\mathbf{U}_0^{-1}(\overline{\mathbf{y}}_2 - \boldsymbol{\mu}))\right\}} \tag{4.12}$$

with

$$\begin{aligned}\boldsymbol{\mu}_n &= \mathbf{T}(\mathbf{T} + n_1^{-1}\boldsymbol{\Sigma})^{-1}\overline{\mathbf{y}}_1 + n_1^{-1}\boldsymbol{\Sigma}(\mathbf{T} + n_1^{-1}\boldsymbol{\Sigma})^{-1}\boldsymbol{\mu}, \\ \mathbf{T}_n &= \mathbf{T} - \mathbf{T}(\mathbf{T} + n_1^{-1}\boldsymbol{\Sigma})^{-1}\mathbf{T}, \\ \mathbf{U}_0 &= \mathbf{T} + n_2^{-1}\boldsymbol{\Sigma}, \\ \mathbf{U}_n &= \mathbf{T}_n + n_2^{-1}\boldsymbol{\Sigma}.\end{aligned}$$

4.2.3 Equality of different likelihood ratio expressions

In Section 4.2.1 and Section 4.2.2 we have seen that under a normal between-source distribution there are two different likelihood ratio formulas in forensic literature. By applying the definition of the posterior $h(\boldsymbol{\theta} \mid \overline{\mathbf{y}}_1)$ we have seen that in theory the likelihood ratio expression in equation (4.10) is the same as the likelihood ratio given in equation (3.15). However, for the ENFSI-LR project it is of great importance that the explicit formulas, given in equation (4.9) and equation (4.12), are the same as well. The ENFSI-LR project has implemented the likelihood ratio given in Bolck and Alberink (2011) and thus it is important that this formula is in harmony with the formula given in other literature (Aitken and Lucy (2004) and Zadora et al. (2014)). Therefore, below it is shown that the the likelihood ratios given in equation (4.9) and equation (4.12) are equal.

Lemma 4.2.3. *The likelihood ratios given in equation* (4.9) *and equation* (4.12) *are the same.*

The proof of this lemma is given in Appendix A.3.

5

Parameter estimation for Gaussian two-level models

In Section 4.2 an explicit formula for the likelihood ratio has been derived under the assumption of a normal between-source distribution h. To compute the likelihood ratio in equation (4.9), the mean vector $\boldsymbol{\mu}$ and covariance matrices $\mathbf{T}$ and $\boldsymbol{\Sigma}$ are required. Since these parameters are unknown, this chapter is devoted to estimation techniques for the parameters within the Gaussian two-level model.

In Section 5.1 we start off with an overview of the available background data, i.e. the data that will be used to estimate the parameters. In the ENFSI-LR project it is decided that the software must contain a "simple" plug-in estimator as default choice. Other estimators will be implemented as optional choices. Section 5.2 and Section 5.3 describe these possible plug-in estimators. Currently, forensic statisticians are discussing whether the so called "weighted mean" or "unweighted mean" should be used as an estimator for the mean. Section 5.2 therefore gives a comparison of both estimators that can help in this decision. It is generally accepted to use the analysis of variance estimators to estimate the covariance matrices $\boldsymbol{\Sigma}$ and $\mathbf{T}$. This method is briefly described in Section 5.3. However, since this estimator for $\mathbf{T}$ depends on $\boldsymbol{\mu}$, the estimator is closely related to the discussion in Section 5.2. To overcome this problem, in Section 5.3 a general formula for the analysis of variance estimator for $\mathbf{T}$ is derived, such that it easily adapts to the choice of the mean estimator. This is also useful for implementation purposes. In Section 5.4 the method of maximum likelihood is suggested to estimate the parameters. The maximum likelihood estimates can be computed iteratively using the EM-algorithm, see Section 5.4.1. In Section 5.5 the suggested estimators are compared using Monte Carlo simulation.

5.1 Background data

In the likelihood ratio given in equation (4.9), the control- and recovered data are modelled as a Gaussian two-level model. To estimate the parameters $\boldsymbol{\mu}$, $\mathbf{T}$ and $\boldsymbol{\Sigma}$ of this model, background data will be used as observations for the control- and recovered data. Let $(\mathbf{Z}_{ij} \mid 1 \leq i \leq m, 1 \leq j \leq n_i)$ denote the background data that consist of m groups (batches) and n_i measured tablets within each batch, such that

$$\mathbf{Z}_{ij} \quad = \quad \text{measurement vector of } p \text{ characteristics within batch } i \text{ on tablet } j.$$

This data set can be *unbalanced*, because the number of measured tablets n_i can differ in each group. The data set would be called *balanced* if the number of measured tablets

is the same in each batch, i.e. $n_i = n_{i'}\ \forall i, i' \in \{1, \ldots, m\}$. In practice, the background data will consist of seized xtc consignments. The data will often be unbalanced, because it is infeasible to measure the same number of tablets in each seized xtc consignment.

Because the background data can be seen as m repeated observations of xtc batches, we model the background data by the Gaussian two-level model that is used to model the control- and recovered data (see Section 3.2),

$$\mathbf{Z}_{ij} \quad = \quad \boldsymbol{\theta}_i + \boldsymbol{\varepsilon}_{ij} \quad \text{for } i = 1, \ldots, m, \quad j = 1, \ldots, n_i \tag{5.1}$$

with

$$\begin{aligned} \boldsymbol{\theta}_i &\overset{\text{iid}}{\sim} \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{T}) \qquad \forall i, \\ \boldsymbol{\varepsilon}_{ij} &\overset{\text{iid}}{\sim} \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}) \qquad \forall i, j, \end{aligned}$$

and

$$\boldsymbol{\theta}_i \quad \perp \quad \boldsymbol{\varepsilon}_{i'j} \quad \forall i, i'.$$

Hence,

$$\mathbf{Z}_{ij} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma} + \mathbf{T}).$$

A schematic representation of this model is given in Figure 5.1, which is an extension of Figure 3.1.

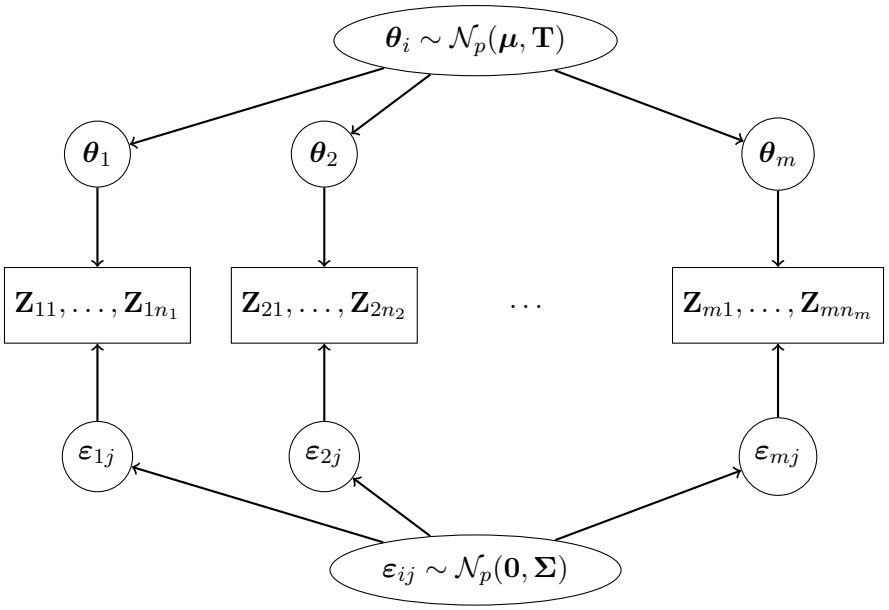


Figure 5.1: Schematic representation of the Gaussian two-level model for the background data.

In Section 5.3 we are also interested in the distribution of the batch means $\overline{\mathbf{Z}}_i, i \in \{1, \ldots, m\}$. Recall that in Section 4.1.2 we have used these m batch means of p features to evaluate the assumption of normality, i.e.

$$\overline{\mathbf{Z}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{Z}_{ij} = \boldsymbol{\theta}_i + \overline{\boldsymbol{\varepsilon}_i} \qquad i \in \{1, \ldots, m\} \tag{5.2}$$

such that by independence,

$$\overline{\mathbf{Z}}_i \sim \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{T} + n_i^{-1}\boldsymbol{\Sigma}). \tag{5.3}$$

To estimate the parameters $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ and $\mathbf{T}$ suitable background data should thus be available. Nowadays, within an increasing number of forensic fields people try to create such background databases. Naturally these databases should contain measurements on the required features and they need to contain a sufficient amount of data. However, as we have already mentioned in Chapter 3, it takes a lot of time and resources to accomplish these goals.

Remarks:

- As mentioned in Section 4.1, the described two-level model is a multivariate random effects model. In the literature it is common to write $\boldsymbol{\theta}_i$ as the sum of its mean $\boldsymbol{\mu}$ and the random group effect $\boldsymbol{\alpha}_i$, i.e. $\boldsymbol{\theta}_i = \boldsymbol{\mu} + \boldsymbol{\alpha}_i$ where $\boldsymbol{\alpha}_i \sim \mathcal{N}_p(\mathbf{0}, \mathbf{T})$. Recognizing this as a random effects model is helpful for a literature study to parameter estimation in such models.

- The notation used to model the background data is the same as the notation used to model the control- and recovered data in Chapter 3. In this section the parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ and the group sizes n_1 and n_2 are the batch means and group sizes of two random batches in the background data respectively. Thus, these quantities do not correspond to the control- and recovered data as modelled in Section 3.2.1.

Background information

In the likelihood ratio in equation (3.15), the densities f and h are conditioned on the background information I. In Section 3.1 we have discussed the example that the total database contains xtc consignments from Europe, but from the background information it follows that we should only consider xtc batches that originate from the Netherlands. In this case a subset of the database as described above should be used and thus the estimates of the parameters will be based on xtc consignments from the Netherlands instead of from Europe. Thus, conditioning on the background information I could influence the background data that should be used. Hence, conditioning on the background information affects the estimated parameters $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ and $\mathbf{T}$.

In theory, the background information can only be used when the evidence E depends in a certain way on this information. For example, if the evidence exists of the weights of the tablets, it may depend on the country of origin of the tablets. But the kind of bags the tablets are found in, does probably not influence the weights and hence such information cannot be used.

It can thus be difficult to determine in what way the evidence depends on the background information and how it can be used. Moreover, earlier it was mentioned that it is hard to create a suitable background database. Hence, one can imagine that a background database based on the background information is even harder to obtain. Despite the fact that in practice this remains a bit of a black box, in theory we condition on the background information I.

5.2 Estimating the mean

In this section plug-in estimators for the mean vector $\boldsymbol{\mu}$ are examined. Between forensic statisticians an active discussion exist about whether the weighted- or the unweighted mean should be used. Therefore, in Section 5.2.1 these estimators are compared based on the mean squared error. It will be shown that both estimators are unbiased. Hence, it is interesting to examine the variances of the estimators and in particular which one has smallest variance. We will see that this depends on the relation between unknown parameters. Section 5.2.2 gives a generalized weighted mean as an alternative to the weighted- and the unweighted mean.

5.2.1 Weighted- versus unweighted mean

Two natural estimators for the mean vector $\boldsymbol{\mu}$ are the *weighted mean* and the *unweighted mean*. The weighted mean is the average over all observations $\mathbf{Z}_{ij}$ (Searle (1992)),

$$\hat{\boldsymbol{\mu}}_w = \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\mathbf{Z}_{ij}, \tag{5.4}$$

where N is the total number of observations, i.e. $N = \sum_{i=1}^{m} n_i$. The unweighted mean is the average over the average of the observations $\mathbf{Z}_{ij}$ in each group (Sahai and Ojeda (2005)):

$$\hat{\boldsymbol{\mu}}_u = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{Z}_{ij} = \frac{1}{m}\sum_{i=1}^{m}\overline{\mathbf{Z}}_i. \tag{5.5}$$

First note that if the data is balanced, i.e. $n_i = n_{i'}\ \forall i, i' \in \{1, \ldots, m\}$, the weighted- and unweighted average are exactly the same. For unbalanced data, some prefer the weighted average and others the unweighted average (Aitken and Lucy (2004), Bolck and Alberink (2011)). The fact that the weighted average is more robust to single outliers in a group and has least variance are arguments in favor of the weighted average. An argument in favor of the unweighted average is that in practical forensic research it is beneficial that groups have equal importance, despite the number of observations. In fact, it is shown in this section that the best choice depends on the situation. For example, if the error ε_{ij} is small on average i.e. if $\overline{\mathbf{Z}}_i \approx \boldsymbol{\theta}_i$, then the unweighted average is the maximum likelihood estimator and we would prefer this one. To make such statements more explicit, the mean squared errors of both estimators will be compared. For ease of notation this comparison is done for one dimension, i.e. $p = 1$, such that the mean vector $\boldsymbol{\mu}$ boils down to a single parameter μ and covariance matrices $\mathbf{T}$ and $\boldsymbol{\Sigma}$ are equal to τ^2 and σ^2, respectively. At the end of this section the results will be extended to the multivariate case ($p > 1$).

First, we show that both estimators are unbiased,

$$\mathrm{E}\left(\hat{\mu}_w\right) = \frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}\mathrm{E}\left(Z_{ij}\right) = \frac{1}{N}\sum_{i}^{m} n_i\mu = \mu$$

and

$$\mathrm{E}\left(\hat{\mu}_u\right) = \frac{1}{m}\sum_{i}^{m}\mathrm{E}\left(\overline{Z}_i\right) = \frac{1}{m}\sum_{i}^{m}\mu = \mu$$

by using equation (5.3). For the variance of the weighted mean it follows that

$$\begin{aligned}\mathrm{Var}\left(\hat{\mu}_w\right) &= \frac{1}{N^2}\sum_{i=1}^{m}\mathrm{Var}\left(\sum_{j=1}^{n_i} Z_{ij}\right)\\ &= \frac{1}{N^2}\sum_{i=1}^{m}\mathrm{Var}\left(n_i\theta_i + \sum_{j=1}^{n_i}\varepsilon_{ij}\right)\end{aligned}$$

by the definition in (5.1). Because of independence it then follows that

$$\begin{aligned}\mathrm{Var}\left(\hat{\mu}_w\right) &= \frac{1}{N^2}\sum_{i=1}^{m}\left\{\mathrm{Var}\left(n_i\theta_i\right) + \sum_{j=1}^{n_i}\mathrm{Var}\left(\varepsilon_{ij}\right)\right\}\\ &= \frac{1}{N^2}\sum_{i=1}^{m}\left\{n_i^2\tau^2 + n_i\sigma^2\right\}.\end{aligned}$$

For the variance of the unweighted mean we have

$$\begin{aligned}\mathrm{Var}\left(\hat{\mu}_u\right) &= \frac{1}{m^2}\sum_{i=1}^{m}\mathrm{Var}\left(\theta_i + \bar{\varepsilon}_i\right)\\ &= \frac{1}{m^2}\sum_{i=1}^{m}\left\{\tau^2 + \frac{\sigma^2}{n_i}\right\}.\end{aligned}$$

Because both estimators are unbiased, the mean squared error of the estimator is equal to its variance. We will therefore examine which estimator has smallest variance. Hence, consider the efficiency of $\hat{\mu}_u$ relative to $\hat{\mu}_w$ (Rice (2007)):

$$\begin{aligned}\mathrm{eff}(\hat{\mu}_u, \hat{\mu}_w) &= \frac{\mathrm{Var}(\hat{\mu}_w)}{\mathrm{Var}(\hat{\mu}_u)}\\ &= \frac{\frac{1}{N^2}\sum_{i=1}^{m}\left\{n_i^2\tau^2 + n_i\sigma^2\right\}}{\frac{1}{m^2}\sum_{i=1}^{m}\left\{\tau^2 + \frac{\sigma^2}{n_i}\right\}}\\ &= \frac{\frac{\tau^2}{N^2}\sum_{i=1}^{m} n_i^2 + \frac{\sigma^2}{N}}{\frac{\tau^2}{m} + \frac{\sigma^2}{m^2}\sum_{i=1}^{m}\frac{1}{n_i}}.\end{aligned} \tag{5.6}$$

By Jensen's inequality it follows that

$$\frac{1}{m}\sum_{i=1}^{m} n_i^2 \geq \left(\frac{1}{m}\sum_{i=1}^{m} n_i\right)^2 = \frac{N^2}{m^2} \tag{5.7}$$

and thus,

$$\frac{\tau^2}{N^2}\sum_{i=1}^{m} n_i^2 \geq \frac{\tau^2}{N^2}\frac{mN^2}{m^2} = \frac{\tau^2}{m}. \tag{5.8}$$

This inequality refers to the first terms in the numerator and denominator of equation (5.6). So, if we can show that the inequality $\frac{\sigma^2}{N} \geq \frac{\sigma^2}{m^2}\sum_{i=1}^{m}\frac{1}{n_i}$ is true (the second terms), then we have shown that the variance of the weighted mean is always bigger

than the variance of the unweighted mean. However, again by using Jensen's inequality it follows that

$$\frac{1}{m}\sum_{i=1}^{m}\frac{1}{n_i} \geq \frac{1}{\frac{1}{m}\sum_{i=1}^{m} n_i} = \frac{m}{N}. \tag{5.9}$$

Here we have used that the function $\phi(x) = \frac{1}{x}$ is convex for $x > 0$, which is all we need since we are only considering positive values. Using the latter inequality it thus follows that

$$\frac{\sigma^2}{m^2}\sum_{i=1}^{m}\frac{1}{n_i} \geq \frac{\sigma^2}{N}. \tag{5.10}$$

Furthermore, whether the term $\frac{\tau^2}{N^2}\sum_{i=1}^{m} n_i^2$ is greater or equal than the denominator, completely depends on the magnitudes of the quantities. Consequently, the efficiency of $\hat{\mu}_u$ relative to $\hat{\mu}_w$ is not always bigger than one. On the other hand, the efficiency of $\hat{\mu}_u$ relative to $\hat{\mu}_w$ is not always smaller than one either. This can be seen from equation (5.10) and by the fact that the term $\frac{\sigma^2}{m^2}\sum_{i=1}^{m}\frac{1}{n_i}$ is not always bigger than the numerator. Thus, it can be concluded that we cannot be conclusive about which estimator has smallest variance. However, certain conditions can be determined for the quotient in equation (5.6) to be greater or smaller than one. To make things easier, multiply the quotient in equation (5.6) with the term m^2N^2 in the numerator and denominator and let $r = \frac{\sigma^2}{\tau^2}$. Then

$$\text{eff}(\hat{\mu}_u, \hat{\mu}_w) \quad = \quad \frac{m^2\sum_{i=1}^{m} n_i^2 + m^2 N r}{mN^2 + rN^2\sum_{i=1}^{m}\frac{1}{n_i}}. \tag{5.11}$$

It can be seen that

$$\text{eff}(\hat{\mu}_u, \hat{\mu}_w) > 1 \quad \text{if} \quad r\left(m^2N - N^2\sum_{i=1}^{m}\frac{1}{n_i}\right) > mN^2 - m^2\sum_{i=1}^{m} n_i^2.$$

Note that the term $m^2N - N^2\sum_{i=1}^{m}\frac{1}{n_i}$ is always negative, because of equation (5.9). Hence,

$$\text{eff}(\hat{\mu}_u, \hat{\mu}_w) > 1 \quad \text{if} \quad r < \frac{mN^2 - m^2\sum_{i=1}^{m} n_i^2}{m^2N - N^2\sum_{i=1}^{m}\frac{1}{n_i}} := c. \tag{5.12}$$

Furthermore, the term $mN^2 - m^2\sum_{i=1}^{m} n_i^2$ is always negative as well because of equation (5.7), such that the constant c is always positive as required. Therefore,

$$\begin{cases} \text{Var}(\hat{\mu}_w) > \text{Var}(\hat{\mu}_u) & \text{if } \sigma^2 < c\tau^2, \\ \text{Var}(\hat{\mu}_w) < \text{Var}(\hat{\mu}_u) & \text{if } \sigma^2 > c\tau^2, \end{cases} \tag{5.13}$$

Thus which variance is the smallest, and correspondingly which estimator is the best choice, depends on the proportion between σ and τ. So if the error is small, i.e. the variance σ^2 is small and can assumed to be smaller than $c\tau^2$, then the variance of the weighted mean is bigger than the variance of the unweighted mean and one should prefer the unweighted mean. This example corresponds with the example which we have seen in the beginning of this section, i.e if $\overline{\mathbf{Z}}_i \approx \boldsymbol{\theta}_i$ then we would prefer the unweighted average.

To decide which estimator is best to use, one should thus have certain prior knowledge about the proportion between the variances σ^2 and τ^2. However, because these quantities are unknown this choice is combined with uncertainty.

For the multivariate case, instead of estimating a single mean μ, a whole vector of means should be estimated (see equation (5.4) and (5.5)). It is easy to see that for the multivariate case, both estimators are still unbiased. But instead of the variance of the estimator, we now have a covariance matrix of the mean vector to consider. Hence, the variance of each of the p components of the mean vector estimator should thus be compared as described in (5.13). In practice the choice between the two estimators based on this condition, would be obviously harder than in the one dimensional case.

5.2.2 Generalized weighted mean

This section suggests a more general estimator for the mean than the weighted- and unweighted mean which are described in Section 5.2.1. We will refer to this more general estimator as the *generalized weighted mean*[1]. We start off with the estimator for the univariate model. At the end of this section the estimator is extended to the multivariate model.

Define the generalized weighted mean as (Rice (2007))

$$\hat{\mu} = \sum_{i=1}^{m} w_i \overline{Z}_i \qquad \text{where} \qquad \sum_{i=1}^{m} w_i = 1. \tag{5.14}$$

Recall from Section 5.1 that $\overline{Z}_i \sim \mathcal{N}(\mu, \tau^2+\sigma^2/n_i)$. From the constraint $w_1+\cdots+w_m = 1$ it then follows that the generalized weighted mean is unbiased,

$$\mathrm{E}(\hat{\mu}) = \sum_{i=1}^{m} w_i \,\mathrm{E}\left(\overline{Z}_i\right) = \sum_{i=1}^{m} w_i \mu = \mu.$$

The variance of $\hat{\mu}$ is equal to

$$\mathrm{Var}(\hat{\mu}) = \sum_{i=1}^{m} w_i^2 \,\mathrm{Var}\left(\overline{Z}_i\right) = \sum_{i=1}^{m} w_i^2 \left(\tau^2 + \sigma^2/n_i\right).$$

Since the variance depends on the choice of the weights w_i, $i = 1, \ldots, m$, the question arises how to choose these weights to minimize $\mathrm{Var}(\hat{\mu})$ subject to the constraint $w_1 + \cdots + w_m = 1$ (Rice (2007)).

Lemma 5.2.1. *The weights $w_1, \ldots, w_m$ that minimize $\mathrm{Var}(\hat{\mu})$ subject tot the constraint $w_1 + \cdots + w_m = 1$ are given by*

$$w_i = \frac{1}{\left(\tau^2 + \sigma^2/n_i\right) \sum_{i=1}^{m} \left(\tau^2 + \sigma^2/n_i\right)^{-1}} \tag{5.15}$$

where $i = 1, \ldots, m$.

Proof. To minimize $\mathrm{Var}(\hat{\mu})$ subject to the constraint $w_1 + \cdots + w_m = 1$ we introduce a Lagrange multiplier λ such that the Lagrange function is equal to:

$$\mathcal{L}_\lambda(w_1, \ldots, w_m, \lambda) = \sum_{i=1}^{m} w_i^2 \left(\tau^2 + \sigma^2/n_i\right) - \lambda \left(\sum_{i=1}^{m} w_i - 1 \right).$$

[1] In the literature this estimator is called the weighted mean. However, in forensic literature the estimator in equation (5.4) is called the weighted mean. Therefore we will refer to this estimator as the generalized weighted mean.

We will minimize the Lagrange function over $\mathbb{R}^m$. For $i = 1, \ldots, m$ we have

$$\frac{\partial \mathcal{L}_\lambda}{\partial w_i} = 2\left(\tau^2 + \sigma^2/n_i\right) w_i - \lambda.$$

Setting these partial derivatives equal to zero, we have the system of equations

$$w_i = \frac{\lambda}{2\left(\tau^2 + \sigma^2/n_i\right)}.$$

Now using the constraint $\sum_{i=1}^m w_i = 1$ gives

$$\sum_{i=1}^m \frac{\lambda}{2\left(\tau^2 + \sigma^2/n_i\right)} = 1.$$

Hence,

$$\lambda = \frac{1}{\sum_{i=1}^m \frac{1}{2}\left(\tau^2 + \sigma^2/n_i\right)^{-1}}.$$

Thus,

$$w_i = \frac{1}{\left(\tau^2 + \sigma^2/n_i\right)\sum_{i=1}^m \left(\tau^2 + \sigma^2/n_i\right)^{-1}}$$

which proves the lemma. □

This lemma shows that for the weights in equation (5.15) the generalized weighted mean is optimal, i.e.

$$\hat{\mu}_{\text{opt}} = \frac{\sum_{i=1}^m \left(\tau^2 + \sigma^2/n_i\right)^{-1} \overline{Z}_i}{\sum_{i=1}^m \left(\tau^2 + \sigma^2/n_i\right)^{-1}}. \tag{5.16}$$

Note that the generalized weighted mean can only be called an estimator if the parameters σ^2 and τ^2 are known. When we hereafter refer to $\hat{\mu}_{\text{opt}}$ as an estimator we are assuming that σ^2 and τ^2 are known.

The weighted- and unweighted mean are special cases of the generalized weighted mean given in equation (5.14). It can be seen that the weighted mean $\hat{\mu}_w$ is the generalized weighted mean with weights $w_i = n_i N^{-1}\ \forall i$. The unweighted mean $\hat{\mu}_u$ is the generalized weighted mean with weights $w_i = m^{-1}\ \forall i$. Since the weights in equation (5.15) yield the minimum variance for $\hat{\mu}$ we can thus conclude that, if the parameters σ^2 and τ^2 are known, $\hat{\mu}_{\text{opt}}$ is the best of these three estimators.

Further it can be noticed that, if the parameters σ^2 and τ^2 are known, the generalized weighted mean with optimal weights is the maximum likelihood estimator. In fact, $\overline{Z}_1, \ldots, \overline{Z}_m$, are independent and identically distributed normal random variables such that their joint density is the product of the marginal densities. The log likelihood is thus equal to

$$-\frac{m}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^m \log\left(\tau^2 + \sigma^2/n_i\right) - \frac{1}{2}\sum_{i=1}^m \frac{(\overline{Z}_i - \mu)^2}{\tau^2 + \sigma^2/n_i}.$$

To find the maximum likelihood estimator for μ we thus want to minimize the term

$$\sum_{i=1}^m \frac{(\overline{Z}_i - \mu)^2}{\tau^2 + \sigma^2/n_i}$$

with respect to μ. Taking the derivative of the latter expression with respect to μ and setting this derivative equal to zero then gives exactly the estimator in equation (5.16). The fact that this generalized weighted mean is the maximum likelihood estimator is not surprising, because in many situations the maximum likelihood estimator is the minimum variance estimator.

Although in theory the generalized weighted mean $\hat{\mu}_{\text{opt}}$ is the minimum variance unbiased estimator, in practice this is not true. The weights in equation (5.15) depend on the parameters τ^2 and σ^2, which are unknown. Therefore, estimated values of these parameters should be substituted which will have influence on the variance of the generalized weighted mean. So in conclusion we theoretically derived that for known σ^2 and τ^2 the generalized weighted average is better than the weighted- and unweighted mean. In practice, however, this result is not necessarily true after substituting estimates for τ^2 and σ^2. In Section 5.5 this issue will be investigated by comparing the mean squared errors of the estimators in a simulation study. Another interesting question is to investigate what the effect of a non-optimal estimator is on the likelihood ratio. Due to time limitations this is not covered in this thesis. Nevertheless, since forensic experts currently not always report the numerical value of the likelihood ratio but only the verbal scale, it is likely that the effect of the estimator on the verbal scale is rather small.

In the one-dimensional case it is a natural choice to minimize the variance of the generalized weighted mean to obtain optimal weights. To extend the generalized weighted mean to the multivariate case, instead of the variance we now have a covariance matrix of the generalized weighted mean to consider. Hence, a choice should be made which object will be minimized. Since it is desired to minimize the variances of the multivariate generalized weighted mean, i.e. the diagonal of the covariance matrix, rather than the covariances of the generalized weighted mean, we choose to minimize the trace of the covariance matrix. Then, again using a Lagrange multiplier it can be found that the optimal weights are (see Appendix A.5)

$$\mathbf{w}_i = \left(\sum_{i=1}^{m} \left(\mathbf{T} + n_i^{-1} \boldsymbol{\Sigma} \right)^{-1} \right)^{-1} \left(\mathbf{T} + n_i^{-1} \boldsymbol{\Sigma} \right)^{-1}$$

for $i = 1, \ldots, m$ such that

$$\hat{\boldsymbol{\mu}}_{\mathbf{opt}} = \left(\sum_{i=1}^{m} (\mathbf{T} + n_i^{-1} \boldsymbol{\Sigma})^{-1} \right)^{-1} \left(\sum_{i=1}^{m} (\mathbf{T} + n_i^{-1} \boldsymbol{\Sigma})^{-1} \overline{\mathbf{Z}}_i \right) \tag{5.17}$$

is the optimal generalized weighted mean.

5.3 Analysis of variance estimators

In random effects models the analysis of variance technique is commonly used to estimate the within covariance matrix $\boldsymbol{\Sigma}$ and the between covariance matrix $\mathbf{T}$. The analysis of variance estimation is based on the following identity:

$$\begin{aligned} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(\mathbf{Z}_{ij} - \overline{\mathbf{Z}} \right) \left(\mathbf{Z}_{ij} - \overline{\mathbf{Z}} \right)' &= \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(\mathbf{Z}_{ij} - \overline{\mathbf{Z}}_i \right) \left(\mathbf{Z}_{ij} - \overline{\mathbf{Z}}_i \right)' \\ &+ \sum_{i=1}^{m} n_i \left(\overline{\mathbf{Z}}_i - \overline{\mathbf{Z}} \right) \left(\overline{\mathbf{Z}}_i - \overline{\mathbf{Z}} \right)'. \end{aligned} \tag{5.18}$$

Here $\overline{\mathbf{Z}}_i$ and $\overline{\mathbf{Z}}$ are the group mean and the overall mean, respectively. In Rice (2007) the latter identity is showed for balanced one-dimensional data. For unbalanced higher dimensional data the idea is roughly the same, this is shown in Appendix A.4 for convenience. The left-hand side of equation (5.18) is called the *total sum of squares*. The first- and second term on the right-hand side of the identity are called the *within group sums of squares* (SS_W) and the *between group sum of squares* (SS_B), respectively. Due to the outer products these quantities are $p \times p$ matrices that represent variation within groups (SS_W) and between groups (SS_B). The idea of analysis of variance estimation is to derive the expected values of SS_B and SS_W. Subsequently, these expected values can be equated to the observed values for SS_B and SS_W. These equations need to be solved for the matrices $\mathbf{\Sigma}$ and $\mathbf{T}$ to obtain the analysis of variance estimators.

For the within group sum of squares it can be shown that

$$\mathrm{E}\left(SS_W\right) = \mathbf{\Sigma}(N-m). \tag{5.19}$$

The derivation of this expectation for one dimensional problems can be found in several places in the literature (e.g. Searle (1992) or Sahai and Ojeda (2005)). The multivariate derivation is an extension and is given in Appendix A.6. The expectation of SS_W can now be equated to the observed value of SS_W, such that the analysis of variance estimator for $\mathbf{\Sigma}$ is found

$$\widehat{\mathbf{\Sigma}} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(\mathbf{z}_{ij}-\overline{\mathbf{z}}_i\right)\left(\mathbf{z}_{ij}-\overline{\mathbf{z}}_i\right)'}{N-m}. \tag{5.20}$$

To compute the expectation of the between group sum of squares, the definition of $\overline{\mathbf{Z}}$ is important. In the literature it is common to use the weighted average as described in Section 5.2.1. The corresponding estimator for $\mathbf{T}$ will thus depend on this choice, see Searle et al. (1992). In Sahai and Ojeda (2005) a remark is given for the analysis of variance estimator for $\mathbf{T}$ using the unweighted average for $\overline{\mathbf{Z}}$. In Section 5.2 we have seen that we cannot be exclusive in our choice between the weighted- and unweighted average for the overall mean. In theory, this choice should depend on the relation between the within- and between variances. In fact, neither the weighted- or unweighted average is the minimum variance estimator, because in theory this is the generalized weighted mean as given in equation (5.20). We have mentioned that all these estimators can be expressed as generalized weighted averages with different weights, see equation (5.14). Therefore below the expectation for SS_W is computed in terms of a general estimator for $\overline{\mathbf{Z}}$. Hence, the estimator for $\mathbf{T}$ will be in terms of this expression. Depending on the choice of the estimator for $\overline{\mathbf{Z}}$, the corresponding weights should be substituted in the estimator for $\mathbf{T}$. In this way, the estimator for $\mathbf{T}$ can be easily adapted to each situation. Obviously, $\widehat{\mathbf{\Sigma}}$ is not affected by this choice and will remain as given in equation (5.20). Let $\overline{\mathbf{Z}}$ be a generalized weighted average

$$\overline{\mathbf{Z}} = \sum_{i=1}^{m} \mathbf{w}_i \overline{\mathbf{Z}}_i.$$

The expectation of the between group sum of squares is equal to

$$\mathrm{E}\left(SS_B\right) = \sum_{i=1}^{m} n_i \left\{ \mathrm{E}\left(\overline{\mathbf{Z}}_i\overline{\mathbf{Z}}_i'\right) - \mathrm{E}\left(\overline{\mathbf{Z}}_i\overline{\mathbf{Z}}'\right) - \mathrm{E}\left(\overline{\mathbf{Z}}\overline{\mathbf{Z}}_i'\right) + \mathrm{E}\left(\overline{\mathbf{Z}}\overline{\mathbf{Z}}'\right) \right\}.$$

Using equation (5.3) we have

$$\begin{aligned} \mathrm{E}\left(\overline{\mathbf{Z}}_i\overline{\mathbf{Z}}_i'\right) &= \mathrm{Cov}\left(\overline{\mathbf{Z}}_i, \overline{\mathbf{Z}}_i\right) + \mathrm{E}\left(\overline{\mathbf{Z}}_i\right)\mathrm{E}\left(\overline{\mathbf{Z}}_i\right)' \\ &= n_i^{-1}\mathbf{\Sigma} + \mathbf{T} + \boldsymbol{\mu}\boldsymbol{\mu}'. \end{aligned}$$

Using the same approach to compute the expectation $\mathrm{E}\left(\overline{\mathbf{Z}}_i\overline{\mathbf{Z}}'\right)$, the covariance between $\overline{\mathbf{Z}}_i$ and $\overline{\mathbf{Z}}$ is thus required:

$$\mathrm{Cov}\left(\overline{\mathbf{Z}}_i, \overline{\mathbf{Z}}\right) = \mathrm{Cov}\left(\overline{\mathbf{Z}}_i, \sum_{r=1}^{m} \mathbf{w}_r \overline{\mathbf{Z}}_r\right)$$

By independence of $\overline{\mathbf{Z}}_i$ and $\overline{\mathbf{Z}}_i'$ for all $i \neq i'$ it follows that

$$\begin{aligned}\mathrm{Cov}\left(\overline{\mathbf{Z}}_i, \overline{\mathbf{Z}}\right) &= \mathrm{Cov}\left(\overline{\mathbf{Z}}_i, \mathbf{w}_i\overline{\mathbf{Z}}_i\right) \\ &= \mathbf{w}_i \,\mathrm{Cov}\left(\overline{\mathbf{Z}}_i, \overline{\mathbf{Z}}_i\right).\end{aligned}$$

Since the generalized weighted average for $\overline{\mathbf{Z}}$ is unbiased, i.e. $\mathrm{E}\left(\overline{\mathbf{Z}}\right) = \boldsymbol{\mu}$, we now have

$$\mathrm{E}\left(\overline{\mathbf{Z}}_i\overline{\mathbf{Z}}'\right) = \mathbf{w}_i\left(\mathbf{T} + n_i^{-1}\boldsymbol{\Sigma}\right) + \boldsymbol{\mu}\boldsymbol{\mu}'.$$

This computation holds for the expectation $\mathrm{E}\left(\overline{\mathbf{Z}}\overline{\mathbf{Z}}_i'\right)$ as well. To compute the expectation $\mathrm{E}\left(\overline{\mathbf{Z}}\overline{\mathbf{Z}}'\right)$, again by independence it follows that

$$\begin{aligned}\mathrm{Cov}\left(\overline{\mathbf{Z}}, \overline{\mathbf{Z}}\right) &= \sum_{i=1}^{m} \mathrm{Cov}\left(\mathbf{w}_i\overline{\mathbf{Z}}_i, \mathbf{w}_i\overline{\mathbf{Z}}_i\right) \\ &= \sum_{i=1}^{m} \mathbf{w}_i^2\left(\mathbf{T} + n_i^{-1}\boldsymbol{\Sigma}\right).\end{aligned}$$

Thus, we have

$$\begin{aligned}\mathrm{E}\left(SS_B\right) &= \sum_{i=1}^{m} n_i\left\{\mathbf{T} + n_i^{-1}\boldsymbol{\Sigma} - 2\mathbf{w}_i\left(\mathbf{T} + n_i^{-1}\boldsymbol{\Sigma}\right) + \sum_{r=1}^{m}\mathbf{w}_r^2\left(\mathbf{T} + n_r^{-1}\boldsymbol{\Sigma}\right)\right\} \\ &= \sum_{i=1}^{m} n_i\left\{\boldsymbol{\Sigma}\left(n_i^{-1} - 2\mathbf{w}_i n_i^{-1} + \sum_{r=1}^{m}\mathbf{w}_r^2 n_r^{-1}\right)\right\} \\ &+ \sum_{i=1}^{m} n_i\left\{\mathbf{T}\left(1 - 2\mathbf{w}_i + \sum_{r=1}^{m}\mathbf{w}_r^2\right)\right\} \\ &= \boldsymbol{\Sigma}\left(m - 2\sum_{i=1}^{m}\mathbf{w}_i + N\sum_{r=1}^{m}\mathbf{w}_r^2 n_r^{-1}\right) + \mathbf{T}\left(N - 2\sum_{i=1}^{m} n_i\mathbf{w}_i + N\sum_{r=1}^{m}\mathbf{w}_r^2\right).\end{aligned}$$

Equating this expectation to the observed value of SS_B then gives the estimator

$$\widehat{\mathbf{T}} = \frac{\sum_{i=1}^{m} n_i\left(\overline{\mathbf{Z}}_i - \overline{\mathbf{Z}}\right)\left(\overline{\mathbf{Z}}_i - \overline{\mathbf{Z}}\right)' - \widehat{\boldsymbol{\Sigma}}\left(m - 2\sum_{i=1}^{m}\mathbf{w}_i + N\sum_{r=1}^{m}\mathbf{w}_r^2 n_r^{-1}\right)}{\left(N - 2\sum_{i=1}^{m} n_i\mathbf{w}_i + N\sum_{r=1}^{m}\mathbf{w}_r^2\right)}. \quad (5.21)$$

Thus, if for example the weighted average is chosen for $\overline{\mathbf{Z}}$, then we have seen in Section 5.2.2 that the weights should equal $\mathbf{w}_i = n_i N^{-1}$. Hence, substituting these weights in equation (5.21) will give the same estimator for $\mathbf{T}$ as is derived in for example Searle et al. (1992).

A difficulty related to these estimators is the fact that for some data it can happen that the variance estimates are negative. In that case it could be that the wrong model is used or it may be an indication that the true value is zero, because it is an unbiased estimator (Searle et al. (1992)). To avoid the possibility of negative estimates, other estimators can be used. An example is maximum likelihood estimation, which is described in the following section.

5.4 Maximum likelihood estimation

In this section the method of maximum likelihood is discussed to estimate the parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\mathbf{T}$. First, the likelihood function will be derived from which it can be seen that there exist no explicit formulas to estimate the parameters. In Section 5.4.1 the EM-algorithm is suggested as an iterative method to solve this problem. Let $\boldsymbol{\Psi} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{T})$ and consider the background data $\mathbf{Z} = (\mathbf{Z}_{ij}, 1 \leq i \leq m, 1 \leq j \leq n_i)$ as described in Section 5.1. The joint density of the background data is

$$f_{\boldsymbol{\Psi}}(\mathbf{z}) \quad = \quad \prod_{i=1}^{m} f_{\boldsymbol{\Psi}}(\mathbf{z}_{i1}, \ldots, \mathbf{z}_{in_i}), \tag{5.22}$$

because the observations are independent between groups. The observations within group i are not independent and thus the joint density of $\mathbf{Z}_{i1}, \ldots, \mathbf{Z}_{in_i}$ is needed.

Lemma 5.4.1. *The joint density of $\boldsymbol{Z}_{i1}, \ldots, \boldsymbol{Z}_{in_i}$ is multivariate normal with mean vector* $\begin{pmatrix} \boldsymbol{\mu} \\ \vdots \\ \boldsymbol{\mu} \end{pmatrix}$ *and covariance matrix* $\boldsymbol{\Sigma}_i = \begin{pmatrix} \boldsymbol{T}+\boldsymbol{\Sigma} & & \boldsymbol{T} \\ & \ddots & \\ \boldsymbol{T} & & \boldsymbol{T}+\boldsymbol{\Sigma} \end{pmatrix}$.

Proof. To show this result an extension of the steps used in the proof of lemma 4.2.1 can be applied, i.e.

$$\begin{pmatrix} \mathbf{Z}_{i1} \\ \mathbf{Z}_{i2} \\ \vdots \\ \mathbf{Z}_{in_1} \end{pmatrix} = A \begin{pmatrix} \boldsymbol{\theta}_i \\ \boldsymbol{\varepsilon}_{i1} \\ \vdots \\ \boldsymbol{\varepsilon}_{in_i} \end{pmatrix} \sim \mathcal{N}_{pn_i} \left(A \begin{pmatrix} \boldsymbol{\mu} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix}, A \begin{pmatrix} \mathbf{T} & \mathbf{0} & \ldots & \ldots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma} & \mathbf{0} & \ldots & \mathbf{0} \\ \vdots & \mathbf{0} & \ddots & \mathbf{0} & \vdots \\ \vdots & \vdots & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} & \boldsymbol{\Sigma} \end{pmatrix} A' \right)$$

with A a $(pn_i \times p(n_i + 1))$ matrix

$$A = \begin{pmatrix} \mathbf{I}_p & \mathbf{I}_p & \mathbf{0} & \ldots & \mathbf{0} \\ \vdots & \mathbf{0} & \ddots & \mathbf{0} & \vdots \\ \vdots & \vdots & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{I}_p & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{I}_p \end{pmatrix}.$$

□

From this result is follows that the joint density of $\mathbf{Z}$ is equal to

$$f_{\boldsymbol{\Psi}}(\mathbf{z}) \quad = \quad \prod_{i=1}^{m} \frac{1}{\sqrt{(2\pi)^{pn_i}|\boldsymbol{\Sigma}_i|}} \exp\left(-\frac{1}{2} (\mathbf{z}_i - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_i^{-1} (\mathbf{z}_i - \boldsymbol{\mu}_i) \right),$$

where $\mathbf{z}_i = (\mathbf{z}'_{i1}, \ldots, \mathbf{z}'_{in_i})'$. Then, the log likelihood function is

$$\begin{aligned} l(\boldsymbol{\Psi}) \quad &= \quad -\frac{p}{2} \sum_{i=1}^{m} n_i \log 2\pi - \frac{1}{2} \sum_{i=1}^{m} \log |\boldsymbol{\Sigma}_i| \\ &- \quad \frac{1}{2} \sum_{i=1}^{m} (\mathbf{Z}_i - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_i^{-1} (\mathbf{Z}_i - \boldsymbol{\mu}_i) \end{aligned} \tag{5.23}$$

In Sahai and Ojeda (2005) it is shown that for $p = 1$ this log likelihood can be rewritten in terms of μ, σ^2 and τ^2 such that partial derivatives can be calculated. However, it

is shown that explicit solutions do not exist for the maximum likelihood estimators and hence iterative procedures are required. For the log likelihood derived in Section 5.2.2 we have the same problem and hence iterative procedures are required as well.[2] Therefore, below the EM-algorithm is suggested as an iterative procedure to obtain the maximum likelihood estimates.

5.4.1 EM-algorithm

The EM-algorithm is a broadly applicable approach to the iterative computation of maximum likelihood estimates that has become popular in the fundamental paper of Dempster, Laird and Rubin (1977). The algorithm is useful in a variety of incomplete data problems. The most evident problems have for example missing data or censored group observations. However, a random effect model as considered here can also be seen as an incomplete data problem and thus the EM-algorithm is applicable for such problems as well. Recall that the random effect model is

$$\mathbf{Z}_{ij} = \boldsymbol{\theta}_i + \boldsymbol{\varepsilon}_{ij} \quad 1 \leq i \leq m,\ 1 \leq j \leq n_i.$$

Hence, the observed background data $\mathbf{Z}$ can be viewed as incomplete, since the random group effects $\boldsymbol{\theta}_i$ are unobservable data. Then, the complete data are

$$(\mathbf{Z}_{ij}, \boldsymbol{\theta}_i) \quad 1 \leq i \leq m,\ 1 \leq j \leq n_i.$$

In case complete data are observed, maximum likelihood estimation is often easy. Hence, the idea of the EM-algorithm is to approach the problem of solving the incomplete-data likelihood by associating it to the complete data log likelihood function. The complete data likelihood function is

$$\begin{aligned} f_{\boldsymbol{\Psi}}(\mathbf{z}, \boldsymbol{\theta}) &= f_{\boldsymbol{\Psi}}(\mathbf{z} \mid \boldsymbol{\theta}) h_{\Psi}(\boldsymbol{\theta}) \\ &= \prod_{i=1}^{m} f_{\boldsymbol{\Psi}}(\mathbf{z}_{i1}, \ldots, \mathbf{z}_{in_i} \mid \boldsymbol{\theta}_i) \cdot \prod_{i=1}^{m} h_{\boldsymbol{\Psi}}(\boldsymbol{\theta}_i). \end{aligned}$$

Contrary to equation (5.22), the observations within group i are independent because we condition on the group effect $\boldsymbol{\theta}_i$, i.e.

$$f_{\boldsymbol{\Psi}}(\mathbf{z}, \boldsymbol{\theta}) = \prod_{i=1}^{m} \prod_{j=1}^{n_i} f_{\boldsymbol{\Psi}}(\mathbf{z}_{ij} \mid \boldsymbol{\theta}_i) \cdot \prod_{i=1}^{m} h_{\boldsymbol{\Psi}}(\boldsymbol{\theta}_i)$$

Then, by using the conditional distribution for $\mathbf{Z}_{ij}$ and the Gaussian between-source distribution it follows that

$$\begin{aligned} f_{\boldsymbol{\Psi}}(\mathbf{z}, \boldsymbol{\theta}) &= \prod_{i=1}^{m} \prod_{j=1}^{n_i} \left\{ (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} (\mathbf{z}_{ij} - \boldsymbol{\theta}_i)' \boldsymbol{\Sigma}^{-1} (\mathbf{z}_{ij} - \boldsymbol{\theta}_i) \right) \right\} \\ &\times \prod_{i=1}^{m} \left\{ (2\pi)^{-\frac{p}{2}} |\mathbf{T}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} (\boldsymbol{\theta}_i - \boldsymbol{\mu})' \mathbf{T}^{-1} (\boldsymbol{\theta}_i - \boldsymbol{\mu}) \right) \right\} \\ &= (2\pi)^{-\frac{p(m+N)}{2}} |\mathbf{T}|^{-\frac{m}{2}} |\boldsymbol{\Sigma}|^{-\frac{N}{2}} \\ &\times \prod_{i=1}^{m} \exp\left(-\frac{1}{2} (\boldsymbol{\theta}_i - \boldsymbol{\mu})' \mathbf{T}^{-1} (\boldsymbol{\theta}_i - \boldsymbol{\mu}) - \frac{1}{2} \sum_{j=1}^{n_i} (\mathbf{z}_{ij} - \boldsymbol{\theta}_i)' \boldsymbol{\Sigma}^{-1} (\mathbf{z}_{ij} - \boldsymbol{\theta}_i) \right). \end{aligned}$$

[2] Note that for the likelihood in Section 6.2.2 we have used the means $\overline{Z}_1, \ldots, \overline{Z}_m$ as data instead of the total background data $\mathbf{Z}$. The reason for this is that the generalized weighted mean is defined in terms of the means $\overline{Z}_i$.

The complete data log likelihood function is then

$$\begin{aligned} l_c(\Psi) &= -\frac{p}{2}(m+N)\log 2\pi - \frac{m}{2}\log|\mathbf{T}| - \frac{1}{2}N\log|\mathbf{\Sigma}| \\ &- \frac{1}{2}\sum_{i=1}^{m}\left\{(\boldsymbol{\theta}_i-\boldsymbol{\mu})'\mathbf{T}^{-1}(\boldsymbol{\theta}_i-\boldsymbol{\mu}) + \sum_{j=1}^{n_i}(\mathbf{Z}_{ij}-\boldsymbol{\theta}_i)'\mathbf{\Sigma}^{-1}(\mathbf{Z}_{ij}-\boldsymbol{\theta}_i)\right\}. \end{aligned}$$

From this complete log likelihood it follows that maximum likelihood estimation is indeed computationally more tractable if we had observed the variables $\boldsymbol{\theta}_i$ in addition to $\mathbf{Z}_{ij}$. In this case we could find closed form maximum likelihood estimators from the latter expression using the maximum likelihood estimators for multivariate Gaussian likelihood functions (Mardia et al. (1979), p.103). The complete data problem thus yields a closed form solution for the maximum likelihood estimator.

This fact is used in the EM-algorithm to approach the problem of solving the incomplete data likelihood in equation (5.23) indirectly by proceeding iteratively in terms of the complete data log likelihood (McLachlan and Krishnan (1997)). Because the complete data log likelihood is unobservable, this log likelihood is replaced by its conditional expectation given the observed data under the distribution determined by the current fit for the parameters. This step is known as the *expectation* step of the algorithm. In the *maximization* step this conditional expectation is maximized such that a new estimate is found. More specifically, for iteration $(k+1)$ (McLachlan and Krishnan (1997)):

E-step : Caclulate $Q(\mathbf{\Psi};\mathbf{\Psi}^{(k)})$ where $Q(\mathbf{\Psi};\mathbf{\Psi}^{(k)}) = \mathrm{E}_{\mathbf{\Psi}^{(k)}}\left(l_c(\mathbf{\Psi}) \mid \mathbf{Z}=\mathbf{z}\right)$ for $\mathbf{\Psi}^{(k)}$ the current estimate.

M-step : Choose $\mathbf{\Psi}^{(k+1)}$ to be any value of $\mathbf{\Psi}$ such that it maximizes $Q(\mathbf{\Psi};\mathbf{\Psi}^{(k)})$, i.e. $Q(\mathbf{\Psi}^{(k+1)};\mathbf{\Psi}^{(k)}) \geq Q(\mathbf{\Psi};\mathbf{\Psi}^{(k)}) \quad \forall\mathbf{\Psi}$.

These steps should be repeated until convergence. In practice this means that a certain stopping criterion should be chosen. A discussion about stopping criteria is given in Section 5.5. The E-step and the M-step for this problem are derived below.

E-step

For the expectation step we need to compute the conditional expectation of the complete log likelihood,

$$\begin{aligned} Q(\mathbf{\Psi};\mathbf{\Psi}^{(k)}) &= \mathrm{E}_{\mathbf{\Psi}^{(k)}}\left(l_c(\mathbf{\Psi}) \mid \mathbf{Z}=\mathbf{z}\right) \\ &= -\frac{p}{2}(m+N)\log 2\pi - \frac{m}{2}\log|\mathbf{T}| - \frac{N}{2}\log|\mathbf{\Sigma}| \\ &- \frac{1}{2}\sum_{i=1}^{m}\mathrm{E}_{\mathbf{\Psi}^{(k)}}\left((\boldsymbol{\theta}_i-\boldsymbol{\mu})'\mathbf{T}^{-1}(\boldsymbol{\theta}_i-\boldsymbol{\mu}) \mid \mathbf{Z}_i=\mathbf{z}_i\right) \\ &- \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\mathrm{E}_{\mathbf{\Psi}^{(k)}}\left((\mathbf{Z}_{ij}-\boldsymbol{\theta}_i)'\mathbf{\Sigma}^{-1}(\mathbf{Z}_{ij}-\boldsymbol{\theta}_i) \mid \mathbf{Z}_i=\mathbf{z}_i\right). \end{aligned}$$

From the latter expression it can be seen that we need the following conditional expectations:

(i) $\mathrm{E}_{\mathbf{\Psi}^{(k)}}\left(\boldsymbol{\theta}_i'\mathbf{T}^{-1}\boldsymbol{\theta}_i \mid \mathbf{Z}_i=\mathbf{z}_i\right)$

(ii) $\mathrm{E}_{\mathbf{\Psi}^{(k)}}\left(\boldsymbol{\theta}_i'\mathbf{\Sigma}^{-1}\boldsymbol{\theta}_i \mid \mathbf{Z}_i=\mathbf{z}_i\right)$

(iii) $\mathrm{E}_{\boldsymbol{\Psi}^{(k)}}\left(\boldsymbol{\theta}_i \mid \mathbf{Z}_i = \mathbf{z}_i\right)$.

To compute the expectations given in (i) and (ii) note that since each term $\boldsymbol{\theta}_i'\mathbf{T}^{-1}\boldsymbol{\theta}_i$ is a scalar it equals the trace of itself

$$\boldsymbol{\theta}_i'\mathbf{T}^{-1}\boldsymbol{\theta}_i = tr(\boldsymbol{\theta}_i'\mathbf{T}^{-1}\boldsymbol{\theta}_i).$$

Now using the cyclic property of the trace it follows that (Searle (1982), p.45)

$$\boldsymbol{\theta}_i'\mathbf{T}^{-1}\boldsymbol{\theta}_i = tr(\mathbf{T}^{-1}\boldsymbol{\theta}_i\boldsymbol{\theta}_i').$$

Since both the expectation and the trace are linear operators (Searle (1982), p.29) they commute and thus

$$\mathrm{E}_{\boldsymbol{\Psi}^{(k)}}\left(\boldsymbol{\theta}_i'\mathbf{T}^{-1}\boldsymbol{\theta}_i \mid \mathbf{Z}_i = \mathbf{z}_i\right) = tr\left(\mathbf{T}^{-1}\,\mathrm{E}_{\Psi^{(k)}}\left(\boldsymbol{\theta}_i\boldsymbol{\theta}_i' \mid \mathbf{Z}_i = \mathbf{z}_i\right)\right).$$

Therefore, instead of the expectations given in (i) and (ii) we need

(iv) $\mathrm{E}_{\boldsymbol{\Psi}^{(k)}}\left(\boldsymbol{\theta}_i\boldsymbol{\theta}_i' \mid \mathbf{Z}_i = \mathbf{z}_i\right)$

to find $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)})$. To compute the expectations given in (iii) and (iv) the conditional distribution of $\boldsymbol{\theta}_i$ given the data $\mathbf{Z}_i$ is required. To find this distribution, the following result is necessary.

Lemma 5.4.2. *The joint density of $\boldsymbol{\theta}_i$ and $\boldsymbol{Z}_i$ is multivariate normal with mean vector* $\boldsymbol{\mu}_i^1 = \begin{pmatrix} \boldsymbol{\mu} \\ \vdots \\ \boldsymbol{\mu} \end{pmatrix}$ *and covariance matrix* $\boldsymbol{\Sigma}_i^1 = \begin{pmatrix} \boldsymbol{T} & \boldsymbol{T} & \dots & \boldsymbol{T} \\ \boldsymbol{T} & \boldsymbol{T}+\boldsymbol{\Sigma} & & \boldsymbol{T} \\ \vdots & & \ddots & \\ \boldsymbol{T} & \boldsymbol{T} & & \boldsymbol{T}+\boldsymbol{\Sigma} \end{pmatrix}$.

Proof. The proof follows from the proof of lemma 5.4.1 with A a $(p(n_i+1)\times p(n_i+1))$ matrix

$$A = \begin{pmatrix} \mathbf{I}_p & \mathbf{0} & \dots & \dots & \mathbf{0} \\ \mathbf{I}_p & \mathbf{I}_p & \mathbf{0} & \dots & \mathbf{0} \\ \vdots & \mathbf{0} & \ddots & \mathbf{0} & \vdots \\ \vdots & \vdots & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{I}_p & \mathbf{0} & \dots & \mathbf{0} & \mathbf{I}_p \end{pmatrix}.$$

□

To find the desired result of the conditional distribution of $\boldsymbol{\theta}_i$ given $\mathbf{Z}_i$, partition the vector $\boldsymbol{\mu}_i^1$ and matrix $\boldsymbol{\Sigma}_i^1$ as follows:

$$\boldsymbol{\mu}_i^1 = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}$$

such that the vector $\boldsymbol{\mu}_1$ consist of a single vector $\boldsymbol{\mu}$ and the vector $\boldsymbol{\mu}_2$ consist of the remaining n_i vectors $\boldsymbol{\mu}$. The matrix $\boldsymbol{\Sigma}_i^1$ can be partitioned as

$$\boldsymbol{\Sigma}_i^1 = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$$

where $\boldsymbol{\Sigma}_{11} = \mathbf{T}$, $\boldsymbol{\Sigma}_{22} = \boldsymbol{\Sigma}_i$ as defined in lemma 5.4.1 and the matrix $\boldsymbol{\Sigma}_{12}$ consist of pn_i matrices $\mathbf{T}$, i.e. $\boldsymbol{\Sigma}_{12} = (\mathbf{T}, \dots, \mathbf{T})$. Then we have that (Rao (1973), p.522)

$$\boldsymbol{\theta}_i \mid \mathbf{Z}_i \sim \mathcal{N}_p\left(\overline{\boldsymbol{\mu}}_i, \overline{\boldsymbol{\Sigma}}_i\right)$$

with

$$\begin{aligned}\overline{\boldsymbol{\mu}}_i &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{z}_i - \boldsymbol{\mu}_2)\\ \overline{\boldsymbol{\Sigma}}_i &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}.\end{aligned}$$

Therefore the conditional expectations given in (iii) and (iv) are equal to

$$\mathrm{E}_{\boldsymbol{\Psi}^{(k)}}(\boldsymbol{\theta}_i \mid \mathbf{Z}_i = \mathbf{z}_i) = \overline{\boldsymbol{\mu}}_i^{(k)}$$

and

$$\mathrm{E}_{\boldsymbol{\Psi}^{(k)}}\left(\boldsymbol{\theta}_i\boldsymbol{\theta}_i' \mid \mathbf{Z}_i = \mathbf{z}_i\right) = \overline{\boldsymbol{\Sigma}}_i^{(k)} + \overline{\boldsymbol{\mu}}_i^{(k)}\overline{\boldsymbol{\mu}}_i^{(k)\prime}.$$

Hence, the conditional expectation of the complete log likelihood is equal to

$$\begin{aligned}Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)}) &= -\frac{p}{2}(m+N)\log 2\pi - \frac{m}{2}\log|\mathbf{T}| - \frac{N}{2}\log|\boldsymbol{\Sigma}|\\ &- \frac{1}{2}\sum_{i=1}^{m}\Big\{tr\left(\mathbf{T}^{-1}\left(\overline{\boldsymbol{\Sigma}}_i^{(k)} + \overline{\boldsymbol{\mu}}_i^{(k)}\overline{\boldsymbol{\mu}}_i^{(k)\prime}\right)\right) - \overline{\boldsymbol{\mu}}_i^{(k)\prime}\mathbf{T}^{-1}\boldsymbol{\mu} - \boldsymbol{\mu}'\mathbf{T}^{-1}\overline{\boldsymbol{\mu}}_i^{(k)}\\ &+ \boldsymbol{\mu}'\mathbf{T}^{-1}\boldsymbol{\mu}\Big\} - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\Big\{\mathbf{z}_{ij}'\boldsymbol{\Sigma}^{-1}\mathbf{z}_{ij} - \mathbf{z}_{ij}'\boldsymbol{\Sigma}^{-1}\overline{\boldsymbol{\mu}}_i^{(k)} - \overline{\boldsymbol{\mu}}_i^{(k)\prime}\boldsymbol{\Sigma}^{-1}\mathbf{z}_{ij}\\ &+ tr\left(\boldsymbol{\Sigma}^{-1}\left(\overline{\boldsymbol{\Sigma}}_i^{(k)} + \overline{\boldsymbol{\mu}}_i^{(k)}\overline{\boldsymbol{\mu}}_i^{(k)\prime}\right)\right)\Big\}. \qquad (5.24)\end{aligned}$$

This expression has the same structure as the complete log likelihood. This is beneficial for the maximization step, since we have explained that the complete data problem yields closed form solutions for the maximum likelihood estimators.

M-step

In the maximization step we need to maximize $Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)})$ given in equation (5.24) with respect to Ψ. First, $Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)})$ will be maximized with respect to $\boldsymbol{\mu}$. Hence, we need to compute

$$\frac{\partial Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)})}{\partial\boldsymbol{\mu}} = \frac{\partial}{\partial\boldsymbol{\mu}}\Big\{-\frac{1}{2}\sum_{i=1}^{m}\Big\{-\overline{\boldsymbol{\mu}}_i^{(k)\prime}\mathbf{T}^{-1}\boldsymbol{\mu} - \boldsymbol{\mu}'\mathbf{T}^{-1}\overline{\boldsymbol{\mu}}_i^{(k)} + \boldsymbol{\mu}'\mathbf{T}^{-1}\boldsymbol{\mu}\Big\}\Big\}.$$

The latter derivative is found by the derivative of scalars with respect to the vector $\boldsymbol{\mu}$. Because the inverse of the covariance matrix $\mathbf{T}$ is symmetric as well it follows that (Searle (1982), p.336)

$$\frac{\partial\overline{\boldsymbol{\mu}}_i^{(k)\prime}\mathbf{T}^{-1}\boldsymbol{\mu}}{\partial\boldsymbol{\mu}} = \frac{\partial\boldsymbol{\mu}'\mathbf{T}^{-1}\overline{\boldsymbol{\mu}}_i^{(k)}}{\partial\boldsymbol{\mu}} = \overline{\boldsymbol{\mu}}_i^{(k)}\mathbf{T}^{-1}$$

and

$$\frac{\partial\boldsymbol{\mu}'\mathbf{T}^{-1}\boldsymbol{\mu}}{\partial\boldsymbol{\mu}} = 2\boldsymbol{\mu}'\mathbf{T}^{-1}.$$

Hence, for the partial derivative of $Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)})$ with respect to $\boldsymbol{\mu}$ it follows that

$$\begin{aligned}\frac{\partial Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)})}{\partial\boldsymbol{\mu}} &= -\frac{1}{2}\sum_{i=1}^{m}\left\{-2\overline{\boldsymbol{\mu}}_i^{(k)}\mathbf{T}^{-1} + 2\boldsymbol{\mu}'\mathbf{T}^{-1}\right\}\\ &= \sum_{i=1}^{m}\left\{\overline{\boldsymbol{\mu}}_i^{(k)}\mathbf{T}^{-1}\right\} - m\boldsymbol{\mu}'\mathbf{T}^{-1}.\end{aligned}$$

Equating this derivative to zero and solving for $\boldsymbol{\mu}$ results in the required estimator

$$\boldsymbol{\mu}^{(k+1)} = \frac{1}{m}\sum_{i=1}^{m}\overline{\boldsymbol{\mu}}_i^{(k)}. \tag{5.25}$$

To maximize $Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)})$ with respect to $\mathbf{T}$, we compute

$$\begin{aligned}\frac{\partial Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)})}{\partial \mathbf{T}^{-1}} &= \frac{\partial}{\partial \mathbf{T}^{-1}}\Bigg\{-\frac{m}{2}\log|\mathbf{T}| - \frac{1}{2}\sum_{i=1}^{m}\Big\{tr\left(\mathbf{T}^{-1}\left(\overline{\boldsymbol{\Sigma}}_i^{(k)} + \overline{\boldsymbol{\mu}}_i^{(k)}\overline{\boldsymbol{\mu}}_i^{(k)\prime}\right)\right)\\ &- \overline{\boldsymbol{\mu}}_i^{(k)\prime}\mathbf{T}^{-1}\boldsymbol{\mu} - \boldsymbol{\mu}'\mathbf{T}^{-1}\overline{\boldsymbol{\mu}}_i^{(k)} + \boldsymbol{\mu}'\mathbf{T}^{-1}\boldsymbol{\mu}\Big\}\Bigg\}. \end{aligned}\tag{5.26}$$

Because the inverse of the matrix $\mathbf{T}$ is symmetric we have that (Searle (1982), p.337)

$$\frac{d\log|\mathbf{T}^{-1}|}{d\mathbf{T}^{-1}} = 2\mathbf{T} - \mathrm{diag}(\mathbf{T}).$$

In addition, the matrix $\overline{\boldsymbol{\Sigma}}_i^{(k)} + \overline{\boldsymbol{\mu}}_i^{(k)}\overline{\boldsymbol{\mu}}_i^{(k)\prime}$ is symmetric because it is the sum of a covariance matrix and an outer product, hence (Searle (1982), p.336)

$$\frac{\partial tr\left(\mathbf{T}^{-1}\left(\overline{\boldsymbol{\Sigma}}_i^{(k)} + \overline{\boldsymbol{\mu}}_i^{(k)}\overline{\boldsymbol{\mu}}_i^{(k)\prime}\right)\right)}{\partial \mathbf{T}^{-1}} = 2\left(\overline{\boldsymbol{\Sigma}}_i^{(k)} + \overline{\boldsymbol{\mu}}_i^{(k)}\overline{\boldsymbol{\mu}}_i^{(k)\prime}\right) - \mathrm{diag}\left(\overline{\boldsymbol{\Sigma}}_i^{(k)} + \overline{\boldsymbol{\mu}}_i^{(k)}\overline{\boldsymbol{\mu}}_i^{(k)\prime}\right).$$

In general it follows that (Searle (1982), p.336)

$$\begin{aligned}\frac{\partial \overline{\boldsymbol{\mu}}_i^{(k)\prime}\mathbf{T}^{-1}\boldsymbol{\mu}}{\partial \mathbf{T}^{-1}} &= \frac{\partial tr\left(\mathbf{T}^{-1}\boldsymbol{\mu}\overline{\boldsymbol{\mu}}_i^{(k)\prime}\right)}{\partial \mathbf{T}^{-1}}\\ &= \boldsymbol{\mu}\overline{\boldsymbol{\mu}}_i^{(k)\prime} + \overline{\boldsymbol{\mu}}_i^{(k)}\boldsymbol{\mu}' - \mathrm{diag}\left(\boldsymbol{\mu}\overline{\boldsymbol{\mu}}_i^{(k)\prime}\right).\end{aligned}$$

This approach can also be applied for the partial derivatives of the terms $\boldsymbol{\mu}'\mathbf{T}^{-1}\overline{\boldsymbol{\mu}}_i^{(k)}$ and $\boldsymbol{\mu}'\mathbf{T}^{-1}\boldsymbol{\mu}$. Then, the partial derivative of $Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)})$ with respect to $\mathbf{T}^{-1}$ is

$$\begin{aligned}\frac{\partial Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)})}{\partial \mathbf{T}^{-1}} &= \frac{1}{2}\left(2m\mathbf{T} - m\cdot\mathrm{diag}(\mathbf{T})\right) - \frac{1}{2}\sum_{i=1}^{m}\Big\{2\left(\overline{\boldsymbol{\Sigma}}_i^{(k)} + \overline{\boldsymbol{\mu}}_i^{(k)}\overline{\boldsymbol{\mu}}_i^{(k)\prime}\right)\\ &- \mathrm{diag}\left(\overline{\boldsymbol{\Sigma}}_i^{(k)} + \overline{\boldsymbol{\mu}}_i^{(k)}\overline{\boldsymbol{\mu}}_i^{(k)\prime}\right) - \boldsymbol{\mu}\overline{\boldsymbol{\mu}}_i^{(k)\prime} - \overline{\boldsymbol{\mu}}_i^{(k)}\boldsymbol{\mu}' + \mathrm{diag}\left(\boldsymbol{\mu}\overline{\boldsymbol{\mu}}_i^{(k)\prime}\right)\\ &- \overline{\boldsymbol{\mu}}_i^{(k)}\boldsymbol{\mu}' - \boldsymbol{\mu}\overline{\boldsymbol{\mu}}_i^{(k)\prime} + \mathrm{diag}\left(\boldsymbol{\mu}\overline{\boldsymbol{\mu}}_i^{(k)\prime}\right) + 2\boldsymbol{\mu}\boldsymbol{\mu}' - \mathrm{diag}(\boldsymbol{\mu}\boldsymbol{\mu}')\Big\}\\ &:= \frac{1}{2}\left(2V - \mathrm{diag}(V)\right),\end{aligned}$$

where

$$V = m\mathbf{T} - \sum_{i=1}^{m}\left\{\overline{\boldsymbol{\Sigma}}_i^{(k)} + \overline{\boldsymbol{\mu}}_i^{(k)}\overline{\boldsymbol{\mu}}_i^{(k)\prime}\right\} + \sum_{i=1}^{m}\left\{\boldsymbol{\mu}\overline{\boldsymbol{\mu}}_i^{(k)\prime}\right\} + \sum_{i=1}^{m}\left\{\overline{\boldsymbol{\mu}}_i^{(k)}\boldsymbol{\mu}'\right\} - m\boldsymbol{\mu}\boldsymbol{\mu}'. \tag{5.27}$$

Equating the latter derivative to zero implies that

$$V = \frac{1}{2}\mathrm{diag}(V),$$

however this is only true for $V = \mathbf{0}$. Thus if we solve equation (5.27) for $\mathbf{T}$, the required estimator is found:

$$\mathbf{T}^{(k+1)} = \frac{1}{m}\sum_{i=1}^{m}\left\{\overline{\boldsymbol{\Sigma}}_i^{(k)} + \overline{\boldsymbol{\mu}}_i^{(k)}\overline{\boldsymbol{\mu}}_i^{(k)\prime} - \boldsymbol{\mu}\overline{\boldsymbol{\mu}}_i^{(k)\prime} - \overline{\boldsymbol{\mu}}_i^{(k)}\boldsymbol{\mu}' + \boldsymbol{\mu}\boldsymbol{\mu}'\right\}. \tag{5.28}$$

Note that the optimal value for $\mathbf{T}$ depends on $\boldsymbol{\mu}$. Hence, it is a natural choice to substitute the optimal value for $\boldsymbol{\mu}$ in the estimator for $\mathbf{T}$. Since the parameters are optimized simultaneously, this means that in iteration $k+1$ the mean $\boldsymbol{\mu}$ will be replaced by the optimal value for $\boldsymbol{\mu}$, i.e. $\boldsymbol{\mu}^{(k+1)}$.

To maximize $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)})$ with respect to $\boldsymbol{\Sigma}$, we have to compute

$$\begin{aligned}\frac{\partial Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)})}{\partial \boldsymbol{\Sigma}^{-1}} &= \frac{\partial}{\partial \boldsymbol{\Sigma}}\Bigg\{ -\frac{N}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\Big\{\mathbf{z}_{ij}'\boldsymbol{\Sigma}^{-1}\mathbf{z}_{ij} - \mathbf{z}_{ij}'\boldsymbol{\Sigma}^{-1}\overline{\boldsymbol{\mu}}_i^{(k)} \\ &- \overline{\boldsymbol{\mu}}_i^{(k)\prime}\boldsymbol{\Sigma}^{-1}\mathbf{z}_{ij} + tr\left(\boldsymbol{\Sigma}^{-1}\left(\overline{\boldsymbol{\Sigma}}_i^{(k)} + \overline{\boldsymbol{\mu}}_i^{(k)}\overline{\boldsymbol{\mu}}_i^{(k)\prime}\right)\right)\Big\}\Bigg\}.\end{aligned}$$

This derivative is of the same form as equation (5.26) and thus the same steps can be applied. Hence,

$$\boldsymbol{\Sigma}^{(k+1)} = \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left\{\overline{\boldsymbol{\Sigma}}_i^{(k)} + \overline{\boldsymbol{\mu}}_i^{(k)}\overline{\boldsymbol{\mu}}_i^{(k)\prime} - \mathbf{z}_{ij}\overline{\boldsymbol{\mu}}_i^{(k)\prime} - \overline{\boldsymbol{\mu}}_i^{(k)}\mathbf{z}_{ij}' + \mathbf{z}_{ij}\mathbf{z}_{ij}'\right\}. \tag{5.29}$$

The new estimate $\boldsymbol{\Psi}^{(k+1)}$ is given by equation (5.25), (5.28) and (5.29).

The described EM-algorithm has several appealing properties relative to other iterative algorithms (McLachlan and Krishnan (1997)). For instance, it is numerically stable with each iteration increasing the likelihood.Further, under general conditions convergence is nearly always to a local maximizer from arbitrary initial values. Additionally, the algorithm is often easily implemented with low cost per iteration. On the other hand, if there is more than one local maximum it does not guarantee convergence to a global maximum and then the estimate depends on the initial values. Further, it may converge slowly in certain situations. Therefore modified versions of the EM-algorithm have been developed that can be used as well.

5.5 Comparison of methods of estimation

To compare the described methods in this chapter a Monte Carlo simulation has been performed. In Section 5.5.1 the simulation setup is briefly discussed. After that, the mean estimators described in Section 5.2 are compared. The remainder of the section covers the comparison of the analysis of variance estimators (Section 5.3) and the EM-algorithm for maximum likelihood estimation (Section 5.4).

5.5.1 Monte Carlo simulation

The number of repeated simulations is fixed at $M = 1000$. This means that M times a background data set is simulated according to the model described in Section 5.1 and subsequently the parameters are estimated according to the methods described in this chapter.

The simulation is performed for two situations which we will refer to as *balanced design* and *unbalanced design*. The values for the quantities m and $n_i, i \in \{1, \ldots, m\}$

are based on xtc data collected by the Netherlands Forensic Institute and are given in the table below.

	m	n	N
balanced design	10	20	200
unbalanced design	10	(39,17,28,13,10,3,31,7,6,46)	200

Table 5.1: The quantities m and $n_i, i \in \{1, \ldots, m\}$ that are chosen for the simulation of the balanced- and unbalanced design.

To simulate the background data, the parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\mathbf{T}$ had been fixed based on real xtc data, see e.g. Bolck et al. (2009) and Bolck and Alberink (2011). By fixing these parameters, the random vectors $\boldsymbol{\varepsilon}_{ij}$ and $\boldsymbol{\theta}_i$ were drawn from multivariate normal distributions as given in Section 5.1 for each $i \in \{1, \ldots, m\}$ and $j \in \{1, \ldots, n_i\}$. Hence the background data $(\mathbf{Z}_{ij}, 1 \leq i \leq m, 1 \leq j \leq n_i)$ was generated by taking the sum of these vectors, see equation (5.1).

To assess the quality of the estimations we consider the estimated values within the mean vector and covariance matrices separately. To do this, the following notation is used

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_p^2 \end{pmatrix}, \qquad \mathbf{T} = \begin{pmatrix} \tau_1^2 & & & \\ \tau_{12} & \tau_2^2 & & \\ \vdots & & \ddots & \\ \tau_{1p} & \cdots & \tau_{(p-1)p} & \tau_p^2 \end{pmatrix}.$$

In this simulation $\boldsymbol{\Sigma}$ is modelled as a diagonal matrix, hence only the diagonal is subtracted from the estimation $\widehat{\boldsymbol{\Sigma}}$ and assessed on the performance. To assess the estimate of $\mathbf{T}$ we only consider the lower triangular of the matrix, because it is symmetric. To give an overview of the estimated parameters in the Monte Carlo simulation box plots are used (Rice (2007)). To assess the performance of the estimates the mean squared error is chosen (Wasserman (2004)). For example, for the estimates of the mean the mean squared error is equal to:

$$\mathrm{E}_{\mu_k} \left(\widehat{\mu}_k - \mu_k\right)^2 \qquad \text{for } k = 1, \ldots, p.$$

This means that for each simulation i, with $i \in \{1, \ldots, M\}$, the squared difference between the estimate and the true value is computed. After M simulations the average over these squared differences is taken as the mean squared error.

One should bear in mind that the number of parameters to compare is

$$\begin{aligned} p + p + \sum_{k=1}^{p} k &= 2p + \frac{1}{2}p(p+1) \\ &= \frac{1}{2}p(p+5). \end{aligned}$$

Hence, we start off with a simulation study for $p = 4$ since this gives a reasonable amount of estimated parameters to compare. Some of the parameters that are used are given below.

μ_1	μ_2	μ_3	μ_4	σ_1^2	σ_2^2	σ_3^2	σ_4^2
0.812	6.920	1.800	2.505	0.007	0.033	0.003	0.028
		τ_1^2	τ_2^2	τ_3^2	τ_4^2		
		0.638	7.936	4.151	0.025		

Table 5.2: Some parameters used in the simulation for $p = 4$ that are based on real xtc data.

5.5.2 Comparing the mean estimators

In Section 5.2 three plug-in estimators for the mean are described. In this section the suggested estimators are compared based on the unbalanced design as specified above. For a balanced design the estimators for the mean are the same.

In Section 5.2.1 it is argued that the choice between the weighted- and the unweighted mean depends on the proportion between the unknown variances, see equation (5.13). Although the variances are unknown, the constant c can be computed since it depends on the quantities m and $n_i, i \in \{1, \ldots, m\}$, see equation (5.12). For the unbalanced design it then follows that $c = 10.294$. This constant can also be found if we plot the efficiency of $\hat{\mu}_u$ relative to $\hat{\mu}_w$ against the quotient $r = \sigma^2/\tau^2$ as given in equation (5.11). Recall from equation (5.13) that the efficiency of $\hat{\mu}_u$ relative to $\hat{\mu}_w$ is greater than one if $r < c$. Hence, in Figure 5.2 we can find the value of c. This is illustrated in the figure below.

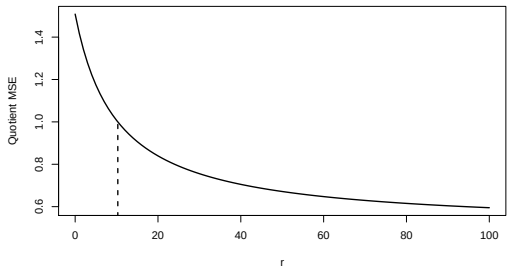


Figure 5.2: Plot of the efficiency of $\hat{\mu}_u$ relative to $\hat{\mu}_w$ (Quotient MSE) against the quotient $r = \sigma^2/\tau^2$ for the unbalanced design. The dashed line shows that the efficiency is 1 for $c = 10.294$.

If $r < c$ then the efficiency quickly increases to a maximum of 1.5, i.e. the variance of the weighted mean is at most 1.5 times bigger than the variance of the unweighted mean. If $r > c$ the quotient of the mean squared errors slowly decreases to zero.

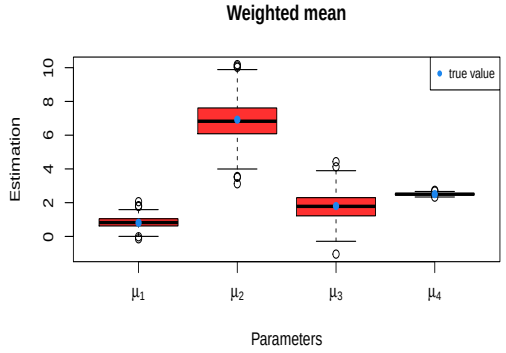
In this simulation study the values for $\mathbf{\Sigma}$ and $\mathbf{T}$ are fixed and hence it is known that $c\tau_k^2 > \sigma_k^2$ for $k \in \{1, 2, 3, 4\}$, see Table 5.2. Thus, the unweighted mean should be preferred over the weighted mean. In practice such a choice can be based on experience. In the table below the value of c is given for two xtc data sets of the NFI which are used in casework.

	m	n	N	c
Data set 1	38	$4 \leq n_i \leq 20$	458	11.356
Data set 2	1372	$1 \leq n_i \leq 20$	1792	11.135

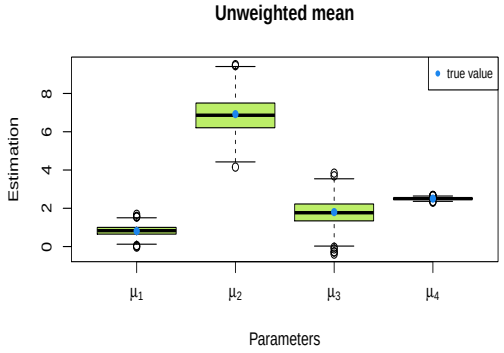
Table 5.3: Two values of c for two background xtc data sets of the NFI.

In forensic xtc comparison it is a valid assumption that the covariance matrix $\mathbf{\Sigma}$ consist of small values. One of the reasons for this small error variances is that the measurement device is very accurate. Since the constant c is approximately 11 for the background xtc data sets, by equation (5.13) it is therefore convincing to use the unweighted mean instead of the weighted mean for xtc comparison problems. However, the decision might be more difficult in other forensic applications if the value for c is smaller, for instance if $c = 1.5$ it is harder to examine whether $c\tau^2 > \sigma^2$ than if $c = 11$.

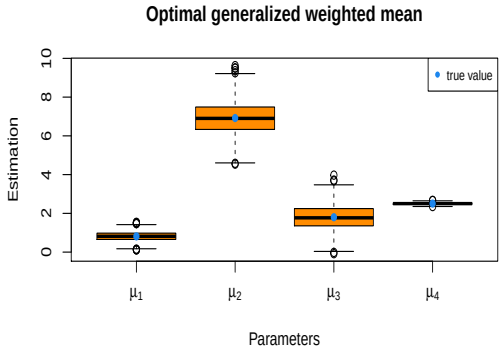
In Section 5.2.2 we have seen that if the covariance matrices $\mathbf{\Sigma}$ and $\mathbf{T}$ are known, the generalized weighted mean $\hat{\mu}_{\text{opt}}$ is the minimum variance estimator. Since in this simulation the values for $\mathbf{\Sigma}$ and $\mathbf{T}$ are known, these can be substituted into the generalized weighted mean. It is therefore interesting to examine the difference between these estimates and the weighted- and unweighted mean that can be used in practice more easily. In Figure 5.3 the results are given.



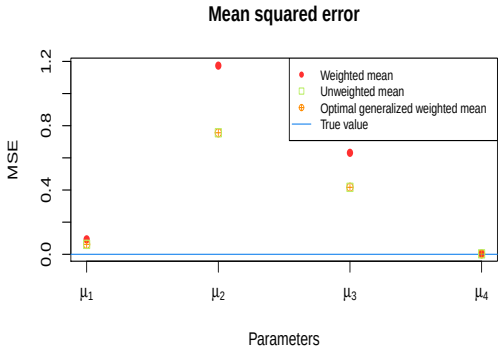
(a) Boxplot for the weighted mean estimator.



(b) Boxplot for the unweighted mean estimator.



(c) Boxplot for the optimal generalized weighted mean.



(d) Comparison of the mean squared errors.

Figure 5.3: Boxplots and means squared errors of the estimated mean vector $\boldsymbol{\mu}$ using the weighted mean, the unweighted mean and the optimal generalized weighted mean in the unbalanced design.

From the figures above it can be seen that the weighted mean has indeed a larger mean squared error (variance) than the unweighted mean. From the true parameter values given in Table 5.2 it can be computed that $\boldsymbol{\sigma}_2^2/\boldsymbol{\tau}_2^2 = 0.004$. In Figure 5.2 it can then be seen that the variance of the weighted mean is 1.5 times bigger than the variance of the unweighted mean. This is confirmed by the mean squared errors for μ_2 in Figure

5.3(d). For μ_4 the weighted variance is 1.4 times bigger than the unweighted variance ($\boldsymbol{\sigma}_4^2/\boldsymbol{\tau}_4^2 = 1.12$), obviously this has less effect since the mean squared errors itself are smaller. The fact that some mean squared errors are larger than others (see μ_2, μ_3 in comparison to μ_1, μ_4 in Figure 5.3(d)) can be confirmed by the theoretical values of the variances of the weighed- and unweighted mean in equation (5.6), which will increase when the parameter values are larger.

It is interesting to note that the optimal generalized weighted mean has more or less the same mean squared error as the unweighted average in this situation. This can be explained by the small values for the parameters in $\boldsymbol{\Sigma}$, see Table 5.2. To illustrate this, if the value for σ^2 is small ($p = 1$) it follows that $\tau^2 + \sigma^2/n_i \approx \tau^2$. Hence,

$$\text{Var}\left(\hat{\mu}_u\right) \approx \frac{1}{m^2}\sum_{i=1}^{m}\tau^2 = \frac{\tau^2}{m}.$$

The variance of the optimal generalized weighted mean is approximately,

$$\text{Var}\left(\hat{\mu}_{\text{opt}}\right) \approx \frac{1}{\left(\sum_{i=1}^{m}\frac{1}{\tau^2}\right)^2}\sum_{i=1}^{m}\left(\frac{1}{\tau^2}\right)^2\tau^2 = \frac{\tau^2}{m}.$$

Thus for small values of σ^2, we have

$$\text{Var}\left(\hat{\mu}_u\right) \approx \text{Var}\left(\hat{\mu}_{\text{opt}}\right)$$

and hence for such situation the unweighted mean is as good as the minimum variance estimator.

In Section 5.4 we have seen that the maximum likelihood estimate for $\boldsymbol{\mu}$ can be found using the EM-algorithm. It can be expected that for this situation this estimator behaves more or less the same as the unweighted mean, since the unweighted mean approaches the optimal generalized weighted mean (maximum likelihood estimator). In the next section we will see that this is indeed true.

5.5.3 Comparing ANOVA estimators with the ML estimators

This section will compare the ANOVA estimators with the maximum likelihood estimators for the balanced- and unbalanced design for $p = 4$.

Starting values and stopping criterion

To use the EM-algorithm starting values and a stopping criterion must be specified. In this problem, two natural choices for the starting values are:

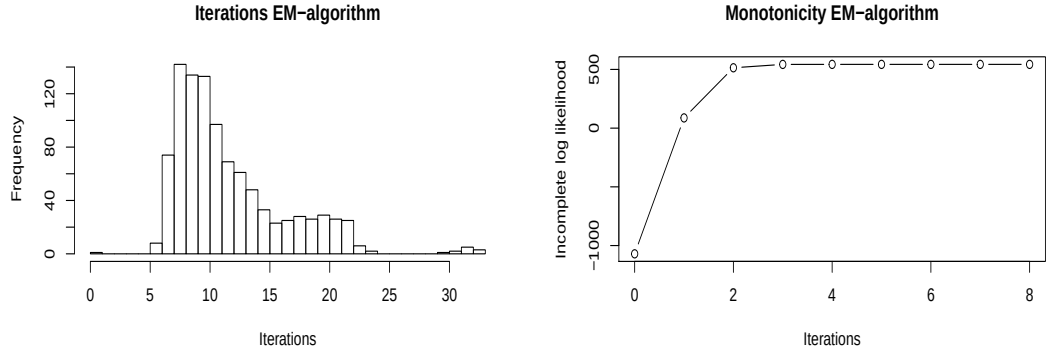
1. A vector with ones for $\boldsymbol{\mu}^{(0)}$ and the identity matrices for $\boldsymbol{\Sigma}^{(0)}$ and $\mathbf{T}^{(0)}$.
2. The analysis of variance estimators for $\boldsymbol{\Sigma}^{(0)}$ and $\mathbf{T}^{(0)}$ with the corresponding mean estimator for $\boldsymbol{\mu}^{(0)}$.

A stopping criterion for the EM-algorithm is usually in terms of either the magnitude of the relative change in the parameter estimates or the (incomplete) log likelihood (McLachlan and Kirshnan (1997)):

1. $\left|\boldsymbol{\Psi}^{k+1} - \boldsymbol{\Psi}^{k}\right| < 10^{-\delta}$
2. $l(\boldsymbol{\Psi}^{(k+1)}) - l(\boldsymbol{\Psi}^{(k)}) < 10^{-\delta}$,

where l is the incomplete log likelihood given in equation (5.23) which increases in each step of the EM-algorithm. Since the EM-algorithm can suffer from very slow convergence, it is important to emphasize that both criteria measure lack of progress and not actual convergence. Thus, selecting a suitable stopping criterion might be difficult. In any event, one should try to balance the number of iterations and the lack of progress.

In the literature such trade-offs are often obtained with δ varying between 2 and 8. Based on the simulation of the balanced design it was found that for identity starting values and a reasonable amount of iterations the relative changes between all estimated parameters was less than 10^{-5}. Figure 5.4(a) shows a histogram of the number of iterations that were needed for the Monte Carlo simulation using this stopping criterion. In Figure 5.4(b) the monotonicity of the log likelihood for one of the repeated simulations is shown. For this simulation it was found that if the stopping criterion based on the likelihood increment would be used ($\delta = 5$), the number of iterations would be the same as for the stopping criterion based on the changes in parameter estimates ($\delta = 5$), that was used to produce Figure 5.4(b). The results for the unbalanced design were more or less the same. Therefore, for this simulation it is chosen to use the relative changes of the parameters for $\delta = 5$ as the stopping criterion.



(a) Number of iterations needed for 1000 Monte Carlo simulations.

(b) Incomplete log likelihood for each iteration for one of the repeated simulations.

Figure 5.4: EM-algorithm for the balanced design using identity matrices as initial values and $\left|\boldsymbol{\Psi}^{k+1} - \boldsymbol{\Psi}^{k}\right| < 10^{-5}$ as stopping criterium.

When the analysis of variance estimators are used as starting values, the stopping criterion is satisfied for less iterations and the same estimates are found. For instance, for the simulation illustrated in Figure 5.4(b) it was then found that only 5 iterations were needed instead of 8. Since only a few more iterations are needed to satisfy the same stopping criterion, for this simulation it is chosen to use the identity matrices as starting values.

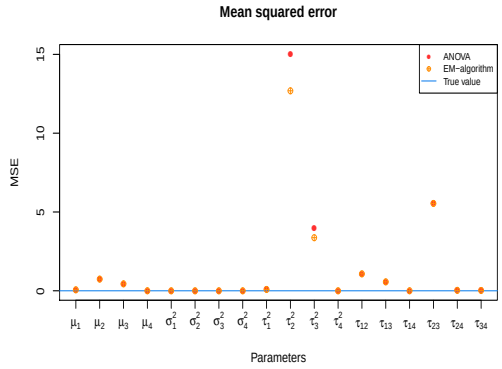
Balanced design

The results of the simulation of the balanced design are given in Figure 5.5. For balanced data the weighted- and unweighted mean are the same and hence it is irrelevant which mean estimator is used in the ANOVA estimator for $\widehat{\mathbf{T}}$. Moreover, these means are equal to the optimal generalized weighted mean. Hence, both the weighted- and unweighted mean are equal to the maximum likelihood estimator. Therefore the results

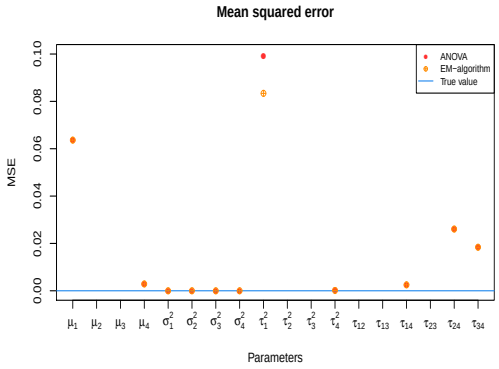
for the mean estimates in Figure 5.5 are exactly the same.

The estimates of the parameters in $\mathbf{\Sigma}$ have very small mean squared errors for both the ANOVA estimates and for the maximum likelihood estimates. In fact, the mean squared errors of the maximum likelihood estimates are smaller than those of the ANOVA estimates. Some differences are negligible or not visible in Figure 5.5(a), therefore Figure 5.5(b) shows the mean squared errors on a smaller scale.

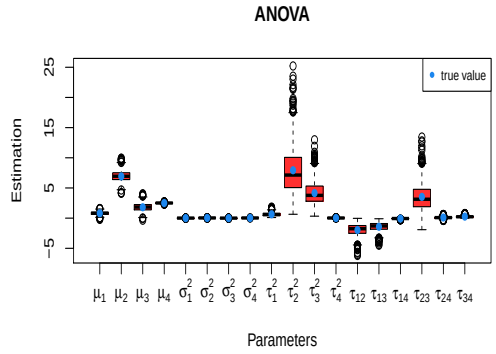
The quality of the estimates of the parameters in $\mathbf{T}$ is very different. It can be noticed that the parameters that have highest mean squared errors depend on the mean estimates with highest mean squared errors $(\tau^2, \tau^3, \tau_{23}, \tau_{12})$. Overall, in this simulation the mean squared errors of the maximum likelihood estimates are lower or equal to the mean squared errors of the ANOVA estimates.



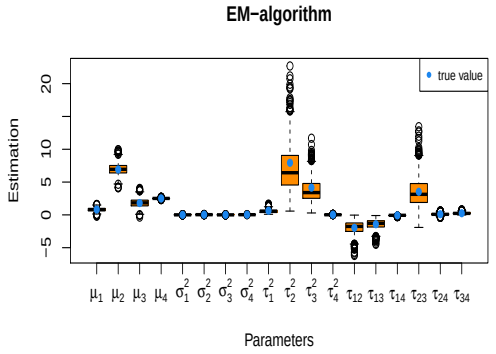
(a) Mean squared errors of the estimates.



(b) Mean squared errors of the estimates on a smaller scale.



(c) Boxplot of the ANOVA estimates.

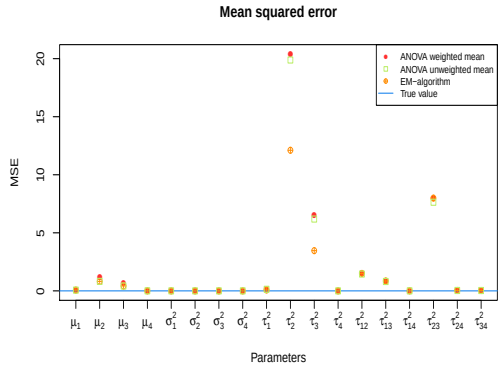


(d) Boxplot of the maximum likelihood estimates.

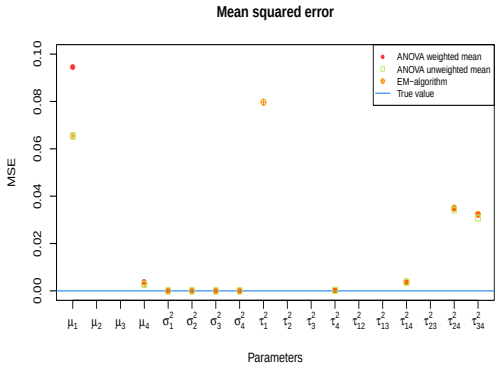
Figure 5.5: Results of the Monte Carlo simulation of the balanced design.

Unbalanced design

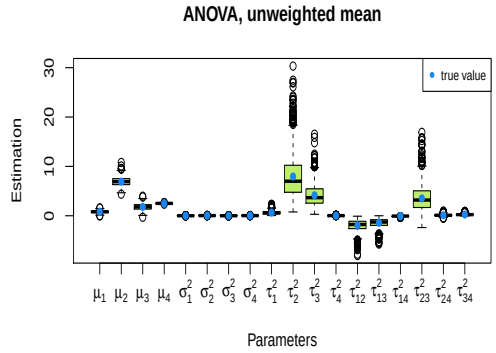
The results of the simulation of the unbalanced design are given in Figure 5.6. In Section 5.5.2 we have seen that for this kind of data, the unweighted mean performs better than the weighted mean. Moreover, we have showed that the unweighted mean approaches the optimal generalized weighted mean. Hence, we expected that the unweighted mean behaves more or less the same as the maximum likelihood estimator. In Figure 5.6(a) and Figure 5.6(b) we see that this is indeed the case.



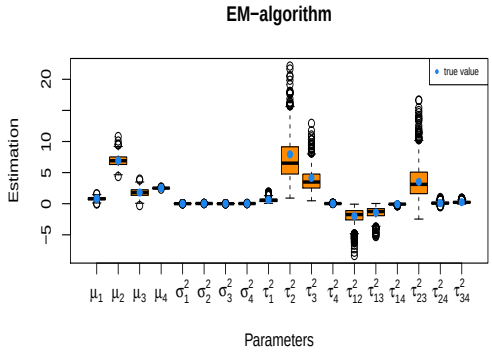
(a) Mean squared errors of the estimates.



(b) Mean squared errors of the estimates on a smaller scale.



(c) Boxplot of the ANOVA estimates.



(d) Boxplot of the maximum likelihood estimates.

Figure 5.6: Results of the Monte Carlo simulation of the unbalanced design.

Since the unweighted mean is preferred over the weighted mean in this situation, it is interesting to compare the ANOVA estimates using the unweighted mean (substitute $w_i = 1$ in equation (5.21)) with the estimates obtained with the EM-algorithm. However, for convenience the results of the ANOVA estimator with the weighted mean are given in Figure 5.6(a) and Figure 5.6(b) as well.

In comparison to the balanced design it can be noticed that the mean squared errors are higher for the unbalanced design. Moreover, the difference between the mean squared errors for the ANOVA estimates and the maximum likelihood estimates is increased (see τ_2^2, τ_3^2). Overall it can again be concluded that the maximum likelihood estimators perform the same or better than the ANOVA estimators.

6

Likelihood ratios in non-Gaussian two-level models

In Chapter 4 we have focused on likelihood ratios for continuous evidence that is modelled using a Gaussian two-level model. This means that the control- and recovered data are modelled such that they vary around their random group means $\boldsymbol{\theta}_l$, $l \in \{1, 2\}$, which are drawn from a (multivariate) normal distribution. In Section 4.1.1 we have mentioned that a multivariate normal distribution for $\boldsymbol{\theta}_l$ will often be inappropriate, because the mean of each feature must be normally distributed. If the between-source distribution is indeed non-normal, other options must be considered. These options can include assuming another parametric distribution for the between-source distribution or using a nonparametric method to estimate the between-source distribution.

In forensic statistics it is common practice to use the nonparametric method called kernel density estimation to estimate the between-source density in such cases. In forensic literature this method is referred to as a kernel distribution for the distribution of between-source variability, see for example Aitken an Lucy (2004). In this thesis we refer to two-level models using such a kernel density estimate as "non-Gaussian two-level models". The purpose of this chapter is to describe the theory behind the kernel density estimator, to discuss some difficulties and finally to derive the likelihood ratio given in Section 3.2 for non-Gaussian two-level models. In Section 6.1 kernel density estimation is explained. In Section 6.2 some attention is paid to the difficulty of choosing a smoothing parameter in the kernel density estimator for multivariate problems. In Section 6.3 the likelihood ratio is given and the difficulty due to the curse of dimensionality is briefly discussed.

6.1 Kernel density estimation

In Chapter 4 and Chapter 5 we used the parametric approach to estimate h assuming that h belongs to the normal family. If the normal family cannot be assumed, a kernel density estimator can be used which assumes no pre-specified functional form for h. To use this estimator, a sample of the random vectors $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m$ is required. In Section 4.1.2 we have seen that the mean vectors $\overline{\mathbf{Z}}_1, \ldots, \overline{\mathbf{Z}}_m$ are used as realizations for $\boldsymbol{\theta}_l$ to evaluate the assumption of normality for the true means $\boldsymbol{\theta}_l$, $l \in \{1, 2\}$. In fact, $\overline{\mathbf{Z}}_i$ is the sum of the two independent random vectors $\boldsymbol{\theta}_i$ and $\overline{\boldsymbol{\varepsilon}}_i$, see equation (5.2). Hence, the distribution of $\overline{\mathbf{Z}}_i$ is the convolution of their individual distributions, i.e. the between-source density h and the normal distribution with zero mean and covariance matrix $n_i^{-1}\boldsymbol{\Sigma}$. The sample $\overline{\mathbf{Z}}_1, \ldots, \overline{\mathbf{Z}}_m$ thus serves as an approximation of an i.i.d. sample $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m$ from the between-source distribution H. In practice, the use of this

approximating sample is a natural and common choice. Note that for larger values of n_i the approximation improves. Hence, this section deals with the following question:

> How do we estimate h using a kernel density estimator based on the sample $\overline{\mathbf{Z}}_1, \ldots, \overline{\mathbf{Z}}_m$?

In forensic comparison problems the number of features is often greater than one, hence we are particularly interested in the multivariate kernel density estimator. In Section 6.1.1 the univariate kernel density estimator is described, because this estimator is easier to visualize. The univariate estimator can be generalized to a multivariate problem quite easily, which is described in Section 6.1.2

6.1.1 Kernel density estimator

Let the number of features equal to one, i.e. $p = 1$, such that the observations $\overline{Z}_1, \ldots, \overline{Z}_m$ are random variables instead of vectors. The oldest non-parametric estimator for the density h is the well-known histogram. This estimator is simple, but results in a stair function with possible discontinuities. A smoother density estimate is preferred, since this will in general approximate the distribution of the underlying variable more accurately. Hence as an alternative to histograms, kernel density estimators can be used. These estimators result in smoother estimates which converge faster to the true density than histograms (Wasserman (2006)).

Empirical distributions assign mass of size $\frac{1}{m}$ to each $\overline{Z}_i$. The idea of kernel density estimators is to (smoothly) spread this mass of size $\frac{1}{m}$ over the neighbourhood of the associated data point. This mass is spread according to a kernel function K, i.e. a non negative function that integrates to one (Silverman (1986)). Examples of such kernel functions are for example,

$$\begin{aligned} \text{Uniform kernel} \quad &: \quad K(u) = \frac{1}{2}\mathbb{1}_{\{|u|\leq 1\}} \\ \text{Normal kernel} \quad &: \quad K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right). \end{aligned}$$

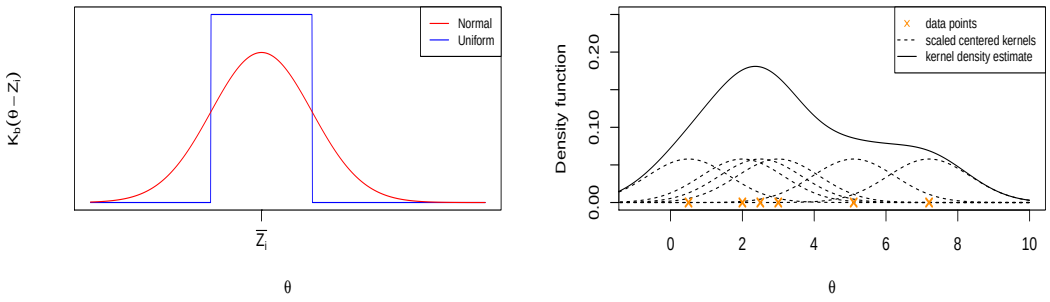
To spread the mass over the neighbourhood of the data point, the kernel function is scaled with bandwidth parameter b and centered around the data point:

$$K_b\left(\theta - \overline{Z}_i\right) = \frac{1}{b} K\left(\frac{\theta - \overline{Z}_i}{b}\right).$$

Hence the mass $\frac{1}{m}$ of each data point $\overline{Z}_i$ is spread according to the function $K_b\left(\theta - \overline{Z}_i\right)$. If a uniform kernel is used, the mass is spread over the finite support $\theta \in [\overline{Z}_i \pm 1]$ and is equal for all points in the neighbourhood of the data point. If a normal kernel is used, the mass is spread over all points θ and the mass is bigger for points closer to the data point, i.e. the mass depends on the distance to the data point. This is illustrated in Figure 6.1(a). If the scaled and centered kernel function is used to spread the mass on each data point, the kernel density estimator for each point θ is the sum over the mass contributed by the data points, i.e. the sum over the mass spread by all centered and scaled kernel functions:

$$\hat{h}(\theta) = \sum_{i=1}^{m} \frac{1}{m} K_b\left(\theta - \overline{Z}_i\right) = \frac{1}{mb} \sum_{i=1}^{m} K\left(\frac{\theta - \overline{Z}_i}{b}\right). \tag{6.1}$$

The kernel density estimate for a normal kernel is illustrated in Figure 6.1(b).

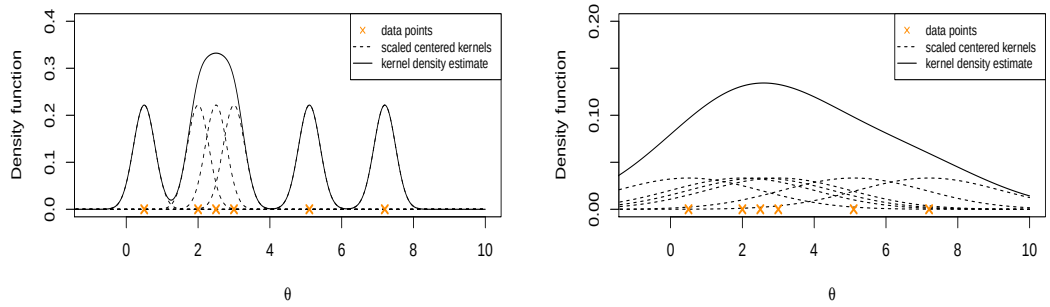


(a) A Normal- and Uniform scaled and centered kernel that spread the mass around a data point.

(b) Example of a kernel density estimate for 6 data points using the optimal bandwidth $b = 1.15$ (see Section 6.2).

Figure 6.1: Examples of scaled and centered kernels and the kernel density estimate.

The bandwidth parameter b scales the kernel K, but it also functions as a smoothing parameter in equation (6.1). If the bandwidth is larger, the mass is spread around the data points more extensively. To see this, consider the Gaussian kernel. The centered and scaled kernel is the normal density with mean $\overline{Z}_i$ and standard deviation b. Hence, the larger b the larger the standard deviation and thus the more mass is given to points further from the center. If points θ further from the data point $\overline{Z}_i$ have more mass, the density estimator $\hat{h}(\theta)$ will be smoother than when the scaled and centered kernels are more peaked. This is illustrated in the figure below.



(a) Example of a kernel density estimate for 6 data points using $b = 0.3$, which results in a less smooth estimate than Figure 6.1(b).

(b) Example of a kernel density estimate for 6 data points using $b = 2$, which results in a smoother estimate than Figure 6.1(b).

Figure 6.2: Examples of over- and under smoothed kernel density estimates.

The choice of the kernel K is not crucial for the estimate, but the choice of b is very important (Wasserman (2006)). The choice of an optimal bandwidth b is discussed in Section 6.2.2.

6.1.2 Multivariate problem

In this section we suppose that the data are p-dimensional so that $\overline{\mathbf{Z}}_i = (\overline{Z}_{i1}, \ldots, \overline{Z}_{ip})$ is a random vector, as given in equation (5.2). The kernel density estimator for the univariate problem can be generalized to p dimensions quite easily. In comparison to the univariate kernel density estimator there are two important modifications. The first modification is that the kernel function should be taken to be a p-variate kernel function. The kernel function can for example be generated from univariate kernels by using a product kernel (Wand and Jones (1995))

$$K(\mathbf{u}) = \prod_{i=1}^{p} K_0(u_i),$$

where K_0 is a univariate kernel and $\mathbf{u} = (u_1, \ldots, u_p)$. If the univariate normal kernel is taken, the product kernel results in the standard p-variate normal density:

$$K(\mathbf{u}) = (2\pi)^{-p/2} \exp\left(-\frac{1}{2}\mathbf{u}'\mathbf{u}\right), \quad \mathbf{u} \in \mathbb{R}^p.$$

The second modification in the multivariate problem concerns the bandwidth parameter b that turns into a symmetric and positive definite $p \times p$ bandwidth matrix $\mathbf{B}$. Now the kernel density estimator can be smoothed in each of the p directions. There are different possible choices for B, which are discussed in section 6.2.1. Using these two modifications, the multivariate density estimator results in

$$\hat{h}(\boldsymbol{\theta}) = m^{-1}|\mathbf{B}|^{-\frac{1}{2}} \sum_{i=1}^{m} K\left(\mathbf{B}^{-\frac{1}{2}}\left(\boldsymbol{\theta} - \overline{\mathbf{Z}}_i\right)\right). \tag{6.2}$$

As we have mentioned in the univariate problem, the choice of the kernel is not as important as the choice for the bandwidth matrix. For more details about the choice of a (multivariate) kernel we refer to for example Wand and Jones (1995). In forensic statistics it is common to use a normal kernel. In the next section we explain more about the choice for the bandwidth matrix $\mathbf{B}$ and the optimal bandwidth selection.

6.2 Smoothing parametrisation

This section will focus on the bandwidth (or smoothing) parametrisation for the kernel density estimator. In the literature some standard suggestions for the bandwidth matrix $\mathbf{B}$ exist. Section 6.2.1 will give some motivation for each of these suggestions, but the objective of Section 6.2.1 is to discuss the bandwidth matrix which is used in forensic applications. Hence, first some theoretical motivation for this bandwidth matrix is given. The section will end with a disadvantage of this particular choice.

In the Section 6.1 it is mentioned that the performance of the kernel density estimator depends on the choice of the value for the smoothing parameter b. The search for an optimal bandwidth is therefore important. For the univariate kernel density estimator this problem is well understood. There exist a number of methods that combine theoretical properties with practical performance. In the multivariate problem, choosing an optimal bandwidth matrix $\mathbf{B}$ can be more challenging. However, the bandwidth matrix which is used in forensic applications is found by the extension of a univariate method.

The objective of Section 6.2.2 is to give a brief overview of optimal bandwidth selection. Detailed theoretical motivation and calculations are not given in this thesis, but are discussed in the literature extensively.

6.2.1 Choice for the bandwidth matrix

Consider the multivariate kernel density estimator given in equation (6.2). The bandwidth matrix $\mathbf{B}$ is defined as a $p \times p$ symmetric and positive definite matrix:

$$\mathbf{B} = \begin{pmatrix} b_1^2 & & & \\ b_{12} & b_2^2 & & \\ \vdots & & \ddots & \\ b_{1p} & \dots & b_{(p-1)p} & b_p^2 \end{pmatrix}. \tag{6.3}$$

Hence, the matrix $\mathbf{B}$ has $\frac{1}{2}p(p+1)$ entries that needs to be determined. This results in a numerous amount of bandwidth parameters even for moderate dimensions. Therefore, often simplifications for the bandwidth matrix are used. Two familiar simplifications are (Wand and Jones (1995)):

1. $\mathbf{B} = b^2\mathbf{I}_{p\times p}$, with $\mathbf{I}_{p\times p}$ the $p \times p$ identity matrix.
2. $\mathbf{B} = \text{diag}(b_1^2, \dots, b_p^2)$.

The advantage of the first suggestion is that only one parameter b has to be determined. However, as a consequence the amount of smoothing is the same in each of the p directions. In comparison to the first option, the second option has p parameters that needs to be determined and hence smoothing is possible in each of the p directions. If the full bandwidth matrix $\mathbf{B}$, given in equation (6.3), is used than smoothing in each possible direction is feasible. The difference between the full bandwidth matrix and the simplified bandwidth matrices as given above can be illustrated with the well-known two-dimensional example as given in the figure below.

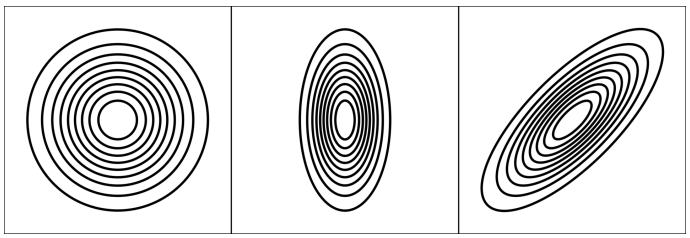


Figure 6.3: Contour plots of two-dimensional kernels for different parametrisation. The left panel: the bandwidth matrix is $\mathbf{B} = b^2\mathbf{I}_{p\times p}$, with $\mathbf{I}_{p\times p}$, hence the amount of smoothing is the same in each direction. Panel in the centre: the bandwidth matrix is $\mathbf{B} = \text{diag}(b_1^2, \dots, b_p^2)$, hence the smoothing is possible in both horizontal and vertical direction. The right panel: full bandwidth matrix $\mathbf{B}$ as defined in equation (6.3), such that smoothing is possible in each direction. Figure by Multivariate Kernel Estimation (2010).

As illustrated in the figure above, the best choice for the bandwidth matrix completely depends on the data. As a simple example, one can think of the (elliptical) contours of bivariate normal densities (Rice (2007)). More specifically, the contour plot of bivariate normal variables which are uncorrelated and have equal variance exactly correspond to the left figure of Figure 6.3 and hence the matrix $\mathbf{B} = b^2\mathbf{I}_{p\times p}$ would suffice. The contour plot of correlated bivariate normal variables, however, correspond to the right figure of Figure 6.3 and hence full bandwidth matrices are needed.

A scatter plot of the data thus might indicate which bandwidth matrix should be used. However, in higher dimensions this is more difficult or even impossible ($p > 3$).

Therefore, choosing a full bandwidth matrix might be the most preferable since a kernel parametrized by this matrix performs well in all cases. In Silverman (1986) a simple way of obtaining a full bandwidth matrix is suggested:

$$\mathbf{B} = b^2\mathbf{S}, \tag{6.4}$$

where $\mathbf{S}$ is the sample covariance matrix. In forensic applications this bandwidth matrix is used as well. Therefore, some theoretical motivation for this matrix is explained below.

Let the data $\overline{\mathbf{Z}}_1, \ldots, \overline{\mathbf{Z}}_m$ have sample mean $\boldsymbol{\mu}_s$ and sample covariance matrix $\mathbf{S}$. The transformation

$$\overline{\mathbf{Z}}_i^* = \mathbf{S}^{-\frac{1}{2}}(\overline{\mathbf{Z}}_i - \boldsymbol{\mu}_s)$$

is called *whitening* or *sphering* data (Koch (2014)). These names originate from the important properties that sphering the data results in uncorrelated variables with zero mean and unit variance. Sphered data have an identity covariance matrix, because

$$\mathrm{Var}\left(\overline{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\left(\overline{\mathbf{Z}}_i - \overline{\boldsymbol{\mu}}\right)\right) = \mathrm{E}\left(\overline{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\left(\overline{\mathbf{Z}}_i - \overline{\boldsymbol{\mu}}\right)\left(\overline{\mathbf{Z}}_i - \overline{\boldsymbol{\mu}}\right)'\overline{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\right) = \overline{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\overline{\boldsymbol{\Sigma}}\overline{\boldsymbol{\Sigma}}^{-\frac{1}{2}} = \mathbf{I}_{p\times p}$$

where $\overline{\boldsymbol{\mu}}$ and $\overline{\boldsymbol{\Sigma}}$ are the true mean and covariance of $\overline{\mathbf{Z}}_i$. Consequently, if the data is transformed the sphered data will look like Figure 6.4.

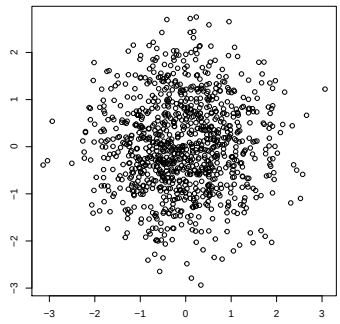


Figure 6.4: Scatter plot of sphered data, i.e. data with zero mean and identity covariance matrix.

From this figure it can be noticed why sphering the data might be useful. In fact, if the data looks like those visualized in Figure 6.4 it can be seen from Figure 6.3 that the bandwidth matrix $\mathbf{B} = b^2\mathbf{I}_{p\times p}$ would be suitable to use in the kernel density estimator. If this bandwidth matrix is used, the kernel density estimator for the transformed data in equation (6.2) boils down to the following formula

$$\hat{h}(\boldsymbol{\theta}^*) = m^{-1}b^{-p}\sum_{i=1}^{m} K\left(b^{-1}\left(\boldsymbol{\theta}^* - \overline{\mathbf{Z}}_i^*\right)\right)$$

where $\boldsymbol{\theta}$ is transformed in exactly the same way as $\overline{\mathbf{Z}}_i$. To find the kernel density estimator for the original data, we have to transform the variables back. To do this recall equation (4.6) for the variable transformation $\overline{\mathbf{Z}}_i = \mathbf{S}^{\frac{1}{2}}\overline{\mathbf{Z}}_i^* + \boldsymbol{\mu}_s$. Then, the kernel density estimator for $\boldsymbol{\theta}$ is

$$\begin{aligned}\hat{h}(\boldsymbol{\theta}) = &= \left|\mathbf{S}^{-\frac{1}{2}}\right| \hat{h}\left(\mathbf{S}^{-\frac{1}{2}}(\boldsymbol{\theta} - \mu_s)\right) \\ &= \left|\mathbf{S}^{-\frac{1}{2}}\right| m^{-1}b^{-p}\sum_{i=1}^{m} K\left(b^{-1}\mathbf{S}^{-\frac{1}{2}}(\boldsymbol{\theta} - \overline{\mathbf{Z}}_i)\right).\end{aligned}$$

Hence, using equation (6.2) it can be seen that the matrix $\mathbf{B}$ for the original data must satisfy the following equation:

$$\mathbf{B}^{-\frac{1}{2}} = b^{-1}\mathbf{S}^{-\frac{1}{2}}.$$

Since $\mathbf{B}^{\frac{1}{2}} = \mathbf{S}^{\frac{1}{2}}b$ it follows that equation (6.4) is indeed true.

Although this method results in a simple bandwidth matrix which depends on a single parameter b, Wand and Jones (1995) advise not to use this method. They claim that in case of multivariate normal data sphering is appropriate, but there is no corresponding theoretical support for estimation of general density shapes. To get some feeling for the problem, below the problem is illustrated using an example. Consider a bivariate normal density

$$\mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0.5\\0.5 & 1\end{pmatrix}\right)$$

and a mixture density of two bivariate normals

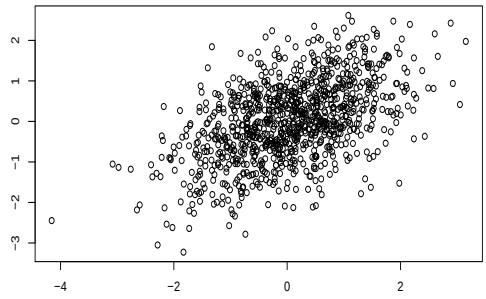
$$\frac{1}{2}\mathcal{N}\left(\begin{pmatrix}-2\\0\end{pmatrix}, \begin{pmatrix}\frac{1}{10} & 0\\0 & 1\end{pmatrix}\right) + \frac{1}{2}\mathcal{N}\left(\begin{pmatrix}2\\0\end{pmatrix}, \begin{pmatrix}\frac{1}{10} & 0\\0 & 1\end{pmatrix}\right),$$

which gives a non-normal shape. Scatter plots of variables which are generated from these two densities are given in Figure 6.5(a) and Figure 6.5(b). The problem of the bandwidth matrix in equation (6.4) comes from the fact that it depends on the sample covariance matrix $\mathbf{S}$. For multivariate normal variables as given in Figure 6.5(a) the sample covariance matrix approximates the true covariance matrix and hence the kernel density estimate performs as expected, see Figure 6.5(c) and Figure 6.5(e). For the mixture density, however, the sample covariance matrix is equal to

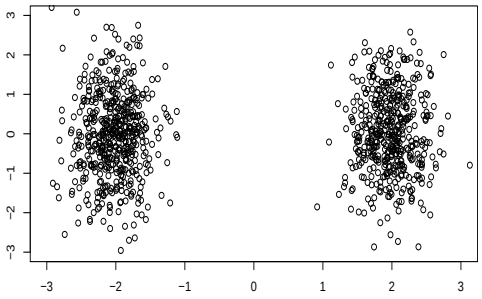
$$\begin{pmatrix}4.12 & 0.063\\0.063 & 0.95\end{pmatrix}.$$

Since the estimated variance of 4.12 is a very poor representation of the normal mixture, the amount of smoothing in the horizontal direction ($h^2 \cdot 4.12$) will not take into account the mixture shape of the density. Consequently, this is harmful for the result of the kernel density estimate which is illustrated in Figure 6.5(d) and Figure 6.5(f).

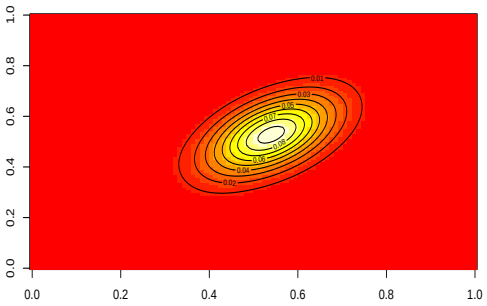
From Figure 6.5(f) it is clear that the two tops of the normal mixture are not captured in the kernel density estimate. Therefore, using the bandwidth matrix $\mathbf{B} = b^2\mathbf{S}$ can be very detrimental for the kernel density estimate of a simple non-Gaussian density. The choice for the bandwidth matrix can thus be crucial in the performance of the estimated density. In Section 4.1 it is mentioned that for xtc problems it can be expected that the density h will have multiple peaks and thus it will not be exceptional that the h will look like a kind of normal mixture. The example described above shows a disadvantage of the choice of the bandwidth matrix $\mathbf{B} = b^2\mathbf{S}$ and hence it might be useful to reconsider this choice in future development. However, in the following section we will see that considering a full bandwidth matrix $\mathbf{B}$ without further assumptions results in some other practical issues.



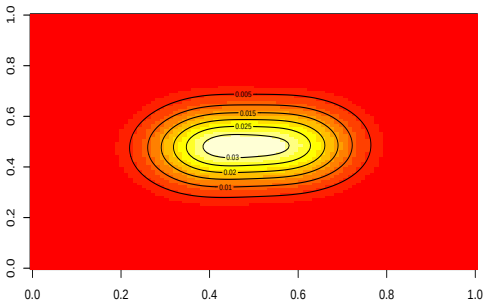
(a) Scatter plot of bivariate normal variables. 0.5.



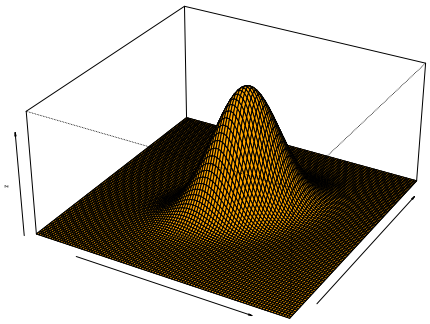
(b) Contour plot of the kernel density estimate of the bivariate normal density.



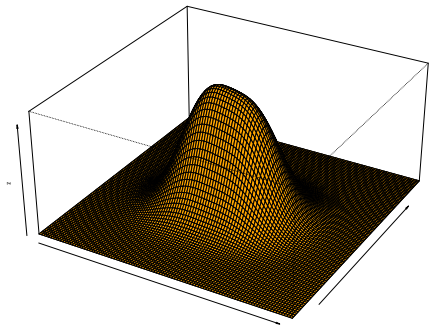
(c) Scatter plot of mixture of bivariate normal variables.



(d) Contour plot of the kernel density estimate of the mixture of bivariate normal densities.



(e) Kernel density estimate of the bivariate normal density.



(f) Kernel density estimate of the mixture of bivariate normal densities.

Figure 6.5: Two examples of kernel density estimates with a normal kernel and $\mathbf{B} = b^2\mathbf{S}$.

6.2.2 Optimal bandwidth selection

In Section 6.1 we have mentioned that the (univariate) kernel density estimate is sensitive to the choice of the bandwidth parameter. In Figure 6.2 we have seen that a small bandwidth b gives a rough (under smoothed) estimate, while a large bandwidth b gives

a smoother (over smoothed) estimate. Hence, in the literature a lot of attention is paid to the optimal choice of the bandwidth parameter.

To find an optimal value for b, the performance of the estimator h is taken into account. The performance can be measured locally in terms of the mean squared error. However, it is more convenient to study the global behaviour of the estimator through the mean integrated squared error (MISE):

$$\text{MISE}(\hat{h}) = \text{E}\left(\int \left(\hat{h}(\theta) - h(\theta)\right)^2 d\theta\right). \tag{6.5}$$

Hence, minimizing this risk with respect to the bandwidth b leads to an optimal bandwidth. Nevertheless, this optimal bandwidth depends on the unknown density h. Therefore there are different approaches to choose b in practice. In general it is recommended to use a plug-in type bandwidth, such as a reference bandwidth, or a cross-validation approach. Details about these approaches can be found in Wand and Jones (1995) or Wasserman (2006) for example.

For the multivariate problem we have seen that the choice for the bandwidth matrix plays an important role for the performance of the kernel density estimator. Moreover, the number of parameters b which has to be optimized depends on the choice for the bandwidth matrix. In Section 5.2.1 we have discussed that in applications of forensic statistics the matrix $\mathbf{B} = b^2\mathbf{S}$ is used. Hence, only one smoothing parameter b has to be found. Given the choice for the bandwidth matrix $\mathbf{B} = b^2\mathbf{S}$, in Silverman (1985) an optimal choice for b is discussed extensively. This optimal choice is based on the extension of the univariate problem, i.e. the multivariate version of the MISE given in equation (6.5) is minimized. This leads to (Silverman (1985))

$$b_{\text{opt}} = \left(p\beta\alpha^{-2}\left\{\int (\nabla^2 h)^2\right\}^{-1} m^{-1}\right)^{\frac{1}{p+4}},$$

where α and β are constants depending on the kernel K and ∇ is the gradient. If a reference bandwidth is used, this means that we have to choose a parametric family which can be substituted for h in the formula for b_{opt}. If a normal reference density h is assumed and a normal kernel is used, Silverman (1985) shows that the optimal bandwidth turns out to be the following simple plug-in formula:

$$b_{\text{opt}} = \left(\frac{4}{(p+2)m}\right)^{\frac{1}{p+4}}. \tag{6.6}$$

This optimal bandwidth can then be used directly in the kernel density estimator with normal kernel and bandwidth matrix $\mathbf{B} = b^2\mathbf{S}$. In forensic statistics this method is applied. However, it is worth noticing that for the determination of b_{opt} a cross-validation approach could be used as well.

In situations where an unconstrained full bandwidth matrix $\mathbf{B}$ would be preferable, e.g. the example in Section 5.2.1, the problem of finding optimal bandwidths becomes more complicated. If such a full matrix would be considered, $\frac{1}{2}p(p+1)$ bandwidth parameters should be determined. In Wand and Jones (1995) plug-in bandwidths and least squares cross validation is discussed to allow selection of an optimal full bandwidth matrix $\mathbf{B}$. However, these methods turns out to be hard to use in practice. Research to improve this method is still ongoing. In Duong and Hazelton (2005) an algorithm is presented that gives a fast and accurate computation for unconstrained bandwidth matrices.

6.3 Likelihood ratio

To find an explicit formula for the likelihood ratio in equation (3.15), the between-source density h is required. In Chapter 4 we assumed a Gaussian two-level model, hence a multivariate normal density will be substituted for h. In this chapter no pre-specified form for h is assumed, consequently h is estimated using the kernel density estimator given in equation (6.2). As mentioned in the previous sections, in forensic statistics it is common to use a normal kernel and a bandwidth matrix as given in equation (6.4). Since it is assumed that the means $\overline{\mathbf{Z}}_i$ are observations for the variable $\boldsymbol{\theta}$, in forensic literature the sample covariance matrix $\mathbf{S}$ in (6.4) is replaced by the true covariance matrix $\mathbf{T}$ of $\boldsymbol{\theta}$.[1] Consequently, the kernel density estimator is equal to

$$\hat{h}(\boldsymbol{\theta}) = \frac{1}{m}\sum_{i=1}^{m}(2\pi)^{-\frac{p}{2}}\left|b_{\text{opt}}^2\mathbf{T}\right|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\left(\boldsymbol{\theta}-\overline{\mathbf{Z}}_i\right)'\left(b_{\text{opt}}^2\mathbf{T}\right)^{-1}\left(\boldsymbol{\theta}-\overline{\mathbf{Z}}_i\right)\right), \tag{6.7}$$

with the optimal bandwidth b_{opt} as given in equation (6.6). The latter estimator $\hat{h}(\boldsymbol{\theta})$ can be substituted into the likelihood ratio given in equation (3.15). To find an explicit formula for the likelihood ratio, three integrals have to be solved. In Chapter 4 we have seen that for a Gaussian two-level model two likelihood ratios are derived in forensic literature, one based on direct integration and one based on a bayesian approach. For a non-Gaussian two-level model two likelihood ratios are derived in the same way. The resulting likelihood ratios are given in Aitken and Lucy (2004) and Bolck and Alberink (2011). Although it is shown in in Section 4.2.3 that the two different likelihood ratios are exactly the same for the Gaussian two-level model, in this thesis we will not show the equality of the likelihood ratios for the non-Gaussian two-level model. In Chapter 7 the likelihood ratio given in Bolck and Alberink (2011) is used for an application to real xtc data. Therefore this formula is stated below:

$$\begin{aligned}\text{LR}(\overline{\mathbf{y}}_1,\overline{\mathbf{y}}_2) &= \frac{\sum_{i=1}^{m}\exp\left(-\frac{1}{2}\left(\overline{\mathbf{y}}_1-\overline{\mathbf{z}}_i\right)'\mathbf{U}_{hx}^{-1}\left(\overline{\mathbf{y}}_1-\overline{\mathbf{z}}_i\right)-\frac{1}{2}\left(\overline{\mathbf{y}}_2-\boldsymbol{\mu}_{hi}\right)'\mathbf{U}_{hx}^{-1}\left(\overline{\mathbf{y}}_2-\boldsymbol{\mu}_{hi}\right)\right)}{\sum_{i=1}^{m}\exp\left(-\frac{1}{2}\left(\overline{\mathbf{y}}_1-\overline{\mathbf{z}}_i\right)'\mathbf{U}_{hx}^{-1}\left(\overline{\mathbf{y}}_1-\overline{\mathbf{z}}_i\right)\right)} \\ &\times \frac{m\left|\mathbf{U}_{hn}\right|^{-\frac{1}{2}}}{\left|\mathbf{U}_{h0}\right|^{-\frac{1}{2}}\sum_{i=1}^{m}\exp\left(-\frac{1}{2}\left(\overline{\mathbf{y}}_2-\overline{\mathbf{z}}_i\right)'\mathbf{U}_{h0}^{-1}\left(\overline{\mathbf{y}}_2-\overline{\mathbf{z}}_i\right)\right)}\end{aligned} \tag{6.8}$$

with

$$\begin{aligned}\boldsymbol{\mu}_{hi} &= b_{\text{opt}}^2\mathbf{T}\left(b_{\text{opt}}^2\mathbf{T}+n_1^{-1}\boldsymbol{\Sigma}\right)^{-1}\overline{\mathbf{y}}_1+n_1^{-1}\boldsymbol{\Sigma}\left(b_{\text{opt}}^2\mathbf{T}+n_1^{-1}\boldsymbol{\Sigma}\right)^{-1}\overline{\mathbf{z}}_i \\ \mathbf{U}_{hx} &= b_{\text{opt}}^2\mathbf{T}+n_1^{-1}\boldsymbol{\Sigma} \\ \mathbf{U}_{h0} &= b_{\text{opt}}^2\mathbf{T}+n_2^{-1}\boldsymbol{\Sigma} \\ \mathbf{U}_{hn} &= n_2^{-1}\boldsymbol{\Sigma}+b_{\text{opt}}^2\mathbf{T}-b_{\text{opt}}^2\mathbf{T}\left(b_{\text{opt}}^2\mathbf{T}+n_1^{-1}\boldsymbol{\Sigma}\right)^{-1}b_{\text{opt}}^2\mathbf{T}.\end{aligned}$$

The likelihood ratio in equation (6.8) depends on the covariance matrix $\mathbf{T}$ and on the covariance matrix $\boldsymbol{\Sigma}$. Chapter 5 covered the estimation of these covariance matrices for the Gaussian two-level model. For the non-Gaussian two-level model the matrices $\mathbf{T}$ and $\boldsymbol{\Sigma}$ can again be estimated using the analysis of variance estimators described in Section 5.3, because this method does not rely on the normality assumption. On the other hand the method of maximum likelihood described in Section 5.4 depends on the assumption of normality and is thus not natural to use to estimate the covariance matrices. Hence, it is suggested to use the analysis of variance estimators for $\boldsymbol{\Sigma}$ and $\mathbf{T}$.

[1] However, bear in mind that the covariance matrix of $\overline{\mathbf{Z}}_i$ is equal to $\mathbf{T}+n_i^{-1}\boldsymbol{\Sigma}$.

Statistical curse of dimensionality

A difficulty that occurs by using smoothing methods such as kernel density estimation is that the accuracy of the estimate decreases quickly if the dimension increases. In fact, if the data has dimension p then a sample size that grows exponentially with p is needed to obtain a required accuracy. This problem is referred to as the statistical curse of dimensionality. To get some feeling for this problem, we consider the following example. Suppose we want a point estimate of an unknown density f and we have chosen the bandwidth such that it minimizes the mean squared error on this point. The mean squared error of the estimate is approximately equal to (Wasserman (2006)):

$$\text{MSE} \approx cm^{-\frac{4}{p+4}},$$

for some $c > 0$ and sample size m. Then, the required sample size for a certain mean squared error is proportionally equal to

$$m \propto \left(\frac{c}{\text{MSE}}\right)^{\frac{p}{4}}.$$

Hence, the sample size grows exponentially with the dimension p. Silverman (1986) shows required sample sizes for a mean squared error less than 0.1 for estimation of a standard multivariate normal density using a normal kernel at the point zero. Some of these values are shown in Table 6.1. This table shows that up to four dimensions a reasonable sample size is needed, but nearly a million observations are needed in 10 dimensions. Furthermore, Silverman (1986) claims that if the mean integrated squared error (global risk) was used, the sample sizes would be approximately 1.7 times higher than is shown in Table 6.1.

Dimension p	Sample size m
1	4
2	19
3	67
4	223
8	43 700
10	842 000

Table 6.1: Required sample size to ensure that the MSE at zero is less than 0.1, when estimating a standard multivariate normal density using a normal kernel and bandwidth that minimizes the MSE at zero (Silverman (1986)).

Thus, in higher dimensions one may be able to compute an estimate but it will not be accurate. Therefore it is suggested that kernel density estimates should not be reported without confidence bands. These bands will be very wide for higher dimensions. A derivation for the confidence bands can for example be found in Wasserman (2006). A visualization for such bands in one dimensions is given in Figure 6.6.

The question of interest is what the impact of the increasing uncertainty in the estimate of h (growing with the dimension p) will be on the likelihood ratio in equation (3.15). When only an estimate of the density h is required, this uncertainty in h can be nicely illustrated by use of a confidence band. These bands can thus be of help when communicating results of the density estimate and its accuracy.[2]

[2]Note that this auxiliary plot only works in an understandable way for the one-dimensional case, however in higher dimensions the bands also exists. Moreover the bands are then even wider due to the curse of dimensionality.

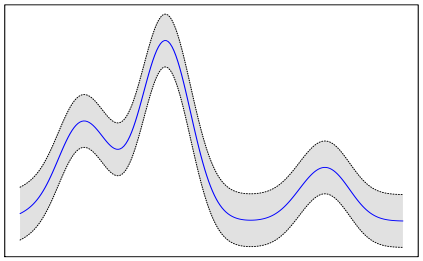


Figure 6.6: A visualization of confidence bands for a kernel density estimate.

However, in this application the density estimator in equation (6.7) will be substituted into the likelihood ratio in equation (3.15). Unfortunately the uncertainty in h now disappears and consequently the likelihood ratio cannot reflect the uncertainty brought by the density estimate of h. The error that is possibly made in the likelihood ratio is thus omitted. For example, the true density h could have been located very close to the upper bound of the confidence band in Figure 6.6. The functions that are integrated over θ in equation (3.15) using the density estimate $\hat{h}$ according to equation (5.8) (the blue line in Figure 6.6), will have a deviating shape compared to the functions that are integrated over θ using the true density h. Hence, the likelihood ratio that will be calculated using the blue line estimate will thus deviate from the likelihood ratio when the true density h would have been used. Important to note is that this error in the likelihood ratio is unavoidable. However, as stated in Wasserman (2006), it is necessary to be aware of this error in the model, especially when the dimensionality of the features increases and the sample size is not in line with Table 6.1. An interesting follow-up question would thus be to investigate this uncertainty in the likelihood ratio.

7

Likelihood ratios to combine discrete- and continuous evidence

In Chapter 3 we have seen that in forensic statistics two types of models exist to compute likelihood ratios in comparison problems. These models are applicable to either discrete- or continuous data (characteristics). This means that up to now, forensic experts could only report two separate likelihood ratios. For discrete data, the likelihood ratio is for example based on the color and logo of the tablets. For continuous data, the likelihood ratio is for example based on the weights and thickness of the tablets. The objective of this chapter is to describe a model that can be used to combine the discrete- and continuous evidence into one likelihood ratio.

In Section 7.1 the discrete- and continuous model described in Chapter 3 are combined such that in Section 7.2 a likelihood ratio is found that is applicable for a combination of discrete- and continuous evidence. In Section 7.3 the described theory will be illustrated using an example based on real xtc data.

7.1 Discrete- and continuous evidence model

Suppose that p_1 discrete features ($p_1 > 1$) and p_2 continuous features ($p_2 > 1$) of xtc tablets are measured by forensic experts. The evidence E is divided into the discrete- and continuous evidence,

$$E = (E_d, E_c).$$

Here, the discrete evidence E_d is given by $E_d = (\mathbf{Y}_1, \mathbf{Y}_2)$ where $\mathbf{Y}_1$ represents the p_1 discrete characteristics of tablets from consignment C_1, see Section 3.1. The evidence E_c is assumed to be $E_c = (\overline{\mathbf{Y}}_1, \overline{\mathbf{Y}}_2)$, where $\overline{\mathbf{Y}}_1$ are the means of the measurements of p_2 continuous characteristics of tablets in consignment C_1, as described in Section 3.2. In the same way, the vectors $\mathbf{Y}_2$ and $\overline{\mathbf{Y}}_2$ correspond to consignment C_2.

To describe a suitable model in this section, we will use the assumptions from the previous sections. This means that the assumptions in Section 3.1.1 are applicable for the discrete part of the evidence, E_d, and the assumptions in Section 3.2.1 are applicable for the continuous part, E_c. Thus, recall from equation (3.2) that $E_d = (\mathbf{Y}_1, \mathbf{Y}_2) \sim g_{\mathbf{Y}_1,\mathbf{Y}_2}$ and $\mathbf{Y}_1$ and $\mathbf{Y}_2$ share their probability mass function,

$$\mathbf{Y}_l \sim g \qquad \text{for} \qquad l \in \{1, 2\}.$$

Remember from Section 3.2 that we assume a two-level model for the continuous evidence. Consequently, the group means are drawn from the between source density,

$$\boldsymbol{\theta}_l \sim h \qquad \text{for} \qquad l \in \{1, 2\}$$

and conditional on $\boldsymbol{\theta}_l$ the means of the measurements are distributed according to a multivariate normal density f as given in equation (3.6),

$$\overline{\mathbf{Y}}_l \mid \boldsymbol{\theta}_l \quad \sim \quad f_{\overline{\mathbf{Y}}_l|\boldsymbol{\theta}_l} \qquad \text{for} \qquad l \in \{1,2\}.$$

In the next section we will see that by conditioning on the discrete vectors, we can express the likelihood ratio in an intuitive way in terms of the probability functions for the continuous- and discrete random vectors.

7.2 Likelihood ratio

The purpose of this section is to find an expression for the likelihood ratio for a combination of discrete- and continuous evidence. To find this expression the techniques described in Section 3.1.2 and Section 3.2.2 will be used. Since we are working with a combination of discrete- and continuous vectors the likelihood ratio can be written as a combination of the definition in equation (3.3) and the approximation in equation (3.7), i.e.

$$\mathrm{LR}(e_d, e_c) = \frac{P(\mathbf{Y}_1 = \mathbf{y}_1, \mathbf{Y}_2 = \mathbf{y}_2, \overline{\mathbf{Y}}_1 \in [\overline{\mathbf{y}}_1 \pm \delta \mathbf{1}_p], \overline{\mathbf{Y}}_2 \in [\overline{\mathbf{y}}_2 \pm \delta \mathbf{1}_p] \mid H_p, I)}{P(\mathbf{Y}_1 = \mathbf{y}_1, \mathbf{Y}_2 = \mathbf{y}_2, \overline{\mathbf{Y}}_1 \in [\overline{\mathbf{y}}_1 \pm \delta \mathbf{1}_p], \overline{\mathbf{Y}}_2 \in [\overline{\mathbf{y}}_2 \pm \delta \mathbf{1}_p] \mid H_d, I)} \tag{7.1}$$

where $e_d = (\mathbf{y}_1, \mathbf{y}_2)$, $e_c = (\overline{\mathbf{y}}_1, \overline{\mathbf{y}}_2)$ and δ some small and positive constant. Below, a useful expression for this likelihood ratio will be derived.

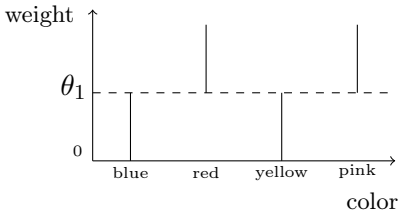
First assume that H_p is true, i.e. the consignments C_1 and C_2 come from the same batch. Recall from Section 3.1 that we will assume that tablets will have the same discrete features if they come from the same production batch. Then, the numerator of equation (7.1) can be written as

$$\begin{aligned} P((E_d, E_c) = (e_d, e_c) \mid I) &= P(\mathbf{Y}_1 = \mathbf{y}_1, \overline{\mathbf{Y}}_1 \in [\overline{\mathbf{y}}_1 \pm \delta \mathbf{1}_p], \overline{\mathbf{Y}}_2 \in [\overline{\mathbf{y}}_2 \pm \delta \mathbf{1}_p] \mid I) \mathbb{1}_{\{\mathbf{y}_1 = \mathbf{y}_2\}} \\ &= P(\overline{\mathbf{Y}}_1 \in [\overline{\mathbf{y}}_1 \pm \delta \mathbf{1}_p], \overline{\mathbf{Y}}_2 \in [\overline{\mathbf{y}}_2 \pm \delta \mathbf{1}_p] \mid \mathbf{Y}_1 = \mathbf{y}_1, I) \\ &\times \; g(\mathbf{y}_1 \mid I) \mathbb{1}_{\{\mathbf{y}_1 = \mathbf{y}_2\}}. \end{aligned}$$

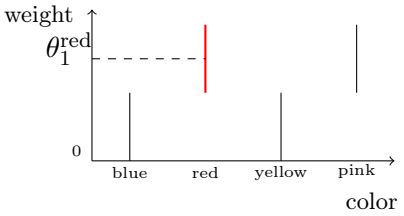
Hence, to find a useful expression for the numerator we have to consider the conditional probability of the continuous evidence given the discrete evidence,

$$P(\overline{\mathbf{Y}}_1 \in [\overline{\mathbf{y}}_1 \pm \delta \mathbf{1}_p], \overline{\mathbf{Y}}_2 \in [\overline{\mathbf{y}}_2 \pm \delta \mathbf{1}_p] \mid \mathbf{Y}_1 = \mathbf{y}_1, I). \tag{7.2}$$

To find an approximation for this probability, we consider the following example. Suppose that we are interested in the weight $\overline{Y}_1$ and color Y_1 of xtc tablets. We assume that xtc tablets can only have four different colors and the weight of the tablets has a true mean θ_1. An example of a rough sketch of a joint feature space of the weight and color of xtc tablets is given in Figure 7.1(a). Now suppose that we are interested in the probability that tablets have weight $\overline{y}_1$ given that their color is red, i.e. $P(\overline{Y}_1 \in [\overline{y}_1 \pm \delta \mathbf{1}_p] \mid Y_1 = \text{red})$. To find this probability we would focus on one specific part of the feature space, see Figure 7.1(b). Hence, the mean of the weight we are interested in will change from the "overall" true mean θ_1 to the mean of the red tablets θ_1^{red}. To approximate the probability in equation (7.2) we will use the same idea.



(a) Rough sketch of a joint feature space of weight and color.



(b) To find the conditional probability of a certain weight given that the color was red, focus on the red part of the feature space.

Figure 7.1: An example of a (conditional) feature space of the weight and color of xtc tablets. The joint density is positive for values in the feature space indicated by the vertical lines.

Recall from Section 3.2.2 that under H_p we can assume that the true means of $\overline{\mathbf{Y}}_1$ and $\overline{\mathbf{Y}}_2$ are the same, i.e. the true mean of the continuous control- and recovered data is $\boldsymbol{\theta} \sim h$. However, since we are interested in the probability of the continuous evidence given that certain discrete features features are observed, the mean we are interested in will be restricted to $\boldsymbol{\theta}^{\mathbf{y}_1}$. In other words, that is the true mean of the continuous features from tablets that have discrete features $\mathbf{y}_1$. Then, by using Lemma 3.2.1, the probability in equation (7.2) can be approximated by

$$(2\delta)^{2p_2} \int_{\boldsymbol{\theta}^{\mathbf{y}_1}} f_{\overline{\mathbf{Y}}_1, \overline{\mathbf{Y}}_2 | \boldsymbol{\theta}^{\mathbf{y}_1}, I}(\overline{\mathbf{y}}_1, \overline{\mathbf{y}}_2 \mid \boldsymbol{\theta}^{\mathbf{y}_1}, I) h(\boldsymbol{\theta}^{\mathbf{y}_1} \mid I) d\boldsymbol{\theta}^{\mathbf{y}_1}.$$

Since $\overline{\mathbf{Y}}_1$ and $\overline{\mathbf{Y}}_2$ are conditionally independent given their mean $\boldsymbol{\theta}^{\mathbf{y}_1}$, we can use the following expression for the numerator in equation (7.1):

$$\begin{aligned} P((E_d, E_c) = (e_d, e_c) \mid H_p, I) &= (2\delta)^{2p_2} g(\mathbf{y}_1 \mid I) \mathbb{1}_{\{\mathbf{y}_1 = \mathbf{y}_2\}} \int_{\boldsymbol{\theta}^{\mathbf{y}_1}} f_{\overline{\mathbf{Y}}_1 | \boldsymbol{\theta}^{\mathbf{y}_1}, I}(\overline{\mathbf{y}}_1 \mid \boldsymbol{\theta}^{\mathbf{y}_1}, I) \\ &\times f_{\overline{\mathbf{Y}}_2 | \boldsymbol{\theta}^{\mathbf{y}_1}, I}(\overline{\mathbf{y}}_2 \mid \boldsymbol{\theta}^{\mathbf{y}_1}, I) h(\boldsymbol{\theta}^{\mathbf{y}_1} \mid I) d\boldsymbol{\theta}^{\mathbf{y}_1}. \end{aligned} \tag{7.3}$$

Now assume that H_d is true, i.e. the consignments C_1 and C_2 come from different sources. Since $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are independent under H_d, the denominator of equation (7.1) can be written as

$$\begin{aligned} P((E_d, E_c) = (e_d, e_c) \mid I) &= P(\overline{\mathbf{Y}}_1 \in [\overline{\mathbf{y}}_1 \pm \delta], \overline{\mathbf{Y}}_2 \in [\overline{\mathbf{y}}_2 \pm \delta] \mid \mathbf{Y}_1 = \mathbf{y}_1, \mathbf{Y}_2 = \mathbf{y}_2, I) \\ &\times g(\mathbf{y}_1 \mid I) g(\mathbf{y}_2 \mid I). \end{aligned}$$

Again, conditioning on $\mathbf{Y}_1$ and $\mathbf{Y}_2$ can be seen as restricting the true means $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ to $\boldsymbol{\theta}_1^{\mathbf{y}_1}$ and $\boldsymbol{\theta}_2^{\mathbf{y}_2}$, such that they are based on tablets with discrete features $\mathbf{y}_1$ and $\mathbf{y}_2$ respectively. Then, by using Lemma 3.2.1 and the independence of $\overline{\mathbf{Y}}_1$ and $\overline{\mathbf{Y}}_2$ given their means $\boldsymbol{\theta}^{\mathbf{y}_1}$ and $\boldsymbol{\theta}^{\mathbf{y}_2}$, the following expression can be used for the denominator of equation (7.1):

$$\begin{aligned} P((E_d, E_c) = (e_d, e_c) \mid H_d, I) &= (2\delta)^{2p_2} \int_{\boldsymbol{\theta}_1^{\mathbf{y}_1}} f_{\overline{\mathbf{Y}}_1 | \boldsymbol{\theta}_1^{\mathbf{y}_1}, I}(\overline{\mathbf{y}}_1 \mid \boldsymbol{\theta}_1^{\mathbf{y}_1}, I) h(\boldsymbol{\theta}_1^{\mathbf{y}_1} \mid I) d\boldsymbol{\theta}_1^{\mathbf{y}_1} \\ &\times \int_{\boldsymbol{\theta}_2^{\mathbf{y}_2}} f_{\overline{\mathbf{Y}}_2 | \boldsymbol{\theta}_2^{\mathbf{y}_2}, I}(\overline{\mathbf{y}_2} \mid \boldsymbol{\theta}_2^{\mathbf{y}_2}, I) h(\boldsymbol{\theta}_2^{\mathbf{y}_2} \mid I) d\boldsymbol{\theta}_2^{\mathbf{y}_2} \\ &\times g(\mathbf{y}_1 \mid I) g(\mathbf{y}_2 \mid I). \end{aligned} \tag{7.4}$$

Thus, by combining equation (7.3) and equation (7.4), the following expression for the likelihood ratio in equation (7.1) can be used:

$$\begin{aligned} \mathrm{LR}(e_d, e_c) &= \frac{\int_{\boldsymbol{\theta}^{\mathbf{y}_1}} f_{\overline{\mathbf{Y}}_1|\boldsymbol{\theta}^{\mathbf{y}_1},I}(\overline{\mathbf{y}}_1 \mid \boldsymbol{\theta}^{\mathbf{y}_1}, I) f_{\overline{\mathbf{Y}}_2|\boldsymbol{\theta}^{\mathbf{y}_1},I}(\overline{\mathbf{y}}_2 \mid \boldsymbol{\theta}^{\mathbf{y}_1}, I) h(\boldsymbol{\theta}^{\mathbf{y}_1} \mid I) d\boldsymbol{\theta}^{\mathbf{y}_1}}{g(\mathbf{y}_2 \mid I) \int_{\boldsymbol{\theta}_1^{\mathbf{y}_1}} f_{\overline{\mathbf{Y}}_1|\boldsymbol{\theta}_1^{\mathbf{y}_1},I}(\overline{\mathbf{y}}_1 \mid \boldsymbol{\theta}_1^{\mathbf{y}_1}, I) h(\boldsymbol{\theta}_1^{\mathbf{y}_1} \mid I) d\boldsymbol{\theta}_1^{\mathbf{y}_1}} \\ &\times \frac{1}{\int_{\boldsymbol{\theta}_2^{\mathbf{y}_2}} f_{\overline{\mathbf{Y}}_2|\boldsymbol{\theta}_2^{\mathbf{y}_2},I}(\overline{\mathbf{y}}_2 \mid \boldsymbol{\theta}_2^{\mathbf{y}_2}, I) h(\boldsymbol{\theta}_2^{\mathbf{y}_2} \mid I) d\boldsymbol{\theta}_2^{\mathbf{y}_2}}, \end{aligned} \tag{7.5}$$

where $e_d = (\mathbf{y}_1, \mathbf{y}_2)$ and $e_c = (\overline{\mathbf{y}}_1, \overline{\mathbf{y}}_2)$. If the discrete part of the evidence is not the same ($\mathbf{y}_1 \neq \mathbf{y}_2$), the likelihood ratio will be equal to zero.

It can be seen that this likelihood ratio is almost the same expression as the product of the discrete likelihood ratio and the continuous likelihood ratio, given in equations (3.5) and (3.15). However, there is a subtle difference in the continuous part of the likelihood ratio since we have restricted the mean vector $\boldsymbol{\theta}$ to the discrete evidence, $\boldsymbol{\theta}^{\mathbf{y}_1}$. This restriction can affect the choice of the between source density and thus the computation of the continuous part of the likelihood ratio. An example of the computation of this likelihood ratio will be given in the next section. If the continuous features are independent of the discrete features, the likelihood ratio will be exactly the product of equation (3.5) and equation (3.15).

7.3 Application on real xtc data

To apply the described theory about likelihood ratios for combined evidence, we will consider an example based on real xtc data. The background data that is used is based on (a subset of) real xtc data and therefore the data is anonymized. Most of the xtc data collected by the NFI does not contain the combination of continuous- and discrete features. Hence, a relative small data set is used in the problem below. However, since this problem is only an illustration and not based on a real lawsuit this is not an issue. This section can therefore also be seen as a motivation to collect more data consisting of both discrete- and continuous features in the future.

7.3.1 Problem definition

Consider a lawsuit about the origin of two seized consignments C_1 and C_2. The prosecutor's hypothesis H_p states that the consignments C_1 and C_2 come from the same production batch. The hypothesis of the defense H_d supposes that the consignments come from different production batches (see Section 2.1.1). The task of a forensic drug expert is to provide the judge with the likelihood ratio. To calculate the likelihood ratio, the forensic expert is able to examine one tablet from both consignments. He established the color and logo of the tablets and measured two continuous post-tabletting features A and B of the tablets in both consignments. The color, logo and features A and B of each consignment is given in the table below.

Consignment	Logo	Color	A	B
C_1	Harley Davidson	Beige	0.238	0.719
C_2	Harley Davidson	Beige	0.213	0.676

Table 7.1: The logo, color and features A and B of the consignments C_1 and C_2 measured by a forensic expert.

In this example we will assume that consignments that originate from the same production batch will have the same discrete features, see Section 2.1.1. Furthermore, we will assume that there is no further background information I available.

To calculate the likelihood ratio, a suitable background database is required to estimate the parameters (see Section 5.1). In this example we will use background data which consist of consignments that are confiscated in the Netherlands. A small subset of the available background data is given in the table below.

Batch	**Tablet**	**Color**	**A**	**B**
1	1	Beige	0.533	0.682
1	2	Beige	0.570	0.674
⋮	⋮	⋮	⋮	⋮
1	20	Beige	0.592	0.677
2	1	Pink	0.433	0.527
⋮	⋮	⋮	⋮	⋮
50	1	Red	0.475	0.361
⋮	⋮	⋮	⋮	⋮
50	8	Red	0.471	0.362

Table 7.2: A subset of the available background data $\mathbf{Z}$.

The available background data $\mathbf{Z} = (\mathbf{Z}_{ij} \mid 1 \leq i \leq 50, 1 \leq j \leq n_i)$ consists of m batches and $1 \leq n_i \leq 20$ measured tablets within each batch. The vector $\mathbf{Z}_{ij}$ contains measurements of the color, feature A and feature B within batch i on tablet j. For example, $\mathbf{Z}_{11} = (\text{beige}, 0.533, 0.682)$. Note that the color is equal within each batch i, while the continuous features A and B can differ for each tablet.

Since the logos of the 50 batches in the background data are not established, the forensic expert will not use the logo of consignment C_1 and C_2 in the likelihood ratio calculation. Therefore, we will only consider the color of the tablets as discrete evidence and according to the model described in Section 3.1.1 the discrete evidence is then given by $E_d = (\text{beige}, \text{beige})$. The forensic expert is able to measure the continuous features A and B of one tablet from each consignment, i.e. $n_1 = n_2 = 1$. Then according to the model described in Section 3.2.1 the continuous evidence is equal to the measurements on the control- and recovered data, $E_c = ((0.238, 0.719), (0.213, 0.676))$. First we will compute two separate likelihood ratios, one based on the discrete evidence and the other based on the continuous evidence. After that, a likelihood ratio based on a combination of the discrete and continuous evidence will be computed using the theory described in Section 7.1 and Section 7.2.

7.3.2 Discrete evidence

The forensic expert has established that the discrete evidence is given by $E_d = (\text{beige}, \text{beige})$. To determine the strength of evidence of two consignments both having beige tablets, we will assume the model described in Section 3.1.1. Since the control- and recovered data are the same and we do not have any background information I, we know from equation (3.5) that the likelihood ratio is equal to

$$\text{LR}(\text{beige}, \text{beige}) = \frac{1}{g(\text{beige})}.$$

To estimate the probability of a consignment containing beige tablets, $g(\text{beige})$, we will use the frequency of batches containing beige tablets in the background data $\mathbf{Z}$. An example of the frequency of some colors in the background data are given below.

Color	**Frequency**	**Estimated probability**
Beige	4	0.08
Pink	11	0.22
Red	2	0.04
Blue	8	0.16
Light green	2	0.04

Table 7.3: Example of the frequency of some colors in the database.

Table 7.3 shows that 4 batches in the database contain beige tablets, hence the required probability is estimated as $\hat{g}(\text{beige}) = 4/50 = 0.08$. Therefore, the likelihood ratio is equal to

$$\text{LR}(\text{beige}, \text{beige}) = \frac{1}{0.08} = 12.5. \tag{7.6}$$

Then, according to Table 2.2, we would conclude that the matching beige color is *more probable* if the consignments come from the same batch than if they come from different batches.

In Section 2.2.1 we described a difficulty in this approach, due to the fast increase of new designs of xtc tablets. Since batches that contain beige tablets occur only four times in the database, seizing two consignments with beige tablets is considered as a rare event. Consequently, the likelihood ratio indicates that the forensic findings are more probable given the prosecutors hypothesis. It could happen that beige colored tablets is a new popular design and more consignments with this color are confiscated by the police later this month. This could affect the strength of evidence. For example, suppose that at the end of this month 6 additional consignments with beige tablets are confiscated. Since beige tablets occur more frequently, the estimated probability will be higher than before and hence the likelihood ratio will be smaller, i.e.

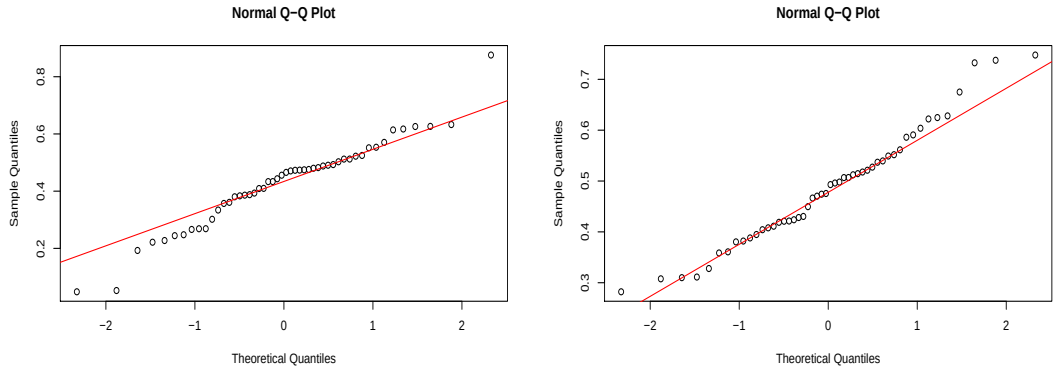
$$\text{LR}(\text{beige}, \text{beige}) = \frac{1}{10/50} = 5.$$

Consequently, the verbal likelihood ratio will turn into: "the matching beige color is *slightly more probable* if the consignments come from the same batch than if they come from different batches". Then, the evidence will be less strong than the reported value in equation (7.6) at the beginning of the month. This problem is considered to be an open problem in forensic statistics. One of the suggested approaches is to use a dynamic database, e.g. a database which includes data of the last three years.

7.3.3 Continuous evidence

To determine the strength of the continuous evidence we have to compute the likelihood ratio given in equation (3.15). To compute this likelihood ratio we want to decide whether it is reasonable to assume a Gaussian between-source distribution or not. To evaluate the assumption of normality, the background data $\mathbf{Z}$ will be used. Since we are focusing on the continuous evidence, we consider the background data $\mathbf{Z}$ with vectors $\mathbf{Z}_{ij}$ containing only the measurements of the continuous features A and B.

The vectors $\overline{\mathbf{Z}}_1, \ldots, \overline{\mathbf{Z}}_{50}$ are the batch means taken over the 50 batches for these two continuous features. To assess whether the observations $\overline{\mathbf{Z}}_1, \ldots, \overline{\mathbf{Z}}_{50}$ are samples from a bivariate normal distribution, first the means $\overline{Z}_{1k}, \ldots, \overline{Z}_{50k}$, $k \in \{1, 2\}$ are examined for (univariate) normality using the techniques described in Section 4.1.2. In Figure 7.2(a) and Figure 7.2(b) the normal QQ-plots of these means are given. In Table 7.4 results from the Shapiro-Wilk test are given. Since the points in the normal QQ-plot are approximately on a straight line and the p-values of the Shapiro-Wilk test are greater than 0.05 (the test-statistic W is close to one) we will assume that the marginals of $\overline{\mathbf{Z}}_1, \ldots, \overline{\mathbf{Z}}_{50}$ are normal distributed.



(a) (b)

Figure 7.2: Normal QQ-plots of the means $\overline{Z}_{1k}, \ldots, \overline{Z}_{50k}$, $k \in \{1, 2\}$, of the features A ($k = 1$) and B ($k = 2$).

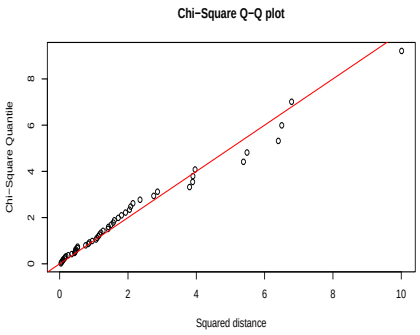


Figure 7.3: Chi-squared QQ plot to assess the multivariate normality of $\overline{\mathbf{Z}}_1, \ldots, \overline{\mathbf{Z}}_{50}$.

Shapiro-Wilk		
Feature	**W**	p**-value**
A	0.962	0.109
B	0.972	0.290
Mardia's MVN		
	Estimate	p**-value**
Skewness	0.439	0.454
Kurtosis	8.693	0.540

Table 7.4: Results from the univariate and multivariate goodness of fit tests to assess the assumption of normality.

Although the marginals are assumed to be normal distributed individually, testing the multivariate structure on normality is required as well. In Section 4.1.2 we have seen that a Chi-squared QQ-plot can be used for this purpose, provided that $m - p = 50 - 2 > 25$. In Figure 7.3 the generalized distances $d_1^2, \ldots, d_m^2$ are plotted against the quantiles of a Chi-squared distribution with 2 degrees of freedom. Since these points are approximatley on a straight line through the origin, this plot indicates multivariate normality. Furthermore, in Table 7.4 results from Mardia's goodness of fit test for multivariate normality are given. This test is based on a multivariate extension of

skewness and kurtosis, see for example Mardia (1980) and Korkmaz et al. (2015). Both the skewness and kurtosis estimates indicate multivariate normality. Hence, according to the Chi-squared QQ-plot and the goodness of fit test we assume that the data $\overline{\mathbf{Z}}_1, \ldots, \overline{\mathbf{Z}}_{50}$ follows a multivariate normal distribution.

Encouraged by these results we will assume that it is appropriate to model the continuous evidence using a Gaussian two-level model as described in Chapter 4. Consequently, we will compute the likelihood ratio which is given in equation (4.12). Since this likelihood ratio depends on the parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\mathbf{T}$, estimation of these parameters is required. The EM-algorithm is used to estimate the parameters, see Section 5.4.

The identity matrices are chosen as starting values and the relative change of the parameters for $\delta = 5$ as the stopping criterion, see Section 5.5.3. Using these starting values and stopping criterion 18 iterations were needed. The incomplete log likelihood is given in Figure 7.4. The estimated parameters are listed below:

$$\widehat{\boldsymbol{\mu}} = \begin{pmatrix} 0.419 \\ 0.495 \end{pmatrix}, \quad \widehat{\boldsymbol{\Sigma}} = \begin{pmatrix} 0.042 & -0.004 \\ -0.004 & 0.023 \end{pmatrix}, \quad \widehat{\mathbf{T}} = \begin{pmatrix} 9.482 \times 10^{-3} & 4.210 \times 10^{-3} \\ 4.210 \times 10^{-3} & 6.007 \times 10^{-3} \end{pmatrix}.$$

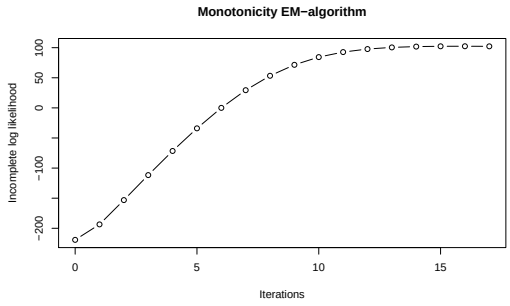


Figure 7.4: Incomplete log likelihood for each iteration of the EM-algorithm.

Using these parameter estimates, the estimated likelihood ratio in equation (4.12) can now be calculated

$$\mathrm{LR}\Big((0.238, 0.719), (0.213, 0.676)\Big) = 1.38. \tag{7.7}$$

Then, according to Table 2.2, we would conclude that the continuous evidence provides no assistance in addressing the issue.

7.3.4 Combination of discrete- and continuous evidence

In Section 7.3.2 and Section 7.3.3 we have seen the computation of two separate likelihood ratios, one based on the discrete evidence and the other based on the continuous evidence. If the discrete and continuous evidence are assumed to be independent, the combined likelihood ratio would be the product of these two separate likelihood ratios. However, in Section 2.1.1 we have argued that since post-tabletting characteristics are formed within one source, it is likely that there is a certain dependency between these characteristics. Consequently, in this example it is also assumed that there is a certain dependence between the color and the continuous features A and B of the tablets. Hence, in a situation like this, a forensic expert could only report two separate likelihood ratios. By using the theory from Section 7.1 and Section 7.2 it will be shown how the evidence can be combined into one likelihood ratio.

The evidence E is divided into the discrete- and continuous evidence, $E = (E_d, E_c)$. The likelihood ratio for this combined evidence is given in equation (7.5). We have seen that this likelihood ratio is almost the same expression as the product of the "discrete likelihood ratio" and the "continuous likelihood ratio". For the discrete part of the likelihood ratio, the likelihood ratio given in equation (7.6) can be used.

However, the continuous part of the likelihood ratio is slightly different than the likelihood ratio given in equation (7.7). In the continuous part of the likelihood ratio in equation (7.5) we restrict the mean vector $\boldsymbol{\theta}$ to the discrete evidence $\boldsymbol{\theta}^{\mathbf{y}_1}$. This means that we condition the true mean vector $\boldsymbol{\theta}$ of continuous features A and B on the discrete evidence $\mathbf{y}_1$, i.e. we consider the continuous features A and B for tablets with a beige color. It is worth noting that this results in the same likelihood ratio as the likelihood ratio in equation (3.15) when the discrete data is seen as the background information I. In Section 5.1 it is explained that conditioning on the background information I can be of influence to the background data that should be used. This automatically affects the estimated parameters. If the background information consists of the discrete features, the resulting likelihood ratio in equation (3.15) is thus exactly the same as the continuous part of the likelihood ratio in equation (7.5).

Hence, to decide whether a Gaussian between-source distribution can be used for $\boldsymbol{\theta}^{\text{beige}}$, we have to use a restricted background data set $\mathbf{Z}^{\text{beige}}$ instead of the total background data set $\mathbf{Z}$. The restricted data set only contains the batches in the background data which have a beige color, i.e. $\mathbf{Z}^{\text{beige}} = (\mathbf{Z}_{ij}, 1 \leq i \leq 4, 1 \leq j \leq n_i)$ where $2 \leq n_i \leq 20$. Consequently, to evaluate the assumption of normality for $\boldsymbol{\theta}^{\text{beige}}$, only four observations $\overline{\mathbf{Z}}_1, \overline{\mathbf{Z}}_2, \overline{\mathbf{Z}}_3, \overline{\mathbf{Z}}_4$ can be used. For such a small sample only very aberrant behaviour will be identified by goodness of fit tests as a lack of fit from normality. Recall that at this moment larger data sets which consist of both discrete- and continuous features are not available. For this example we will therefore assume that the between-source density of $\boldsymbol{\theta}^{\text{beige}}$ is normal. Note that when larger sets are available in the future, the normality test will be more decisive.

Hence, we will again assume that it is appropriate to model the continuous evidence using a Gaussian two-level model. The computation of the continuous part of the likelihood ratio in equation (7.5) is now equal to the likelihood ratio given in equation (4.1) except for the fact that the parameters are restricted to the discrete evidence, $\boldsymbol{\mu}^{\text{beige}}, \boldsymbol{\Sigma}^{\text{beige}}$ and $\mathbf{T}^{\text{beige}}$. To estimate these parameters based on the restricted background data $\mathbf{Z}^{\text{beige}}$, again the EM-algorithm will be used with identity starting values and the relative changes in the parameters as stopping criterion. To obtain the estimated parameters which are given below, 8 iterations were needed.

$$\begin{aligned} \widehat{\boldsymbol{\mu}}^{\text{beige}} &= \begin{pmatrix} 0.359 \\ 0.572 \end{pmatrix}, \quad \widehat{\boldsymbol{\Sigma}}^{\text{beige}} = \begin{pmatrix} 4.751 \times 10^{-4} & 2.053 \times 10^{-5} \\ 2.053 \times 10^{-5} & 1.306 \times 10^{-4} \end{pmatrix}, \\ \widehat{\mathbf{T}}^{\text{beige}} &= \begin{pmatrix} 0.046 & 0.005 \\ 0.005 & 0.003 \end{pmatrix}. \end{aligned}$$

By using these parameter estimates, the likelihood ratio of the continuous features of beige tablets is equal to 17.2. By using the discrete likelihood ratio in equation (7.6), the likelihood ratio of the combined evidence (equation (7.5)) can now be calculated,

$$\text{LR}\Big(E_d, E_c\Big) = 12.5 \times 17.2 = 215. \tag{7.8}$$

Thus, for this example we have seen that based on the evidence which is used, different conclusions are obtained:

(i) **Discrete evidence**: Based on the color the forensic expert would conclude that the discrete evidence is *more probable* if the consignments come from the same batch than if they come from different batches.

(ii) **Continuous evidence**: Based on the features A and B the forensic expert would conclude that the continuous evidence provides *no assistance in addressing the issue.*

(iii) **Discrete- and continuous evidence**: Based on both the color and the features A and B, the forensic expert would conclude that the discrete and continuous evidence is *much more probable* if the consignments come from the same batch than if they come from different batches.

In this section we have seen the calculation of the likelihood ratio for a combination of discrete- and continuous evidence. We mentioned that the available (combined) database is small and more data should be collected in the future. However, for many discrete features a bigger combined data set is not feasible, because the number of categories is too big (recall that the number of new xtc designs is increasing fast). Hence, discrete features with less possible categories are more appropriate to use in a likelihood ratio for combined evidence. Two examples of such features are shape or diameter[1].

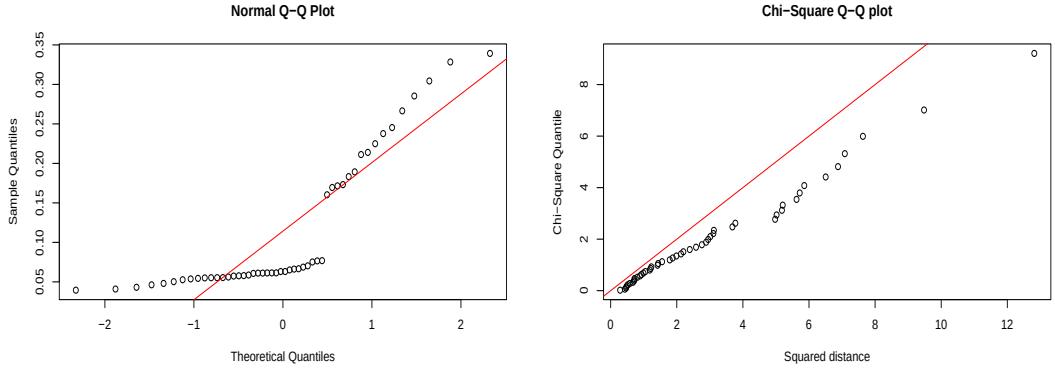
7.3.5 A non-Gaussian between-source distribution

In Section 7.3.3 we have seen that for the problem described in Section 7.3.1, it can be assumed that the between-source distribution is a normal distribution. However, in many practical cases this assumption is not valid. Therefore, the non-Gaussian two-level model described in Chapter 6 is important. Below, the use of the non-Gaussian two-level model in practice will be briefly illustrated by an extension of the example described in Section 7.3.1.

Suppose that an additional continuous feature C is measured by forensic experts. In the background data, each vector $\mathbf{Z}_{ij}$ now contains measurements of the features A, B and C. Likewise in Section 7.3.3, to assess whether the between-source distribution is a multivariate normal distribution, we consider the batch means $\overline{\mathbf{Z}}_1, \ldots, \overline{\mathbf{Z}}_{50}$ for these three continuous features. In Figure 7.5(a) a QQ plot of the marginal distribution of the batch means of feature C is given. In Figure 7.5(b) a QQ-plot of the squared generalized distances is given. Both visual tests shows that multivariate normality cannot be assumed, since the plotted points deviate from the straight reference lines. Hence, a non-Gaussian model will be assumed to model the continuous evidence and the likelihood ratio from equation (6.8) can be applied.

If it is desired to combine the discrete- and continuous evidence as described in Section 7.3.4, the restricted background data $\mathbf{Z}^{\text{beige}}$ still consist of four batches. Due to this small sample size, the multivariate normality assumption for the between-source distribution will again not be rejected. On the other hand, the use of kernel density estimation for such a small sample for three dimensions can be questioned as well (see Table 6.1). Therefore, both reasons should be seen as motivation to collect more combined data in the future.

[1]Currently it is being discussed whether diameter should be treated as a continuous variable or a discrete variable. Forensic experts have strong suspicions that tabletting machines can only produce tablet of 4, 5 or 6 millimeter. If this is indeed true, the measurements can be categorized



(a) Normal QQ-plot of $\overline{Z}_{1k}, \ldots, \overline{Z}_{50k}$ of the feature C ($k = 3$).

(b) Chi-squared QQ-plot to assess the multivariate normality of $\overline{\mathbf{Z}}_1, \ldots, \overline{\mathbf{Z}}_{50}$.

Figure 7.5: QQ-plots to assess the assumption of multivariate normality.

Conclusion

The ENFSI-LR project aims to construct a GUI around software that helps forensic experts to calculate a likelihood ratio. In the past two years forensic statisticians developed software called SAILR for this project. During this development some problems occurred that had to be tackled. This thesis is written to investigate some of these problems. Two main problems are proving the equality of likelihood ratio formulas arising from the Gaussian two-level model and investigating various possibilities for parameter estimation within the Gaussian two-level model.

Before these problems were addressed in this thesis, in Chapter 2 an introduction to the likelihood ratio approach in forensic evidence evaluation is given. Chapter 3 describes the underlying discrete- and continuous models in the likelihood ratio approach. The focus in this thesis was mainly on the continuous two-level model. This two-level model is distinguished in a Gaussian two-level model and a non-Gaussian two-level model.

In Chapter 4 the equality of two likelihood ratio formulas arising from the Gaussian two-level models is proven. Since only one of these formulas is implemented in the software, the equality of these formulas is important for the validation of the software and agreement upon likelihood ratios within the ENFSI-LR project.

In Chapter 5 estimation techniques are explored for the parameters $\boldsymbol{\mu}$, $\mathbf{T}$ and $\boldsymbol{\Sigma}$ within the Gaussian two-level model. In the ENFSI-LR project it is decided that the software must contain a "simple" default choice and some optional choices to estimate the parameters. As default choice for the estimators of the covariance matrices $\mathbf{T}$ and $\boldsymbol{\Sigma}$, forensic statisticians have decided that ANOVA estimators will be used. As a default choice for the mean estimator, forensic statisticians are discussing whether the weighted- or the unweighted mean should be used. Therefore, this chapter compares both estimators for the mean. As an alternative to the weighted- and unweighted mean, in this thesis a generalized weighted mean with optimal weights is suggested. A disadvantage of these optimal weights is that they depend on unknown parameters. In case these parameters are known, the generalized weighted mean would be the estimator that has minimum variance among all estimators in it class, which includes the weighted- and unweighted mean.

As optional choice to estimate the parameters, this thesis suggests the EM-algorithm as an iterative method to find the maximum likelihood estimates. Based on the simulation study at the end of Chapter 5, it can be concluded that the maximum likelihood estimators performs the same as or better than the ANOVA estimators. Furthermore, for this simulation we have seen that the stopping criterion is reached within a reasonable amount of iterations. Moreover, if ANOVA estimates are used as starting values this number will be even less. Additionally, the use of the maximum likelihood estimators avoids the choice between the mean estimators. Thus, on the basis of the results in Chapter 5, we would suggest to use the EM-algorithm with the ANOVA estimates as starting values to find the maximum likelihood estimates.

In Chapter 6 the likelihood ratio approach under the assumption of a non-Gaussian two-level model is described. This model is also embedded in the ENFSI-LR software, because of its importance in practice. In this chapter we illuminated some possible difficulties resulting of the use of this method.

In Chapter 7 an extension of the discrete- and continuous models is introduced in this thesis, such that the discrete- and continuous evidence can be combined into one likelihood ratio. We have seen that this combined likelihood ratio results in an intuitive approach, but demands for more data sets that contain a combination of discrete- and continuous features. This extended model can be used in future likelihood ratio calculation, in future development of the software and as a motivation to collect more combined data sets.

Recommendations and future development

- Currently, the ENFSI-LR software does not contain a tool to test the assumption of multivariate normality for the between-source distribution. Including such tools in the software can help forensic experts in the decision between a Gaussian two-level model and a non-Gaussian two-level model. In Section 4.1 some first ideas to assess the assumption of multivariate normality are given. In Section 7.3 these ideas are applied to real xtc data. In future development, these ideas can be further explored and implemented in the software.

- In Chapter 5 estimation techniques for the parameters in the Gaussian two-level model are explored. In this thesis the EM-algorithm is suggested as an alternative iterative method to find the maximum likelihood estimates. In Bolck and Alberink (2011) it is suggested to explore the method of restricted maximum likelihood estimation (REML) or a Bayesian approach, see for example Searle et al. (1992). Due to computational difficulties useful REML estimators are not derived in this thesis. This problem might be solved in further research.

- In Section 6.3 we have seen that in forensic statistics it is common to use the bandwidth matrix $B = b_{\text{opt}}^2 \mathbf{T}$ in the kernel density estimator for h. The advantage of this bandwidth matrix is that it is a simple way of obtaining a full bandwidth matrix. However, in Section 6.2.1 we have seen a simple example where this choice for the bandwidth matrix can be very detrimental for the estimate of h. In such a situation an unconstrained full bandwidth matrix would be preferable. It can be recommended to explore methods to find such a matrix, even though such methods can be hard and research to improve methods is ongoing. An example of an algorithm resulting in a fast and accurate computation for unconstrained bandwidth matrices is given in Duong and Hazelton (2005).

- In Section 6.3 we have mentioned the difficulty of the statistical curse of dimensionality that occurs in kernel density estimation. We have seen that the sample size m (the number of groups in the background data) grows exponentially with dimension p to obtain a required accuracy. Silverman (1986) shows in an example that for 10 dimensions a sample size of 842000 would be needed to obtain a mean squared error less than 0.1. Thus, in higher dimensions the kernel density estimate will not be very accurate. Hence, it is suggested that the estimates should not be reported without confidence bands. However, the uncertainty brought by the kernel density estimate is not reflected in the likelihood ratio. Since in xtc comparison problems often 15 pre-tabletting features are compared (see Section

2.1.1) this dimensionality problem most likely occurs in the likelihood ratio approach. Therefore an interesting follow-up question would be to investigate this uncertainty in the likelihood ratio.

Appendix A

Detailed calculations and proofs

A.1 Proof of the matrix identities (M_1) and (M_2)

Assuming that each of the stated inverses exist, the following statements will be proved:

$$\begin{aligned}(M_1) &: (\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1} = \mathbf{A}-\mathbf{A}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{A} = \mathbf{B}-\mathbf{B}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B}\\ (M_2) &: (\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1} = \mathbf{A}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B} = \mathbf{B}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{A}\end{aligned}$$

Proof of (M_1)

To prove that

$$(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1} = \mathbf{A}-\mathbf{A}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{A},$$

it is sufficient to show that

$$(\mathbf{A}^{-1}+\mathbf{B}^{-1})(\mathbf{A}-\mathbf{A}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{A}) = \mathbf{I}$$

where $\mathbf{I}$ is the identity matrix. Indeed this is true because

$$\begin{aligned}(\mathbf{A}^{-1}+\mathbf{B}^{-1})(\mathbf{A}-\mathbf{A}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{A}) &= (\mathbf{A}^{-1}+\mathbf{B}^{-1})\mathbf{A}-(\mathbf{A}^{-1}+\mathbf{B}^{-1})\mathbf{A}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{A}\\ &= \mathbf{I}+\mathbf{B}^{-1}\mathbf{A}-(\mathbf{A}+\mathbf{B})^{-1}\mathbf{A}-\mathbf{B}^{-1}\mathbf{A}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{A}\\ &= \mathbf{I}-(\mathbf{A}+\mathbf{B})^{-1}\mathbf{A}-\mathbf{B}^{-1}(\mathbf{A}(\mathbf{A}+\mathbf{B})^{-1}-\mathbf{I})\mathbf{A}\\ &= \mathbf{I}-(\mathbf{A}+\mathbf{B})^{-1}\mathbf{A}-\mathbf{B}^{-1}((\mathbf{A}-(\mathbf{A}+\mathbf{B}))(\mathbf{A}+\mathbf{B})^{-1})\mathbf{A}\\ &= \mathbf{I}-(\mathbf{A}+\mathbf{B})^{-1}\mathbf{A}-\mathbf{B}^{-1}(-\mathbf{B}(\mathbf{A}+\mathbf{B})^{-1})\mathbf{A}\\ &= \mathbf{I}-(\mathbf{A}+\mathbf{B})^{-1}\mathbf{A}+(\mathbf{A}+\mathbf{B})^{-1}\mathbf{A}\\ &= \mathbf{I}.\end{aligned}$$

When we interchange the role of $\mathbf{A}$ and $\mathbf{B}$ it can be shown that $(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1} = \mathbf{B}-\mathbf{B}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B}$.

Proof of (M_2)

To prove that

$$(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1} = \mathbf{A}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B},$$

it is sufficient to show that

$$(\mathbf{A}^{-1}+\mathbf{B}^{-1})\mathbf{A}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B} = \mathbf{I}$$

where $\mathbf{I}$ is the identity matrix. Indeed this is true because

$$\begin{aligned}(\mathbf{A}^{-1}+\mathbf{B}^{-1})\mathbf{A}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B} &= (\mathbf{A}+\mathbf{B})^{-1}\mathbf{B}+\mathbf{B}^{-1}\mathbf{A}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B}\\ &= (\mathbf{I}+\mathbf{B}^{-1}\mathbf{A})(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B}\\ &= \mathbf{B}^{-1}(\mathbf{B}+\mathbf{A})(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B}\\ &= \mathbf{B}^{-1}\mathbf{I}\mathbf{B}\\ &= \mathbf{B}^{-1}\mathbf{B}\\ &= \mathbf{I}.\end{aligned}$$

When we interchange the role of $\mathbf{A}$ and $\mathbf{B}$ it can be shown that $(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1} = \mathbf{B}-\mathbf{B}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B}$.

A.2 Proof of the likelihood ratio in equation (4.12)

First the numerator of equation (4.10) will be computed. The computation of the denominator of equation (4.10) is similar to the computation of the numerator. We know that

$$\begin{aligned}\overline{\mathbf{y}}_2 \mid \boldsymbol{\theta} &\sim \mathcal{N}_p(\boldsymbol{\theta}, n_2^{-1}\boldsymbol{\Sigma})\\ \boldsymbol{\theta} \mid \overline{\mathbf{y}}_1 &\sim \mathcal{N}_p(\boldsymbol{\mu}_n, \mathbf{T}_n)\end{aligned}$$

The numerator of equation (4.10) can be written as follows

$$\begin{aligned}\int_{\boldsymbol{\theta}} f(\overline{\mathbf{y}}_2|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\overline{\mathbf{y}}_1)d\boldsymbol{\theta} &= \int_{\boldsymbol{\theta}} |2\pi n_2^{-1}\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\Big(-\frac{1}{2}(\overline{\mathbf{y}}_2-\boldsymbol{\theta})'(n_2^{-1}\boldsymbol{\Sigma})^{-1}(\overline{\mathbf{y}}_2-\boldsymbol{\theta})\Big)|2\pi\mathbf{T}_n|^{-\frac{1}{2}}\\ &\quad\times\exp\Big(-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu}_n)'\mathbf{T}_n^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu}_n)\Big)d\boldsymbol{\theta}\\ &= |2\pi n_2^{-1}\boldsymbol{\Sigma}|^{-\frac{1}{2}}|2\pi\mathbf{T}_n|^{-\frac{1}{2}}\exp\Big(-\frac{1}{2}\Big(\overline{\mathbf{y}}_2'(n_2^{-1}\boldsymbol{\Sigma})^{-1}\overline{\mathbf{y}}_2+\boldsymbol{\mu}_n'\mathbf{T}_n^{-1}\boldsymbol{\mu}_n\Big)\Big)\\ &\quad\times\int_{\boldsymbol{\theta}}\exp\Big(-\frac{1}{2}\Big(\boldsymbol{\theta}'((n_2^{-1}\boldsymbol{\Sigma})^{-1}+\mathbf{T}_n^{-1})\boldsymbol{\theta}-(\overline{\mathbf{y}}_2'(n_2^{-1}\boldsymbol{\Sigma})^{-1}+\boldsymbol{\mu}_n'\mathbf{T}_n^{-1})\boldsymbol{\theta}\\ &\quad-\boldsymbol{\theta}'((n_2^{-1}\boldsymbol{\Sigma})^{-1}\overline{\mathbf{y}}_2+\mathbf{T}_n^{-1}\boldsymbol{\mu}_n)\Big)\Big)d\boldsymbol{\theta}.\end{aligned}$$

Let

$$\begin{aligned}\widetilde{\Sigma}^{-1} &= (n_2^{-1}\boldsymbol{\Sigma})^{-1}+\mathbf{T}_n^{-1},\\ \widetilde{\Sigma}^{-1}\widetilde{\boldsymbol{\mu}} &= (n_2^{-1}\boldsymbol{\Sigma})^{-1}\overline{\mathbf{y}}_2+\mathbf{T}_n^{-1}\boldsymbol{\mu}_n.\end{aligned}$$

Using a multivariate normal distribution for $\boldsymbol{\theta}$ with parameters $\widetilde{\boldsymbol{\mu}}$ and $\widetilde{\Sigma}$, the numerator of equation (4.10) is equal to

$$\begin{aligned}\int_{\boldsymbol{\theta}} f(\overline{\mathbf{y}}_2|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\overline{\mathbf{y}}_1)d\boldsymbol{\theta} &= |2\pi n_2^{-1}\boldsymbol{\Sigma}|^{-\frac{1}{2}}|2\pi\mathbf{T}_n|^{-\frac{1}{2}}\exp\Big(-\frac{1}{2}\Big(\overline{\mathbf{y}}_2'(n_2^{-1}\boldsymbol{\Sigma})^{-1}\overline{\mathbf{y}}_2+\boldsymbol{\mu}_n'\mathbf{T}_n^{-1}\boldsymbol{\mu}_n\Big)\Big)\\ &\quad\times|2\pi((n_2^{-1}\boldsymbol{\Sigma})^{-1}+\mathbf{T}_n^{-1})^{-1}|^{\frac{1}{2}}\\ &\quad\times\exp\Big(\frac{1}{2}\Big(((n_2^{-1}\boldsymbol{\Sigma})^{-1}\overline{\mathbf{y}}_2+\mathbf{T}_n^{-1}\boldsymbol{\mu}_n)'\widetilde{\Sigma}'((n_2^{-1}\boldsymbol{\Sigma})^{-1}\overline{\mathbf{y}}_2+\mathbf{T}_n^{-1}\boldsymbol{\mu}_n)\Big)\Big).\end{aligned}$$

In the latter equation it is used that $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$ for two matrices $\mathbf{A}$ and $\mathbf{B}$. Using properties of the determinant and the definition of $\mathbf{U}_n$ on page 37 it follows that

$$\begin{aligned}
\int_{\boldsymbol{\theta}} f(\overline{\mathbf{y}}_2|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\overline{\mathbf{y}}_1)d\boldsymbol{\theta} &= (2\pi)^{-\frac{p}{2}} \frac{|n_2^{-1}\boldsymbol{\Sigma}\mathbf{U}_n^{-1}\mathbf{T}_n|^{\frac{1}{2}}}{|n_2^{-1}\boldsymbol{\Sigma}\mathbf{T}_n|^{\frac{1}{2}}} \exp\Big(-\frac{1}{2}\Big(\overline{\mathbf{y}}_2'(n_2^{-1}\boldsymbol{\Sigma})^{-1}\overline{\mathbf{y}}_2 + \boldsymbol{\mu}_n'\mathbf{T}_n^{-1}\boldsymbol{\mu}_n\Big)\Big) \\
&\quad \times \exp\Big(\frac{1}{2}\Big((\overline{\mathbf{y}}_2'((n_2^{-1}\boldsymbol{\Sigma})^{-1})' + \boldsymbol{\mu}_n'(\mathbf{T}_n^{-1})')\widetilde{\Sigma}((n_2^{-1}\boldsymbol{\Sigma})^{-1}\overline{\mathbf{y}}_2 + \mathbf{T}_n^{-1}\boldsymbol{\mu}_n)\Big)\Big) \\
&= (2\pi)^{-\frac{p}{2}}|\mathbf{U}_n^{-1}|^{\frac{1}{2}}\exp\Big(-\frac{1}{2}\Big(\overline{\mathbf{y}}_2'(n_2^{-1}\boldsymbol{\Sigma})^{-1}\overline{\mathbf{y}}_2 + \boldsymbol{\mu}_n'\mathbf{T}_n^{-1}\boldsymbol{\mu}_n\Big)\Big) \\
&\quad \times \exp\Big(\frac{1}{2}\Big(\overline{\mathbf{y}}_2'(n_2^{-1}\boldsymbol{\Sigma})^{-1}\widetilde{\Sigma}(n_2^{-1}\boldsymbol{\Sigma})^{-1}\overline{\mathbf{y}}_2 + \boldsymbol{\mu}_n'\mathbf{T}_n^{-1}\widetilde{\Sigma}(n_2^{-1}\boldsymbol{\Sigma})^{-1}\overline{\mathbf{y}}_2 \\
&\quad + \overline{\mathbf{y}}_2'(n_2^{-1}\boldsymbol{\Sigma})^{-1}\widetilde{\Sigma}\mathbf{T}_n^{-1}\boldsymbol{\mu}_n + \boldsymbol{\mu}_n'\mathbf{T}_n^{-1}\widetilde{\Sigma}\mathbf{T}_n^{-1}\boldsymbol{\mu}_n\Big)\Big) \\
&= (2\pi)^{-\frac{p}{2}}|\mathbf{U}_n|^{-\frac{1}{2}}\exp\Big(-\frac{1}{2}\Big(\overline{\mathbf{y}}_2'((n_2^{-1}\boldsymbol{\Sigma})^{-1} - (n_2^{-1}\boldsymbol{\Sigma})^{-1}\widetilde{\Sigma}(n_2^{-1}\boldsymbol{\Sigma})^{-1})\overline{\mathbf{y}}_2 \\
&\quad + \boldsymbol{\mu}_n'\mathbf{T}_n^{-1}\widetilde{\Sigma}(n_2^{-1}\boldsymbol{\Sigma})^{-1}\overline{\mathbf{y}}_2 + \overline{\mathbf{y}}_2'(n_2^{-1}\boldsymbol{\Sigma})^{-1}\widetilde{\Sigma}\mathbf{T}_n^{-1}\boldsymbol{\mu}_n + \boldsymbol{\mu}_n'(\mathbf{T}_n^{-1} - \mathbf{T}_n^{-1}\widetilde{\Sigma}\mathbf{T}_n^{-1})\boldsymbol{\mu}_n\Big)\Big).
\end{aligned}$$

The first two equations are true since covariance matrices are symmetric. Because $\widetilde{\Sigma}$ consists of covariance matrices it is symmetric too. Finally, the matrix identities (M_1) and (M_2) will be used:

$$\begin{aligned}
\int_{\boldsymbol{\theta}} f(\overline{\mathbf{y}}_2|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\overline{\mathbf{y}}_1)d\boldsymbol{\theta} &= (2\pi)^{-\frac{p}{2}}|\mathbf{U}_n|^{-\frac{1}{2}}\exp\Big(-\frac{1}{2}\Big(\overline{\mathbf{y}}_2'(n_2^{-1}\boldsymbol{\Sigma} + \mathbf{T}_n)^{-1}\overline{\mathbf{y}}_2 + \boldsymbol{\mu}_n'(n_2^{-1}\boldsymbol{\Sigma} + \mathbf{T}_n)^{-1}\overline{\mathbf{y}}_2 \\
&\quad + \overline{\mathbf{y}}_2'(n_2^{-1}\boldsymbol{\Sigma} + \mathbf{T}_n)^{-1}\boldsymbol{\mu}_n + \boldsymbol{\mu}_n'(n_2^{-1}\boldsymbol{\Sigma} + \mathbf{T}_n)^{-1}\boldsymbol{\mu}_n\Big)\Big) \\
&= (2\pi)^{-\frac{p}{2}}|\mathbf{U}_n|^{-\frac{1}{2}}\exp\Big(-\frac{1}{2}(\overline{\mathbf{y}}_2 - \boldsymbol{\mu}_n)'\mathbf{U}_n^{-1}(\overline{\mathbf{y}}_2 - \boldsymbol{\mu}_n)\Big).
\end{aligned}$$

Since $\overline{\mathbf{Y}}_2 \sim \mathcal{N}_p(\boldsymbol{\mu}, n_2^{-1}\boldsymbol{\Sigma} + \mathbf{T})$, the denominator is equal to

$$\int_{\boldsymbol{\theta}} f(\overline{\mathbf{y}}_2|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta} = (2\pi)^{-\frac{p}{2}}|\mathbf{U}_0|^{-\frac{1}{2}}\exp\Big(-\frac{1}{2}(\overline{\mathbf{y}}_2 - \boldsymbol{\mu})'\mathbf{U}_0^{-1}(\overline{\mathbf{y}}_2 - \boldsymbol{\mu})\Big).$$

Dividing the numerator by this denominator gives the likelihood ratio of expression (4.12).

A.3 Proof of Lemma 4.2.3

Consider the following part of equation (4.9):

$$\frac{1}{|n_2^{-1}\boldsymbol{\Sigma} + \mathbf{T}|^{-\frac{1}{2}}}\exp\left\{\frac{1}{2}(\overline{\mathbf{y}}_2 - \boldsymbol{\mu})'(n_2^{-1}\boldsymbol{\Sigma} + \mathbf{T})^{-1}(\overline{\mathbf{y}}_2 - \boldsymbol{\mu})\right\}.$$

It is immediately clear that this term equals the following part of equation (4.12):

$$|\mathbf{U}_0|^{\frac{1}{2}}\exp\left\{\frac{1}{2}(\overline{\mathbf{y}}_2 - \boldsymbol{\mu})'\mathbf{U}_0^{-1}(\overline{\mathbf{y}}_2 - \boldsymbol{\mu})\right\}.$$

Thus, in order to show the equality of the likelihood ratio formulas it is sufficient to show that equation (A.1)

$$\begin{aligned}
&\frac{\left|(n_1^{-1}+n_2^{-1})\boldsymbol{\Sigma}\right|^{-\frac{1}{2}}\left|\left(\mathbf{T}+\frac{\boldsymbol{\Sigma}}{n_1+n_2}\right)\right|^{-\frac{1}{2}}}{\left|\mathbf{T}+n_1^{-1}\boldsymbol{\Sigma}\right|^{-\frac{1}{2}}}\exp\left\{\frac{1}{2}(\overline{\mathbf{y}}_1-\boldsymbol{\mu})'(\mathbf{T}+n_1^{-1}\boldsymbol{\Sigma})^{-1}(\overline{\mathbf{y}}_1-\boldsymbol{\mu})\right\}\\
\times\quad &\exp\left\{-\frac{1}{2}(\overline{\mathbf{y}}_1-\overline{\mathbf{y}}_2)'\Big((n_1^{-1}+n_2^{-1})\boldsymbol{\Sigma}\Big)^{-1}(\overline{\mathbf{y}}_1-\overline{\mathbf{y}}_2)\right\}\\
\times\quad &\exp\left\{-\frac{1}{2}\Big(\frac{n_1\overline{\mathbf{y}}_1+n_2\overline{\mathbf{y}}_2}{n_1+n_2}-\boldsymbol{\mu}\Big)'\Big(\mathbf{T}+\frac{\boldsymbol{\Sigma}}{n_1+n_2}\Big)^{-1}\Big(\frac{n_1\overline{\mathbf{y}}_1+n_2\overline{\mathbf{y}}_2}{n_1+n_2}-\boldsymbol{\mu}\Big)\right\} \qquad \text{(A.1)}
\end{aligned}$$

and equation (A.2)

$$|\mathbf{U}_n|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\overline{\mathbf{y}}_2-\boldsymbol{\mu}_n)'\mathbf{U}_n^{-1}(\overline{\mathbf{y}}_2-\boldsymbol{\mu}_n)\right\} \qquad \text{(A.2)}$$

are the same.

Equality of equation (A.1) **and equation** (A.2)

In order to show that equation (A.1) coincides with equation (A.2), the following two statements for square matrices $\mathbf{A}$ and $\mathbf{B}$ will be used frequently:

$$\begin{aligned}
(\mathrm{M}_1)\quad &:\quad (\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}=\mathbf{A}-\mathbf{A}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{A}=\mathbf{B}-\mathbf{B}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B}\\
(\mathrm{M}_2)\quad &:\quad (\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}=\mathbf{A}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B}=\mathbf{B}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{A}
\end{aligned}$$

A proof of these identities is given in Appendix A.1. All of the stated inverses are assumed to exist for covariances matrices $\boldsymbol{\Sigma}$ and $\mathbf{T}$. The equality of equation (A.1) and equation (A.2) will be shown in several steps.

Step 1

The first step is to show the following equality:

$$\left|\mathbf{U}_n\right|^{-\frac{1}{2}}=\frac{\left|n_1^{-1}\boldsymbol{\Sigma}+n_2^{-1}\boldsymbol{\Sigma}\right|^{-\frac{1}{2}}\left|(n_1+n_2)^{-1}\boldsymbol{\Sigma}+\mathbf{T}\right|^{-\frac{1}{2}}}{\left|n_1^{-1}\boldsymbol{\Sigma}+\mathbf{T}\right|^{-\frac{1}{2}}}. \qquad \text{(A.3)}$$

We will focus on the right hand side of equation (A.3). First, use that $|\mathbf{A}|^{-1}=|\mathbf{A}^{-1}|$ and $|\mathbf{AB}|=|\mathbf{BA}|$, for matrices $\mathbf{A}$ and $\mathbf{B}$. Then, the right hand side of equation (A.3) is equal to

$$\left|\left(\frac{\boldsymbol{\Sigma}}{n_1}+\frac{\boldsymbol{\Sigma}}{n_2}\right)\left(\frac{\boldsymbol{\Sigma}}{n_1}+\mathbf{T}\right)^{-1}\left(\frac{\boldsymbol{\Sigma}}{n_1+n_2}+\mathbf{T}\right)\right|^{-\frac{1}{2}}.$$

Completing this product, it can be seen that

$$\frac{\boldsymbol{\Sigma}}{n_1}\left(\frac{\boldsymbol{\Sigma}}{n_1}+\mathbf{T}\right)^{-1}\mathbf{T}=(n_1\boldsymbol{\Sigma}^{-1}+\mathbf{T}^{-1})^{-1}=\mathbf{T}_n,$$

where in the first and second equality the statements (M_2) and (M_1) respectively are applied. The remaining terms should thus equal $n_2^{-1}\boldsymbol{\Sigma}$. For that reason, write the right hand side of equation (A.3) as

$$\left|\mathbf{T}_n+\frac{\boldsymbol{\Sigma}}{n_2}\left(\frac{\boldsymbol{\Sigma}}{n_1}+\mathbf{T}\right)^{-1}\left[\frac{n_2}{n_1}\frac{\boldsymbol{\Sigma}}{n_1+n_2}+\frac{\boldsymbol{\Sigma}}{n_1+n_2}+\mathbf{T}\right]\right|^{-\frac{1}{2}}.$$

And since the terms in the square brackets can be simplified to $\left(\frac{\boldsymbol{\Sigma}}{n_1}+\mathbf{T}\right)$, equation (A.3) is indeed true.

Step 2

The next step in showing that the equations (A.1) and (A.2) are equal, is to prove that the exponential terms in these equations are the same.This means that the following four equations have to be shown:

$$\overline{\mathbf{y}}_2'\mathbf{U}_n^{-1}\overline{\mathbf{y}}_2 = \overline{\mathbf{y}}_2'\left(\frac{\boldsymbol{\Sigma}}{n_1}+\frac{\boldsymbol{\Sigma}}{n_2}\right)^{-1}\overline{\mathbf{y}}_2 + \frac{n_2\overline{\mathbf{y}}_2'}{n_1+n_2}\left(\frac{\boldsymbol{\Sigma}}{n_1+n_2}+\mathbf{T}\right)^{-1}\frac{n_2\overline{\mathbf{y}}_2}{n_1+n_2} \quad \text{(A.4)}$$

$$\begin{aligned}-\overline{\mathbf{y}}_2'\mathbf{U}_n^{-1}\boldsymbol{\mu}_n &= -\overline{\mathbf{y}}_2'\left(\frac{\boldsymbol{\Sigma}}{n_1}+\frac{\boldsymbol{\Sigma}}{n_2}\right)^{-1}\overline{\mathbf{y}}_1 + \frac{n_2\overline{\mathbf{y}}_2'}{n_1+n_2}\left(\frac{\boldsymbol{\Sigma}}{n_1+n_2}+\mathbf{T}\right)^{-1}\\ &\times \left(\frac{n_1\overline{\mathbf{y}}_1}{n_1+n_2}-\boldsymbol{\mu}\right) \quad \text{(A.5)}\end{aligned}$$

$$\begin{aligned}-\boldsymbol{\mu}_n'\mathbf{U}_n^{-1}\overline{\mathbf{y}}_2 &= -\overline{\mathbf{y}}_1'\left(\frac{\boldsymbol{\Sigma}}{n_1}+\frac{\boldsymbol{\Sigma}}{n_2}\right)^{-1}\overline{\mathbf{y}}_2 + \left(\frac{n_1\overline{\mathbf{y}}_1}{n_1+n_2}-\boldsymbol{\mu}\right)'\left(\frac{\boldsymbol{\Sigma}}{n_1+n_2}+\mathbf{T}\right)^{-1}\\ &\times \frac{n_2\overline{\mathbf{y}}_2}{n_1+n_2} \quad \text{(A.6)}\end{aligned}$$

$$\begin{aligned}\boldsymbol{\mu}_n'\mathbf{U}_n^{-1}\boldsymbol{\mu}_n &= \overline{\mathbf{y}}_1'\left(\frac{\boldsymbol{\Sigma}}{n_1}+\frac{\boldsymbol{\Sigma}}{n_2}\right)^{-1}\overline{\mathbf{y}}_1 + \left(\frac{n_1\overline{\mathbf{y}}_1}{n_1+n_2}-\boldsymbol{\mu}\right)'\left(\frac{\boldsymbol{\Sigma}}{n_1+n_2}+\mathbf{T}\right)^{-1}\\ &\times \left(\frac{n_1\overline{\mathbf{y}}_1}{n_1+n_2}-\boldsymbol{\mu}\right) - (\overline{\mathbf{y}}_1-\boldsymbol{\mu})'\left(\frac{\boldsymbol{\Sigma}}{n_1}+\mathbf{T}\right)^{-1}(\overline{\mathbf{y}}_1-\boldsymbol{\mu}) \quad \text{(A.7)}\end{aligned}$$

Step 2.1

To prove equation (A.4), it is sufficient to prove that

$$\left((\mathbf{T}^{-1}+n_1\boldsymbol{\Sigma}^{-1})^{-1}+\frac{\boldsymbol{\Sigma}}{n_2}\right)^{-1} = \left(\frac{\boldsymbol{\Sigma}}{n_1}+\frac{\boldsymbol{\Sigma}}{n_2}\right)^{-1} + \left(\frac{n_2}{n_1+n_2}\right)^2\left(\frac{\boldsymbol{\Sigma}}{n_1+n_2}+\mathbf{T}\right)^{-1},$$

where the definition of $\mathbf{U}_n^{-1}$ is used in the left hand side of the equation. First, apply statement (M_1) on the left hand side of this equation

$$\left((\mathbf{T}^{-1}+n_1\boldsymbol{\Sigma}^{-1})^{-1}+\frac{\boldsymbol{\Sigma}}{n_2}\right)^{-1} = n_2\boldsymbol{\Sigma}^{-1} - n_2\boldsymbol{\Sigma}^{-1}\Big(\mathbf{T}^{-1}+(n_1+n_2)\boldsymbol{\Sigma}^{-1}\Big)^{-1}n_2\boldsymbol{\Sigma}^{-1}.$$

If (M_1) is applied on the right hand side of the latter equation, then

$$\left((\mathbf{T}^{-1}+n_1\boldsymbol{\Sigma}^{-1})^{-1}+\frac{\boldsymbol{\Sigma}}{n_2}\right)^{-1} = n_2\boldsymbol{\Sigma}^{-1} - n_2\boldsymbol{\Sigma}^{-1}\left[\frac{\boldsymbol{\Sigma}}{n_1+n_2} - \frac{\boldsymbol{\Sigma}}{n_1+n_2}\left(\frac{\boldsymbol{\Sigma}}{n_1+n_2}+\mathbf{T}\right)^{-1}\frac{\boldsymbol{\Sigma}}{n_1+n_2}\right]n_2\boldsymbol{\Sigma}^{-1}.$$

Completing the product gives

$$\left((\mathbf{T}^{-1}+n_1\boldsymbol{\Sigma}^{-1})^{-1}+\frac{\boldsymbol{\Sigma}}{n_2}\right)^{-1} = n_2\boldsymbol{\Sigma}^{-1} - \frac{n_2^2}{n_1+n_2}\boldsymbol{\Sigma}^{-1} + \left(\frac{n_2}{n_1+n_2}\right)^2\left(\frac{\boldsymbol{\Sigma}}{n_1+n_2}+\mathbf{T}\right)^{-1}.$$

Now notice that

$$n_2\boldsymbol{\Sigma}^{-1} - \frac{n_2^2}{n_1+n_2}\boldsymbol{\Sigma}^{-1} = \left(\frac{n_1+n_2}{n_1n_2}\boldsymbol{\Sigma}\right)^{-1} = \left(\frac{\boldsymbol{\Sigma}}{n_1}+\frac{\boldsymbol{\Sigma}}{n_2}\right)^{-1}$$

and thus equation (A.4) is indeed true.

Step 2.2

To prove equation (A.5), it has to be shown that

$$\mathbf{U}_n^{-1}\boldsymbol{\mu}_n = \left(\frac{\boldsymbol{\Sigma}}{n_1}+\frac{\boldsymbol{\Sigma}}{n_2}\right)^{-1}\overline{\mathbf{y}}_1 - \frac{n_2}{n_1+n_2}\left(\frac{\boldsymbol{\Sigma}}{n_1+n_2}+\mathbf{T}\right)^{-1}\left(\frac{n_1\overline{\mathbf{y}}_1}{n_1+n_2}-\boldsymbol{\mu}\right),$$

with

$$\begin{aligned}\mathbf{U}_n^{-1} &= \left(\frac{\boldsymbol{\Sigma}}{n_1}+\frac{\boldsymbol{\Sigma}}{n_2}\right)^{-1} + \left(\frac{n_2}{n_1+n_2}\right)^2\left(\frac{\boldsymbol{\Sigma}}{n_1+n_2}+\mathbf{T}\right)^{-1} && \text{(step 2.1)}\\ \boldsymbol{\mu}_n &= \mathbf{T}\left(\mathbf{T}+\frac{\boldsymbol{\Sigma}}{n_1}\right)^{-1}\overline{\mathbf{y}}_1 + \frac{\boldsymbol{\Sigma}}{n_1}\left(\mathbf{T}+\frac{\boldsymbol{\Sigma}}{n_1}\right)^{-1}\boldsymbol{\mu}. && \text{(by definition)}\end{aligned}$$

Thus, the following two equations have to be true

$$\begin{aligned}\mathbf{U}_n^{-1}\mathbf{T}(\mathbf{T}+\frac{\boldsymbol{\Sigma}}{n_1})^{-1}\overline{\mathbf{y}}_1 &= (\frac{\boldsymbol{\Sigma}}{n_1}+\frac{\boldsymbol{\Sigma}}{n_2})^{-1}\overline{\mathbf{y}}_1 - \frac{n_2}{n_1+n_2}(\frac{\boldsymbol{\Sigma}}{n_1+n_2}+\mathbf{T})^{-1}\\ &\quad\times\ \frac{n_1\overline{\mathbf{y}}_1}{n_1+n_2}. && \text{(A.8)}\\ \mathbf{U}_n^{-1}\frac{\boldsymbol{\Sigma}}{n_1}(\mathbf{T}+\frac{\boldsymbol{\Sigma}}{n_1})^{-1}\boldsymbol{\mu} &= \frac{n_2}{n_1+n_2}(\frac{\boldsymbol{\Sigma}}{n_1+n_2}+\mathbf{T})^{-1}\boldsymbol{\mu}. && \text{(A.9)}\end{aligned}$$

To prove equation (A.8), first apply (M_1) on the left hand side

$$\mathbf{U}_n^{-1}\mathbf{T}\left(\mathbf{T}+\frac{\boldsymbol{\Sigma}}{n_1}\right)^{-1}\overline{\mathbf{y}}_1 = \mathbf{U}_n^{-1}\overline{\mathbf{y}}_1 - \mathbf{U}_n^{-1}(\mathbf{T}^{-1}+n_1\boldsymbol{\Sigma}^{-1})^{-1}\mathbf{T}^{-1}\overline{\mathbf{y}}_1.$$

Thus, it has to be shown that

$$\mathbf{U}_n^{-1}-\mathbf{U}_n^{-1}(\mathbf{T}^{-1}+n_1\boldsymbol{\Sigma}^{-1})^{-1}\mathbf{T}^{-1} = \left(\frac{\boldsymbol{\Sigma}}{n_1}+\frac{\boldsymbol{\Sigma}}{n_2}\right)^{-1} - \frac{n_2}{n_1+n_2}\left(\frac{\boldsymbol{\Sigma}}{n_1+n_2}+\mathbf{T}\right)^{-1}\frac{n_1}{n_1+n_2}.$$

This is the same as showing that the following equation is true

$$\begin{aligned}\left(\frac{n_2}{n_1+n_2}\right)^2\left(\frac{\boldsymbol{\Sigma}}{n_1+n_2}+\mathbf{T}\right)^{-1} - \mathbf{U}_n^{-1}\left(\mathbf{T}^{-1}+n_1\boldsymbol{\Sigma}^{-1}\right)^{-1}\mathbf{T}^{-1} &= -\frac{n_2}{n_1+n_2}\left(\frac{\boldsymbol{\Sigma}}{n_1+n_2}+\mathbf{T}\right)^{-1}\\ &\quad\times\ \frac{n_1}{n_1+n_2} && \text{(A.10)}\end{aligned}$$

To confirm the latter equation, first apply (M_2) on the second term in the left hand side

$$-\mathbf{U}_n^{-1}(\mathbf{T}^{-1}+n_1\boldsymbol{\Sigma}^{-1})^{-1}\mathbf{T}^{-1} = -\mathbf{U}_n^{-1}\frac{\boldsymbol{\Sigma}}{n_1}\left(\mathbf{T}+\frac{\boldsymbol{\Sigma}}{n_1}\right)^{-1}.$$

Using the expression for $\mathbf{U}_n^{-1}$, the left hand side of equation (A.10) can then be written as

$$\left(\frac{n_2}{n_1+n_2}\right)^2\left(\frac{\boldsymbol{\Sigma}}{n_1+n_2}+\mathbf{T}\right)^{-1} - \frac{n_2}{n_1+n_2}\left(\mathbf{T}+\frac{\boldsymbol{\Sigma}}{n_1}\right)^{-1} - \left(\frac{n_2}{n_1+n_2}\right)^2\left(\frac{\boldsymbol{\Sigma}}{n_1+n_2}+\mathbf{T}\right)^{-1}\frac{\boldsymbol{\Sigma}}{n_1}\left(\mathbf{T}+\frac{\boldsymbol{\Sigma}}{n_1}\right)^{-1}.$$

The latter expression can be written in the same form as the right hand side of equation (A.10),

$$-\frac{n_2}{n_1+n_2}\left(\frac{\boldsymbol{\Sigma}}{n_1+n_2}+\mathbf{T}\right)^{-1}\left[-\frac{n_2}{n_1}+\frac{n_1+n_2}{n_1}\left(\frac{\boldsymbol{\Sigma}}{n_1+n_2}+\mathbf{T}\right)\left(\mathbf{T}+\frac{\boldsymbol{\Sigma}}{n_1}\right)^{-1} - \frac{n_2}{n_1^2}\boldsymbol{\Sigma}\left(\mathbf{T}+\frac{\boldsymbol{\Sigma}}{n_1}\right)^{-1}\right]\frac{n_1}{n_1+n_2}.$$

Now notice that

$$-\frac{n_2}{n_1}+\frac{n_1+n_2}{n_1}\left(\frac{\boldsymbol{\Sigma}}{n_1+n_2}+\mathbf{T}\right)\left(\mathbf{T}+\frac{\boldsymbol{\Sigma}}{n_1}\right)^{-1}-\frac{n_2}{n_1^2}\boldsymbol{\Sigma}\left(\mathbf{T}+\frac{\boldsymbol{\Sigma}}{n_1}\right)^{-1}=\mathbf{I}.$$

Hence, equation (A.10) is true en thus equation (A.8) is true. To prove equation (A.9), the left hand side can be written as

$$\mathbf{U}_n^{-1}\frac{\boldsymbol{\Sigma}}{n_1}\left(\mathbf{T}+\frac{\boldsymbol{\Sigma}}{n_1}\right)^{-1}\boldsymbol{\mu}=\frac{n_2}{n_1+n_2}\left(\mathbf{T}+\frac{\boldsymbol{\Sigma}}{n_1}\right)^{-1}\boldsymbol{\mu}+\left(\frac{n_2}{n_1+n_2}\right)^2\left(\mathbf{T}+\frac{\boldsymbol{\Sigma}}{n_1+n_2}\right)^{-1}\frac{\boldsymbol{\Sigma}}{n_1}\left(\mathbf{T}+\frac{\boldsymbol{\Sigma}}{n_1}\right)^{-1}\boldsymbol{\mu}.$$

The right hand side of the latter equation can be written in the same form as the right hand side of equation (A.9),

$$\begin{aligned}\mathbf{U}_n^{-1}\frac{\boldsymbol{\Sigma}}{n_1}\left(\mathbf{T}+\frac{\boldsymbol{\Sigma}}{n_1}\right)^{-1}\boldsymbol{\mu} &= \frac{n_2}{n_1+n_2}\left(\frac{\boldsymbol{\Sigma}}{n_1+n_2}+\mathbf{T}\right)^{-1}\left[\left(\frac{\boldsymbol{\Sigma}}{n_1+n_2}+\mathbf{T}\right)\left(\frac{\boldsymbol{\Sigma}}{n_1}+\mathbf{T}\right)^{-1}\right.\\ &+ \left.\frac{n_2}{n_1+n_2}\frac{\boldsymbol{\Sigma}}{n_1}\left(\mathbf{T}+\frac{\boldsymbol{\Sigma}}{n_1}\right)^{-1}\right]\boldsymbol{\mu}.\end{aligned}$$

Now it follows that

$$\begin{aligned}\left(\frac{\boldsymbol{\Sigma}}{n_1+n_2}+\mathbf{T}\right)\left(\frac{\boldsymbol{\Sigma}}{n_1}+\mathbf{T}\right)^{-1}+\frac{n_2}{n_1+n_2}\frac{\boldsymbol{\Sigma}}{n_1}\left(\mathbf{T}+\frac{\boldsymbol{\Sigma}}{n_1}\right)^{-1} &= \left[\left(\frac{\boldsymbol{\Sigma}}{n_1+n_2}+\mathbf{T}\right)+\frac{n_2}{n_1+n_2}\frac{\boldsymbol{\Sigma}}{n_1}\right]\\ &\times \left(\frac{\boldsymbol{\Sigma}}{n_1}+\mathbf{T}\right)^{-1}\\ &= \mathbf{I}.\end{aligned}$$

and hence equation (A.9) is indeed true.

Step 2.3

The proof of equation (A.6) is similar to step 2.2.

Step 2.4

To prove equation (A.7), equation (A.6) will be used. From equation (A.6), we know that

$$\boldsymbol{\mu}_n'\mathbf{U}_n^{-1}=\overline{\mathbf{y}}_1'\left[\left(\frac{\boldsymbol{\Sigma}}{n_1}+\frac{\boldsymbol{\Sigma}}{n_2}\right)^{-1}-\frac{n_1n_2}{(n_1+n_2)^2}\left(\frac{\boldsymbol{\Sigma}}{n_1+n_2}+\mathbf{T}\right)^{-1}\right]+\boldsymbol{\mu}'\left(\frac{\boldsymbol{\Sigma}}{n_1+n_2}+\mathbf{T}\right)^{-1}\frac{n_2}{n_1+n_2}.$$

If the latter equation is multiplied on both sides with $\boldsymbol{\mu}_n$, it follows that the following four equations should be true

1. $\overline{\mathbf{y}}_1'[(\frac{\boldsymbol{\Sigma}}{n_1}+\frac{\boldsymbol{\Sigma}}{n_2})^{-1}-\frac{n_1n_2}{(n_1+n_2)^2}(\frac{\boldsymbol{\Sigma}}{n_1+n_2}+\mathbf{T})^{-1}]\mathbf{T}(\mathbf{T}+\frac{\boldsymbol{\Sigma}}{n_1})^{-1}\overline{\mathbf{y}}_1=\overline{\mathbf{y}}_1'(\frac{\boldsymbol{\Sigma}}{n_1}+\frac{\boldsymbol{\Sigma}}{n_2})^{-1}\overline{\mathbf{y}}_1+\frac{n_1\overline{\mathbf{y}}_1'}{n_1+n_2}(\frac{\boldsymbol{\Sigma}}{n_1+n_2}+\mathbf{T})^{-1}\frac{n_1\overline{\mathbf{y}}_1}{n_1+n_2}-\overline{\mathbf{y}}_1'(\frac{\boldsymbol{\Sigma}}{n_1}+\mathbf{T})^{-1}\overline{\mathbf{y}}_1.$
2. $\overline{\mathbf{y}}_1'[(\frac{\boldsymbol{\Sigma}}{n_1}+\frac{\boldsymbol{\Sigma}}{n_2})^{-1}-\frac{n_1n_2}{(n_1+n_2)^2}(\frac{\boldsymbol{\Sigma}}{n_1+n_2}+\mathbf{T})^{-1}]\frac{\boldsymbol{\Sigma}}{n_1}(\mathbf{T}+\frac{\boldsymbol{\Sigma}}{n_1})^{-1}\boldsymbol{\mu}=-\frac{n_1\overline{\mathbf{y}}_1'}{n_1+n_2}(\frac{\boldsymbol{\Sigma}}{n_1+n_2}+\mathbf{T})^{-1}\boldsymbol{\mu}+\overline{\mathbf{y}}_1'(\frac{\boldsymbol{\Sigma}}{n_1}+\mathbf{T})^{-1}\boldsymbol{\mu}.$
3. $\boldsymbol{\mu}'(\frac{\boldsymbol{\Sigma}}{n_1+n_2}+\mathbf{T})^{-1}\frac{n_2}{n_1+n_2}\mathbf{T}(\mathbf{T}+\frac{\boldsymbol{\Sigma}}{n_1})^{-1}\overline{\mathbf{y}}_1=-\boldsymbol{\mu}'(\frac{\boldsymbol{\Sigma}}{n_1+n_2}+\mathbf{T})^{-1}\frac{n_1\overline{\mathbf{y}}_1}{n_1+n_2}+\boldsymbol{\mu}'(\frac{\boldsymbol{\Sigma}}{n_1}+\mathbf{T})^{-1}\overline{\mathbf{y}}_1$
4. $\boldsymbol{\mu}'(\frac{\boldsymbol{\Sigma}}{n_1+n_2}+\mathbf{T})^{-1}\frac{n_2}{n_1+n_2}\frac{\boldsymbol{\Sigma}}{n_1}(\mathbf{T}+\frac{\boldsymbol{\Sigma}}{n_1})^{-1}\boldsymbol{\mu}=-\boldsymbol{\mu}'(\frac{\boldsymbol{\Sigma}}{n_1+n_2}+\mathbf{T})^{-1}\boldsymbol{\mu}-\boldsymbol{\mu}'(\frac{\boldsymbol{\Sigma}}{n_1}+\mathbf{T})^{-1}\boldsymbol{\mu}$

To prove equation (1) (ignoring the terms $\overline{\mathbf{y}}_1$), the left-hand side can be written as

$$\left(\frac{\mathbf{\Sigma}}{n_1+n_2}+\mathbf{T}\right)^{-1}\left[\left(\frac{\mathbf{\Sigma}}{n_1+n_2}+\mathbf{T}\right)\left(\frac{\mathbf{\Sigma}}{n_1}+\frac{\mathbf{\Sigma}}{n_2}\right)^{-1}\mathbf{T}-\frac{n_1n_2}{(n_1+n_2)^2}\mathbf{T}\right]\left(\mathbf{T}+\frac{\mathbf{\Sigma}}{n_1}\right)^{-1}$$

and thus the left-hand side is equal to

$$\left(\frac{\mathbf{\Sigma}}{n_1+n_2}+\mathbf{T}\right)^{-1}\left[\frac{n_1n_2}{n_1+n_2}\mathbf{T}\mathbf{\Sigma}^{-1}\mathbf{T}\right]\left(\mathbf{T}+\frac{\mathbf{\Sigma}}{n_1}\right)^{-1}.$$

The right-hand side of equation (1) can be written in the same form as the left hand side

$$\begin{aligned}&\left(\frac{\mathbf{\Sigma}}{n_1+n_2}+\mathbf{T}\right)^{-1}\left[\left(\frac{\mathbf{\Sigma}}{n_1+n_2}+\mathbf{T}\right)\left(\frac{n_1n_2}{n_1+n_2}\mathbf{\Sigma}^{-1}\right)\left(\frac{\mathbf{\Sigma}}{n_1}+\mathbf{T}\right)\right.\\ +\quad&\left.\frac{n_1^2}{(n_1+n_2)^2}\left(\frac{\mathbf{\Sigma}}{n_1}+\mathbf{T}\right)-\left(\frac{\mathbf{\Sigma}}{n_1+n_2}+\mathbf{T}\right)\right]\left(\mathbf{T}+\frac{\mathbf{\Sigma}}{n_1}\right)^{-1}.\end{aligned}$$

By simplifying the products in the square brackets it follows that this is indeed the same as the left-hand side. To prove equation (2) a similar trick is used. The left-hand side of equation (2) can be written as (ignoring the terms $\overline{\mathbf{y}}_1$ and $\boldsymbol{\mu}$)

$$\left(\frac{\mathbf{\Sigma}}{n_1+n_2}+\mathbf{T}\right)^{-1}\left[\frac{n_2}{n_1+n_2}\left(\frac{\mathbf{\Sigma}}{n_1+n_2}+\mathbf{T}\right)-\frac{n_1n_2}{(n_1+n_2)^2}\frac{\mathbf{\Sigma}}{n_1}\right]\left(\mathbf{T}+\frac{\mathbf{\Sigma}}{n_1}\right)^{-1}.$$

This can be simplified to

$$\left(\frac{\mathbf{\Sigma}}{n_1+n_2}+\mathbf{T}\right)^{-1}\frac{n_2}{n_1+n_2}\mathbf{T}\left(\mathbf{T}+\frac{\mathbf{\Sigma}}{n_1}\right)^{-1}.$$

The right-hand side of equation (2) can be written as

$$\left(\frac{\mathbf{\Sigma}}{n_1+n_2}+\mathbf{T}\right)^{-1}\left[-\frac{n_1}{n_1+n_2}\left(\frac{\mathbf{\Sigma}}{n_1}+\mathbf{T}\right)+\left(\frac{\mathbf{\Sigma}}{n_1+n_2}+\mathbf{T}\right)\right]\left(\mathbf{T}+\frac{\mathbf{\Sigma}}{n_1}\right)^{-1}$$

and thus this is equal to the left-hand side of equation (2). For equation (3), write the right-hand side as

$$\left(\frac{\mathbf{\Sigma}}{n_1+n_2}+\mathbf{T}\right)^{-1}\frac{n_1}{n_1+n_2}\left[\left(\frac{\mathbf{\Sigma}}{n_1+n_2}+\mathbf{T}\right)\frac{n_1+n_2}{n_1}-\left(\mathbf{T}+\frac{\mathbf{\Sigma}}{n_1}\right)\right]\left(\mathbf{T}+\frac{\mathbf{\Sigma}}{n_1}\right)^{-1}.$$

If the product in the square brackets is simplified it follows that equation (3) is indeed true. To prove that equation (4), notice that the right-hand side can be written as

$$\left(\frac{\mathbf{\Sigma}}{n_1+n_2}+\mathbf{T}\right)^{-1}\left[\left(\frac{\mathbf{\Sigma}}{n_1}+\mathbf{T}\right)-\left(\frac{\mathbf{\Sigma}}{n_1+n_2}+\mathbf{T}\right)\right]\left(\frac{\mathbf{\Sigma}}{n_1}+\mathbf{T}\right)^{-1}.$$

A.4 Proof of the identity in equation (5.18)

The left-hand side of equation (5.18) can be written as

$$\begin{aligned}
\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\mathbf{Z}_{ij}-\overline{\mathbf{Z}})(\mathbf{Z}_{ij}-\overline{\mathbf{Z}})' &= \sum_{i=1}^{m}\sum_{j=1}^{n_i}((\mathbf{Z}_{ij}-\overline{\mathbf{Z}}_i)+(\overline{\mathbf{Z}}_i-\overline{\mathbf{Z}}))((\mathbf{Z}_{ij}-\overline{\mathbf{Z}}_i)+(\overline{\mathbf{Z}}_i-\overline{\mathbf{Z}}))' \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n_i}(\mathbf{Z}_{ij}-\overline{\mathbf{Z}}_i)(\mathbf{Z}_{ij}-\overline{\mathbf{Z}}_i)' + \sum_{i=1}^{m}\sum_{j=1}^{n_i}(\overline{\mathbf{Z}}_i-\overline{\mathbf{Z}})(\overline{\mathbf{Z}}_i-\overline{\mathbf{Z}})' \\
&\quad +2\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\mathbf{Z}_{ij}-\overline{\mathbf{Z}}_i)(\overline{\mathbf{Z}}_i-\overline{\mathbf{Z}})' \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n_i}(\mathbf{Z}_{ij}-\overline{\mathbf{Z}}_i)(\mathbf{Z}_{ij}-\overline{\mathbf{Z}}_i)' + \sum_{i=1}^{m} n_i(\overline{\mathbf{Z}}_i-\overline{\mathbf{Z}})(\overline{\mathbf{Z}}_i-\overline{\mathbf{Z}})' \\
&\quad +2\sum_{i=1}^{m}\left\{\sum_{j=1}^{n_i}(\mathbf{Z}_{ij}-\overline{\mathbf{Z}}_i)\right\}(\overline{\mathbf{Z}}_i-\overline{\mathbf{Z}})'.
\end{aligned}$$

Since

$$\sum_{j=1}^{n_i}(\mathbf{Z}_{ij}-\overline{\mathbf{Z}}_i) = \sum_{j=1}^{n_i}\mathbf{Z}_{ij} - \sum_{j=1}^{n_i}\frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{Z}_{ij} = \sum_{j=1}^{n_i}\mathbf{Z}_{ij} - \sum_{j=1}^{n_i}\mathbf{Z}_{ij} = 0$$

the identity in equation (5.18) is indeed true.

A.5 Proof of the generalized weighted mean in equation (5.17)

Define the multivariate generalized weighted mean as

$$\hat{\boldsymbol{\mu}} = \sum_{i=1}^{m}\mathbf{w}_i\overline{\mathbf{Z}}_i \qquad \text{where} \qquad \sum_{i=1}^{m}\mathbf{w}_i = 1.$$

Recall from Section 5.1 that $\overline{\mathbf{Z}}_i \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{T} + n_i^{-1}\boldsymbol{\Sigma})$. The covariance matrix of $\hat{\boldsymbol{\mu}}$ is equal to

$$\text{Var}(\hat{\boldsymbol{\mu}}) = \sum_{i=1}^{m}\mathbf{w}_i\left(\mathbf{T} + n_i^{-1}\boldsymbol{\Sigma}\right)\mathbf{w}_i'.$$

The trace of the covariance matrix is equal to

$$tr\left(\text{Var}(\hat{\boldsymbol{\mu}})\right) = \sum_{i=1}^{m} tr\left(\mathbf{w}_i\left(\mathbf{T} + n_i^{-1}\boldsymbol{\Sigma}\right)\mathbf{w}_i'\right).$$

To minimize $tr\left(\text{Var}(\hat{\boldsymbol{\mu}})\right)$ subject to the constraint $\mathbf{w}_1 + \cdots + \mathbf{w}_m = 1$ we introduce a Lagrange multiplier λ such that the Lagrange function is equal to:

$$\mathcal{L}_\lambda(\mathbf{w}_1, \ldots, \mathbf{w}_m, \lambda) = \sum_{i=1}^{m} tr\left(\mathbf{w}_i\left(\mathbf{T} + n_i^{-1}\boldsymbol{\Sigma}\right)\mathbf{w}_i'\right) - \lambda\left(\sum_{i=1}^{m}\mathbf{w}_i - 1\right).$$

We will minimize the Lagrange function over $\mathbb{R}^m$. For $i = 1, \ldots, m$ we have

$$\frac{\partial\mathcal{L}_\lambda}{\partial\mathbf{w}_i} = 2\mathbf{w}_i\left(\mathbf{T} + n_i^{-1}\boldsymbol{\Sigma}\right) - \lambda.$$

Setting these partial derivatives equal to zero, we have the system of equations

$$\mathbf{w}_i = \lambda \left(\mathbf{T} + n_i^{-1}\mathbf{\Sigma}\right)^{-1}.$$

Now using the constraint $\sum_{i=1}^{m} \mathbf{w}_i = 1$ gives

$$\sum_{i=1}^{m} \lambda \left(\mathbf{T} + n_i^{-1}\mathbf{\Sigma}\right)^{-1} = 1.$$

Hence,

$$\lambda = \left(\sum_{i=1}^{m} \left(\mathbf{T} + n_i^{-1}\mathbf{\Sigma}\right)^{-1}\right)^{-1}.$$

Thus,

$$\mathbf{w}_i = \left(\sum_{i=1}^{m} \left(\mathbf{T} + n_i^{-1}\mathbf{\Sigma}\right)^{-1}\right)^{-1} \left(\mathbf{T} + n_i^{-1}\mathbf{\Sigma}\right)^{-1}.$$

A.6 Proof of the expectation in equation (5.19)

The expectation of the within group sums of squares SS_W is equal to

$$\begin{aligned}
\mathrm{E}(SS_W) &= \sum_{i=1}^{m}\sum_{j=1}^{n_i} \mathrm{E}\left[(\mathbf{Z}_{ij} - \overline{\mathbf{Z}}_i)(\mathbf{Z}_{ij} - \overline{\mathbf{Z}}_i)'\right] \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n_i} \mathrm{E}\left[((\boldsymbol{\theta}_i + \boldsymbol{\varepsilon}_{ij}) - (\boldsymbol{\theta}_i + \overline{\boldsymbol{\varepsilon}}_i))((\boldsymbol{\theta}_i + \boldsymbol{\varepsilon}_{ij}) - (\boldsymbol{\theta}_i + \overline{\boldsymbol{\varepsilon}}_i))'\right] \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n_i} \mathrm{E}\left[(\boldsymbol{\varepsilon}_{ij} - \overline{\boldsymbol{\varepsilon}}_i)(\boldsymbol{\varepsilon}_{ij} - \overline{\boldsymbol{\varepsilon}}_i)'\right] \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n_i} \mathrm{E}\left[\boldsymbol{\varepsilon}_{ij}\boldsymbol{\varepsilon}_{ij}' + \overline{\boldsymbol{\varepsilon}}_i\overline{\boldsymbol{\varepsilon}}_i' - \boldsymbol{\varepsilon}_{ij}\overline{\boldsymbol{\varepsilon}}_i' - \overline{\boldsymbol{\varepsilon}}_i\boldsymbol{\varepsilon}_{ij}'\right],
\end{aligned}$$

Because $\boldsymbol{\varepsilon}_{ij}$ is a normal random vector with zero mean and covariance matrix $\mathbf{\Sigma}$ it follows that

$$\mathrm{E}\left(\boldsymbol{\varepsilon}_{ij}\boldsymbol{\varepsilon}_{ij}'\right) = \mathbf{\Sigma}.$$

Furthermore by independence of the errors we have,

$$\begin{aligned}
\mathrm{E}\left(\overline{\boldsymbol{\varepsilon}}_i\overline{\boldsymbol{\varepsilon}}_i'\right) &= \frac{1}{n_i^2}\mathrm{E}\left[\left(\sum_j \boldsymbol{\varepsilon}_{ij}\right)\left(\sum_j \boldsymbol{\varepsilon}_{ij}\right)'\right] \\
&= \frac{1}{n_i^2}\cdot n_i\mathbf{\Sigma}.
\end{aligned}$$

For the same reason,

$$\begin{aligned}
\mathrm{E}\left(\boldsymbol{\varepsilon}_{ij}\overline{\boldsymbol{\varepsilon}}_i'\right) &= \frac{1}{n_i}\mathrm{E}\left[\boldsymbol{\varepsilon}_{ij}\left(\sum_{j'} \boldsymbol{\varepsilon}_{ij'}\right)'\right] \\
&= \frac{1}{n_i}\mathbf{\Sigma}.
\end{aligned}$$

Thus,

$$\begin{aligned} \mathrm{E}(SS_W) &= \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(\mathbf{\Sigma} - \frac{2}{n_i}\mathbf{\Sigma} + \frac{1}{n_i}\mathbf{\Sigma}\right) \\ &= \mathbf{\Sigma}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(1 - \frac{1}{n_i}\right) \\ &= \mathbf{\Sigma}(N-m). \end{aligned}$$

Bibliography

Blik op nieuws (2016) *Verdachte na DNA match aangehouden in misbruikzaak 60 jarige vrouw.* Available at: `http://www.blikopnieuws.nl/nieuws/242515/verdachte-na-dna-match-aangehouden-in-misbruikzaak-60-jarige-vrouw.html`. Accessed: 22 September 2016.

Adams, R.A., Essex, C. (2010) *Calculus: A complete course.* Seventh edition, Pearson Canada.

Aitken, C.G.G., Taroni, F. (2004) *Statistics and the evaluation of evidence for forensic scientists.* John Wiley & Sons.

Aitken, C.G.G., Lucy, D. (2004) *Evaluation of trace evidence in the form of multivariate data.* Appl. Statist, **53**, Part 1, 109-122.

Bolck, A., Weyermann, C., Dujourdy L., Esseiva, P., van den Berg, J. (2009). *Different likelihood ratio approaches to evaluate the strength of evidence of MDMA tablet comparisons.* Forensic Science International, **191**, 42-51.

Bolck, A., Alberink, I. (2011). *Variation in likelihood ratios for forensic evidence evaluation of XTC tablets comparison.* Journal of Chemometrics, **25**, 41-49.

Bolck, A., Stoel, R.D., Alberink, I., Sjerps, M. (2012) *LR models for evidence evaluation.* Chinese Journal of Forensic Science, **4**, 43-53

Champod, C. (2013) *Overview and meaning of identification/individualization.* Encyclopedia of Forensic Sciences, second ed. Elsevier, 303-309.

D'Agostino, R.B. Stephens, M.A. (1986) *Goodness-of-Fit techniques.* Marcel Dekker, Inc.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) *Maximum likelihood from incomplete data via the EM algorithm.* Journal of the Royal Statistical Society. Series B (Methodological), **39**, No. 1, 1-38.

Doornik, J.A., Hansen, H. (2008) *An omnibus test for univariate and multivariate normality.* Oxford bulletin of economics and statistics, **70**.

Duong, T. Hazelton, M.L. (2005) *Cross-validation bandwidth matrices for multivariate kernel density estimation.* Scandinavian Journal of Statistics, **32**, No. 3, 485-506.

EMCDDA (2016a) *European drug report.* Trends and developments. Available at: `http://www.emcdda.europa.eu/system/files/publications/2637/TDAT16001ENN.pdf`. Accessed: 25 July 2016.

EMCDDA (2016b) *Recent changes in Europe's MDMA/ecstasy market.* Results from an EMDDA trendspotter study. Available at: `http://www.emcdda.europa.eu/system/files/publications/2473/TD0116348ENN.pdf` Accessed: 25 July 2016.

Evett, I.W. (1998) *Towards a uniform framework for reporting opinions in forensic science casework.* Science & Justice, **38**, No. 3, 198-202.

Evett, I.W., Jackson, G., Lambert, J.A., McCrossan, S. (2000) *The impact of the principles of evidence interpretation on the structure and content of statements.* Science & Justice, **40**, 233-239.

Finkelstein, M.O., Fairley, W.B. (1970) *A Bayesian approach to identification evidence.* Harvard Law Review, **83**, No. 3, 489-517.

Friedman, R.D. (1996) *Assessing evidence.* Michigan Law Review, **94**, No. 6, 1810-1838

Gnanadesikan, R. (1977) *Methods for statical data analysis of multivariate observations.* John Wiley & Sons.

Hoff, P.D. (2009) *A first course in Bayesian statistical methods.* Springer.

Johnson, R.A. Wichern, D.W. (2007) *Applied multivariate statistical analysis.* Pearson.

Koch, I. (2014) *Analysis of multivariate and high dimensional data.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Koper, C., van den Boom, C., Wiarda, W., Schrader, M., de Joode, P., van der Peijl, G., Bolck, A. (2007) *Elemental analysis of 3,4 -methulenedioxymethamphetamine (MDMA): A tool to determine the synthesis method and trace links.* Forensic Science International, **171**, 171-179.

Lindley, D.V. (1977) *A problem in forensic science.* Biometrika, **64**, No. 2, pp. 207-213

Korkmaz, S., Goksuluk, D. and Zararsiz, G. (2015) *MVN: An R Package for Assessing Multivariate Normality.* MVN version 4.0.

Malkovich, J.F., Afifi, A.A. (1973) *On tests for multivariate normality.* Journal of the American Statistical Association, **68**, No. 341.

Mardia, K.V., Kent, J.T., Bibby, J.M. (1979) *Multivariate analysis.* Academic Press Limited.

Mardia, K.V. (1980) *9 tests of univariate and multivariate normality.*

McLachlan, G.J., Krishnan, T. (1997) *The EM Algorithm and extensions.* John Wiley & Sons.

Milliet, Q., Weyermann, C., Esseiva, P. (2009) *The profiling of MDMA tablets: A study of the combination of physical characteristics and organic impurities as sources of information.* Forensic Science International, **187**, 58-65.

Multivariate kernel density estimation (2010) *Wikipedia.* Avalaible at: `https://en.wikipedia.org/wiki/Multivariate_kernel_density_estimation` Accessed: 4 October 2016.

NFI (2014) *De reeks waarschijnlijkheidstermen van het NFI en het Bayesiaanse model voor interpretatie van bewijs.* Vakbijlage. Available at: `https://www.forensischinstituut.nl/binaries/nfi-vakbijlage-waarschijnlijkheidstermen-versie-2.1-oktober-2014_tcm35-56319.pdf` Accessed: 23 September 2016.

NOS (2016) *Database met glas in strijd tegen zware misdrijven.* Available at: `http://nos.nl/artikel/2108097-database-met-glas-in-strijd-tegen-zware-misdrijven.html` Accessed: 22 July 2016.

Nordgaard, A., Rasmusson, B. (2012) *The likelihood ratio as value of evidence – more than a question of numbers.* Law, Probability and Risk, **0**, 1-13.

Rao, C.R. (1973) *Linear statistical inference and its applications.* Second Edition, John Wiley & Sons, Chapter 8.

Rice, A. (2007) *Mathematical statistics and data analysis.* Brook/Cole.

Robert, C.P. (2007) *The Bayesian choice.* From decision-theoretic foundations to computational implementation. Second Edition, Springer.

Sahai, H., Ojeda, M. (2005) *Analysis of variance for random models.* Unbalanced Data. Theory, Methods, Applications and Data Analysis. Volume II. Birkhäuser, Chapter 11.

Searle, S.R. (1982) *Matrix algebra useful for statistics.* John Wiley & Sons.

Searle, S.R., Casella, G., McCulloch, C.E. (1992) *Variance components.* John Wiley & Sons.

Silverman, B.W. (1986) *Density estimation for statistics and data analysis.* Monographs on Statistics and Applied Probability, **26**, Chapman & Hall.

Sjerps, M. (2004) *Forensische statistiek.* NAW 5/5, No. 3, 106-111.

Shapiro, S.S., Wilk, M.B. (1965) *An analysis of variance test for normality (complete samples).* Biometrika, **52**, No. 3/4, 591-611.

Wand, M.P., Jones, M.C. (1995) *Kernel smoothing.* Monographs on Statistics and Applied Probability, **60**, Chapman & Hall.

Wasserman, L.A. (2004) *All of statistics.* A concise course in statistical inference. Springer.

Wasserman, L.A. (2006) *All of nonparametric statistics.* Springer.

Weyermann, C., Marquis, R., Delaporte, C., Esseiva, P., Lock, E., Aalberg, L., Bozenko Jr., J.S., Dieckmann, S., Dujourdy, L., Zrcek, F. (2008) *Drug intelligence based on MDMA tablets data.* I. Organic impurities profiling. Forensic Science International, **177**, 11-16.

Zadora, G., Martyna, A., Ramos, D. Aitken, C. (2014) *Statistical analysis in forensic science.* Evidential value of multivariate physicochemical data. John Wiley & Sons.