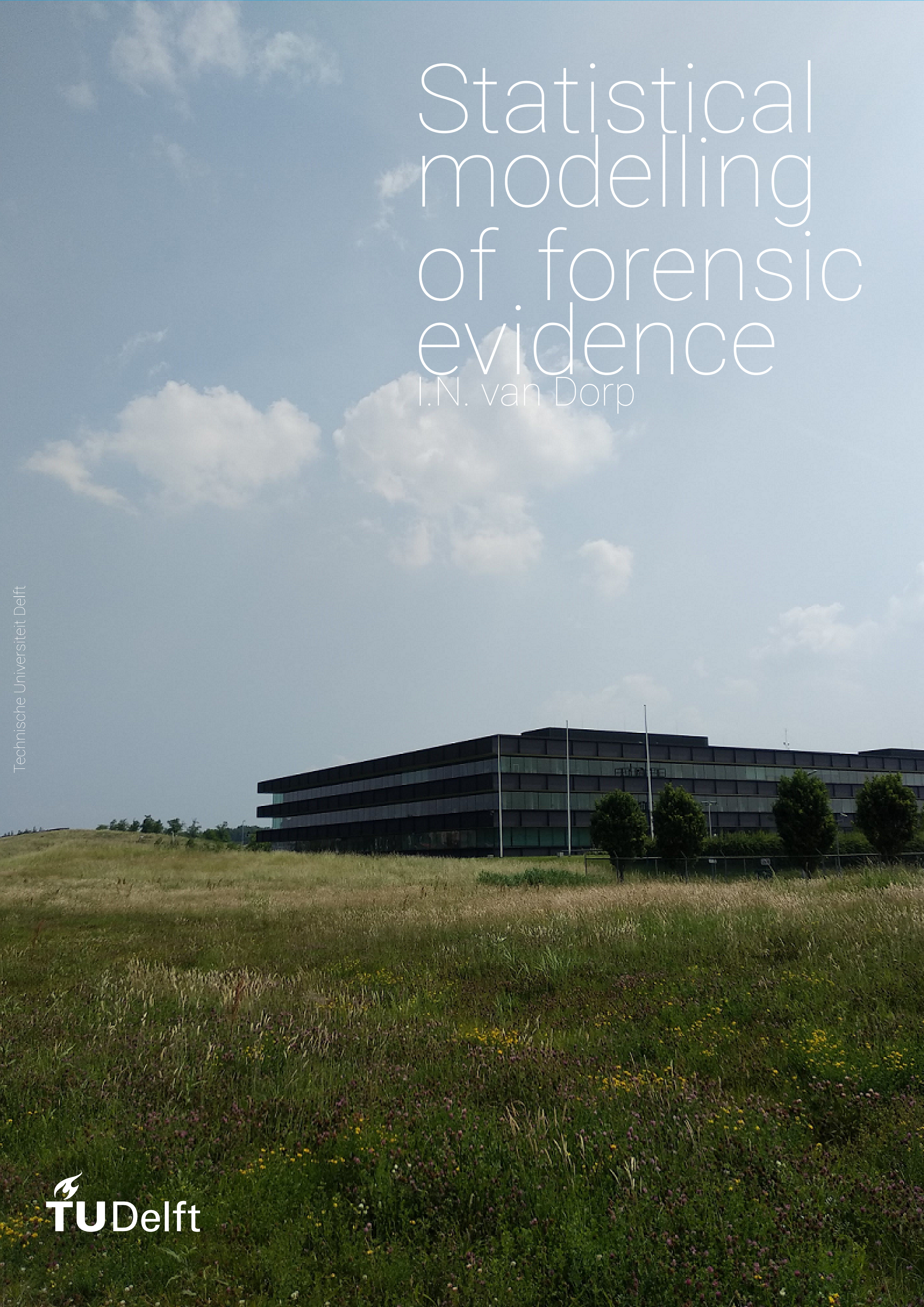

Statistical modelling of forensic evidence

I.N. van Dorp

Technische Universiteit Delft

Statistical modelling of forensic evidence

by

I.N. van Dorp

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Tuesday August 28, 2018 at 3:00 PM.

Student number:	4135644	
Project duration:	November 8, 2017 – August 28, 2018	
Thesis committee:	Prof. dr. ir. G. Jongbloed,	TU Delft, supervisor
	Dr. C. Kraaikamp,	TU Delft
	A.J. Leegwater, MSc.	NFI

Nederlands Forensisch Instituut
Ministerie van Justitie en Veiligheid

This thesis was updated on October 27th, 2019

Abstract

The evaluation of evidence found at a crime scene is primarily conducted through comparison of two competing statistical hypotheses. In forensic science, there is currently no consensus on the formulation of the competing hypotheses. A main point of discussion is the difference between common source and specific source problems. In a common source problem, all evidence is assumed to come from unknown sources, whereas the specific source problem states that one of the sources is fixed. Since the value of evidence is affected by the choice of hypotheses, this thesis tries to shed more light on the statistical framework underlying the common and specific source problem. Both a frequentist and a Bayesian approach can be followed to quantify the value of evidence, resulting in the likelihood ratio and the Bayes Factor, respectively. The theoretical framework is put into practice through two frequently used models in forensic science, namely the continuous two-level normal-normal model and the discrete one-level Bernoulli model. Since calculation of the Bayes Factor for the two-level normal-normal model cannot be done analytically, Markov Chain Monte Carlo methods are proposed and the theoretical convergence properties of the resulting methods are discussed. An explicit expression of the Bayes Factor does exist for the one-level Bernoulli model. For both models, more conservative values of the Bayes Factor are observed within the common source problem than in the specific source problem. Two approaches are considered to calculate the posterior probability of guilt and explicit bounds are derived for the difference between both techniques applied to the one-level Bernoulli model. The opportunities and challenges of a copula-based method and permutation tests are discussed as alternatives to the models generally used in evaluating forensic evidence.

Preface

This thesis has been submitted as the final requirement to obtain the degree Master of Science in Applied Mathematics at Delft University of Technology. The research was conducted in collaboration with the Netherlands Forensic Institute (NFI) in the period from November 2017 until August 2018.

During the past nine months, I have been given the opportunity to get acquainted with the applications of statistics in forensic science. The subjects covered in this thesis are the results of countless consultations and I owe a debt of gratitude to many people. First and foremost, I would like to thank Geurt Jongbloed for his guidance during my graduation project. I think it is fair to say that none of us could have predicted the course of events, but you helped me to overcome every obstacle, both personally and intellectually, and gave me the freedom to choose my own research path.

Secondly, I want to thank Jeannette Leegwater for introducing me to the challenges of forensic statistics and for always asking critical questions about the practical application of the theoretical models. Moreover, your keen eye was able to spot the smallest grammatical errors, which certainly helped to improve my thesis. I think we have learned a lot from each other and I have enjoyed our weekly meetings during this project.

Many thanks to my colleagues from the NFI, especially to Ivo Alberink, for suggesting some interesting problems to consider in my research, and to my roommates Paulien Kegge and Rebecca Hutasoit, who were always in for a cup of tea and a nice talk.

I also want to thank Cor Kraaikamp, for being part of my thesis committee and for taking the time to read and value this report.

Moreover, I would like to express my gratitude to some people who have not contributed to this thesis directly, but who supported me during the process the last nine months. A big thanks to Niels for his endless patience and love when I needed it. Thanks to my parents Ellen and Peter, and my brother Daniel, for believing that I can do anything and for always supporting me in my decisions. And last but not least, thanks to my friends for always having my back and making my student life the amazing time that it was.

Lastly, I would like to name Annabel Bolck, who has given me the possibility to carry out my graduation project at the NFI. Although we did not get the chance to work together, I owe it to her that I have been able to perform my research in the field of forensic statistics.

I.N. van Dorp
Delft, August 2018

Contents

1

Introduction

When evidence is found at a crime scene, the main interest is in the information the corresponding traces provide about the origin of the evidence. The evaluation of evidence is realised through a close cooperation between the law and forensic science. In the Netherlands, most of the forensic investigations are performed by the Netherlands Forensic Institute (NFI). Forensic experts assist legal decision makers in reconstructing past events, where the main focus is on quantifying the *value of evidence*. Considering just the traces without a framework to evaluate the evidence, there is little to say about their value. Over the past decades, forensic scientist have cautiously studied the process of evidence evaluation, where the following position has been reached:

"In court as elsewhere, the data cannot 'speak for themselves'. They have to be interpreted in the light of two competing hypotheses put forward, against a background of knowledge and experience about the world." [39]

This quote illustrates the main idea of forensic evidence evaluation. The formulation of hypotheses is of great importance, since it is closely related to the value of evidence. Moreover, the hypotheses should be formulated such that background knowledge sheds light on the interpretation of the evidence. However, currently there is no consensus on the formulation of the competing hypotheses. The primary dispute is on the difference between *common source* and *specific source* problems, as formulated by [33]. In a common source problem, all evidence is assumed to come from unknown sources, whereas the specific source problem states that one of the sources is fixed.

To quantify the value of evidence, statistical models are proposed by [32] based on the framework specified by the competing hypotheses. This means that changing the hypotheses will also change the value of evidence. Therefore, the difference between the common source and specific source problem will be a recurring subject in this research. The proposed models can be used for evidence that can be expressed in terms of continuous data as well as discrete data. To illustrate the forensic context, two examples will be used throughout this thesis: for the continuous evidence, the elemental composition of glass fragments is considered, whereas the discrete evidence is illustrated by DNA profiles.

Thesis outline

This research was started by addressing the different forensic identification of source questions posed by [33]. Existing literature on this subject was studied and summarized to provide a clear statistical framework, with explicit mentioning of the underlying assumptions. Both the common and specific source model from [31–33] are presented in Chapter 2, where also general expressions for the corresponding likelihood functions are restated. These likelihood functions will play an important role in quantifying the value of evidence.

Chapter 3 describes how the value of evidence is defined. In forensic statistics, there are two commonly used approaches to quantify the value of evidence: frequentists will say that the evidential value is given by the likelihood ratio, whereas Bayesians believe that the Bayes Factor should be considered. Both the likelihood ratio and the Bayes Factor from [32] will be restated for the common source and specific source problem, and two interesting relationships between the statistics will be presented from [31, 32] for ease of reference. Moreover, some explanation will be given about how the value of evidence is interpreted in practice.

After the theoretical framework is completed for both the common and specific source problem, a commonly used continuous model in forensic science is introduced in Chapter 4. This two-level normal-normal model is a hierarchical model that assumes normality in two levels. Since conjugate priors are proposed, the possibilities of analytical calculation of the Bayes Factors are investigated. The NFI prefers analytical expressions for the value of evidence because approximation methods can lead to numerical errors that could be difficult to quantify. Moreover, the results of a stochastic approximation can be hard to justify in court, which is undesirable for a transparent lawsuit.

Unfortunately, the Bayes Factor of the two-level normal-normal cannot be computed analytically, and therefore Chapter 5 considers alternative methods to compute the value of evidence. Here, Monte Carlo integration is combined with Gibbs sampling and the theoretical convergence of the resulting Markov chain is discussed for both the common and specific source problem.

The proposed Markov Chain Monte Carlo method is put into practice in Chapter 6. The two-level normal-normal model is applied to both simulated and real datasets. First, only one-dimensional data are considered to investigate some interesting properties of the model, such as the influence of the prior distributions on the Bayes Factor. Later on, the model is applied to higher dimensional data and the results of the described model are compared with a (frequentist) approach currently used at the NFI.

In the frequentist approach, estimators are needed to evaluate the likelihood ratio corresponding to the two-level normal-normal model. In forensic statistics, two estimators are commonly used to estimate the overall mean. The effect of choosing either of the mean estimators is explored in Chapter 7.

Because the two-level normal-normal model is quite a restrictive model and does not always provide the best fit to the data, copulas are considered in Chapter 8 as a possible alternative approach. Unfortunately, the most general form of the copula model is found to be too difficult to apply in practice. Some alternative approaches from literature are considered to show how copulas can be used to model forensic evidence.

Since using the theoretical framework for the common and specific source problem in a discrete setting is not straightforward, Chapter 9 explains this procedure in more detail. The resulting models are consistent with the approach usually taken in the literature, although the relation with the general framework as described in Chapter 2 has not been made before.

In Chapter 10 a frequently used model for discrete evidence in forensic science is considered: the one-level Bernoulli model. The discrete setup from the previous chapter is used to derive the Bayes Factor for both the common and specific source problem. This derivation sheds more light on the debate about which traces should be added to the background material in the evidence evaluation process. A benefit of the one-level Bernoulli model is that for certain priors a closed form expression of the Bayes Factor exists. Under this model, the difference between the two identification of source problems is shown to be significant when a rare type trace is recovered.

When evidence is found in a criminal case, the court is mainly interested in the probability of guilt given all available evidence. Therefore, Chapter 11 considers two approaches to calculate this posterior probability of guilt. The theoretical differences and similarities of these two approaches are presented. Using the common source one-level Bernoulli problem, the posterior probability of guilt is evaluated and an upper and lower bound for the difference between the two approaches is derived.

Finally, an alternative to the ‘standard’ forensic approach to quantify the value of evidence is given in Chapter 12. Here, the distributional assumptions on the evidence are dropped and permutation tests are used to consider the exchangeability of the evidence sets. Different test statistics are applied to both simulated and real data and their performance is measured through the evaluation of type I and type II errors.

2

Forensic identification of source

This chapter contains a literature study on the forensic identification of source, primarily based on the work from [31–33]. The aim of this chapter is to give an overview for ease of reference and to provide some extra explanation of the underlying theory.

The main task of a forensic scientist is to provide help in deciding between two competing forensic hypotheses. One of these hypotheses will be presented by the prosecution, denoted H_p, and one by the defence, denoted H_d. The formulation of the forensic hypotheses depends upon the source identification question of interest [31]. Although other types of identification problems may be encountered in forensic science, in this research the focus will be on the common source and specific source identification question. In the *specific source* identification question, one is interested whether a trace originates from a fixed specific source, whereas the *common source* identification problem seeks to answer the question if two traces of unknown origin share the same common source, which is not known [32].

To test these types of hypotheses, an alternative to the traditional hypothesis testing methods has to be provided. In the traditional textbook tests, the two competing hypotheses specify parametric models for the data up to the point of a finite dimensional vector space for the indexing parameter θ. Then the hypothesis that $\theta \in \Theta_0$ would be tested against the hypothesis $\theta \in \Theta_1$ for two disjoint subsets $\Theta_i \subset \Theta$, $i = 0, 1$, of the parameter space Θ. Of course, alternative tests, such as goodness of fit tests, exists. However, in forensic science the interest is not only in the parameters, but also in the process of generating the evidence. Therefore, the parameter space now consists of a set of possible *sampling models* from which a selection has to be made [32]. This means that there is uncertainty about both the parameter and the corresponding parametric model.

The setup from [32] is used as a starting point of this research. To model the forensic evidence, three main components need to be specified. The first is the statement of the sampling models, which provide information about the exchangeability of the evidence given each hypothesis. Secondly, the class of parametric models approximating the true sampling distributions used in the sampling models has to be determined. Finally, a prior belief structure has to be chosen for the parameters characterizing the class of probability models specified by each of the parametric models. This can be either an estimate of the parameter to fit the frequentist framework or a prior distribution under the Bayesian paradigm.

This chapter will be devoted to uncovering the assumptions made in the construction of the models for both the common source and specific source problems. Moreover, the sampling models as discussed in Chapter 3 from [32] will be reformulated to clarify the underlying mathematical models. Finally, following [32], the sampling models will be used to derive the likelihood functions for the evidence given each of the hypotheses. These will be used later on in quantifying the value of evidence.

2.1. Common source problem

Suppose two traces are found, possibly at two different crime scenes, and one is interested if the traces come from the same unknown source. Several features can be measured from samples of each trace, and these features are the same for both traces. The measurements of each sample can be represented by a vector of values with length the number of features. The collection of these sample measurements of the two traces forms the unknown source evidence. In this common source identification problem, the following hypotheses could be considered [32, 33]:

H_p: The two sets of unknown source evidence (e_{u_1} and e_{u_2}) both originate from the same unknown source.
H_d: The two sets of unknown source evidence (e_{u_1} and e_{u_2}) originate from two different unknown sources.

To give meaning to these forensic hypotheses in a testing setting, a sampling scheme is needed, as proposed in [32]. Suppose the discovered traces consist of glass fragments and one is interested if these fragments originate from the same (unknown) window.

Before making any statement about this, the forensic expert has to consider some background material of windows for comparison. The windows used for this comparison can be seen as a random sample from the total population of windows. In the most general scenario, all types of glass could be considered in the total population of windows, ranging from car window glass to centuries old stained glass. The elemental composition can differ significantly in each type of glass [43]. In practice, the forensic expert will have some idea about the type and dating of the glass fragments corresponding to the traces. Therefore, only the relevant sources as determined by the forensic expert will be considered. For each window, the elemental composition of a sample of glass fragments can be evaluated and the most discriminative elements in glass can be used as features [2]. This means that for each source, only the k relevant features as determined by the forensic expert will be considered. With these assumptions on the data, the first modelling assumption is:

Assumption 1 The total population of (alternative) sources follows a certain k-dimensional distribution $G(\cdot|\boldsymbol{\theta}_a)$ indexed by parameter $\boldsymbol{\theta}_a$. The sources $\mathbf{A}_i$ used for the background material are random samples from the total population of (alternative) sources, i.e., $\mathbf{A}_i \overset{\text{iid}}{\sim} G(\cdot|\boldsymbol{\theta}_a)$.

If the windows used for the background material are known, the background material can be generated by randomly sampling some glass fragments from one of the windows and by measuring each feature per fragment. For each fragment, it is assumed that there is only dependency on the variation of the most discriminative elements within the window. For example, if the elemental composition of the glass depends on the year the window was manufactured, the variation of the most discriminative elements is assumed to be independent of the year of manufacturing, and the time dependency should be entirely captured by the distribution G and the parameter $\boldsymbol{\theta}_a$. This leads to the second modelling assumption:

Assumption 2 Given source $\mathbf{A}_i = \mathbf{a}_i$, the background samples $\mathbf{Y}_{ij}$ follow a certain k-dimensional distribution $F_a(\cdot|\mathbf{a}_i, \boldsymbol{\theta}_a)$ indexed by $\mathbf{a}_i$ and parameter $\boldsymbol{\theta}_a$. The background samples $\mathbf{Y}_{ij}|\mathbf{A}_i = \mathbf{a}_i$ are random samples from within the source, i.e., $\mathbf{Y}_{ij}|\mathbf{A}_i = \mathbf{a}_i \overset{\text{iid}}{\sim} F_a(\cdot|\mathbf{a}_i, \boldsymbol{\theta}_a)$.

The distribution G gives the general elemental composition of each window, while the distribution F_a models the variation of the elements within each window. The distribution G is in forensic science usually referred to as the *between-source distribution* and F_a is called the *within-source distribution*. Since the samples are assumed to be organised at both a between-source level and a within-source level, one usually speaks of a *two-level model*.

For the common source problem, the unknown source evidence is assumed to be generated similarly to the background material, although the prosecution will argue that only one time a source was sampled, whereas according to the defence two sources were sampled.

2.1.1. Sampling models

In [32], the sampling models for the common source problem at first seem to be described as simple random sampling models, where a finite total population of sources is considered and sampling is done without replacement. However, sampling distributions are described later on, which is consistent with the definition of G and F_a in the modelling assumptions. While the simple random sampling models are easy to interpret and intuitively clear, here it is chosen to give the formal definition of the sampling models.

Following [32], let e_a denote the background material, e_{u_1} the traces from the first unknown source, and e_{u_2} the traces from the second unknown source. Under Assumption 1 and 2, sampling models implied by the prosecution hypothesis and the defence hypothesis can be constructed for each set of evidence.

The prosecution will argue that the background material e_a has been generated according to sampling model M_a and that the recovered evidence, e_{u_1} and e_{u_2}, has been generated according to sampling model M_p. On the other hand, the defence will claim that the background material e_a has been generated according to sampling model M_a, but that the recovered evidence, e_{u_1} and e_{u_2}, has been generated according to sampling model M_d. Thus, both the prosecution and the defence agree on the generation of the background material e_a. All three sampling models will be described here [31, 32]:

M_a:

- Sample n_a sources from the total population of sources,

$$\mathbf{A}_i \overset{\text{iid}}{\sim} G(\cdot|\boldsymbol{\theta}_a), \qquad \text{for } i = 1, 2, \ldots, n_a.$$

- For each of the sampled sources, sample n_i elements from within the ith source,

$$\mathbf{Y}_{ij}|\mathbf{A}_i = \mathbf{a}_i \overset{\text{iid}}{\sim} F_a(\cdot|\mathbf{a}_i, \boldsymbol{\theta}_a) \qquad \text{for } j = 1, 2, \ldots, n_i$$

and set

$$\mathbf{Y}_i = \begin{bmatrix} \mathbf{Y}_{i1} & \mathbf{Y}_{i2} & \cdots & \mathbf{Y}_{in_i} \end{bmatrix} = \begin{bmatrix} Y_{i,11} & Y_{i,12} & \cdots & Y_{i,1n_i} \\ Y_{i,21} & Y_{i,22} & \cdots & Y_{i,2n_i} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{i,k1} & Y_{i,k2} & \cdots & Y_{i,kn_i} \end{bmatrix}. \tag{2.1}$$

Here, $Y_{i,jl}$ denotes the jth measurement of the lth sample from source $\mathbf{a}_i$. Since there are k measurements and n_i samples, $\mathbf{Y}_i$ is a $k \times n_i$ matrix for $i = 1, 2, \ldots, n_a$. Each measurement per sample corresponds to one feature, so this is equivalent to saying that there are n_i observations of k features.

- Let $\mathbf{y}_i = \begin{bmatrix} \mathbf{y}_{i1} & \mathbf{y}_{i2} & \cdots & \mathbf{y}_{in_i} \end{bmatrix}$ denote the realisations of the measurements on the samples from the ith source. The realisations from all background sources can be represented by the following block matrix structure:

$$\mathbf{y}_a = \left[\begin{array}{c|c|c|c} \mathbf{y}_1 & \mathbf{y}_2 & \cdots & \mathbf{y}_{n_a} \end{array} \right], \tag{2.2}$$

where each block corresponds to a sampled source. This results in a $k \times \left(\sum_{i=1}^{n_a} n_i\right)$ dimensional matrix. The background material is then given by the composed measurement vector $e_a = (\mathbf{y}_{ij}, 1 \le i \le n_a, 1 \le j \le n_i)$.

M_p:

- Sample a single source from the total population of sources,

$$\mathbf{P} \sim G(\cdot|\boldsymbol{\theta}_a).$$

- Sample the first set of n_{u_1} elements from within source $\mathbf{P} = \mathbf{p}$,

$$\mathbf{Y}_{u_1 j}|\mathbf{P} = \mathbf{p} \overset{\text{iid}}{\sim} F_a(\cdot|\mathbf{p}, \boldsymbol{\theta}_a) \qquad \text{for } j = 1, 2, \ldots, n_{u_1}.$$

- Keep $\mathbf{p}$ fixed and sample the second set of n_{u_2} elements from within source $\mathbf{p}$, independently of the first sample

$$\mathbf{Y}_{u_2 j}|\mathbf{P} = \mathbf{p} \overset{\text{iid}}{\sim} F_a(\cdot|\mathbf{p}, \boldsymbol{\theta}_a) \qquad \text{for } j = 1, 2, \ldots, n_{u_2}.$$

- Set

$$\mathbf{Y}_{u_i} = \begin{bmatrix} \mathbf{Y}_{u_i 1} & \mathbf{Y}_{u_i 2} & \cdots & \mathbf{Y}_{u_i n_{u_i}} \end{bmatrix} = \begin{bmatrix} Y_{u_i,11} & Y_{u_i,12} & \cdots & Y_{u_i,1n_{u_i}} \\ Y_{u_i,21} & Y_{u_i,22} & \cdots & Y_{u_i,2n_{u_i}} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{u_i,k1} & Y_{u_i,k2} & \cdots & Y_{u_i,kn_{u_i}} \end{bmatrix} \qquad \text{for } i = 1, 2, \tag{2.3}$$

where $Y_{u_i,jl}$ denotes the jth measurement of the lth sample from source $\mathbf{p}$. This results in a $k \times n_{u_i}$ matrix for $i = 1, 2$.

- Let $\mathbf{y}_{u_i} = \begin{bmatrix}\mathbf{y}_{u_i1} & \mathbf{y}_{u_i2} & \cdots & \mathbf{y}_{u_in_{u_i}}\end{bmatrix}$ denote the realisations of the measurements on the samples from source $\mathbf{p}$, for $i = 1,2$. The material originating from the first unknown source is then given by the composed measurement vector $e_{u_1} = (\mathbf{y}_{u_1j}, 1 \leq j \leq n_{u_1})$ and the material originating from the second unknown source is $e_{u_2} = (\mathbf{y}_{u_2j}, 1 \leq j \leq n_{u_2})$.

M_d:

- Independently sample two sources from the total population of sources,

$$\mathbf{D}_1 \sim G(\cdot|\boldsymbol{\theta}_a) \qquad \text{and} \qquad \mathbf{D}_2 \sim G(\cdot|\boldsymbol{\theta}_a).$$

- Sample n_{u_1} elements from within the first source,

$$\mathbf{Y}_{u_1j}|\mathbf{D}_1 = \mathbf{d}_1 \overset{\text{iid}}{\sim} F_a(\cdot|\mathbf{d}_1, \boldsymbol{\theta}_a) \qquad \text{for } j = 1,2,\ldots,n_{u_1}.$$

- Sample n_{u_2} elements from within the second source,

$$\mathbf{Y}_{u_2j}|\mathbf{D}_2 = \mathbf{d}_2 \overset{\text{iid}}{\sim} F_a(\cdot|\mathbf{d}_2, \boldsymbol{\theta}_a) \qquad \text{for } j = 1,2,\ldots,n_{u_2}.$$

- Set

$$\mathbf{Y}_{u_i} = \begin{bmatrix}\mathbf{Y}_{u_i1} & \mathbf{Y}_{u_i2} & \cdots & \mathbf{Y}_{u_in_{u_i}}\end{bmatrix} = \begin{bmatrix} Y_{u_i,11} & Y_{u_i,12} & \cdots & Y_{u_i,1n_{u_i}} \\ Y_{u_i,21} & Y_{u_i,22} & \cdots & Y_{u_i,2n_{u_i}} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{u_i,k1} & Y_{u_i,k2} & \cdots & Y_{u_i,kn_{u_i}} \end{bmatrix} \qquad \text{for } i = 1,2, \tag{2.4}$$

  where $Y_{u_i,jl}$ denotes the jth measurement of the lth sample from source $\mathbf{d}_i$. This results in a $k \times n_{u_i}$ matrix for $i = 1,2$.

- Let $\mathbf{y}_{u_i} = \begin{bmatrix}\mathbf{y}_{u_i1} & \mathbf{y}_{u_i2} & \cdots & \mathbf{y}_{u_in_{u_i}}\end{bmatrix}$ denote the realisations of the measurements on the samples from source $\mathbf{d}_i$, for $i = 1,2$. The material originating from the first unknown source is then given by the composed measurement vector $e_{u_1} = (\mathbf{y}_{u_1j}, 1 \leq j \leq n_{u_1})$ and the material originating from the second unknown source is $e_{u_2} = (\mathbf{y}_{u_2j}, 1 \leq j \leq n_{u_2})$.

Note that under the prosecution model M_p, e_{u_1} and e_{u_2} are conditionally independent given the common source $\mathbf{P}$, i.e., both evidence sets consist of random samples from the same common source. Under the defence model M_d, e_{u_1} and e_{u_2} are unconditionally independent, since both evidence sets contain random samples from different sources.

2.1.2. Likelihood functions

Let $e = \{e_{u_1}, e_{u_2}, e_a\}$ denote the collection of three datasets containing the available forensic evidence. In Section 2.1.1, the three sampling models for the common source problem were explained and the resulting evidence was given in both matrix notation and as composed measurement vectors.

The background material e_a is according to both hypotheses obtained from the hierarchical sampling model M_a. Using the same notation as in equation (2.1) and (2.2), $\mathbf{y}_a$ is a block matrix (also known as partitioned matrix) where each block represents a source, each column within a block corresponds to a sample from that source, and each row corresponds to a measurement on a sample. Given the source, each sample can be seen as a random vector that is independent of and identically distributed to the other samples within that source. In this modelling stage it is assumed that $\boldsymbol{\theta}_a$ is known, either from an estimate (frequentist) or as the result of a prior distribution (Bayesian). More details about the parameter $\boldsymbol{\theta}_a$ will be given in Chapter 3. The sampling model M_a can be represented by setting

$$\mathbf{A}_i \overset{\text{iid}}{\sim} G(\cdot|\boldsymbol{\theta}_a) \qquad \text{and} \qquad \mathbf{Y}_{ij}|\mathbf{A}_i = \mathbf{a}_i \overset{\text{iid}}{\sim} F_a(\cdot|\mathbf{a}_i, \boldsymbol{\theta}_a) \qquad \text{for } i = 1,2,\ldots,n_a \text{ and } j = 1,2,\ldots,n_i. \tag{2.5}$$

Here, G is the probability distribution function indexed by parameter $\boldsymbol{\theta}_a$ used to sample source $\mathbf{A}_i$ from the total population of sources. Given the source $\mathbf{a}_i$, the n_i samples $\mathbf{Y}_{ij}$ are obtained from the probability distribution F_a corresponding to the probability measure under model M_a.

Let f_a denote the probability density function corresponding to F_a and let g be the probability density function corresponding to G. Following [32], the likelihood functions corresponding to the common source problem will be derived in this section. The probability density function for vector $\mathbf{y}_{ij}$ for fixed i and j can be defined by

$$f_a(\mathbf{y}_{ij}|\boldsymbol{\theta}_a) = \int f_a(\mathbf{y}_{ij},\mathbf{a}_i|\boldsymbol{\theta}_a)\; d\mathbf{a}_i = \int f_a(\mathbf{y}_{ij}|\mathbf{a}_i,\boldsymbol{\theta}_a) g(\mathbf{a}_i|\boldsymbol{\theta}_a)\; d\mathbf{a}_i. \tag{2.6}$$

Note that for fixed i, $\mathbf{y}_{ij}$ are conditionally independent and identically distributed for $j = 1,2,\dots,n_i$. However, dependency still exists if all samples in e_a are considered. This means that within each block in $\mathbf{y}_a$ the columns are conditionally independent, but that the set of all columns in the matrix is dependent. Using equation (2.6), the joint probability density function of $\mathbf{y}_{ij}$, $j = 1,2,\dots,n_i$, can be defined for fixed i as

$$f_i(\mathbf{y}_{i1},\mathbf{y}_{i2},\dots,\mathbf{y}_{in_i}|\boldsymbol{\theta}_a) = \int \prod_{j=1}^{n_i} f_a(\mathbf{y}_{ij}|\mathbf{a}_i,\boldsymbol{\theta}_a) g(\mathbf{a}_i|\boldsymbol{\theta}_a)\; d\mathbf{a}_i. \tag{2.7}$$

Let f denote the likelihood structure of the evidence. Since the sources $\mathbf{a}_i$ are sampled randomly from the total population of sources for $i = 1,2,\dots,n_a$, the likelihood function for e_a can now be expressed as

$$\begin{aligned} f(e_a|\boldsymbol{\theta}_a,H_p) = f(e_a|\boldsymbol{\theta}_a,H_d) &= \prod_{i=1}^{n_a} f_i(\mathbf{y}_{i1},\mathbf{y}_{i2},\dots,\mathbf{y}_{in_i}|\boldsymbol{\theta}_a) \\ &= \prod_{i=1}^{n_a}\left(\int \prod_{j=1}^{n_i} f_a(\mathbf{y}_{ij}|\mathbf{a}_i,\boldsymbol{\theta}_a) g(\mathbf{a}_i|\boldsymbol{\theta}_a)\; d\mathbf{a}_i\right). \end{aligned} \tag{2.8}$$

Note that the likelihood function for e_a is the same given both the prosecution and defence hypothesis and therefore the notational dependence on H_p or H_d is typically dropped, i.e., $f(e_a|\boldsymbol{\theta}_a,H_p) = f(e_a|\boldsymbol{\theta}_a,H_d) = f(e_a|\boldsymbol{\theta}_a)$ [31, 32].

Now consider both datasets with unknown source evidence e_{u_i}, for $i = 1,2$, which consist of n_{u_1} and n_{u_2} random samples from the first and second unknown source, respectively. Using the same notation as in equations (2.3) and (2.4), $\mathbf{y}_{u_1}$ and $\mathbf{y}_{u_2}$ are matrices where each column corresponds to a sample from one of the unknown sources, and each row corresponds to a measurement on a sample. Given the source, each sample can be seen as a random vector that is independent of and identically distributed to the other samples within that source. The sampling model M_p can be represented by

$$\mathbf{P} \sim G(\cdot|\boldsymbol{\theta}_a) \qquad \text{and} \qquad \mathbf{Y}_{u_ij}|\mathbf{P} = \mathbf{p} \overset{\text{iid}}{\sim} F_a(\cdot|\mathbf{p},\boldsymbol{\theta}_a) \qquad \text{for } i = 1,2 \text{ and } j = 1,2,\dots,n_{u_i}.$$

Given H_p, e_{u_1} and e_{u_2} consist of samples drawn from the same randomly selected source $\mathbf{p}$ from the total population of sources. Conditional on the selected source, the samples are independent. Therefore, the joint likelihood function for e_{u_1} and e_{u_2} given H_p is defined as

$$\begin{aligned} f(e_{u_1},e_{u_2}|\boldsymbol{\theta}_a,H_p) &= f_u(\mathbf{y}_{u_11},\dots,\mathbf{y}_{u_1n_{u_1}},\mathbf{y}_{u_21},\dots,\mathbf{y}_{u_2n_{u_2}}|\boldsymbol{\theta}_a) \\ &= \int\left(\prod_{j=1}^{n_{u_1}} f_a(\mathbf{y}_{u_1j}|\mathbf{p},\boldsymbol{\theta}_a)\right)\left(\prod_{j=1}^{n_{u_2}} f_a(\mathbf{y}_{u_2j}|\mathbf{p},\boldsymbol{\theta}_a)\right) g(\mathbf{p}|\boldsymbol{\theta}_a)\; d\mathbf{p}. \end{aligned}$$

Assuming exchangeability of $\mathbf{y}_{u_11},\dots,\mathbf{y}_{u_1n_{u_1}}$ and $\mathbf{y}_{u_21},\dots,\mathbf{y}_{u_2n_{u_2}}$, this expression can be simplified by setting $\mathbf{y}_u = (\mathbf{y}_{u_11},\dots,\mathbf{y}_{u_1n_{u_1}},\mathbf{y}_{u_21},\dots,\mathbf{y}_{u_2n_{u_2}})$ and $n_u = n_{u_1} + n_{u_2}$:

$$f(e_{u_1},e_{u_2}|\boldsymbol{\theta}_a,H_p) = \int \prod_{j=1}^{n_u} f_a(\mathbf{y}_{uj}|\mathbf{p},\boldsymbol{\theta}_a) g(\mathbf{p}|\boldsymbol{\theta}_a)\; d\mathbf{p} \tag{2.9}$$

Conversely, given H_d, e_{u_1} and e_{u_2} consist of samples drawn from two independent randomly selected sources from the total population of sources. The sampling model M_d can be represented by

$$\mathbf{D}_1 \sim G(\cdot|\boldsymbol{\theta}_a) \qquad \text{and} \qquad \mathbf{Y}_{u_1j}|\mathbf{D}_1 = \mathbf{d}_1 \overset{\text{iid}}{\sim} F_a(\cdot|\mathbf{d}_1,\boldsymbol{\theta}_a) \qquad \text{for } j = 1,2,\dots,n_{u_1},$$

$$\mathbf{D}_2 \sim G(\cdot|\boldsymbol{\theta}_a) \qquad \text{and} \qquad \mathbf{Y}_{u_2j}|\mathbf{D}_2 = \mathbf{d}_2 \overset{\text{iid}}{\sim} F_a(\cdot|\mathbf{d}_2,\boldsymbol{\theta}_a) \qquad \text{for } j = 1,2,\dots,n_{u_2}.$$

This means that the samples from e_{u_1} and e_{u_2} are independent unconditional on the source, which leads to the following joint likelihood function for e_{u_1} and e_{u_2}:

$$\begin{aligned} f(e_{u_1}, e_{u_2}|\boldsymbol{\theta}_a, H_d) &= f(e_{u_1}|\boldsymbol{\theta}_a, H_d) f(e_{u_2}|\boldsymbol{\theta}_a, H_d) \\ &= f_{u_1}(\mathbf{y}_{u_1 1}, \dots, \mathbf{y}_{u_1 n_{u_1}}|\boldsymbol{\theta}_a) f_{u_2}(\mathbf{y}_{u_2 1}, \dots, \mathbf{y}_{u_2 n_{u_2}}|\boldsymbol{\theta}_a) \\ &= \left(\int \prod_{j=1}^{n_{u_1}} f_a(\mathbf{y}_{u_1 j}|\mathbf{d}_1, \boldsymbol{\theta}_a) g(\mathbf{d}_1|\boldsymbol{\theta}_a)\, d\mathbf{d}_1\right)\left(\int \prod_{j=1}^{n_{u_2}} f_a(\mathbf{y}_{u_2 j}|\mathbf{d}_2, \boldsymbol{\theta}_a) g(\mathbf{d}_2|\boldsymbol{\theta}_a)\, d\mathbf{d}_2\right). \end{aligned} \tag{2.10}$$

The likelihood functions obtained in equation (2.9) and (2.10) will play an important role in Chapter 3 to quantify the value of evidence for the common source problem.

2.2. Specific source problem

Suppose a trace is found at a crime scene and one is interested if a specific suspect can be linked to this trace. Several features from samples of the trace can be measured, and the same features can be measured from samples of the suspect. The measurements of each sample can be represented by a vector of values with length the number of features. The collection of the sample measurements of the trace forms the unknown source evidence and the collection of the sample measurements of the suspect forms the specific source evidence.

Now suppose a trace is found on a suspect and one is interested if a specific crime scene can be linked to this trace. After measuring several features from both samples of the trace and samples of the crime scene, the collection of the sample measurements of the trace forms the unknown source evidence and the collection of the sample measurements of the crime scene forms the specific source evidence.

Both scenarios describe a specific source identification problem, where the following hypotheses could be considered [32, 33]:

H_p : The unknown source evidence e_u and the specific source evidence e_s both originate from the specific source.

H_d : The unknown source evidence e_u does not originate from the specific source, but from some other source in the alternative source population.

To give meaning to these forensic hypotheses in a testing setting, again a sampling scheme is needed, as proposed in [32]. Suppose that glass fragments are found on the clothes of the suspect. Then the question of interest is whether or not the glass fragments on the suspect originate from the (known) window at the crime scene. Again, the forensic expert has to consider some background material of windows for comparison. As in the common source problem, several assumptions are made when modelling the specific source problem. The assumptions on the data as well as Assumptions 1 and 2 remain unchanged, but the specific source problem leads to two more modelling assumptions:

Assumption 3 The specific source is known and does not depend on the total population of alternative sources or a between-source distribution.

Assumption 4 Given the specific source, the samples $\mathbf{Y}_{sj}$ follow a certain k-dimensional distribution $F_s(\cdot|\boldsymbol{\theta}_s)$ indexed by parameter $\boldsymbol{\theta}_s$. The specific source samples $\mathbf{Y}_{sj}$ are random samples from within the specific source, i.e., $\mathbf{Y}_{sj} \overset{\text{iid}}{\sim} F_s(\cdot|\boldsymbol{\theta}_s)$.

Note that in the specific source problem, the background material is generated from sources sampled from the total population of alternative sources [32, 33], i.e., other sources than the specific source. This means that it is assumed that the specific source is not contained in the total population of alternative sources.

In the specific source problem, the prosecution will argue that the unknown source evidence is generated similarly to the specific source evidence, whereas according to the defence it is generated similarly to the background material, which was described in Assumptions 1 and 2.

2.2.1. Sampling models

In [32], the sampling models for the specific source problem at first seem to be described as simple random sampling models. However, sampling distributions are described later on, which is consistent with the definition of G, F_a and F_s in the modelling assumptions. Again here it is chosen to give the formal definition of the sampling models.

Following [32], let e_a denote the background material, e_u the traces from the unknown source, and e_s the traces from the specific source. Under Assumptions 1-4, sampling models implied by the prosecution hypothesis and the defence hypothesis can be constructed for each set of evidence.

The prosecution will argue that the background material e_a and the specific source evidence e_s have been generated according to model M_a and M_s, respectively, and that the recovered evidence e_u has been generated according to model M_p. On the other hand, the defence will claim that the background material e_a and the specific source evidence e_s have been generated according to model M_a and M_s, respectively, but that the recovered evidence e_u has been generated according to model M_d. Thus, both the prosecution and the defence agree on the generation of the background material e_a and the specific source evidence e_s. All four sampling models will be described here [31, 32]:

M_s:

- Sample n_s elements from within the specific source,

$$\mathbf{Y}_{sj} \overset{\text{iid}}{\sim} F_s(\cdot|\boldsymbol{\theta}_s) \qquad \text{for } j = 1, 2, \ldots, n_s$$

  and set

$$\mathbf{Y}_s = \begin{bmatrix} \mathbf{Y}_{s1} & \mathbf{Y}_{s2} & \cdots & \mathbf{Y}_{sn_s} \end{bmatrix} = \begin{bmatrix} Y_{s,11} & Y_{s,12} & \cdots & Y_{s,1n_s} \\ Y_{s,21} & Y_{s,22} & \cdots & Y_{s,2n_s} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{s,k1} & Y_{s,k2} & \cdots & Y_{s,kn_s} \end{bmatrix}. \tag{2.11}$$

  Here, $Y_{s,jl}$ denotes the jth measurement of the lth sample from the specific source. Since there are k measurements and n_s samples, $\mathbf{Y}_s$ is a $k \times n_s$ matrix. Each measurement per sample corresponds to one feature, so this is equivalent to saying that there are n_s observations of k features.

- Let $\mathbf{y}_s = \begin{bmatrix} \mathbf{y}_{s1} & \mathbf{y}_{s2} & \cdots & \mathbf{y}_{sn_s} \end{bmatrix}$ denote the realisations of the measurements on the samples from the specific source. The material originating from the specific source is then given by the composed measurement vector $e_s = (\mathbf{y}_{sj}, 1 \le j \le n_s)$.

M_a:

- Sample n_a sources from the total population of sources,

$$\mathbf{A}_i \overset{\text{iid}}{\sim} G(\cdot|\boldsymbol{\theta}_a), \qquad \text{for } i = 1, 2, \ldots, n_a.$$

- For each of the sampled sources, sample n_i elements from within the ith source,

$$\mathbf{Y}_{ij}|\mathbf{A}_i = \mathbf{a}_i \overset{\text{iid}}{\sim} F_a(\cdot|\mathbf{a}_i, \boldsymbol{\theta}_a) \qquad \text{for } j = 1, 2, \ldots, n_i$$

  and set

$$\mathbf{Y}_i = \begin{bmatrix} \mathbf{Y}_{i1} & \mathbf{Y}_{i2} & \cdots & \mathbf{Y}_{in_i} \end{bmatrix} = \begin{bmatrix} Y_{i,11} & Y_{i,12} & \cdots & Y_{i,1n_i} \\ Y_{i,21} & Y_{i,22} & \cdots & Y_{i,2n_i} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{i,k1} & Y_{i,k2} & \cdots & Y_{i,kn_i} \end{bmatrix}.$$

  Here, $Y_{i,jl}$ denotes the jth measurement of the lth sample from source $\mathbf{a}_i$. Since there are k measurements and n_i samples, $\mathbf{Y}_i$ is a $k \times n_i$ matrix for $i = 1, 2, \ldots, n_a$. Each measurement per sample corresponds to one feature, so this is equivalent to saying that there are n_i observations of k features.

- Let $\mathbf{y}_i = \begin{bmatrix} \mathbf{y}_{i1} & \mathbf{y}_{i2} & \cdots & \mathbf{y}_{in_i} \end{bmatrix}$ denote the realisations of the measurements on the samples from the ith source. The realisations from all background sources can be represented by the following block matrix structure:

$$\mathbf{y}_a = \left[\begin{array}{c|c|c|c} \mathbf{y}_1 & \mathbf{y}_2 & \cdots & \mathbf{y}_{n_a} \end{array}\right],$$

where each block corresponds to a sampled source. This results in a $k \times \left(\sum_{i=1}^{n_a} n_i\right)$ dimensional matrix. The background material is then given by the composed measurement vector $e_a = (\mathbf{y}_{ij}, 1 \leq i \leq n_a, 1 \leq j \leq n_i)$. Note that this sampling model is identical to the sampling model M_a in the common source problem.

M_p :

- Sample n_u elements from within the specific source,

$$\mathbf{Y}_{uj} \overset{\text{iid}}{\sim} F_s(\cdot|\boldsymbol{\theta}_s) \qquad \text{for } j = 1, 2, \ldots, n_u$$

and set

$$\mathbf{Y}_u = \begin{bmatrix} \mathbf{Y}_{u1} & \mathbf{Y}_{u2} & \cdots & \mathbf{Y}_{un_u} \end{bmatrix} = \begin{bmatrix} Y_{u,11} & Y_{u,12} & \cdots & Y_{u,1n_u} \\ Y_{u,21} & Y_{u,22} & \cdots & Y_{u,2n_u} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{u,k1} & Y_{u,k2} & \cdots & Y_{u,kn_u} \end{bmatrix}. \tag{2.12}$$

Here, $Y_{u,jl}$ denotes the jth measurement of the lth sample from the specific source. This results in a $k \times n_u$ matrix.

- Let $\mathbf{y}_u = \begin{bmatrix} \mathbf{y}_{u1} & \mathbf{y}_{u2} & \cdots & \mathbf{y}_{un_u} \end{bmatrix}$ denote the realisations of the measurements on the samples from the specific source. The material originating from the unknown source is then given by the composed measurement vector $e_u = (\mathbf{y}_{uj}, 1 \leq j \leq n_u)$.

M_d :

- Sample a single source from the total population of alternative sources,

$$\mathbf{D} \overset{\text{iid}}{\sim} G(\cdot|\boldsymbol{\theta}_a).$$

- Sample n_u elements from within source $\mathbf{D} = \mathbf{d}$,

$$\mathbf{Y}_{uj}|\mathbf{D} = \mathbf{d} \overset{\text{iid}}{\sim} F_a(\cdot|\mathbf{d}, \boldsymbol{\theta}_a) \qquad \text{for } j = 1, 2, \ldots, n_u$$

and set

$$\mathbf{Y}_u = \begin{bmatrix} \mathbf{Y}_{u1} & \mathbf{Y}_{u2} & \cdots & \mathbf{Y}_{un_u} \end{bmatrix} = \begin{bmatrix} Y_{u,11} & Y_{u,12} & \cdots & Y_{u,1n_u} \\ Y_{u,21} & Y_{u,22} & \cdots & Y_{u,2n_u} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{u,k1} & Y_{u,k2} & \cdots & Y_{u,kn_u} \end{bmatrix}. \tag{2.13}$$

Here, $Y_{u,jl}$ denotes the jth measurement of the lth sample from source $\mathbf{d}$. This results in a $k \times n_u$ matrix.

- Let $\mathbf{y}_u = \begin{bmatrix} \mathbf{y}_{u1} & \mathbf{y}_{u2} & \cdots & \mathbf{y}_{un_u} \end{bmatrix}$ denote the realisations of the measurements on the samples from source $\mathbf{d}$. The material originating from the unknown source is then given by the composed measurement vector $e_u = (\mathbf{y}_{uj}, 1 \leq j \leq n_u)$.

Note that under the prosecution model, e_u and e_s are conditionally independent given the parameters of the specific source, i.e., both evidence sets consist of randomly selected elements from the same specific source, and both e_u and e_s are unconditionally independent of e_a. However, under the defence model, e_u and e_s are unconditionally independent, just like e_a and e_s, since in both cases the evidence sets contain randomly selected elements from different sources, whereas e_u and e_a are only conditionally independent given the corresponding sources.

2.2.2. Likelihood functions

Let $e = \{e_u, e_s, e_a\}$ denote the collection of three datasets containing the available forensic evidence. In Section 2.2.1, the four sampling models for the specific source problem were explained and the resulting evidence was given in both matrix notation and as composed measurement vectors.

Since the sampling model M_a of the specific source problem is identical to the one from the common source problem, the matrix structure of $\mathbf{y}_a$ is the same as in equation (2.1) and (2.2). The sampling model M_a can be represented by equation (2.5). The derivation of the likelihood function for e_a is also equivalent to the derivation given for the common source problem, resulting in equation (2.8).

Now consider the dataset e_s, which consists of n_s random samples from the specific source. Using the same notation as in equation (2.11), $\mathbf{y}_s$ is a matrix where each column corresponds to a sample from the specific source, and each row corresponds to a measurement on a sample. Each sample can be seen as a random vector that is independent of and identically distributed to the other samples within the specific source. In this modelling stage it is assumed that $\boldsymbol{\theta}_s$ is known, either from an estimate (frequentist) or as the result of a prior distribution (Bayesian). More details about the parameter $\boldsymbol{\theta}_s$ will be given in Chapter 3. The sampling model M_s can be represented by

$$\mathbf{Y}_{sj} \overset{\text{iid}}{\sim} F_s(\cdot|\boldsymbol{\theta}_s) \qquad \text{for } j = 1, 2, \ldots, n_s.$$

Here, the n_s samples $\mathbf{Y}_{sj}$ are obtained from the probability distribution F_s corresponding to the probability measure under model M_s. Note that, in contrast to the common source problem, there is no distribution to sample the specific source since this source is assumed to be known. Let f_s denote the probability density function corresponding to F_s and let f denote the likelihood structure of the evidence. Again following [32], the likelihood functions corresponding to the specific source problem will be derived in this section. The likelihood function for e_s can be defined by

$$f(e_s|\boldsymbol{\theta}_s, H_p) = f(e_s|\boldsymbol{\theta}_s, H_d) = \prod_{j=1}^{n_s} f_s(\mathbf{y}_{sj}|\boldsymbol{\theta}_s).$$

Finally, consider the dataset with the unknown source evidence e_u, which consists of n_u random samples from the unknown source. Using the same notation as in equations (2.12) and (2.13), $\mathbf{y}_u$ is a matrix where each column corresponds to a sample from the unknown source, and each row corresponds to a measurement on a sample.

Given H_p, e_u consists of samples drawn from the specific source. Each column of $\mathbf{y}_u$ can be seen as the realisation of a random vector that is independent of and identically distributed to each of the other columns within the specific source. The sampling model M_p can be represented by

$$\mathbf{Y}_{uj} \overset{\text{iid}}{\sim} F_s(\cdot|\boldsymbol{\theta}_s) \qquad \text{for } j = 1, 2, \ldots, n_u.$$

Therefore, the likelihood function for e_u given H_p is defined as

$$f(e_u|\boldsymbol{\theta}_s, H_p) = \prod_{j=1}^{n_u} f_s(\mathbf{y}_{uj}|\boldsymbol{\theta}_s). \tag{2.14}$$

Conversely, given H_d, e_u consists of samples drawn from a randomly selected source from the total population of alternative sources. Given the source, each column of $\mathbf{y}_u$ can be seen as the realisation of a random vector that is independent of and identically distributed to each of the other columns within that source. The sampling model M_d can be represented by

$$\mathbf{D} \sim G(\cdot|\boldsymbol{\theta}_a) \qquad \text{and} \qquad \mathbf{Y}_{uj}|\mathbf{D} = \mathbf{d} \overset{\text{iid}}{\sim} F_a(\cdot|\mathbf{d}, \boldsymbol{\theta}_a) \qquad \text{for } j = 1, 2, \ldots, n_u.$$

Conditional on the selected source $\mathbf{d}$, the samples in e_u are independent. Therefore, the likelihood function for e_u given H_d is defined as

$$f(e_u|\boldsymbol{\theta}_a, H_d) = f_u(\mathbf{y}_{u1}, \mathbf{y}_{u2}, \ldots, \mathbf{y}_{un_u}|\boldsymbol{\theta}_a) = \int \prod_{j=1}^{n_u} f_a(\mathbf{y}_{uj}|\mathbf{d}, \boldsymbol{\theta}_a) g(\mathbf{d}|\boldsymbol{\theta}_a)\, d\mathbf{d}. \tag{2.15}$$

The likelihood functions obtained in equation (2.14) and (2.15) will play an important role in Chapter 3 to quantify the value of evidence for the specific source problem.

2.3. Common source versus specific source

Now that both the common and specific source problem are made precise, the question arises which problem should be considered in forensic casework. Although this choice mainly depends on the assumptions a forensic expert makes based on the context of the evidence, the differences between the two identification of source problems will be briefly discussed here.

The main difference between the common and specific source problem is whether the unknown source evidence is compared to evidence originating from either a fixed or a random source. Assuming a random source, only a single background population is under consideration where all sources are assumed to be part of. Assuming a fixed source results in two 'background' populations: one corresponding to the random sources and one related to the specific source [33]. This was already highlighted in the third modelling assumption stated in the previous section. Of course, one could argue that all evidence is generated from an overall distribution and that the specific evidence under consideration is also a realisation of a random source, which would be an argument in favor of the common source problem.

For the common source problem it is not necessary to specify a suspected source, whereas for the specific source problem the suspected source needs to be identified. The common source problem often leads to a more conservative value of evidence[1] than the specific source problem, since in the common source problem an extra level of uncertainty is considered. The court will be mostly interested in answering a specific source question, which would help provide a decision between guilt and innocence of a suspect. However, if there is not enough specific source evidence available, the corresponding specific source distribution cannot be modelled and the common source framework should be used instead [33].

[1] See next chapter for an explanation of the value of evidence. In Chapter 6 and 10 this effect is illustrated for both continuous and discrete evidence.

3 Value of evidence

This chapter contains a literature study on the value of evidence, primarily based on the work from [31–33]. The aim of this chapter is to give an overview for ease of reference and to provide some extra explanation of the underlying theory.

The decision between two competing forensic hypotheses should be made as transparent and objective as possible. Multiple people take various roles in the decision making process, each within their own field of expertise. The people involved can roughly be divided into two groups: legal experts, such as judges and jurors, and forensic experts. Each group has its own responsibility in the decision making process. The forensic experts are tasked to provide the *value of evidence*, whereas the legal experts have to determine the probability of occurrence of the two hypotheses without considering the recovered evidence.

Let e denote the entire set of available forensic evidence and I the relevant background information for both hypotheses. To decide which hypothesis is most probable after observing all evidence, the ratio

$$\frac{\mathbb{P}(H_p|e,I)}{\mathbb{P}(H_d|e,I)}$$

should be considered. This ratio is referred to as the *posterior odds* [34]. If the posterior odds are larger than one, this means that given all evidence the prosecution hypothesis is more probable than the defence hypothesis. However, neither the legal experts nor the forensic experts can determine these probabilities directly. Therefore, Bayes' theorem is used to split the posterior odds into two parts.

Theorem 1 (Bayes' theorem)**.** *Let A and B be events where $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$. Then*

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

Bayes' theorem follows from the definition of conditional probabilities. [36]

Applying Bayes' theorem and the definition of conditional probability to the posterior odds gives

$$\frac{\mathbb{P}(H_p|e,I)}{\mathbb{P}(H_d|e,I)} = \frac{\mathbb{P}(e,I|H_p)\mathbb{P}(H_p)}{\mathbb{P}(e,I)}\frac{\mathbb{P}(e,I)}{\mathbb{P}(e,I|H_d)\mathbb{P}(H_d)} = \frac{\mathbb{P}(e|H_p,I)}{\mathbb{P}(e|H_d,I)}\cdot\frac{\mathbb{P}(I|H_p)}{\mathbb{P}(I|H_d)}\cdot\frac{\mathbb{P}(H_p)}{\mathbb{P}(H_d)}.$$

Then, using Bayes' theorem once more results in the odds form of Bayes' theorem

$$\frac{\mathbb{P}(H_p|e,I)}{\mathbb{P}(H_d|e,I)} = \frac{\mathbb{P}(e|H_p,I)}{\mathbb{P}(e|H_d,I)}\cdot\frac{\mathbb{P}(H_p|I)}{\mathbb{P}(H_d|I)},$$

or in words

Posterior odds = Value of Evidence × Prior odds.

In general, explicit mention of the background information I is omitted for ease of notation [2]. The forensic experts are responsible for determining the value of evidence, whereas the legal experts have to report the *prior odds.* The prior odds summarize the personal belief regarding the validity of the prosecution and defence hypotheses before observing the evidence. This personal prior belief is then updated by the value of evidence to arrive at the posterior odds [32].

In the statistics community, there are two commonly used approaches to evaluate the value of evidence. Following [32], let e_u denote the unknown source evidence, where $e_u = \{e_{u_1}, e_{u_2}\}$ for the common source problem. The evidence is assumed to be generated according to the sampling models as discussed in Chapter 2. Frequentists will say that the value of evidence is given by the *likelihood ratio*

$$LR(\theta; e_u) = \frac{f(e_u|\theta, H_p)}{f(e_u|\theta, H_d)},$$

i.e., the ratio of the likelihood functions derived in the previous chapter, which is a function of the unknown parameter θ and e_u. To improve readability, θ is used in this chapter as if it is a one-dimensional parameter, but bear in mind that the same results hold for the parameter vector $\boldsymbol{\theta}$. Suppose that θ_0 is the true value of the parameter θ. Then $LR(\theta_0; e_u)$ denotes the 'true' likelihood ratio, which is a single point of the likelihood ratio function for $\theta = \theta_0$. In practice, the true parameter is unknown and there are many ad-hoc solutions to evaluate the likelihood ratio, such as taking some estimate of the unknown parameter based on the background material and substituting it into the likelihood ratio function [2, 32, 34].

Instead of using the likelihood ratio with some plug-in estimate of θ, it is also possible to consider a Bayesian approach. Here, the uncertainty is incorporated into the value of evidence by constructing the *Bayes Factor*. Following the notation of [31–34] the Bayes Factor is given by

$$BF(e) = \frac{\int f(e|\theta, H_p)\ d\Pi(\theta|H_p)}{\int f(e|\theta, H_d)\ d\Pi(\theta|H_d)},$$

where $\Pi(\theta)$ denotes the proper prior probability measure on the parameter space Θ, with corresponding prior density $\pi(\theta)$, when it exists. The Bayes Factor is, in contrast to the likelihood ratio, a function of the entire set of evidence and represents the ratio of the marginal likelihood of observing all evidence under the prosecution model to the marginal likelihood of observing all evidence under the defence model [32]. In the next sections, the likelihood ratio and Bayes Factor will be made precise for the common source problem and for the specific source problem.

3.1. Common source problem

Let $e = \{e_{u_1}, e_{u_2}, e_a\}$ denote the collection of the available forensic evidence. For the common source problem, the likelihood ratio becomes

$$LR_{CS}(\theta_a; e_{u_1}, e_{u_2}) = \frac{f(e_{u_1}, e_{u_2}|\theta_a, H_p)}{f(e_{u_1}, e_{u_2}|\theta_a, H_d)} = \frac{f(e_{u_1}, e_{u_2}|\theta_a, H_p)}{f(e_{u_1}|\theta_a, H_d) f(e_{u_2}|\theta_a, H_d)}.$$

Note that the likelihood ratio is a function of the unknown source evidence and θ_a only, although the parameter θ_a is based on the background material.

It is reasonable to assume that the marginal likelihood of the background material e_a is the same given both hypotheses, i.e., $f(e_a|H_p) = f(e_a|H_d)$. Under this assumption, the Bayes Factor can be written as

$$\begin{aligned}
BF_{CS}(e) &= \frac{\int f(e|\theta_a, H_p)\ d\Pi(\theta_a|H_p)}{\int f(e|\theta_a, H_d)\ d\Pi(\theta_a|H_d)} \\
&= \frac{\int f(e_{u_1}, e_{u_2}|\theta_a, H_p) f(e_a|\theta_a, H_p)\ d\Pi(\theta_a|H_p)}{\int f(e_{u_1}|\theta_a, H_d) f(e_{u_2}|\theta_a, H_d) f(e_a|\theta_a, H_d)\ d\Pi(\theta_a|H_d)} \\
&= \frac{\displaystyle\int f(e_{u_1}, e_{u_2}|\theta_a, H_p) \frac{f(e_a|\theta_a, H_p)\ d\Pi(\theta_a|H_p)}{f(e_a|H_p)}}{\displaystyle\int f(e_{u_1}|\theta_a, H_d) f(e_{u_2}|\theta_a, H_d) \frac{f(e_a|\theta_a, H_d)\ d\Pi(\theta_a|H_d)}{f(e_a|H_d)}} \qquad (3.1) \\
&= \frac{\int f(e_{u_1}, e_{u_2}|\theta_a, H_p)\ d\Pi(\theta_a|e_a, H_p)}{\int f(e_{u_1}|\theta_a, H_d) f(e_{u_2}|\theta_a, H_d)\ d\Pi(\theta_a|e_a, H_d)}
\end{aligned}$$

$$= \frac{\int f(e_{u_1}, e_{u_2} | \theta_a, H_p) \; d\Pi(\theta_a | e_a)}{\int f(e_{u_1} | \theta_a, H_d) f(e_{u_2} | \theta_a, H_d) \; d\Pi(\theta_a | e_a)}. \tag{3.2}$$

which was found in Derivation (3.7) from [32] and is included here for ease of reference. Here, in (3.1) the assumption on the marginal likelihood of e_a is used, and step (3.2) follows from the assumption that the prior belief of θ_a, given the background material, is the same according to both hypotheses. The rest of the derivation follows from standard Bayesian analysis.

3.1.1. Relation between Bayes Factor and likelihood ratio

Using the expressions found for the Bayes Factor and the likelihood ratio, it is possible to relate the two statistics [31, 32]. This might help to bring the Bayesian and the frequentist framework closer together. Moreover, the computational complexity of the Bayes Factor could be reduced for some common source problems.

Starting from equation (3.2) and using some standard Bayesian analysis, it was found in Derivation (5.1) from [32] that

$$\begin{aligned}
BF_{CS}(e) &= \frac{\int f(e_{u_1}, e_{u_2} | \theta_a, H_p) \; d\Pi(\theta_a | e_a)}{\int f(e_{u_1} | \theta_a, H_d) f(e_{u_2} | \theta_a, H_d) \; d\Pi(\theta_a | e_a)} \\
&= \frac{1}{f(e_{u_1}, e_{u_2} | e_a, H_d)} \int f(e_{u_1}, e_{u_2} | \theta_a, H_p) \; d\Pi(\theta_a | e_a) \\
&= \int \frac{f(e_{u_1}, e_{u_2} | \theta_a, H_p)}{f(e_{u_1}, e_{u_2} | e_a, H_d)} \times \frac{f(e_{u_1}, e_{u_2} | \theta_a, H_d)}{f(e_{u_1}, e_{u_2} | \theta_a, H_d)} \; d\Pi(\theta_a | e_a) \\
&= \int \frac{f(e_{u_1}, e_{u_2} | \theta_a, H_p)}{f(e_{u_1}, e_{u_2} | \theta_a, H_d)} \times \frac{f(e_{u_1}, e_{u_2} | \theta_a, H_d) \; d\Pi(\theta_a | e_a)}{f(e_{u_1}, e_{u_2} | e_a, H_d)} \\
&= \int LR_{CS}(\theta_a; e_{u_1}, e_{u_2}) \; d\Pi(\theta_a | e_{u_1}, e_{u_2}, e_a, H_d),
\end{aligned} \tag{3.3}$$

where in the second-last equation the assumption is used that the prior belief of θ_a, given the background material, is the same according to both hypotheses. This alternative expression of the Bayes Factor shows that the frequentist likelihood ratio can be adapted to the Bayesian framework by imposing a prior for θ_a given the entire set of evidence under the defence model. Furthermore, only one integral has to be evaluated to compute the Bayes Factor.

Similarly, following Derivation (5.3) from [32], for the reciprocal of the Bayes Factor in equation (3.2) it holds that

$$\begin{aligned}
\frac{1}{BF_{CS}(e)} &= \frac{1}{f(e_{u_1}, e_{u_2} | e_a, H_p)} \int f(e_{u_1}, e_{u_2} | \theta_a, H_d) \; d\Pi(\theta_a | e_a) \\
&= \int \frac{f(e_{u_1}, e_{u_2} | \theta_a, H_d)}{f(e_{u_1}, e_{u_2} | \theta_a, H_p)} \times \frac{f(e_{u_1}, e_{u_2} | \theta_a, H_p) \; d\Pi(\theta_a | e_a)}{f(e_{u_1}, e_{u_2} | e_a, H_p)} \\
&= \int \frac{1}{LR_{CS}(\theta_a; e_{u_1}, e_{u_2})} \; d\Pi(\theta_a | e_{u_1}, e_{u_2}, e_a, H_p),
\end{aligned}$$

which gives

$$BF_{CS}(e) = \left[\int \frac{1}{LR_{CS}(\theta_a; e_{u_1}, e_{u_2})} \; d\Pi(\theta_a | e_{u_1}, e_{u_2}, e_a, H_p) \right]^{-1}. \tag{3.4}$$

Note that in both expressions the prior for θ_a given the available forensic evidence explicitly depends on the hypothesis.

3.2. Specific source problem

Let $e = \{e_u, e_s, e_a\}$ denote the collection of the available forensic evidence. For the specific source problem, the likelihood ratio becomes

$$LR_{SS}(\boldsymbol{\theta}; e_u) = \frac{f(e_u | \theta_s, H_p)}{f(e_u | \theta_a, H_d)},$$

where $\boldsymbol{\theta} = (\theta_a, \theta_s)$. Again, the likelihood ratio is a function of the unknown source evidence and $\boldsymbol{\theta}$ only, although the parameters θ_a and θ_s are based on the background material and the specific source evidence, respectively.

In the formulation of the specific source problem, the specific source is not contained in the total population of alternative sources and therefore it is reasonable to assume that the prior for θ_s is independent of the prior for θ_a [32]. Under this assumption, the Bayes Factor can be written as

$$\begin{aligned}
BF_{SS}(e) &= \frac{\int f(e|\boldsymbol{\theta}, H_p)\ d\Pi(\boldsymbol{\theta}|H_p)}{\int f(e|\boldsymbol{\theta}, H_d)\ d\Pi(\boldsymbol{\theta}|H_d)} \\
&= \frac{\int f(e_u|\theta_s, H_p) f(e_s|\theta_s, H_p) f(e_a|\theta_a, H_p)\ d\Pi(\theta_a, \theta_s|H_p)}{\int f(e_u|\theta_a, H_d) f(e_s|\theta_s, H_d) f(e_a|\theta_a, H_d)\ d\Pi(\theta_a, \theta_s|H_d)} \\
&= \frac{\int f(e_u|\theta_s, H_p) f(e_s|\theta_s, H_p)\ d\Pi(\theta_s|H_p)}{\int f(e_s|\theta_s, H_d)\ d\Pi(\theta_s|H_d)} \times \frac{\int f(e_a|\theta_a, H_p)\ d\Pi(\theta_a|H_p)}{\int f(e_u|\theta_a, H_d) f(e_a|\theta_a, H_d)\ d\Pi(\theta_a|H_d)} \\
&= (1) \times (2)
\end{aligned} \tag{3.5}$$

which was found in Derivation (3.7) from [32] and is included here for ease of reference.
Under the assumption that the marginal likelihood of the background material e_a as well as of the specific source evidence is the same according to both hypotheses, i.e., $f(e_a|H_p) = f(e_a|H_d)$ and $f(e_s|H_p) = f(e_s|H_d)$, these fractions can be simplified significantly:

$$\begin{aligned}
(1) &= \frac{\int f(e_u|\theta_s, H_p) f(e_s|\theta_s, H_p)\ d\Pi(\theta_s|H_p)}{\int f(e_s|\theta_s, H_d)\ d\Pi(\theta_s|H_d)} \times \frac{f(e_s|H_d)}{f(e_s|H_p)} \\
&= \int f(e_u|\theta_s, H_p) \frac{f(e_s|\theta_s, H_p)\ d\Pi(\theta_s|H_p)}{f(e_s|H_p)} \times \frac{f(e_s|H_d)}{\int f(e_s|\theta_s, H_d)\ d\Pi(\theta_s|H_d)} \\
&= \int f(e_u|\theta_s, H_p)\ d\Pi(\theta_s|e_s, H_p) \times \frac{f(e_s|H_d)}{f(e_s|H_d)} \\
&= \int f(e_u|\theta_s, H_p)\ d\Pi(\theta_s|e_s, H_p).
\end{aligned}$$

Similarly,

$$\begin{aligned}
(2) &= \frac{\int f(e_a|\theta_a, H_p)\ d\Pi(\theta_a|H_p)}{\int f(e_u|\theta_a, H_d) f(e_a|\theta_a, H_d)\ d\Pi(\theta_a|H_d)} \times \frac{f(e_a|H_d)}{f(e_a|H_p)} \\
&= \frac{\int f(e_a|\theta_a, H_p)\ d\Pi(\theta_a|H_p)}{f(e_a|H_p)} \times \frac{f(e_a|H_d)}{\int f(e_u|\theta_a, H_d) f(e_a|\theta_a, H_d)\ d\Pi(\theta_a|H_d)} \\
&= \frac{f(e_a|H_p)}{f(e_a|H_p)} \times \left(\int f(e_u|\theta_a, H_d) \frac{f(e_a|\theta_a, H_d)\ d\Pi(\theta_a|H_d)}{f(e_a|H_d)} \right)^{-1} \\
&= \left(\int f(e_u|\theta_a, H_d)\ d\Pi(\theta_a|e_a, H_d) \right)^{-1}.
\end{aligned}$$

So for the specific source problem, the Bayes Factor reduces to

$$BF_{SS}(e) = \frac{\int f(e_u|\theta_s, H_p)\ d\Pi(\theta_s|e_s, H_p)}{\int f(e_u|\theta_a, H_d)\ d\Pi(\theta_a|e_a, H_d)} = \frac{\int f(e_u|\theta_s, H_p)\ d\Pi(\theta_s|e_s)}{\int f(e_u|\theta_a, H_d)\ d\Pi(\theta_a|e_a)}, \tag{3.6}$$

where the assumption is used that the prior belief of θ_a and θ_s, given the background material and the specific source evidence respectively, is the same according to both hypotheses. This derivation is slightly different than the one given in [32], where the notational dependence on the hypothesis is not made explicit and therefore some assumptions that follow from the sampling models are not mentioned explicitly.

3.2.1. Relation between Bayes Factor and likelihood ratio

For the specific source problem it is also possible to relate the Bayes Factor and the likelihood ratio. In the derivation given in [32], the notational dependence on the hypothesis is again not made explicit. Here, the hypotheses will be incorporated in the derivation and the required assumptions will be mentioned explicitly.

Starting from equation (3.5), it follows that

$$\begin{aligned}
BF_{SS}(e) &= \frac{\int f(e_u|\theta_s, H_p) f(e_s|\theta_s, H_p) f(e_a|\theta_a, H_p)\ d\Pi(\theta_a, \theta_s|H_p)}{\int f(e_u|\theta_a, H_d) f(e_s|\theta_s, H_d) f(e_a|\theta_a, H_d)\ d\Pi(\theta_a, \theta_s|H_d)} \\
&= \frac{1}{f(e_u, e_s, e_a|H_d)} \int f(e_u|\theta_s, H_p) f(e_s|\theta_s, H_p) f(e_a|\theta_a, H_p)\ d\Pi(\theta_a, \theta_s|H_p) \\
&= \int \frac{f(e_u|\theta_s, H_p) f(e_s|\theta_s, H_p) f(e_a|\theta_a, H_p)}{f(e_u, e_s, e_a|H_d)} \times \frac{f(e_u|\theta_a, H_d)}{f(e_u|\theta_a, H_d)}\ d\Pi(\theta_a, \theta_s|H_p) \\
&= \int \frac{f(e_u|\theta_s, H_p)}{f(e_u|\theta_a, H_d)} \times \frac{f(e_u|\theta_a, H_d) f(e_s|\theta_s, H_p) f(e_a|\theta_a, H_p)\ d\Pi(\theta_a, \theta_s|H_p)}{f(e_u, e_s, e_a|H_d)} \\
&= \int LR_{SS}(\theta_a, \theta_s; e_u) \times \frac{f(e_u|\theta_a, H_d) f(e_s|\theta_s, H_d) f(e_a|\theta_a, H_d)\ d\Pi(\theta_a, \theta_s|H_d)}{f(e_u, e_s, e_a|H_d)} \\
&= \int LR_{SS}(\theta_a, \theta_s; e_u)\ d\Pi(\theta_a, \theta_s|e_u, e_s, e_a, H_d). \qquad (3.7)
\end{aligned}$$

Here, in the second-last step it is used that the sampling models for the specific source evidence and the background material are the same according to both hypotheses, i.e., $f(e_s|\theta_s, H_p) = f(e_s|\theta_s, H_d)$ and $f(e_a|\theta_a, H_p) = f(e_a|\theta_a, H_d)$ [31, 32]. Moreover, the prior belief of θ_a and θ_s is the same given each hypothesis. Again, the likelihood ratio can be adapted to the Bayesian framework by imposing a prior for (θ_s, θ_a) given the entire set of evidence under the defence model. However, the computational complexity of the Bayes Factor does not reduce, since still two integrals have to be evaluated.

Following Derivation (5.6) from [32], for the reciprocal of the Bayes Factor from equation (3.5) it holds that

$$\begin{aligned}
\frac{1}{BF_{SS}(e)} &= \frac{1}{f(e_u, e_s, e_a|H_p)} \int f(e_u|\theta_a, H_d) f(e_s|\theta_s, H_d) f(e_a|\theta_a, H_d)\ d\Pi(\theta_a, \theta_s|H_d) \\
&= \int \frac{f(e_u|\theta_a, H_d) f(e_s|\theta_s, H_d) f(e_a|\theta_a, H_d)}{f(e_u, e_s, e_a|H_p)} \times \frac{f(e_u|\theta_s, H_p)}{f(e_u|\theta_s, H_p)}\ d\Pi(\theta_a, \theta_s|H_d) \\
&= \int \frac{f(e_u|\theta_a, H_d)}{f(e_u|\theta_s, H_p)} \times \frac{f(e_u|\theta_s, H_p) f(e_s|\theta_s, H_d) f(e_a|\theta_a, H_d)\ d\Pi(\theta_a, \theta_s|H_d)}{f(e_u, e_s, e_a|H_p)} \\
&= \int \frac{1}{LR_{SS}(\theta_a, \theta_s, e_u)} \times \frac{f(e_u|\theta_s, H_p) f(e_s|\theta_s, H_p) f(e_a|\theta_a, H_p)\ d\Pi(\theta_a, \theta_s|H_p)}{f(e_u, e_s, e_a|H_p)} \\
&= \int \frac{1}{LR_{SS}(\theta_a, \theta_s; e_u)}\ d\Pi(\theta_a, \theta_s|e_u, e_s, e_a, H_p),
\end{aligned}$$

so that

$$BF_{SS}(e) = \left[\int \frac{1}{LR_{SS}(\theta_a, \theta_s; e_u)}\ d\Pi(\theta_a, \theta_s|e_u, e_s, e_a, H_p) \right]^{-1}. \qquad (3.8)$$

Note that in both expressions the prior for (θ_a, θ_s) given the available forensic evidence explicitly depends on the hypothesis.

In summary, for both the common and specific source problem, the Bayes Factor and the likelihood ratio can be related by the following general expressions:

$$BF(e) = \int LR(\theta; e_u)\ d\Pi(\theta|e, H_d) \qquad \text{and} \qquad BF(e) = \left[\int \frac{1}{LR(\theta; e_u)}\ d\Pi(\theta|e, H_p) \right]^{-1}.$$

These relations will play an important role in Chapters 5, 10 and 11, where two commonly used models in forensic science will be considered.

3.3. Interpreting the value of evidence

In the previous sections, different methods to obtain the value of evidence were explained. This numerical value indicates how many times more probable the evidence is if the prosecution hypothesis is true compared to the defence hypothesis. Such a quantitative expression can be hard to interpret for some people. Moreover, people tend to understand the same verbal probability expression differently which can easily lead to miscommunication. A well-known example is the *prosecutor's fallacy*: the probability of observing the evidence given the hypothesis is wrongly interpreted as the probability of the hypothesis given the evidence [45]. Another example is given by the *weak evidence* or *boomerang effect*: weak evidence supporting a prosecution hypothesis is sometimes wrongly interpreted as evidence supporting the defence hypothesis [27].

As explained before, a lot of people are involved in the decision making process regarding two competing forensic hypotheses. Especially legal experts might be less familiar with interpreting the value of evidence than forensic experts. Since clear communication is crucial for an objective and fair lawsuit, it is important to provide a connection between numerical and verbal values of evidence. To this end, the Association of Forensic Science Providers have proposed a scale for the translation of numerical values of evidence into verbal formats [5], which can be found in Table 3.1.

Value of evidence	Verbal equivalent
$>1-10$	Weak support for proposition
$10-100$	Moderate support
$100-1,000$	Moderately strong support
$1,000-10,000$	Strong support
$10,000-1,000,000$	Very strong support
$>1,000,000$	Extremely strong support

Table 3.1: Standards for numerical and verbal expression of the value of evidence.

Now suppose that glass fragments are found on the clothes of a suspect and one is interested if these fragments originate from the broken window at a crime scene. In the quantitative framework, a forensic expert might give the following statement in court:

"In my opinion, the correspondence between the glass fragments found on the clothes of the accused and fragments from the window at the crime scene is 5,900 times more likely if the fragments originated from the same window than if the fragments originated from different windows."

Translating this to the verbal format would result in:

"... the correspondence between the glass fragments found on the clothes of the accused and fragments from the window at the crime scene offers strong support to the proposition that the fragments originated from the same window ..."

Although this research will mainly consider the numerical value of evidence, it is important to keep in mind that in practice also the verbal expression of support is used.

4

Two-level normal-normal model

One commonly used model in forensic science is the *two-level normal-normal model.* In this model, both the between-source variation and the within-source variation of the evidence are assumed to follow a (multivariate) normal distribution. The two-level normal-normal model is briefly discussed in [32], but only applied in a Markov Chain Monte Carlo simulation study. Since conjugate priors are proposed, this chapter will consider the possibilities of analytical calculation of the Bayes Factor for both the common source and the specific source problem.

Adapting the notation from the definition of the common and specific source problem, the two-level normal-normal model can be used as sampling model for the forensic evidence. This model is also known as the hierarchical *simple random effects model.* As before, let $\mathbf{Y}_{ij}$ denote the k-dimensional column vector of measurements on the jth sample from the ith source for $j = 1, 2, \ldots, n_i$ and $i = 1, 2, \ldots, n_a$. Then the two-level normal-normal model is obtained by setting $\boldsymbol{\theta}_a = (\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a, \boldsymbol{\Sigma}_w)$ and

$$\mathbf{A}_i \overset{\text{iid}}{\sim} \mathcal{N}_k\left(\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a\right) \qquad \text{and} \qquad \mathbf{Y}_{ij}|\mathbf{A}_i = \mathbf{a}_i \overset{\text{iid}}{\sim} \mathcal{N}_k\left(\mathbf{a}_i, \boldsymbol{\Sigma}_w\right) \qquad \text{for } i = 1, 2, \ldots, n_a \text{ and } j = 1, 2, \ldots, n_i. \tag{4.1}$$

Here, $\boldsymbol{\mu}_a$ denotes a k-dimensional column vector, where for each $i = 1, 2, \ldots, k$, $(\boldsymbol{\mu}_a)_i$ denotes the overall mean of the ith measurement on every sample from all sources. This is equivalent to saying that $(\boldsymbol{\mu}_a)_i$ is the overall mean of the ith feature from all observations of this feature. Then $\mathbf{a}_i$ denotes the k-dimensional column vector that is the realisation of the random vector $\mathbf{A}_i$. The matrices $\boldsymbol{\Sigma}_a$ and $\boldsymbol{\Sigma}_w$ denote the covariance matrices corresponding to the total population of sources and the samples within the sources, respectively.

Equivalently, this model can be represented as simple random effects model by writing

$$\mathbf{Y}_{ij} = \mathbf{A}_i + \mathbf{W}_{ij} = \boldsymbol{\mu}_a + \mathbf{B}_i + \mathbf{W}_{ij} \qquad \text{for } i = 1, 2, \ldots, n_a \text{ and } j = 1, 2, \ldots, n_i,$$

where $\mathbf{B}_i \overset{\text{iid}}{\sim} \mathcal{N}_k(\mathbf{0}_k, \boldsymbol{\Sigma}_a)$, $\mathbf{W}_{ij} \overset{\text{iid}}{\sim} \mathcal{N}_k(\mathbf{0}_k, \boldsymbol{\Sigma}_w)$, and $\mathbf{0}_k$ denotes the k-dimensional column vector with all elements equal to zero.

For the common source problem, all evidence sets are assumed to follow this two-level model as explained in Section 2.1. However, for the specific source problem, only the evidence sets obtained by sampling models M_a and M_d follow a two-level model. The other two sampling models, M_s and M_p, assume that the evidence is sampled from a specific and known source, which is not randomly selected from the total population of alternative sources and therefore no sampling distribution has to be specified for the between-source distribution. This means that for the sampling models M_s and M_p, $\boldsymbol{\theta}_s = (\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s)$ and only a one-level normal model is needed:

$$\mathbf{Y}_{sj} \overset{\text{iid}}{\sim} \mathcal{N}_k\left(\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s\right) \qquad \text{for } j = 1, 2, \ldots, n_s, \tag{4.2}$$

where $\mathbf{Y}_{sj}$ denotes the k-dimensional column vector of measurements on the jth sample from the specific source.

A fully Bayesian approach is obtained by specifying priors for all elements of $\boldsymbol{\theta}_a$. For the common source problem and for the sampling models M_a and M_d of the specific source problem, the two-level normal-normal model is completely defined by (4.1) with priors for $\boldsymbol{\theta}_a = (\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a, \boldsymbol{\Sigma}_w)$ given by

$$\boldsymbol{\mu}_a \sim \mathcal{N}_k(\boldsymbol{\mu}_\pi, \lambda\boldsymbol{\Sigma}_b), \qquad \boldsymbol{\Sigma}_a \sim \mathcal{W}_k^{-1}(\boldsymbol{\Sigma}_b, \nu_b), \qquad \text{and} \qquad \boldsymbol{\Sigma}_w \sim \mathcal{W}_k^{-1}(\boldsymbol{\Sigma}_e, \nu_e)$$

as proposed in [32]. Here, λ is a scalar and $\mathcal{W}_k^{-1}$ denotes the inverse Wishart distribution. For the sampling models M_s and M_p of the specific source problem, the one-level normal model is completely defined by (4.2) with priors for $\boldsymbol{\theta}_s = (\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s)$ given by

$$\boldsymbol{\mu}_s \sim \mathcal{N}_k(\boldsymbol{\mu}_\pi, \boldsymbol{\Sigma}_b) \qquad \text{and} \qquad \boldsymbol{\Sigma}_s \sim \mathcal{W}_k^{-1}(\boldsymbol{\Sigma}_e, \nu_e)$$

as proposed by [32]. The argumentation for choosing the multivariate normal distribution and the inverse Wishart distribution as priors for respectively the mean and covariance of the multivariate normal distribution given in [32] is that they are conjugate priors for the multivariate normal distribution (see also [16]). The hyperparameters of these prior distributions are estimated from the evidence.

Usually, this conjugacy has a lot of advantages, since the posterior is given by a known distribution and therefore an analytical solution of the Bayes Factor should exist. However, in the next section it will be shown that in this formulation the conjugacy does not hold and no analytical solution of the Bayes Factor can be derived.

4.1. Common source Bayes Factor

For simplicity, suppose only one feature is measured for the common source problem. Then the two-level normal-normal model reduces to a one-dimensional problem, which is schematically given in Figure 4.1.

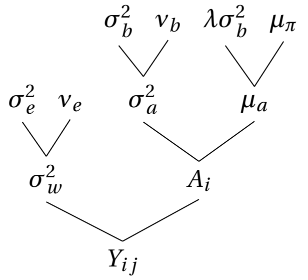

Figure 4.1: Common source hierarchical structure.

The integration problem can be summarized by

$$Y_{ij}|a_i, \sigma_w^2 \overset{\text{iid}}{\sim} N\left(a_i, \sigma_w^2\right) \tag{4.3}$$

$$A_i|\mu_a, \sigma_a^2 \overset{\text{iid}}{\sim} N\left(\mu_a, \sigma_a^2\right) \tag{4.4}$$

$$\mu_a \sim N\left(\mu_\pi, \lambda\sigma_b^2\right) \tag{4.5}$$

$$\sigma_a^2 \sim \text{Scale-inv-}\chi^2\left(\nu_b, \sigma_b^2\right) \tag{4.6}$$

$$\sigma_w^2 \sim \text{Scale-inv-}\chi^2\left(\nu_e, \sigma_e^2\right), \tag{4.7}$$

where the last two distributions are the univariate specialisation of the inverse Wishart distribution [16].

In Section 3.1, an explicit expression of the Bayes Factor for the common source problem was given. Since $\pi(\boldsymbol{\theta}_a|e_a) = \pi(\mu_a, \sigma_a^2, \sigma_w^2|e_a)$ is difficult to compute for the two-level normal-normal model, the expression for the Bayes Factor as given in equation (3.2) cannot be used directly. Therefore, the original definition of the Bayes Factor

$$BF_{CS}(e) = \frac{\int f(e_{u_1}, e_{u_2}|\boldsymbol{\theta}_a, H_p) f(e_a|\boldsymbol{\theta}_a, H_p)\, d\Pi(\boldsymbol{\theta}_a)}{\int f(e_{u_1}|\boldsymbol{\theta}_a, H_d) f(e_{u_2}|\boldsymbol{\theta}_a, H_d) f(e_a|\boldsymbol{\theta}_a, H_d)\, d\Pi(\boldsymbol{\theta}_a)}$$

will be evaluated.

First, consider the numerator of the Bayes Factor. Using the likelihood found in Section 2.1.2, this integral can be written as

$$\int \left(\int \prod_{j=1}^{n_u} f_a(y_{uj}|p, \boldsymbol{\theta}_a) g(p|\boldsymbol{\theta}_a)\, dp \right) \left(\prod_{i=1}^{n_a} \int \prod_{j=1}^{n_i} f_a(y_{ij}|a_i, \boldsymbol{\theta}_a) g(a_i|\boldsymbol{\theta}_a)\, da_i \right) d\Pi(\boldsymbol{\theta}_a). \tag{4.8}$$

Since p is sampled similarly as all the a_i, the unknown source evidence can be merged with the background material to simplify this expression to

$$\int \left(\prod_{i=1}^{n_a+1} \int \prod_{j=1}^{n_i} f_a(y_{ij}|a_i, \boldsymbol{\theta}_a) g(a_i|\boldsymbol{\theta}_a)\, da_i \right) d\Pi(\boldsymbol{\theta}_a),$$

where $y_{(n_a+1)j} := y_{uj}$ for all j, $n_{n_a+1} := n_u$ and $a_{n_a+1} := p$. Now applying the two-level normal-normal model, the integral becomes

$$\int_0^\infty \int_0^\infty \int_{-\infty}^\infty \left(\prod_{i=1}^{n_a+1} \int_{-\infty}^\infty \prod_{j=1}^{n_i} f_a(y_{ij}|a_i,\sigma_w^2) g(a_i|\mu_a,\sigma_a^2)\ da_i \right) \pi(\mu_a)\pi(\sigma_a^2)\pi(\sigma_w^2)\ d\mu_a\ d\sigma_a^2\ d\sigma_w^2. \tag{4.9}$$

The integral between the brackets in equation (4.9) will be evaluated first. From **Derivation 1** in Appendix A it follows that

$$\prod_{i=1}^{n_a+1} \int_{-\infty}^\infty \prod_{j=1}^{n_i} f_a(y_{ij}|a_i,\sigma_w^2) g(a_i|\mu_a,\sigma_a^2)\ da_i =$$

$$\prod_{i=1}^{n_a+1} \left(2\pi\sigma_w^2\right)^{-n_i/2} \left(\frac{\sigma_w^2}{n_i\sigma_a^2 + \sigma_w^2} \right)^{1/2} \exp\left[-\frac{\sum_{j=1}^{n_i} y_{ij}^2}{2\sigma_w^2} - \frac{\mu_a^2}{2\sigma_a^2} \right] \exp\left[\frac{1}{2} \frac{\left(\sigma_a^2 \sum_{j=1}^{n_i} y_{ij} + \mu_a\sigma_w^2\right)^2}{\sigma_a^2\sigma_w^2\left(n_i\sigma_a^2 + \sigma_w^2\right)} \right]$$

Clearly, this cannot be expressed as a normal distribution with mean μ_a and variance σ_a^2 or σ_w^2. Therefore, the choice for the prior distributions seems to be rather arbitrary and one could argue that using a non-informative prior would make more sense [4]. Similar problems arise when considering the denominator of the Bayes Factor for the common source problem. Note that in this definition of the model it is not allowed to change the order of integration to

$$\int_0^\infty \int_{-\infty}^\infty g(a_i|\mu_a,\sigma_a^2)\pi(\mu_a)\pi(\sigma_a^2)\ d\mu_a\ d\sigma_a^2,$$

which would lead to at least semi-conjugacy in the first level (see Section 3.4 in [16]).

4.2. Specific source Bayes Factor

Again, suppose only one feature is measured, but now consider the specific source problem. Then the two-level normal-normal model reduces to a one-dimensional problem for the specific source evidence. Given the defence hypothesis, the model for the unknown source evidence can still be represented by Figure 4.1. However, given the prosecution hypothesis the model for the unknown source evidence simplifies significantly and is schematically given in Figure 4.2.

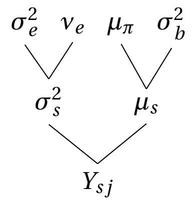

Figure 4.2: Specific source hierarchical structure.

This integration problem for the specific source evidence can be summarized by

$$Y_{sj}|\mu_s,\sigma_s^2 \overset{\text{iid}}{\sim} N\left(\mu_s,\sigma_s^2\right) \tag{4.10}$$

$$\mu_s \sim N\left(\mu_\pi,\sigma_b^2\right) \tag{4.11}$$

$$\sigma_s^2 \sim \text{Scale-inv-}\chi^2\left(\nu_e,\sigma_e^2\right), \tag{4.12}$$

where the last distribution is the univariate specialisation of the inverse Wishart distribution [16].

In Section 3.2, an explicit expression of the Bayes Factor for the specific source problem was given. Again, since both $\pi(\boldsymbol{\theta}_a|e_a) = \pi(\mu_a,\sigma_a^2,\sigma_w^2|e_a)$ and $\pi(\boldsymbol{\theta}_s|e_s) = \pi(\mu_s,\sigma_s^2|e_s)$ are difficult to compute, the expression for the Bayes Factor as given in equation (3.6) cannot be used directly. Therefore, the original definition of the Bayes Factor will be used, i.e.,

$$BF_{SS}(e) = \frac{\int f(e_u|\boldsymbol{\theta}_s,H_p) f(e_s|\boldsymbol{\theta}_s,H_p) f(e_a|\boldsymbol{\theta}_a,H_p)\ d\Pi(\boldsymbol{\theta}_a,\boldsymbol{\theta}_s)}{\int f(e_u|\boldsymbol{\theta}_a,H_d) f(e_s|\boldsymbol{\theta}_s,H_d) f(e_a|\boldsymbol{\theta}_a,H_d)\ d\Pi(\boldsymbol{\theta}_a,\boldsymbol{\theta}_s)}.$$

Using the likelihood as found in Section 2.2.2, the numerator of the Bayes Factor can be written as

$$\int \prod_{j=1}^{n_u} f_s(y_{uj}|\boldsymbol{\theta}_s) \prod_{j=1}^{n_s} f_s(y_{sj}|\boldsymbol{\theta}_s)\ d\Pi(\boldsymbol{\theta}_s) \times \int \left(\prod_{i=1}^{n_a} \int \prod_{j=1}^{n_i} f_a(y_{ij}|a_i,\boldsymbol{\theta}_a) g(a_i|\boldsymbol{\theta}_a)\ da_i \right) d\Pi(\boldsymbol{\theta}_a),$$

which can be simplified to

$$\int \prod_{j=1}^{n_u+n_s} f_s(y_{tj}|\boldsymbol{\theta}_s)\ d\Pi(\boldsymbol{\theta}_s) \times \int \left(\prod_{i=1}^{n_a} \int \prod_{j=1}^{n_i} f_a(y_{ij}|a_i,\boldsymbol{\theta}_a) g(a_i|\boldsymbol{\theta}_a)\ da_i \right) d\Pi(\boldsymbol{\theta}_a)$$

by setting $y_t = (y_{u1}, \ldots, y_{un_u}, y_{s1}, \ldots, y_{sn_s})$. Note that the integration with respect to $\Pi(\boldsymbol{\theta}_a)$ was already considered in the previous section, where it was found that conjugacy did not hold. Therefore, only the integration with respect to $\Pi(\boldsymbol{\theta}_s)$ will be considered here. Applying the two-level normal-normal model, the integral becomes

$$\int_0^\infty \int_{-\infty}^\infty \prod_{j=1}^{n_u+n_s} f_s(y_{tj}|\mu_s, \sigma_s^2)\pi(\mu_s)\pi(\sigma_s^2)\, d\mu_s\, d\sigma_s^2. \tag{4.13}$$

Now the priors specified for this problem are semi-conjugate [16]. The marginal likelihood can only be computed analytically if the prior for μ_s is adjusted to

$$\mu_s|\sigma_s^2 \sim N\left(\mu_\pi, \frac{\sigma_s^2}{\kappa_\pi}\right)$$

for some scalar κ_π, so that conjugacy is achieved. Then it follows from **Derivation 2** in Appendix A that (4.13) is equal to

$$\frac{\Gamma\left(\frac{\nu_e+n_u+n_s}{2}\right)}{\Gamma\left(\frac{\nu_e}{2}\right)} \sqrt{\frac{\kappa_\pi}{\kappa_\pi + n_u + n_s}} \frac{(\nu_e\sigma_e^2)^{\nu_e/2}}{\left((\nu_e + n_u + n_s)\sigma_n^2\right)^{(\nu_e+n_u+n_s)/2}} \frac{1}{\pi^{(n_u+n_s)/2}},$$

where

$$\sigma_n^2 = \frac{1}{\nu_e + n_u + n_s}\left(\nu_e\sigma_e^2 + \sum_{j=1}^{n_u+n_s}(y_{tj} - \bar{y}_t)^2 + \frac{(n_u + n_s)\kappa_\pi}{\kappa_\pi + n_u + n_s}(\mu_\pi - \bar{y}_t)^2\right) \qquad \text{and} \qquad \bar{y}_t = \frac{1}{n_u + n_s}\sum_{j=1}^{n_u+n_s} y_{tj}.$$

The denominator of the Bayes Factor is of similar form as the numerator and is given by

$$\int \prod_{j=1}^{n_s} f_s(y_{sj}|\boldsymbol{\theta}_s)\, d\Pi(\boldsymbol{\theta}_s) \times \int \left(\int \prod_{j=1}^{n_u} f_a(y_{uj}|p, \boldsymbol{\theta}_a) g(p|\boldsymbol{\theta}_a)\, dp\right)\left(\prod_{i=1}^{n_a}\int \prod_{j=1}^{n_i} f_a(y_{ij}|a_i, \boldsymbol{\theta}_a) g(a_i|\boldsymbol{\theta}_a)\, da_i\right) d\Pi(\boldsymbol{\theta}_a).$$

Again, the prior density functions corresponding to $\Pi(\boldsymbol{\theta}_s)$ are semi-conjugate and can be adjusted to achieve conjugacy, but the integration with respect to $\Pi(\boldsymbol{\theta}_a)$ does not result in an analytical solution. Therefore, the specific source Bayes Factor cannot be computed analytically and the choice for the prior distributions is questionable.

5 Convergence of the Gibbs sampler

Since the Bayes Factor of the two-level normal-normal model cannot be computed analytically, as seen in Chapter 4, alternative methods have to be considered. This chapter will describe how the two-level normal-normal model can be put into practice and considers the theoretical convergence properties of the proposed method. The details of the study should be sufficient to reproduce this procedure for other models following the discussed common source or specific source framework. In the next chapter, the actual results of applying the two-level normal-normal model will be discussed.

For the computation of the Bayes Factor it was chosen to use Monte Carlo integration combined with Gibbs sampling, as suggested in [32]. To apply these methods, one of the relationships between the likelihood ratio and the Bayes Factor as given in Chapter 3 is used:

$$BF(e) = \int LR(\theta; e_u)\ d\Pi(\theta|e, H_d) = \mathbb{E}_{\Theta|e,H_d}\left[LR(\Theta; e_u)\right].$$

Here, the subscript notation indicates under which probability measure the expectation is calculated. Note that the likelihood ratio as well as the probability measure $\Pi(\theta|e, H_d)$ are different for the common and specific source problem.

In classical *Monte Carlo integration* the generic problem of evaluating an integral of the form

$$\mathbb{E}_f[h(X)] = \int h(x)f(x)\ dx$$

is considered. Using a sample $\left(X^{(1)}, X^{(2)}, \ldots, X^{(m)}\right)$ generated from the density f, this integral can be approximated by the empirical average

$$\hat{h}_m = \frac{1}{m}\sum_{i=1}^{m} h(x^{(i)})$$

since $\hat{h}_m$ converges almost surely to $\mathbb{E}_f[h(X)]$ by the Strong Law of Large Numbers, provided the expectation is finite [37]. Applying this to the two-level normal-normal model, the Bayes Factor can be approximated by

$$\widehat{BF}(e) = \frac{1}{m}\sum_{i=1}^{m} LR(\theta^{(i)}; e_u)$$

where $\left(\Theta^{(1)}, \Theta^{(2)}, \ldots, \Theta^{(m)}\right)$ is a sample from $\pi(\theta|e, H_d)$. The problem here is that the density $\pi(\theta|e, H_d)$ is not known. However, *Gibbs sampling* can be incorporated to obtain the sample.

As explained in [38], the general multistage Gibbs sampler works as follows. Suppose that a sample is needed from $p(\mathbf{x})$, where $\mathbf{X} = (X_1, \ldots, X_h)$ is a random vector for some $h > 1$. Moreover, suppose that the corresponding conditional densities $p_1, \ldots, p_h$ are known, where p_i denotes the conditional density of X_i given all X_j's except for X_i, so that it is possible to simulate

$$X_i|x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_h \sim p_i(x_i|x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_h) \qquad \text{for } i = 1, 2, \ldots, h.$$

The densities $p_1, \dots, p_h$ are called the *full conditionals*. The associated Gibbs sampler is given by the transition from $\mathbf{X}^{(t)}$ to $\mathbf{X}^{(t+1)}$ as explained in Algorithm 1. In the next sections, the full conditionals and the approximation procedure will be made precise for the common source and the specific source problem in the two-level normal-normal setting. In this chapter, only one-dimensional problems are considered. Chapter 6 will explain how the approximation procedure can be adapted for higher dimensional problems.

Algorithm 1: The general Multistage Gibbs Sampler

```
for iteration $t = 1, 2, \dots,$ given $\mathbf{x}^{(t)} = (x_1^{(t)}, \dots, x_h^{(t)})$ do
    Generate $X_1^{(t+1)} \sim p_1(x_1 | x_2^{(t)}, \dots, x_h^{(t)})$
    Generate $X_2^{(t+1)} \sim p_2(x_2 | x_1^{(t+1)}, x_3^{(t)}, \dots, x_h^{(t)})$
    $\vdots$
    Generate $X_h^{(t+1)} \sim p_h(x_h | x_1^{(t+1)}, x_2^{(t+1)}, \dots, x_{h-1}^{(t+1)})$
end
```

5.1. Common source problem

To approximate the Bayes Factor, both the likelihood ratio and the full conditionals for the Gibbs sampler corresponding to the two-level normal-normal common source problem need to be derived. The likelihood ratio for the common source problem was given in Chapter 3:

$$
\begin{aligned}
LR_{CS}(\boldsymbol{\theta}_a; e_{u_1}, e_{u_2}) &= \frac{f(e_{u_1}, e_{u_2} | \boldsymbol{\theta}_a, H_p)}{f(e_{u_1} | \boldsymbol{\theta}_a, H_d) f(e_{u_2} | \boldsymbol{\theta}_a, H_d)} \\
&= \frac{\int \prod_{j=1}^{n_{u_1}+n_{u_2}} f_a(y_{uj} | p, \boldsymbol{\theta}_a) g(p | \boldsymbol{\theta}_a) \, \mathrm{d}p}{\left(\int \prod_{j=1}^{n_{u_1}} f_a(y_{u_1 j} | d_1, \boldsymbol{\theta}_a) g(d_1 | \boldsymbol{\theta}_a) \, \mathrm{d}d_1 \right) \left(\int \prod_{j=1}^{n_{u_2}} f_a(y_{u_2 j} | d_2, \boldsymbol{\theta}_a) g(d_2 | \boldsymbol{\theta}_a) \, \mathrm{d}d_2 \right)},
\end{aligned}
$$

where $\boldsymbol{\theta}_a = (\mu_a, \sigma_a^2, \sigma_w^2)$ and $\mathbf{y}_u = (y_{u_1 1}, y_{u_1 2} \dots, y_{u_1 n_{u_1}}, y_{u_2 1}, y_{u_2 2}, \dots, y_{u_2 n_{u_2}})$. These likelihoods can be calculated explicitly and from **Derivation 1** in Appendix A it follows that

$$
\begin{aligned}
LR_{CS}(\boldsymbol{\theta}_a; e_{u_1}, e_{u_2}) = {} & \left(\frac{\sigma_a^2 \left(n_{u_1} \sigma_a^2 + \sigma_w^2 \right) \left(n_{u_2} \sigma_a^2 + \sigma_w^2 \right)}{(n_{u_1} + n_{u_2}) \sigma_a^2 + \sigma_w^2} \right)^{\frac{1}{2}} \exp\left[\frac{\mu_a^2}{2\sigma_a^2} \right] \exp\left[\frac{1}{2} \frac{\left(\sigma_a^2 \sum_j y_{uj} + \mu_a \sigma_w^2 \right)^2}{\sigma_a^2 \sigma_w^2 \left((n_{u_1} + n_{u_2}) \sigma_a^2 + \sigma_w^2 \right)} \right] \\
& \times \exp\left[-\frac{1}{2} \frac{\left(\sigma_a^2 \sum_j y_{u_1 j} + \mu_a \sigma_w^2 \right)^2}{\sigma_a^2 \sigma_w^2 \left(n_{u_1} \sigma_a^2 + \sigma_w^2 \right)} \right] \exp\left[-\frac{1}{2} \frac{\left(\sigma_a^2 \sum_j y_{u_2 j} + \mu_a \sigma_w^2 \right)^2}{\sigma_a^2 \sigma_w^2 \left(n_{u_2} \sigma_a^2 + \sigma_w^2 \right)} \right]. \qquad (5.1)
\end{aligned}
$$

This is the same likelihood ratio as for example given in [7], which has been proven in [22].

The full conditionals for the Gibbs sampler can be found from

$$
\begin{aligned}
\pi(a_i | \mu_a, \sigma_a^2, \sigma_w^2, \mathbf{y}, H_d) &\propto \pi(a_i | \mu_a, \sigma_a^2) \prod_{j=1}^{n_i} f_a(y_{ij} | a_i, \sigma_w^2) \qquad \text{for } i = 1, 2, \dots, n_a + 2 \\
\pi(\mu_a | a_1, a_2, \dots, a_{n_a+2}, \sigma_a^2, \sigma_w^2, \mathbf{y}, H_d) &\propto \pi(\mu_a) \prod_{i=1}^{n_a+2} g(a_i | \mu_a, \sigma_a^2) \\
\pi(\sigma_a^2 | a_1, a_2, \dots, a_{n_a+2}, \mu_a, \sigma_w^2, \mathbf{y}, H_d) &\propto \pi(\sigma_a^2) \prod_{i=1}^{n_a+2} g(a_i | \mu_a, \sigma_a^2) \qquad (5.2) \\
\pi(\sigma_w^2 | a_1, a_2, \dots, a_{n_a+2}, \mu_a, \sigma_a^2, \mathbf{y}, H_d) &\propto \pi(\sigma_w^2) \prod_{i=1}^{n_a+2} \prod_{j=1}^{n_i} f_a(y_{ij} | a_i, \sigma_w^2)
\end{aligned}
$$

where $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_{n_a}, \mathbf{y}_{u_1}, \mathbf{y}_{u_2})$, $a_{n_a+1} = d_1$ and $a_{n_a+2} = d_2$, since in the defence model the unknown source evidence comes from two different sources. Note that the latent variables a_i have to be incorporated into the full conditionals. It follows from (5.2) that

$$
A_i | \mu_a, \sigma_a^2, \sigma_w^2, \mathbf{y}, H_d \sim N\left(\frac{\sigma_a^2 \sigma_w^2}{\sigma_w^2 + n_i \sigma_a^2} \left[\frac{1}{\sigma_w^2} \sum_{j=1}^{n_i} y_{ij} + \frac{\mu_a}{\sigma_a^2} \right], \frac{\sigma_a^2 \sigma_w^2}{\sigma_w^2 + n_i \sigma_a^2} \right)
$$

$$\begin{aligned}
\mu_a|a_1,a_2,\dots,a_{n_a+2},\sigma_a^2,\sigma_w^2,\mathbf{y},H_d &\sim N\left(\frac{\lambda\sigma_b^2\sigma_a^2}{\sigma_a^2+(n_a+2)\lambda\sigma_b^2}\left[\frac{1}{\sigma_a^2}\sum_{i=1}^{n_a+2}a_i+\frac{\mu_\pi}{\lambda\sigma_b^2}\right],\frac{\lambda\sigma_b^2\sigma_a^2}{\sigma_a^2+(n_a+2)\lambda\sigma_b^2}\right)\\
\sigma_a^2|a_1,a_2,\dots,a_{n_a+2},\mu_a,\sigma_w^2,\mathbf{y},H_d &\sim \text{Scale-inv-}\chi^2\left(\nu_b+n_a+2,\frac{\nu_b\sigma_b^2+\sum_{i=1}^{n_a+2}(a_i-\mu_a)^2}{\nu_b+n_a+2}\right)\\
\sigma_w^2|a_1,a_2,\dots,a_{n_a+2},\mu_a,\sigma_a^2,\mathbf{y},H_d &\sim \text{Scale-inv-}\chi^2\left(\nu_e+\sum_{i=1}^{n_a+2}n_i,\frac{\nu_e\sigma_e^2+\sum_{i=1}^{n_a+2}\sum_{j=1}^{n_i}(y_{ij}-a_i)^2}{\nu_e+\sum_{i=1}^{n_a+2}n_i}\right).
\end{aligned}$$

So although the priors for the two-level normal-normal model as proposed in [32] do not result in an analytical expression of the Bayes Factor, they are conveniently chosen to easily derive the full conditionals for the Gibbs sampler. Since the values of σ_b^2, σ_e^2 and μ_π are unknown, in practice the estimates $\hat{\sigma}_b^2$, $\hat{\sigma}_e^2$ and $\hat{\mu}_\pi$ will be plugged into the full conditionals. The total approximation procedure of the common source Bayes Factor is described in the next algorithm:

Algorithm 2: Common source approximation of Bayes Factor

```
Data: $e_{u_1}, e_{u_2}, e_a$
Result: Approximation of common source Bayes Factor
Initialization
    $n_{\text{gibbs}} = 20{,}000, \quad n_{\text{burn}} = 5{,}000, \quad \mu_a^{(1)} = 1, \quad \sigma_a^{2\,(1)} = 1, \quad \sigma_w^{2\,(1)} = 1, \quad a_i^{(1)} = 1$ for $i = 1,2,\dots,n_a+2$
Approximate hyperparameters $\hat{\mu}_\pi$, $\hat{\sigma}_b^2$ and $\hat{\sigma}_e^2$
for iteration $t = 1,2,\dots,n_{gibbs}$ do
    Generate $a_i^{(t+1)} \sim \pi(a_i|\mu_a^{(t)},\sigma_a^{2\,(t)},\sigma_w^{2\,(t)},\mathbf{y},H_d)$ for $i = 1,2,\dots,n_a+2$
    Generate $\mu_a^{(t+1)} \sim \pi(\mu_a|a_1^{(t+1)},a_2^{(t+1)},\dots,a_{n_a+2}^{(t+1)},\sigma_a^{2\,(t)},\sigma_w^{2\,(t)},\mathbf{y},H_d)$
    Generate $\sigma_a^{2\,(t+1)} \sim \pi(\sigma_a^2|a_1^{(t+1)},a_2^{(t+1)},\dots,a_{n_a+2}^{(t+1)},\mu_a^{(t+1)},\sigma_w^{2\,(t)},\mathbf{y},H_d)$
    Generate $\sigma_w^{2\,(t+1)} \sim \pi(\sigma_a^2|a_1^{(t+1)},a_2^{(t+1)},\dots,a_{n_a+2}^{(t+1)},\mu_a^{(t+1)},\sigma_a^{2\,(t+1)},\mathbf{y},H_d)$
    if $t > n_{burn}$ then
        Calculate $\log\left\{LR_{CS}(\boldsymbol{\theta}_a^{(t)};\mathbf{y}_{u_1},\mathbf{y}_{u_2})\right\}$ and $LR_{CS}(\boldsymbol{\theta}_a^{(t)};\mathbf{y}_{u_1},\mathbf{y}_{u_2})$ with formula (5.1)
    end
end
$\widehat{BF}_{CS} = \frac{1}{n_{\text{gibbs}}-n_{\text{burn}}}\sum_{t=n_{\text{burn}}+1}^{n_{\text{gibbs}}}\exp\left[\log\left\{LR_{CS}(\boldsymbol{\theta}_a^{(t)};\mathbf{y}_{u_1},\mathbf{y}_{u_2})\right\}\right]$
$\widehat{BF}_{CS} = \frac{1}{n_{\text{gibbs}}-n_{\text{burn}}}\sum_{t=n_{\text{burn}}+1}^{n_{\text{gibbs}}}LR_{CS}(\boldsymbol{\theta}_a^{(t)};\mathbf{y}_{u_1},\mathbf{y}_{u_2})$
```

The motivation to calculate the approximate Bayes Factor using the logarithm of the likelihood ratio is that very large or very small values often result in 'infinity' or 'not a number' because of machine precision. By calculating the logarithm of the likelihood ratio and transform it later on, this can be prevented in some cases. Both approximations give the same result if the likelihood ratio can be calculated directly. In the next section, the convergence of the Gibbs sampler will be discussed.

5.1.1. Convergence properties

The two-level normal-normal model is frequently studied in literature, usually under the name *one-way* or *simple random effects model.* The formulation of the hierarchical structure is slightly different in the sense that in the literature the priors on the variance parameters are specified by

$$\sigma_a^{-2} \sim \Gamma(a_1, b_1) \qquad \sigma_w^{-2} \sim \Gamma(a_2, b_2).$$

Lemma 2 and 3 show that the two-level normal-normal as discussed here can be reformulated to coincide with the problem in the literature by setting $a_1 = \nu_b/2$, $b_1 = \nu_b\sigma_b^2/2$, $a_2 = \nu_e/2$ and $b_2 = \nu_e\sigma_e^2/2$ in the Gamma priors above. This means that convergence results from the literature can immediately be applied.

Lemma 2. *If a random variable X follows a* Scale-inv-$\chi^2\left(\nu,\tau^2\right)$*-distribution, then this is equivalent to saying that X follows a* $\Gamma^{-1}\left(\frac{\nu}{2},\frac{\nu\tau^2}{2}\right)$*-distribution.*

Proof. This follows directly from the probability density function of the Scale-inv-$\chi^2\left(\nu,\tau^2\right)$-distribution

$$f(x) = \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)}\left(\tau^2\right)^{\nu/2} x^{-(\nu/2+1)} \exp\left[-\frac{\nu\tau^2}{2x}\right], \quad x > 0$$

and the probability density function of the $\Gamma^{-1}(\alpha,\beta)$-distribution

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} \exp\left[-\frac{\beta}{x}\right], \quad x > 0.$$

□

Lemma 3. *If a random variable X follows a $\Gamma^{-1}(\alpha,\beta)$-distribution, then X^{-1} follows a $\Gamma(\alpha,\beta)$-distribution.*

Proof. Let F denote the cumulative distribution function of X and let G denote the cumulative distribution function of X^{-1}, with corresponding densities f and g, respectively. Since both distributions are only defined for $x > 0$, it follows that

$$G(x) = \mathbb{P}\left(X^{-1} \leq x\right) = 1 - \mathbb{P}\left(X \leq x^{-1}\right) = 1 - F\left(x^{-1}\right).$$

This implies that

$$g(x) = \frac{\mathrm{d}}{\mathrm{d}x}\left(1 - F\left(x^{-1}\right)\right) = x^{-2} f\left(x^{-1}\right) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp\left[-\beta x\right],$$

which is indeed the density of the $\Gamma(\alpha,\beta)$-distribution. □

To ensure that the results from Algorithm 2 are reliable, one has to make sure that the Markov Chain obtained by the Gibbs sampler has converged from the starting value to stationarity. A *burn-in* of 5,000 is chosen to realise this. It is assumed here that the burn-in is large enough to accomplish stationarity. Estimating the sufficient burn-in is difficult and depends heavily on the prior variance on σ_a^2, see for example [20].

Another subject of interest is the rate of convergence of the Gibbs sampler. Therefore, the concept of *geometric ergodicity* needs to be introduced.

Definition 4. *Let E be a subset of Euclidean k-space, $\mathscr{E}$ be the corresponding Borel σ-algebra, and $P : E \times \mathscr{E} \to [0,1]$ a Markov transition function defining a discrete time, time homogeneous Markov chain $\{X_n : n = 0,1,\ldots\}$. Assume that this Markov chain is μ-irreducible (where μ is Lebesgue measure on $\mathscr{E}$), aperiodic and positive Harris. Let $P^n : E \times \mathscr{E} \to [0,1]$, $n = 2,3,\ldots$, denote the n-step Markov transition functions, and π the invariant probability measure. The Markov chain is called* geometrically ergodic *if there exists a π-integrable function $M : E \to \mathbb{R}_{>0}$, and a constant $0 < r < 1$ such that*

$$||P^n(x,\cdot) - \pi|| \leq M(x) r^n$$

for all $x \in E$ and $n = 0,1,2,\ldots$, where $||\cdot||$ represents total variation distance. [18]

The Gibbs Markov chain considered in [18] coincides with the one in Algorithm 2 and satisfies the assumptions given in Definition 4. The probability measure $\pi(\theta|e, H_d)$ is the invariant measure for the chain. Geometric ergodicity ensures quick convergence of the Markov chain to its stationary distribution, which is crucial for achieving effective simulation results in finite time and it is a key sufficient condition for the existence of a central limit theorem [19]. The next theorem will give sufficient conditions to establish geometric ergodicity of the Gibbs sampler.

Theorem 5. *The Gibbs sampler corresponding to the two-level normal-normal common source problem is geometrically ergodic whenever*

$$n' > (\sqrt{5} - 2) n'' \qquad \text{and} \qquad \frac{\nu_b}{2} \geq \frac{3(n_a + 2) - 2}{2(n_a + 2) - 2},$$

where $n' = \min(n_1, n_2, \ldots, n_{n_a}, n_{u_1}, n_{u_2})$ and $n'' = \max(n_1, n_2, \ldots, n_{n_a}, n_{u_1}, n_{u_2})$.

Proof. See [18]. □

If the conditions in Theorem 5 are met, this means that the algorithm to approximate the common source Bayes Factor converges fast and reliable approximations are obtained. In practice, the rate of convergence will also depend on the burn-in. There are many possibilities to check if the chosen burn-in is sufficient, such as considering autocorrelation or trace plots [32].

5.2. Specific source problem

Since the specific source problem corresponds to a different statistical model than the common source problem, the likelihood ratio as well as the full conditionals for the Gibbs sampler need to be derived again in order to approximate the Bayes Factor. The likelihood ratio for the specific source problem was given in Chapter 3:

$$LR_{SS}(\boldsymbol{\theta}; e_u) = \frac{f(e_u|\boldsymbol{\theta}_s, H_p)}{f(e_u|\boldsymbol{\theta}_a, H_d)} = \frac{\prod_{j=1}^{n_u} f_s(y_{uj}|\boldsymbol{\theta}_s)}{\int \prod_{j=1}^{n_u} f_a(y_{uj}|d,\boldsymbol{\theta}_a)g(d|\boldsymbol{\theta}_a)\,\mathrm{d}d},$$

where $\boldsymbol{\theta}_a = (\mu_a, \sigma_a^2, \sigma_w^2)$ and $\boldsymbol{\theta}_s = (\mu_s, \sigma_s^2)$ for the two-level normal-normal model. This likelihood ratio can be calculated explicitly using **Derivation 1** in Appendix A:

$$\begin{aligned} LR_{SS}(\boldsymbol{\theta}; e_u) = {} & \left(\frac{\sigma_w^2}{\sigma_s^2}\right)^{\frac{n_u}{2}} \left(\frac{\sigma_w^2}{n_u\sigma_a^2 + \sigma_w^2}\right)^{-\frac{1}{2}} \exp\left[\frac{\sum_j y_{uj}^2}{2\sigma_w^2} + \frac{\mu_a^2}{2\sigma_a^2}\right] \exp\left[-\frac{1}{2\sigma_s^2}\sum_j (y_{uj} - \mu_s)^2\right] \\ & \times \exp\left[-\frac{1}{2}\frac{\left(\sigma_a^2 \sum_j y_{uj} + \mu_a\sigma_w^2\right)^2}{\sigma_a^2\sigma_w^2\left(n_u\sigma_a^2 + \sigma_w^2\right)}\right]. \end{aligned} \tag{5.3}$$

Note that this is a different likelihood ratio than for example given in [7], since a different model is proposed for the evidence.

The Gibbs sampler is used to sample from $\pi(\boldsymbol{\theta}_a, \boldsymbol{\theta}_s|e_a, e_u, e_s, H_d) = \pi(\boldsymbol{\theta}_a|e_a, e_u, H_d)\pi(\boldsymbol{\theta}_s|e_s.H_d)$. The two posterior densities are independent, which means that the full conditionals from equation (5.2) hold with $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_{n_a}, \mathbf{y}_u)$ and $a_{n_a+1} = d$, augmented with the full conditionals on the specific source parameters that can be found from

$$\begin{aligned} \pi(\mu_s|\sigma_s^2, \mathbf{y}_s, H_d) &\propto \pi(\mu_s)\prod_{j=1}^{n_s} f_s(y_{sj}|\mu_s, \sigma_s^2) \\ \pi(\sigma_s^2|\mu_s, \mathbf{y}_s, H_d) &\propto \pi(\sigma_s^2)\prod_{j=1}^{n_s} f_s(y_{sj}|\mu_s, \sigma_s^2). \end{aligned} \tag{5.4}$$

It follows from (5.2) and (5.4) that

$$\begin{aligned} A_i|\mu_a, \sigma_a^2, \sigma_w^2, \mathbf{y}, H_d &\sim N\left(\frac{\sigma_a^2\sigma_w^2}{\sigma_w^2 + n_i\sigma_a^2}\left[\frac{1}{\sigma_w^2}\sum_{j=1}^{n_i} y_{ij} + \frac{\mu_a}{\sigma_a^2}\right], \frac{\sigma_a^2\sigma_w^2}{\sigma_w^2 + n_i\sigma_a^2}\right) \\ \mu_a|a_1, a_2, \ldots, a_{n_a+1}, \sigma_a^2, \sigma_w^2, \mathbf{y}, H_d &\sim N\left(\frac{\lambda\sigma_b^2\sigma_a^2}{\sigma_a^2 + (n_a+1)\lambda\sigma_b^2}\left[\frac{1}{\sigma_a^2}\sum_{i=1}^{n_a+1} a_i + \frac{\mu_\pi}{\lambda\sigma_b^2}\right], \frac{\lambda\sigma_b^2\sigma_a^2}{\sigma_a^2 + (n_a+1)\lambda\sigma_b^2}\right) \\ \sigma_a^2|a_1, a_2, \ldots, a_{n_a+1}, \mu_a, \sigma_w^2, \mathbf{y}, H_d &\sim \text{Scale-inv-}\chi^2\left(\nu_b + n_a + 1, \frac{\nu_b\sigma_b^2 + \sum_{i=1}^{n_a+1}(a_i - \mu_a)^2}{\nu_b + n_a + 1}\right) \\ \sigma_w^2|a_1, a_2, \ldots, a_{n_a+1}, \mu_a, \sigma_a^2, \mathbf{y}, H_d &\sim \text{Scale-inv-}\chi^2\left(\nu_e + \sum_{i=1}^{n_a+1} n_i, \frac{\nu_e\sigma_e^2 + \sum_{i=1}^{n_a+1}\sum_{j=1}^{n_i}(y_{ij} - a_i)^2}{\nu_e + \sum_{i=1}^{n_a+1} n_i}\right) \\ \mu_s|\sigma_s^2, \mathbf{y}_s, H_d &\sim N\left(\frac{\sigma_b^2\sigma_s^2}{\sigma_s^2 + n_s\sigma_b^2}\left[\frac{1}{\sigma_s^2}\sum_{j=1}^{n_s} y_{sj} + \frac{\mu_\pi}{\sigma_b^2}\right], \frac{\sigma_b^2\sigma_s^2}{\sigma_s^2 + n_s\sigma_b^2}\right) && (5.5) \\ \sigma_s^2|\mu_s, \mathbf{y}_s, H_d &\sim \text{Scale-inv-}\chi^2\left(\nu_e + n_s, \frac{\nu_e\sigma_e^2 + \sum_{j=1}^{n_s}(y_{sj} - \mu_s)^2}{\nu_e + n_s}\right). && (5.6) \end{aligned}$$

Again, the estimates $\hat{\sigma}_b^2$, $\hat{\sigma}_e^2$ and $\hat{\mu}_\pi$ will be plugged into the full conditionals, since it is assumed that the real values of σ_b^2, σ_e^2 and μ_π are unknown. The total approximation procedure of the specific source Bayes Factor is described in Algorithm 3. The convergence of the Gibbs sampler will be discussed in the next section.

Algorithm 3: Specific source approximation of Bayes Factor

```
Data: $e_u, e_s, e_a$
Result: Approximation of specific source Bayes Factor
Initialization
    $n_{\text{gibbs}} = 20{,}000, \quad n_{\text{burn}} = 5{,}000, \quad \mu_a^{(1)} = 1, \quad \sigma_a^{2\,(1)} = 1, \quad \sigma_w^{2\,(1)} = 1, \quad \mu_s^{(1)} = 1, \quad \sigma_s^{2\,(1)} = 1,$
    $a_i^{(1)} = 1$ for $i = 1, 2, \ldots, n_a + 1$
Approximate hyperparameters $\hat{\mu}_\pi$, $\hat{\sigma}_b^2$ and $\hat{\sigma}_e^2$
for iteration $t = 1, 2, \ldots, n_{gibbs}$ do
    Generate $a_i^{(t+1)} \sim \pi(a_i | \mu_a^{(t)}, \sigma_a^{2\,(t)}, \sigma_w^{2\,(t)}, \mathbf{y}, H_d)$ for $i = 1, 2, \ldots, n_a + 1$
    Generate $\mu_a^{(t+1)} \sim \pi(\mu_a | a_1^{(t+1)}, a_2^{(t+1)}, \ldots, a_{n_a+1}^{(t+1)}, \sigma_a^{2\,(t)}, \sigma_w^{2\,(t)}, \mathbf{y}, H_d)$
    Generate $\sigma_a^{2\,(t+1)} \sim \pi(\sigma_a^2 | a_1^{(t+1)}, a_2^{(t+1)}, \ldots, a_{n_a+1}^{(t+1)}, \mu_a^{(t+1)}, \sigma_w^{2\,(t)}, \mathbf{y}, H_d)$
    Generate $\sigma_w^{2\,(t+1)} \sim \pi(\sigma_a^2 | a_1^{(t+1)}, a_2^{(t+1)}, \ldots, a_{n_a+1}^{(t+1)}, \mu_a^{(t+1)}, \sigma_a^{2\,(t+1)}, \mathbf{y}, H_d)$
    Generate $\mu_s^{(t+1)} \sim \pi(\mu_s | \sigma_s^{2\,(t)}, \mathbf{y}_s, H_d)$
    Generate $\sigma_s^{2\,(t+1)} \sim \pi(\sigma_s^2 | \mu_s^{(t+1)}, \mathbf{y}_s, H_d)$
    if $t > n_{burn}$ then
        Calculate $\log\left\{LR_{SS}(\boldsymbol{\theta}_a^{(t)}, \boldsymbol{\theta}_s^{(t)}; \mathbf{y}_u)\right\}$ and $LR_{SS}(\boldsymbol{\theta}_a^{(t)}, \boldsymbol{\theta}_s^{(t)}; \mathbf{y}_u)$ with formula (5.3)
    end
end
$\widehat{BF}_{SS} = \frac{1}{n_{\text{gibbs}} - n_{\text{burn}}} \sum_{t=n_{\text{burn}}+1}^{n_{\text{gibbs}}} \exp\left[\log\left\{LR_{SS}(\boldsymbol{\theta}_a^{(t)}, \boldsymbol{\theta}_s^{(t)}; \mathbf{y}_u)\right\}\right]$
$\widehat{BF}_{SS} = \frac{1}{n_{\text{gibbs}} - n_{\text{burn}}} \sum_{t=n_{\text{burn}}+1}^{n_{\text{gibbs}}} LR_{SS}(\boldsymbol{\theta}_a^{(t)}, \boldsymbol{\theta}_s^{(t)}; \mathbf{y}_u)$
```

5.2.1. Convergence properties

The Gibbs sampler for the specific source problem generates two independent Markov chains: one corresponding to $\pi(\boldsymbol{\theta}_a | e_a, e_u, H_d)$ and the other corresponding to $\pi(\boldsymbol{\theta}_s | e_s, H_d)$. The convergence of the first Markov chain was already discussed in Section 5.1.1. Therefore, only the convergence of the second chain will be discussed in this section. The following theory is needed to prove geometric ergodicity:

Definition 6. *A positive function w is* unbounded off compact sets *if for every $\gamma > 0$ the level set $\{x : w(x) \leq \gamma\}$ is compact. [18]*

Definition 7. *A Markov chain $\{X_t : t = 0, 1, \ldots\}$ is* Feller continuous *if*

$$\mathbb{E}[f(X_{t+1}) | X_t = x_n] \to \mathbb{E}[f(X_{t+1}) | X_t = x] \quad \text{if } x_n \to x \text{ in } \mathcal{X}$$

for all bounded continuous $f : \mathcal{X} \to \mathbb{R}$. [6]

Proposition 8. *Suppose the Markov chain $\{X_t : t = 0, 1, \ldots\}$ is Feller continuous. If for some positive function $w : \mathcal{X} \to [1, \infty]$ that is unbounded off compact sets*

$$\mathbb{E}[w(X_{t+1}) | X_t = x] \leq \rho w(x) + L \quad \text{for all } x \in \mathcal{X} \tag{5.7}$$

for some $\rho < 1$ and $L < \infty$, then the Markov chain is geometrically ergodic. [18]

Theorem 9. *The Gibbs sampler corresponding to the model*

$$\begin{aligned} Y_{sj} | \mu_s, \sigma_s^2 &\overset{iid}{\sim} N\left(\mu_s, \sigma_s^2\right), \quad j = 1, \ldots, n_s \\ \mu_s &\sim N\left(\mu_\pi, \sigma_b^2\right) \\ \sigma_s^{-2} &\sim \Gamma\left(\frac{\nu_e}{2}, \frac{\nu_e \sigma_e^2}{2}\right) \end{aligned}$$

is geometrically ergodic whenever $\nu_e + n_s > 2$.

Proof. The proof is based on [18]. To show Feller continuity of the chain, let $f : \mathcal{X} \to \mathbb{R}$ be bounded and continuous, and consider

$$\begin{aligned}&\mathbb{E}\left[f(\mu_s^{(t+1)},\sigma_s^{2\,(t+1)}) \mid (\mu_s^{(t)},\sigma_s^{2\,(t)}) = (\mu_n,\sigma_n^2),\, \mathbf{y}_s\right] = \\ &\int\int f\left(\mu_s^{(t+1)},\sigma_s^{2\,(t+1)}\right)\pi\left(\mu_s^{(t+1)} \mid \sigma_s^{2\,(t)}=\sigma_n^2,\, \mathbf{y}_s\right)\pi\left(\sigma_s^{2\,(t+1)} \mid \mu_s^{(t+1)},\, \mathbf{y}_s\right)\, d\mu_s^{(t+1)}\, d\sigma_s^{2\,(t+1)},\end{aligned}$$

where the last two functions correspond to the distributions given in (5.5) and (5.6), respectively. Using the Dominated Convergence Theorem, the Feller continuity is found from this expression.

Next, define the functions

$$\begin{aligned}w_1(\mu_s,\sigma_s^2) &= \sigma_s^2 \\ w_2(\mu_s,\sigma_s^2) &= (\mu_s-\bar{y}_s)^2, \quad \text{where } \bar{y}_s = \frac{1}{n_s}\sum_{j=1}^{n_s} y_{sj} \\ w_3(\mu_s,\sigma_s^2) &= e^{c/\sigma_s^2}, \quad 0 < c < \frac{\nu_e\sigma_e^2}{2}.\end{aligned}$$

Consider the function $w(\mu_s,\sigma_s^2) = \sum_{i=1}^3 B_i w_i(\mu_s,\sigma_s^2)$, where the B_i are positive constants to be determined. Clearly, w is positive and continuous by construction. To show that w is unbounded off compacts sets, Definition 6 states that

$$K_\gamma = \left\{(\mu_s,\sigma_s^2) : w\left(\mu_s,\sigma_s^2\right) \le \gamma\right\}$$

needs to be compact for every $\gamma > 0$. Since w is continuous, the level set K_γ is closed and it suffices to show that K_γ is bounded. This can be done by showing that for $(\mu_s,\sigma_s^2) \in K_\gamma$, $|\mu_s|$ is bounded and σ_s^2 is bounded away from both 0 and ∞. Therefore, note that for $(\mu_s,\sigma_s^2) \in K_\gamma$

$$w_i(\mu_s,\sigma_s^2) \le \frac{1}{B_i} w(\mu_s,\sigma_s^2) \le \frac{\gamma}{B_i} \qquad \text{for } i = 1,2,3.$$

Using this property, it follows that

$$\frac{c}{\log(\gamma/B_3)} \le \sigma_s^2 \le \frac{\gamma}{B_1} \qquad \text{and} \qquad \bar{y}_s - \sqrt{\frac{\gamma}{B_2}} \le \mu_s \le \bar{y}_s + \sqrt{\frac{\gamma}{B_2}},$$

which gives the desired bounds for γ large enough, i.e., $\gamma > B_3$. Since closed subsets of a compact set are again compact, this shows that K_γ is compact for every $\gamma > 0$ and therefore w is unbounded off compact sets.

In order to use Proposition 8, it remains to show that (5.7) holds. Therefore the conditional expectation

$$\begin{aligned}\mathbb{E}\left[w\left(\mu_s^{(t+1)},\sigma_s^{2\,(t+1)}\right) \mid \mu_s^{(t)},\sigma_s^{2\,(t)},\mathbf{y}_s\right] &= \mathbb{E}\left[\mathbb{E}\left[w\left(\mu_s^{(t+1)},\sigma_s^{2\,(t+1)}\right) \mid \mu_s^{(t+1)},\mathbf{y}_s\right] \mid \mu_s^{(t)},\sigma_s^{2\,(t)},\mathbf{y}_s\right] \\ &= B_1\cdot\mathbb{E}\left[\mathbb{E}[w_1\left(\mu_s^{(t+1)},\sigma_s^{2\,(t+1)}\right) \mid \mu_s^{(t+1)},\mathbf{y}_s] \mid \mu_s^{(t)},\sigma_s^{2\,(t)},\mathbf{y}_s\right] \\ &\quad + B_2\cdot\mathbb{E}\left[w_2\left(\mu_s^{(t+1)},\sigma_s^{2\,(t+1)}\right) \mid \mu_s^{(t)},\sigma_s^{2\,(t)},\mathbf{y}_s\right] \\ &\quad + B_3\cdot\mathbb{E}\left[\mathbb{E}[w_3\left(\mu_s^{(t+1)},\sigma_s^{2\,(t+1)}\right) \mid \mu_s^{(t+1)},\mathbf{y}_s] \mid \mu_s^{(t)},\sigma_s^{2\,(t)},\mathbf{y}_s\right]\end{aligned}$$

will be considered. The third term is easy to bound using the moment generating function of the Gamma distribution and equation (5.6):

$$\begin{aligned}\mathbb{E}\left[w_3\left(\mu_s^{(t+1)},\sigma_s^{2\,(t+1)}\right) \mid \mu_s^{(t+1)},\mathbf{y}_s\right] = \mathbb{E}\left[e^{c/\sigma_s^{2\,(t+1)}} \;\middle|\; \mu_s^{(t+1)},\mathbf{y}_s\right] &= \left(\frac{\frac{1}{2}\nu_e\sigma_e^2 + \frac{1}{2}\sum_{j=1}^{n_s}\left(y_{sj}-\mu_s^{(t+1)}\right)^2}{\frac{1}{2}\nu_e\sigma_e^2 + \frac{1}{2}\sum_{j=1}^{n_s}\left(y_{sj}-\mu_s^{(t+1)}\right)^2 - c}\right)^{\frac{\nu_e+n_s}{2}} \\ &\le \left(\frac{\nu_e\sigma_e^2}{\nu_e\sigma_e^2-2c}\right)^{\frac{\nu_e+n_s}{2}} = \text{const}.\end{aligned}$$

It follows from (5.7) that constants are irrelevant, so it is not necessary to keep track of them since it is always possible to choose an $L < \infty$ larger than the sum of all the constants. Note that w_3 is only included in w to ensure that it is unbounded off compact sets.

Next, using (5.6) and the identity $\sum_{j=1}^{n_s}(y_{sj}-\mu_s)^2 = \sum_{j=1}^{n_s}(y_{sj}-\bar{y}_s)^2 + n_s(\mu_s-\bar{y}_s)^2$ it is found that

$$\begin{aligned}
\mathbb{E}\left[w_1\left(\mu_s^{(t+1)},\sigma_s^{2\,(t+1)}\right)\mid \mu_s^{(t+1)},\mathbf{y}_s\right] = \mathbb{E}\left[\sigma_s^{2\,(t+1)}\mid \mu_s^{(t+1)},\mathbf{y}_s\right] &= \frac{\nu_e\sigma_e^2+\sum_{j=1}^{n_s}\left(y_{sj}-\mu_s^{(t+1)}\right)^2}{\nu_e+n_s-2}\\
&= \frac{\nu_e\sigma_e^2}{\nu_e+n_s-2}+\frac{\sum_{j=1}^{n_s}\left(y_{sj}-\bar{y}_s\right)^2}{\nu_e+n_s-2}+\frac{n_s\left(\mu_s^{(t+1)}-\bar{y}_s\right)^2}{\nu_e+n_s-2}\\
&= \text{const.}+\frac{n_s}{\nu_e+n_s-2}\,w_2\left(\mu_s^{(t+1)},\sigma_s^{2\,(t+1)}\right).
\end{aligned}$$

In order to calculate $\mathbb{E}\left[\mathbb{E}\left[w_1\left(\mu_s^{(t+1)},\sigma_s^{2\,(t+1)}\right)\,\middle|\,\mu_s^{(t+1)},\mathbf{y}_s\right]\,\middle|\,\mu_s^{(t)},\sigma_s^{2\,(t)},\mathbf{y}_s\right]$, the following conditional expectation is required:

$$\begin{aligned}
\mathbb{E}\left[w_2\left(\mu_s^{(t+1)},\sigma_s^{2\,(t+1)}\right)\mid \mu_s^{(t)},\sigma_s^{2\,(t)},\mathbf{y}_s\right] &= \mathbb{E}\left[\left(\mu_s^{(t+1)}-\bar{y}_s\right)^2\,\middle|\,\mu_s^{(t)},\sigma_s^{2\,(t)},\mathbf{y}_s\right]\\
&= \mathrm{Var}\left(\mu_s^{(t+1)}-\bar{y}_s\mid \mu_s^{(t)},\sigma_s^{2\,(t)},\mathbf{y}_s\right)+\left(\mathbb{E}\left[\mu_s^{(t+1)}-\bar{y}_s\mid \mu_s^{(t)},\sigma_s^{2\,(t)},\mathbf{y}_s\right]\right)^2\\
&= \frac{\sigma_b^2\sigma_s^{2\,(t)}}{\sigma_s^{2\,(t)}+n_s\sigma_b^2}+\left(\frac{\sigma_b^2}{\sigma_s^{2\,(t)}+n_s\sigma_b^2}n_s\bar{y}_s+\frac{\sigma_s^{2\,(t)}}{\sigma_s^{2\,(t)}+n_s\sigma_b^2}\mu_\pi-\bar{y}_s\right)^2\\
&= \frac{1}{\sigma_b^{-2}+n_s\sigma_s^{-2\,(t)}}+\left(\frac{\sigma_s^{2\,(t)}}{\sigma_s^{2\,(t)}+n_s\sigma_b^2}\right)^2(\mu_\pi-\bar{y}_s)^2\\
&\leq n_s^{-1}\sigma_s^{2\,(t)}+(\mu_\pi-\bar{y}_s)^2 = n_s^{-1}w_1\left(\mu_s^{(t)},\sigma_s^{2\,(t)}\right)+\text{ const.}
\end{aligned}$$

Therefore,

$$\mathbb{E}\left[w_1\left(\mu_s^{(t+1)},\sigma_s^{2\,(t+1)}\right)\mid \mu_s^{(t)},\sigma_s^{2\,(t)},\mathbf{y}_s\right]\leq \text{const.}+\delta_1 w_1\left(\mu_s^{(t)},\sigma_s^{2\,(t)}\right)$$

and

$$\mathbb{E}\left[w_2\left(\mu_s^{(t+1)},\sigma_s^{2\,(t+1)}\right)\mid \mu_s^{(t)},\sigma_s^{2\,(t)},\mathbf{y}_s\right]\leq \text{const.}+\delta_2 w_1\left(\mu_s^{(t)},\sigma_s^{2\,(t)}\right)$$

where

$$\delta_1=\frac{1}{\nu_e+n_s-2}<\infty \qquad \text{and} \qquad \delta_2=\frac{1}{n_s}<\infty.$$

Now, there exist an $\epsilon>0$ and a $0<\rho<1$ such that $\epsilon(\delta_1+\delta_2)<\rho$. Therefore,

$$\mathbb{E}\left[\epsilon\left(w_1\left(\mu_s^{(t+1)},\sigma_s^{2\,(t+1)}\right)+w_2\left(\mu_s^{(t+1)},\sigma_s^{2\,(t+1)}\right)\right)+w_3\left(\mu_s^{(t+1)},\sigma_s^{2\,(t+1)}\right)\mid \mu_s^{(t)},\sigma_s^{2\,(t)}\right]$$

$$\leq \epsilon(\delta_1+\delta_2)\,w_1\left(\mu_s^{(t)},\sigma_s^{2\,(t)}\right)+\text{const.}<\rho\,w_1\left(\mu_s^{(t)},\sigma_s^{2\,(t)}\right)+\text{const.},$$

which implies geometric ergodicity by Proposition 8. □

From Theorem 5 it follows that the Gibbs sampler corresponding to $\pi(\boldsymbol{\theta}_a|e_a,e_u,H_d)$ is geometrically ergodic whenever

$$n'>(\sqrt{5}-2)n'' \qquad \text{and} \qquad \frac{\nu_b}{2}\geq\frac{3(n_a+1)-2}{2(n_a+1)-2},$$

where $n'=\min(n_1,n_2,\ldots,n_{n_a},n_u)$ and $n''=\max(n_1,n_2,\ldots,n_{n_a},n_u)$. Moreover, Theorem 9 shows that the Gibbs sampler corresponding to $\pi(\boldsymbol{\theta}_s|e_s,H_d)$ is geometrically ergodic as long as $\nu_e+n_s>2$. If all these conditions are satisfied, fast convergence of the algorithm to approximate the specific source Bayes Factor is ensured. In the next chapter, the approximation of both the common source and specific source Bayes Factor will be put into practice. The conditions from Theorem 5 and 9 will be considered to ensure reliable approximations.

6

Application of the two-level normal-normal model

To assess the performance of the algorithms discussed in the previous chapter, the two-level normal-normal model was applied to both simulated and real datasets. First, the simulation procedure will be explained and simulated data in only one dimension will be considered to investigate the effect of the hyperparameters. Then the two-level normal-normal model is applied to three-dimensional simulated data. Finally, real datasets containing measurements on glass, MDMA tablets and knives will be used, and the results are compared with methods currently used at the Netherlands Forensic Institute (NFI).

6.1. Data simulation

Data needs to be simulated according to both the common source and specific source sampling models. For simplicity, first only one feature will be considered. The simulation procedure for the data corresponding to the common source problem is given in the following algorithm:

Algorithm 4: Common source data simulation (basic setup)

```
Specify hyperparameters for $k = 1$ features
    $\sigma_b^2 \in \left\{\frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1, 2, 3, 4\right\}$;   $\mu_\pi = 5$;   $\sigma_e^2 = 1$;   $\nu_b = 4$;   $\nu_e = \nu_b$;   $\lambda = 10$.
Generate parameters $\mu_a, \sigma_a^2, \sigma_w^2$ according to the prior distributions (4.5), (4.6) and (4.7).
Generate background material $Y_{ij}$ with $n_a = 100$ sources and $n_i = 5$ measurements per source
    according to (4.4) and (4.3).
Generate a set of separate background material $Y_{ij}^{\text{hyp}}$ for hyperparameter fitting with $n_a^{\text{hyp}} = 100$
    sources and $n_i^{\text{hyp}} = 5$ measurements per source according to (4.4) and (4.3).
Generate first unknown source material with $n_{u_1} = 5$ measurements where $P \sim N(\mu_a, \sigma_a^2)$ and
    $Y_{u_1 j} | P = p \overset{\text{iid}}{\sim} N(p, \sigma_w^2)$, $j = 1, \dots, n_{u_1}$.
if $H_p$ is assumed to be true then
    Generate second unknown source material with $n_{u_2} = 5$ measurements where
        $Y_{u_2 j} | P = p \overset{\text{iid}}{\sim} N(p, \sigma_w^2)$, $j = 1, \dots, n_{u_2}$.
else
    Generate second unknown source material with $n_{u_2} = 5$ measurements where $D \sim N(\mu_a, \sigma_a^2)$ and
        $Y_{u_2 j} | D = d \overset{\text{iid}}{\sim} N(d, \sigma_w^2)$, $j = 1, \dots, n_{u_2}$.
end
```

This simulation procedure will be referred to as the common source ‘basic setup’. In the next section, the basic setup will be altered to investigate the effect of the hyperparameters. Note that in the basic setup the background material is split. The effect of splitting the background material will also be further evaluated. Since the simulated data is balanced, the first condition from Theorem 5 is automatically satisfied. To satisfy the second condition, ν_b is set to a value larger than 3. Therefore, if the approximation procedure from Algorithm 2 is applied to this data, the corresponding Gibbs sampler is geometrically ergodic and reliable approximation results should be obtained.

To study the Bayes Factor corresponding to the specific source problem, data needs to be simulated according to the specific source sampling models. Again, only one feature will be considered for simplicity. The simulation procedure from Algorithm 4 is altered to the specific source setup and given in Algorithm 5.

Algorithm 5: Specific source data simulation (basic setup)

```
Specify hyperparameters and generate parameters and background material according to steps 1-4 in
  Algorithm 4.
Generate parameters $\mu_s$, $\sigma_s^2$ according to the prior distributions (4.11) and (4.12).
Generate specific source material with $n_s = 5$ measurements where $Y_{sj} \overset{\text{iid}}{\sim} N(\mu_s, \sigma_s^2), j = 1, \dots, n_s$.
if $H_p$ is assumed to be true then
    Generate unknown source material with $n_u = 5$ measurements where $Y_{uj} \overset{\text{iid}}{\sim} N(\mu_s, \sigma_s^2), j = 1, \dots, n_u$.
else
    Generate unknown source material with $n_u = 5$ measurements where $D \sim N(\mu_a, \sigma_a^2)$ and
      $Y_{uj} | D = d \overset{\text{iid}}{\sim} N(d, \sigma_w^2), j = 1, \dots, n_u$.
end
```

Since for the specific source problem $n_s = 5$ and $\nu_e = 4$, the condition of Theorem 9 is satisfied, which means that the Gibbs sampler corresponding to $\pi(\boldsymbol{\theta}_s | e_s, H_d)$ is geometrically ergodic. Again, since the simulated data is balanced and $\nu_b > 3$, the conditions from Theorem 5 are also satisfied. Therefore, the Gibbs sampler corresponding to $\pi(\boldsymbol{\theta}_a | e_a, e_u, H_d)$ is also geometrically ergodic and fast convergence of the approximation procedure is ensured.

6.2. One-dimensional problem

For the one-dimensional problem, data are simulated according to the sampling models corresponding to H_p or H_d as described in Algorithms 4 and 5. Seeds are used to ensure reproducible results and to be able to study the effect of changing the basic setup. For each setting of the parameter values, only one run of Algorithms 4 and 5 will be considered. The Bayes Factor is then approximated using Algorithm 2 or 3, depending on which problem is considered. Pretending that the values of the hyperparameters are unknown, estimates for σ_b^2, σ_e^2 and μ_π are obtained from the separate background material for the hyperparameter fitting as described in [2] and given by:

$$\hat{\mu}_\pi = \bar{y}, \qquad \hat{\sigma}_e^2 = \frac{1}{n_a^{\text{hyp}}} \sum_{i=1}^{n_a^{\text{hyp}}} s_{iw}^2, \qquad \hat{\sigma}_b^2 = s_b^2 - \frac{1}{n_i^{\text{hyp}}} \hat{\sigma}_e^2, \tag{6.1}$$

where

$$\bar{y}_i = \frac{1}{n_i^{\text{hyp}}} \sum_{j=1}^{n_i^{\text{hyp}}} y_{ij}^{\text{hyp}}, \qquad \bar{y} = \frac{1}{n_a^{\text{hyp}}} \sum_{i=1}^{n_a^{\text{hyp}}} \bar{y}_i, \qquad s_{iw}^2 = \frac{1}{n_i^{\text{hyp}} - 1} \sum_{j=1}^{n_i^{\text{hyp}}} \left(y_{ij}^{\text{hyp}} - \bar{y}_i \right)^2, \qquad s_b^2 = \frac{1}{n_a^{\text{hyp}} - 1} \sum_{i=1}^{n_a^{\text{hyp}}} (\bar{y}_i - \bar{y})^2.$$

Different values of σ_b^2 are evaluated and since the between-source variation increases when σ_b^2 increases, it is expected that a more extreme Bayes Factor is obtained for larger values of σ_b^2. Other interesting properties of the model are also investigated, such as the influence of using separate background material for the hyperparameter estimation. Two sets of background material are used in the basic setup, whereas in the situation denoted by '$Y_{\text{hyp}} = Y$' only one set of background material is simulated and both hyperparameter estimation and the Gibbs sampler use the same set of background material. The impact of the parameters λ and $\nu_b = \nu_e$ is considered by changing their values to 1 and 6, respectively. Lastly, in the situations '$Y_{u_2} = Y_{u_1}$' and '$Y_u = Y_s$' the unknown source material Y_{u_2} and Y_u is not simulated but is set exactly equal to Y_{u_1} and Y_s, respectively. It is expected that this artificial choice will result in the largest Bayes Factors. The results for the common source and specific source approximate Bayes Factors are given in Figures 6.1-6.4, where the data are simulated assuming that either the prosecution hypothesis or the defence hypothesis is true.

Considering Figures 6.1 and 6.2 it may seem as if some lines are missing in the graphs, because the results for the basic setup, '$Y_{\text{hyp}} = Y$' and '$\lambda = 1$' almost coincide. This means that both the separate background material and the parameter λ have little impact on the Bayes Factor. Since both sets of background material are simulated similarly, approximately the same hyperparameter estimations should be obtained. Moreover, the estimates are only used in the second level of the hierarchical model, which explains the small impact on the Bayes Factor. The parameter λ is used in the simulation of *all* evidence sets and the effect is therefore limited.

Increasing the parameter $\nu_b = \nu_e$ results in a significantly smaller Bayes Factor when the evidence is generated according to the prosecution model. This can be explained by the fact that the scaled inverse chi-square distributions become more centered around σ_b^2 and σ_e^2, meaning that the values of both the within- and the between-source variation are less spread out. Consequently, it is harder to indicate same sources which is beneficial for the defence and reduces the Bayes Factor. The approximate Bayes Factor for '$Y_{u_1} = Y_{u_2}$' is slightly smaller than for the basic setup. Looking at formula (5.1), this is probably caused by the fact that the absolute value of the sum of Y_{u_2}, when it is sampled, is slightly larger than the sum of Y_{u_1}.

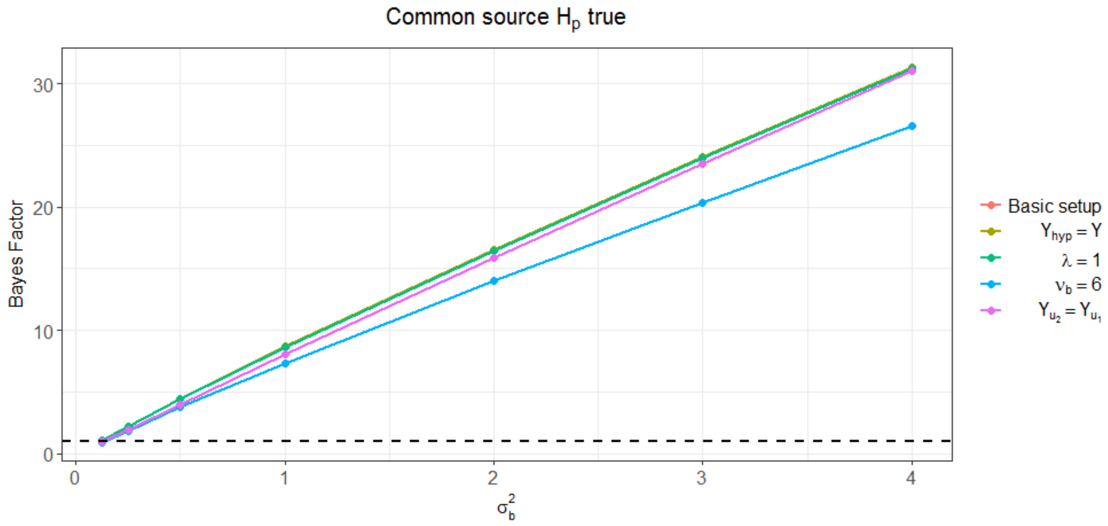

Figure 6.1: Approximate values of the Bayes Factor for the common source problem with unknown source evidence generated according to the prosecution model. The dashed line indicates the value 1.

A horizontal line is added to Figure 6.1 to indicate where the Bayes Factor is equal to 1. When σ_b^2 is greater than or equal to $\sigma_e^2 = 1$, the model is able to correctly determine that the evidence was generated according to the prosecution model. However, when σ_b^2 is significantly smaller than σ_e^2, false negatives are obtained. This should not be surprising, since in this situation the within-source variation is much larger than the between-source variation, making it nearly impossible to distinguish different sources.

In Figure 6.2 the effect of a small within-source variation can also be seen. When the data are generated according to the defence model, it is expected that the Bayes Factor decreases as σ_b^2 increases. However, for $\sigma_b^2 \leq \frac{1}{2}$ increasing values are obtained. A larger value of $\nu_b = \nu_e$ is again beneficial for the defence, increasing the Bayes Factor when $\sigma_b^2 > \frac{1}{2}$ and reducing the Bayes Factor when the model incorrectly behaves like the data are generated according to the prosecution model. For all values of σ_b^2 a good result is obtained in the sense that the approximate Bayes Factor is smaller than 1.

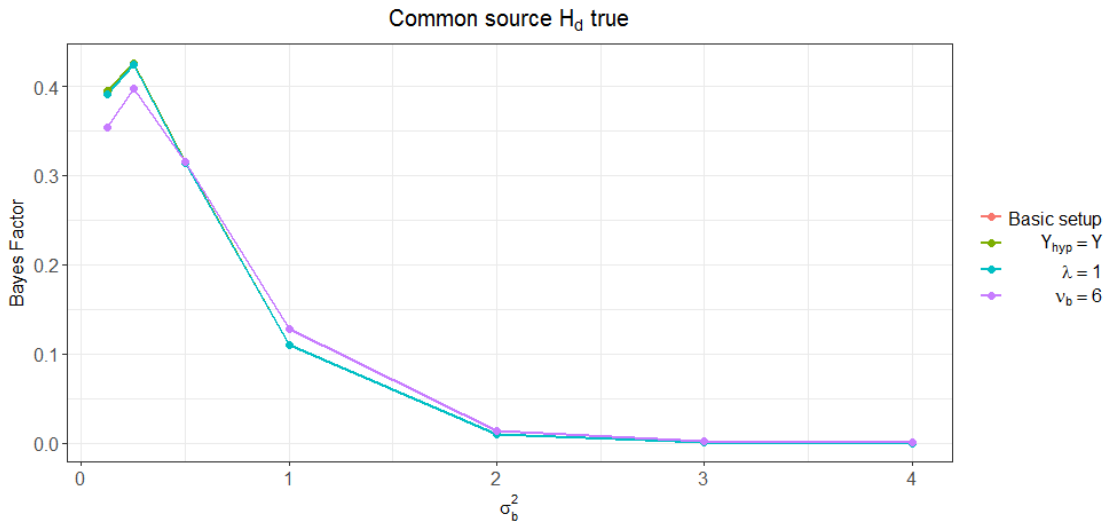

Figure 6.2: Approximate values of the Bayes Factor for the common source problem with unknown source evidence generated according to the defence model.

Figure 6.3 gives the results of the approximate specific source Bayes Factor when the unknown source evidence is generated according to the prosecution model. It is immediately visible that changing the basic setup has a significant effect on the results. Using the same background material for the hyperparameter estimation and the Gibbs sampler leads to a smaller Bayes Factor, but this is presumably only because a smaller estimate of σ_b^2 is obtained when using Y instead of Y_{hyp}, while the estimates for σ_e^2 and μ_π are similar. Since $\hat{\sigma}_b^2$ is now used in the first level of the hierarchical model for the unknown source evidence, it has a lot more influence. A smaller estimated value of σ_b^2 results in less diffuse sources, making the model less confident.

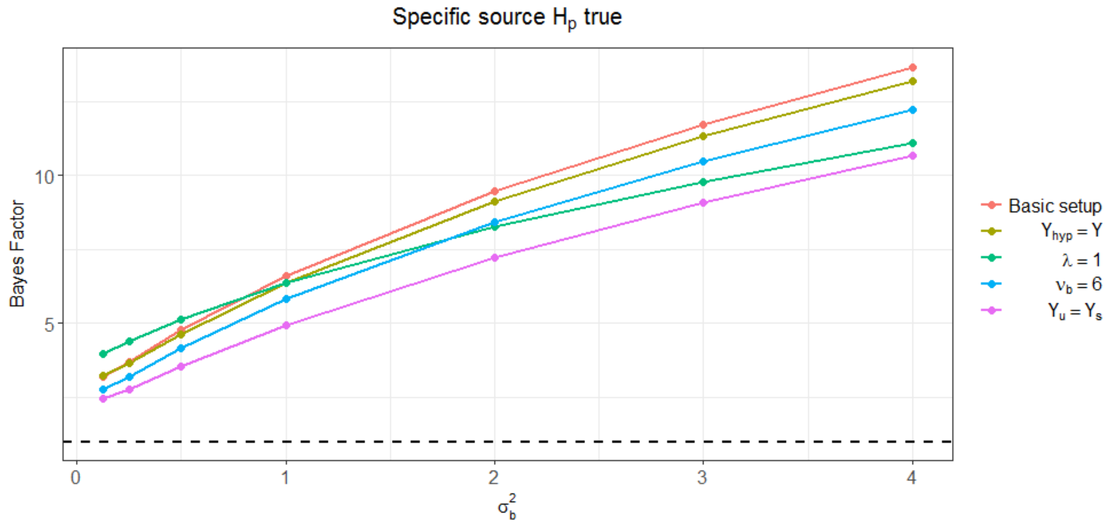

Figure 6.3: Approximate values of the Bayes Factor for the specific source problem with unknown source evidence generated according to the prosecution model. The dashed line indicates the value 1.

The impact of using Y instead of Y_{hyp} is negligible when the evidence is simulated according to the defence model, because then the estimate $\hat{\sigma}_b^2$ is only used in the second level of the hierarchical model for the unknown source evidence (see Figure 6.4). Setting the parameter λ equal to 1 leads to less extreme Bayes Factors when $\sigma_b^2 \geq \sigma_e^2$ in both Figure 6.3 and 6.4. When $\lambda = 1$, the means of the background material and the mean of the specific source material are simulated similarly and lie more closely together, which makes the model less confident about the difference between the unknown source evidence and the background material. Increasing the parameter $\nu_b = \nu_e$ gives a less extreme approximate Bayes factor, which can be explained by the same reasoning as was done for the common source problem.

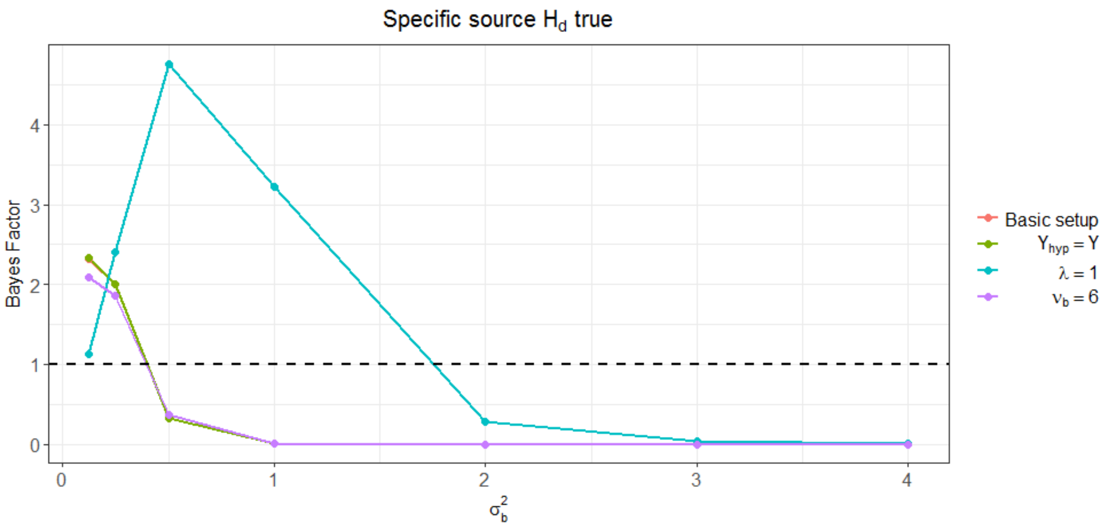

Figure 6.4: Approximate values of the Bayes Factor for the specific source problem with unknown source evidence generated according to the defence model.

A critical note should be added to the use of seeds in the simulation of the evidence. Here, only one setting of the seeds is considered. For different seeds, the approximation of the Bayes Factor might give incorrect results in the sense that a value smaller than 1 is found when the data are generated according to the prosecution model, or a value larger than 1 when the data are generated according to the defence model. This does not necessarily mean that the model does not work: for some seeds the simulated data just gives more support to agree with the other hypothesis, for example when the parameters are sampled from the tails of the prior distributions.

6.3. Higher dimensional problems

In practice, multiple features of the evidence have to be taken into account, leading to higher dimensional problems. The simulation procedure explained in Algorithms 4 and 5 can easily be adapted using the multivariate distributions given in Chapter 4. The estimates for the hyperparameters given in equation (6.1) remain valid and can be adjusted to the higher dimensional setup by replacing the squares with outer vector products, i.e., $\mathbf{u} \otimes \mathbf{v} = \mathbf{u}\mathbf{v}^T$. For both the common and specific source problem, the likelihood ratio and the full conditionals for the Gibbs sampler can be found in Appendix A so that Algorithms 2 and 3 can be used to approximate the Bayes Factor.

The higher dimensional model was tested for 3 features, where the initial hyperparameters are specified as follows:

$$\boldsymbol{\Sigma}_b = \begin{bmatrix} 6 & 3 & 5 \\ 3 & 4 & 2 \\ 5 & 2 & 9 \end{bmatrix}, \quad \boldsymbol{\Sigma}_e = \begin{bmatrix} 3 & 0 & 3 \\ 0 & 1 & -2 \\ 3 & -2 & 8 \end{bmatrix}, \quad \boldsymbol{\mu}_\pi = \begin{bmatrix} 3 \\ 5 \\ 4 \end{bmatrix}, \quad \nu_b = \nu_e = 27, \quad \lambda = 10. \tag{6.2}$$

Note that not every randomly chosen matrix suffices for $\boldsymbol{\Sigma}_b$ and $\boldsymbol{\Sigma}_e$; the proposed model requires that the matrices are symmetric and positive definite.

Increasing the number of features leads to a far more computationally intensive approximation procedure. For the three-dimensional problem with $n_{\text{gibbs}} = 20{,}000$ and $n_{\text{burn}} = 5{,}000$ the computation already takes about 10 minutes. This is mainly caused by the fact that in each iteration of the Gibbs sampler several matrix inverses have to be calculated. Because of this computational intensity, the simulation study is only repeated for three different evidence sets, keeping the initial hyperparameters unchanged. The results can be found in Table 6.1.

	Scenario	$\widehat{BF}_{CS}$	$\widehat{BF}_{SS}$
1.	H_p true	1.349179	$5.65137 \cdot 10^{17}$
	H_d true	$2.336728 \cdot 10^{-5}$	$2.372235 \cdot 10^{-21}$
2.	H_p true	32.65359	$2.905801 \cdot 10^{25}$
	H_d true	$1.683076 \cdot 10^{-11}$	$3.234398 \cdot 10^{-33}$
3.	H_p true	28.04373	$6.38302 \cdot 10^{-12}$
	H_d true	$1.743259 \cdot 10^{-39}$	$5.133723 \cdot 10^{-223}$

Table 6.1: Approximate Bayes Factors for three different evidence sets for the three-dimensional common and specific source problem.

For the common source problem, the results of the approximate Bayes Factor look good: a Bayes Factor larger than 1 is found when the evidence is simulated assuming that H_p is true and a Bayes Factor smaller than 1 when the evidence is simulated assuming that H_d is true. Note that more extreme values are found when the evidence is simulated according to the defence model; for example, in the third simulation a value of the order 10^{-39} is found when the evidence is simulated assuming that H_d is true, but only a value of 28 for the competing hypothesis. It is unclear why these results are so unbalanced. More thorough research on the (hyper)parameters might give more insights.

The values found for the approximate Bayes Factor of the specific source problem are more balanced. However, the model fails to support the true hypothesis in the third simulation when the evidence is generated according to the prosecution model. On the other hand, an extremely small approximate Bayes Factor is found when the evidence is generated according to the defence model. This suggests that for this scenario the method tends to agree with the defence no matter how the evidence is generated, which can be caused by extreme values of the simulated parameters. Moreover, there might not be enough specific source evidence available to accurately model the specific source distribution. Note that the common source problem leads to more conservative values for the Bayes Factor than the specific source problem.

Little is known about the convergence of the Gibbs sampler for the higher dimensional problem. Since the hyperparameters are chosen randomly and geometric ergodicity is not proven, it is hard to explain peculiarities in the results. Other ad-hoc methods might be used to ensure convergence of the Gibbs sampling algorithm. For example, in [32] a *thinning interval* is used to obtain approximately independent Gibbs samples. When a thinning interval of size i is chosen, only every ith value is used from the Gibbs sampler. The approximate independence suggests that the Central Limit Theorem applies to the resulting samples.

6.4. Application to real datasets

Lastly, the two-level normal-normal model is applied to real data. To this end, one open source dataset from the Federal Bureau of Investigation (FBI) is used as well as two anonimised datasets provided by the Netherlands Forensic Institute (NFI). It is interesting to compare the results of the described model with the methods currently used at the NFI. In many casework, customised models are developed to ensure the best fit for the evidence. A more unified framework is provided by SAILR, Software for the Analysis and Implementation of Likelihood Ratios. This is a graphical user-friendly interface (GUI) that helps forensic experts calculate numerical likelihood ratios for a selection of statistical models.

The two-level normal-normal model implemented in SAILR for feature-based comparison is similar to the common source model described in Chapter 4, where a normal distribution for both the within- and between-source variation is assumed. However, SAILR computes the likelihood ratio in a frequentist way, using the following estimates for $\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a, \boldsymbol{\Sigma}_w$ [23]:

$$\hat{\boldsymbol{\mu}}_a = \bar{\mathbf{y}} = \frac{1}{\sum_{i=1}^{n_a} n_i} \sum_{i=1}^{n_a} \sum_{j=1}^{n_i} \mathbf{y}_{ij}, \quad \hat{\boldsymbol{\Sigma}}_w = \frac{1}{\sum_{i=1}^{n_a} n_i - n_a} \sum_{i=1}^{n_a} \sum_{j=1}^{n_i} (\mathbf{y}_{ij} - \bar{\mathbf{y}}_i)(\mathbf{y}_{ij} - \bar{\mathbf{y}}_i)^T, \quad \hat{\boldsymbol{\Sigma}}_a = \frac{MS^2_{\text{between}} - \hat{\boldsymbol{\Sigma}}_w}{\kappa}, \tag{6.3}$$

where

$$\bar{\mathbf{y}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{y}_{ij}, \quad MS^2_{\text{between}} = \frac{1}{n_a - 1} \sum_{i=1}^{n_a} n_i (\bar{\mathbf{y}}_i - \bar{\mathbf{y}})(\bar{\mathbf{y}}_i - \bar{\mathbf{y}})^T, \quad \kappa = \frac{1}{n_a - 1} \left(\sum_{i=1}^{n_a} n_i - \frac{\sum_{i=1}^{n_a} n_i^2}{\sum_{i=1}^{n_a} n_i} \right).$$

In the next sections, a short description of each dataset will be provided and the results of the approximate common source Bayes Factor will be compared to the likelihood ratio from SAILR. For the approximate Bayes Factor 100,000 iterations of the Gibbs sampler are computed with a burn-in of 5,000. The parameters ν_b and ν_e are both set to 27 and the chosen value for λ is 10. No separate background material is used in the Gibbs sampler to keep both models as closely related as possible.

6.4.1. Glass

The first dataset consists of measurements made on a group of glass fragments from 16 different window panes. This data was collected by Dr. JoAnn Buscaglia of the FBI Laboratory Division and is publicly available online through the Journal of the Royal Statistical Society [2]. The composition of the chemical elements calcium (Ca), potassium (K), silicion (Si) and iron (Fe) are measured for each of the fragments. The values used for the analysis are the natural logarithms of the ratios Ca/K, Ca/Si and Ca/Fe as suggested in [2] and [32].

From the dataset it is known which fragments belong to which window pane. Therefore, it is possible to construct two scenarios corresponding to situations where either H_p or H_d is true. Following [32], the evidence sets are composed as follows:

e_a: Five measurements from each of 14 windows, excluding windows number 10 and 48;

e_{u_1}: Two measurements from window number 10;

e_{u_2}: Three different measurements from window number 10 (H_p true);
Five measurements from window number 48 (H_d true).

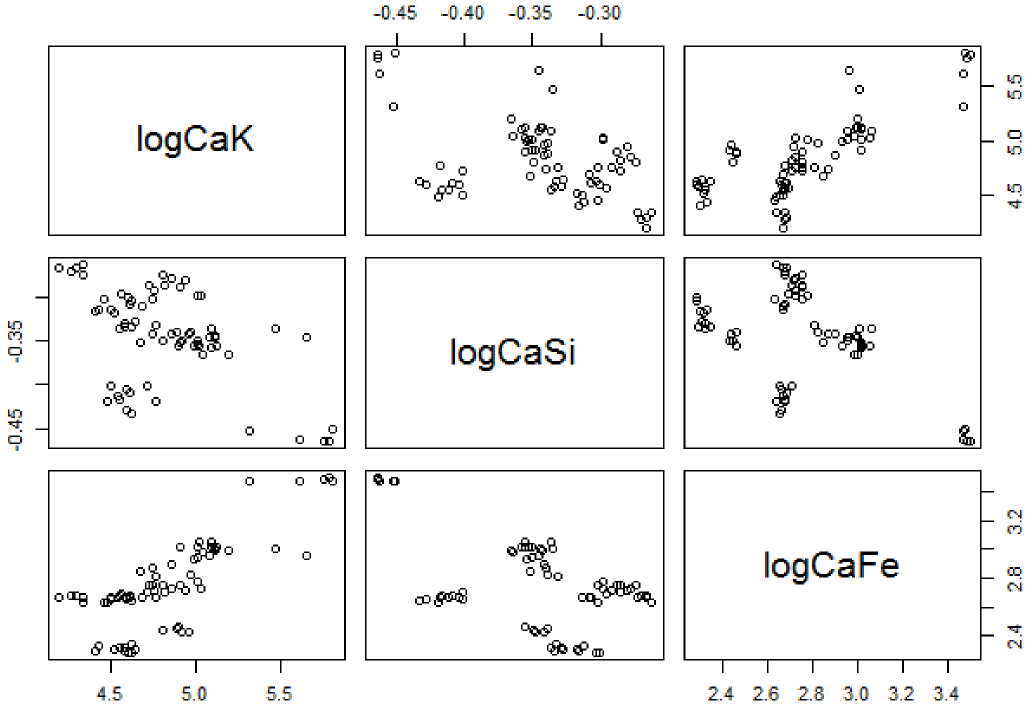

Figure 6.5: Pairwise plot of the three features of the background material from the glass dataset.

This means for SAILR that the recovered data is e_{u_2}, the control data is e_{u_1} and the background data is e_a. The results can be found in Table 6.2. Although the assumption of two-level normality is not verified, both methods result in a value of evidence supporting the true hypothesis for each scenario. In practice, evaluating the assumption of two-level multivariate normality is difficult, since both the number of sources and the number of measurements per source are limited. Note that the value of evidence calculated with the approximate Bayes Factor results in more conservative values than the likelihood ratio obtained from SAILR.

Scenario	$\widehat{BF}_{CS}$	LR_{SAILR}
H_p true	38.189	88.433
H_d true	$3.886 \cdot 10^{-18}$	$4.182 \cdot 10^{-21}$

Table 6.2: Values of evidence for the two-level normal-normal model applied to the glass dataset.

6.4.2. MDMA tablets

A commonly used dataset within the NFI comes from the CHAMP (Collaborative Harmonization of Methods for Profiling of Amphetamine Type Stimulants) project. The data contains measurements on four characteristics of MDMA tablets: diameter, thickness, weight and purity. Since the last feature is not measured for each tablet, it is not used in this analysis. The dataset is expanded with street samples consisting of 160 consignments with 2 or more tablet measurements where it is not known whether there are links between the consignments.

Several tablets per batch are measured and since for each tablet is indicated from which batch or consignment it originates, again scenarios can be constructed corresponding to situations where either H_p or H_d is true. The evidence sets are composed as follows:

e_a: Four measurements of each of 77 different street sample consignments;

e_{u_1}: 42 measurements of batch 9 of the CHAMP data;

e_{u_2}: Five measurements of batch 9 of the CHAMP data (H_p true);
Three measurements of the CHAMP data, from batch 1, 7 and 12, and two from two different street sample consignments (H_d true).

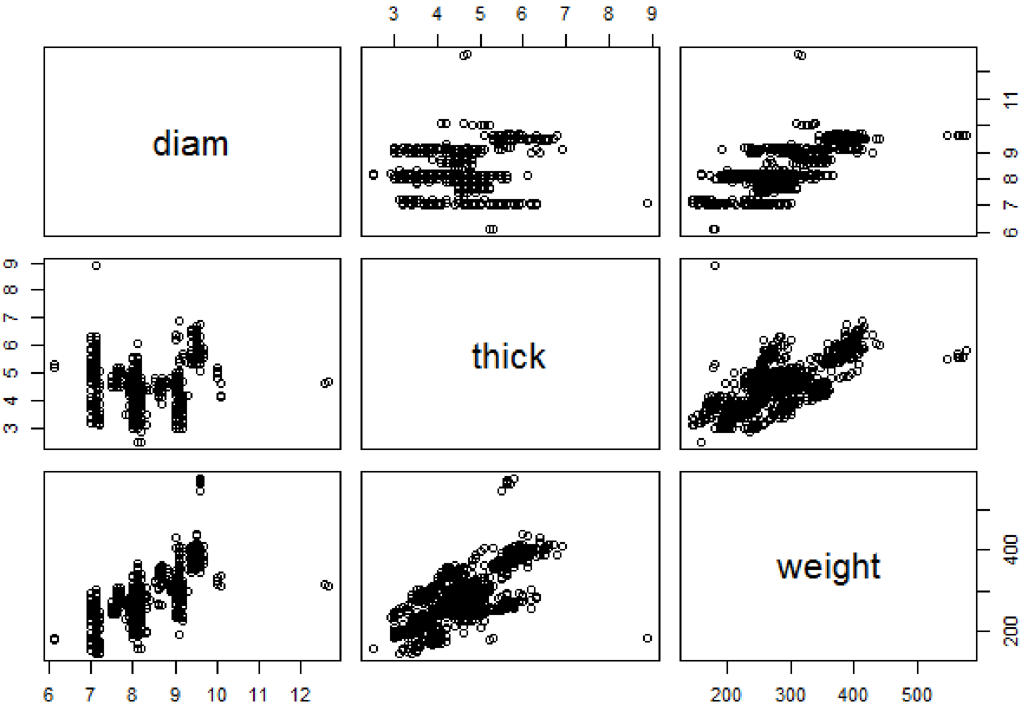

Figure 6.6: Pairwise plot of the three features of the background material from the MDMA dataset.

For the background material, the consignments with less than four tablet measurements are removed and for the consignments with more measurements, only the first four are used. This is only done because the hyperparameter estimation for the approximate Bayes Factor simplifies significantly when an equal amount of measurements per source is obtained. Still a reasonably large background dataset is left, but of course the analysis could be repeated using all background data from the street samples. Here, it is chosen to test the hypothesis that *all* tablets in e_{u_1} and e_{u_2} originate from the same source or different sources. Another approach would be to test each tablet from e_{u_2} separately, but because of the computational intensity of the approximate Bayes Factor this has not been done yet.

The results can be found in Table 6.3. In the scenario when H_p is true, both methods find a similar value of evidence. However, the values of evidence differ extremely in the scenario that H_d is true. Looking at Figure 6.6 this might be caused by the discrete behaviour of the diameter feature, which suggests that the restrictive prior distributions are not appropriate for this dataset.

Scenario	$\widehat{BF}_{CS}$	LR_{SAILR}
H_p true	2591.901	2488.846
H_d true	$3.487 \cdot 10^{-8}$	$2.493 \cdot 10^{-88}$

Table 6.3: Values of evidence for the two-level normal-normal model applied to the MDMA dataset.

6.4.3. Knives

The last dataset consists of the chemical composition of several knives, measured by seven features indicated with the letters *A*-*G*. The data is provided by the department of chemical and physical traces at the NFI (Peter Zoon and colleagues). To reduce the computational intensity, only three of the seven features are used in the analysis. It is chosen to work with the features *A*, *C* and *E*, because they do not show discrete behaviour as some of the other features do, see Figure 6.7.

This time two traces are measured without knowing the actual source of the traces. Therefore, it is not possible to decide beforehand what the true hypothesis is. The evidence sets for the two scenarios are composed as follows:

e_a: 50 measurements of each of 15 different knives;

e_{u_1}: 10 measurements of one knife;

e_{u_2}: One measurement of one knife (scenario 1);
One measurement of one knife (scenario 2).

The results are given in Table 6.4. As in the glass dataset, the value of evidence calculated with the approximate Bayes Factor results in more conservative values than the likelihood ratio obtained from SAILR. Both methods suggest that in scenario 1 the measurements are from the same knife, although the support is minimal, and that in scenario 2 the measurements are from different knives.

Scenario	$\widehat{BF}_{CS}$	LR_{SAILR}
1.	7.630	14.290
2.	$3.428 \cdot 10^{-4}$	$6.679 \cdot 10^{-6}$

Table 6.4: Values of evidence for the two-level normal-normal model applied to the knives dataset.

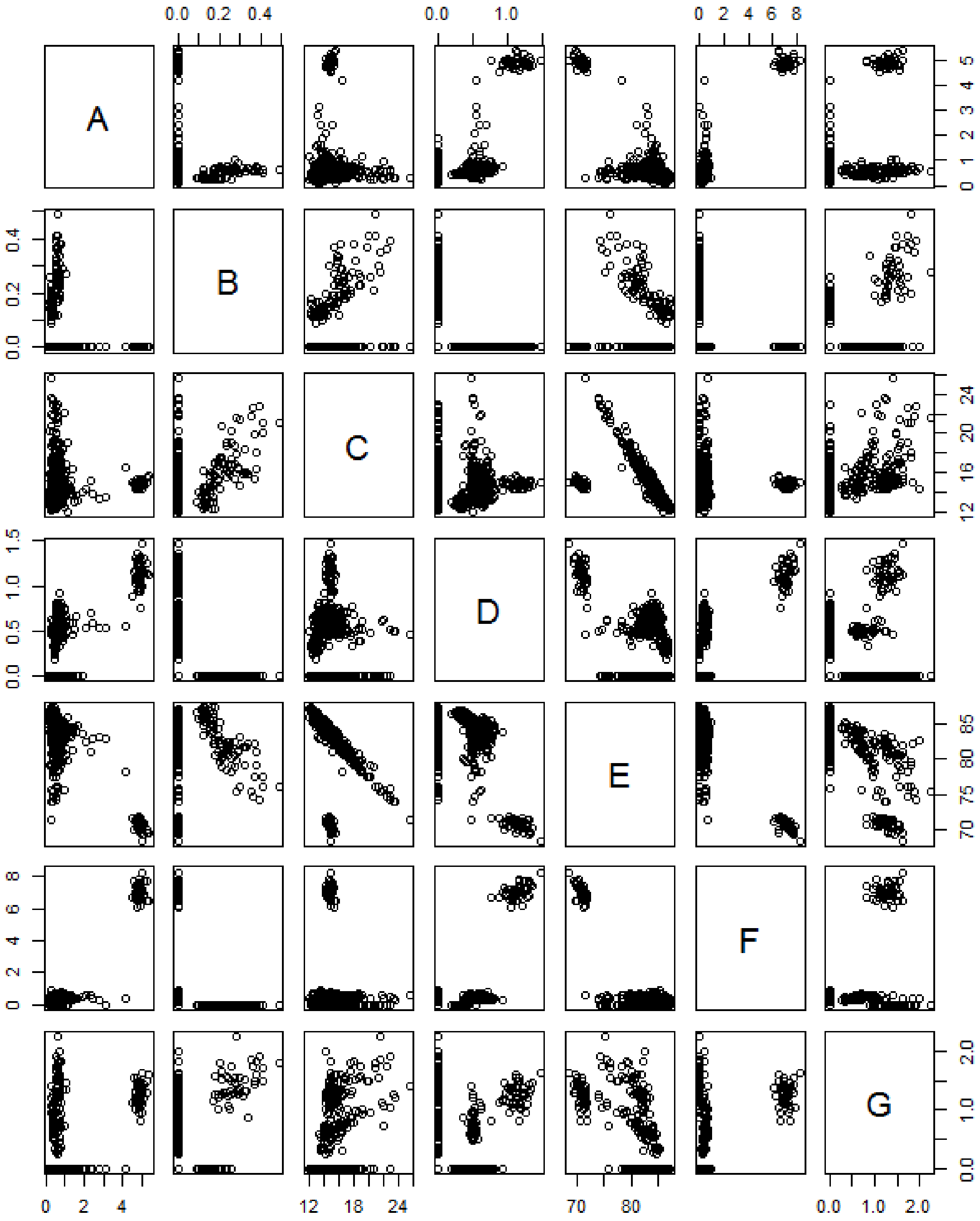

Figure 6.7: Pairwise plot of the seven features of the background material from the knives dataset.

7

Estimating the overall mean

Instead of using a Bayesian approach and a Gibbs sampler to calculate the Bayes Factor for the two-level normal-normal model, it is possible to use a frequentistic approach and calculate the likelihood ratio with estimates of all unknown parameters. The latter was already discussed in Section 6.4. In forensic statistics, two estimators are used to estimate the overall mean $\boldsymbol{\mu}_a$ from the background material: the weighted and unweighted mean. The effect of choosing either of these mean estimators has already been investigated by ir. F.S. Kool for the univariate case. This research has been extended to the multivariate case and has resulted in the submission of the paper “Overall mean estimation of trace evidence in a two-level normal-normal model” by Kool, F.S.; Van Dorp, I.N.; Bolck, A.; Leegwater, A.J.; and Jongbloed, G. to be published in the journal *Forensic Science International.* This chapter contains the submitted paper and therefore the notation is slightly different than in the rest of this thesis.

Overall mean estimation of trace evidence in a two-level normal-normal model

Kool, F.S.; Van Dorp, I.N.; Bolck, A.; Leegwater, A.J.; and Jongbloed, G.

Abstract

In the evaluation of measurements on characteristics of forensic trace evidence, Aitken and Lucy (2004) model the data as a two-level model using assumptions of normality where likelihood ratios are used as a measure for the strength of evidence. A two-level model assumes two sources of variation: the variation within measurements in a group (first level) and the variation between different groups (second level). Estimates of the variation within groups, the variation between groups and the overall mean are required in this approach. This paper discusses three estimators for the overall mean. In forensic science, two of these estimators are known as the weighted and unweighted mean. For an optimal choice between these estimators, the within- and between-group covariance matrices are required. In this paper a generalization to the latter two mean estimators is suggested, which is referred to as the generalized weighted mean. The weights of this estimator can be chosen such that they minimize the variance of the generalized weighted mean. These optimal weights lead to a “toy estimator”, because they depend on the unknown within- and between-group covariance matrices. Using these optimal weights with estimates for the within- and between-group covariance matrices leads to the third estimator, the optimal “plug-in” generalized weighted mean estimator. The three estimators and the toy estimator are compared through a simulation study. Under conditions generally encountered in practice, we show that the unweighted mean can be preferred over the weighted mean. Moreover, in these situations the unweighted mean and the optimal generalized weighted mean behave similarly. An artificial choice of parameters is used to provide an example where the optimal generalized weighted mean outperforms both the weighted and unweighted mean. Finally, the three mean estimators are applied to real XTC data to illustrate the impact of the choice of overall mean estimator.

7.1. Introduction

The likelihood ratio is a generally accepted measure for the strength of evidence in many forensic comparison problems. Modelling the data as a two-level random effects model using assumptions of normality is a well-known approach in likelihood ratio calculation [2, 24]. The use of a two-level model leads to a likelihood ratio which depends on the unknown parameters of the two-level model. Within the Likelihood Paradigm [40] estimates of these parameters are required to estimate the likelihood ratio. Alternatively, it is possible to assign priors to all parameters following a full-Bayesian approach [4, 9, 47]. In this paper, different methods are described to estimate one of the parameters: the overall mean vector of the two-level model. Two currently used estimators in forensic statistics, the weighted and unweighted mean, are compared. There is still discussion which of these mean estimators should be used when the data are unbalanced, i.e., when the number of data points differs per group [2, 41]. Moreover, a general class of estimators for the overall mean, referred to as generalized weighted mean, is suggested. This class contains the two aforementioned estimators as special cases. The choice of the mean estimator is important for the commonly used analysis of variance estimator to estimate the between-source covariance matrix, which is another parameter to be estimated in the two-level model [41].

In Section 7.2 the likelihood ratio approach in the setting of a two-level model is described, yielding an explicit expression for the likelihood ratio in terms of the model parameters. Section 7.3 covers the explanation of the estimators and their relative efficiencies in terms of (partly) unknown parameters. In Section 7.4 a comparison of the estimation techniques is given through a simulation study and in Section 7.5 the estimators are applied to real XTC data. In this paper the results are given for the multivariate case. The results for the univariate case are obtained by replacing the (traces of the) covariance matrices with the corresponding variances.

7.2. Likelihood ratio approach

In forensic comparison problems it is investigated whether a control item (e.g. XTC tablets from consignment C_1) and a recovered item (e.g. XTC tablets from consignment C_2) originate from the same unknown source[1]. Very generally stated, a prosecutor's hypothesis (H_p) and a hypothesis of the defence (H_d) may be as follows:

$$\begin{cases} H_p: & \text{The control and recovered item originate from the same source.} \\ H_d: & \text{The control and recovered item originate from different sources.} \end{cases}$$

Comparison of the control and recovered item given the two hypotheses involves evidence E. This evidence concerns certain characteristics or features of the two items. The likelihood ratio approach refers to a well-known probabilistic framework based on Bayes' rule to evaluate the strength of the evidence in such forensic comparison problems. In this approach, the likelihood ratio is the ratio of the probability of evidence E given the two hypotheses H_p and H_d:

$$\mathrm{LR} = \frac{P(E \mid H_p)}{P(E \mid H_d)}. \tag{7.1}$$

This likelihood ratio expresses how much more likely it is to find the evidence under the prosecutor's hypothesis than under the hypothesis of the defence. Therefore, the likelihood ratio can be seen as a measure to quantify the strength of evidence.

7.2.1. Model

Various types of models exist to compute the likelihood ratio in equation (7.1). In this paper, the focus will be on a feature-based two-level random effects model using assumptions of normality which is applicable to continuous data [3, 24].

Consider the situation that several continuous features of the control and recovered item are measured by forensic experts, e.g. the diameter, thickness and weight of the XTC tablets in consignment C_1 and C_2. Let k denote the number of features and let n_1 be the number of measurements of these features on the control item, e.g. the number of tablets that is measured in consignment C_1. The composed continuous random

[1] In the context of [32], this problem is known as a common source problem. The model corresponds to the situation where the sources are assumed to be random realizations from a probability distribution. For more details about the difference between common and specific source problems, see [33].

vector $\mathbf{Y}_1$ represents the n_1 measurement vectors of the features on the control item,

$$\mathbf{Y}_1 = (\mathbf{Y}_{11}, \ldots, \mathbf{Y}_{1n_1}) = \left(\begin{bmatrix} Y_{11,1} \\ Y_{11,2} \\ \vdots \\ Y_{11,k} \end{bmatrix}, \ldots, \begin{bmatrix} Y_{1n_1,1} \\ Y_{1n_1,2} \\ \vdots \\ Y_{1n_1,k} \end{bmatrix} \right).$$

This vector can be referred to as control data. The control data will be compared to the recovered data $\mathbf{Y}_2$, i.e., the composed random vector which represents the n_2 measurements of the features on the recovered item. Thus, the composed random vectors $\mathbf{Y}_l = (\mathbf{Y}_{lj}, 1 \le j \le n_l), l = 1,2$, represent for example the diameters, thicknesses and weights of the tablets from consignments C_1 and C_2. To compare the control and recovered item, the means of the control and recovered data can be used as the evidence, i.e.,

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

where

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

denotes the mean over the n_l measurements.

The data are modelled using a (two-level) random effects model under the assumption of normality [3, 24]. The use of such a two-level model is appropriate, because the data are organized at more than one level: the measurements (first level) are nested within the items (second level), such as the control and recovered item. The variation between the n_l measurements within the same item is known as the within-source variation. The variation between the items is known as the between-source variation. It is assumed that both the within- and between-source variation are multivariate normally distributed. This means that within a source, the control and recovered data are independent and normally distributed around their group means $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, i.e.

$$\overline{\mathbf{Y}}_l \mid \boldsymbol{\theta}_l \sim \mathcal{N}_k\left(\boldsymbol{\theta}_l, n_l^{-1}\boldsymbol{\Sigma}\right) \qquad \text{for } l = 1,2$$

and the between-source variation is modelled by independent normally distributed random variables

$$\boldsymbol{\theta}_l \sim \mathcal{N}_k(\boldsymbol{\mu}, \mathbf{T}) \qquad \text{for } l = 1,2.$$

7.2.2. Likelihood ratio

In the literature, explicit likelihood ratio formulas under the normality assumptions in the two-level model are derived [2, 24, 51]. In this paper we will use the following likelihood ratio of the observed evidence $E = (\overline{\mathbf{y}}_1, \overline{\mathbf{y}}_2)$ [8]:

$$\mathrm{LR}(\overline{\mathbf{y}}_1, \overline{\mathbf{y}}_2 | \boldsymbol{\mu}) = \frac{|\mathbf{U}_0|^{\frac{1}{2}}}{|\mathbf{U}_n|^{\frac{1}{2}}} \exp\left[\frac{1}{2}\left((\overline{\mathbf{y}}_2 - \boldsymbol{\mu})^T \mathbf{U}_0^{-1} (\overline{\mathbf{y}}_2 - \boldsymbol{\mu}) - (\overline{\mathbf{y}}_2 - \boldsymbol{\mu}_n)^T \mathbf{U}_n^{-1} (\overline{\mathbf{y}}_2 - \boldsymbol{\mu}_n)\right)\right] \tag{7.2}$$

where

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

The explicit likelihood ratio formulas depend on the unknown overall mean $\boldsymbol{\mu}$, the between-source covariance matrix $\mathbf{T}$ and the within-source covariance matrix $\boldsymbol{\Sigma}$ of the described two-level model. Hence, in the Likelihood Paradigm, estimates of these parameters are required to estimate the likelihood ratio. In Section 7.3, estimators for the overall mean $\boldsymbol{\mu}$ are described. Estimators for the covariance matrices $\mathbf{T}$ and $\boldsymbol{\Sigma}$ are for example the multivariate analysis of variance estimators [41, 44]. Next to the computation of the likelihood ratio, the choice of the mean estimator $\hat{\boldsymbol{\mu}}$ is important for the analysis of variance estimator of the between-source covariance matrix $\mathbf{T}$, because this quantity depends on the mean $\boldsymbol{\mu}$ [41]. As an alternative to these approaches, maximum likelihood estimators can be used [44]. However, in the two-level normal-normal setup no explicit formulas exist for these estimators. Therefore, iterative methods are required [41]. Another option is to use a full-Bayesian approach with priors assigned to all parameters [4, 9, 47]. In this paper, we will focus on the non-Bayesian approach with $\mathbf{\Sigma}$ and $\mathbf{T}$ fixed, and we will compare several estimators for $\boldsymbol{\mu}$.

7.2.3. Background data

To estimate the parameters of the two-level model, background data that represent the population are required. The background data consist of measurements of the continuous features on a random sample of m items or groups, which represent the population. In each of the m groups, n_i $(i = 1, \dots, m)$ measurements are taken. The background data are denoted as $\{\mathbf{Z}_{ij} \mid 1 \leq i \leq m, 1 \leq j \leq n_i\}$, where $\mathbf{Z}_{ij}$ represents the vector of measured features within group i of measurement j. The background data are modelled by the extension of the two-level model described in Section 7.2.2 [2],

$$\begin{aligned} \mathbf{Z}_{ij} \mid \boldsymbol{\theta}_i &\overset{\text{iid}}{\sim} \mathcal{N}_k(\boldsymbol{\theta}_i, \mathbf{\Sigma}), \quad 1 \leq j \leq n_i, \\ \boldsymbol{\theta}_i &\overset{\text{iid}}{\sim} \mathcal{N}_k(\boldsymbol{\mu}, \mathbf{T}), \quad 1 \leq i \leq m. \end{aligned}$$

Under these assumptions, the background data are in fact modelled by a random effects model [44], i.e.,

$$\mathbf{Z}_{ij} = \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_{ij} \qquad \text{for } 1 \leq i \leq m, \quad 1 \leq j \leq n_i,$$

with $\boldsymbol{\mu}$ the overall mean,

$$\boldsymbol{\alpha}_i \overset{\text{iid}}{\sim} \mathcal{N}_k(\mathbf{0}_k, \mathbf{T}), \qquad 1 \leq i \leq m,$$

the random group effect and, independent of the $\boldsymbol{\alpha}_i$'s,

$$\boldsymbol{\varepsilon}_{ij} \overset{\text{iid}}{\sim} \mathcal{N}_k(\mathbf{0}_k, \mathbf{\Sigma}), \qquad 1 \leq j \leq n_i,$$

the random noise vectors or within-source variation.

7.3. Estimating the overall mean

First, the weighted mean and the unweighted mean are discussed as estimators for the overall mean $\boldsymbol{\mu}$. In Section 7.3.2 it is shown that what is the best estimator (the estimator with smallest variance) depends on the ratio of the traces of the within- and between-source covariance matrices. To derive this, a multivariate generalization of variance is given in Section 7.3.1. In Section 7.3.3 a generalization of the weighted and unweighted mean estimators is suggested, which is referred to as the generalized weighted mean. The weights of this estimator can be chosen such that they minimize the variance of the generalized weighted mean. These optimal weights lead to what we will call a "toy estimator". We use the term "toy estimator", because the optimal weights depend on the unknown within- and between-source covariance matrices $\mathbf{\Sigma}$ and $\mathbf{T}$. Hence, in practice only an estimate of the optimal weights can be obtained and the resulting estimator will be referred to as the optimal "plug-in" generalized weighted mean estimator.

7.3.1. Multivariate generalization of variance

A natural choice for the multivariate concept of variance for unbiased estimators is to consider the expected value of the squared Euclidean distance between the estimator and the true parameter of interest, i.e.,

$$\text{Var}(\hat{\boldsymbol{\mu}}) := \text{E}\left[||\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}||^2\right].$$

Note that we will prefer the unbiased estimator with minimal expected distance to the true parameter. For unbiased estimators it follows that

$$\begin{aligned} \text{Var}(\hat{\boldsymbol{\mu}}) &= \text{E}\left[||\hat{\boldsymbol{\mu}} - \text{E}[\hat{\boldsymbol{\mu}}]||^2\right] = \text{E}\left[(\hat{\boldsymbol{\mu}} - \text{E}[\hat{\boldsymbol{\mu}}])^T(\hat{\boldsymbol{\mu}} - \text{E}[\hat{\boldsymbol{\mu}}])\right] \\ &= \text{E}\left[\sum_{i=1}^{k}(\hat{\mu}_i - \text{E}[\hat{\mu}_i])^2\right] = \sum_{i=1}^{k}\text{Var}(\hat{\mu}_i) = tr(\mathbf{\Sigma}), \end{aligned}$$

where $\mathbf{\Sigma}$ denotes the covariance matrix of $\hat{\boldsymbol{\mu}}$. Any further mention of variance will refer to this definition.

7.3.2. Weighted versus unweighted mean

The group means of the background data are defined as the average of the observations $\mathbf{Z}_{ij}$ in each group,

$$\overline{\mathbf{Z}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{Z}_{ij}, \qquad 1 \leq i \leq m, \tag{7.3}$$

such that $\overline{\mathbf{Z}}_i \sim \mathcal{N}_k(\boldsymbol{\mu}, \mathbf{T} + n_i^{-1}\boldsymbol{\Sigma})$. These group means are used to approximate $\boldsymbol{\theta}_i$. Two estimators for the overall mean $\boldsymbol{\mu}$ are the weighted mean and the unweighted mean. The weighted mean is the average over all observations $\mathbf{Z}_{ij}$ in the background data [44],

$$\hat{\boldsymbol{\mu}}_w = \frac{1}{N}\sum_{i=1}^{m} n_i\overline{\mathbf{Z}}_i = \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\mathbf{Z}_{ij}, \tag{7.4}$$

where N is the total number of observations, i.e. $N = \sum_{i=1}^{m} n_i$. The weighted mean is unbiased, since

$$\mathrm{E}[\hat{\boldsymbol{\mu}}_w] = \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\mathrm{E}\left[\mathbf{Z}_{ij}\right] = \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\mathrm{E}\left[\boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_{ij}\right] = \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\boldsymbol{\mu} = \boldsymbol{\mu}.$$

The variance of the weighted mean is equal to

$$\mathrm{Var}(\hat{\boldsymbol{\mu}}_w) = \frac{tr(\mathbf{T})}{N^2}\sum_{i=1}^{m} n_i^2 + \frac{tr(\boldsymbol{\Sigma})}{N},$$

see Appendix 7.7.1. The unweighted mean is the mean of the group means [41],

$$\hat{\boldsymbol{\mu}}_u = \frac{1}{m}\sum_{i=1}^{m}\overline{\mathbf{Z}}_i. \tag{7.5}$$

The unweighted mean is also unbiased, since

$$\mathrm{E}[\hat{\boldsymbol{\mu}}_u] = \frac{1}{m}\sum_{i=1}^{m}\mathrm{E}\left[\overline{\mathbf{Z}}_i\right] = \frac{1}{m}\sum_{i=1}^{m}\boldsymbol{\mu} = \boldsymbol{\mu}$$

and its variance is equal to

$$\mathrm{Var}(\hat{\boldsymbol{\mu}}_u) = \frac{tr(\mathbf{T})}{m} + \frac{tr(\boldsymbol{\Sigma})}{m^2}\sum_{i=1}^{m}\frac{1}{n_i},$$

see Appendix 7.7.1.

First note that if the data are balanced, i.e., $n_i = n$ for all $i = 1, \ldots, m$, the weighted and unweighted mean are exactly the same. For unbalanced data where the number of measurements differs per group, there is a dispute whether to use the weighted mean or the unweighted mean [2, 41]. The weighted mean fits naturally with a designed experiment or other reasons where the unequal number of measurements reflects the composition of the population or the importance of the groups. In that case it is important that groups with more measurements have more weight in the estimation of the overall mean, which is an argument in favor of the weighted mean. In cases where the number of measurements is more or less randomly chosen or determined by factors independent of the population composition (e.g. sampling costs) the number of measurements is not important. It is then beneficial that groups have equal importance, despite the number of observations, which is an argument in favor of the unweighted mean. In fact, below it is shown that the best choice between these estimators depends on the situation.

Since both estimators are unbiased, it will be examined which estimator has smallest variance. Hence, consider the efficiency of $\hat{\boldsymbol{\mu}}_w$ relative to $\hat{\boldsymbol{\mu}}_u$ [36]:

$$\mathrm{eff}(\hat{\boldsymbol{\mu}}_u, \hat{\boldsymbol{\mu}}_w) = \frac{\mathrm{Var}(\hat{\boldsymbol{\mu}}_w)}{\mathrm{Var}(\hat{\boldsymbol{\mu}}_u)} = \frac{\frac{tr(\mathbf{T})}{N^2}\sum_{i=1}^{m} n_i^2 + \frac{tr(\boldsymbol{\Sigma})}{N}}{\frac{tr(\mathbf{T})}{m} + \frac{tr(\boldsymbol{\Sigma})}{m^2}\sum_{i=1}^{m}\frac{1}{n_i}}. \tag{7.6}$$

Multiplying the numerator and denominator in equation (7.6) with the term m^2N^2 and setting $r = \frac{tr(\boldsymbol{\Sigma})}{tr(\mathbf{T})}$ results in

$$\text{eff}(\hat{\boldsymbol{\mu}}_u, \hat{\boldsymbol{\mu}}_w) = \frac{m^2\sum_{i=1}^m n_i^2 + m^2Nr}{mN^2 + rN^2\sum_{i=1}^m \frac{1}{n_i}}. \tag{7.7}$$

Using Jensen's inequality it can be shown that the efficiency can have larger and smaller values than one, see Appendix 7.7.2. Therefore, one cannot be conclusive about which estimator has smallest variance. From Appendix 7.7.2, it follows that

$$\text{eff}(\hat{\boldsymbol{\mu}}_u, \hat{\boldsymbol{\mu}}_w) > 1 \quad \text{iff} \quad r < \frac{m^2\sum_{i=1}^m n_i^2 - mN^2}{N^2\sum_{i=1}^m \frac{1}{n_i} - m^2N} =: c. \tag{7.8}$$

Note that both the numerator and the denominator of c are positive because of inequalities (7.17) and (7.18) (see Appendix 7.7.2), hence the constant c is always positive. Therefore,

$$\begin{cases} \text{Var}(\hat{\boldsymbol{\mu}}_w) > \text{Var}(\hat{\boldsymbol{\mu}}_u) & \text{if } tr(\boldsymbol{\Sigma}) < c \cdot tr(\mathbf{T}), \\ \text{Var}(\hat{\boldsymbol{\mu}}_w) < \text{Var}(\hat{\boldsymbol{\mu}}_u) & \text{if } tr(\boldsymbol{\Sigma}) > c \cdot tr(\mathbf{T}). \end{cases} \tag{7.9}$$

From the inequalities in (7.9) it follows that the best choice of the estimator depends on two factors. The first is the ratio between the trace of the within-source covariance matrix $\boldsymbol{\Sigma}$ and the trace of the between-source covariance matrix $\mathbf{T}$. For example, if the trace of the within-source covariance matrix $\boldsymbol{\Sigma}$ is small, i.e. $\overline{\mathbf{Z}}_i \approx \boldsymbol{\theta}_i$, the unweighted mean virtually equals the maximum likelihood estimator based on the (unobservable) $\boldsymbol{\theta}_i$'s and we would prefer the unweighted mean. This example corresponds to the first inequality in expression (7.9). Since the parameters $\boldsymbol{\Sigma}$ and $\mathbf{T}$ are unknown, this factor relies on prior knowledge or on experience of the forensic expert. The second factor that affects the choice between the weighted and unweighted mean is the value of the constant c, which depends on the number of groups m and the number of measurements within each group n_i. The following lemma gives more insight in the possible values of the constant c.

Lemma 10. *The constant*

$$c = \frac{m^2\sum_{i=1}^m n_i^2 - mN^2}{N^2\sum_{i=1}^m \frac{1}{n_i} - m^2N}$$

is always greater than or equal to 1.

The proof of this lemma can be found in Appendix 7.7.3. This lemma illustrates that when $tr(\boldsymbol{\Sigma}) < tr(\mathbf{T})$ the unweighted mean will always have a smaller variance than the weighted mean. In many forensic comparison problems it is realistic to assume that within-source variation is smaller than between-source variation. For instance, in XTC comparison problems this is due to the fact that the errors that cause the within-group variation (e.g. measurement errors, production errors, inhomogeneity within a batch) are often smaller than the between-group variation (mainly based on the preference of the producers). Consequently, in many XTC comparison problems it can be assumed that the trace of the within-source covariance matrix $\boldsymbol{\Sigma}$ is smaller than the trace of the between-source covariance matrix $\mathbf{T}$, i.e., $tr(\boldsymbol{\Sigma}) < tr(\mathbf{T})$. Since $c \geq 1$ always holds, the unweighted mean should in these situations be preferred over the weighted mean.

7.3.3. Generalized weighted mean

This section suggests a more general estimator for the mean compared to the weighted and unweighted mean described in Section 7.3.2. This general estimator will be referred to as the generalized weighted mean[2]. Define the generalized weighted mean as [36]

$$\hat{\boldsymbol{\mu}} = \sum_{i=1}^m \mathbf{W}_i\overline{\mathbf{Z}}_i \qquad \text{where } \mathbf{W}_i \text{ is a } k \times k \text{ matrix such that } \sum_{i=1}^m \mathbf{W}_i = \mathbf{I}_k. \tag{7.10}$$

Here, $\mathbf{I}_k$ denotes the $k \times k$-dimensional identity matrix. The weighted and unweighted mean are special cases of the generalized weighted mean given in equation (7.10). It can be seen that the weighted mean $\hat{\boldsymbol{\mu}}_w$ is the generalized weighted mean with weight matrices $\mathbf{W}_i = \frac{n_i}{N}\mathbf{I}_k$ for $1 \leq i \leq m$. The unweighted mean $\hat{\boldsymbol{\mu}}_u$ is the

[2] In the literature this estimator is called the weighted mean. However, in forensic literature the estimator in equation (7.4) is called the weighted mean. Therefore, we will refer to this estimator as generalized weighted mean.

generalized weighted mean with weight matrices $\mathbf{W}_i = \frac{1}{m}\mathbf{I}_k$ for $1 \le i \le m$.

Since the weight matrices $\mathbf{W}_1,\ldots,\mathbf{W}_m$ in equation (7.10) add up to the identity matrix, it follows that the generalized weighted mean is unbiased, i.e.,

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

The covariance matrix of $\hat{\boldsymbol{\mu}}$ is equal to

$$\mathrm{Cov}(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\mu}}) = \sum_{i=1}^{m} \mathbf{W}_i \,\mathrm{Cov}\left(\overline{\mathbf{Z}}_i,\overline{\mathbf{Z}}_i\right)\mathbf{W}_i^T = \sum_{i=1}^{m} \mathbf{W}_i(\mathbf{T}+n_i^{-1}\boldsymbol{\Sigma})\mathbf{W}_i^T \tag{7.11}$$

so that its variance is given by

$$\mathrm{Var}(\hat{\boldsymbol{\mu}}) = \sum_{i=1}^{m} tr\left(\mathbf{W}_i(\mathbf{T}+n_i^{-1}\boldsymbol{\Sigma})\mathbf{W}_i^T\right).$$

Since the variance depends on the choice of the weight matrices $\mathbf{W}_1,\ldots,\mathbf{W}_m$, the question arises how to choose these weights to minimize $\mathrm{Var}(\hat{\boldsymbol{\mu}})$ subject to the constraint $\sum_{i=1}^m \mathbf{W}_i = \mathbf{I}_k$.[3]

Lemma 11. *The weights* $\mathbf{W}_1,\ldots,\mathbf{W}_m$ *that minimize* $\mathrm{Var}(\hat{\boldsymbol{\mu}})$ *subject to the constraint* $\sum_{i=1}^m \mathbf{W}_i = \mathbf{I}_k$ *are given by*

$$\mathbf{W}_i = \left(\sum_{j=1}^{m}(\mathbf{T}+n_j^{-1}\boldsymbol{\Sigma})^{-1}\right)^{-1}(\mathbf{T}+n_i^{-1}\boldsymbol{\Sigma})^{-1} \tag{7.12}$$

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

The proof of this lemma is given in Appendix 7.7.4. This lemma shows that the weights in equation (7.12) minimize the variance of the generalized weighted mean. Hence, these optimal weights lead to the following "toy estimator":

$$\hat{\boldsymbol{\mu}}_{\text{opt}} = \left(\sum_{i=1}^{m}(\mathbf{T}+n_i^{-1}\boldsymbol{\Sigma})^{-1}\right)^{-1}\left(\sum_{i=1}^{m}(\mathbf{T}+n_i^{-1}\boldsymbol{\Sigma})^{-1}\overline{\mathbf{Z}}_i\right). \tag{7.13}$$

Since the weights in equation (7.12) yield minimum variance for $\hat{\boldsymbol{\mu}}$ we can thus conclude that, if the parameters $\boldsymbol{\Sigma}$ and $\mathbf{T}$ are known, $\hat{\boldsymbol{\mu}}_{\text{opt}}$ is the best of these three estimators.

However, in practice this result is not immediately useful because the optimal weights depend on the unknown parameters $\boldsymbol{\Sigma}$ and $\mathbf{T}$. If estimated values for these parameters are substituted in the optimal weights, this will influence the variance of the toy estimator in equation (7.13) and the resulting estimator will be biased. For example, the multivariate analysis of variance estimators [41] for $\boldsymbol{\Sigma}$ and $\mathbf{T}$ could be used, which are given by

$$\begin{aligned}
\hat{\boldsymbol{\Sigma}} &= \frac{1}{N-m}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\mathbf{z}_{ij}-\overline{\mathbf{z}}_{i\cdot})(\mathbf{z}_{ij}-\overline{\mathbf{z}}_{i\cdot})^T \qquad \text{where } \overline{\mathbf{z}}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{z}_{ij},\\
\hat{\mathbf{T}} &= \frac{\mathrm{MS}^2_{\text{between}}-\hat{\boldsymbol{\Sigma}}}{\kappa} \qquad \text{where } \kappa = \frac{1}{m-1}\left(N-\frac{\sum_{i=1}^m n_i^2}{N}\right),\\
\mathrm{MS}^2_{\text{between}} &= \frac{1}{m-1}\sum_{i=1}^{m} n_i(\overline{\mathbf{z}}_{i\cdot}-\overline{\mathbf{z}})(\overline{\mathbf{z}}_{i\cdot}-\overline{\mathbf{z}})^T \text{ and } \overline{\mathbf{z}} = \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\mathbf{z}_{ij}.
\end{aligned} \tag{7.14}$$

The performance of the plug-in estimator $\hat{\boldsymbol{\mu}}_{\text{plug}}$ based on these estimates for $\boldsymbol{\Sigma}$ and $\mathbf{T}$ will be further evaluated in the following sections. Introducing the toy estimator gives more theoretical insight in the various estimators for the overall mean $\boldsymbol{\mu}$. In the results of the simulation study in Section 7.4 this will be further explored.

[3] If only diagonal matrices would be considered, a similar analysis shows that the matrix with weights $\mathbf{w}_i = (w_{i1},\ldots,w_{ik})^T$ on the diagonal that minimizes $\mathrm{Var}(\hat{\boldsymbol{\mu}})$ subject to the constraint $\sum_{i=1}^m \mathbf{w}_i = \mathbf{1}_k$ is found from

$$\mathbf{w}_i = \left(\sum_{j=1}^{m}\frac{1}{\mathrm{diag}(\mathbf{T}+n_j^{-1}\boldsymbol{\Sigma})}\right)^{-1}\frac{1}{\mathrm{diag}(\mathbf{T}+n_i^{-1}\boldsymbol{\Sigma})}, \qquad 1 \le i \le m,$$

where all vector products and divisions are elementwise. Choosing the diagonal matrix with these weights results in a better mean estimator in terms of variance than $\hat{\boldsymbol{\mu}}_w$ and $\hat{\boldsymbol{\mu}}_u$, but it will not be as good as $\hat{\boldsymbol{\mu}}_{\text{opt}}$, which is the optimal mean estimator.

7.4. Simulation study

In this section the mean estimators of Section 7.3 are compared in a simulation study. In Section 7.4.1 the performance of the weighted and unweighted mean estimators is compared using Monte Carlo simulation. In Section 7.4.2, this comparison is extended with the optimal generalized weighted mean estimator. Since this is a toy estimator and cannot be computed in practice, the optimal generalized weighted mean with estimates for the within- and between-source covariance matrices will also be considered, which will be called the optimal "plug-in" generalized weighted mean estimator. Finally, in Section 7.4.3 an artificial choice of parameters is used to show some examples where the optimal generalized weighted mean outperforms both the weighted and unweighted mean.

7.4.1. Weighted versus unweighted mean

In expression (7.9) we have seen that the best choice between the weighted and unweighted mean depends on the ratio of the traces of the within- and between-source covariance matrices $\mathbf{\Sigma}$ and $\mathbf{T}$. However, the covariance matrices $\mathbf{\Sigma}$ and $\mathbf{T}$ are unknown. Hence, to use expression (7.9) in practice, one should have prior knowledge about the ratio between $tr(\mathbf{\Sigma})$ and $tr(\mathbf{T})$. In many comparison problems the trace of the within-source covariance matrix can be assumed to be smaller than the trace of the between-source covariance matrix. Furthermore, in Lemma 10 it is shown that the value of the constant c will always be larger than one. Therefore, it is expected that in most cases the unweighted mean has a smaller variance than the weighted mean.

Given this result, it is interesting to compare the performance of the weighted and the unweighted mean in estimating the true mean $\boldsymbol{\mu}$. To this end, we perform two Monte Carlo simulations. In these simulations, the values for the number of groups m are set to $m = 10$ and $m = 1200$, respectively, and the number of measurements in each group $n_i, 1 \leq i \leq m$, is drawn randomly, where values $1 \leq n_i \leq 20$ are used. Given these values of n_i and m, a background data set is generated M times according to the model described in Section 7.2.3. To simulate the background data set in both situations, the parameters $\boldsymbol{\mu}$, $\mathbf{\Sigma}$ and $\mathbf{T}$ are fixed based on diameter (in millimeters), thickness (in millimeters) and weight (in milligrams) observations in real XTC tablet comparisons. These values are given by:

$$\boldsymbol{\mu} = \begin{bmatrix} 8.242 \\ 4.528 \\ 276.0 \end{bmatrix}, \quad \mathbf{\Sigma} = \begin{bmatrix} 0.002013 & 0.0007271 & 0.01408 \\ 0.0007271 & 0.03046 & 0.6133 \\ 0.01408 & 0.6133 & 90.61 \end{bmatrix}, \quad \mathbf{T} = \begin{bmatrix} 0.6026 & 0.06689 & 31.56 \\ 0.06689 & 0.6371 & 32.90 \\ 31.56 & 32.90 & 3562 \end{bmatrix}. \tag{7.15}$$

The results of the two Monte Carlo simulations are given for each element of the three-dimensional estimated mean vector and can be found in the box plots in Figure 7.1. From these figures it can be seen that the estimated values of the two mean estimators are close. As can be expected, if there are more observations in the background data (1200 groups), the estimates are more accurate compared to the estimates using fewer observations (10 groups).

The mean squared error (MSE) [36] is chosen as a measure of performance for the estimators. The MSE measures the average of the squared values of the errors, i.e. the Euclidean distance between the estimate and the true value $\boldsymbol{\mu}$:

$$\mathrm{E}\left[||\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}||^2\right].$$

Hence, a mean squared error of zero means that the estimator estimates the true mean $\boldsymbol{\mu}$ perfectly. The estimators can be compared by using their MSEs, where the smallest MSE is preferred. For the unbiased weighted and unweighted mean, the MSE equals the variance of the estimators. Hence, minimizing the mean squared error is equivalent to minimizing the variance and the estimators with the lowest MSE are thus the most efficient.

To compute the MSE based on the Monte Carlo simulation, for each simulation i, with $1 \leq i \leq M$, the squared Euclidean distance between the estimate and the true value is computed. After M simulations the average over these squared distances is taken as the (numerically approximated) mean squared error. The resulting mean squared errors are given in Table 7.1.

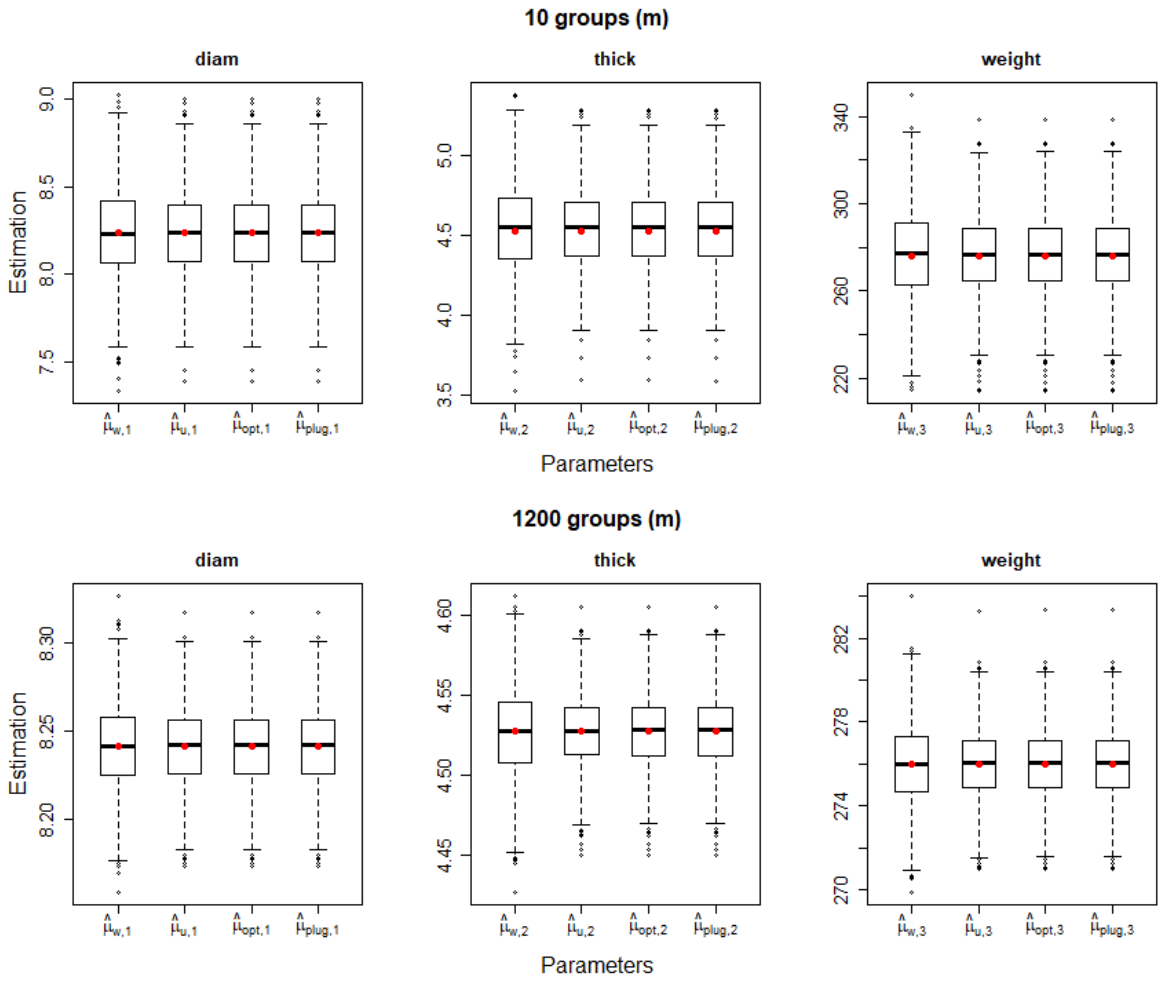

Figure 7.1: Box plots of estimated values from $\hat{\boldsymbol{\mu}}_w$, $\hat{\boldsymbol{\mu}}_u$, $\hat{\boldsymbol{\mu}}_{\text{opt}}$ and $\hat{\boldsymbol{\mu}}_{\text{plug}}$ for two Monte Carlo simulations ($M = 1000$) with parameters given as in equation (7.15). The red dot indicates the true overall mean value.

m	MSE $\hat{\boldsymbol{\mu}}_w$	MSE $\hat{\boldsymbol{\mu}}_u$	MSE $\hat{\boldsymbol{\mu}}_{\text{opt}}$	MSE $\hat{\boldsymbol{\mu}}_{\text{plug}}$
10	406	352	352	352
1200	3.93	2.96	2.96	2.96

Table 7.1: Mean squared errors of the estimated mean using the weighted mean $\hat{\boldsymbol{\mu}}_w$, the unweighted mean $\hat{\boldsymbol{\mu}}_u$, the toy estimator $\hat{\boldsymbol{\mu}}_{\text{opt}}$ and the plug-in estimator $\hat{\boldsymbol{\mu}}_{\text{plug}}$ for two Monte Carlo simulations ($M = 1000$) with parameters as given in equation (7.15).

From these MSE values it is clear that the performance of the estimators increases when the number of groups m is higher. Since the MSEs of the unweighted mean are smaller than the MSEs of the weighted mean, the unweighted mean should be preferred over the weighted mean. For both simulations the constant c can be computed and equals $c = 4.48$ for 10 groups and $c = 3.43$ for 1200 groups and with $tr(\boldsymbol{\Sigma}) = 90.6$ and $tr(\mathbf{T}) = 3563$ it can be seen that $tr(\boldsymbol{\Sigma}) < c \cdot tr(\mathbf{T})$. Consequently, from expression (7.9) it follows that the variance of the unweighted mean is smaller than the variance of the weighted mean.

Since the values for the overall mean $\boldsymbol{\mu}$ and the covariance matrices $\boldsymbol{\Sigma}$ and $\mathbf{T}$ are fixed, it is possible to determine the true value of the likelihood ratio for this problem. Therefore, five measurements for both the control and recovered data are generated, assuming that the prosecutor's hypothesis is true, i.e., that the control and recovered item originate from the same source. Using equation (7.2) with the parameters given in equation (7.15), the true value of the likelihood ratio is found. Keeping the covariance matrices $\boldsymbol{\Sigma}$ and $\mathbf{T}$ fixed, the likelihood ratios based on $\hat{\boldsymbol{\mu}}_w$ and $\hat{\boldsymbol{\mu}}_u$ can also be calculated. The approximated mean squared error for the likelihood ratio values is then computed by

$$\frac{1}{M}\sum_{i=1}^{M}\left[\mathrm{LR}(\bar{\mathbf{y}}_1,\bar{\mathbf{y}}_2|\boldsymbol{\mu})-\mathrm{LR}(\bar{\mathbf{y}}_1,\bar{\mathbf{y}}_2|\hat{\boldsymbol{\mu}}(i))\right]^2$$

for each Monte Carlo simulation i, with $1 \leq i \leq M$. The resulting mean squared errors can be found in Table 7.2.

Clearly, the MSE values of the likelihood ratios reduce significantly when the number of groups m is higher. Moreover, the performance of the unweighted mean is significantly better than the performance of the weighted mean. Combining this observation with the fact that the unweighted mean is more efficient than the weighted mean, the unweighted mean should in this situation be preferred over the weighted mean.

m	MSE $\mathrm{LR}(\bar{\mathbf{y}}_1,\bar{\mathbf{y}}_2\|\hat{\boldsymbol{\mu}}_w)$	MSE $\mathrm{LR}(\bar{\mathbf{y}}_1,\bar{\mathbf{y}}_2\|\hat{\boldsymbol{\mu}}_u)$	MSE $\mathrm{LR}(\bar{\mathbf{y}}_1,\bar{\mathbf{y}}_2\|\hat{\boldsymbol{\mu}}_{\mathrm{opt}})$	MSE $\mathrm{LR}(\bar{\mathbf{y}}_1,\bar{\mathbf{y}}_2\|\hat{\boldsymbol{\mu}}_{\mathrm{plug}})$
10	$3.97\cdot10^6$	$2.61\cdot10^6$	$2.60\cdot10^6$	$2.61\cdot10^6$
1200	$6.34\cdot10^3$	$4.99\cdot10^3$	$4.96\cdot10^3$	$4.96\cdot10^3$

Table 7.2: Mean squared errors of the estimated likelihood ratio using the weighted mean $\hat{\boldsymbol{\mu}}_w$, the unweighted mean $\hat{\boldsymbol{\mu}}_u$, the toy estimator $\hat{\boldsymbol{\mu}}_{\mathrm{opt}}$ and the plug-in estimator $\hat{\boldsymbol{\mu}}_{\mathrm{plug}}$ for two Monte Carlo simulations ($M = 1000$) with parameters as given in equation (7.15). The true likelihood ratio is equal to $1.11\cdot10^3$.

7.4.2. Generalized weighted mean

In the Monte Carlo simulations in Section 7.4.1 the values for the covariance matrices $\boldsymbol{\Sigma}$ and $\mathbf{T}$ are fixed, see equation (7.15). Substituting these values into the toy estimator in equation (7.13), the toy estimator yields the minimum variance estimator. It is therefore interesting to examine the difference between this estimator and the weighted and unweighted mean that can be used in practice more easily. We will also consider the plug-in estimator based on the multivariate analysis of variance estimates for $\boldsymbol{\Sigma}$ and $\mathbf{T}$, given by (7.14). Note that the plug-in estimator is a biased estimator, which motivates the use of the mean squared error to compare the mean estimators and not only the variance. To compare the performance of the toy estimator and the plug-in estimator with the performance of the weighted and unweighted mean, the simulations as described in Section 7.4.1 based on the same values of m and corresponding n_i's are used. The results of these Monte Carlo simulations are given in Table 7.1 and 7.2 and Figure 7.1.

An interesting observation from Table 7.1 and 7.2 is that the optimal generalized weighted mean has (approximately) the same mean squared errors as the unweighted mean in this simulation. This can be explained by the small value for the parameter $\boldsymbol{\Sigma}$ in comparison to the value for $\mathbf{T}$, see equation (7.15). Consequently, it follows that $\mathbf{T}+n_i^{-1}\boldsymbol{\Sigma}\approx\mathbf{T}$. Hence,

$$\mathrm{Cov}\left(\hat{\boldsymbol{\mu}}_u,\hat{\boldsymbol{\mu}}_u\right)\approx\frac{1}{m^2}\sum_{i=1}^{m}\mathbf{T}=\frac{\mathbf{T}}{m}\quad\text{and therefore}\quad\mathrm{Var}(\hat{\boldsymbol{\mu}}_u)\approx\frac{tr(\mathbf{T})}{m}.$$

The weight matrices for the optimal generalized weighted mean are approximately equal to

$$\mathbf{W}_i\approx\left(\sum_{i=1}^{m}\mathbf{T}^{-1}\right)^{-1}\mathbf{T}^{-1}=\frac{1}{m}\mathbf{I}_k$$

so that the variance of the optimal generalized weighted mean is approximately

$$\mathrm{Var}\left(\hat{\boldsymbol{\mu}}_{\mathrm{opt}}\right)\approx\sum_{i=1}^{m}tr\left(\frac{1}{m}\mathbf{I}_k\mathbf{T}\frac{1}{m}\mathbf{I}_k\right)=\frac{tr(\mathbf{T})}{m}.$$

Thus, if the within-source variation is small relative to the between-source variation it follows that

$$\mathrm{Var}\left(\hat{\boldsymbol{\mu}}_u\right)\approx\mathrm{Var}\left(\hat{\boldsymbol{\mu}}_{\mathrm{opt}}\right).$$

Hence, for such situations the unweighted mean is as good as the minimum variance estimator $\hat{\boldsymbol{\mu}}_{\mathrm{opt}}$. Note that the plug-in estimator $\hat{\boldsymbol{\mu}}_{\mathrm{plug}}$ also behaves similarly to the minimum variance estimator.

7.4.3. Artificial choice of parameters

For the covariance matrices $\mathbf{\Sigma}$ and $\mathbf{T}$ from equation (7.15), we have seen that the within-source variation $\mathbf{\Sigma}$ is very small so that $\mathbf{T} + n_i^{-1}\mathbf{\Sigma} \approx \mathbf{T}$ and therefore the unweighted mean is approximately as good as the minimum variance estimator $\hat{\boldsymbol{\mu}}_{\text{opt}}$. It is interesting to consider some situations where $\hat{\boldsymbol{\mu}}_{\text{opt}}$ outperforms both the weighted and unweighted mean estimator. To this end, the following artificial choice of parameters was made:

$$\boldsymbol{\mu} = \begin{bmatrix} 3 \\ 5 \\ 4 \end{bmatrix}, \quad \mathbf{\Sigma} = \begin{bmatrix} 0.3 & 0.0 & 0.3 \\ 0.0 & 0.1 & -0.2 \\ 0.3 & -0.2 & 0.8 \end{bmatrix}, \quad \mathbf{T} = \begin{bmatrix} 0.6 & 0.3 & 0.5 \\ 0.3 & 0.4 & 0.2 \\ 0.5 & 0.2 & 0.9 \end{bmatrix}. \tag{7.16}$$

Again a Monte Carlo simulation study is performed for $m = 1200$ groups, as was described in Section 7.4.1. The values of r and c for the simulated data set are equal to 0.632 and 3.49 respectively, so that the inequality $tr(\mathbf{\Sigma}) < c \cdot tr(\mathbf{T})$ holds. The mean squared errors of both the mean estimates and the likelihood ratio values are given in Table 7.3.

	$\hat{\boldsymbol{\mu}}_w$	$\hat{\boldsymbol{\mu}}_u$	$\hat{\boldsymbol{\mu}}_{\text{opt}}$	$\hat{\boldsymbol{\mu}}_{\text{plug}}$
MSE $\hat{\boldsymbol{\mu}}$	$2.13 \cdot 10^{-3}$	$1.76 \cdot 10^{-3}$	$1.73 \cdot 10^{-3}$	$1.73 \cdot 10^{-3}$
MSE LR$(\bar{\mathbf{y}}_1, \bar{\mathbf{y}}_2 \| \hat{\boldsymbol{\mu}})$	$4.54 \cdot 10^{-3}$	$3.59 \cdot 10^{-3}$	$3.56 \cdot 10^{-3}$	$3.56 \cdot 10^{-3}$

Table 7.3: Mean squared errors of the estimated mean and likelihood ratio using the weighted mean $\hat{\boldsymbol{\mu}}_w$, the unweighted mean $\hat{\boldsymbol{\mu}}_u$, the toy estimator $\hat{\boldsymbol{\mu}}_{\text{opt}}$ and the plug-in estimator $\hat{\boldsymbol{\mu}}_{\text{plug}}$ for a Monte Carlo simulation ($M = 1000$) with parameters as given in equation (7.16) and $m = 1200$ groups. The true likelihood ratio is equal to 2.47.

Indeed, the optimal generalized weighted mean performs better than the other overall mean estimators, although the performance is comparable to that of the unweighted mean estimator and the plug-in mean estimator.

Another interesting situation is when the inequality $tr(\mathbf{\Sigma}) < c \cdot tr(\mathbf{T})$ does not hold. Therefore, the parameter $\mathbf{\Sigma}$ is multiplied by 10 whereas the other parameters as well as the sampled n_i's remain unchanged. Again a Monte Carlo simulation study is performed for $m = 1200$ groups, but we now have $r = 6.32$ and $c = 3.49$ so that $tr(\mathbf{\Sigma}) > c \cdot tr(\mathbf{T})$. This means that the weighted mean should perform better than the unweighted mean. Note that the values of r and c do not influence $\hat{\boldsymbol{\mu}}_{\text{opt}}$ and that this is still the minimum variance unbiased estimator. The results of the simulation study can be found in Table 7.4.

	$\hat{\boldsymbol{\mu}}_w$	$\hat{\boldsymbol{\mu}}_u$	$\hat{\boldsymbol{\mu}}_{\text{opt}}$	$\hat{\boldsymbol{\mu}}_{\text{plug}}$
MSE $\hat{\boldsymbol{\mu}}$	$2.96 \cdot 10^{-3}$	$3.31 \cdot 10^{-3}$	$2.67 \cdot 10^{-3}$	$2.67 \cdot 10^{-3}$
MSE LR$(\bar{\mathbf{y}}_1, \bar{\mathbf{y}}_2 \| \hat{\boldsymbol{\mu}})$	$1.65 \cdot 10^{-2}$	$1.71 \cdot 10^{-2}$	$1.45 \cdot 10^{-2}$	$1.45 \cdot 10^{-2}$

Table 7.4: Mean squared errors of the estimated mean and likelihood ratio using the weighted mean $\hat{\boldsymbol{\mu}}_w$, the unweighted mean $\hat{\boldsymbol{\mu}}_u$, the toy estimator $\hat{\boldsymbol{\mu}}_{\text{opt}}$ and the plug-in estimator $\hat{\boldsymbol{\mu}}_{\text{plug}}$ for a Monte Carlo simulation ($M = 1000$) with parameters as given in equation (7.16), where $\mathbf{\Sigma}$ is multiplied by 10, and $m = 1200$ groups. The true likelihood ratio is equal to 4.58.

As expected, the weighted mean now performs better than the unweighted mean, but the optimal generalized weighted mean is still the best of all estimators. Again, the performance of the toy estimator and the plug-in estimator is similar.

7.5. Estimating the overall mean of XTC data

In this section, the different estimators will be applied to real XTC data to illustrate the impact of the choice of overall mean estimator. The XTC data come from the CHAMP (Collaborative Harmonization of Methods for Profiling of Amphetamine Type Stimulants) project. Instead of generating the control and recovered data $\mathbf{Y}_1$ and $\mathbf{Y}_2$ based on the parameters given in equation (7.15), it is also possible to apply the three mean estimators to real XTC trace evidence. Since the true mean $\boldsymbol{\mu}$ and the true likelihood ratio LR$(\bar{\mathbf{y}}_1, \bar{\mathbf{y}}_2|\boldsymbol{\mu})$ are now unknown, we cannot say anything about mean squared errors. Therefore, this application is purely meant to indicate the difference in results when using the weighted, unweighted or optimal plug-in generalized weighted mean. The latter will again be based on the multivariate analysis of variance estimates as given in equation (7.14). In fact, these are the same estimates as used to obtain the parameters $\mathbf{\Sigma}$ and $\mathbf{T}$ in equation (7.15) from the real XTC data.

The control data $\mathbf{Y}_1$ now consists of 42 measurements of the diameter, thickness and weight of tablets from consignment C_1, and the recovered data $\mathbf{Y}_2$ consists of 5 measurements on tablets that also come from consignment C_1. This means that the prosecutor's hypothesis is true and likelihood ratio values larger than 1 are expected. It is assumed that the origin of consignment C_1 is unknown, i.e., it is not known which production process produced the tablets, so that indeed the described two-level model applies to this situation. The background data consists of 186 consignments with two or more tablet measurements where it is not known whether there are links between the consignments. For this data set, we have $c = 11.0$ and $r = tr(\mathbf{\Sigma})/tr(\mathbf{T}) = 0.0254$, so that the inequality $tr(\mathbf{\Sigma}) < c \cdot tr(\mathbf{T})$ holds. The following estimates for the overall mean $\boldsymbol{\mu}$ are obtained:

$$\hat{\boldsymbol{\mu}}_w = \begin{bmatrix} 8.242 \\ 4.528 \\ 276.0 \end{bmatrix}, \qquad \hat{\boldsymbol{\mu}}_u = \begin{bmatrix} 8.240 \\ 4.211 \\ 260.0 \end{bmatrix}, \qquad \hat{\boldsymbol{\mu}}_{\text{plug}} = \begin{bmatrix} 8.240 \\ 4.212 \\ 260.1 \end{bmatrix}.$$

Using the same estimates from equations (7.14) for $\mathbf{\Sigma}$ and $\mathbf{T}$ and the likelihood ratio formula from equation (7.2), the likelihood ratio values can be calculated for each of the overall mean estimates:

$$\text{LR}(\bar{\mathbf{y}}_1, \bar{\mathbf{y}}_2 | \hat{\boldsymbol{\mu}}_w) = 1455, \qquad \text{LR}(\bar{\mathbf{y}}_1, \bar{\mathbf{y}}_2 | \hat{\boldsymbol{\mu}}_u) = 2073, \qquad \text{LR}(\bar{\mathbf{y}}_1, \bar{\mathbf{y}}_2 | \hat{\boldsymbol{\mu}}_{\text{plug}}) = 2072.$$

This shows that there is a significant difference in likelihood ratio values when using $\hat{\boldsymbol{\mu}}_w$ instead of $\hat{\boldsymbol{\mu}}_u$ or $\hat{\boldsymbol{\mu}}_{\text{plug}}$. The analysis in the previous sections showed that, since $tr(\mathbf{\Sigma}) < c \cdot tr(\mathbf{T})$, both $\hat{\boldsymbol{\mu}}_u$ and $\hat{\boldsymbol{\mu}}_{\text{plug}}$ outperform $\hat{\boldsymbol{\mu}}_w$. Hence, it would be strongly discouraged to use the weighted mean when reporting likelihood ratio values for this evidence set.

7.6. Conclusion

In this paper three estimators for the mean are presented, which can be used if the evidence is modelled as a two-level model using assumptions of multivariate normality: the weighted mean, the unweighted mean and a generalized weighted mean estimator. The choice of the estimator of the overall mean is important for the estimation of the between-source covariance matrix and for the calculation of the likelihood ratio. There is no consensus on which of these two estimators to use when the data are unbalanced. In this paper a relation is found which can be used to find the most efficient estimator and thus to decide whether the weighted or the unweighted mean should be used. The unweighted mean is preferred over the weighted mean if $tr(\mathbf{\Sigma}) < c \cdot tr(\mathbf{T})$, where the constant c depends on the number of groups in the background data and the number of measurements in each group. It is argued that in many forensic comparison problems the within-source variation can be assumed to be smaller than the between-source variation. Moreover, it is proven that the value of c will never be smaller than one. Therefore, it is expected that in practice the unweighted mean will often be preferred over the weighted mean. Of course, there might also be contextual reasons to prefer one of the overall mean estimators over the other.

The weights of the generalized weighted mean are derived such that they minimize the variance of this estimator. These optimal weights lead to a toy estimator, because they depend on the unknown within- and between-source covariance matrices. If these parameters would be known, the derived toy estimator has smaller (or equal) variance than the weighted and the unweighted mean. Using the optimal weights with estimates for the within- and between-source covariance matrices leads to a plug-in estimator. When comparing the multivariate mean estimators in a simulation study where the unweighted mean should be preferred over the weighted mean, the unweighted mean and plug-in estimator perform similarly to the toy estimator which yields minimum variance. Using an artificial choice of parameters provides some examples where the toy estimator outperforms both the weighted and unweighted mean, regardless of the number of groups and number of measurements in the background data. Applying the weighted mean, the unweighted mean and the plug-in mean estimator to real data shows the impact that the choice of estimator has on the value of evidence.

7.7. Appendix

7.7.1. The variance of $\hat{\boldsymbol{\mu}}_w$ and $\hat{\boldsymbol{\mu}}_u$

The covariance matrix of the weighted mean $\hat{\boldsymbol{\mu}}_w$ is found by setting $\mathbf{W}_i = \frac{n_i}{N}\mathbf{I}_k$ in equation (7.11) so that

$$\mathrm{Cov}(\hat{\boldsymbol{\mu}}_w, \hat{\boldsymbol{\mu}}_w) = \sum_{i=1}^{m} \frac{n_i}{N}\mathbf{I}_k(\mathbf{T} + n_i^{-1}\boldsymbol{\Sigma})\frac{n_i}{N}\mathbf{I}_k = \frac{\mathbf{T}}{N^2}\sum_{i=1}^{m} n_i^2 + \frac{\boldsymbol{\Sigma}}{N}.$$

By linearity of the trace, we have

$$\mathrm{Var}(\hat{\boldsymbol{\mu}}_w) = \frac{tr(\mathbf{T})}{N^2}\sum_{i=1}^{m} n_i^2 + \frac{tr(\boldsymbol{\Sigma})}{N}.$$

Similarly, the covariance matrix of the unweighted mean $\hat{\boldsymbol{\mu}}_u$ can be found by setting $\mathbf{W}_i = \frac{1}{m}\mathbf{I}_k$ in equation (7.11) so that

$$\mathrm{Cov}(\hat{\boldsymbol{\mu}}_u, \hat{\boldsymbol{\mu}}_u) = \sum_{i=1}^{m} \frac{1}{m}\mathbf{I}_k(\mathbf{T} + n_i^{-1}\boldsymbol{\Sigma})\frac{1}{m}\mathbf{I}_k = \frac{\mathbf{T}}{m} + \frac{\boldsymbol{\Sigma}}{m^2}\sum_{i=1}^{m} \frac{1}{n_i}$$

and linearity of the trace gives

$$\mathrm{Var}(\hat{\boldsymbol{\mu}}_u) = \frac{tr(\mathbf{T})}{m} + \frac{tr(\boldsymbol{\Sigma})}{m^2}\sum_{i=1}^{m} \frac{1}{n_i}.$$

7.7.2. The efficiency of $\hat{\boldsymbol{\mu}}_u$ relative to $\hat{\boldsymbol{\mu}}_w$

To find the relation of the efficiency as given in expression (7.9), Jensen's inequality can be used. Consider the random variable U, uniformly distributed on $n_1, \ldots, n_m$, ordered integers ≥ 1. By Jensen's inequality it follows that

$$\frac{1}{m}\sum_{i=1}^{m}\frac{1}{n_i} = \mathrm{E}\left[\frac{1}{U}\right] \geq \frac{1}{\mathrm{E}[U]} = \frac{1}{\frac{1}{m}\sum_{i=1}^{m} n_i} = \frac{m}{N}. \tag{7.17}$$

Here is used that the function $\phi(x) = \frac{1}{x}$ is convex for $x > 0$, which is sufficient since only positive values are considered. Applying Jensen's inequality to the function $\phi(x) = x^2$ it follows that

$$\frac{1}{m}\sum_{i=1}^{m} n_i^2 = \mathrm{E}\left[U^2\right] \geq (\mathrm{E}[U])^2 = \left(\frac{1}{m}\sum_{i=1}^{m} n_i\right)^2 = \frac{N^2}{m^2}. \tag{7.18}$$

Moreover, from inequality (7.18) it follows that

$$\frac{tr(\mathbf{T})}{N^2}\sum_{i=1}^{m} n_i^2 \geq \frac{tr(\mathbf{T})}{N^2}\frac{mN^2}{m^2} = \frac{tr(\mathbf{T})}{m},$$

which refers to the first terms in the numerator and denominator of equation (7.6). On the other hand, from inequality (7.17) it follows that

$$\frac{tr(\boldsymbol{\Sigma})}{m^2}\sum_{i=1}^{m}\frac{1}{n_i} \geq \frac{tr(\boldsymbol{\Sigma})}{N},$$

which refers to the second terms in the denominator and numerator of equation (7.6).

7.7.3. Proof of Lemma 10

Multiplying both the numerator and the denominator by $\frac{1}{m^3}$ and using $N = \sum_{i=1}^{m} n_i$, the expression for c can be re-written to

$$c = \frac{\frac{1}{m}\sum_{i=1}^{m} n_i^2 - \left(\frac{1}{m}\sum_{i=1}^{m} n_i\right)^2}{\left(\frac{1}{m}\sum_{i=1}^{m} n_i\right)^2\left(\frac{1}{m}\sum_{i=1}^{m}\frac{1}{n_i}\right) - \frac{1}{m}\sum_{i=1}^{m} n_i}.$$

To simplify notation a bit, consider the random variable U as defined in Appendix 7.7.2. Then we can write

$$c^{-1} = \frac{(\mathrm{E}[U])^2\,\mathrm{E}[U^{-1}] - \mathrm{E}[U]}{\mathrm{Var}(U)}.$$

Consider the convex function ϕ on $[1, n_m]$ defined by

$$\phi(y) = y^{-1}.$$

Since ϕ is a convex function, the tangent lines to ϕ are below the graph of ϕ. The idea of the proof is to find a parabola that can be added to the tangent lines so that it will always be above the graph of ϕ, see Figure 7.2. It follows that for fixed $u \in [1, n_m]$ we have for any $y \in [1, n_m]$

$$\phi(y) \leq \phi(u) + \phi'(u)(y-u) + \frac{(y-u)^2}{u^2}. \tag{7.19}$$

Indeed,

$$\frac{1}{y} \leq \frac{1}{u} - \frac{1}{u^2}(y-u) + \frac{1}{u^2}(y-u)^2,$$

which can be re-written to

$$\frac{(y-1)(u-y)^2}{u^2 y} \geq 0$$

and holds as long as $u \geq 1$ and $y \geq 1$.

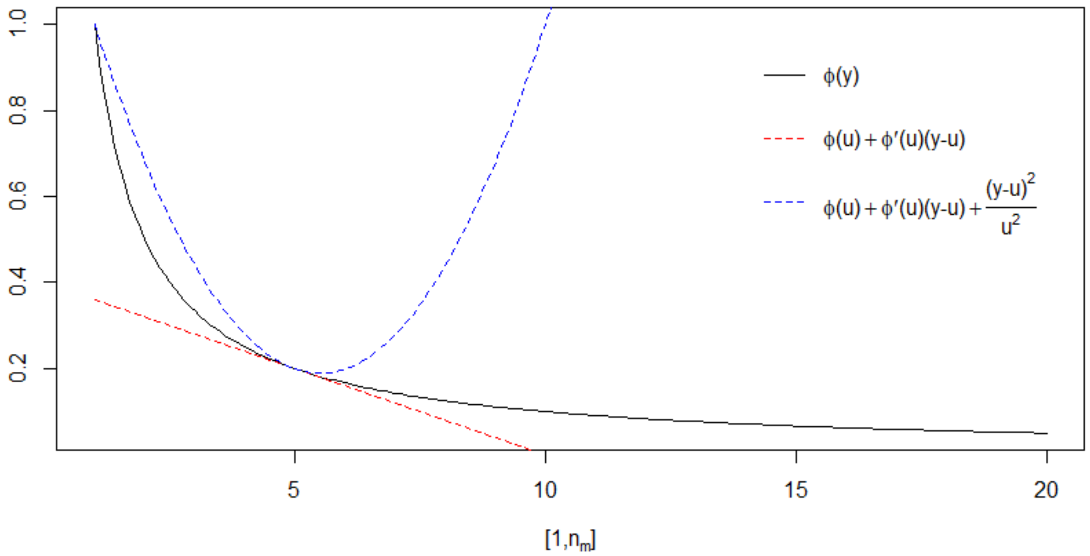

Figure 7.2: Illustration of equation (7.19) for $u = 5$.

Choosing $u = \mathrm{E}[U] \geq 1$ and substituting the random variable $U \geq 1$ for y results in

$$\frac{1}{U} \leq \frac{1}{\mathrm{E}[U]} - \frac{1}{(\mathrm{E}[U])^2}(U - \mathrm{E}[U]) + \frac{1}{(\mathrm{E}[U])^2}(U - \mathrm{E}[U])^2.$$

Now taking expectations, we get

$$\mathrm{E}[U^{-1}] \leq \frac{1}{\mathrm{E}[U]} + \frac{1}{(\mathrm{E}[U])^2}\mathrm{Var}(U).$$

Hence,

$$(\mathrm{E}[U])^2\,\mathrm{E}[U]^{-1} - \mathrm{E}[U] \leq \mathrm{Var}(U),$$

which implies that $c^{-1} \leq 1$, i.e., $c \geq 1$.

7.7.4. Proof of Lemma 11

To minimize Var($\hat{\boldsymbol{\mu}}$) subject to the constraint $\mathbf{W}_1 + \cdots + \mathbf{W}_m = \mathbf{I}_k$ a k^2-dimensional Lagrange multiplier $\boldsymbol{\lambda} = (\lambda_{11}, \lambda_{12}, \ldots, \lambda_{kk})$ is introduced such that the Lagrange function is equal to:

$$\mathscr{L}_{\boldsymbol{\lambda}}(\mathbf{W}_1,\ldots,\mathbf{W}_m,\lambda_{11},\lambda_{12},\ldots,\lambda_{kk}) = f(\mathbf{W}_1,\ldots,\mathbf{W}_m) - \sum_{j=1}^{k}\sum_{l=1}^{k}\lambda_{jl}g_{jl}(\mathbf{W}_1,\ldots,\mathbf{W}_m),$$

where

$$f(\mathbf{W}_1,\ldots,\mathbf{W}_m) = \sum_{i=1}^{m} tr\left(\mathbf{W}_i(\mathbf{T}+n_i^{-1}\mathbf{\Sigma})\mathbf{W}_i^T\right)$$

and

$$g_{jl}(\mathbf{W}_1,\ldots,\mathbf{W}_m) = \left[\sum_{i=1}^{m}\mathbf{W}_i - \mathbf{I}_k\right]_{jl}.$$

I.e., $g_{jl}(\mathbf{W}_1,\ldots,\mathbf{W}_m)$ is equal to the matrix element with index jl. Let

$$\frac{\partial}{\partial \mathbf{W}_i} = \begin{bmatrix} \frac{\partial}{\partial w_{i,11}} & \frac{\partial}{\partial w_{i,12}} & \cdots & \frac{\partial}{\partial w_{i,1k}} \\ \frac{\partial}{\partial w_{i,21}} & \frac{\partial}{\partial w_{i,22}} & \cdots & \frac{\partial}{\partial w_{i,2k}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial}{\partial w_{i,k1}} & \frac{\partial}{\partial w_{i,k2}} & \cdots & \frac{\partial}{\partial w_{i,kk}} \end{bmatrix}$$

denote the derivative with respect to the matrix $\mathbf{W}_i$. Then we have

$$\frac{\partial}{\partial \mathbf{W}_i}\left(\sum_{j=1}^{m} tr\left(\mathbf{W}_j(\mathbf{T}+n_j^{-1}\mathbf{\Sigma})\mathbf{W}_j^T\right)\right) = \frac{\partial}{\partial \mathbf{W}_i} tr\left(\mathbf{W}_i(\mathbf{T}+n_i^{-1}\mathbf{\Sigma})\mathbf{W}_i^T\right) = 2\mathbf{W}_i(\mathbf{T}+n_i^{-1}\mathbf{\Sigma})$$

since $(\mathbf{T}+n_i^{-1}\mathbf{\Sigma})$ is symmetric and $\dfrac{\partial\, tr(\mathbf{XAX}^T)}{\partial \mathbf{X}} = \mathbf{X}(\mathbf{A}+\mathbf{A}^T)$ [35]. Clearly,

$$\frac{\partial}{\partial w_{i,jl}} g_{jl}(\mathbf{W}_1,\ldots,\mathbf{W}_m) = 1$$

and zero for all other indices. Therefore, it follows that

$$\frac{\partial \mathscr{L}_{\boldsymbol{\lambda}}}{\partial \mathbf{W}_i} = 2\mathbf{W}_i(\mathbf{T}+n_i^{-1}\mathbf{\Sigma}) - \begin{bmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1k} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{k1} & \lambda_{k2} & \cdots & \lambda_{kk} \end{bmatrix} := 2\mathbf{W}_i(\mathbf{T}+n_i^{-1}\mathbf{\Sigma}) - \mathbf{\Lambda}$$

and the Lagrange function will be minimized over $\mathbb{R}^{k\times k}$. Setting the derivative equal to the $k \times k$ zero matrix results in

$$\mathbf{W}_i = \frac{1}{2}\mathbf{\Lambda}(\mathbf{T}+n_i^{-1}\mathbf{\Sigma})^{-1}, \qquad 1 \le i \le m.$$

Now using the constraint $\sum_{i=1}^{m}\mathbf{W}_i = \mathbf{I}_k$ gives

$$\sum_{i=1}^{m}\frac{1}{2}\mathbf{\Lambda}(\mathbf{T}+n_i^{-1}\mathbf{\Sigma})^{-1} = \mathbf{I}_k.$$

Hence,

$$\frac{1}{2}\mathbf{\Lambda} = \left(\sum_{i=1}^{m}(\mathbf{T}+n_i^{-1}\mathbf{\Sigma})^{-1}\right)^{-1}.$$

Thus,

$$\mathbf{W}_i = \left(\sum_{j=1}^{m}(\mathbf{T}+n_j^{-1}\mathbf{\Sigma})^{-1}\right)^{-1}(\mathbf{T}+n_i^{-1}\mathbf{\Sigma})^{-1}, \qquad 1 \le i \le m$$

which proves the lemma.

8

Copula models

In the sampling models described in Chapter 2 it is assumed that the sources of the forensic evidence are generated from a certain multivariate distribution. The sources are represented by a k-dimensional vector consisting of k relevant features and this multivariate distribution models the dependencies and distribution of these features simultaneously. However, considering for example Figures 6.6 and 6.7 from Chapter 6, it is not immediately evident that using the two-level normal-normal model provides the best fit to each type of data. One could argue that it would make more sense to first model the marginal distribution of each of the k features and then consider the dependencies between the features separately. This could be accomplished by replacing the multivariate distribution of the sources with a *copula.*

Definition 12. *A k-*dimensional copula *(or k-*copula*) is a function C from $[0,1]^k$ to $[0,1]$ with the following properties:*

1. *For every $\mathbf{u}$ in $[0,1]^k$, $C(\mathbf{u}) = 0$ if at least one coordinate of $\mathbf{u}$ is 0, and if all coordinates of $\mathbf{u}$ are 1 except u_i, then $C(\mathbf{u}) = u_i$;*
2. *For every $\mathbf{a}$ and $\mathbf{b}$ in $[0,1]^k$ such that $\mathbf{a} \leq \mathbf{b}$, i.e., $a_i \leq b_i$ for all $i = 1,2,\dots,k$, $V_C([\mathbf{a},\mathbf{b}]) \geq 0$.*

Here, $V_C([\mathbf{a},\mathbf{b}])$ denotes the C-volume of $[\mathbf{a},\mathbf{b}] = [a_1,b_1] \times [a_2,b_2] \times \cdots \times [a_k,b_k]$ given by

$$V_C([\mathbf{a},\mathbf{b}]) = \sum sgn(\mathbf{t})C(\mathbf{t}),$$

where the sum is taken over all vertices $\mathbf{t}$ of $[\mathbf{a},\mathbf{b}]$, and $sgn(\mathbf{t})$ is given by

$$sgn(\mathbf{t}) = \begin{cases} 1, & \text{if } t_k = a_k \text{ for an even number of } k\text{'s}, \\ -1, & \text{if } t_k = a_k \text{ for an odd number of } k\text{'s}. \end{cases}$$ *[30]*

Theorem 13 (Sklar's theorem in k dimensions)**.** *Let $\overline{\mathbb{R}}$ denote the extended real line $[-\infty,\infty]$. Let H be a k-dimensional distribution function with margins $F_1, F_2, \dots, F_k$. Then there exists a k-copula C such that for all $\mathbf{x}$ in $\overline{\mathbb{R}}^k$,*

$$H(x_1,x_2,\dots,x_k) = C\big(F_1(x_1),F_2(x_2),\dots,F_k(x_k)\big). \tag{8.1}$$

If $F_1,F_2,\dots,F_k$ are all continuous, then C is unique; otherwise, C is uniquely determined on Range(F_1) × Range(F_2) × ⋯ × Range(F_k). Conversely, if C is a k-copula and $F_1,F_2,\dots,F_k$ are distribution functions, then the function H defined by (8.1) is a k-dimensional distribution function with margins $F_1,F_2,\dots,F_k$. [30]

Let $\mathbf{A}_i = (A_{i1},A_{i2},\dots,A_{ik})^T$ denote the k-dimensional vector corresponding to the ith source. Suppose that for fixed m

$$A_{im} \overset{\text{iid}}{\sim} G_m(\cdot), \qquad \text{for } i = 1,2,\dots,n_a,$$

i.e., $G_m(\cdot)$ denotes the marginal distribution for feature m. Then, by Sklar's theorem, there exists a k-dimensional copula C to model the dependencies between the features, so that

$$\mathbf{A}_i \overset{\text{iid}}{\sim} C\big(G_1(\cdot),G_2(\cdot),\dots,G_k(\cdot)\big), \qquad \text{for } i = 1,2,\dots,n_a.$$

Then the elements within source i can be sampled from the within-source distribution as before:

$$\mathbf{Y}_{ij}|\mathbf{A}_i = \mathbf{a}_i \overset{\text{iid}}{\sim} F_a(\cdot|\mathbf{a}_i, \boldsymbol{\theta}_a), \qquad \text{for } j = 1,2,\dots,n_i.$$

This theoretical framework looks nice, but difficulties arise when putting this statistical model into practice. For example, assuming that c is the copula density corresponding to C and g_m is the probability density function corresponding to the marginal distribution G_m, the likelihoods from equation (2.9) and (2.10) for the common source problem would become

$$f(e_{u_1}, e_{u_2}|\boldsymbol{\theta}_a, H_p) = \int \prod_{j=1}^{n_u} f_a(\mathbf{y}_{uj}|\mathbf{p}, \boldsymbol{\theta}_a) h(\mathbf{p})\, d\mathbf{p}$$

and

$$f(e_{u_1}, e_{u_2}|\boldsymbol{\theta}_a, H_d) = \left(\int \prod_{j=1}^{n_{u_1}} f_a(\mathbf{y}_{u_1 j}|\mathbf{d}_1, \boldsymbol{\theta}_a) h(\mathbf{d}_1)\, d\mathbf{d}_1\right)\left(\int \prod_{j=1}^{n_{u_2}} f_a(\mathbf{y}_{u_2 j}|\mathbf{d}_2, \boldsymbol{\theta}_a) h(\mathbf{d}_2)\, d\mathbf{d}_2\right),$$

where

$$h(\mathbf{x}) = h(x_1, x_2, \dots, x_k) = c(G_1(x_1), G_2(x_2), \dots, G_k(x_k)) \prod_{m=1}^{k} g_m(x_m).$$

These integrals will in general be very difficult to evaluate. Moreover, a lot of distributions have to be fitted and a lot of parameters have to be estimated, either in a frequentist or Bayesian way; besides the within-source distribution, all k marginal distributions are needed as well as an appropriate copula to model their dependencies, and all these distributions will have unknown parameters. Since the A_{im} are latent variables, it might be hard to accurately determine the marginal distribution of the features. Furthermore, if the number of features is large it can be difficult to find a suitable copula outside the family of Archimedean copulas and one might even need to consider vine copula structures, which would make the model even more complex.

To overcome some of these problems, several approaches have been suggested in literature. Two of them will be discussed in the following sections.

8.1. Gaussian copula model

First of all, in [29] a *Gaussian copula* is used and maximum likelihood estimates for the copula and marginal distribution parameters are obtained through a two-stage estimation procedure.

Definition 14. *For a given correlation matrix $\boldsymbol{\Sigma} \in [-1,1]^{k\times k}$, the* Gaussian copula *with parameter matrix $\boldsymbol{\Sigma}$ is given by*

$$C(u_1, u_2, \dots, u_k; \boldsymbol{\Sigma}) = \boldsymbol{\Phi}_{\boldsymbol{\Sigma}}(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \dots, \Phi^{-1}(u_k)), \qquad u_m \in (0,1), \quad m = 1,2,\dots,k,$$

where $\boldsymbol{\Phi}_{\boldsymbol{\Sigma}}$ denotes the k-dimensional normal cumulative distribution function with mean zero and correlation matrix $\boldsymbol{\Sigma}$, and Φ^{-1} denotes the inverse of the standard univariate normal cumulative distribution function. Its density is given by

$$c(u_1, u_2, \dots, u_k; \boldsymbol{\Sigma}) = |\boldsymbol{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2}\mathbf{q}^T\boldsymbol{\Sigma}^{-1}\mathbf{q} + \frac{1}{2}\mathbf{q}^T\mathbf{q}\right] = |\boldsymbol{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2}\mathbf{q}^T(\boldsymbol{\Sigma}^{-1} - \mathbf{I}_k)\mathbf{q}\right],$$

where $\mathbf{q} = (q_1, q_2, \dots, q_k)^T$ with $q_m = \Phi^{-1}(u_m), m = 1,2,\dots,k,$ and $\mathbf{I}_k$ denotes the k-dimensional identity matrix. [50]

Let $G_m(\cdot|\psi_m)$ be some marginal distribution and let $u_m = G_m(p_m|\psi_m)$ for all $m = 1,2,\dots,k$. Using the Gaussian copula as proposed in [29], the likelihood of the evidence given H_p in the common source problem would become

$$\mathscr{L}(\boldsymbol{\psi}, \mathbf{p}, \boldsymbol{\Sigma}|\mathbf{y}_u) = |\boldsymbol{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2}\mathbf{q}^T(\boldsymbol{\Sigma}^{-1} - \mathbf{I}_k)\mathbf{q}\right] \prod_{j=1}^{n_u} f_a(\mathbf{y}_{uj}|\mathbf{p}) \prod_{m=1}^{k} g_m(p_m|\psi_m). \tag{8.2}$$

Note that here it is assumed that the within-source distribution F_a only depends on the source $\mathbf{p}$. When k is large, it can become difficult to estimate $\boldsymbol{\Sigma}$, since there are too many parameters. This can be solved by considering $\mathbf{\Sigma}$ as a function of a single parameter ρ corresponding to the Pearson correlation. Then $\mathbf{\Sigma} := \mathbf{\Sigma}(\rho)$ can have either a uniform correlation structure

$$\mathbf{\Sigma}(\rho) = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \ddots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}$$

which could be used for a intra-class correlation model, or a serial correlation structure

$$\mathbf{\Sigma}(\rho) = \begin{bmatrix} 1 & \rho & \cdots & \rho^{k-1} \\ \rho & 1 & \cdots & \rho^{k-2} \\ \vdots & \ddots & \ddots & \vdots \\ \rho^{k-1} & \rho^{k-2} & \cdots & 1 \end{bmatrix}$$

corresponding to a first order autoregressive correlation model [52] .

The likelihood (8.2) can be written as $\mathscr{L}(\boldsymbol{\psi}, \mathbf{p}, \mathbf{\Sigma}|\mathbf{y}_u) = \mathscr{L}_C \times \mathscr{L}_{\mathbf{y}_u,\mathbf{p}}$, where

$$\mathscr{L}_C = |\mathbf{\Sigma}(\rho)|^{-1/2} \exp\left[-\frac{1}{2}\mathbf{q}^T(\mathbf{\Sigma}(\rho)^{-1} - \mathbf{I}_k)\mathbf{q}\right] \qquad \text{and} \qquad \mathscr{L}_{\mathbf{y}_u,\mathbf{p}} = \prod_{j=1}^{n_u} f_a(\mathbf{y}_{uj}|\mathbf{p}) \prod_{m=1}^{k} g_m(p_m|\psi_m).$$

In the first stage of the estimation procedure, $\boldsymbol{\psi}$ is estimated by integrating the p_m out of $\mathscr{L}_{\mathbf{y}_u,\mathbf{p}}$ and optimizing the result,

$$\mathscr{L}_{\mathbf{y}_u} = \int \prod_{j=1}^{n_u} f_a(\mathbf{y}_{uj}|\mathbf{p}) \prod_{m=1}^{k} g_m(p_m|\psi_m)\, d\mathbf{p},$$

which can be obtained analytically for some choices of g_m and f_a. Then it is possible to use $\mathscr{L}_{\mathbf{y}_u,\mathbf{p}}$ with $\boldsymbol{\psi}$ fixed at $\hat{\boldsymbol{\psi}}$ to obtain $\hat{\mathbf{p}}$. In the second stage, ρ is estimated by optimizing $\mathscr{L}_C$ with $\mathbf{q}$ replaced by $\hat{\mathbf{q}}$, where

$$\hat{q}_m = \Phi^{-1}(G_m(\hat{p}_m|\hat{\psi}_m)).$$

Plugging the estimates $\hat{\boldsymbol{\psi}}$, $\hat{\mathbf{p}}$, $\hat{\mathbf{q}}$ and $\mathbf{\Sigma}(\hat{\rho})$ in the likelihood from equation (8.2) results in an estimate of the likelihood $f(e_{u_1}, e_{u_2}|H_p)$. This method can be repeated to estimate the other likelihood functions discussed in Chapter 3.

Note that this is quite a computational intensive procedure and would become even more complicated when the maximum likelihood estimates would be replaced by a Bayesian estimation procedure. Extending the model to a version where the within-source distribution also depends on a parameter $\boldsymbol{\theta}_a$ would require some extra optimization steps of $\mathscr{L}_{\mathbf{y}_u}$ in the first stage. However, the model lacks the possibility of incorporating background material into the estimation procedure; something that is highly valued in forensic science. Moreover, the model proposed here would result in different marginal distributions G_m under the prosecution or defence model, which contradicts the first modelling assumption described in Chapter 2.

8.2. Score-based copula model

A more simplified approach is considered in [46], where instead of features only *scores* are used. The measurements of two traces can be converted to scores using some similarity measure d. Let s_m denote the score obtained by comparing the mth feature from both traces for $m = 1, 2, \ldots, k$, i.e., $s_m = d(\mathbf{y}_{u_1,m}, \mathbf{y}_{u_2,m})$ for the common source problem, where $\mathbf{y}_{u_1,m}$ and $\mathbf{y}_{u_2,m}$ denote the vectors of the n_{u_1} or n_{u_2} observations, respectively, of feature m. These scores will have a different distribution f_{score} depending on which hypothesis is assumed to be true, so that the likelihood ratio would become

$$\frac{f_{\text{score}}(s_1, s_2, \ldots, s_k|H_p)}{f_{\text{score}}(s_1, s_2, \ldots, s_k|H_d)}.$$

Let $G_m(\cdot)$ denote the marginal distribution for score m. Both multivariate distributions can be decomposed using Sklar's theorem into

$$\frac{c(G_1(s_1|H_p), G_2(s_2|H_p), \ldots, G_k(s_k|H_p)|H_p)}{c(G_1(s_1|H_d), G_2(s_2|H_d), \ldots, G_k(s_k|H_d)|H_d)} \prod_{m=1}^{k} \frac{g_m(s_m|H_p)}{g_m(s_m|H_d)}. \tag{8.3}$$

Note that the last term is actually the product of k individual likelihood ratios, i.e.,

$$\prod_{m=1}^{k} \frac{g_m(s_m|H_p)}{g_m(s_m|H_d)} = \prod_{m=1}^{k} LR_m(s_m).$$

Equation (8.3) can be split into two parts, which will be evaluated separately: the copula fraction will be called the *correction factor* and the product of individual likelihood ratios is the *Naive Bayes part*.

For the correction factor, [46] proposes to use modified empirical distribution functions $\widehat{G}_m(\cdot)$ based on background material to model the marginal distributions given each hypothesis. Therefore, assume that the background material is transformed to n vectors of feature scores where for each vector it is known if either the prosecution hypothesis or the defence hypothesis is true. Let n_p and n_d denote the number of score vectors where H_p or H_d is true, respectively. If $s_{m,1}, s_{m,2}, \dots, s_{m,n_p}$ are all scores in the background material for feature m where H_p is true, the modified empirical distribution function for feature m is given by

$$\widehat{G}_m(s|H_p) = \frac{1}{n_p + 1} \sum_{i=1}^{n_p} I(s_{m,i} \leq s)$$

and a similar definition holds for $\widehat{G}_m(s|H_d)$. The copula densities $c(\cdot|H_p)$ and $c(\cdot|H_d)$ are determined by considering a finite set of possible copulas and choosing the one that gives the best result according to some self-defined performance measure.

The Naive Bayes part is computed using the method of *Pool Adjacent Violators (PAV)*. For every feature $m = 1, 2, \dots, k$, the scores from the background material are sorted and a posterior probability of 1 is assigned to scores where H_p is true and 0 where H_d is true. The PAV algorithm then searches for non-monotonic adjacent groups of probabilities and replaces it with the average of that group. Repeating this step until the whole sequence of probabilities is monotonically increasing results in an estimate $\hat{p}_m(\cdot)$ of the posterior probability $\mathbb{P}(H_p|\cdot)$ for feature m. Assuming that the prior $\mathbb{P}(H_p)$ is known, the Naive Bayes part can then be found from

$$\widehat{LR}_m(s_m) = \frac{\hat{p}_m(s_m)}{1 - \hat{p}_m(s_m)} \frac{1 - \mathbb{P}(H_p)}{\mathbb{P}(H_p)}.$$

Note that this model does not have the possibility to be adjusted to a Bayesian setup and that only a selection of possible copulas is considered, which are two severe restrictions of the score-based copula model.

The Netherlands Forensic Institute also uses a score-based approach for some casework. Instead of computing a score for every feature, the measurements are transformed to one single score so that the likelihood ratio becomes

$$\frac{f_{\text{score}}(d(\mathbf{y}_{u_1}, \mathbf{y}_{u_2})|H_p)}{f_{\text{score}}(d(\mathbf{y}_{u_1}, \mathbf{y}_{u_2})|H_d)} = \frac{f_{\text{score}}(s|H_p)}{f_{\text{score}}(s|H_d)}$$

for some similarity measure $d(\cdot,\cdot)$, where only the within-source score distribution $f_{\text{score}}(\cdot|H_p)$ and between-source score distribution $f_{\text{score}}(\cdot|H_d)$ need to be fitted. Further research should decide if modelling the feature scores separately would give better results and thus if the score-based copula approach is worth the extra work.

9

Discrete evidence

Forensic scientists often have to work with discrete evidence, for example when DNA traces are found at a crime scene. In this case, certain features of the DNA profile of the trace are compared with a suspect's DNA profile. If the features from the trace and the suspect are the same, one speaks of a match. Only certain features of the DNA profile are considered and a match does not necessarily mean that the suspect is the perpetrator: there may be other people in the population of potential perpetrators having the same features in their DNA profile. Therefore the features of the DNA profile of the trace are compared with a DNA database to be able to say something about the rarity of the features. This is reflected by the *random match probability*, i.e., the proportion of that profile among the population of potential perpetrators [12].

The proposed framework from Chapter 2 using the sampling models can be adopted for discrete evidence. One important difference with continuous evidence is that there is no variation of the features within the source. It is therefore useless to consider more than one sample per source, since all samples will be identical. The modelling assumptions still hold, but Assumption 2 will be reformulated to clarify the discrete setup:

Assumption 2* Given source $\mathbf{A}_i = \mathbf{a}_i$, the background samples $\mathbf{Y}_{ij}$ are identical to $\mathbf{A}_i$ and follow the k-dimensional distribution $F_a(\cdot|\mathbf{a}_i) = H(\mathbf{a}_i - \cdot\,)$. Here, $H(\mathbf{a}_i - \cdot\,)$ denotes the Heaviside function, which splits to product form for higher dimensional $\mathbf{a}_i$. The background samples $\mathbf{Y}_{ij}|\mathbf{A}_i = \mathbf{a}_i$ are random samples from within the source, i.e., $\mathbf{Y}_{ij}|\mathbf{A}_i = \mathbf{a}_i \overset{\text{iid}}{\sim} F_a(\cdot|\mathbf{a}_i)$.

The Heaviside function $H(\mathbf{a}_i - \cdot\,)$ is a degenerate distribution function that can be seen as a cumulative distribution function placing probability mass 1 at $\mathbf{a}_i$ and 0 elsewhere. This is the limiting case of a normal within-source distribution where the variance goes to zero. It might seem cumbersome to use this setup to enforce equality of the source and the background sample, but by doing so all expressions derived for the likelihood ratio and the Bayes Factor remain valid. Moreover, this shows that the sampling models in [32] indeed hold for all types of forensic evidence. Note that using the Heaviside functions in the two-level model is equivalent to considering a one-level model, where only the between-source distribution is of interest.

The likelihood functions for both the common source problem and the specific source problem simplify significantly for discrete evidence. Let $\mathbf{Y}_{i1}$ denote the k-dimensional column vector of features from the ith source for $i = 1, 2, \dots, n_a$. For $\mathbf{a}_i = (a_{i1}, a_{i2}, \dots, a_{ik})$, the probability mass function corresponding to $H(\mathbf{a}_i - \cdot\,) = \prod_{j=1}^{k} H(a_{ij} - \cdot\,)$ is the Dirac delta function $\delta(\mathbf{a}_i - \cdot\,) = \prod_{j=1}^{k} \delta(a_{ij} - \cdot\,)$. The delta function is defined such that any nonlinear multivariate real function can be expressed with delta functions and integrals as

$$f(x_1, x_2, \dots, x_n) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(\mu_1, \dots, \mu_n)\delta(\mu_1 - x_1) \cdots \delta(\mu_n - x_n)\, d\mu_1 \cdots\, d\mu_n,$$

see [13]. This definition will be used to derive the simplified likelihood functions for the common source and specific source problem in the next sections.

9.1. Common source problem

For the common source problem, the likelihood function from equation (2.8) reduces for discrete background material to

$$\begin{aligned} f(e_a|\boldsymbol{\theta}_a, H_p) = f(e_a|\boldsymbol{\theta}_a, H_d) &= \prod_{i=1}^{n_a}\left(\int f_a(\mathbf{y}_{i1}|\mathbf{a}_i)g(\mathbf{a}_i|\boldsymbol{\theta}_a)\, d\mathbf{a}_i\right)\\ &= \prod_{i=1}^{n_a}\left(\int \delta(\mathbf{a}_i-\mathbf{y}_{i1})g(\mathbf{a}_i|\boldsymbol{\theta}_a)\, d\mathbf{a}_i\right)\\ &= \prod_{i=1}^{n_a} g(\mathbf{y}_{i1}|\boldsymbol{\theta}_a). \end{aligned} \tag{9.1}$$

Given the prosecution hypothesis, the unknown source evidence $e_u = \{e_{u_1}, e_{u_2}\}$ consists of the samples $\mathbf{Y}_{u_1 1}$ and $\mathbf{Y}_{u_2 1}$ that are both identical to $\mathbf{P} \sim G(\cdot|\boldsymbol{\theta}_a)$. Therefore, $\mathbf{Y}_{u_1 1}$ is also identical to $\mathbf{Y}_{u_2 1}$ and the likelihood function from equation (2.9) for the unknown source evidence becomes

$$\begin{aligned} f(e_{u_1}, e_{u_2}|\boldsymbol{\theta}_a, H_p) &= f(e_{u_2}|e_{u_1}, \boldsymbol{\theta}_a, H_p) f(e_{u_1}|\boldsymbol{\theta}_a, H_p)\\ &= \int f_a(\mathbf{y}_{u_1 1}|\mathbf{p})g(\mathbf{p}|\boldsymbol{\theta}_a)\, d\mathbf{p}\\ &= \int \delta(\mathbf{p}-\mathbf{y}_{u_1 1})g(\mathbf{p}|\boldsymbol{\theta}_a)\, d\mathbf{p}\\ &= g(\mathbf{y}_{u_1 1}|\boldsymbol{\theta}_a), \end{aligned}$$

where the property $f(e_{u_2}|e_{u_1}, \boldsymbol{\theta}_a, H_p) = 1$ is used, since for the discrete common source problem the evidence of the second unknown source is equal to the evidence of the first unknown source by definition of sampling model M_p.

Given the defence hypothesis, the first unknown source evidence e_{u_1} consists of a sample $\mathbf{Y}_{u_1 1}$ that is identical to $\mathbf{D}_1 \sim G(\cdot|\boldsymbol{\theta}_a)$ and the second unknown source evidence is a sample $\mathbf{Y}_{u_2 1}$ that is identical to $\mathbf{D}_2 \sim G(\cdot|\boldsymbol{\theta}_a)$, where $\mathbf{D}_1 \neq \mathbf{D}_2$. This means that the likelihood function from equation (2.10) shortens to

$$\begin{aligned} f(e_{u_1}, e_{u_2}|\boldsymbol{\theta}_a, H_d) &= f(e_{u_1}|\boldsymbol{\theta}_a, H_d) f(e_{u_2}|\boldsymbol{\theta}_a, H_d)\\ &= \left(\int f_a(\mathbf{y}_{u_1 1}|\mathbf{d}_1)g(\mathbf{d}_1|\boldsymbol{\theta}_a)\, d\mathbf{d}_1\right)\left(\int f_a(\mathbf{y}_{u_2 1}|\mathbf{d}_2)g(\mathbf{d}_2|\boldsymbol{\theta}_a)\, d\mathbf{d}_2\right)\\ &= \left(\int \delta(\mathbf{d}_1-\mathbf{y}_{u_1 1})g(\mathbf{d}_1|\boldsymbol{\theta}_a)\, d\mathbf{d}_1\right)\left(\int \delta(\mathbf{d}_2-\mathbf{y}_{u_2 1})g(\mathbf{d}_2|\boldsymbol{\theta}_a)\, d\mathbf{d}_2\right)\\ &= g(\mathbf{y}_{u_1 1}|\boldsymbol{\theta}_a)g(\mathbf{y}_{u_2 1}|\boldsymbol{\theta}_a). \end{aligned}$$

These derivations are consistent with the approach usually taken to model discrete forensic evidence. The problem is equivalent to

$$\mathbf{Y}_{u_1 1} \sim G(\cdot|\boldsymbol{\theta}_a)$$

and $\mathbf{Y}_{u_2 1}$ is equal to $\mathbf{Y}_{u_1 1}$ with probability 1 according to the prosecution hypothesis, and

$$\mathbf{Y}_{u_1 1} \sim G(\cdot|\boldsymbol{\theta}_a) \qquad \text{and} \qquad \mathbf{Y}_{u_2 1} \sim G(\cdot|\boldsymbol{\theta}_a) \qquad \text{independently}$$

according to the defence hypothesis.

9.2. Specific source problem

For the specific source problem, two extra modelling assumptions on the specific source evidence were formulated in Chapter 2. Assumption 3 remains unchanged in the discrete setup, but Assumption 4 cancels: Assumption 2* and 3 completely define the discrete specific source evidence, which is assumed to be known without uncertainty. Note that the parameter $\boldsymbol{\theta}_s$ is not defined in this setting.

Since the sampling model M_a of the specific source problem is identical to the one in the common source problem, the derivation of the likelihood function for e_a is also equivalent to the derivation given for the common source problem, resulting in equation (9.1).

Given the prosecution hypothesis, the unknown source evidence e_u consists of a sample $\mathbf{Y}_{u1}$ that is identical to the specific source evidence. This means that the likelihood is simply

$$f(e_u|H_p) = I(\mathbf{y}_{u1} = \mathbf{y}_{s1}),$$

where $I(\cdot)$ denotes the indicator function. Note that in practical discrete evidence evaluation the situation $\mathbf{y}_{u1} \neq \mathbf{y}_{s1}$ will never occur: it does not make sense to determine the value of evidence if it is already observed that the discrete features do not match (which would give a value of 0).

Given the defence hypothesis, the unknown source evidence e_u consists of one sample $\mathbf{Y}_{u1}$ that is identical to $\mathbf{D} \sim G(\cdot|\boldsymbol{\theta}_a)$. Therefore, the likelihood function from equation (2.15) becomes

$$f(e_u|\boldsymbol{\theta}_a, H_d) = \int f_a(\mathbf{y}_{u1}|\mathbf{d})g(\mathbf{d}|\boldsymbol{\theta}_a)\ d\mathbf{d} = \int \delta(\mathbf{d} - \mathbf{y}_{u1})g(\mathbf{d}|\boldsymbol{\theta}_a)\ d\mathbf{d} = g(\mathbf{y}_{u1}|\boldsymbol{\theta}_a).$$

These derivations are equivalent to considering the problem where it is assumed that $\mathbf{Y}_{u1} = \mathbf{y}_{s1}$ with probability 1 according to the prosecution hypothesis, and

$$\mathbf{Y}_{u1} \sim G(\cdot|\boldsymbol{\theta}_a)$$

according to the defence hypothesis.

In the next chapter, the derived likelihood functions will be used to quantify the value of evidence in a discrete setting.

10

One-level Bernoulli model

One frequently used model for discrete evidence in forensic science is the *one-level Bernoulli model.* Suppose a DNA trace is found at a crime scene originating from an unknown source, in particular from an unknown individual. It is natural to assume that each individual in the total population of sources matches with this DNA profile with probability θ_a. If every DNA profile from every source would be known, this probability would be equal to the number of individuals with this DNA profile divided by the size of the total population. However, in practice it is impossible to consider every individual in the total population of sources. Therefore θ_a is either estimated in a frequentist way or the Bayesian framework with a prior distribution on θ_a is used. Note that this framework immediately reduces the complex DNA profiles to one-dimensional Bernoulli random variables with probability of success equal to θ_a. Of course, it is also possible to consider certain characteristics of the DNA profile separately, leading to higher dimensional random variables.

Here, the Bayes Factor will be calculated for both the common source and the specific source problem in the one-dimensional situation. Suppose that

$$A_i \overset{\text{iid}}{\sim} Ber(\theta_a) \qquad \text{and} \qquad Y_{i1}|A_i = a_i \sim H(a_i - \cdot\,) \qquad \text{for } i = 1,2,\dots,n_a.$$

Since this enforces that Y_{i1} is identical to A_i for $i = 1,2,\dots,n_a$, it is equivalent to set

$$Y_{i1} \overset{\text{iid}}{\sim} Ber(\theta_a) \qquad \text{for } i = 1,2,\dots,n_a.$$

For a fully Bayesian setup, a prior distribution needs to be specified for θ_a. One common choice is a beta prior, because of the known conjugacy with the Bernoulli distribution. This leads to convenient closed form expressions of the Bayes Factor, as will be seen in the next sections. Therefore, set

$$\Theta_a \sim Beta(\alpha, \beta), \qquad \text{where } \alpha > 0, \beta > 0.$$

Conditional on the background material e_a, the parameters of the prior can be updated. The background material can be represented by $e_a = (y_{i1}, 1 \le i \le n_a)$ and $s_a = \sum_{i=1}^{n_a} y_{i1}$, so that

$$\Theta_a|e_a \sim Beta(\alpha + s_a, \beta + n_a - s_a).$$

The updated distribution of θ_a will be needed to calculate the common source and specific source Bayes Factor in this chapter.

10.1. Common source Bayes Factor

Recall from equation (3.2) that the Bayes Factor of the common source problem is given by

$$BF_{CS}(e) = \frac{\int f(e_{u_1}, e_{u_2}|\theta_a, H_p)\; d\Pi(\theta_a|e_a)}{\int f(e_{u_1}|\theta_a, H_d) f(e_{u_2}|\theta_a, H_d)\; d\Pi(\theta_a|e_a)}.$$

By definition of the problem, only the situation when $y_{u_1 1} = 1$ and $y_{u_2 1} = 1$ will be considered: the DNA profile of one of the two traces will be seen as ‘success’ and there is no use in reporting the value of evidence if the other trace does not match with this DNA profile. Note that this does not mean that the random variables $Y_{u_1 1}$ and $Y_{u_2 1}$ are the same. Using the simplified expressions of the likelihood functions for discrete evidence as found in Chapter 9, the Bayes Factor reduces to

$$BF_{CS}(e) = \frac{\int g(y_{u_1 1}|\theta_a)\pi(\theta_a|e_a)\ d\theta_a}{\int g(y_{u_1 1}|\theta_a)g(y_{u_2 1}|\theta_a)\pi(\theta_a|e_a)\ d\theta_a}.$$

The numerator can be calculated as follows:

$$\begin{aligned}\int g(y_{u_1 1}|\theta_a)\pi(\theta_a|e_a)\ d\theta_a &= \int \theta_a \frac{\Gamma(\alpha+\beta+n_a)}{\Gamma(\alpha+s_a)\Gamma(\beta+n_a-s_a)}\theta_a^{\alpha+s_a-1}(1-\theta_a)^{\beta+n_a-s_a-1}\ d\theta_a \\ &= \frac{\Gamma(\alpha+\beta+n_a)}{\Gamma(\alpha+s_a)\Gamma(\beta+n_a-s_a)}\int \theta_a^{\alpha+s_a}(1-\theta)^{\beta+n_a-s_a-1}\ d\theta_a \\ &= \frac{\Gamma(\alpha+\beta+n_a)}{\Gamma(\alpha+s_a)\Gamma(\beta+n_a-s_a)}\frac{\Gamma(\alpha+s_a+1)\Gamma(\beta+n_a-s_a)}{\Gamma(\alpha+\beta+n_a+1)} \\ &= \frac{(\alpha+\beta+n_a-1)!(\alpha+s_a)!}{(\alpha+s_a-1)!(\alpha+\beta+n_a)!} \\ &= \frac{\alpha+s_a}{\alpha+\beta+n_a},\end{aligned}$$

where in the second-last equation the identity $\Gamma(n) = (n-1)!$ is used.

Similarly, for the denominator of the Bayes Factor,

$$\begin{aligned}\int g(y_{u_1 1}|\theta_a)g(y_{u_2 1}|\theta_a)\pi(\theta_a|e_a)\ d\theta_a &= \int \theta_a\cdot\theta_a\cdot \frac{\Gamma(\alpha+\beta+n_a)}{\Gamma(\alpha+s_a)\Gamma(\beta+n_a-s_a)}\theta_a^{\alpha+s_a-1}(1-\theta_a)^{\beta+n_a-s_a-1}\ d\theta_a \\ &= \frac{\Gamma(\alpha+\beta+n_a)}{\Gamma(\alpha+s_a)\Gamma(\beta+n_a-s_a)}\int \theta_a^{\alpha+s_a+1}(1-\theta)^{\beta+n_a-s_a-1}\ d\theta_a \\ &= \frac{\Gamma(\alpha+\beta+n_a)}{\Gamma(\alpha+s_a)\Gamma(\beta+n_a-s_a)}\frac{\Gamma(\alpha+s_a+2)\Gamma(\beta+n_a-s_a)}{\Gamma(\alpha+\beta+n_a+2)} \\ &= \frac{(\alpha+\beta+n_a-1)!(\alpha+s_a+1)!}{(\alpha+s_a-1)!(\alpha+\beta+n_a+1)!} \\ &= \frac{(\alpha+s_a)(\alpha+s_a+1)}{(\alpha+\beta+n_a)(\alpha+\beta+n_a+1)}.\end{aligned}$$

Therefore, the Bayes Factor of the one-level Bernoulli common source problem is equal to

$$BF_{CS}(e) = \frac{\alpha+\beta+n_a+1}{\alpha+s_a+1}.$$

This expression corresponds to the value of evidence proposed in for example [14] and [48].

10.1.1. Alternative calculation: adding traces to background material

Since the Bayes Factor of the one-level Bernoulli model can be calculated analytically, it is also possible to use the expressions found in Section 3.1.1 to do the calculation. To use these, the likelihood ratio corresponding to this common source problem has to be derived first. Luckily, the likelihood ratio is easily found using the simplified likelihood functions for discrete evidence:

$$LR_{CS}(\theta_a; e_{u_1}, e_{u_2}) = \frac{f(e_{u_1}, e_{u_2}|\theta_a, H_p)}{f(e_{u_1}|\theta_a, H_d)f(e_{u_2}|\theta_a, H_d)} = \frac{g(y_{u_1 1}|\theta_a)}{g(y_{u_1 1}|\theta_a)g(y_{u_2 1}|\theta_a)} = \frac{\theta_a}{\theta_a^2} = \frac{1}{\theta_a}.$$

To use the relation between the likelihood ratio and the Bayes Factor from equation (3.3) for the calculation of the common source Bayes Factor, the prior distribution of θ_a given H_d and the entire evidence set is needed. Under the defence model, the unknown source evidence $e_u = \{e_{u_1}, e_{u_2}\}$ consists of two i.i.d. Bernoulli distributed random variables. Updating the prior for θ_a given the entire set of evidence thus results in

$$\Theta_a|e_a, e_{u_1}, e_{u_2}, H_d \sim Beta(\alpha+s_a+2, \beta+n_a-s_a).$$

This means that given H_d, both traces e_{u_1} and e_{u_2} are added to the background material to determine the distribution of $\theta_a|e_a, e_{u_1}, e_{u_2}$. Then, the Bayes Factor is found by

$$\begin{aligned}
BF_{CS}(e) &= \int LR_{CS}(\theta_a; e_{u_1}, e_{u_2})\pi(\theta_a|e_a, e_{u_1}, e_{u_2}, H_d)\ d\theta_a \\
&= \int \frac{1}{\theta_a}\frac{\Gamma(\alpha+\beta+n_a+2)}{\Gamma(\alpha+s_a+2)\Gamma(\beta+n_a-s_a)}\theta_a^{\alpha+s_a+1}(1-\theta_a)^{\beta+n_a-s_a-1}\ d\theta_a \\
&= \frac{\Gamma(\alpha+\beta+n_a+2)}{\Gamma(\alpha+s_a+2)\Gamma(\beta+n_a-s_a)}\int \theta_a^{\alpha+s_a}(1-\theta_a)^{\beta+n_a-s_a-1}\ d\theta_a \\
&= \frac{\Gamma(\alpha+\beta+n_a+2)}{\Gamma(\alpha+s_a+2)\Gamma(\beta+n_a-s_a)}\frac{\Gamma(\alpha+s_a+1)\Gamma(\beta+n_a-s_a)}{\Gamma(\alpha+\beta+n_a+1)} \\
&= \frac{\alpha+\beta+n_a+1}{\alpha+s_a+1}.
\end{aligned}$$

Similarly, to use equation (3.4) for the calculation of the Bayes Factor, the prior distribution of θ_a given H_p and the entire evidence set is needed. Under the prosecution model, the unknown source evidence consists of one Bernoulli distributed random variable. Updating the prior for θ_a given the entire evidence set gives

$$\Theta_a|e_a, e_{u_1}, e_{u_2}, H_p \sim Beta(\alpha+s_a+1, \beta+n_a-s_a).$$

This means that given H_p, only one trace is added to the background material to determine the distribution of $\theta_a|e_a, e_{u_1}, e_{u_2}$. Hence, the Bayes Factor can also be calculated by

$$\begin{aligned}
\frac{1}{BF_{CS}(e)} &= \int \frac{1}{LR_{CS}(\theta_a; e_{u_1}, e_{u_2})}\pi(\theta_a|e_a, e_{u_1}, e_{u_2}, H_p)\ d\theta_a \\
&= \int \theta_a \frac{\Gamma(\alpha+\beta+n_a+1)}{\Gamma(\alpha+s_a+1)\Gamma(\beta+n_a-s_a)}\theta_a^{\alpha+s_a}(1-\theta_a)^{\beta+n_a-s_a-1}\ d\theta_a \\
&= \frac{\Gamma(\alpha+\beta+n_a+1)}{\Gamma(\alpha+s_a+1)\Gamma(\beta+n_a-s_a)}\int \theta_a^{\alpha+s_a+1}(1-\theta_a)^{\beta+n_a-s_a-1}\ d\theta_a \\
&= \frac{\Gamma(\alpha+\beta+n_a+1)}{\Gamma(\alpha+s_a+1)\Gamma(\beta+n_a-s_a)}\frac{\Gamma(\alpha+s_a+2)\Gamma(\beta+n_a-s_a)}{\Gamma(\alpha+\beta+n_a+2)} \\
&= \frac{\alpha+s_a+1}{\alpha+\beta+n_a+1},
\end{aligned}$$

which gives

$$BF_{CS}(e) = \frac{\alpha+\beta+n_a+1}{\alpha+s_a+1}.$$

Note that for this problem both derivations of the Bayes Factor using the likelihood ratio result in significantly shorter calculations, since only one integral needs to be evaluated.

10.2. Specific source Bayes Factor

The Bayes Factor of the specific source problem was given in equation (3.6) and equals

$$BF_{SS}(e) = \frac{\int f(e_u|\theta_s, H_p)\ d\Pi(\theta_s|e_s)}{\int f(e_u|\theta_a, H_d)\ d\Pi(\theta_a|e_a)}.$$

By definition of the problem, only the situation when $y_{u1} = 1$ and $y_{s1} = 1$ will be considered: the DNA profile of the specific source will be seen as ‘success’ and there is no use in reporting the value of evidence if the unknown source trace does not match with this DNA profile. Plugging in the simplified expressions of the likelihood functions for discrete evidence from Chapter 9, the Bayes Factor becomes

$$BF_{SS}(e) = \frac{\int I(y_{u1} = y_{s1})\ d\Pi(\theta_s|e_s)}{\int g(y_{u1}|\theta_a)\pi(\theta_a|e_a)\ d\theta_a} = \frac{\int 1\ d\Pi(\theta_s|e_s)}{\int g(y_{u1}|\theta_a)\pi(\theta_a|e_a)\ d\theta_a} = \frac{1}{\int g(y_{u1}|\theta_a)\pi(\theta_a|e_a)\ d\theta_a}.$$

The denominator is equivalent to the numerator of the common source Bayes Factor, which gives

$$\int g(y_{u1}|\theta_a)\pi(\theta_a|e_a)\ d\theta_a = \int \theta_a \frac{\Gamma(\alpha+\beta+n_a)}{\Gamma(\alpha+s_a)\Gamma(\beta+n_a-s_a)}\theta_a^{\alpha+s_a-1}(1-\theta_a)^{\beta+n_a-s_a-1}\ d\theta_a = \frac{\alpha+s_a}{\alpha+\beta+n_a}.$$

The Bayes Factor for the one-level Bernoulli specific source problem thus equals

$$BF_{SS}(e) = \frac{\alpha+\beta+n_a}{\alpha+s_a}.$$

This expression is similar to the one obtained in [47] and [49], where the posterior mean of $\theta_a|e_a$ is used as a plug-in estimate for θ_a in the likelihood ratio to arrive at the value of evidence. Note that

$$BF_{CS}(e) = \frac{\alpha+\beta+n_a+1}{\alpha+s_a+1} < \frac{\alpha+\beta+n_a}{\alpha+s_a} = BF_{SS}(e) \quad \Longleftrightarrow \quad s_a < \beta+n_a$$

and since $s_a \leq n_a$ and $\beta > 0$ per definition, this will always be true. This means that the common source problem leads to a more conservative Bayes Factor than the specific source problem.

10.2.1. Alternative calculation: adding traces to background material

For the one-level Bernoulli specific source problem, the Bayes Factor can also be calculated using the relation with the likelihood ratio as can be found in Section 3.2.1. The likelihood ratio corresponding the specific source problem equals

$$LR_{SS}(\theta_a,\theta_s;e_u) = \frac{f(e_u|\theta_s,H_p)}{f(e_u|\theta_a,H_d)} = \frac{I(y_{u1}=y_{s1})}{g(y_{u1}|\theta_a)} = \frac{1}{\theta_a}.$$

To use equation (3.7), the prior distribution of (θ_a,θ_s) given H_d and the entire evidence set is needed. Since θ_s and θ_a are assumed to be independent, it is possible to consider both priors separately. Under the defence model, the distribution of $\theta_s|e_s,e_u,e_a$ does not depend on e_u and e_a, and the distribution of $\theta_a|e_s,e_u,e_a$ does not depend on e_s. The prior for θ_a can therefore be updated to

$$\Theta_a|e_u,e_a,H_d \sim Beta(\alpha+s_a+1,\beta+n_a-s_a).$$

Hence, only the trace e_u is added to the background material to determine the distribution of $\theta_a|e_a,e_u$. This gives for the specific source Bayes Factor:

$$\begin{aligned}
BF_{SS}(e) &= \int\int LR_{SS}(\theta_a,\theta_s;e_u)\pi(\theta_a,\theta_s|e_s,e_u,e_a,H_d)\ d\theta_a\ d\theta_s \\
&= \int\int \frac{1}{\theta_a}\pi(\theta_a|e_u,e_a,H_d)\pi(\theta_s|e_s,H_d)\ d\theta_a\ d\theta_s \\
&= \int \pi(\theta_s|e_s,H_d)\ d\theta_s \int \frac{1}{\theta_a}\pi(\theta_a|e_u,e_a,H_d)\ d\theta_a \\
&= 1\cdot\int \frac{1}{\theta_a}\frac{\Gamma(\alpha+\beta+n_a+1)}{\Gamma(\alpha+s_a+1)\Gamma(\beta+n_a-s_a)}\theta_a^{\alpha+s_a}(1-\theta_a)^{\beta+n_a-s_a-1}\ d\theta_a \\
&= \frac{\Gamma(\alpha+\beta+n_a+1)}{\Gamma(\alpha+s_a+1)\Gamma(\beta+n_a-s_a)}\int \theta_a^{\alpha+s_a-1}(1-\theta_a)^{\beta+n_a-s_a-1}\ d\theta_a \\
&= \frac{\Gamma(\alpha+\beta+n_a+1)}{\Gamma(\alpha+s_a+1)\Gamma(\beta+n_a-s_a)}\frac{\Gamma(\alpha+s_a)\Gamma(\beta+n_a-s_a)}{\Gamma(\alpha+\beta+n_a)} \\
&= \frac{\alpha+\beta+n_a}{\alpha+s_a}.
\end{aligned}$$

Similarly, to use equation (3.8) for the calculation of the Bayes Factor, the prior distribution of (θ_a,θ_s) given H_p and the entire evidence set is needed. Again, both priors for θ_a and θ_s can be considered separately. Under the prosecution model, the distribution of $\theta_s|e_s,e_u,e_a$ does not depend on e_a, and the distribution of $\theta_a|e_s,e_u,e_a$ does not depend on e_s and e_u. The distribution of $\theta_a|e_a$ was given before and does not depend on the hypothesis. This means that none of the traces is added to the background material. The derivation of the Bayes Factor becomes

$$\begin{aligned}
\frac{1}{BF_{SS}(e)} &= \int\int \frac{1}{LR_{SS}(\theta_a,\theta_s;e_u)}\pi(\theta_a,\theta_s|e_s,e_u,e_a,H_p)\ d\theta_a\ d\theta_s \\
&= \int\int \theta_a\ \pi(\theta_a|e_a,H_p)\pi(\theta_s|e_s,e_u,H_p)\ d\theta_a\ d\theta_s
\end{aligned}$$

$$
\begin{aligned}
&= \int \pi(\theta_s|e_s, e_u, H_p)\, d\theta_s \int \theta_a\, \pi(\theta_a|e_a, H_p)\, d\theta_a \\
&= 1 \cdot \int \theta_a \frac{\Gamma(\alpha+\beta+n_a)}{\Gamma(\alpha+s_a)\Gamma(\beta+n_a-s_a)} \theta_a^{\alpha+s_a-1}(1-\theta_a)^{\beta+n_a-s_a-1}\, d\theta_a \\
&= \frac{\Gamma(\alpha+\beta+n_a)}{\Gamma(\alpha+s_a)\Gamma(\beta+n_a-s_a)} \int \theta_a^{\alpha+s_a}(1-\theta_a)^{\beta+n_a-s_a-1}\, d\theta_a \\
&= \frac{\Gamma(\alpha+\beta+n_a)}{\Gamma(\alpha+s_a)\Gamma(\beta+n_a-s_a)} \frac{\Gamma(\alpha+s_a+1)\Gamma(\beta+n_a-s_a)}{\Gamma(\alpha+\beta+n_a+1)} \\
&= \frac{\alpha+s_a}{\alpha+\beta+n_a},
\end{aligned}
$$

so that

$$BF_{SS}(e) = \frac{\alpha+\beta+n_a}{\alpha+s_a}.$$

Note that for the specific source one-level Bernoulli problem the calculation of the Bayes Factor does not simplify using the alternative expressions.

The difference between the common source and specific source Bayes Factor might look small and therefore the discussion about which traces to add to the background material might sound irrelevant. However, in forensic casework the number of DNA matches s_a in a database is very often equal to zero (which is called a *rare type match problem* [12]). Considering Figure 10.1, it can be seen that for small values of s_a the difference between the common source and specific source Bayes Factor is significant. Therefore, it is important to indicate which identification of source problem is considered when reporting the value of evidence.

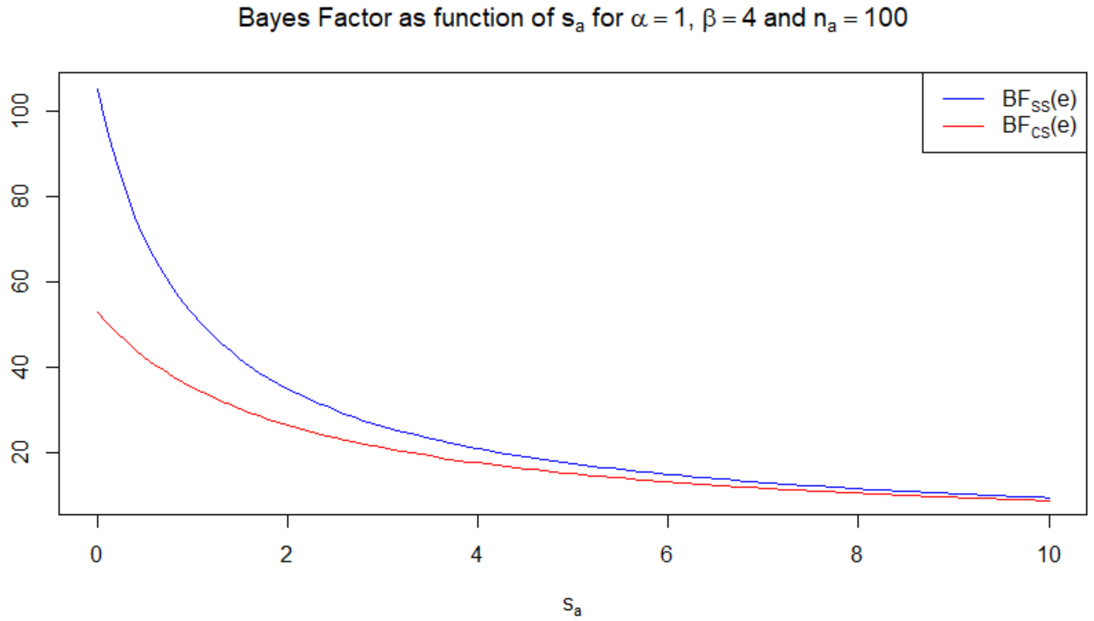

Figure 10.1: Common source and specific source Bayes Factor corresponding to the one-level Bernoulli problem as function of s_a, where $\alpha = 1$, $\beta = 4$ and $n_a = 100$.

11

Posterior probability of guilt

When evidence is found in a criminal case, the court is mainly interested in the probability that the prosecution hypothesis is true given all available evidence, i.e., $\mathbb{P}(H_p|e)$. This probability will be called the *posterior probability of guilt*. In Chapter 3 the posterior probability of guilt was already used in the posterior odds. Here, the posterior odds will be viewed both dependent and independent of the parameter θ. This will have a significant effect on the outcome of the posterior probability of guilt.

Two approaches can be taken when calculating the posterior probability of guilt:

(1) Using the Bayes Factor, the posterior odds can be expressed, independent of the parameter θ, as

$$\frac{\mathbb{P}(H_p|e)}{\mathbb{P}(H_d|e)} = BF(e) \times \frac{\mathbb{P}(H_p)}{\mathbb{P}(H_d)}.$$

Under the assumption that $\mathbb{P}(H_d|e) = 1 - \mathbb{P}(H_p|e)$, the posterior probability of guilt becomes

$$\mathbb{P}_{BF}(H_p|e) := \frac{\mathbb{P}(H_p) \times BF(e)}{\mathbb{P}(H_d) + \mathbb{P}(H_p) \times BF(e)} = \frac{\mathbb{P}(H_p) \times \int LR(\theta; e_u)\ d\Pi(\theta|e, H_d)}{\mathbb{P}(H_d) + \mathbb{P}(H_p) \times \int LR(\theta; e_u)\ d\Pi(\theta|e, H_d)},$$

where the last equality follows from the relation between the Bayes Factor and the likelihood ratio as discussed in Chapter 3.

(2) Using the likelihood ratio, the posterior odds can be expressed as a function of θ, i.e.,

$$\frac{\mathbb{P}(H_p|e,\theta)}{\mathbb{P}(H_d|e,\theta)} = LR(\theta; e_u) \times \frac{\mathbb{P}(H_p)}{\mathbb{P}(H_d)}.$$

Under the assumption that $\mathbb{P}(H_d|e,\theta) = 1 - \mathbb{P}(H_p|e,\theta)$, this can be written as

$$\mathbb{P}(H_p|e,\theta) = \frac{\mathbb{P}(H_p) \times LR(\theta; e_u)}{\mathbb{P}(H_d) + \mathbb{P}(H_p) \times LR(\theta; e_u)}.$$

Therefore, the posterior probability of guilt can be found by

$$\mathbb{P}_{LR}(H_p|e) := \int \mathbb{P}(H_p|e,\theta)\ d\Pi(\theta|e) = \int \frac{\mathbb{P}(H_p) \times LR(\theta; e_u)}{\mathbb{P}(H_d) + \mathbb{P}(H_p) \times LR(\theta; e_u)}\ d\Pi(\theta|e).$$

Intuitively, one would say that it does not matter which approach is taken, since in the end the same probabilities are calculated. However, mathematically there is a difference between both expressions for the posterior probability of guilt: the first approach considers the ratio of two integrals, while the second approach evaluates the integral of the same ratio, and these are in general not the same. Moreover, the integration in approach (1) is with respect to a different measure than the integration in approach (2).

The Netherlands Forensic Institute often uses the likelihood ratio in their casework and is interested if it is possible to combine this with a Bayesian prior on the parameter, as is done in the second approach. The question then arises, “What is the difference in the posterior probability of guilt obtained from the first and second approach, and how big is this difference?". This chapter will be devoted to answering these questions for the common source one-level Bernoulli problem, since for this problem the Bayes Factor can be calculated analytically, as seen in Chapter 10. The analysis can easily be repeated to evaluate the specific source one-level Bernoulli problem by using the insights from the previous chapter.

11.1. Theoretical differences and similarities

In many situations, the first approach to calculate the posterior probability of guilt will be straightforward. After the Bayes Factor is obtained from either an analytical calculation, a numerical approach or Markov Chain Monte Carlo methods, the probability follows directly from the formula given in approach (1). The second approach needs more clarification: here, the frequentist likelihood ratio is combined with a Bayesian prior. This means that the unknown parameter θ is first seen as a fixed quantity, with no specific value assigned. After the likelihood ratio is obtained, the probability $\mathbb{P}(H_p|e,\theta)$ is averaged over all possible values for θ as assigned by a prior distribution.

Since the unknown source evidence is generated according to different sampling models corresponding to each of the two hypotheses, the prior on $\theta|e$ also depends on the hypothesis. Abbreviating $\mathbb{P}(H_p)=\pi_1$ and using $\mathbb{P}(H_p)+\mathbb{P}(H_d)=1$, the law of total probability gives

$$\pi(\theta|e)=\pi_1\cdot\pi(\theta|e,H_p)+(1-\pi_1)\cdot\pi(\theta|e,H_d)$$

so that the posterior probability of guilt from the second approach splits into two integrals:

$$\mathbb{P}_{LR}(H_p|e)=\pi_1\int\frac{\pi_1\cdot LR(\theta;e_u)}{1-\pi_1+\pi_1\cdot LR(\theta;e_u)}\,d\Pi(\theta|e,H_p)+(1-\pi_1)\int\frac{\pi_1\cdot LR(\theta;e_u)}{1-\pi_1+\pi_1\cdot LR(\theta;e_u)}\,d\Pi(\theta|e,H_d).$$

These integrals can be seen as expectations with respect to different measures, i.e.,

$$\begin{aligned}\mathbb{P}_{LR}(H_p|e)&=\pi_1\cdot\mathbb{E}_{\Theta|e,H_p}\left[\frac{\pi_1\cdot LR(\Theta;e_u)}{1-\pi_1+\pi_1\cdot LR(\Theta;e_u)}\right]+(1-\pi_1)\cdot\mathbb{E}_{\Theta|e,H_d}\left[\frac{\pi_1\cdot LR(\Theta;e_u)}{1-\pi_1+\pi_1\cdot LR(\Theta;e_u)}\right]\\&=:\pi_1\cdot E_1+(1-\pi_1)\cdot E_2.\end{aligned}$$

Here, the subscript notation is used to indicate under which probability measure the expectation is calculated.

To compare the first approach with this convex combination of expectations, the Bayes Factor also has to be viewed as an expectation. Recall that the alternative expressions of the Bayes Factor from Chapter 3 were given by

$$BF(e)=\frac{1}{\int LR(\theta;e_u)^{-1}\,d\Pi(\theta|e,H_p)}=\frac{1}{\mathbb{E}_{\Theta|e,H_p}\left[LR(\Theta;e_u)^{-1}\right]}$$

and

$$BF(e)=\int LR(\theta;e_u)\,d\Pi(\theta|e,H_d)=\mathbb{E}_{\Theta|e,H_d}\left[LR(\Theta;e_u)\right].$$

Therefore, the posterior probability of guilt from approach (2) can be expressed as

$$\mathbb{P}_{BF}(H_p|e)=\frac{\pi_1\cdot\left(\mathbb{E}_{\Theta|e,H_p}\left[LR(\Theta;e_u)^{-1}\right]\right)^{-1}}{1-\pi_1+\pi_1\cdot\left(\mathbb{E}_{\Theta|e,H_p}\left[LR(\Theta;e_u)^{-1}\right]\right)^{-1}}=\frac{\pi_1}{(1-\pi_1)\cdot\mathbb{E}_{\Theta|e,H_p}\left[LR(\Theta;e_u)^{-1}\right]+\pi_1}\tag{11.1}$$

and

$$\mathbb{P}_{BF}(H_p|e)=\frac{\pi_1\cdot\mathbb{E}_{\Theta|e,H_d}\left[LR(\Theta;e_u)\right]}{1-\pi_1+\pi_1\cdot\mathbb{E}_{\Theta|e,H_d}\left[LR(\Theta;e_u)\right]}.\tag{11.2}$$

The goal is to relate equation (11.1) and (11.2) to E_1 and E_2, respectively. To achieve this, Jensen's inequality for conditional expectations will be used:

Theorem 15 (Jensen's inequality for conditional expectations)**.** *Let $\phi : \mathbb{R} \to \mathbb{R}$ be a convex function and let ξ be an integrable random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that $\phi(\xi)$ is also integrable. Then*

$$\phi\big(\mathbb{E}[\xi|\mathcal{G}]\big) \le \mathbb{E}\left[\phi(\xi)|\mathcal{G}\right] \quad a.s.$$

for any σ-field $\mathcal{G}$ on Ω contained in $\mathcal{F}$. [10]

Now define $\phi_1(x) := \frac{\pi_1}{(1-\pi_1)x+\pi_1}$ with $0 < \pi_1 < 1$ and $x > 0$ (since $BF^{-1} > 0$). Then ϕ_1 is a convex function. Hence, Jensen's conditional inequality and equation (11.1) give

$$\begin{aligned}
\mathbb{P}_{BF}(H_p|e) &= \frac{\pi_1}{(1-\pi_1)\cdot \mathbb{E}_{\Theta|e,H_p}\left[LR(\Theta;e_u)^{-1}\right]+\pi_1} = \phi_1\left(\mathbb{E}_{\Theta|e,H_p}\left[LR(\Theta;e_u)^{-1}\right]\right) \\
&\le \mathbb{E}_{\Theta|e,H_p}\left[\phi_1\left(LR(\Theta;e_u)^{-1}\right)\right] = \mathbb{E}_{\Theta|e,H_p}\left[\frac{\pi_1}{(1-\pi_1)\cdot LR(\Theta;e_u)^{-1}+\pi_1}\right] \\
&= \mathbb{E}_{\Theta|e,H_p}\left[\frac{\pi_1 \cdot LR(\Theta;e_u)}{(1-\pi_1)+\pi_1\cdot LR(\Theta;e_u)}\right] = E_1.
\end{aligned}$$

Similarly, define $\phi_2(x) := \frac{\pi_1 x}{(1-\pi_1)+\pi_1 x}$ with $0 < \pi_1 < 1$ and $x > 0$ (since $BF > 0$). Then ϕ_2 is a concave function and thus $-\phi_2$ is convex. Applying Jensen's conditional inequality to $-\phi_2$ gives

$$-\phi_2\big(\mathbb{E}[\xi|\mathcal{G}]\big) \le \mathbb{E}\left[-\phi_2(\xi)|\mathcal{G}\right] = -\mathbb{E}\left[\phi_2(\xi)|\mathcal{G}\right] \quad \text{a.s.} \qquad \Longleftrightarrow \qquad \phi_2\big(\mathbb{E}[\xi|\mathcal{G}]\big) \ge \mathbb{E}\left[\phi_2(\xi)|\mathcal{G}\right] \quad \text{a.s.}$$

Using this last inequality and equation (11.2), it follows that

$$\begin{aligned}
\mathbb{P}_{BF}(H_p|e) &= \frac{\pi_1 \cdot \mathbb{E}_{\Theta|e,H_d}\left[LR(\Theta;e_u)\right]}{1-\pi_1+\pi_1\cdot \mathbb{E}_{\Theta|e,H_d}\left[LR(\Theta;e_u)\right]} = \phi_2\left(\mathbb{E}_{\Theta|e,H_d}\left[LR(\Theta;e_u)\right]\right) \\
&\ge \mathbb{E}_{\Theta|e,H_d}\left[\phi_2\left(LR(\Theta;e_u)\right)\right] = \mathbb{E}_{\Theta|e,H_d}\left[\frac{\pi_1\cdot LR(\Theta;e_u)}{1-\pi_1+\pi_1\cdot LR(\Theta;e_u)}\right] = E_2.
\end{aligned}$$

This shows that $\mathbb{P}_{LR}(H_p|e)$ is a convex combination of an element that is always greater than or equal to $\mathbb{P}_{BF}(H_p|e)$ and an element that is always less than or equal to $\mathbb{P}_{BF}(H_p|e)$. Equality is only obtained when both ϕ_1 and ϕ_2 are linear functions. Asymptotically, this is reached when $BF \to \infty$, since

$$\lim_{x\to 0}\phi_1(x) = 1 \qquad \text{and} \qquad \lim_{x\to\infty}\phi_2(x) = 1.$$

Therefore, the two approaches to calculate the posterior probability of guilt will only give approximately the same result if the Bayes Factor is very large. In this case, both E_1 and E_2 will be approximately equal to $\mathbb{P}_{BF}(H_p|e)$, implying that $\mathbb{P}_{LR}(H_p|e)$ will be approximately equal to 1. Since

$$\mathbb{P}_{BF}(H_p|e) = \frac{\pi_1}{(1-\pi_1)BF(e)^{-1}+\pi_1},$$

it can easily be seen that $\mathbb{P}_{BF}(H_p|e)$ will also be approximately equal to 1 when the Bayes Factor is very large. However, the fact that the posterior probability of guilt will approach 1 for large Bayes Factor should already be clear from intuition. Note that the theoretical results in this section are valid independently of the model under consideration.

11.2. Common source one-level Bernoulli problem

Recall from Chapter 10 that for the one-dimensional common source Bernoulli problem, the following setup is used to represent the background material:

$$Y_{i1} \overset{\text{iid}}{\sim} Ber(\theta_a) \quad \text{for } i = 1,2,\ldots,n_a \quad \text{and} \quad \Theta_a \sim Beta(\alpha,\beta) \quad \text{where } \alpha > 0, \beta > 0.$$

This resulted in a Bayes Factor of

$$BF_{CS}(e) = \frac{\alpha+\beta+n_a+1}{\alpha+s_a+1},$$

where $s_a = \sum_{i=1}^{n_a} y_{i1}$. If $n_a \to \infty$ and $n_a >> s_a$, also $BF_{CS}(e) \to \infty$ and according to the previous section the two approaches to calculate the posterior probability of guilt should approximately give the same result.

Abbreviating $\mathbb{P}(H_p) = \pi_1$ and using $\mathbb{P}(H_p) + \mathbb{P}(H_d) = 1$, the posterior probability of guilt from approach (1) becomes

$$\mathbb{P}_{BF}(H_p|e_{u_1}, e_{u_2}, e_a) = \frac{\pi_1(\alpha+\beta+n_a+1)}{(1-\pi_1)(\alpha+s_a+1)+\pi_1(\alpha+\beta+n_a+1)}.$$

However, calculating the posterior probability of guilt using approach (2) is a lot harder. Since the likelihood ratio corresponding to this problem is $1/\theta_a$ (see Chapter 10), the integral

$$\mathbb{P}_{LR}(H_p|e_{u_1}, e_{u_2}, e_a) = \int \frac{\pi_1 \times \frac{1}{\theta_a}}{1-\pi_1+\pi_1 \times \frac{1}{\theta_a}} \pi(\theta_a|e_{u_1}, e_{u_2}, e_a)\ d\theta_a$$

has to be evaluated. The prior $\pi(\theta_a|e_{u_1}, e_{u_2}, e_a)$ depends on the hypothesis. Given H_p, the unknown source evidence $e_u = \{e_{u_1}, e_{u_2}\}$ consists of one $Ber(\theta_a)$ random variable, whereas given H_d, $e_u = \{e_{u_1}, e_{u_2}\}$ consists of two i.i.d. $Ber(\theta_a)$ random variables. Therefore,

$$\Theta_a|e_{u_1}, e_{u_2}, e_a, H_p \sim Beta(\alpha+s_a+1, \beta+n_a-s_a)$$

and

$$\Theta_a|e_{u_1}, e_{u_2}, e_a, H_d \sim Beta(\alpha+s_a+2, \beta+n_a-s_a).$$

The law of total probability gives

$$\pi(\theta_a|e_{u_1}, e_{u_2}, e_a) = \pi_1 \cdot \pi(\theta_a|e_{u_1}, e_{u_2}, e_a, H_p) + (1-\pi_1)\cdot \pi(\theta_a|e_{u_1}, e_{u_2}, e_a, H_d)$$

so that the posterior probability of guilt splits into:

$$\begin{aligned}\mathbb{P}_{LR}(H_p|e_{u_1}, e_{u_2}, e_a) &= \pi_1 \int \frac{\pi_1}{(1-\pi_1)\theta_a+\pi_1} \pi(\theta_a|e_{u_1}, e_{u_2}, e_a, H_p)\ d\theta_a \\ &\quad + (1-\pi_1) \int \frac{\pi_1}{(1-\pi_1)\theta_a+\pi_1} \pi(\theta_a|e_{u_1}, e_{u_2}, e_a, H_d)\ d\theta_a \\ &=: I_1 + I_2.\end{aligned}$$

Another way to see that $\mathbb{P}_{LR}(H_p|e_{u_1}, e_{u_2}, e_a)$ is approximately equal to 1 for large n_a follows from this expression. Using the property that the $Beta(\alpha, \beta)$ distribution becomes a degenerate distribution with all mass located at $x = 0$ for $\beta/\alpha \to \infty$ implies that for $n_a \to \infty$

$$\begin{aligned}\mathbb{P}_{LR}(H_p|e_{u_1}, e_{u_2}, e_a) &\longrightarrow \pi_1 \int \frac{\pi_1}{(1-\pi_1)\theta_a+\pi_1} \delta(\theta_a)\ d\theta_a + (1-\pi_1) \int \frac{\pi_1}{(1-\pi_1)\theta_a+\pi_1} \delta(\theta_a)\ d\theta_a \\ &= \pi_1 \cdot \frac{\pi_1}{(1-\pi_1)\cdot 0+\pi_1} + (1-\pi_1) \cdot \frac{\pi_1}{(1-\pi_1)\cdot 0+\pi_1} = 1,\end{aligned}$$

where $\delta(\cdot)$ denotes the Dirac delta function.

To determine the size of the difference between the two approaches, the expression for $\mathbb{P}_{LR}(H_p|e_{u_1}, e_{u_2}, e_a)$ has to be further investigated. This can be done by evaluating the two integrals separately:

$$\begin{aligned}I_1 &= \pi_1 \int \frac{\pi_1}{(1-\pi_1)\theta_a+\pi_1} \frac{\Gamma(\alpha+\beta+n_a+1)}{\Gamma(\alpha+s_a+1)\Gamma(\beta+n_a-s_a)} \theta_a^{\alpha+s_a}(1-\theta_a)^{\beta+n_a-s_a-1}\ d\theta_a \\ &= \pi_1^2 \frac{\Gamma(\alpha+\beta+n_a+1)}{\Gamma(\alpha+s_a+1)\Gamma(\beta+n_a-s_a)} \int \frac{1}{1-(1-\theta_a)(1-\pi_1)} \theta_a^{\alpha+s_a}(1-\theta_a)^{\beta+n_a-s_a-1}\ d\theta_a \\ &= C_1 \times \int \sum_{k=0}^{\infty} (1-\theta_a)^k (1-\pi_1)^k \theta_a^{\alpha+s_a}(1-\theta_a)^{\beta+n_a-s_a-1}\ d\theta_a, \qquad |(1-\theta_a)(1-\pi_1)| < 1 \\ &= C_1 \times \sum_{k=0}^{\infty} (1-\pi_1)^k \int \theta_a^{\alpha+s_a}(1-\theta_a)^{\beta+n_a-s_a-1+k}\ d\theta_a \\ &= C_1 \times \sum_{k=0}^{\infty} (1-\pi_1)^k \frac{\Gamma(\alpha+s_a+1)\Gamma(\beta+n_a+k-s_a)}{\Gamma(\alpha+\beta+n_a+k+1)} \\ &= \pi_1^2 \frac{\Gamma(\alpha+\beta+n_a+1)}{\Gamma(\beta+n_a-s_a)} \sum_{k=0}^{\infty} (1-\pi_1)^k \frac{\Gamma(\beta+n_a+k-s_a)}{\Gamma(\alpha+\beta+n_a+k+1)}\end{aligned}$$

$$= \pi_1^2 \cdot {}_2F_1(1, \beta + n_a - s_a; \alpha + \beta + n_a + 1; 1 - \pi_1),$$

where integration and summation can be swapped because of Tonelli's theorem for non-negative functions. Here, ${}_2F_1(a, b; c; z)$ denotes the *Gauss hypergeometric series* as described in [1]:

$${}_2F_1(a, b; c; z) = \frac{\Gamma(c)}{\Gamma(a)\Gamma(b)} \sum_{k=0}^{\infty} \frac{\Gamma(a+k)\Gamma(b+k)}{\Gamma(c+k)} \frac{z^k}{k!}.$$

The series converges for $|z| < 1$ and it can be analytically continued onto the entire complex plane cut along $[1, \infty]$ [21].

Similarly,

$$\begin{aligned}
I_2 &= (1-\pi_1) \int \frac{\pi_1}{(1-\pi_1)\theta_a + \pi_1} \frac{\Gamma(\alpha+\beta+n_a+2)}{\Gamma(\alpha+s_a+2)\Gamma(\beta+n_a-s_a)} \theta_a^{\alpha+s_a+1}(1-\theta_a)^{\beta+n_a-s_a-1}\, d\theta_a \\
&= \pi_1(1-\pi_1) \frac{\Gamma(\alpha+\beta+n_a+2)}{\Gamma(\alpha+s_a+2)\Gamma(\beta+n_a-s_a)} \int \frac{1}{1-(1-\theta_a)(1-\pi_1)} \theta_a^{\alpha+s_a+1}(1-\theta_a)^{\beta+n_a-s_a-1}\, d\theta_a \\
&= C_2 \times \int \sum_{k=0}^{\infty} (1-\theta_a)^k (1-\pi_1)^k \theta_a^{\alpha+s_a+1}(1-\theta_a)^{\beta+n_a-s_a-1}\, d\theta_a, \qquad |(1-\theta_a)(1-\pi_1)| < 1 \\
&= C_2 \times \sum_{k=0}^{\infty} (1-\pi_1)^k \int \theta_a^{\alpha+s_a+1}(1-\theta_a)^{\beta+n_a-s_a-1+k}\, d\theta_a \\
&= C_2 \times \sum_{k=0}^{\infty} (1-\pi_1)^k \frac{\Gamma(\alpha+s_a+2)\Gamma(\beta+n_a+k-s_a)}{\Gamma(\alpha+\beta+n_a+k+2)} \\
&= \pi_1(1-\pi_1) \frac{\Gamma(\alpha+\beta+n_a+2)}{\Gamma(\beta+n_a-s_a)} \sum_{k=0}^{\infty} (1-\pi_1)^k \frac{\Gamma(\beta+n_a+k-s_a)}{\Gamma(\alpha+\beta+n_a+k+2)} \\
&= \pi_1(1-\pi_1) \cdot {}_2F_1(1, \beta+n_a-s_a; \alpha+\beta+n_a+2; 1-\pi_1).
\end{aligned}$$

This gives for the posterior probability of guilt:

$$\begin{aligned}
\mathbb{P}_{LR}(H_p | e_{u_1}, e_{u_2}, e_a) = \pi_1^2 \cdot {}_2F_1(1, \beta+n_a-s_a; \alpha+\beta+n_a+1; 1-\pi_1) \\
+ \pi_1(1-\pi_1) \cdot {}_2F_1(1, \beta+n_a-s_a; \alpha+\beta+n_a+2; 1-\pi_1)
\end{aligned} \tag{11.3}$$

However, this expression only has an analytical solution for certain values of α, β, n_a and s_a.

From identity (15.3.4) from [1] it can be derived that

$${}_2F_1(a, b; c; z) = (1-z)^{-a} {}_2F_1\left(a, c-b; c; \frac{z}{z-1}\right) \quad \Longleftrightarrow \quad (1-z)^a {}_2F_1(a, b; c; z) = {}_2F_1\left(a, c-b; c; \frac{z}{z-1}\right).$$

This can be used to write equation (11.3) as

$$\begin{aligned}
\mathbb{P}_{LR}(H_p | e_{u_1}, e_{u_2}, e_a) = \pi_1 \cdot {}_2F_1\left(1, \alpha+s_a+1; \alpha+\beta+n_a+1; -\frac{1-\pi_1}{\pi_1}\right) \\
+ (1-\pi_1) \cdot {}_2F_1\left(1, \alpha+s_a+2; \alpha+\beta+n_a+2; -\frac{1-\pi_1}{\pi_1}\right),
\end{aligned} \tag{11.4}$$

which shows that the posterior probability of guilt from approach (2) results in a convex combination of Gauss hypergeometric functions.

11.3. Gauss hypergeometric functions and continued fractions

To be able to say something about the difference between the posterior probability of guilt from approach (1) and (2), the properties of the Gauss hypergeometric functions need to be further investigated. In [21] it is explained that Gauss hypergeometric functions of the form ${}_2F_1(1, b; c; -z)$ can be represented by a *continued fraction*. The Gauss hypergeometric functions obtained in equation (11.4) are exactly of this form. Therefore, some theory of continued fractions and its relation to the Gauss hypergeometric functions will be explained in this section.

A continued fraction is an infinite structure which works as an alternative to infinite series. The standard form of a continued fraction is

$$a_0 + \cfrac{b_1}{a_1 + \cfrac{b_2}{a_2 + \cfrac{b_3}{a_3 + \cfrac{b_4}{a_4 + \ddots}}}}$$

which can be converted to *simple form*

$$c_0 + \cfrac{1}{c_1 + \cfrac{1}{c_2 + \cfrac{1}{c_3 + \cfrac{1}{c_4 + \ddots}}}}$$

by setting

$$c_0 = a_0, \quad c_1 = \frac{a_1}{b_1}, \quad c_2 = \frac{a_2 b_1}{b_2}, \quad c_3 = \frac{a_3 b_2}{b_1 b_3}, \quad c_4 = \frac{a_4 b_1 b_3}{b_2 b_4}.$$

The convergence of a continued fraction can be discussed in terms of the approximants or *convergents*. Each convergent f_n is obtained by truncating the continued fraction after n fraction terms. Then the continued fraction converges to the value f if and only if $\lim_{n\to\infty} f_n = f$ [25]. The first five convergents for a continued fraction in simple form are given by:

$$f_0 := \frac{c_0}{1}, \quad f_1 := \frac{c_1 c_0 + 1}{c_1}, \quad f_2 := \frac{c_2(c_1 c_0 + 1) + c_0}{c_2 c_1 + 1}, \quad f_3 := \frac{c_3(c_2(c_1 c_0 + 1) + c_0) + c_1 c_0 + 1}{c_3(c_2 c_1 + 1) + c_1},$$

$$f_4 := \frac{c_4(c_3(c_2(c_1 c_0 + 1) + c_0) + c_1 c_0 + 1) + c_2(c_1 c_0 + 1) + c_0}{c_4(c_3(c_2 c_1 + 1) + c_1) + c_2 c_1 + 1}.$$

The continued fraction corresponding to the Gauss hypergeometric functions of the form ${}_2F_1(1,b;c;-z)$ is given by

$${}_2F_1(1,b;c;-z) = \cfrac{c-1}{c-1+\cfrac{bz}{c+\cfrac{(c-b)z}{c+1+\cfrac{c(b+1)z}{c+2+\ddots}}}}$$

This continued fraction converges and has positive elements when $z > 0$ and $c > b > 1$. Moreover, its convergents have a very interesting property: the even convergents form an increasing sequence and approximate the value of ${}_2F_1(1,b;c;-z)$ from below, while the odd convergents form a decreasing sequence and approximate ${}_2F_1(1,b;c;-z)$ from above [21]. Therefore, convergent f_3 gives an upper bound for ${}_2F_1(1,b;c;-z)$ and convergent f_4 gives a lower bound for ${}_2F_1(1,b;c;-z)$.

Applying this theory to the Gauss hypergeometric functions from equation (11.4) gives for

$${}_2F_1\left(1, \alpha + s_a + 1; \alpha + \beta + n_a + 1; -\frac{1-\pi_1}{\pi_1}\right)$$

the lower and upper bound

$$L_1 = \frac{c_4 c_3 c_2 + c_4 + c_2}{c_4 c_3 c_2 + c_4 c_3 + c_4 + c_2 + 1} \quad \text{and} \quad U_1 = \frac{c_3 c_2 + 1}{c_3 c_2 + c_3 + 1},$$

respectively, with

$$c_2 = \frac{\pi_1(\alpha+\beta+n_a+1)}{(1-\pi_1)(\alpha+s_a+1)}, \quad c_3 = \frac{(\alpha+\beta+n_a+2)(\alpha+s_a+1)}{\beta+n_a-s_a}, \quad c_4 = \frac{\pi_1(\alpha+\beta+n_a+3)(\beta+n_a-s_a)}{(1-\pi_1)(\alpha+s_a+1)(\alpha+s_a+2)(\alpha+\beta+n_a+1)}.$$

Similarly, for

$${}_2F_1\left(1, \alpha+s_a+2; \alpha+\beta+n_a+2; -\frac{1-\pi_1}{\pi_1}\right)$$

the lower and upper bound are given by

$$L_2 = \frac{c_4c_3c_2+c_4+c_2}{c_4c_3c_2+c_4c_3+c_4+c_2+1} \quad \text{and} \quad U_2 = \frac{c_3c_2+1}{c_3c_2+c_3+1},$$

respectively, with

$$c_2 = \frac{\pi_1(\alpha+\beta+n_a+2)}{(1-\pi_1)(\alpha+s_a+2)}, \quad c_3 = \frac{(\alpha+\beta+n_a+3)(\alpha+s_a+2)}{\beta+n_a-s_a}, \quad c_4 = \frac{\pi_1(\alpha+\beta+n_a+4)(\beta+n_a-s_a)}{(1-\pi_1)(\alpha+s_a+2)(\alpha+s_a+3)(\alpha+\beta+n_a+2)}.$$

Therefore, the posterior probability of guilt from approach (2) is bounded by

$$\pi_1L_1+(1-\pi_1)L_2 < \mathbb{P}_{LR}(H_p|e_{u_1},e_{u_2},e_a) < \pi_1U_1+(1-\pi_1)U_2.$$

11.4. Difference between Bayes Factor and likelihood ratio approach

To answer the question of interest, the difference in the posterior probability of guilt obtained from the approach using the Bayes Factor and the likelihood ratio should be considered. Using the bounds from the previous section leads to

$$L_{\text{bound}} < \mathbb{P}_{BF}(H_p|e_{u_1},e_{u_2},e_a) - \mathbb{P}_{LR}(H_p|e_{u_1},e_{u_2},e_a) < U_{\text{bound}},$$

where

$$L_{\text{bound}} = \frac{\pi_1(\alpha+\beta+n_a+1)}{(1-\pi_1)(\alpha+s_a+1)+\pi_1(\alpha+\beta+n_a+1)} - \pi_1U_1-(1-\pi_1)U_2$$

and

$$U_{\text{bound}} = \frac{\pi_1(\alpha+\beta+n_a+1)}{(1-\pi_1)(\alpha+s_a+1)+\pi_1(\alpha+\beta+n_a+1)} - \pi_1L_1-(1-\pi_1)L_2.$$

These expressions for L_{bound} and U_{bound} are explicit, but not very tractable. Therefore, Maple™ is used to evaluate the difference between the lower and upper bound. Since in general $n_a >> \alpha, \beta, s_a$, the highest power of n_a in both the numerator and denominator will mainly determine the size of this difference. Neglecting lower order terms, it is found that

$$\begin{aligned} U_{\text{bound}} - L_{\text{bound}} &\approx \frac{1}{n_a^3}\Bigg[(2\alpha+2s_a+4)\pi_1-(\alpha+s_a+9)(\alpha+s_a+2)+(3\alpha+3s_a+15)(\alpha+s_a+2)\frac{1}{\pi_1} \\ &\qquad -(3\alpha+3s_a+11)(\alpha+s_a+2)\frac{1}{\pi_1^2}+(\alpha+s_a+3)(\alpha+s_a+2)\frac{1}{\pi_1^3}\Bigg] \\ &=: \frac{C(\alpha,s_a,\pi_1)}{n_a^3} = \mathcal{O}\left(\frac{1}{n_a^3}\right) \end{aligned}$$

where $C(\alpha, s_a, \pi_1) > 0$ for $0 < \pi_1 < 1$ and all $\alpha > 0$ (see Figure 11.1).

Hence, the difference between the posterior probability of guilt obtained from the two approaches is of order $1/n_a^3$, where n_a denotes the number of sources in the background material e_a. This means that, if a reasonably large set of background material is used compared to the number of matches, the difference becomes negligible and it does not really matter which approach is used. If one is interested in the size of the difference, $C(\alpha, s_a, \pi_1)/n_a^3$ will provide an accurate estimate (see Figure 11.2).

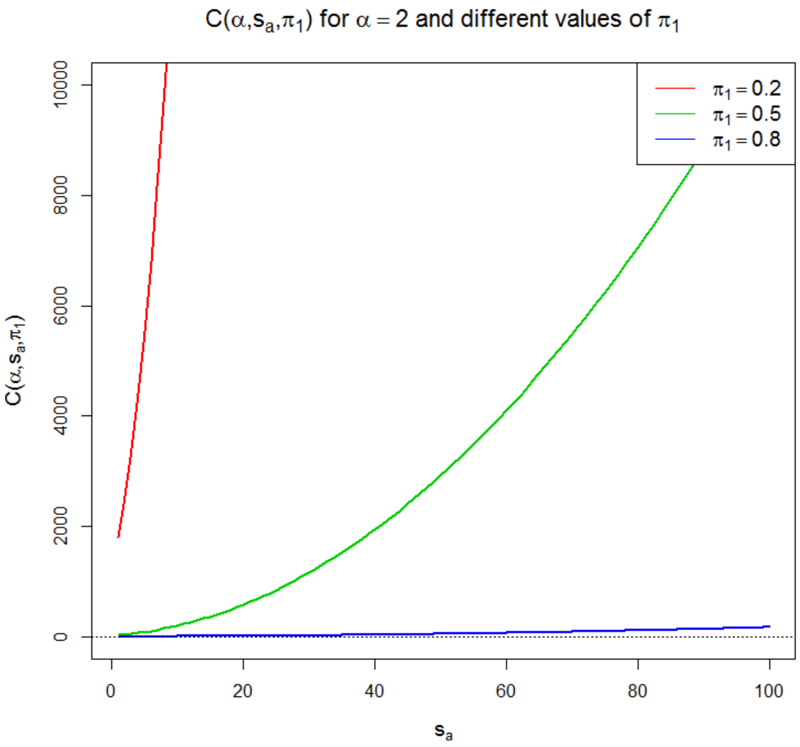

Figure 11.1: $C(\alpha, s_a, \pi_1)$ as function of s_a, for $\alpha = 2$ and different values of π_1.

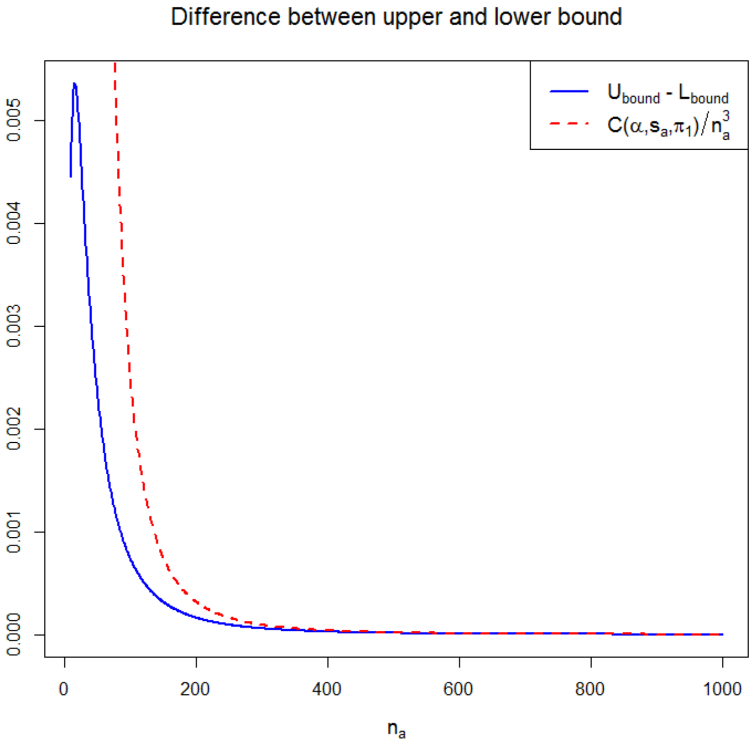

Figure 11.2: Exact difference between upper and lower bound together with the estimated difference, for $\alpha = 2$, $\beta = 5$, $s_a = 10$ and $\pi_1 = 0.3$.

In Figure 11.3 the posterior probability of guilt obtained from both approaches is visualized as function of n_a. For $\mathbb{P}_{LR}(H_p|e)$ the graph is split into two parts: the dotted line up to $n_a = 180$ shows the result of the hypergeometric functions using Euler's integral representation [21]

$$ {}_2F_1(1,b;c;-z) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)} \int_0^1 \frac{t^{b-1}(1-t)^{c-b-1}}{1+zt}\, dt, $$

evaluated using numerical contour integrals, whereas the solid line is obtained by the continued fraction expansion of the hypergeometric functions, as discussed in the previous section. This distinction is made because the continued fraction expansion sometimes fails to converge to the correct value for small n_a, and Euler's integral representation gives numerical errors for large n_a.

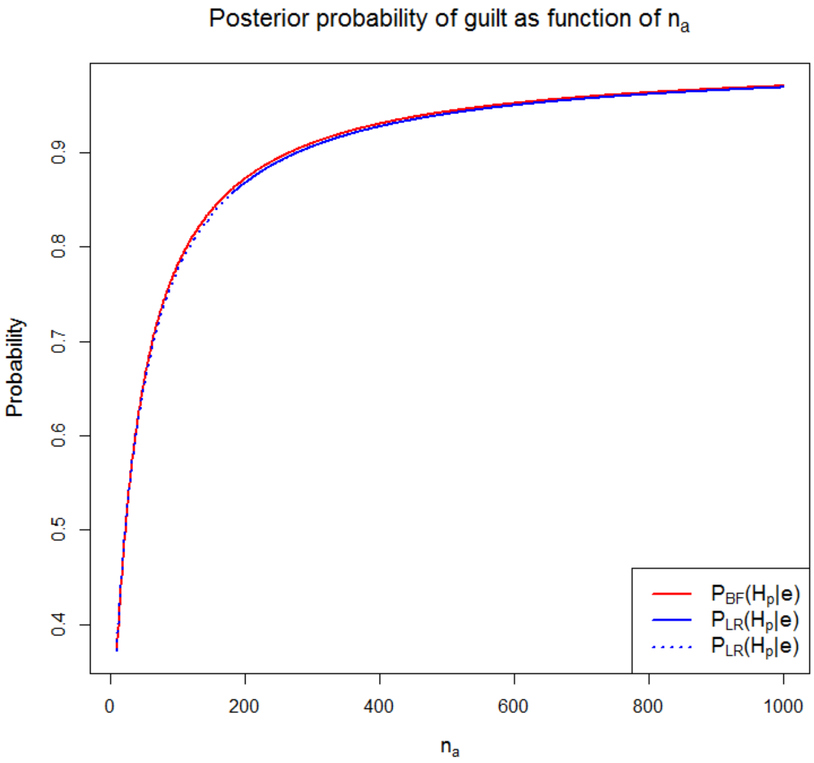

Figure 11.3: Posterior probability of guilt from both the Bayes Factor and likelihood ratio approach, for $\alpha = 2$, $\beta = 5$, $s_a = 10$ and $\pi_1 = 0.3$.

Although both graphs for the posterior probability of guilt look very similar, Figure 11.4 illustrates how they still differ. The upper and lower bound on the difference are also added to the plot and seem to be very accurate. For small n_a, the lower bound could be more strict. Since the odd convergents from the hypergeometric functions form a decreasing sequence, a stricter lower bound on the difference could be obtained by considering convergent f_5 for both hypergeometric functions in formula (11.4) and replace U_1 and U_2 in L_{bound} by this fifth convergent. The spikes in the difference around $n_a = 170$ are most likely caused by numerical errors.

11.5. Conclusions

To conclude this chapter, a short summary of the most important findings regarding the posterior probability of guilt obtained from the Bayes Factor and likelihood ratio approach will be given. From a theoretical point of view, it is clear that both probabilities will never be exactly the same. Only for problems where the Bayes Factor is very large, the probabilities will be approximately the same and will both approach 1.

Whereas the calculation of the posterior probability of guilt using the Bayes Factor is straightforward, the approach using the likelihood ratio might lead to some difficulties. For the common source one-level Bernoulli problem, $\mathbb{P}_{LR}(H_p|e)$ can be represented by a convex combination of hypergeometric functions. Extreme care has to be taken when evaluating these functions, since numerical errors can occur easily and continued fraction expansions might fail to converge to the correct value.

The difference between the posterior probability of guilt from both methods for the common source one-level Bernoulli problem was shown to be of order $1/n_a^3$. However, this means that for small values of n_a the difference is still noticeable. The upper and lower bound on the difference are accurate, and the lower bound could be made even sharper by considering the fifth convergent of the hypergeometric functions.

Taking all these arguments into account, it should be clear that using the likelihood ratio as described here to calculate the posterior probability of guilt can easily lead to errors. Moreover, combining a frequentist likelihood ratio with a Bayesian prior in this manner is questionable from a theoretical perspective. Therefore, if a Bayesian prior on the parameter is required, using the Bayes Factor to calculate $\mathbb{P}(H_p|e)$ would be preferred in most situations. Only when the number of sources in the background material is reasonably large compared to the number of matches, the likelihood ratio approach as described in this chapter should be used. Note that even in this case one still has to be careful when evaluating the hypergeometric functions for the common source one-level Bernoulli model. For the sake of completeness, the bounds on the difference between both posterior probabilities of guilt could be reported.

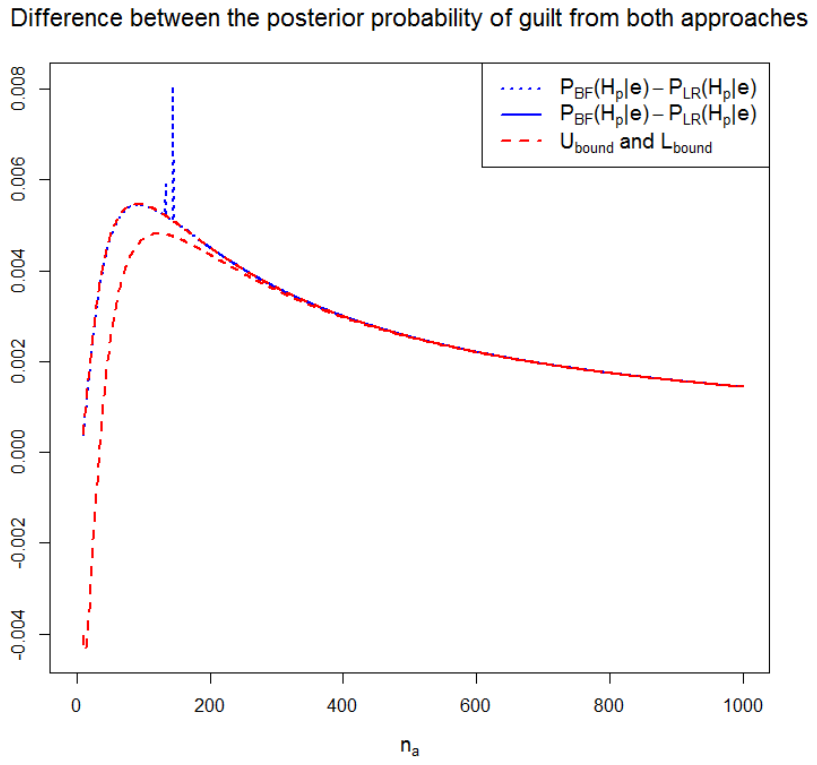

Figure 11.4: Difference between the posterior probability of guilt from the Bayes Factor and likelihood ratio approach, together with bounds on the difference, for $\alpha = 2$, $\beta = 5$, $s_a = 10$ and $\pi_1 = 0.3$.

12 Permutation tests

Until now, the 'standard' forensic approach to quantify the value of evidence has been considered, which was explained in Chapter 3 and proceeds via a two step procedure. In the first step it is determined whether it is probable that two traces share the same source, and the value of this belief is expressed in the numerator of the likelihood ratio or Bayes Factor. The second step is the assessment of the probability that the trace of interest comes from a randomly selected source, which is used in the denominator of the likelihood ratio or Bayes Factor. These probabilities depend heavily on distributional assumptions and are usually not straightforward to calculate. Therefore, it is argued in [11] that this two step approach is neither desirable nor necessary, since the first step could be replaced by a much simpler nonparametric test based on the concept of *exchangeability*.

Definition 16. *A finite set $X_1,\ldots,X_n$ of random quantities is said to be* exchangeable *if every permutation of $(X_1,\ldots,X_n)$ has the same joint distribution as $(X_1,\ldots,X_n)$. An infinite collection of random variables is exchangeable if every finite subcollection is exchangeable. [42]*

Both the common and specific source problem, as discussed in Chapter 2, boil down to the question which of the evidence sets are exchangeable. If the prosecution hypothesis is true, the (first) unknown source evidence is assumed to be generated similarly as the second set of unknown source evidence or the specific source evidence, respectively, for the common and specific source problem. This means that given H_p the two sets of evidence are exchangeable, whereas given H_d they are not. The concept of exchangeability can easily be evaluated through the use of *permutation testing*.

The basic idea of permutation testing is that changing the labelling of the measurements to assign them to one of the two evidence sets does not significantly change the value of a predetermined *test statistic* if the prosecution hypothesis is true. In [17] the "five steps" to construct a permutation test are given:

1. Identify the hypothesis and alternative(s) of interest. For example, in the forensic setting with two traces $\mathbf{X} = (X_1, X_2, \ldots, X_{n_x})$ and $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_{n_y})$ where it is assumed that $X_i \overset{\text{iid}}{\sim} F(\cdot)$ for $i = 1,2,\ldots,n_x$ and $Y_i \overset{\text{iid}}{\sim} G(\cdot)$ for $i = 1,2,\ldots,n_y$, the competing hypotheses might be $H_p : F = G$ and $H_d : F \neq G$.
2. Choose a relevant test statistic to confirm or disprove the prosecution hypothesis H_p.
3. Compute the test statistic based on the original labelling of the observations.
4. Recompute the test statistic for all possible rearrangements of the labels. Repeat this computation until you obtain the distribution of the test statistic for all rearrangements.
5. Accept or reject the prosecution hypothesis using this permutation distribution as a guide.

In practice, the number of possible permutations might be too large to evaluate in step 4. Therefore, a *random* or *Monte Carlo permutation test* can be considered, where only a random sample of all possible permutations is taken. The decision of accepting or rejecting the hypothesis in step 5 is usually based on a *p-value* . This is the probability of observing a test statistic from a permutation that is at least as unusual as the one observed from the two traces. If the p-value is smaller than a pre-defined significance level α (usually 0.05 or 0.01), the prosecution hypothesis is rejected.

It is important to mention that permutation tests exist, but were not designed for forensic purposes. Therefore, they should not be used for the interpretation of evidence in court, but only for laboratory based pre-screening. If the permutation test shows strong evidence against the prosecution hypothesis, then it might not be worthwhile completing further statistical interpretation in light of this evidence [11]. Since the obtained p-value does not fit in the Bayesian framework and cannot be combined with the prior odds, extreme care has to be taken when drawing conclusions. However, permutation tests were already discussed in [49] for the evaluation of DNA evidence.

Note that for Monte Carlo permutation testing only an approximate p-value is obtained. It is possible to construct an approximate 95% confidence interval for the true p-value based on the approximate p-value $\hat{p}$. Under the prosecution model, each sampled permutation has a probability equal to p of observing a test statistic at least as extreme as the test statistic based on the original labelling. Let B denote the total number of samples with a test statistic at least as extreme as the one based on the original labelling. For a Monte Carlo permutation test of size M, B follows a binomial distribution with parameters M and p. This means that

$$\hat{p} = \frac{B}{M} \sim N\left(p, \frac{p(1-p)}{M}\right)$$

by the Central Limit Theorem. As a result, an approximate 95% confidence interval is given by

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{M}}.$$

This ensures that as long as M is large or $\hat{p}$ is small, a good approximation of the true p-value is obtained.

There are many possibilities when choosing a test statistic in permutation testing. In this chapter, four test statistics will be discussed and applied to both simulated and real data. The real glass and MDMA data from Section 6.4 is used, and the simulated data is generated as described in Section 6.3. The knives data is not used since only one measurement is available from both traces, which makes it unsuitable for permutation testing.

Throughout this chapter, the measurements of the k features of the traces will be denoted by $\mathbf{x}$ and $\mathbf{y}$, where n_x and n_y denote the number of measurements for each trace:

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_{n_x} \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n_x} \\ x_{21} & x_{22} & \cdots & x_{2n_x} \\ \vdots & \vdots & \ddots & \vdots \\ x_{k1} & x_{k2} & \cdots & x_{kn_x} \end{bmatrix}$$

and

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 & \mathbf{y}_2 & \cdots & \mathbf{y}_{n_y} \end{bmatrix} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n_y} \\ y_{21} & y_{22} & \cdots & y_{2n_y} \\ \vdots & \vdots & \ddots & \vdots \\ y_{k1} & y_{k2} & \cdots & y_{kn_y} \end{bmatrix}.$$

This notation is slightly different than in the previous chapters, but it is used to avoid multiple indices and to provide a general framework for both the common and specific source problem. Note that for the common source problem $\mathbf{x} := \mathbf{y}_{u_1}$ and $\mathbf{y} := \mathbf{y}_{u_2}$, whereas for the specific source problem $\mathbf{x} := \mathbf{y}_s$ and $\mathbf{y} := \mathbf{y}_u$.

12.1. Feature mean difference

Let $\bar{\mathbf{x}} = (\bar{x}_1, \ldots, \bar{x}_k)^T$ and $\bar{\mathbf{y}} = (\bar{y}_1, \ldots, \bar{y}_k)^T$ denote the k-dimensional vectors of feature means, where

$$\bar{x}_i = \frac{1}{n_x}\sum_{j=1}^{n_x} x_{ij} \quad \text{and} \quad \bar{y}_i = \frac{1}{n_y}\sum_{j=1}^{n_y} y_{ij} \quad \text{for } i = 1,2,\ldots,k.$$

The first test statistic considers the norm of the difference in feature means $\bar{\mathbf{x}}-\bar{\mathbf{y}}$. Three different vector norms will be evaluated:

- Using the ℓ_1-norm, the test statistic becomes

$$T_1 = ||\bar{\mathbf{x}} - \bar{\mathbf{y}}||_1 = \sum_{i=1}^{k} |\bar{x}_i - \bar{y}_i|.$$

- Using the ℓ_2-norm, the test statistic becomes

$$T_2 = ||\bar{\mathbf{x}} - \bar{\mathbf{y}}||_2 = \sqrt{\sum_{i=1}^{k} (\bar{x}_i - \bar{y}_i)^2}.$$

- Using the ℓ_∞-norm, the test statistic becomes

$$T_\infty = ||\bar{\mathbf{x}} - \bar{\mathbf{y}}||_\infty = \max\left(|\bar{x}_1 - \bar{y}_1|, \ldots, |\bar{x}_k - \bar{y}_k|\right).$$

If the two traces originate from the same source, the difference in feature means will be small and therefore test statistics close to zero will be expected. The approximate p-value is calculated by

$$\hat{p} = \frac{1}{M} \sum_{i=1}^{M} I(T(i) > T(obs)),$$

where $I(\cdot)$ denotes the indicator function, $T(i)$ the test statistic corresponding to the ith sampled permutation and $T(obs)$ the test statistic based on to the original labelling.

12.2. Hotelling's T^2

Let $\mathbf{x}_j$ and $\mathbf{y}_j$ denote the k-dimensional measurement vectors, and again let $\bar{\mathbf{x}} = (\bar{x}_1, \ldots, \bar{x}_k)^T$ and $\bar{\mathbf{y}} = (\bar{y}_1, \ldots, \bar{y}_k)^T$ be the vectors of feature means. In [11] it is suggested to use *Hotelling's* T^2 as test statistic, which is defined as

$$T^2 = \frac{n_x n_y}{n_x + n_y} \left(\bar{\mathbf{x}} - \bar{\mathbf{y}}\right)^T S_k^{-1} \left(\bar{\mathbf{x}} - \bar{\mathbf{y}}\right),$$

where S_k denotes the estimator of the pooled covariance matrix

$$S_k = \frac{\sum_{j=1}^{n_x} (\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})^T + \sum_{j=1}^{n_y} (\mathbf{y}_j - \bar{\mathbf{y}})(\mathbf{y}_j - \bar{\mathbf{y}})^T}{n_x + n_y - 2}.$$

A restriction to the Hotelling's T^2 test statistic is that the total number of measurements, $n_x + n_y$, should be larger than the number of features plus one to ensure that the inverse of S_k exists. Again, the difference in feature means will be small when the two traces originate from the same source and therefore a small test statistic will be expected. Hence, the approximate p-value is calculated by

$$\hat{p} = \frac{1}{M} \sum_{i=1}^{M} I\left(T^2(i) > T^2(obs)\right).$$

12.3. Average intra- and inter-measurement similarity

Let $\mathbf{x}_j$ and $\mathbf{y}_j$ denote the k-dimensional measurement vectors and let $\bar{x}_j = \frac{1}{k}\sum_{i=1}^{k} x_{ij}$ and $\bar{y}_j = \frac{1}{k}\sum_{i=1}^{k} y_{ij}$ denote the corresponding measurement means. Since taking measurement means only makes sense if the features have the same unit, this test statistic can only be used in certain settings, such as the glass dataset from Section 6.4. The test statistic is based on [28] and compares the average intra-measurement similarity with the average inter-measurement similarity. Therefore, let

$$r(\mathbf{x}_u, \mathbf{y}_v) = \frac{\sum_{i=1}^{k} (x_{iu} - \bar{x}_u)(y_{iv} - \bar{y}_v)}{\sqrt{\sum_{i=1}^{k} (x_{iu} - \bar{x}_u)^2} \sqrt{\sum_{i=1}^{k} (y_{iv} - \bar{y}_v)^2}}$$

denote the *Pearson correlation coefficient* between measurements $\mathbf{x}_u$ and $\mathbf{y}_v$. The Fisher transformation of the Pearson correlation coefficient between two measurements is used as similarity measure:

$$f(\mathbf{x}_u, \mathbf{y}_v) = 0.5 \ln \left[\frac{1 + r(\mathbf{x}_u, \mathbf{y}_v)}{1 - r(\mathbf{x}_u, \mathbf{y}_v)}\right].$$

The motivation for this transformation is that for small values of r, f is nearly equal to r, but as r increases to unity, f approaches infinity [15]. This means that the difference between almost perfect correlation and lower correlations, i.e., the difference between intra-measurement correlations and inter-measurement correlations from different sources, is better noticeable in $f(\mathbf{x}_u, \mathbf{y}_v)$ than in $r(\mathbf{x}_u, \mathbf{y}_v)$. The sign of the correlation coefficient remains unchanged after the transformation.

Let $I_x = \{1, 2, \dots, n_x\}$ denote the set of all measurement indices corresponding to $\mathbf{x}$ and let $I_y = \{1, 2, \dots, n_y\}$ be the set of all measurement indices corresponding to $\mathbf{y}$. Then the test statistic is given by

$$W_0 = \frac{\sum_{(u,v)\in I_x, u\neq v} f(\mathbf{x}_u, \mathbf{x}_v) + \sum_{(u,v)\in I_y, u\neq v} f(\mathbf{y}_u, \mathbf{y}_v)}{n_x(n_x - 1) + n_y(n_y - 1)} - \frac{\sum_{u\in I_x, v\in I_y} f(\mathbf{x}_u, \mathbf{y}_v)}{n_x n_y}. \tag{12.1}$$

The first term in (12.1) represents the intra-measurement similarity and the second term represents the inter-measurement similarity. Since intra-measurement correlations will be positive and close to one, the first term will be positive. If the two traces originate from the same source, the intra-measurement and inter-measurement similarities behave the same and therefore the test statistic is expected to be close to zero. Otherwise, the intra-measurement similarity will be greater than the inter-measurement similarity, which leads to larger values of W_0. Note that this also holds if the second term in (12.1) is negative. The approximate p-value is calculated by

$$\hat{p} = \frac{1}{M}\sum_{i=1}^{M} I(W_0(i) > W_0(obs)).$$

12.4. Ranks of interpoint distances

Let $\mathbf{x}_j$ and $\mathbf{y}_j$ denote the k-dimensional measurement vectors and define the pooled measurement sample $\mathbf{z} = \begin{bmatrix} \mathbf{z}_1 & \mathbf{z}_2 & \cdots & \mathbf{z}_{n_x+n_y} \end{bmatrix}$, where $\mathbf{z}_j = \mathbf{x}_j$ for $j = 1, 2, \dots, n_x$ and $\mathbf{z}_{n_x+j} = \mathbf{y}_j$ for $j = 1, 2, \dots, n_y$. To overcome the limitations of the Hotelling's T^2 test statistic, [26] proposed a test statistic based on interpoint distances. Therefore, choose one of the measurements from $\mathbf{x}$ at random and denote this measurement by $\mathbf{x}_r$. The following steps have to be followed to compute the test statistic:

1. Compute the distance between $\mathbf{x}_r$ and the other measurements of the pooled sample $\mathbf{z}$, obtaining the vector $\mathbf{L}_r$ of the $n_x + n_y - 1$ interpoint distances $l_{rj}, j \neq r$, where

$$l_{rj} = ||\mathbf{x}_r - \mathbf{z}_j||_2 = \sqrt{\sum_{i=1}^{k} (x_{ir} - z_{ij})^2} \quad \text{for } j = 1, 2, \dots, n_x + n_y, j \neq r.$$

2. Compute the ranks r_{rj} of l_{rj}.
3. Compute the statistic $T_r = \sum_{j=n_x+1}^{n_x+n_y} r_{rj}$. Large values of T_r are evidence against the prosecution hypothesis.
4. The test statistic P_0 is obtained by computing the p-value of the T_r statistic.

Note that this approach is similar to the one-sided Wilcoxon rank sum test applied to the two samples of interpoint distances of observations with fixed $\mathbf{x}_r$, where the first sample is l_{rj} for $j = 1, \dots, n_x, j \neq r$, and the second sample is l_{rj} for $j = n_x + 1, \dots, n_y$. It is a one-sided test since under the defence model it is expected that the interpoint distances between $\mathbf{x}_r$ and the measurements from $\mathbf{y}$ are greater than the interpoint distances between $\mathbf{x}_r$ and the measurements from $\mathbf{x}$.

A permutation test can be performed by randomly permuting the pooled sample $\mathbf{z}$ and by computing the test statistic using the rth measurement from the permuted pooled sample $\mathbf{z}^*$ following the steps described above. Given the prosecution hypothesis that the distributions to be compared are the same, i.e., the traces originate from the same source, the measurements of $\mathbf{z}$ are exchangeable which justifies the use of a permutation test. The approximate p-value is then found from

$$\hat{p} = \frac{1}{M}\sum_{i=1}^{M} I(P_0(i) < P_0(obs)).$$

12.5. Results

To see how the different test statistics perform, several permutation tests have been applied to both real and simulated data. The real data consists of the glass and MDMA datasets as discussed in Section 6.4. Both the error of rejecting the prosecution hypothesis when it is true (*type I error*) and the error of accepting the prosecution hypothesis when the defence hypothesis is true (*type II error*) will be addressed in this section.

12.5.1. Type I error

The following setup is used to consider the type I error in the three different datasets:

Glass: $\mathbf{x}$ consists of two measurements from window number 10 with three features;
$\mathbf{y}$ consists of three measurements from window number 10 with three features.

MDMA: $\mathbf{x}$ consists of 42 measurements from batch 9 of the CHAMP data with three features;
$\mathbf{y}$ consists of five measurements from batch 9 of the CHAMP data with three features.

Simulated: $\mathbf{x}$ consists of five measurements generated by

$$\boldsymbol{\mu}_a \sim \mathcal{N}_3(\boldsymbol{\mu}_\pi, \lambda\boldsymbol{\Sigma}_b), \qquad \boldsymbol{\Sigma}_a \sim \mathcal{W}_3^{-1}(\boldsymbol{\Sigma}_b, \nu_b), \qquad \boldsymbol{\Sigma}_w \sim \mathcal{W}_3^{-1}(\boldsymbol{\Sigma}_e, \nu_e)$$

$$\mathbf{P} \sim \mathcal{N}_3(\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a), \qquad \mathbf{X}_j | \mathbf{P} = \mathbf{p} \overset{\text{iid}}{\sim} \mathcal{N}_3(\mathbf{p}, \boldsymbol{\Sigma}_w), \qquad \text{for} \quad j = 1, 2, \ldots, 5,$$

where $\boldsymbol{\mu}_\pi, \boldsymbol{\Sigma}_b, \boldsymbol{\Sigma}_e, \nu_b, \nu_e$ and λ are as given in (6.2);
$\mathbf{y}$ consists of five measurements generated by

$$\mathbf{Y}_j | \mathbf{P} = \mathbf{p} \overset{\text{iid}}{\sim} \mathcal{N}_3(\mathbf{p}, \boldsymbol{\Sigma}_w), \qquad \text{for} \quad j = 1, 2, \ldots, 5.$$

A Monte Carlo permutation test of size $M = 10,000$ is performed for every test statistic as described in the previous sections. Note that the number of possible permutations for the glass and MDMA datasets are smaller than $10,000$. Since every permutation is sampled with equal probability, it is allowed to do Monte Carlo permutation testing to avoid implementing all possible permutations. This means that instead of an exact p-value an approximate p-value is obtained. The results are given in Table 12.1, where also an approximate 95% confidence interval for $\hat{p}$ can be found. Unfortunately, the test statistic W_0 cannot be computed for the MDMA data, since two measurements are exactly equal which gives a Pearson correlation of 1, and therefore the Fisher transformation cannot be applied.

Dataset	Observed test statistic		$\hat{p}$	Confidence interval	Decision
Glass	T_1	0.0775	0.5058	[0.4960, 0.5156]	Accept
	T_2	0.0977	0.5058	[0.4960, 0.5156]	Accept
	T_∞	0.0760	0.5058	[0.4960, 0.5156]	Accept
	T^2	37.2308	0.2916	[0.2827, 0.3005]	Accept
	W_0	-0.1799	0.5972	[0.5876, 0.6068]	Accept
	P_0	0.2500	0.0000	-	Reject
MDMA	T_1	0.6634	0.9338	[0.9289, 0.9387]	Accept
	T_2	0.9357	0.9325	[0.9276, 0.9374]	Accept
	T_∞	0.5809	0.9178	[0.9124, 0.9232]	Accept
	T^2	2.0494	0.7503	[0.7418, 0.7588]	Accept
	W_0	-	-	-	-
	P_0	0.5422	0.5027	[0.4929, 0.5125]	Accept
Simulated	T_1	1.5183	0.0822	[0.0768, 0.0876]	Accept
	T_2	2.2278	0.0837	[0.0783, 0.0891]	Accept
	T_∞	1.1884	0.1460	[0.1391, 0.1529]	Accept
	T^2	8.1948	0.2111	[0.2031, 0.2192]	Accept
	W_0	-0.2179	0.7087	[0.6998, 0.7176]	Accept
	P_0	0.7937	0.7671	[0.7588, 0.7754]	Accept

Table 12.1: Results of several Monte Carlo permutation tests of size $M = 10,000$ applied to three different datasets, with significance level $\alpha = 0.05$.

All permutation tests result in the correct decision of accepting the prosecution hypothesis, except for the test based on the ranks of interpoint distances applied to the glass dataset. This is caused by the fact that the number of measurements in **x** is equal to two, which means that only one reference interpoint distance can be calculated. Looking at the measurements corresponding to window number 10, the measurements that are assigned to **x** lie closer to each other than the ones assigned to **y**. The reference interpoint distance will therefore always have rank 1 and the observed test statistic T_r will be equal to 9. Permuting **z** can only result in smaller test statistics T_r which implies larger p-values and therefore the p-value based on the original labelling can never be exceeded.

12.5.2. Type II error

To be able to say something about the type II error, **x** remains the same as described previously for every dataset, but the values of **y** are changed. For the glass dataset, **y** is iteratively set equal to the first three measurements from every window in the background material. Similarly, in the MDMA dataset **y** is now given by the first five measurements from every other batch of the CHAMP data than batch 9. Batches with less than five measurements are discarded from the analysis. Finally, background data is simulated by

$$\mathbf{A}_i \overset{\text{iid}}{\sim} \mathcal{N}_3(\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a), \qquad \mathbf{Y}_{ij}|\mathbf{A}_i = \mathbf{a}_i \overset{\text{iid}}{\sim} \mathcal{N}_3(\mathbf{a}_i, \boldsymbol{\Sigma}_w), \qquad \text{for } i = 1,2,\ldots,100 \text{ and } j = 1,2,\ldots,5$$

and **y** is iteratively set equal to the five measurements corresponding to source i. For each possible composition of **y** a Monte Carlo permutation test of size $M = 10{,}000$ is performed for every test statistic and the resulting approximate p-values are saved. The averages of these approximate p-values are given in Table 12.2 together with the percentage of approximate p-values exceeding $\alpha = 0.05$. The behaviour of the approximate p-values is also visualized by the boxplots given in Figures 12.1, 12.2 and 12.3. Note that in these figures the names of the test statistics are only given to indicate which one was used in the permutation tests, but that the data represent approximate p-values.

Test statistic used in permutation tests	Average $\hat{p}$			Percentage of type II errors		
	Glass (15 tests)	MDMA (65 tests)	Simulated (100 tests)	Glass (15 tests)	MDMA (65 tests)	Simulated (100 tests)
T_1	0.0600	0.0524	0.1258	26.7%	10.8%	52.0%
T_2	0.0467	0.0465	0.1144	26.7%	9.2%	48.0%
T_∞	0.0736	0.0529	0.1383	33.3%	10.8%	53.0%
T^2	0.0540	0.0178	0.0496	40.0%	3.1%	25.0%
W_0	0.3877	-	0.2569	73.3%	-	69.0%
P_0	0.0000	0.1620	0.4352	0%	24.6%	76.0%

Table 12.2: Average approximate p-values of several Monte Carlo permutation tests of size $M = 10{,}000$ applied to three different datasets.

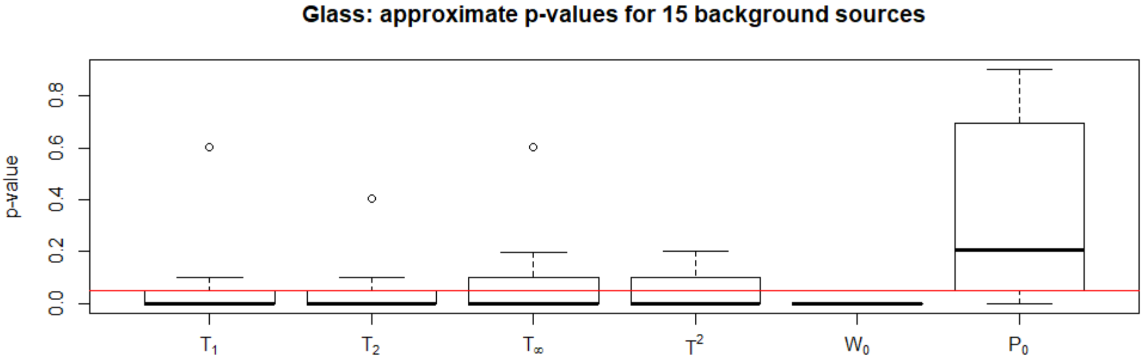

Figure 12.1: Approximate p-values from Monte Carlo permutation tests of size $M = 10{,}000$ for 15 possible compositions of **y** using six different test statistics. The red line indicates $\alpha = 0.05$.

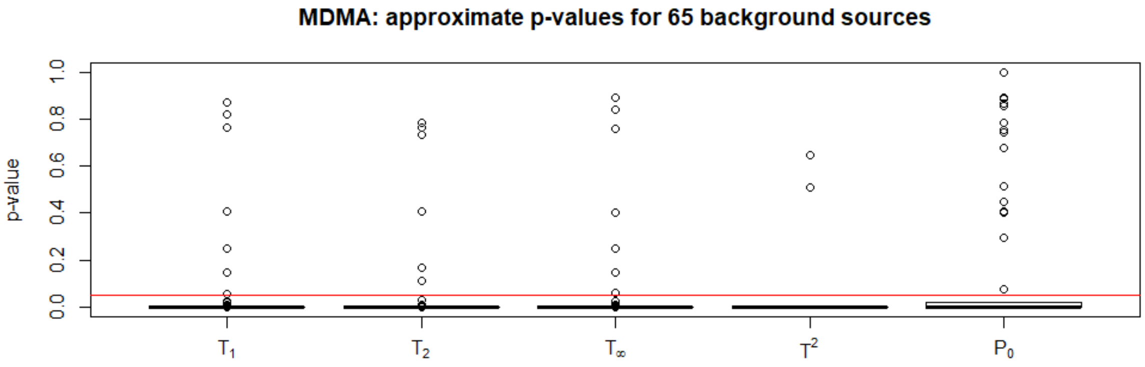

Figure 12.2: Approximate p-values from Monte Carlo permutation tests of size $M = 10,000$ for 65 possible compositions of $\mathbf{y}$ using five different test statistics. The red line indicates $\alpha = 0.05$.

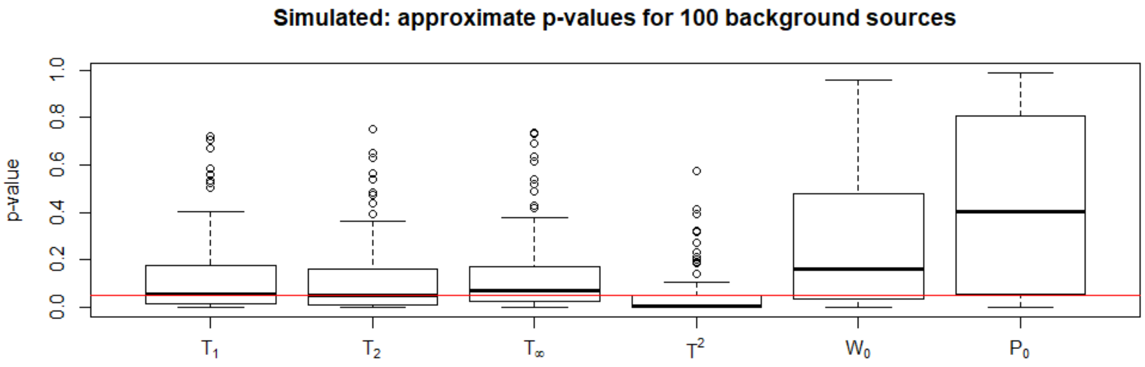

Figure 12.3: Approximate p-values from Monte Carlo permutation tests of size $M = 10,000$ for 100 possible compositions of $\mathbf{y}$ using six different test statistics. The red line indicates $\alpha = 0.05$.

It is possible that the selection of measurements for the composition of $\mathbf{y}$ also influences the results. Therefore, the six test statistics were further examined using the glass dataset. The glass background material consists of 15 sources, each containing five measurements. Therefore there are $\binom{5}{3} = 10$ possibilities to assign three measurements to $\mathbf{y}$. For every source and for every possible composition of $\mathbf{y}$ permutation tests were performed to compute the approximate p-value using each of the six test statistics. The results per test statistic can be found in the boxplots in Appendix B.1. Clearly, both the source and the composition of $\mathbf{y}$ have a lot of influence on the results. The windows corresponding to the labels 3, 4 and 13 result in large approximate p-values for almost every test statistic, indicating that the measurements from these windows are similar to the measurements in $\mathbf{x}$.

Taking both the type I and type II error into account, it seems that Hotelling's T^2 test statistic performs best for these datasets. Of course, the test statistics evaluated here are only a selection of all possibilities. Other test statistics might be considered based on contextual assumptions of the traces. To obtain accurate results from the permutation tests, a sufficient amount of measurements from the traces should be available. Due to the simplicity of the tests and the absence of distributional assumptions, permutation tests can be a useful approach to laboratory based pre-screening of evidence. However, if the value of evidence is required in court, the permutation tests do not offer a solution and other forensic approaches need to be considered.

13 Conclusion

In this research, several statistical models are considered for both continuous and discrete forensic evidence. First, a literature study was carried out to give an overview of the general framework provided by [32] based on the concept of sampling models. These sampling models describe how the different sets of evidence are assumed to be generated. The setup for both the common and specific source problem is clarified and the most important differences between the two problems are highlighted. Using this framework, two statistics to quantify the value of evidence are presented: the likelihood ratio and the Bayes Factor. A summary of the general expressions derived to quantify the evidential value is given and relationships between the two statistics in both the common source and specific source setting from [32] are presented.

The general framework is put into practice by considering the two-level normal-normal model, which is often used for continuous forensic evidence. Using conjugate priors for the (multivariate) normal distribution, an attempt is made to evaluate the Bayes Factor analytically. However, for both the common and specific source problem no explicit expression for the Bayes Factor exists. Therefore, the Bayes Factor is approximated using an appropriate Markov Chain Monte Carlo procedure. The theoretical convergence properties of the methods are discussed for the one-dimensional two-level normal-normal model. For both the common and specific source problem, the resulting Markov chain is proven to be geometrically ergodic under certain constraints.
The two-level normal-normal model is applied to both simulated and real data. Using one-dimensional simulated data, the impact of the choice of hyperparameters on the value of evidence is investigated and explained where possible. Moreover, within the common source problem more conservative values for the Bayes Factor are observed than in the specific source problem. Applying the model to real data, both the likelihood ratio and the Bayes Factor are compared in evaluating the value of evidence.
The likelihood ratio corresponding to the two-level normal-normal model is based on, among others, an estimate of the overall mean. In forensic statistics, the weighted as well as the unweighted mean are commonly used to estimate the overall mean. Under conditions generally encountered in practice, it is shown that the unweighted mean can be preferred over the weighted mean in the sense of mean squared error. Furthermore, a generalisation to the two estimators is given which can be constructed such that an approximation is achieved with a smaller mean squared error than the other candidates. However, this approximation depends on unknown model parameters so that in practice only a ‘toy estimator’ is obtained.
As an alternative to the two-level normal-normal model, copula models are proposed to model the dependencies between features separately. Unfortunately, the general copula model is found to be too difficult to use in practice. Two other existing copula models, using either a Gaussian copula or scores, are briefly discussed to give some idea about how copula theory could be used in forensic evidence evaluation.

For discrete evidence, the general framework is slightly adapted to remove within-source variation from the model. By doing so, the framework from [32] is shown to be suitable for all types of forensic evidence, both continuous and discrete. The one-level Bernoulli model is chosen to exemplify the discrete setup, since this model is frequently used for discrete forensic evidence. The relation between the likelihood ratio and the Bayes Factor is used to indicate which traces should be added to the background material when quantifying the value of evidence. For both the common and specific source problem, the Bayes Factor is calculated analytically. Again, it turns out that within the common source problem more conservative values of evidence are observed than in the specific source problem. The difference between the two problems is found to be most noticeable in case a rare type trace is recovered, i.e., when the number of matches in a database is very small.

Although the main focus in this research is on the value of evidence, the court is primarily interested in the probability that the prosecution hypothesis is true given all available evidence. This posterior probability of guilt is achieved by combining the value of evidence with the prior odds. Two approaches are considered to calculate the posterior probability: using the Bayes Factor, the probability directly follows from an explicit formula, whereas using the likelihood ratio, an integral over the model parameters needs to be evaluated. The theoretical differences and similarities of both approaches are considered and it is found that only when the Bayes Factor is large, the two approaches result in approximately the same probability. For the one-level Bernoulli problem, this means that the two approaches are approximately equal when the size of the background material is large. Moreover, explicit lower and upper bounds are derived for the difference between the two approaches and the difference is shown to be of order $1/n_a^3$, where n_a denotes the size of the background material.

Permutation tests are introduced as a nonparametric alternative to the general framework to evaluate forensic evidence. Using the concept of exchangeability, two evidence sets are compared according to a certain test statistic. Different test statistics are considered and both simulated and real data is used for the analysis. Since the permutation tests neither take background material into account nor quantify the value of evidence explicitly, this approach should only be used for laboratory based pre-screening. If the permutation test shows strong evidence against the prosecution hypothesis, then it might not be worthwhile to complete further statistical analysis using the general framework.

Future work and recommendations

Statistical modelling of forensic evidence is an extremely broad subject and this research has only covered a part of it. There already exists a lot of literature on forensic statistics, but from the subjects addressed in this thesis the following future work and recommendations can be formulated:

- In this research, the main focus is on the two-level normal-normal model and the one-level Bernoulli model. Both models are quite restrictive and are made even more strict by the priors imposed on the model parameters. In practice, it is unlikely that recovered evidence exactly fits in this framework. Therefore, other (nonparametric) models, for instance based on kernel density estimates, are often considered. Alternatively, non-informative priors could be used for the model parameters. It is interesting to explore what impact such models would have on the (difference between) common source and specific source problems.

- The Bayes Factor can often not be computed analytically, but iterative methods can be considered. Besides Monte Carlo integration, other numerical approximation methods of the Bayes Factor, such as Bernstein von Mises or Laplace approximation [32], could be used. New approximation methods will also bring new questions about convergence, which could be investigated from both a theoretical and a practical point of view.

- Until recently, only common source problems were considered in forensic statistics. The introduction of the specific source problem has raised a lot of questions about which approach should be used in practice. To be able to accurately model the specific source distribution, there needs to be a sufficient amount of specific source evidence available. More application to real evidence should decide if using the specific source framework is attainable in practice or that the more conservative common source setup should be preferred.

- Although there already exist many statistical models to evaluate forensic evidence, it is important to keep in mind that alternatives still remain. This is illustrated in this research by considering copula models and permutation tests. Even though the copula models were found to be too complicated to apply to feature-based evidence, the possibilities for score-based evidence could be further investigated. And where the permutation test might be too simplistic to present in court, their nonparametric approach is interesting and could be further explored for the development of new methods to evaluate the value of evidence.

A Detailed calculations

A.1. Two-level normal-normal model

A.1.1. Derivation 1

For the two-level normal-normal model, it follows that

$$\begin{aligned}
&\int_{-\infty}^{\infty}\prod_{j=1}^{n_i} f_a(y_{ij}|a_i,\sigma_w^2)g(a_i|\mu_a,\sigma_a^2)\ da_i\\
&\qquad=\int_{-\infty}^{\infty}\left(2\pi\sigma_w^2\right)^{-n_i/2}\exp\left[-\frac{1}{2\sigma_w^2}\sum_{j=1}^{n_i}(y_{ij}-a_i)^2\right]\left(2\pi\sigma_a^2\right)^{-1/2}\exp\left[-\frac{1}{2\sigma_a^2}(a_i-\mu_a)^2\right]\ da_i\\
&\qquad=\left(2\pi\sigma_w^2\right)^{-n_i/2}\left(2\pi\sigma_a^2\right)^{-1/2}\exp\left[-\frac{\sum_{j=1}^{n_i}y_{ij}^2}{2\sigma_w^2}-\frac{\mu_a^2}{2\sigma_a^2}\right]\int_{-\infty}^{\infty}\exp\left[-\frac{a_i^2}{2}\left(\frac{n_i}{\sigma_w^2}+\frac{1}{\sigma_a^2}\right)+a_i\left(\frac{\sum_{j=1}^{n_i}y_{ij}}{\sigma_w^2}+\frac{\mu_a}{\sigma_a^2}\right)\right]\ da_i\\
&\qquad=\left(2\pi\sigma_w^2\right)^{-n_i/2}\left(2\pi\sigma_a^2\right)^{-1/2}\exp\left[-\frac{\sum_{j=1}^{n_i}y_{ij}^2}{2\sigma_w^2}-\frac{\mu_a^2}{2\sigma_a^2}\right]\exp\left[\frac{1}{2}\frac{\left(\sigma_a^2\sum_{j=1}^{n_i}y_{ij}+\mu_a\sigma_w^2\right)^2}{\sigma_a^2\sigma_w^2\left(n_i\sigma_a^2+\sigma_w^2\right)}\right]\left(2\pi\frac{\sigma_a^2\sigma_w^2}{n_i\sigma_a^2+\sigma_w^2}\right)^{1/2}\\
&\qquad=\left(2\pi\sigma_w^2\right)^{-n_i/2}\left(\frac{\sigma_w^2}{n_i\sigma_a^2+\sigma_w^2}\right)^{1/2}\exp\left[-\frac{\sum_{j=1}^{n_i}y_{ij}^2}{2\sigma_w^2}-\frac{\mu_a^2}{2\sigma_a^2}\right]\exp\left[\frac{1}{2}\frac{\left(\sigma_a^2\sum_{j=1}^{n_i}y_{ij}+\mu_a\sigma_w^2\right)^2}{\sigma_a^2\sigma_w^2\left(n_i\sigma_a^2+\sigma_w^2\right)}\right].
\end{aligned}$$

A.1.2. Derivation 2

The probability density function of the scaled inverse chi-squared distribution with parameters ν and σ^2 is given by [16]

$$p(x)=\frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)}\sigma^{\nu}x^{-(\nu/2+1)}e^{-\nu\sigma^2/(2x)}.$$

Therefore, it follows that

$$\begin{aligned}
&\int_0^{\infty}\int_{-\infty}^{\infty}\prod_{j=1}^{n_u+n_s} f_s(y_{tj}|\mu_s,\sigma_s^2)\pi(\mu_s)\pi(\sigma_s^2)\ d\mu_s\ d\sigma_s^2\\
&\qquad\qquad=\int_0^{\infty}\int_{-\infty}^{\infty}\left(2\pi\sigma_s^2\right)^{-(n_u+n_s)/2}\exp\left[-\frac{1}{2\sigma_s^2}\sum_{j=1}^{n_u+n_s}(y_{tj}-\mu_s)^2\right]\left(2\pi\sigma_s^2/\kappa_\pi\right)^{-1/2}\exp\left[-\frac{\kappa_\pi}{2\sigma_s^2}(\mu_s-\mu_\pi)^2\right]\\
&\qquad\qquad\qquad\times\frac{\left(\sigma_e^2\nu_e/2\right)^{\nu_e/2}}{\Gamma\left(\nu_e/2\right)}\left(\sigma_s^2\right)^{-\nu_e/2-1}\exp\left[-\frac{\nu_e\sigma_e^2}{2\sigma_s^2}\right]\ d\mu_s\ d\sigma_s^2\\
&\qquad\qquad=(2\pi)^{-(n_u+n_s+1)/2}\kappa_\pi^{1/2}\frac{\left(\sigma_e^2\nu_e/2\right)^{\nu_e/2}}{\Gamma\left(\nu_e/2\right)}\int_0^{\infty}\left(\sigma_s^2\right)^{-(\nu_e+n_u+n_s+1)/2-1}\exp\left[-\frac{1}{2\sigma_s^2}\left(\nu_e\sigma_e^2+\sum_{j=1}^{n_u+n_s}y_{tj}^2+\kappa_\pi\mu_\pi^2\right)\right]\\
&\qquad\qquad\quad\times\left(\int_{-\infty}^{\infty}\exp\left[-\frac{\mu_s^2}{2}\frac{\kappa_\pi+n_u+n_s}{\sigma_s^2}+\mu_s\frac{\sum_{j=1}^{n_u+n_s}y_{tj}+\kappa_\pi\mu_\pi}{\sigma_s^2}\right]\ d\mu_s\right)\ d\sigma_s^2
\end{aligned}$$

$$
\begin{aligned}
&= (2\pi)^{-(n_u+n_s+1)/2}\kappa_\pi^{1/2}\frac{\left(\sigma_e^2\nu_e/2\right)^{\nu_e/2}}{\Gamma(\nu_e/2)}\int_0^\infty \left(\sigma_s^2\right)^{-(\nu_e+n_u+n_s+1)/2-1}\exp\left[-\frac{1}{2\sigma_s^2}\left(\nu_e\sigma_e^2+\sum_{j=1}^{n_u+n_s}y_{tj}^2+\kappa_\pi\mu_\pi^2\right)\right]\\
&\quad\times\exp\left[\frac{1}{2\sigma_s^2}\frac{\left(\sum_{j=1}^{n_u+n_s}y_{tj}+\kappa_\pi\mu_\pi\right)^2}{\kappa_\pi+n_u+n_s}\right]\left(2\pi\frac{\sigma_s^2}{\kappa_\pi+n_u+n_s}\right)^{1/2}\,d\sigma_s^2\\
&= (2\pi)^{-(n_u+n_s)/2}\sqrt{\frac{\kappa_\pi}{\kappa_\pi+n_u+n_s}}\frac{\left(\sigma_e^2\nu_e/2\right)^{\nu_e/2}}{\Gamma(\nu_e/2)}\int_0^\infty\left(\sigma_s^2\right)^{-(\nu_e+n_u+n_s)/2-1}\\
&\quad\times\exp\left[-\frac{1}{2\sigma_s^2}\left(\nu_e\sigma_e^2+\sum_{j=1}^{n_u+n_s}y_{tj}^2+\kappa_\pi\mu_\pi^2-\frac{\left(\sum_{j=1}^{n_u+n_s}y_{tj}+\kappa_\pi\mu_\pi\right)^2}{\kappa_\pi+n_u+n_s}\right)\right]\,d\sigma_s^2\\
&= (2\pi)^{-(n_u+n_s)/2}\sqrt{\frac{\kappa_\pi}{\kappa_\pi+n_u+n_s}}\frac{\left(\sigma_e^2\nu_e/2\right)^{\nu_e/2}}{\Gamma(\nu_e/2)}\int_0^\infty\left(\sigma_s^2\right)^{-(\nu_e+n_u+n_s)/2-1}\\
&\quad\times\exp\left[-\frac{1}{2\sigma_s^2}\left(\nu_e\sigma_e^2+\sum_{j=1}^{n_u+n_s}(y_{tj}-\bar{y}_t)^2+\frac{(n_u+n_s)\kappa_\pi}{\kappa_\pi+n_u+n_s}(\mu_\pi-\bar{y}_t)^2\right)\right]\,d\sigma_s^2\\
&= (2\pi)^{-(n_u+n_s)/2}\sqrt{\frac{\kappa_\pi}{\kappa_\pi+n_u+n_s}}\frac{\left(\sigma_e^2\nu_e/2\right)^{\nu_e/2}}{\Gamma(\nu_e/2)}\frac{\Gamma((\nu_e+n_u+n_s)/2)}{\left((\nu_e+n_u+n_s)\sigma_n^2/2\right)^{(\nu_e+n_u+n_s)/2}}\\
&= \frac{\Gamma\left(\frac{\nu_e+n_u+n_s}{2}\right)}{\Gamma\left(\frac{\nu_e}{2}\right)}\sqrt{\frac{\kappa_\pi}{\kappa_\pi+n_u+n_s}}\frac{(\nu_e\sigma_e^2)^{\nu_e/2}}{\left((\nu_e+n_u+n_s)\sigma_n^2\right)^{(\nu_e+n_u+n_s)/2}}\frac{1}{\pi^{(n_u+n_s)/2}},
\end{aligned}
$$

where

$$
\sigma_n^2=\frac{1}{\nu_e+n_u+n_s}\left(\nu_e\sigma_e^2+\sum_{j=1}^{n_u+n_s}(y_{tj}-\bar{y}_t)^2+\frac{(n_u+n_s)\kappa_\pi}{\kappa_\pi+n_u+n_s}(\mu_\pi-\bar{y}_t)^2\right)\qquad\text{and}\qquad\bar{y}_t=\frac{1}{n_u+n_s}\sum_{j=1}^{n_u+n_s}y_{tj}.
$$

A.1.3. Likelihood ratios for higher dimensional model

A similar calculation as **Derivation 1** is needed to derive the likelihood ratios for the higher dimensional models with k features:

$$
\begin{aligned}
&\int_{\mathbb{R}^k}\prod_{j=1}^{n_i}f_a(\mathbf{y}_{ij}|\mathbf{a}_i,\mathbf{\Sigma}_w)g(\mathbf{a}_i|\boldsymbol{\mu}_a,\mathbf{\Sigma}_a)\,d\mathbf{a}_i\\
&=\int_{\mathbb{R}^k}\prod_{j=1}^{n_i}(2\pi)^{-k/2}|\mathbf{\Sigma}_w|^{-1/2}\exp\left[-\frac{1}{2}(\mathbf{y}_{ij}-\mathbf{a}_i)^T\mathbf{\Sigma}_w^{-1}(\mathbf{y}_{ij}-\mathbf{a}_i)\right](2\pi)^{-k/2}|\mathbf{\Sigma}_a|^{-1/2}\exp\left[-\frac{1}{2}(\mathbf{a}_i-\boldsymbol{\mu}_a)^T\mathbf{\Sigma}_a^{-1}(\mathbf{a}_i-\boldsymbol{\mu}_a)\right]\,d\mathbf{a}_i\\
&=(2\pi)^{-k(n_i+1)/2}|\mathbf{\Sigma}_w|^{-n_i/2}|\mathbf{\Sigma}_a|^{-1/2}\exp\left[-\frac{1}{2}\sum_{j=1}^{n_i}\left(\mathbf{y}_{ij}^T\mathbf{\Sigma}_w^{-1}\mathbf{y}_{ij}\right)-\frac{1}{2}\boldsymbol{\mu}_a^T\mathbf{\Sigma}_a^{-1}\boldsymbol{\mu}_a\right]\\
&\quad\times\int_{\mathbb{R}^k}\exp\left[-\frac{1}{2}\mathbf{a}_i^T\left(\mathbf{\Sigma}_a^{-1}+n_i\mathbf{\Sigma}_w^{-1}\right)\mathbf{a}_i+\left(\sum_{j=1}^{n_i}\mathbf{y}_{ij}^T\mathbf{\Sigma}_w^{-1}+\boldsymbol{\mu}_a^T\mathbf{\Sigma}_a^{-1}\right)\mathbf{a}_i\right]\,d\mathbf{a}_i\\
&=(2\pi)^{-k(n_i+1)/2}|\mathbf{\Sigma}_w|^{-n_i/2}|\mathbf{\Sigma}_a|^{-1/2}\exp\left[-\frac{1}{2}\sum_{j=1}^{n_i}\left(\mathbf{y}_{ij}^T\mathbf{\Sigma}_w^{-1}\mathbf{y}_{ij}\right)-\frac{1}{2}\boldsymbol{\mu}_a^T\mathbf{\Sigma}_a^{-1}\boldsymbol{\mu}_a\right](2\pi)^{k/2}\left|\left(\mathbf{\Sigma}_a^{-1}+n_i\mathbf{\Sigma}_w^{-1}\right)^{-1}\right|^{1/2}\\
&\quad\times\exp\left[\frac{1}{2}\left(\sum_{j=1}^{n_i}\mathbf{y}_{ij}^T\mathbf{\Sigma}_w^{-1}+\boldsymbol{\mu}_a^T\mathbf{\Sigma}_a^{-1}\right)\left(\mathbf{\Sigma}_a^{-1}+n_i\mathbf{\Sigma}_w^{-1}\right)^{-1}\left(\mathbf{\Sigma}_w^{-1}\sum_{j=1}^{n_i}\mathbf{y}_{ij}+\mathbf{\Sigma}_a^{-1}\boldsymbol{\mu}_a\right)\right]\\
&=(2\pi)^{-kn_i/2}|\mathbf{\Sigma}_w|^{-n_i/2}|\mathbf{\Sigma}_a|^{-1/2}\left|\mathbf{\Sigma}_a^{-1}+n_i\mathbf{\Sigma}_w^{-1}\right|^{-1/2}\exp\left[-\frac{1}{2}\sum_{j=1}^{n_i}\left(\mathbf{y}_{ij}^T\mathbf{\Sigma}_w^{-1}\mathbf{y}_{ij}\right)-\frac{1}{2}\boldsymbol{\mu}_a^T\mathbf{\Sigma}_a^{-1}\boldsymbol{\mu}_a\right]\\
&\quad\times\exp\left[\frac{1}{2}\left(\sum_{j=1}^{n_i}\mathbf{y}_{ij}^T\mathbf{\Sigma}_w^{-1}+\boldsymbol{\mu}_a^T\mathbf{\Sigma}_a^{-1}\right)\left(\mathbf{\Sigma}_a^{-1}+n_i\mathbf{\Sigma}_w^{-1}\right)^{-1}\left(\mathbf{\Sigma}_w^{-1}\sum_{j=1}^{n_i}\mathbf{y}_{ij}+\mathbf{\Sigma}_a^{-1}\boldsymbol{\mu}_a\right)\right]
\end{aligned}
$$

Using this derivation with the appropriate evidence sets and sources, the likelihood ratios can be easily computed. For the multidimensional common source problem, the likelihood ratio becomes:

$$\begin{aligned} LR_{CS}(\boldsymbol{\theta}_a; e_{u_1}, e_{u_2}) = {} & |\boldsymbol{\Sigma}_a|^{\frac{1}{2}} \left|\boldsymbol{\Sigma}_a^{-1} + n_{u_1}\boldsymbol{\Sigma}_w^{-1}\right|^{\frac{1}{2}} \left|\boldsymbol{\Sigma}_a^{-1} + n_{u_2}\boldsymbol{\Sigma}_w^{-1}\right|^{\frac{1}{2}} \left|\boldsymbol{\Sigma}_a^{-1} + (n_{u_1} + n_{u_2})\boldsymbol{\Sigma}_w^{-1}\right|^{-\frac{1}{2}} \exp\left[\frac{1}{2}\boldsymbol{\mu}_a^T\boldsymbol{\Sigma}_a^{-1}\boldsymbol{\mu}_a\right] \\ & \times \exp\left[\frac{1}{2}\left(\sum_j \mathbf{y}_{uj}^T\boldsymbol{\Sigma}_w^{-1} + \boldsymbol{\mu}_a^T\boldsymbol{\Sigma}_a^{-1}\right)\left(\boldsymbol{\Sigma}_a^{-1} + (n_{u_1} + n_{u_2})\boldsymbol{\Sigma}_w^{-1}\right)^{-1}\left(\boldsymbol{\Sigma}_w^{-1}\sum_j \mathbf{y}_{uj} + \boldsymbol{\Sigma}_a^{-1}\boldsymbol{\mu}_a\right)\right] \\ & \times \exp\left[-\frac{1}{2}\left(\sum_j \mathbf{y}_{u_1 j}^T\boldsymbol{\Sigma}_w^{-1} + \boldsymbol{\mu}_a^T\boldsymbol{\Sigma}_a^{-1}\right)\left(\boldsymbol{\Sigma}_a^{-1} + n_{u_1}\boldsymbol{\Sigma}_w^{-1}\right)^{-1}\left(\boldsymbol{\Sigma}_w^{-1}\sum_j \mathbf{y}_{u_1 j} + \boldsymbol{\Sigma}_a^{-1}\boldsymbol{\mu}_a\right)\right] \\ & \times \exp\left[-\frac{1}{2}\left(\sum_j \mathbf{y}_{u_2 j}^T\boldsymbol{\Sigma}_w^{-1} + \boldsymbol{\mu}_a^T\boldsymbol{\Sigma}_a^{-1}\right)\left(\boldsymbol{\Sigma}_a^{-1} + n_{u_2}\boldsymbol{\Sigma}_w^{-1}\right)^{-1}\left(\boldsymbol{\Sigma}_w^{-1}\sum_j \mathbf{y}_{u_2 j} + \boldsymbol{\Sigma}_a^{-1}\boldsymbol{\mu}_a\right)\right], \end{aligned}$$

where $\mathbf{y}_u = (\mathbf{y}_{u_1 1}, \ldots, \mathbf{y}_{u_1 n_{u_1}}, \mathbf{y}_{u_2 1}, \ldots, \mathbf{y}_{u_2 n_{u_2}})$. This likelihood ratio is the same as the one given in for example [7], which was proven in [22].

For the multidimensional specific source problem, the likelihood ratio is given by:

$$\begin{aligned} LR_{SS}(\boldsymbol{\theta}; e_u) = {} & |\boldsymbol{\Sigma}_s|^{-\frac{n_u}{2}} |\boldsymbol{\Sigma}_w|^{\frac{n_u}{2}} |\boldsymbol{\Sigma}_a|^{\frac{1}{2}} \left|\boldsymbol{\Sigma}_a^{-1} + n_u\boldsymbol{\Sigma}_w^{-1}\right|^{\frac{1}{2}} \exp\left[\frac{1}{2}\sum_j \mathbf{y}_{uj}^T\boldsymbol{\Sigma}_w^{-1}\mathbf{y}_{uj} + \tfrac{1}{2}\boldsymbol{\mu}_a^T\boldsymbol{\Sigma}_a^{-1}\boldsymbol{\mu}_a\right] \\ & \times \exp\left[-\frac{1}{2}\left(\sum_j \mathbf{y}_{uj}^T\boldsymbol{\Sigma}_w^{-1} + \boldsymbol{\mu}_a^T\boldsymbol{\Sigma}_a^{-1}\right)\left(\boldsymbol{\Sigma}_a^{-1} + n_u\boldsymbol{\Sigma}_w^{-1}\right)^{-1}\left(\boldsymbol{\Sigma}_w^{-1}\sum_j \mathbf{y}_{uj} + \boldsymbol{\Sigma}_a^{-1}\boldsymbol{\mu}_a\right)\right] \\ & \times \exp\left[-\frac{1}{2}\sum_j (\mathbf{y}_{uj} - \boldsymbol{\mu}_s)^T\boldsymbol{\Sigma}_s^{-1}(\mathbf{y}_{uj} - \boldsymbol{\mu}_s)\right]. \end{aligned}$$

This is a different likelihood ratio than for example given in [7], since a different model is proposed for the evidence.

A.1.4. Full conditionals for higher dimensional model

To sample from $\pi(\boldsymbol{\theta}_a | e_a, e_{u_1}, e_{u_2}, H_d)$ for the common source problem, the following full conditionals are needed:

$$\begin{aligned} \mathbf{A}_i | \boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a, \boldsymbol{\Sigma}_w, \mathbf{y}, H_d & \sim \mathcal{N}_k\left(\left[\boldsymbol{\Sigma}_a^{-1} + n_i\boldsymbol{\Sigma}_w^{-1}\right]^{-1}\left[\boldsymbol{\Sigma}_w^{-1}\sum_j \mathbf{y}_{ij} + \boldsymbol{\Sigma}_a^{-1}\boldsymbol{\mu}_a\right], \left[\boldsymbol{\Sigma}_a^{-1} + n_i\boldsymbol{\Sigma}_w^{-1}\right]^{-1}\right) \\ \boldsymbol{\mu}_a | \mathbf{a}_1, \ldots, \mathbf{a}_{n_a+2}, \boldsymbol{\Sigma}_a, \boldsymbol{\Sigma}_w, \mathbf{y}, H_d & \sim \mathcal{N}_k\left(\left[\lambda^{-1}\boldsymbol{\Sigma}_b^{-1} + (n_a+2)\boldsymbol{\Sigma}_a^{-1}\right]^{-1}\left[\boldsymbol{\Sigma}_a^{-1}\sum_i \mathbf{a}_i + \lambda^{-1}\boldsymbol{\Sigma}_b^{-1}\boldsymbol{\mu}_\pi\right], \left[\lambda^{-1}\boldsymbol{\Sigma}_b^{-1} + (n_a+2)\boldsymbol{\Sigma}_a^{-1}\right]^{-1}\right) \\ \boldsymbol{\Sigma}_a | \mathbf{a}_1, \ldots, \mathbf{a}_{n_a+2}, \boldsymbol{\mu}_a, \boldsymbol{\Sigma}_w, \mathbf{y}, H_d & \sim \mathcal{W}_k^{-1}\left(\boldsymbol{\Sigma}_b + \sum_i(\mathbf{a}_i - \boldsymbol{\mu}_a)(\mathbf{a}_i - \boldsymbol{\mu}_a)^T, \nu_b + n_a + 2\right) \\ \boldsymbol{\Sigma}_w | \mathbf{a}_1, \ldots, \mathbf{a}_{n_a+2}, \boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a, \mathbf{y}, H_d & \sim \mathcal{W}_k^{-1}\left(\boldsymbol{\Sigma}_e + \sum_i\sum_j(\mathbf{y}_{ij} - \mathbf{a}_i)(\mathbf{y}_{ij} - \mathbf{a}_i)^T, \nu_e + \sum_i n_i\right) \end{aligned}$$

To sample from $\pi(\boldsymbol{\theta}_a | e_a, e_u, H_d)$ and $\pi(\boldsymbol{\theta}_s | e_s, H_d)$ for the specific source problem, the following full conditionals are needed:

$$\begin{aligned} \mathbf{A}_i | \boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a, \boldsymbol{\Sigma}_w, \mathbf{y}, H_d & \sim \mathcal{N}_k\left(\left[\boldsymbol{\Sigma}_a^{-1} + n_i\boldsymbol{\Sigma}_w^{-1}\right]^{-1}\left[\boldsymbol{\Sigma}_w^{-1}\sum_j \mathbf{y}_{ij} + \boldsymbol{\Sigma}_a^{-1}\boldsymbol{\mu}_a\right], \left[\boldsymbol{\Sigma}_a^{-1} + n_i\boldsymbol{\Sigma}_w^{-1}\right]^{-1}\right) \\ \boldsymbol{\mu}_a | \mathbf{a}_1, \ldots, \mathbf{a}_{n_a+1}, \boldsymbol{\Sigma}_a, \boldsymbol{\Sigma}_w, \mathbf{y}, H_d & \sim \mathcal{N}_k\left(\left[\lambda^{-1}\boldsymbol{\Sigma}_b^{-1} + (n_a+1)\boldsymbol{\Sigma}_a^{-1}\right]^{-1}\left[\boldsymbol{\Sigma}_a^{-1}\sum_i \mathbf{a}_i + \lambda^{-1}\boldsymbol{\Sigma}_b^{-1}\boldsymbol{\mu}_\pi\right], \left[\lambda^{-1}\boldsymbol{\Sigma}_b^{-1} + (n_a+1)\boldsymbol{\Sigma}_a^{-1}\right]^{-1}\right) \\ \boldsymbol{\Sigma}_a | \mathbf{a}_1, \ldots, \mathbf{a}_{n_a+1}, \boldsymbol{\mu}_a, \boldsymbol{\Sigma}_w, \mathbf{y}, H_d & \sim \mathcal{W}_k^{-1}\left(\boldsymbol{\Sigma}_b + \sum_i(\mathbf{a}_i - \boldsymbol{\mu}_a)(\mathbf{a}_i - \boldsymbol{\mu}_a)^T, \nu_b + n_a + 1\right) \\ \boldsymbol{\Sigma}_w | \mathbf{a}_1, \ldots, \mathbf{a}_{n_a+1}, \boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a, \mathbf{y}, H_d & \sim \mathcal{W}_k^{-1}\left(\boldsymbol{\Sigma}_e + \sum_i\sum_j(\mathbf{y}_{ij} - \mathbf{a}_i)(\mathbf{y}_{ij} - \mathbf{a}_i)^T, \nu_e + \sum_i n_i\right) \\ \boldsymbol{\mu}_s | \boldsymbol{\Sigma}_s, \mathbf{y}_s, H_d & \sim \mathcal{N}_k\left(\left[\boldsymbol{\Sigma}_b^{-1} + n_s\boldsymbol{\Sigma}_s^{-1}\right]^{-1}\left[\boldsymbol{\Sigma}_s^{-1}\sum_j \mathbf{y}_{sj} + \boldsymbol{\Sigma}_b^{-1}\boldsymbol{\mu}_\pi\right], \left[\boldsymbol{\Sigma}_b^{-1} + n_s\boldsymbol{\Sigma}_s^{-1}\right]^{-1}\right) \\ \boldsymbol{\Sigma}_s | \boldsymbol{\mu}_s, \mathbf{y}_s, H_d & \sim \mathcal{W}_k^{-1}\left(\boldsymbol{\Sigma}_e + \sum_j(\mathbf{y}_{sj} - \boldsymbol{\mu}_s)(\mathbf{y}_{sj} - \boldsymbol{\mu}_s)^T, \nu_e + n_s\right) \end{aligned}$$

B

Additional figures

B.1. Boxplots of approximate p-values for glass data

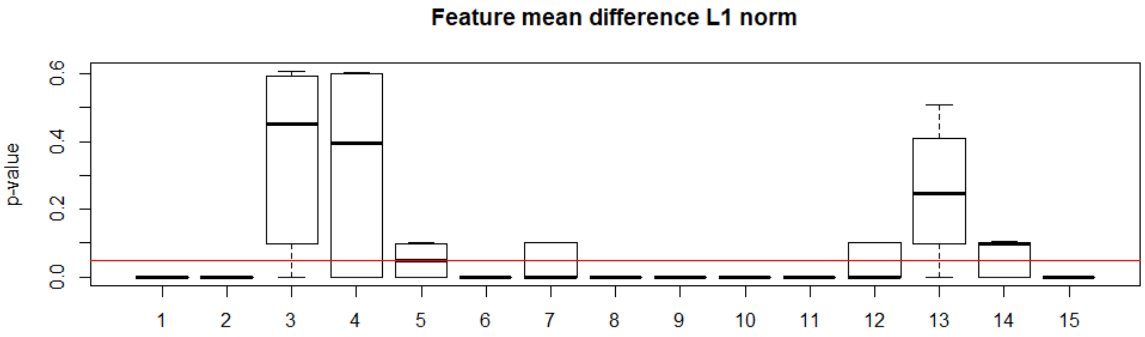

Figure B.1: Approximate p-values from Monte Carlo permutation tests of size $M = 10,000$ for 10 possible compositions of $\mathbf{y}$ for each of 15 windows using the test statistic T_1. The red line indicates $\alpha = 0.05$.

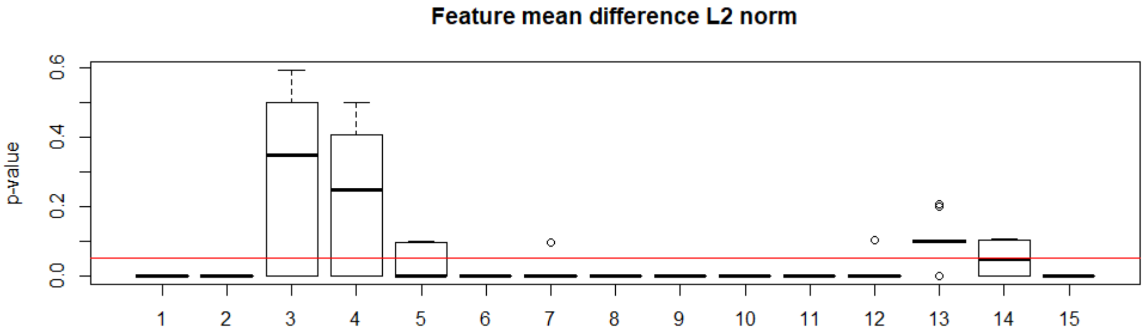

Figure B.2: Approximate p-values from Monte Carlo permutation tests of size $M = 10,000$ for 10 possible compositions of $\mathbf{y}$ for each of 15 windows using the test statistic T_2. The red line indicates $\alpha = 0.05$.

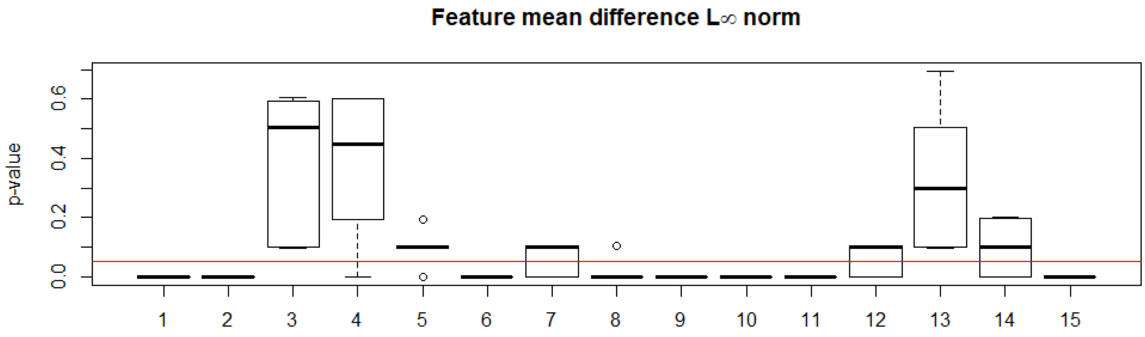

Figure B.3: Approximate p-values from Monte Carlo permutation tests of size $M = 10{,}000$ for 10 possible compositions of $\mathbf{y}$ for each of 15 windows using the test statistic T_∞. The red line indicates $\alpha = 0.05$.

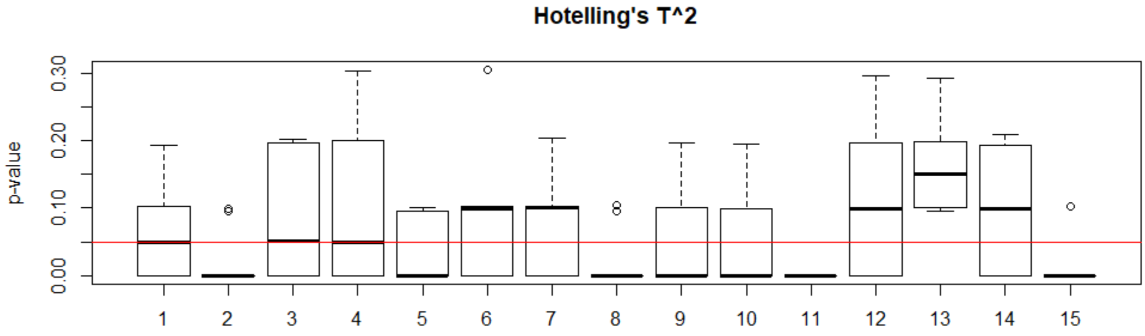

Figure B.4: Approximate p-values from Monte Carlo permutation tests of size $M = 10{,}000$ for 10 possible compositions of $\mathbf{y}$ for each of 15 windows using the test statistic T^2. The red line indicates $\alpha = 0.05$.

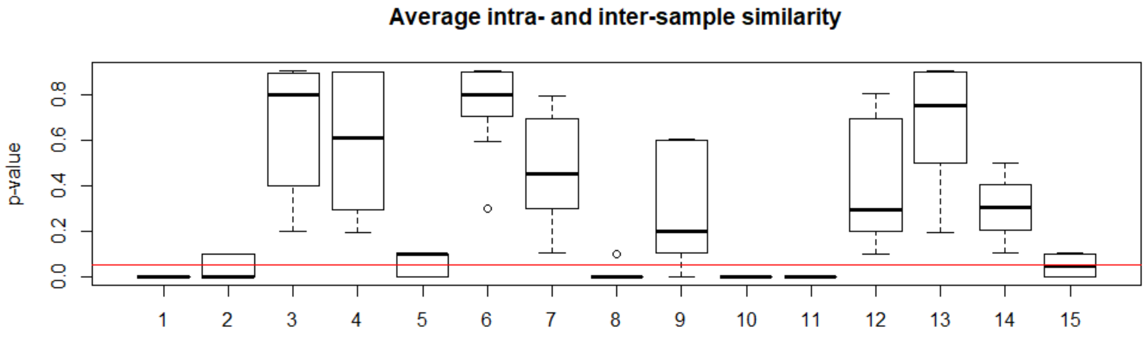

Figure B.5: Approximate p-values from Monte Carlo permutation tests of size $M = 10{,}000$ for 10 possible compositions of $\mathbf{y}$ for each of 15 windows using the test statistic W_0. The red line indicates $\alpha = 0.05$.

The boxplot for the test statistic P_0 based on the ranks of interpoint distances is not given, since all approximate p-values are equal to 0.

Bibliography

[1] M. Abramowitz and I.A. Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, volume 55. Courier Corporation, 1964.

[2] C.G.G. Aitken and D. Lucy. Evaluation of trace evidence in the form of multivariate data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 53(1):109–122, 2004.

[3] C.G.G. Aitken and F. Taroni. *Statistics and the evaluation of evidence for forensic scientists*, volume 16. Wiley Online Library, 2004.

[4] I. Alberink, A. Bolck, and S. Menges. Posterior likelihood ratios for evaluation of forensic trace evidence given a two-level model on the data. *Journal of Applied Statistics*, 40(12):2579–2600, 2013.

[5] Association of Forensic Science Providers. Standards for the formulation of evaluative forensic science expert opinion. *Science & Justice*, 49:161–164, 2009.

[6] K.B. Athreya and S.N. Lahiri. *Measure theory and probability theory*. Springer Science & Business Media, 2006.

[7] A. Bolck and I. Alberink. Variation in likelihood ratios for forensic evidence evaluation of XTC tablets comparison. *Journal of Chemometrics*, 25(1):41–49, 2011.

[8] A. Bolck, C. Weyermann, L. Dujourdy, P. Esseiva, and J. van den Berg. Different likelihood ratio approaches to evaluate the strength of evidence of MDMA tablet comparisons. *Forensic Science International*, 191(1-3):42–51, 2009.

[9] S. Bozza, F. Taroni, R. Marquis, and M. Schmittbuhl. Probabilistic evaluation of handwriting evidence: likelihood ratio for authorship. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 57(3): 329–341, 2008.

[10] Z. Brzezniak and T. Zastawniak. *Basic stochastic processes: a course through exercises*. Springer Science & Business Media, 2000.

[11] G.P. Campbell and J.M. Curran. The interpretation of elemental composition measurements from forensic glass evidence III. *Science and Justice*, 49(1):2–7, 2009.

[12] G. Cereda. Bayesian approach to LR assessment in case of rare type match. *Statistica Neerlandica*, 71(2): 141–164, 2017.

[13] E. Chikayama. Decomposition of multivariate function using the Heaviside step function. *SpringerPlus*, 3(1):704, 2014.

[14] A.P. Dawid and J. Mortera. Coherent analysis of forensic identification evidence. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 425–443, 1996.

[15] R.A. Fisher. *Statistical Methods For Research Workers*. Oliver And Boyd: Edinburgh, 1936.

[16] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin. *Bayesian data analysis*, volume 2. CRC press Boca Raton, FL, 2014.

[17] P.I. Good. *Resampling methods*. Springer, 2006.

[18] J.P. Hobert and C.J. Geyer. Geometric ergodicity of Gibbs and block Gibbs samplers for a hierarchical random effects model. *Journal of Multivariate Analysis*, 67(2):414–430, 1998.

[19] A.A. Johnson. *Geometric ergodicity of Gibbs samplers*. University of Minnesota, 2009.

[20] G.L. Jones, J.P. Hobert, et al. Sufficient burn-in for Gibbs samplers for a hierarchical random effects model. *The Annals of Statistics*, 32(2):784–817, 2004.

[21] D. Karp and S.M. Sitnik. Inequalities and monotonicity of ratios for generalized hypergeometric function. *Journal of Approximation Theory*, 161(1):337–352, 2009.

[22] F.S. Kool. Feature-based models for forensic likelihood ratio calculation: Supporting research for the ENFSI-LR project, 2016. URL `https://repository.tudelft.nl/islandora/object/uuid:5c088097-b0f0-4342-9737-202c81e7212d/`.

[23] F.S. Kool, D. Steenhuis, R. Neijmeijer, D. Kruise, and A. Bolck. User Manual SAILR - Version 1.3.0, June 2017.

[24] D.V. Lindley. A problem in forensic science. *Biometrika*, 64(2):207–213, 1977.

[25] L. Lorentzen and H. Waadeland. *Continued fractions*, volume 1. Atlantis Press, 2008.

[26] M. Marozzi. Tests for comparison of multiple endpoints with application to omics data. *Statistical applications in genetics and molecular biology*, 17(1), 2018.

[27] K.A. Martire, R.I. Kemp, M. Sayle, and B.R. Newell. On the interpretation of likelihood ratios in forensic science evidence: Presentation formats and the weak evidence effect. *Forensic science international*, 240:61–68, 2014.

[28] E. McIntee, E. Viglino, S. Kumor, C. Rinke, L. Ni, and M.E. Sigman. Non-parametric permutation test for the discrimination of float glass samples based on LIBS spectra. *Journal of Chemometrics*, 24(6):312–319, 2010.

[29] D.R. Musgrove, J. Hughes, and L.E. Eberly. Hierarchical copula regression models for areal data. *Spatial Statistics*, 17:38–49, 2016.

[30] R.B. Nelsen. *An introduction to copulas*. Springer Science & Business Media, 2007.

[31] Danica M Ommen, Christopher P Saunders, and Cedric Neumann. The characterization of Monte Carlo errors for the quantification of the value of forensic evidence. *Journal of Statistical Computation and Simulation*, 87(8):1608–1643, 2017.

[32] D.M. Ommen. *Approximate Statistical Solutions to the Forensic Identification of Source Problem*. PhD thesis, 2017.

[33] D.M. Ommen and C.P. Saunders. Building a unified statistical framework for the forensic identification of source problems. *Law, Probability and Risk*, 17(2):179–197, 2018.

[34] D.M. Ommen, C.P. Saunders, and C. Neumann. An argument against presenting interval quantifications as a surrogate for the value of evidence. *Science & Justice*, 56(5):383–387, 2016.

[35] K.B. Petersen and M.S. Pedersen. The Matrix Cookbook, 2012. URL `https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf`.

[36] J. Rice. *Mathematical statistics and data analysis*. Nelson Education, 2006.

[37] C.P. Robert. *Monte Carlo methods*. Wiley Online Library, 2004.

[38] C.P. Robert and G. Casella. *Introducing Monte Carlo methods with R*, volume 18. Springer, 2010.

[39] B. Robertson and G.A. Vignaux. Probability—the logic of the law. *Oxford Journal of Legal Studies*, 13(4):457–478, 1993.

[40] R.M. Royall. Statistical evidence: A likelihood paradigm, 1997.

[41] H. Sahai and M.M. Ojeda. *Analysis of Variance for Random Models, Volume 2: Unbalanced Data*, volume 2. Springer Science & Business Media, 2004.

[42] M.J. Schervish. *Theory of statistics*. Springer Science & Business Media, 2012.

[43] SCHOTT North America, Inc. Technical glasses: Physical and technical properties. *SCHOTT Download Library*, 2014.

[44] S.R. Searle, G. Casella, and C.E. McCulloch. *Variance components.* John Wiley & Sons, New York, 1992.

[45] M. Sjerps and A.D. Kloosterman. Statistical aspects of interpreting dna profiling in legal cases. *Statistica neerlandica*, 57(3):368–389, 2003.

[46] N. Susyanto. *Semiparametric copula models for biometric score level fusion.* PhD thesis, 2016.

[47] F. Taroni, S. Bozza, A. Biedermann, P. Garbolino, and C.G.G. Aitken. *Data analysis in forensic science: a Bayesian decision perspective*, volume 88. John Wiley & Sons, 2010.

[48] F. Taroni, S. Bozza, A. Biedermann, and C.G.G. Aitken. Dismissal of the illusion of uncertainty in the assessment of a likelihood ratio. *Law, probability and risk*, 15(1):1–16, 2015.

[49] B.S. Weir. *Genetic data analysis II.* Sinauer Associates, Sunderland, 1996.

[50] P. Xue-Kun Song. Multivariate dispersion models generated from Gaussian copula. *Scandinavian Journal of Statistics*, 27(2):305–320, 2000.

[51] G. Zadora, A. Martyna, D. Ramos, and C.G.G. Aitken. *Statistical analysis in forensic science: evidential values of multivariate physicochemical data.* John Wiley & Sons, 2014.

[52] I. Žežula. On multivariate Gaussian copulas. *Journal of Statistical Planning and Inference*, 139(11):3942–3946, 2009.

Index