2017

Approximate Statistical Solutions to the Forensic Identification of Source Problem

Danica M. Ommen
South Dakota State University

APPROXIMATE STATISTICAL SOLUTIONS TO THE FORENSIC IDENTIFICATION OF SOURCE PROBLEM

BY

DANICA M. OMMEN

A dissertation submitted in partial fulfillment of the requirements for the

Doctor of Philosophy

Major in Computational Science & Statistics

South Dakota State University

2017

APPROXIMATE STATISTICAL SOLUTIONS TO THE

FORENSIC IDENTIFICATION OF SOURCE PROBLEM

DANICA M. OMMEN

This dissertation is approved as a creditable and independent investigation by a candidate for the Doctor of Philosophy in Computational Science and Statistics degree and is acceptable for meeting the dissertation requirements for this degree. Acceptance of this does not imply that the conclusions reached by the candidate are necessarily the conclusions of the major department.

Christopher Saunders, Ph.D.   Date
Dissertation Advisor

Kurt Cogswell, Ph.D.   Date
Head, Department of Mathematics & Statistics

Dean, Graduate School   Date

This dissertation is dedicated with love to my guardian angels:

Charles Crotty & Ross Ommen, my grandpas,

and Daniel Ommen, my uncle.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my committee: Dr. Christopher Saunders, Dr. Cedric Neumann, Dr. Kurt Cogswell, and Dr. Allen Jones. To my advisor, committee chair, mentor, and friend Dr. Christopher Saunders: I cannot express how greatly I appreciate your sound advice, everlasting patience, and eternal correctness (for which I owe you a lifetime of beer). I am so thankful for everything you have endured and sacrificed to make this dissertation possible. Also, thank you for starting me down the path of forensic statistics. You helped me find my place in this world doing "real math" that makes a difference.

I am very grateful to Drs. Neumann and Saunders for kindly employing me via their National Institute of Justice funded grant while I completed my degree. Your support provided me with the opportunity to attend the numerous conferences and events which have helped build a strong foundation for my career in forensic statistics. In addition, I had the opportunity to participate in the Visiting Scientist program at the Counterterrorism and Forensic Science Research Unit of the FBI Laboratory Division. I owe thanks to my mentor Dr. JoAnn Buscaglia for providing me with plenty of stimulating, real-life examples and datasets. Thank you for giving me an experience at the FBI that was incredibly beneficial for the work in this dissertation.

Finally, a large amount of the successful completion of this dissertation can be attributed to the love and support of my parents, Steve and Char Ommen, and my boyfriend, Matt Jensen. You guys were always there to pick me up after a difficult day, and the first to celebrate with me on the good days. Thanks for helping with cleaning, laundry, cooking, and all the everyday things that can make life as a

graduate student overwhelming without your assistance. I love you!

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

ABSTRACT

APPROXIMATE STATISTICAL SOLUTIONS TO THE

FORENSIC IDENTIFICATION OF SOURCE PROBLEM

DANICA M. OMMEN

2017

Currently in forensic science, the statistical methods for solving the identification of source problems are inherently subjective and generally ad-hoc. The formal Bayesian decision framework provides the most statistically rigorous foundation for these problems to date. However, computing a solution under this framework, which relies on a Bayes Factor, tends to be computationally intensive and highly sensitive to the subjective choice of prior distributions for the parameters. Therefore, this dissertation aims to develop statistical solutions to the forensic identification of source problems which are less subjective, but which retain the statistical rigor of the Bayesian solution.

First, this dissertation focuses on computational issues during the subjective quantification of the Bayes Factor, and on characterizing the numerical error associated with the resulting quantification. Secondly, the asymptotic properties of the Bayes Factor for a fixed set of unknown source evidence are considered as the number of control samples increases. Under the formal Bayesian paradigm, Doob's Consistency Theorem implies that a Bayesian believes in the existence of a value of evidence analogous to a true likelihood ratio in the Frequentist paradigm. Finally, two approximations

to the value of evidence for the forensic identification of source problems are derived relative to the existence of a true likelihood ratio.

The first approximation is derived as a result of the Bernstein-von Mises Theorem. This Bernstein-von Mises approximation eliminates the determination of prior distributions for the parameters. Under suitable conditions, the Bernstein-von Mises approximation converges in probability to the Bayes Factor as the size of the control samples increases. However the Bernstein-von Mises approximation suffers from similar computational issues as the Bayes Factor. The second approximation is derived as a result of various theorems regarding the asymptotic properties of M-estimators. This Neyman-Pearson approximation requires no prior distributions, and is generally less computationally intractable. Under suitable conditions, the Neyman-Pearson approximation converges in probability to the true likelihood ratio as the number of control samples increases. In addition, the Neyman-Pearson approximation can replace the Bayes Factor in the forensic identification of source problems, and result in decisions that are approximately equivalent to using the Bayes Factor.

CHAPTER 1

Introduction & Overview

1.1   Introduction

The forensic identification of source problem seeks an answer to the question of where a collection of forensic evidence originated. The point of origin may be a person, as is the case for DNA and handwriting evidence, or a specific object or collection of objects, as is the case with firearms and glass evidence. This type of problem is typically of interest to the criminal justice system. In the quest for the answer to this problem, the evidence interpretation expert is expected to summarize the observed evidence relative to two competing propositions, often referred to as the prosecution and defense propositions, for how the evidence was generated. This can be done in a number of different ways, but the most fully developed structure for solving this type of problem is a Bayesian decision framework.

Under the Bayesian decision framework, the forensic statistician is tasked with providing the value of evidence as the summary of the observed evidence relative the two competing hypotheses (as well as the necessary prior belief for the nuisance parameters). Traditionally, the value of evidence is given in the form of the Bayes Factor. The Bayes Factor is used to convert the prior odds to posterior odds. The prior odds summarizes the decision-maker's relative personal belief concerning the validity of the prosecution and defense hypotheses before observing the evidence. The value of evidence then allows the decision-maker to update the personal prior belief following the

observation of the evidence and arrive at the posterior odds concerning the relative validity of the two hypotheses after observing the evidence.

Typically when considering forensic evidence, the forensic scientist is concerned with source or sub-source level propositions or hypotheses, although activity level propositions might be considered in some cases. However, the court system is typically concerned with offense level propositions concerning the guilt or innocence of the defendant (for a detailed description of the hierarchy of propositions see Cook et al. [17] or the ENFSI guidelines [24]). The focus of this dissertation is on two different classes of source level identification problems and the corresponding forms of the value of evidence under each problem.

Historically, the forensic identification of source problems have been approached from the perspective of Bayesian hypothesis testing to aid the final decision-making process. Recently, there has been some push-back on the Bayesian decision-making process in forensic science [47, 79]. The criticisms are that the prior odds belong to the decision-maker, while the Bayes Factor belongs to the forensic expert, leading to an invalid updating of the prior odds to arrive at the posterior odds for the decision-maker [47]. Also, there has been a recent push in the United States to increase the rigor of statistics in forensic science [58]. One of the main efforts towards this goal, and one of the main areas of debate (see the virtual special issue in Science and Justice entitled "Measuring and Reporting the Precision of Forensic Likelihood Ratios" [9, 13, 18, 50, 54, 72, 74]), is characterizing the various aspects of uncertainty in quantifying the value of evidence.

1.2 Historical Review

Dating back to the 1970s and 1980s, there were a series of papers regarding a problem in forensic science in which a window was broken to gain entry into a residence and a suspect identified in relation to the crimes committed therein. On the suspect are found fragments of glass whose source is unknown. The intention is to compare these fragments to fragments of glass collected from the crime scene window to determine whether they match. This comparison is made by way of the refractive index measurement taken on each individual glass fragment. A solution to this problem, which is an example of the forensic identification of source problem, was provided by (at least) three authors during this time-frame, Lindley [46], Evett [25], and Grove [39]. The details of each solution will be discussed briefly below.

Lindley's solution, detailed in his seminal paper "A Problem in Forensic Science" in 1977, is often referenced as the foundation of Bayesian model selection techniques for general applications, and in particular for applications in forensic science. For a more detailed historical discussion relating to Lindley's solution, the interested reader is referred to Good [38]. To illustrate his technique in smooth mathematical form, Lindley assumes that the refractive index measured on each of the glass fragments follows a normal distribution with an unknown mean and a known variance [46]. Then the parameter for the mean is attached a normal distribution with known mean and variance as a prior. Standard Bayesian techniques are used to show that the corresponding normal densities can be used to form a Bayes Factor for deciding whether the mean parameter for the unknown source glass fragments takes the same value as the mean parameter for the glass fragments collected from the scene of the crime, or whether they are different values [46]. Lindley then uses a clever transformation of variables to arrive at a closed-form solution of the Bayes Factor under the (unrealistic) assumption of normality. However, it should be noted that in the case of non-normality, a

closed-form expression of the Bayes Factor often does not exist. Therefore, Lindley suggests a Taylor series expansion for non-normal densities to cover this more realistic setting [46], but this result in a numerical approximation (often very complicated and nearly as intractable) of the Bayes Factor instead of an exact value. Nearing the conclusion of this discussion, Lindley assumes that the identification of a common source for the glass fragments implies the guilt of the accused suspect [46].

Lindley then compares his approach to the two-staged approach developed by Parker in 1966 [56] and used by Evett in his 1977 paper "The Interpretation of Refractive Index Measurements" which considers the same problem of comparing glass fragments under the assumption of normality [25]. Initially, Evett's approach consisted of two-different stages. The first stage is a comparison stage under which the mean refractive index for the glass with unknown source is compared to that of the glass from the crime scene using a standard normal Frequentist hypothesis test [25]. If under the first stage, the fragments are concluded to be dissimilar, the second stage is forgone. However, if the fragments are concluded to be similar, the second stage consists of determining the probability that the glass fragments at the scene of the crime would be deemed similar to glass fragments found in the general population [25]. When the probability of coincidence is smaller, greater significance should be placed on the result obtained from the first stage [26]. Evett provides a series of follow-up papers [26, 28, 29, 30] with generalizations of the result to make it more applicable to casework where normality is often unrealistic. These solutions become more and more computationally complex and intractable as the assumptions become more general (and more practical).

The third approach to the forensic identification of source problem arrived a few short years later than Lindley and Evett. Grove added his solution in his 1980 paper "The Interpretation of Forensic Evidence using a Likelihood Ratio" which again considered

the comparison of glass fragments under the assumptions of normality [39]. Grove suggests that Lindley's assumption that the mean parameter for the unknown source evidence be attributed a prior distribution which is the same prior distribution as the mean parameter for the crime scene evidence is unappealing to "orthodox non-Bayesians" [39]. Grove's solution is to consider replacing this parameter with its maximum likelihood estimate in the Bayes Factor suggested by Lindley. Under various simplifying assumptions, it can be shown that Grove's likelihood ratio is quite similar to Lindley's Bayes Factor [39]. While Grove contributed to the advancement of the solution techniques for the forensic identification of source problem, he admits that these solutions are actually answering to the wrong question. The problem of interest to the court system is on proving whether the accused is guilty of committing the crime, and Grove shows that his likelihood ratio for determining whether the two sets of glass fragments match does not provide any direct information regarding the likelihood ratio for determining guilt/innocence [39].

It should be noted that this may be the point where the lines blurred between a Bayes Factor and a likelihood ratio, and the two phrases started to be used interchangeably in forensic science. Grove's likelihood ratio is somewhere in-between a "plug-in" likelihood ratio, typically used in DNA applications [19, 14], and the Bayes Factor presented by Lindley. Additionally, Seheult showed that Lindley's Bayes Factor provides equivalent results to the traditional Neyman-Pearson hypothesis test (which uses the likelihood ratio test statistic), further obscuring the difference between the two quantities. The differences between the Bayes Factor and likelihood ratio are presented in Section 3.3 along with various relationships between the two quantities, presented in Section 5. An interesting issue was first revealed by Shafer in his commentary "Lindley's Paradox" in 1982 [66]. Lindley's paradox occurs when two different methods for assessing the exact same evidence under the exact same circumstances provide opposing conclusions. This is precisely the result that Shafer shows

for the Bayes Factor presented by Lindley and the two-stage approach presented by Evett [25]. This difference between the Bayes Factor and the two-stage approach occurs when "the prior density under the alternative hypothesis is very diffuse relative to the discriminating power of the evidence" [66].

Criticism from Lindley [46] of two-stage approach led many European researchers associated with the Forensic Science Service to abandon the two-staged approach, and research efforts shifted towards approximating the Bayesian solution presented by Lindley (because of its intractability under practical assumptions). Even Evett considered the Bayesian approach to the glass problem in his 1986 paper [27]. Evett's solution in this case takes the form of a "plug-in" likelihood ratio where estimates of nuisance parameters calculated from background data are used in place of the actual values for the parameters. Evett discloses in this paper that this solution is far from optimal, and that further efforts were needed to improve the result [27]. Chan and Aitken [16] also provide an estimate of the Bayes Factor, but using informative kernel density estimates of prior distributions for the non-normal nuisance parameter based on a set of training data. In this paper, the estimated Bayes Factors have closed-form solutions in all of the scenarios considered. In order to evaluate the performance of the estimated Bayes Factor, a simulation study was designed. When the actual prior distribution for the nuisance parameter was non-normal, the kernel density method performed better than using a normal prior density for the nuisance parameter [16].

The methods described above, and many other similar methods, relied on the assumptions that there was a large background database, or training set, available from which reliable estimates of the associated parameters could be obtained. Issues with these methods started surfacing in the presence of uncertainty in the background population [6]. Due to this added complication, many ad-hoc solutions of the forensic

identification of source problem surfaced [14]. As the number and range of methods for quantifying the value of evidence increased, the call for further research to increase the statistical rigor of these methods escalated [51, 24, 58].

1.3   Contributions & Chapter Summaries

To begin, a review of standard results from multivariate statistics, Bayesian statistics, and Markov Chain Monte Carlo methods is provided. Most of the preliminary results needed to complete the proofs in this dissertation are provided in Chapter 2.

In Chapter 3, the forensic identification of source problem is considered under an institutional decision process [12, 35], which attempts to avoid any issues with the current process related to the mismatch of the creator of Bayes Factor to the creator of the prior and posterior odds. Also, in Chapter 3, the forensic identification of source problem is developed under a Bayesian model selection framework, as opposed to hypothesis testing, since the assumptions of the problem have become more generalized since the initial papers by Lindley [46], Evett [25], and Grove [39]. It should also be of note that historically, the forensic identification of source problem considered the "matching" of two sets of glass fragments to a common, but unknown, source. This is what will be called the *common source* problem in Section 3.1. In these initial problems, the source of the glass fragments was considered to be a random source typical of the background population. An alternative approach to the problem is provided in Section 3.2, called the *specific source* problem. In this problem, the source of the glass fragments would be considered a fixed source related to the specific source under investigation. The distinction between the common source and specific source problems is important because the treatment of the two problems is different: the forensic hypotheses, the sampling models, the Bayes Factor, and the likelihood ratio

are all different between the two identification of source problems.

In Chapter 4, a single asymptotic approximation of the standard error associated with quantifying the common source and specific source Bayes Factors using Monte Carlo integration is derived. While many other possible numerical approximation techniques for quantifying the Bayes Factor have been proposed, see for example Kass and Raftery [44], Han and Carlin [41], and Carlin and Chib [15], Monte Carlo integration was chosen since it provided the most straight-forward method of quantifying the numerical standard error. A simulation study is designed to explore the reasonableness of these approximate Monte Carlo standard errors. In addition, the Bayes Factors were computed using a number of different choices for prior distributions on the parameters. It is a well-known issue that Bayes Factors can be sensitive to the choice of prior distribution for the associated parameters [44]. This simulation study revealed that the specific source Bayes Factors are much more sensitive to this choice than the common source Bayes Factors, which is an issue since the specific source Bayes Factor is more appropriate for use in the court system. Therefore, the conclusion of Chapter 4 is that methods less-sensitive to the subjective prior choice be developed for use in the court system.

In Chapter 5, advancements are made towards developing methods which are less sensitive to the subjective choice of prior. First, two alternative forms for the Bayes Factor are developed which relate the Bayes Factor to the likelihood ratio function. Then using these alternative forms and Doob's Consistency Theorem, it can be shown that for almost-every prior parameter value, and for almost-every infinite sequence of data, the Bayes Factor converges to the likelihood ratio. This result implies that a number of other results can reasonably be applied to the forensic identification of source problems.

As a consequence, in Chapter 6, the Bernstein-von Mises Theorem is used to propose

an approximation to the Bayes Factor which reduces the sensitivity of the resulting value of evidence to the choice of prior distributions. Furthermore, it will be shown that the difference between this Bernstein-von Mises approximation and the Bayes Factor will converge to zero as the amount of relevant evidence gathered increases, see Section 6.1. However, this approximation does not avoid the use of computationally intense MCMC techniques. To avoid these techniques, one could use the Laplace approximation of the Bayes Factor, see Section 6.3 for details. Still, this approximation is lacking since it has the potential for sensitivity to the choice of subjective prior distributions.

Finally, the Neyman-Pearson approximation is presented in Chapter 7. This approximation requires no MCMC techniques and avoids any specification of prior distributions. Under reasonable circumstances, the Neyman-Pearson approximation will converge to the true value of the likelihood ratio using a Delta method technique and properties of M-estimators. This Neyman-Pearson approximation can also replace the Bayes Factor in the Bayesian decision-making process, and lead to an approximately equivalent result with less computational hassle.

Related to the topic of quantifying uncertainty in the value of evidence, there are many researchers who are exploring the use of confidence or credible intervals for the computed likelihood ratio [9, 13, 18, 50, 54, 72, 74]. In Ommen et al. [54], it is argued that if an interval is use to characterize some aspect of uncertainty in a computed value of evidence, then a single-number summary of that interval should be available from which a logical and coherent decision is possible. In Chapter 8, a method of computing credible intervals for the likelihood ratio is derived from which the Bayes Factor (or an approximation of it, including the Neyman-Pearson approximation) is the natural choice for the summary of the interval to use as the value of evidence in the Bayesian decision-making process.

CHAPTER 2

Introductory Results from Statistics

The purpose of this chapter is to introduce a common notation to be used throughout this dissertation. Additionally, the following sections will provide definitions and results from standard multivariate and Bayesian statistics, as well as asymptotic statistics, empirical process theory, and Markov Chain Monte Carlo (MCMC) methods that will be used and referenced throughout this dissertation.

2.1 Multivariate Statistics

In order to define the models and distributions needed to calculate the value of evidence, we will need some results from multivariate statistics [43, 33, 4]. The following notation will be used to denote the expected value of a random column vector $X = \begin{pmatrix} X_1 & X_2 & \cdots & X_k \end{pmatrix}^T$,

$$E[X] = \begin{pmatrix} E[X_1] & E[X_2] & \cdots & E[X_k] \end{pmatrix}^T \equiv \begin{pmatrix} \mu_1 & \mu_2 & \cdots & \mu_k \end{pmatrix}^T = \mu, \qquad (2.1)$$

where $E[X_i] = \int x_i dF(x_i)$ and $F(x_i)$ is the probability distribution for X_i with respect to an appropriate dominating measure λ. Note that F is typically referred to as the cumulative distribution function for the random variable X_i. The covariance matrix

for X will be denoted by

$$\Sigma \equiv E[(X - \mu)(X - \mu)^T] \tag{2.2}$$

$$= \begin{pmatrix} \text{Var}[X_1] & \text{Cov}(X_1, X_2) & \cdots & \text{Cov}(X_1, X_k) \\ \text{Cov}(X_1, X_2) & \text{Var}[X_2] & \cdots & \text{Cov}(X_2, X_k) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_1, X_k) & \text{Cov}(X_2, X_k) & \cdots & \text{Var}[X_k] \end{pmatrix} \tag{2.3}$$

where $\text{Var}[X_i] \equiv E[(X_i - \mu_i)^2]$ and $\text{Cov}(X_i, X_j) \equiv E[(X_i - \mu_i)(X_j - \mu_j)]$. The corresponding formula for the covariance of two random vectors X and Y with corresponding means μ_x and μ_y is

$$\text{Cov}(X, Y) = E[(X - \mu_x)(Y - \mu_y)^T] \tag{2.4}$$

Many of the datasets used in the examples will be modeled using multivariate normal distributions. Following Ferguson [31], the multivariate normal distribution is defined in Definition 2.1 below. The multivariate normal distribution for random vectors is the analog to the univariate normal distribution [31].

> **Definition 2.1:** The k-variate random column vector $X = \begin{pmatrix} X_1 & X_2 & \cdots & X_k \end{pmatrix}^T$
> has k-variate normal distribution if, for all vectors $b \in \mathbb{R}^k$, then $U = b^T X$
> is either univariate normal or constant.

A k-variate normal vector X with mean vector μ and covariance matrix Σ will be denoted by $X \sim \mathcal{N}_k(\mu, \Sigma)$. It should be noted that using this definition, a multivariate normal random vector does not always have a proper probability density function defined. When the probability density function exists (i.e. Σ is positive definite), it is given by

$$f(x|\mu, \Sigma) = (2\pi)^{-r/2}|\Sigma|^{-1/2}e^{-1/2(x-\mu)^T\Sigma^{-1}(x-\mu)} \tag{2.5}$$

for $x \in \mathbb{R}^k$, where $|\Sigma|$ denotes the determinant of the covariance matrix.

Theorem 2.1 (The Cramer-Wold Device):
A random k-dimensional vector X is completely determined by its one-dimensional linear projections, $\alpha^T X$, for any given vector $\alpha \in \mathbb{R}^k$.

The Cramer-Wold device has been reproduced from Flury [33], and in combination with Definition 2.1 for multivariate normality results in the following equivalent definition of multivariate normal random vectors from Izenman [43]. The following definition will be useful in determining the distribution of vectors which are composed of other multivariate normal random vectors, which will be the case for certain types of forensic evidence considered as examples in the following Chapters.

Definition 2.2: The random k-dimensional vector X has a *k-variate normal distribution* if, and only if, every linear functional of X has a univariate normal distribution.

Consider a column vector $Y = \begin{pmatrix} Y_1^T & Y_2^T & \cdots & Y_n^T \end{pmatrix}^T$ in which each individual column vector Y_i composing this larger vector is distributed as a multivariate normal

$$Y_i \overset{iid}{\sim} \mathcal{N}_k(\mu, \Sigma)$$

for $i = 1, 2, \cdots, n$. Then under Definition 2.2, the larger vector Y will also have a multivariate normal distribution given by

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \sim \mathcal{N}_{kn} \left(\begin{pmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{pmatrix}, \begin{pmatrix} \Sigma & 0_{k \times k} & \cdots & 0_{k \times k} \\ 0_{k \times k} & \Sigma & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0_{k \times k} \\ 0_{k \times k} & \cdots & 0_{k \times k} & \Sigma \end{pmatrix} \right). \tag{2.6}$$

The derivation of this well-known result from Miller [49] is reproduced for clarity below.

Derivation (2.6): Since each of the Y_i is normally distributed, by Definition 2.1 and the Cramer-Wold Device, Y is also normally distributed since any linear projection of Y will be univariate normal. We just need to find

its mean vector and covariance matrix. Clearly, the expected value will be the kn-dimensional column vector

$$E[Y] = \begin{pmatrix} E[Y_1] \\ E[Y_2] \\ \vdots \\ E[Y_n] \end{pmatrix} = \begin{pmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{pmatrix}.$$

The $\text{Var}[Y_i] = \Sigma$, so we need to find $\text{Cov}(Y_i, Y_j)$.

$$
\begin{aligned}
\text{Cov}(Y_i, Y_j) &= E[(Y_i - \mu)(Y_j - \mu)^T] \\
&= E[Y_i Y_j^T - \mu Y_j^T - Y_i \mu^T + \mu \mu^T] \\
&= E[Y_i Y_j^T] - \mu E[Y_j^T] - E[Y_i]\mu^T + E[\mu \mu^T] \\
&= E[Y_i Y_j^T] - \mu \mu^T - \mu \mu^T + E[\mu \mu^T] \\
&= E[Y_i Y_j^T] - \mu \mu^T - \mu \mu^T + \mu \mu^T \\
&= E[Y_i Y_j^T] - \mu \mu^T \\
&= E[Y_i]\, E[Y_j^T] - \mu \mu^T \\
&= \mu \mu^T - \mu \mu^T = 0_{k \times k}
\end{aligned}
$$

Another typical model that will be used for defining the evidence in forensic science is the hierarchical simple random effects model as given by Miller [49]. Let y_{ij} denote the k-dimensional column vector of measurements on the jth component from the ith source for $j = 1, 2, \cdots, m_i$ and $i = 1, 2, ..., n$. Then the hierarchical *simple random effects model* is given by

$$y_{ij} = \mu_a + a_i + w_{ij}$$

where $a_i \overset{iid}{\sim} \mathcal{N}_k(0_k, \Sigma_a)$ and $w_{ij} \overset{iid}{\sim} \mathcal{N}_k(0_k, \Sigma_w)$ are independent from each other. The vector μ_a denotes the overall mean of the measurements from all of the sources, Σ_a denotes the between-source covariance matrix, and Σ_w denotes the within-source covariance matrix.

It will be useful to think about the evidence which follows the hierarchical simple random effects model as a combined multivariate normal random vector. The following

theorem (a corollary to the results provided in Miller [49]) describes the properties of this combined vector.

Theorem 2.2:

If y_{ij} follows a simple random effects model, then the km_i-dimensional column vector $Y_i = \begin{pmatrix} y_{i1}^T & y_{i2}^T & \cdots & y_{im_i}^T \end{pmatrix}^T$ for $i = 1, 2, \cdots, n$, has the following distribution:

$$Y_i \sim \mathcal{N}_{m_i}(\mu_c, \Sigma_c)$$

where $\Sigma_c = \begin{pmatrix} \Sigma_a + \Sigma_w & \Sigma_a & \cdots & \Sigma_a \\ \Sigma_a & \Sigma_a + \Sigma_w & \ddots & \vdots \\ \vdots & \ddots & \ddots & \Sigma_a \\ \Sigma_a & \cdots & \Sigma_a & \Sigma_a + \Sigma_w \end{pmatrix}$ *and* $\mu_c = \begin{pmatrix} \mu_a \\ \mu_a \\ \vdots \\ \mu_a \end{pmatrix}$.

Proof: First, consider $y_{ij} = \mu_a + a_i + w_{ij}$. Since y_{ij} is a linear combination of normally distributed random variables, then y_{ij} is also normally distributed by Definition 2.1. We need to find the mean vector and the covariance matrix.

$$E[y_{ij}] = E[\mu_a + a_i + w_{ij}] = \mu_a + E[a_i] + E[w_{ij}] = \mu_a + 0_k + 0_k = \mu_a$$

$$\mathrm{Var}[y_{ij}] = \mathrm{Var}[\mu_a + a_i + w_{ij}] = 0_{k \times k} + \mathrm{Var}[a_i] + \mathrm{Var}[w_{ij}] = \Sigma_a + \Sigma_w$$

For $Y_i = \begin{pmatrix} y_{i1}^T & y_{i2}^T & \cdots & y_{im_i}^T \end{pmatrix}^T$ for $i = 1, 2, \cdots, n$, since each of the y_{ij} is normally distributed, then by Definition 2.1, Y_i is also normally distributed since any linear projection of Y_i will be univariate normal. We just need to find its mean vector, μ_c, and covariance matrix, Σ_c.

$$\mu_c = E[Y_i] = \begin{pmatrix} E[y_{i1}] \\ E[y_{i2}] \\ \vdots \\ E[y_{im_i}] \end{pmatrix} = \begin{pmatrix} \mu_a \\ \mu_a \\ \vdots \\ \mu_a \end{pmatrix}$$

$$\begin{aligned} \mathrm{Cov}(y_{ij}, y_{ik}) &= E[(y_{ij} - \mu_a)(y_{ik} - \mu_a)^T] \\ &= E[(\mu_a + a_i + w_{ij} - \mu_a)(\mu_a + a_i + w_{ik} - \mu_a)^T] \\ &= E[(a_i + w_{ij})(a_i + w_{ik})^T] \end{aligned}$$

$$
\begin{aligned}
&= \quad E[a_i a_i^T + a_i w_{ik}^T + w_{ij} a_i^T + w_{ij} w_{ik}^T] \\
&= \quad E[a_i a_i^T] + E[a_i w_{ik}^T] + E[w_{ij} a_i^T] + E[w_{ij} w_{ik}^T] \\
&= \quad \text{Var}[a_i] + E[a_i] \ E[w_{ik}^T] + E[w_{ij}] \ E[a_i^T] + E[w_{ij}] \ E[w_{ik}^T] \\
&= \quad \Sigma_a + 0_k \cdot 0_k^T + 0_k \cdot 0_k^T + 0_k \cdot 0_k^T \\
&= \quad \Sigma_a
\end{aligned}
$$

Therefore, using Equation 2.2 we have that

$$
\Sigma_c = \begin{pmatrix}
\Sigma_a + \Sigma_w & \Sigma_a & \cdots & \Sigma_a \\
\Sigma_a & \Sigma_a + \Sigma_w & \ddots & \vdots \\
\vdots & \ddots & \ddots & \Sigma_a \\
\Sigma_a & \cdots & \Sigma_a & \Sigma_a + \Sigma_w
\end{pmatrix}.
$$

■

Another interesting result involving the multivariate normal distribution is its relationship to the Wishart distribution. The Wishart distribution is a distribution for random matrices and is the multivariate analog to the univarite Chi-Squared distribution [43]. The definition of a Wishart random matrix is reproduced from Izenman [43] below.

> **Definition 2.3:** Given n independent random vectors X_i where $X_i \sim \mathcal{N}_k(0, \Sigma)$ for $i = 1, 2, \ldots, n$ with $n \geq k$, the random positive-definite, symmetric, $k \times k$ matrix $W = \sum_{i=1}^{n} X_i X_i^T$ has a *Wishart distribution* with n degrees of freedom and associated center matrix Σ. This will be denoted by $W \sim \mathcal{W}_k(n, \Sigma)$.

It should be noted that if $n < k$ the probability density function for W does not exist. A matrix follows the inverse Wishart distribution if its inverse follows a Wishart distribution [4].

> **Definition 2.4:** Given $W \sim \mathcal{W}_r(n, \Sigma)$ and $B = W^{-1}$, then the $k \times k$ positive definite matrix B follows the *inverse Wishart distribution*, denoted

$B \sim \mathcal{W}_k^{-1}(n, \Sigma^{-1})$ where Σ^{-1} is a positive definite $k \times k$ matrix and $n \geq k$.

Similar to the Wishart distribution, if $n < k$ the probability density function for B does not exist. When the probability density function does exist, it is given by

$$f(b|n, \Sigma^{-1}) = \frac{|\Sigma^{-1}|^{n/2} \; |b|^{(n+k+1)/2} \; e^{-1/2 \; tr(\Sigma^{-1}b^{-1})}}{2^{nk/2} \; \Gamma_k(\frac{n}{2})} \tag{2.7}$$

for b a $k \times k$ positive-definite symmetric matrix, and where $\Gamma_k(\bullet)$ is the multivariate Gamma function

$$\Gamma_k(\bullet) = \pi^{k(k-1)/4} \prod_{i-1}^{k} \Gamma\left(\bullet + \frac{1-i}{2}\right). \tag{2.8}$$

Both the multivariate normal and inverse Wishart distributions will be useful as prior distributions (prior distributions will be defined in the following subsection). Other typical distributions used in forensic science include the multinomial and Dirichlet distributions. Details of these distributions can be found in Wasserman [78].

2.2 Bayesian Statistics

Currently in forensic science there are three different frameworks that are considered, the classical or Frequentist framework, the Bayesian framework, and the Likelihood framework. These distinct frameworks each use probability to characterize different types of uncertainty within the forensic identification of source problems. This section focuses on the Bayesian probabilistic framework since it is fully developed for the class of statistical problems in forensic science and frequently recommended [24, 51, 2]. Following the typical conventions of the Bayesian framework in forensic science, $\mathbb{P}$ will be used to denote the probability measure characterizing all types of uncertainty. The main tool of Bayesian statistics is Bayes Theorem. For events A and B, the

Bayes Rule states that

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\ \mathbb{P}(A)}{\mathbb{P}(B)} \qquad (2.9)$$

where $\mathbb{P}(A)$ denotes the probability of the event A occurring.

The following section will define a set of standard notation to be used within the Bayesian framework. The underlying probability space for the data, denoted $(\Omega, \mathcal{A}, P_0)$, consists of the sample space Ω, its corresponding σ-field $\mathcal{A}$, and the probability measure P_0. Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ denote an independent and identically distributed set of observations of a random vector $\mathbf{X}$ defined on this probability space. For this reason, P_0 is often referred to as the sampling distribution of $\mathbf{X}$. To be clear, the p-dimensional random vector $\mathbf{X}$ is a measurable function from $(\Omega, \mathcal{A}, P_0)$ into $(\mathbb{R}^p, \mathcal{B}, \lambda)$ where $\mathbb{R}^p$ represents the p-dimensional vectors of real numbers, $\mathcal{B}$ its corresponding σ-field, and λ is an appropriate dominating measure. We will denote the cumulative distribution function associated with $\mathbf{X}$ under P_0 as F_0 such that $F_0(x) = P_0(\omega \in \Omega : X(\omega) \leq x)$ for any $x \in \mathbb{R}$. Recall from Dudley [22] and Ash [5] that there is a one-to-one mapping between P_0 and F_0 by means of the Caratheodory Extension Theorem. When F_0 is absolutely continuous with respect to λ, then the corresponding probability density function exists, and is denoted f_0 [5]. Typically, the actual sampling distribution, P_0, of the data will be unknown (and correspondingly, F_0 is unknown).

Suppose that the unknown sampling distribution for the data is assumed to be in a class of probability measures, $\mathcal{P}$, such that the indexing parameter θ is an element of a finite dimensional vector space $\Theta \subseteq \mathbb{R}^p$, given by $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. Therefore, there exists a $\theta \in \Theta$ such that $P_\theta = P_0$. It may be preferable to work with the class of cumulative distribution functions corresponding to $\mathcal{P}$, denoted $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$ in certain situations. In many cases, $\mathcal{P}$ is only implicitly defined by explicit definition of $\mathcal{F}$. In most applications, $\mathcal{F}$ will be chosen so that the corresponding probability density functions exist. Let $f(\mathbf{x}|\theta)$ denote the probability density function, when it

exists, corresponding to the distribution function F_θ.

First, consider some basic definitions from Robert [59]. In the Bayesian paradigm, any uncertainty about the values for the parameters should be characterized using a measure of belief. A *belief* is a formal measure that follows the basic axioms of probability. There are two basic types of measures of belief, a prior belief and a posterior belief. In this case, since the parameter space Θ is a finite-dimensional Euclidean vector space, then any probability measure on Θ will be equivalent to the corresponding cumulative distribution function. Therefore, we will adopt the standard Bayesian abuse of notation and treat these two measures synonymously [59].

> **Definition 2.5:** The *prior distribution* for the parameter is the probability measure $\Pi(\theta)$ on the parameter space Θ, and when its corresponding density function exists it will be denoted $\pi(\theta)$.

The prior distribution is considered the belief about the value of the parameter before observing any data, and is often thought of as the unconditional distribution on the parameter space since it doesn't depend on the observation of any data [59]. On a similar note, F_θ is the conditional distribution on the sample space since it depends on the value of θ in the parameter space. The unconditional distribution on the sample space (which doesn't depend on θ) is often referred to as the marginal distribution for the data [59].

> **Definition 2.6:** When $f(\mathbf{x}|\theta)$ exists, the *marginal density* of $\mathbf{X}$ is
>
> $$f(\mathbf{x}) = \int_\Theta f(\mathbf{x}|\theta) \, d\Pi(\theta)$$
>
> corresponding to the *marginal distribution* of $\mathbf{X}$ denoted F (with corresponding probability measure P_x).

Another name for the marginal distribution is the prior predictive distribution. This

name indicates more explicitly that the unconditional distribution for the data is implicitly dependent on the prior distribution for the indexing parameter. If the prior distribution for the parameter were to change, then so would the marginal distribution for the data.

Next, the posterior measure on the parameter space can be explicitly defined using Bayes Rule [59]. Recall from Equation 2.9 that the conditional probability of an event A given an event B is the conditional probability of B given A times the probability of A divided by the probability of B. In a similar fashion, the posterior measure of θ given an observation of the data is defined in terms of the probability measure for the data, the prior measure for the parameter, and the marginal measure for the data.

> **Definition 2.7:** The *posterior distribution* for the parameter is the probability measure $\Pi(\theta)$ on the parameter space Θ conditional on the observation $\mathbf{x}$ given by
> $$\Pi(\theta|\mathbf{x}) = \frac{F_\theta(\mathbf{x}) \ \Pi(\theta)}{F(\mathbf{x})}.$$
> When the densities $f(\mathbf{x}|\theta)$, $\pi(\theta)$, and $f(\mathbf{x})$ exist, then the *posterior density* of θ given $\mathbf{x}$ is
> $$\pi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta) \ \pi(\theta)}{f(\mathbf{x})}.$$

The posterior distribution is often thought of as the updated belief about the value of the parameter after the observation of the data [59]. Alternatively, it can be considered the conditional distribution on the parameter space given the data that has been observed. Another conditional distribution concerning the data is the posterior predictive distribution. The posterior predictive distribution can be thought of as the conditional distribution of potentially unobserved data given the observed data.

> **Definition 2.8:** Suppose that $\mathbf{x}$ and $\mathbf{y}$ are independent observations of

a random variable $\mathbf{X}$ with respect to the sampling distribution F_θ, then the *posterior predictive density* of $\mathbf{y}$ given $\mathbf{x}$ is

$$f(\mathbf{y}|\mathbf{x}) = \int_\Theta f(\mathbf{y}|\theta) \, d\Pi(\theta|\mathbf{x}).$$

Analogous to the dependence of the prior predictive distribution on the prior distribution, the posterior predictive distribution depends on the posterior distribution for the parameters (and therefore, the prior distribution for the parameters, as well). Again, if the prior distribution for the parameter were to change, then the posterior predictive distribution for the unobserved data given the observed data would also change. The posterior predictive density can take the alternate form in terms of the prior density for the parameter θ as well as the sampling densities for both $\mathbf{x}$ and $\mathbf{y}$ [59].

$$f(\mathbf{y}|\mathbf{x}) = \frac{\int_\Theta f(\mathbf{y}|\theta)f(\mathbf{x}|\theta) \, d\Pi(\theta)}{\int_\Theta f(\mathbf{x}|\theta) \, d\Pi(\theta)} \tag{2.10}$$

2.3 Asymptotic Statistics

A majority of the results in this section will be based upon different concepts of convergence. A number of good references for these different forms of convergence include van der Vaart [75] and Serfling [65]. The definition of the three most commonly used forms of convergence are provided below. The strongest of these convergences is most familiar to people and is called *almost sure convergence*.

Definition 2.9: A sequence of random vectors X_n defined on $(\Omega, \mathcal{A}, P)$ is said to *converge almost surely* to the random vector X defined on $(\Omega, \mathcal{A}, P)$ if as $n \to \infty$

$$P(\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)) = 1 \quad \text{or} \quad P(\omega : \lim_{n\to\infty} ||X_n(\omega) - X(\omega)|| = 0) = 1$$

and will be denoted

$$X_n \xrightarrow{as} X.$$

Almost sure convergence can also be referred to as *convergence with probability one*, *strong convergence*, or *convergence almost everywhere*. One of the most useful theorems involving almost sure convergence is called the Strong Law of Large Numbers (SLLN). It says that the average of a set of random vectors will converge almost surely to its expectation as the number of random vectors in the set increases without bound. The result, reproduced from Serfling [65], is formalized below.

Theorem 2.3 (Strong Law of Large Numbers):
Let $X_1, X_2, \ldots, X_n$ be an independent and identically distributed random variables with expected value $E[X_i]$. Then as $n \to \infty$

$$\overline{X}_n \equiv \frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{as} E[X_i]$$

holds if and only if $E[X_i]$ is a finite constant.

More general statements of the SLLN which relax the assumptions of independence and identical distributions can be found in Serfling [65] among others, but will not typically be used for the results in this dissertation. The SLLN will be used to prove consistency properties of Monte Carlo integration methods in Section 2.4. The next form of convergence is called *convergence in probability*.

Definition 2.10: A sequence of random vectors X_n defined on $(\Omega, \mathcal{A}, P)$ is said to *converge in probability* to the random vector X defined on $(\Omega, \mathcal{A}, P)$ if as $n \to \infty$

$$\lim_{n \to \infty} P(\omega : ||X_n(\omega) - X(\omega)|| > \varepsilon) = 0, \quad \forall \, \varepsilon > 0$$

or

$$\lim_{n \to \infty} P(\omega : ||X_n(\omega) - X(\omega)|| < \varepsilon) = 1, \quad \forall \, \varepsilon > 0$$

and it will be denoted by

$$X_n \xrightarrow{P} X.$$

A similar result to the SLLN will be presented using convergence in distribution. It is called the Weak Law of Large Numbers (WLLN), and will be formalized in the following theorem [65].

Theorem 2.4 (Weak Law of Large Numbers):
Let $X_1, X_2, \ldots, X_n$ be an independent and identically distributed random variables with expected value $E[X_i]$. Then as $n \to \infty$

$$\overline{X}_n \equiv \frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{P} E[X_i]$$

holds if and only if $E[X_i]$ is a finite constant.

More general statements of the WLLN which relax the assumptions of independence and identical distributions can be found in Serfling [65] among others, but will not be used for the results in this dissertation.

It will be useful for the proofs of the theorems provided in the remainder of the dissertation to provide a short-hand notation for certain sequences with specific convergence in probability properties. Following the notation of van der Vaart [75] and Serfling [65], denote a sequence which converges in probability (with respect to the probability measure P) to zero by $o_P(1)$ and a sequence which is bounded in probability (with respect to the probability measure P) by $O_P(1)$. For example, let U_n and V_n be sequences of random variables, then as $n \to \infty$

$$U_n = o_P(1) \quad \text{if and only if} \quad U_n \xrightarrow{P} 0$$

and

$$U_n = o_P(V_n) \quad \text{if and only if} \quad U_n/V_n \xrightarrow{P} 0.$$

Similarly, for every $\varepsilon > 0$, there exists $0 \le M \le \infty$ and $0 \le N \le \infty$ such that

$$U_n = O_P(1) \quad \text{if and only if} \quad P(|U_n| > M) < \varepsilon, \quad \forall n > N$$

and

$$U_n = O_P(V_n) \quad \text{if and only if} \quad P(|U_n/V_n| > M) < \varepsilon, \quad \forall n > N.$$

It should be noted that this notation for stochastic, or probabilistic, convergence is difference from the 'little-o' and 'big-O' notation for functional orders of magnitude, see Serfling [65].

Next, the weakest form of convergence under consideration is *convergence in distribution.*

Definition 2.11: A sequence of random vectors X_n defined on $(\Omega_n, \mathcal{A}_n, P_n)$ is said to *converge in distribution* to the random vector X defined on $(\Omega_0, \mathcal{A}_0, P_0)$ if as $n \to \infty$

$$\lim_{n \to \infty} P_n(X_n < x) = P_0(X < x), \quad \forall \text{ continuity points } x \text{ under } P_0$$

and, if $X_1, \ldots, X_n,$ and X have corresponding distribution functions $F_1, \ldots, F_n,$ and F

$$\lim_{n \to \infty} F_n(x) = F(x), \quad \forall \text{ continuity points } x \text{ of } F,$$

and it is denoted by

$$X_n \overset{d}{\longrightarrow} X \quad \text{or} \quad X_n \rightsquigarrow X.$$

Convergence in distribution can also be referred to as *weak convergence* or *convergence in law.* There are many equivalent definitions of convergence in distribution which are given in the Portmanteau Theorem of van der Vaart [75] and reproduced in the following theorem.

Theorem 2.5 (Portmanteau):

For any random vectors X_n and X on $(\Omega, \mathcal{A}, P)$, the following statements are equivalent:

1. $P(X_n \leq x) \to P(X \leq x)$ $\quad \forall$ *continuity points* x *of* $x \mapsto P(X \leq x)$;

2. $E[f(X_n)] \to E[f(X)]$ $\quad \forall$ *bounded continuous functions* f;

3. $E[f(X_n)] \to E[f(X)]$ $\quad \forall$ *bounded Lipschitz functions* f;

4. $\liminf E[f(X_n)] \geq E[f(X)]$ $\quad \forall$ *non-negative continuous functions* f;

5. $\liminf P(X_n \in G) \geq P(X \in G)$ $\quad \forall$ *open set* G;

6. $\limsup P(X_n \in F) \leq P(X \in F)$ $\quad \forall$ *closed set* F;

7. $P(X_n \in B) \to P(X \in B)$ $\quad \forall$ *Borel set* B *such that* $P(X \in \delta B) = 0$ *where* δB *is the boundary of* B.

For more general versions of the Portmanteau Theorem using outer- and inner-probability, please see van der Vaart and Wellner [76]. If a sequence converges in distribution, then that sequence is bounded in probability [65]. One of the most important results involving convergence in distribution is the Central Limit Theorem (CLT). The CLT provides the conditions under which a sequence of random vectors has an approximately normal distribution. The formal result is reproduced from Serfling [65] below.

Theorem 2.6 (Central Limit Theorem):
Let $X_1, X_2, \ldots, X_n$ be an independent and identically distributed random sample with finite expected value μ and covariance matrix Σ. Then as $n \to \infty$

$$\sqrt{n}\left(\overline{X}_n - \mu\right) \rightsquigarrow \mathcal{N}_k(0, \Sigma).$$

Alternatively, this may be denoted $\overline{X}_n = \mathcal{AN}_k\left(\mu, \frac{1}{n}\Sigma\right)$.

More general versions of the CLT are provided in Serfling [65]. The relationship between the three types of convergence can be illustrated in the following result from van der Vaart [75]:

Theorem 2.7 (Relationship between Modes of Convergence):

Let X_n and X be random vectors, then as $n \to \infty$

1. *$X_n \xrightarrow{as} X$ implies $X_n \xrightarrow{P} X$*

2. *$X_n \xrightarrow{P} X$ implies $X_n \rightsquigarrow X$*

The next two results (reproduced from van der Vaart [75]) apply regardless of which type of convergence is being used. These theorems play a large role in proving many major and minor results throughout the remainder of the dissertation. The first result, the Continuous Mapping Theorem (CMT), allows the convergence properties of a sequence of random vectors to extend to continuous functions of the random vectors. The next result, Slutsky's Theorem, considers the convergence properties of a sequence of random vectors resulting from combining two different converging sequences of random vectors. Slutsky's Theorem can be viewed as a direct result of the CMT.

Theorem 2.8 (Continuous Mapping Theorem):

Let $g : \mathbb{R}^k \mapsto \mathbb{R}^m$ be a continuous function at every point of a set C such that $P(X \in C) = 1$ and let X_n and X be random vectors. If $X_n \to X$, then as $n \to \infty$

$$g(X_n) \to g(X)$$

where the convergence can be in any manner (almost sure, in probability, or in distribution).

Theorem 2.9 (Slutsky's Theorem):

Let X_n, X, and Y_n be random variables and let c be a constant. If $X_n \to X$ and $Y_n \to c$, then as $n \to \infty$

1. *$X_n + Y_n \to X + c$*

2. *$Y_n X_n \to cX$*

3. *$Y_n^{-1} X_n \to c^{-1} X$, provided that $c \neq 0$*

where the convergence can be in any manner (almost sure, in probability, or in distribution).

The next two results provide the conditions under which a sequence of (Bayesian) posterior probability measures converges to a fixed measure. These results are important since, under the Bayesian setting, the recommended practice is to base all inferences about unknown parameters on the posterior distribution for that parameter. The first result which stems from Doob in 1948 [75] is a completely Bayesian result, whereas the second result is a mixture of Bayesian and Frequentist ideas. The following statement of Doob's Consistency Theorem is from Ghosal and van der Vaart [37] and van der Vaart [75]. Doob's Consistency Theorem will be used in Chapter 5 to show, in most settings, that a Bayesian believes that an analog to the Frequentist likelihood ratio exists.

First, some notational concepts will be presented to facilitate the understanding of the theorem assumptions. Let $(\Theta, \mathcal{C}, \Pi_0)$ denote the parameter space, where Π_0 is referred to as the prior probability measure on Θ. Let $x_1, x_2, \ldots, x_n$ be an independent and identically distributed sequence of observations of a random variable X on the probability space $(\mathcal{X}, \mathcal{A}, P_\theta)$, where P_θ is the conditional measure on $\mathcal{X}$ given $\theta \in \Theta$. The condition that that $P_\theta \neq P_{\theta'}$ whenever $\theta \neq \theta'$ means that the probability measure P_θ is identifiable. Then, let Π_n denote the sequence of posterior measures on $(\Theta, \mathcal{C})$ given the observation of the sequence $x_1, x_2, \ldots, x_n$ for $n = 1, 2, \ldots, \infty$. For a sequence of posterior measures to be strongly consistent under θ, it means that the posterior measure converges in distribution to a measure δ_θ which is degenerate at θ for P_θ^∞-almost every sequence $x_1, x_2, \ldots \equiv x_\infty$ [75], where P_θ^∞ is the limiting joint probability measure on the product space $(\mathcal{X}^\infty, \mathcal{A}^\infty)$ where $\mathcal{X}^\infty$ denotes the infinite Cartesian product of the sample space and $\mathcal{A}^\infty$ the infinite Cartesian product of the corresponding σ-fields. In terms of the probability measures, strong consistency of posterior measures means that

$$P_\theta^\infty(x_\infty : \Pi_n(\theta_n | x_1, x_2, \ldots, x_n) \rightsquigarrow \delta_\theta) = 1. \tag{2.11}$$

Theorem 2.10 (Doob's Consistency Theorem):
Suppose that the sample space $(\mathcal{X}, \mathcal{A})$ is a subset of Euclidean space with its Borel sigma-field. Suppose that the random vectors $X_1, \cdots, X_n$ are independent and identically distributed according to the probability measure P_θ, and that $P_\theta \neq P_{\theta'}$ whenever $\theta \neq \theta'$. Then for every prior probability measure Π on Θ the sequence of posterior measures is strongly consistent for Π-almost every θ.

It is important to note that the Doob's Consistency Theorem is a fully Bayesian result since the parameter θ has an associated distribution characterizing the probability for each possible value this parameter could take. In addition, the degenerate measure to which the posterior distribution of θ converges is considered a random measure with respect to the prior distribution on θ. In the Frequentist paradigm, the parameter θ takes the fixed, non-random value θ_0, and it is the distribution given this fixed value of θ which generated the observed data. This is the view taken for the statement of the Bernstein-von Mises Theorem below which is restated from van der Vaart [75]. The Bernstein-von Mises Theorem will be used in Chapter 6 and Chapter 8 to motivate point-based and interval-based estimations of the value of forensic evidence.

Theorem 2.11 (Bernstein-von Mises Theorem):
Let the experiment $(P_\theta : \theta \in \Theta)$ be differentiable in quadratic mean at θ_0 with nonsingular Fisher information matrix I_{θ_0}, and suppose that for every $\varepsilon > 0$ there exists a sequence of tests ϕ_n such that

$$P_{\theta_0}^n(\phi_n) \to 0 \quad and \quad \sup_{||\theta - \theta_0|| \geq \varepsilon} P_\theta^n(1 - \phi_n) \to 0.$$

Furthermore, let the prior measure be absolutely continuous in a neighborhood of θ_0 with a continuous positive density at θ_0. Then the corresponding posterior distributions satisfy

$$||P_{\sqrt{n}(\overline{\Theta}_n - \theta_0)|X_1, \cdots, X_n} - N(\Delta_{n,\theta_0}, I_{\theta_0}^{-1})|| \xrightarrow{P_{\theta_0}^n} 0.$$

Please refer to van der Vaart [75] for a detailed proof of this theorem.

First, the assumptions of the theorem and the notation are discussed. In the context above, the experiment notation $(P_\theta : \theta \in \Theta)$ means that P is a collection of probability measures on the space $(\mathcal{X}, \mathcal{A})$ indexed by the parameter θ which ranges over $\Theta \subseteq \mathbb{R}^k$, and that P_θ has the corresponding density function p_θ with respect to the measure μ. It is assumed that $X_1, \ldots, X_n$ is a random sample from the distribution P_{θ_0}. An experiment $(P_\theta : \theta \in \Theta)$ is *differentiable in quadratic mean* if there exists a measurable vector-valued function $\dot{\ell}_{\theta_0}$ such that

$$\int \left[\sqrt{p_\theta} - \sqrt{p_{\theta_0}} - \frac{1}{2}(\theta - \theta_0)^T \dot{\ell}_{\theta_0} \sqrt{p_{\theta_0}} \right]^2 d\mu = o(||\theta - \theta_0||^2)$$

as $\theta \to \theta_0$ [75]. Please see Lemma 7.6 of van der Vaart [75] for conditions under which an experiment is differentiable in quadratic mean. The next set of assumptions means that there exists a sequence of estimators of θ_0 (denoted ϕ_n) that is uniformly consistent on Θ. For a detailed description of this assumption and for conditions under which this assumption is met, please see Section 10 of van der Vaart, especially Lemmas 10.4 and 10.6 [75]. Finally, the last assumption is that the prior measure is absolutely continuous in a neighborhood of the true value of the parameter. A measure μ is absolutely continuous with respect to a measure λ if $\lambda(A) = 0$ implies that $\mu(A) = 0$ for every measurable set A [75].

Next, the final results of the theorem and the corresponding notation are discussed to facilitate further understanding. The notation $P_{\sqrt{n}(\overline{\Theta}_n - \theta_0)|X_1,\cdots,X_n}$ refers to the posterior distribution of the rescaled parameter $\sqrt{n}(\overline{\Theta}_n - \theta_0)$ given $X_1, \cdots, X_n$ and $N(\Delta_{n,\theta_0}, I_{\theta_0}^{-1})$ is the distribution of a normal random variable with mean Δ_{n,θ_0} and covariance matrix $I_{\theta_0}^{-1}$, where Δ_{n,θ_0} is the locally sufficient statistic [75] for the rescaled version of the parameter and $I_{\theta_0}^{-1}$ is the inverse Fisher's information matrix. It should be noted that Δ_{n,θ_0} can be replaced by any consistent estimator for the rescaled

parameter [75]. There is also an equivalent result to the Bernstein von-Mises theorem given in van der Vaart which considers the posterior distribution of the unscaled parameter. That result is given by

$$\left|\left| \Pi_n \left(\overline{\Theta}_n | X_1, \cdots, X_n\right) - \mathcal{N}\left(\hat{\theta}_n, \tfrac{1}{n}I_{\theta_0}^{-1}\right) \right|\right|_{TV} \xrightarrow{P_{\theta_0}^n} 0. \tag{2.12}$$

Finally, the metric $||\cdot||_{TV}$ denotes the total variation norm (see van der Vaart [75] or van de Geer [73]). For two measures P and Q on $(\mathcal{X}, \mathcal{A})$, the total variation distance between P and Q is given by $||P - Q||_{TV} = \sup_{A \in \mathcal{A}} |P(A) - Q(A)|$.

While the previous two theorems consider convergence results related to Bayesian posterior distributions, the following two theorems consider asymptotic results related to M-estimation. Using methods based on M-estimators, such as maximum likelihood estimates, is an alternative for inferences concerning unknown parameters (typically within the Frequentist scope of statistics). The first result considers situations in which the M-estimator will be consistent, and the second considers asymptotic normality results for consistent M-estimators. These two results will be used in Chapter 7 to motivate a prior-free approximate solution to the forensic identification of source problem. The first theorem is restated from Theorem 5.7 of van der Vaart [75] for clarity.

Theorem 2.12 (Consistency of M-Estimators):
Let M_n be random functions and let M be a fixed function of θ such that

$$\text{Assumption 1:} \quad \sup_{\theta : d(\theta, \theta_0) \geq \epsilon} M(\theta) \leq M(\theta_0), \quad \forall \varepsilon > 0;$$

$$\text{Assumption 2:} \quad \sup_{\theta \in \Theta} \left| M_n(\theta) - M(\theta) \right| \xrightarrow{P} 0;$$

$$\text{Assumption 3:} \quad M_n(\hat{\theta}_n) \geq M_n(\theta_0) - o_p(1).$$

Then $\hat{\theta}_n \xrightarrow{P} \theta_0$ (i.e. $\hat{\theta}_n$ is consistent).

Please refer to van der Vaart [75] for a discussion of the assumptions and detailed proof of this theorem. Or, for a more general version of this theorem, please see Theorem 3.2.3 from van der Vaart and Wellner [76]. It is important to note that the Consistency of M-Estimators theorem is a fully Frequentist result. The next theorem is also a fully Frequentist result which has been reproduced from Theorem 5.23 of van der Vaart [75] for clarity.

Theorem 2.13 (The Linearization of M-Estimators):

For each θ in an open subset of Euclidean space let $x \mapsto m_\theta(x)$ be a measurable functions such that $\theta \mapsto m_\theta(x)$ is differentiable at θ_0 for P-almost every x with derivative $\dot{m}_{\theta_0}(x)$ and such that for every θ_1 and θ_2 in a neighborhood of θ_0 and a measurable function $\dot{m}$ with $P\dot{m}^2 < \infty$

$$|m_{\theta_1}(x) - m_{\theta_2}(x)| \leq \dot{m}(x)||\theta_1 - \theta_2||.$$

Furthermore, assume that the map $\theta \mapsto Pm_\theta$ admits a second-order Taylor expansion at a point of maximum θ_0 with nonsingular symmetric second derivative matrix W_{θ_0}. If $P_n m_{\hat{\theta}_n} \geq \sup_\theta P_n m_\theta - o_p(n^{-1})$ and $\hat{\theta}_n \xrightarrow{P} \theta_0$, then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -W_{\theta_0}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{m}_{\theta_0}(X_i) + o_p(1).$$

In particular, the sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically normal with mean zero and covariance matrix $W_{\theta_0}^{-1} P\dot{m}_{\theta_0}\dot{m}_{\theta_0}^T W_{\theta_0}^{-1}$.

Please refer to van der Vaart [75] for a detailed proof of this theorem. Or, for a more general version of this theorem, please see Theorem 3.2.16 from van der Vaart and Wellner [76].

2.4 Monte Carlo Integration

Monte Carlo integration is a method which numerically approximates an integral quantity which is computationally intractable (typically because the integral does not

have a closed form solution). For a comprehensive review of Monte Carlo numerical approximation methods, see Han and Carlin [41]. This section will review three different Monte Carlo integration techniques discussed in Kass and Raftery [44] for estimating integral quantities of the form

$$h(x) = \int f(x|\theta) \, g(\theta) \, d\theta,$$

where f is the likelihood function (sampling density) for the data, x, indexed by the parameter, θ, and g is a density for θ. Let n denote the Monte Carlo sample size. The first technique, called the *Arithmetic Mean* estimate, is defined by

$$\hat{h}_1(x) = \frac{1}{n} \sum_{i=1}^{n} f(x|\theta^{(i)}) \qquad (2.13)$$

where $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(n)}$ is an independent sample of size n drawn from the distribution corresponding to the density $g(\theta)$. This is the most computationally simple form of Monte Carlo integration that will be considered. The following Lemma formalizes the result from Geweke [36] that $\hat{h}_1(x)$ is a strongly consistent estimate for $h(x)$.

Lemma 2.14:
Let $\hat{h}_1(x)$ denote the Arithmetic Mean estimate of $h(x)$. Then as $n \to \infty$,

$$\hat{h}_1(x) \xrightarrow{as} h(x).$$

Proof: The proof of this Lemma is reproduced from Geweke [36] for clarity. By the Strong Law of Large Numbers

$$\hat{h}_1(x) = \frac{1}{n} \sum_{i=1}^{n} f(x|\theta^{(i)}) \xrightarrow{as} E_g[f(x|\theta)]$$

as $n \to \infty$. The expectation is taken with respect to the distribution from which the

$\{\theta^{(i)}\}$ are sampled with corresponding density $g(\theta)$. Therefore,

$$E_g[f(x|\theta)] = \int f(x|\theta)g(\theta)\ d\theta = h(x).$$

Therefore, $\hat{h}_1(x)$ is a strongly consistent estimate for $h(x)$ since

$$\hat{h}_1(x) \xrightarrow{as} h(x).$$

$\blacksquare$

The approximate Monte Carlo Standard Error (MCSE) for the estimate, $\hat{h}_1(x)$, gives an approximation to the numerical precision of the Monte Carlo integration estimates [70]. The approximate MCSE for $\hat{h}_1(x)$ is defined by

$$\epsilon_1 = \sqrt{\frac{\sum_{i=1}^{n}[f(x|\theta^{(i)}) - \hat{h}_1(x)]^2}{n(n-1)}}. \tag{2.14}$$

The derivation of this form for the approximate MCSE is a well-known result which is reproduced from Geweke [36] for clarity below.

Derivation (2.14): By the Central Limit Theorem,

$$\sqrt{n}\left[\hat{h}_1(x) - h(x)\right] \rightsquigarrow \mathcal{N}(0, \sigma^2).$$

Therefore, the approximate MCSE is an unbiased estimate of $\sigma/\sqrt{n}$. It is well-known that the sample variance, S^2 is an unbiased estimate of σ^2. The sample variance for this problem is given by

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left[f(x|\theta^{(i)}) - \hat{h}_1(x)\right]^2.$$

Therefore, the approximate MCSE is the unbiased estimate of $\sigma/\sqrt{n}$ given by

$$\epsilon_1 = \sqrt{\frac{\sum_{i=1}^{n}[f(x|\theta^{(i)}) - \hat{h}_1(x)]^2}{n(n-1)}}.$$

□

The second and third techniques are specific versions of importance sampling which takes the form

$$\hat{h}(x) = \frac{\sum_{i=1}^{n} w_i \, f(x|\theta^{(i)})}{\sum_{i=1}^{n} w_i} \tag{2.15}$$

with $w_i = g(\theta^{(i)})/I(\theta^{(i)})$ where $I(\theta)$ is the importance sampling function from which the $\theta^{(i)}$'s are drawn. The following lemma shows that the importance sampling estimates are strongly consistent.

Lemma 2.15:

Let $\hat{h}(x)$ denote the importance sampling estimate of $h(x)$. Then

$$\hat{h}(x) \xrightarrow{as} h(x), \quad as \ n \to \infty.$$

Proof: The proof is reproduced from Geweke [36] for clarity. By the Strong Law of Large Numbers, the following two results hold:

$$\frac{1}{n} \sum_{i=1}^{n} w_i \, f(x|\theta^{(i)}) \xrightarrow{as} E_I \left[\frac{f(x|\theta)g(\theta)}{I(\theta)} \right]$$

and

$$\frac{1}{n} \sum_{i=1}^{n} w_i \xrightarrow{as} E_I \left[\frac{g(\theta)}{I(\theta)} \right].$$

Since the expectations are taken with respect to $I(\theta)$, then

$$\begin{aligned} E_I \left[\frac{f(x|\theta)g(\theta)}{I(\theta)} \right] &= \int \frac{f(x|\theta)g(\theta)}{I(\theta)} I(\theta)d\theta \\ &= \int f(x|\theta)g(\theta)d\theta \\ &= h(x) \end{aligned}$$

and

$$E_I \left[\frac{g(\theta)}{I(\theta)} \right] = \int \frac{g(\theta)}{I(\theta)} I(\theta) d\theta$$
$$= \int g(\theta) d\theta$$
$$= 1.$$

By Slutsky's Lemma, we have that

$$\hat{h}(x) = \frac{\frac{1}{n} \sum_{i=1}^{n} w_i \ f(x|\theta^{(i)})}{\frac{1}{n} \sum_{i=1}^{n} w_i} \xrightarrow{as} \frac{E_I[f(x|\theta)g(\theta)/I(\theta)]}{E_I[g(\theta)/I(\theta)]} = h(x)$$

which proves the strong consistency of $\hat{h}(x)$ as an estimate of $h(x)$ as $n \to \infty$. ■

The general formula for the approximate MCSE of importance sampling estimates from Tanner [70] is

$$\epsilon = \frac{\sqrt{\sum_{i=1}^{n} [f(x|\theta^{(i)}) - \hat{m}(x)]^2 w_i^2}}{\sum_{i=1}^{n} w_i}. \tag{2.16}$$

The details of the derivation for this result is provided in Geweke [36]. Due to the length of this derivation, it will omitted.

For the second technique, called the *Harmonic Mean* estimate, the importance sampling function is the posterior density for θ, $I_2(\theta) = \pi(\theta|x) = f(x|\theta)\pi(\theta)/m(x)$ when g is the prior density for θ, denoted $\pi(\theta)$. (First, note that in this case, a prior density can be any density for θ which does not depend on the data, x, but may depend on the observation of data distinct from x. A posterior is the resulting density for θ with respect to the given prior density which does depend on the observation of the data, x. Secondly, note that when g is not the prior density for θ, the importance sampling function may take a different form, or the Harmonic Mean estimate may not exist.) After some simplification from substituting $I_2(\theta)$ into Equation 2.15, the resulting

estimate is defined by

$$\hat{h}_2(x) = \left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{f(x|\theta^{(i)})}\right]^{-1} \tag{2.17}$$

where $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(n)}$ is an independent sample of size n drawn from $I_2(\theta) = \pi(\theta|x)$. Since the Harmonic Mean estimate is an importance sampling estimate, then it will be strongly consistent as well. This result is formalized in the following lemma.

Lemma 2.16:

Let $\hat{h}_2(x)$ denote the Harmonic Mean estimate of $h(x)$. Then as $n \to \infty$,

$$\hat{h}_2(x) \xrightarrow{as} h(x).$$

Proof: By the Strong Law of Large Numbers

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{f(x|\theta^{(i)})} \xrightarrow{as} E_{I_2}\left[\frac{1}{f(x|\theta)}\right].$$

Because the expectation is taken with respect to the distribution from which the $\{\theta^{(i)}\}$ are sampled with corresponding density $I_2(\theta) = \pi(\theta|x)$, then

$$\begin{aligned}
E_{I_2}\left[\frac{1}{f(x|\theta)}\right] &= \int \frac{1}{f(x|\theta)}\pi(\theta|x)d\theta \\
&= \int \frac{1}{f(x|\theta)}\frac{f(x|\theta)\pi(\theta)}{f(x)}d\theta \\
&= \int \frac{\pi(\theta)}{f(x)}d\theta \\
&= \frac{1}{f(x)}\int \pi(\theta)d\theta \\
&= \frac{1}{f(x)} \\
&= \frac{1}{h(x)}.
\end{aligned}$$

Since $\pi(\theta)$ is a proper density function for θ, then it integrates to one. Also, for the Harmonic Mean estimate $g(\theta) = \pi(\theta)$ which means that $f(x) = \int f(x|\theta)\pi(\theta)d\theta =$

$h(x)$. Now, since we have that

$$\frac{1}{n} \sum_{i=1}^{n} \frac{1}{f(x|\theta^{(i)})} \xrightarrow{as} \frac{1}{h(x)}$$

as $n \to \infty$, then by the Continuous Mapping Theorem

$$\hat{h}_2(x) = \left[\frac{1}{n} \sum_{i=1}^{n} \frac{1}{f(x|\theta^{(i)})} \right]^{-1} \xrightarrow{as} h(x).$$

Therefore, $\hat{h}_2(x)$ is a strongly consistent estimate for $h(x)$. ∎

The approximate MCSE for $\hat{h}_2(x)$ can be derived by substituting the corresponding weights $w_i = h(x)/f(x|\theta)$ into Equation 2.16 above. The resulting formula for the approximate MCSE is given by

$$\epsilon_2 = \frac{\sqrt{\sum_{i=1}^{n} \left[1 - \frac{\hat{h}_2(x)}{f(x|\theta^{(i)})} \right]^2}}{\sum_{i=1}^{n} f(x|\theta^{(i)})^{-1}}. \tag{2.18}$$

The full derivation of Equation 2.18 will now be reproduced for clarity.

Derivation (2.18): For the Harmonic Mean estimate, the importance sampling function is $I(\theta) = f(x|\theta)\pi(\theta)/h(x)$ [44]. Therefore,

$$
\begin{aligned}
w_i &= \pi(\theta^{(i)})/I(\theta^{(i)}) \\
&= \frac{\pi(\theta^{(i)})}{f(x|\theta^{(i)})\pi(\theta^{(i)})/h(x)} \\
&= \frac{\pi(\theta^{(i)})h(x)}{f(x|\theta^{(i)})\pi(\theta^{(i)})} \\
&= \frac{h(x)}{f(x|\theta^{(i)})}.
\end{aligned}
$$

Substituting w_i into the Equation 2.16 gives:

$$
\begin{aligned}
\epsilon_2 &= \frac{\sqrt{\sum_{i=1}^{n}[f(x|\theta^{(i)}) - \hat{h}(x)]^2 w_i^2}}{\sum_{i=1}^{n} w_i} \\[2ex]
&= \frac{\sqrt{\sum_{i=1}^{n}\left[f(x|\theta^{(i)}) - \hat{h}(x)\right]^2 \left[\dfrac{h(x)}{f(x|\theta^{(i)})}\right]^2}}{\sum_{i=1}^{n}\dfrac{h(x)}{f(x|\theta^{(i)})}} \\[2ex]
&= \frac{h(x)\sqrt{\sum_{i=1}^{n}\left[f(x|\theta^{(i)}) - \hat{h}(x)\right]^2 \left[\dfrac{1}{f(x|\theta^{(i)})}\right]^2}}{h(x)\sum_{i=1}^{n}\dfrac{1}{f(x|\theta^{(i)})}} \\[2ex]
&= \frac{\sqrt{\sum_{i=1}^{n}\left[f(x|\theta^{(i)}) - \hat{h}(x)\right]^2 \left[\dfrac{1}{f(x|\theta^{(i)})}\right]^2}}{\sum_{i=1}^{n}\dfrac{1}{f(x|\theta^{(i)})}} \\[2ex]
&= \frac{\sqrt{\sum_{i=1}^{n}\left[\dfrac{f(x|\theta^{(i)}) - \hat{h}(x)}{f(x|\theta^{(i)})}\right]^2}}{\sum_{i=1}^{n} f(x|\theta^{(i)})^{-1}} \\[2ex]
&= \frac{\sqrt{\sum_{i=1}^{n}\left[1 - \dfrac{\hat{h}(x)}{f(x|\theta^{(i)})}\right]^2}}{\sum_{i=1}^{n} f(x|\theta^{(i)})^{-1}}.
\end{aligned}
$$

$\square$

The third technique has an importance sampling function which is a mixture of the prior and posterior distributions for θ denoted $I_3(\theta) = \delta\pi(\theta) + (1 - \delta)\pi(\theta|x)$ for the mixing proportion, $0 < \delta < 1$ when $g(\theta) = \pi(\theta)$. The resulting estimator, called the

Mixture estimate, is defined by

$$\hat{h}_3(x) = \frac{\sum_{i=1}^{n} w_i \, f(x|\theta^{(i)})}{\sum_{i=1}^{n} w_i} \tag{2.19}$$

with $w_i = \pi(\theta^{(i)})/I_3(\theta^{(i)})$ where the $\theta^{(i)}$ are drawn from I_3 in the following way. Let $Z_i \sim Bernoulli(\delta)$, $\theta_+^{(i)}$ be the samples drawn from $\pi(\theta)$, and $\theta_*^{(i)}$ be the samples drawn $\pi(\theta|x)$. Then $\theta^{(i)} = Z_i \, \theta_+^{(i)} + (1 - Z_i) \, \theta_*^{(i)}$. The strong consistency of $\hat{h}_3(x)$ as an estimate of $h(x)$ is a direct result of the consistency derived for the importance sampling estimates. Similar to the derivation for the Harmonic mean MCSE, we can derive the approximate MCSE for $\hat{h}_3(x)$ using the weights $w_i = \pi(\theta^{(i)})/[\delta\pi(\theta^{(i)}) + (1-\delta)\pi(\theta^{(i)}|x)]$. One issue with this estimate arises if either the posterior or prior density is unknown in closed form. In this case, an iterative scheme is needed in order to compute the estimate. Let k denote the iteration for the Monte Carlo integral estimate and let i denote the iteration for the parameter samples.

$$
\begin{aligned}
\hat{h}_3^{(k+1)}(x) &= \frac{\sum_{i=1}^{n} f(x|\theta^{(i)})/[\delta\hat{h}_3^{(k)}(x) + (1-\delta)f(x|\theta^{(i)})]}{\sum_{i=1}^{n}[\delta\hat{h}_3^{(k)}(x) + (1-\delta)f(x|\theta^{(i)})]^{-1}} \\
&\equiv \frac{\sum_{i=1}^{n} f(x|\theta^{(i)})/v_i^{(k+1)}(x,\delta)}{\sum_{i=1}^{n} 1/v_i^{(k+1)}(x,\delta)}
\end{aligned}
\tag{2.20}
$$

The full derivation of Equation 2.20 is provided below. See Newton and Raftery [53] for full details on the iterative scheme.

> **Derivation (2.20):** For the mixture estimate, the weighting function is $w_i = \pi(\theta^{(i)})/[\delta\pi(\theta^{(i)}) + (1-\delta)\pi(\theta^{(i)}|x)]$. The weighting function w_i cannot be calculated if either $\pi(\theta)$ or $\pi(\theta|x)$ are unknown. In this case, w_i can be evaluated given the form below.
>
> $$
> \begin{aligned}
> w_i &= \frac{\pi(\theta^{(i)})}{\delta\pi(\theta^{(i)}) + (1-\delta)\pi(\theta^{(i)}|x)} \\
> &= \left[\frac{\delta\pi(\theta^{(i)}) + (1-\delta)\pi(\theta^{(i)}|x)}{\pi(\theta^{(i)})}\right]^{-1}
> \end{aligned}
> $$

$$= \left[\delta + (1-\delta)\frac{\pi(\theta^{(i)}|x)}{\pi(\theta^{(i)})}\right]^{-1}$$

Note that $\pi(\theta^{(i)}|x) = \dfrac{f(x|\theta^{(i)})\pi(\theta^{(i)})}{h(x)}$ which means that $\dfrac{\pi(\theta^{(i)}|x)}{\pi(\theta^{(i)})} = \dfrac{f(x|\theta^{(i)})}{h(x)}$.
However, since $h(x)$ is unknown, it is approximated using the previous iteration's estimate. Therefore, the new formula for $w_i^{(k)}$, the approximation of w_i at iteration k, is

$$\begin{aligned}
w_i^{(k)} &= \left[\delta + (1-\delta)\frac{f(x|\theta^{(i)})}{\hat{h}^{(k-1)}(x)}\right]^{-1} \\
&= \left[\delta\hat{h}^{(k-1)}(x) + (1-\delta)f(x|\theta^{(i)})\right]^{-1} \\
&= \left[v_i^{(k-1)}(x,\delta)\right]^{-1}
\end{aligned}$$

Now, using the weighting function, we can derive the formula for the mixture estimate under an iterative scheme.

$$\begin{aligned}
\hat{h}_3^{(k)}(x) &= \frac{\sum_{i=1}^n w_i^{(k)} f(x|\theta^{(i)})}{\sum_{i=1}^n w_i^{(k)}} \\
&= \frac{\sum_{i=1}^n f(x|\theta^{(i)})/v_i^{(k-1)}(x,\delta)}{\sum_{i=1}^n 1/v_i^{(k-1)}(x,\delta)}
\end{aligned}$$

$\square$

If an iterative scheme is necessary to compute $\hat{h}_3(x)$ in K iterations, then the corresponding approximate MCSE formula with $w_i = 1/v_i^{(K)}(x,\delta)$ is given by

$$\epsilon_3 = \frac{\sqrt{\sum_{i=1}^n [f(x|\theta^{(i)}) - \hat{h}_3^{(K)}(x)]^2 [v_i^{(K)}(x,\delta)]^{-2}}}{\sum_{i=1}^n [v_i^{(K)}(x,\delta)]^{-1}} \tag{2.21}$$

where $v_i^{(K)}(x,\delta) = \delta\hat{h}_3^{(K-1)}(x) + (1-\delta)f(x|\theta^{(i)})$. The full derivation of Equation 2.21 is provided below.

Derivation (2.21): The MCSE formula for the mixture estimate using an iterative scheme can be derived by substituting the weighting function

derived above,

$$
\begin{aligned}
w_i^{(k)} &= \left[\delta\hat{h}^{(k-1)}(x) + (1-\delta)f(x|\theta^{(i)})\right]^{-1} \\
&= \left[v_i^{(k-1)}(x,\delta)\right]^{-1},
\end{aligned}
$$

into the general form of the MCSE given by Tanner [70],

$$
\epsilon = \frac{\sqrt{\sum_{i=1}^{n}[f(x|\theta^{(i)}) - \hat{h}(x)]^2 w_i^2}}{\sum_{i=1}^{n} w_i}.
$$

Therefore, if the iterative scheme for the mixture estimate takes K iterations, then the MCSE formula is given by

$$
\begin{aligned}
\epsilon_3 &= \frac{\sqrt{\sum_{i=1}^{n}\left[f(x|\theta^{(i)}) - \hat{h}_3^{(K)}(x)\right]^2 \left[w_i^{(K)}\right]^2}}{\sum_{i=1}^{n} w_i^{(K)}} \\
&= \frac{\sqrt{\sum_{i=1}^{n}\left[f(x|\theta^{(i)}) - \hat{h}_3^{(K)}(x)\right]^2 / \left[\delta\hat{h}_3^{(K-1)}(x) + (1-\delta)f(x|\theta^{(i)})\right]^2}}{\sum_{i=1}^{n} 1/\left[\delta\hat{h}_3^{(K-1)}(x) + (1-\delta)f(x|\theta^{(i)})\right]} \\
&= \frac{\sqrt{\sum_{i=1}^{n}\left[f(x|\theta^{(i)}) - \hat{h}_3^{(K)}(x)\right]^2 / \left[v_i^{(K)}(x,\delta)\right]^2}}{\sum_{i=1}^{n} 1/v_i^{(K)}(x,\delta)}
\end{aligned}
$$

$\square$

2.5   Gibbs Sampling

During the Monte Carlo integration methods, it is necessary to sample from various distributions for the parameter θ, even if the distribution is not known in closed form. In this case, a Gibbs sampling algorithm may be used to sample values from the approximate distribution of interest. A Gibbs sampler is a Markov chain where the limiting distribution is the true distribution of interest [70]. Suppose that the joint density is given by $f(\theta, x_1, x_2, \cdots, x_k)$ where $x_1, x_2, \cdots, x_k$ are considered ob-

servations of k different random variables, and there is interest in sampling from the marginal distribution

$$g(\theta) = \int \cdots \int f(\theta, x_1, x_2, \cdots, x_k) \, dx_1 \cdots dx_2$$

which does not have a closed-form solution. The idea of a Gibbs sampler is to create a Markov chain whose value converges to the target distribution using only univariate conditional distributions [60]. The Gibbs sampling algorithm will be presented in Algorithm 1 without derivation. The corresponding theorems (and proofs) related to the Gibbs sampler convergence can be found in Chapter 9 and Chapter 10 of Robert and Casella's textbook [60]. In subsequent chapters of this dissertation, the Gibbs sampling algorithm will be used in many of the practical examples, especially for computing the Bayes Factor. Please see Chapter 4 for further details.

Algorithm 1: Gibbs Sampler

Input: all univariate full conditional distributions $f_\theta, f_{X_1}, f_{X_2}, \ldots, f_{X_k}$
Initialize $x_0 = (x_{1_0}, x_{2_0}, \cdots, x_{k_0})$
for $i = 1, 2, ..., n$ **do**
  Simulate $\theta_{i-1} \sim f_\theta(\cdot | x_{1_{i-1}}, x_{2_{i-1}}, \cdots, x_{k_{i-1}})$
  Simulate $x_{1_i} \sim f_{X_1}(\cdot | \theta_{i-1}, x_{2_{i-1}}, x_{3_{i-1}}, \cdots, x_{k_{i-1}})$
  Simulate $x_{2_i} \sim f_{X_2}(\cdot | \theta_{i-1}, x_{1_i}, x_{3_{i-1}}, \cdots, x_{k_{i-1}})$
  $\vdots$
  Simulate $x_{k_i} \sim f_{X_k}(\cdot | \theta_{i-1}, x_{1_i}, x_{2_i}, x_{3_i}, \cdots, x_{k-1_i})$
end
Output: sample $(\theta_i, x_{1_i}, x_{2_i}, \cdots, x_{k_i})$ for $i = 0, 1, \cdots, n$ from approximate
        joint distribution $f(\theta, x_1, x_2, \cdots, x_k)$

For the examples and applications in the following chapters of this dissertation, the Gibbs sampling algorithm is used as implemented in the 'MCMCglmm' package in R [40]. These examples typically use standard distributions like the multivariate normal and the inverse Wishart. For more complicated distributions, the Gibbs sampler may need to be implemented on a case-by-case basis.

CHAPTER 3

Forensic Identification of Source

Disclaimer: Portions of this chapter are reproduced from Ommen et al.
[54] and Ommen et al. [55].

Two different types of identification of source problems will be considered in this dissertation, the common source and the specific source identification problems, although other types of identification problems are encountered in forensic science (see Kwan [45], for instance). The distinction between these two different types of identification of source problems is important because the methods used to solve the problem and the interpretation of the solution will be different. This can be done in a number of different ways, but the most fully developed structure for solving this type of problem is a modified Bayesian decision framework.

The term 'modified' is used to mean that an institutional decision analysis is being used [35]. Institutional decision analysis is used in businesses, government, and research organizations as a "process of transparently setting up a probability model, utility function, and an inferential framework leading to cost estimates and decision recommendations" [35]. The appealing feature of an institutional decision analysis for forensic evidence interpretation is that multiple people take various roles in the decision process. For instance, judges and jurors set up their own prior beliefs concerning the relative validity of the prosecution and defense models and make the final decision, and the forensic evidence interpretation expert is tasked with providing the prior belief structures necessary to construct a value of evidence. This is in contrast

to personal decision analysis in which a single person completes the entire decision-making process based on her/his beliefs alone. In an institutional setting, it may be useful to determine a community of prior beliefs, although not considered in this dissertation. A community of priors is a set of prior distributions which covers a wide range of possible beliefs that anyone involved in the decision making process could hold [12].

Under the Bayesian decision framework, the forensic statistician is tasked with providing the value of evidence as the summary of the observed evidence relative the two competing hypotheses (as well as the necessary prior belief for the nuisance parameters). Traditionally, the value of evidence is given in the form of the Bayes Factor. For more general definitions of the value of evidence see Aitken and Taroni [3], Berger [10], or Fruhwirth-Schnatter [34]. The Bayes Factor is used to convert the prior odds to posterior odds as follows:

$$
\underbrace{\frac{\mathbb{P}\left(H_p|e, I\right)}{\mathbb{P}\left(H_d|e, I\right)}}_{\text{Posterior Odds}} = \underbrace{\frac{\mathbb{P}\left(e|H_p, I\right)}{\mathbb{P}\left(e|H_d, I\right)}}_{\text{Value of Evidence}} \times \underbrace{\frac{\mathbb{P}\left(H_p|I\right)}{\mathbb{P}\left(H_d|I\right)}}_{\text{Prior Odds}}, \tag{3.1}
$$

where $\mathbb{P}$ is a probability measure characterizing all types of uncertainty, e is the observed evidence, H_p and H_d are the prosecution and defense hypotheses (also commonly termed propositions), and I is the relevant background information common to both hypotheses [46]. The prior odds summarizes the decision-maker's relative personal belief concerning the validity of the prosecution and defense hypotheses before observing the evidence. The value of evidence then allows the decision-maker to update the personal prior belief following the observation of the evidence and arrive at the posterior odds concerning the relative validity of the two hypotheses after observing the evidence. The statistics community tends to use the term "Bayes Factor" when referring to the value of evidence, whereas the forensic science community typically uses the term "likelihood ratio" [38]. However, the Bayes Factor and likelihood

ratio are two distinct quantities. Section 3.3 will discuss in detail the likelihood ratio, the Bayes Factor, and the comparison between the two for both identification of source problems.

In the remainder of this dissertation, the value of the evidence is based on three possible subsets of evidence. The sets of evidence acquired can either originate from a known source (typically referred to as control, or known, material [3]) or from an unknown source (typically referred to as recovered, or questioned, material [3]). The entire set of evidence (consisting of all three subsets) will be denoted by e. In Aitken and Taroni [3], they make a distinction between the objects and the measurements/observations on that object. That distinction is not made here in an attempt to simplify the notation, but note that this set-up can easily be extended to accommodate that distinction. The two competing hypotheses provide a statement of the potential source(s) for the recovered evidence, and these statements of these two hypotheses determine the type of source identification problem under consideration. Depending on the type of identification problem, a different number of evidence sets of each type, known or unknown source, are acquired. For the common source identification problem, there are two sets of unknown source evidence and one set of evidence from many known sources, typically called the background population. For the specific source identification problem, there is one set of unknown source evidence, one set of evidence from a single known source of interest, and one set of evidence from many known sources (the background population). Subsections 3.1 and 3.2 will discuss the evidence and statement of the forensic hypotheses in detail.

3.1 Identification of Common Source

In forensic science, it is often of interest to determine whether two different crimes are related. For instance, consider the situation in which, on separate occasions, two different banks were robbed and a bank robbery note was left at each scene. The two bank robbery notes are collected and then compared to see if they were written by the same person, suggesting that the bank robberies were committed by the same perpetrator. This is an example of the common source identification problem since the question of interest is whether or not the two bank robberies were committed by the same (unknown) person, but without specifying which person.

The forensic hypotheses for the common source problem are typically stated as follows:

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.

In the case of the bank robberies, the forensic hypotheses will typically be stated as "The two bank robbery notes were written by the same (unknown) person" versus "The two bank robbery notes were written by two different (unknown) people."

For the common source problem, the evidence consists of those materials recovered which originate from the first unknown source, denoted e_{u_1}, those materials recovered which originate from the second unknown source, denoted e_{u_2}, and those control materials which originate from the population of alternative sources, denoted e_a, and sometimes referred to as the background population. The entire set of evidence will be denoted $e = \{e_{u_1}, e_{u_2}, e_a\}$. In the case of the bank robberies, the two bank robbery notes (and the measurements or observations made on those notes) serve as e_{u_1} and

e_{u_2}. The alternative source population e_a might, for instance, be the collection of 'London Letters' collected for the FBI 500 dataset [63].

In other application areas, this type of problem might be solved using traditional hypothesis testing methods to determine which hypothesis, prosecution or defense, is supported by the evidence. In order to apply the hypothesis testing methods, the statistical models for the evidence need to be specified by the hypotheses up to the point of a finite dimensional vector space for the nuisance parameters. Then, the sampling distribution as specified by one set of parameters would be implied by the prosecution hypothesis, and the sampling distribution as specified by another set of parameters would be implied by the defense hypothesis. That is, in hypothesis testing problems, the parameter space is a set of indexing variables for a single specified sampling distribution. However, in forensic identification of source problems, the forensic hypotheses provide no clear idea as to what the sampling distributions for the evidence are (at least in most scenarios outside of simple DNA analysis). In these applications, the parameter space consists of a set of possible sampling models from which a selection is to be made. These models only concern the exchangeability of the observations. Since "the sampling distribution is unknown to a larger extent than simply depending on an unknown (finite dimensional) parameter," the forensic identification of source problems are more well-suited to methods of Bayesian model selection [59].

The full Bayesian models for the forensic identification of source problems consist of three separate parts. The first is the statement of the sampling models which provide information about which samples are exchangeable under each of the hypotheses. The second is the statement of the class of parametric models approximating the true sampling distributions implied by each of the sampling models. Finally, the third is a statement specifying the prior belief structure for the parameters characterizing the

class of probability models specified by each of the parametric models. The extent of the information that can be obtained by the forensic hypotheses are summarized in the following sampling models for the common source problem.

M_a: e_a is a sample generated by first randomly selecting n_a sources from the population of alternative sources, and then randomly selecting n_i elements from within the ith source for $i = 1, 2, \ldots, n_a$

M_p: e_{u_1} and e_{u_2} are two samples generated by first randomly selecting a single source from the population of alternative sources, and then e_{u_1} is generated by randomly selecting the first set of n_{u_1} elements from within that source, and finally e_{u_2} is generated by randomly selecting the second set of n_{u_2} elements from within the same source

M_d: e_{u_1} and e_{u_2} are two samples generated by first randomly selecting two different sources from the population of alternative sources, and then e_{u_1} is generated by randomly selecting n_{u_1} elements from within the first source, and finally e_{u_2} is generated by randomly selecting n_{u_2} elements from within the second source

For the common source problem, the prosecution hypothesis implies that the control evidence has been generated according to model M_a and that the recovered evidence has been generated according to model M_p, whereas the defense hypothesis implies that the control evidence has been generated according to model M_a, but that the recovered evidence has been generated according to model M_d. The model selection problem is then a selection between M_p and M_d for the unknown source evidence. In this scenario, the *distributional* models which generated e_{u_1} and e_{u_2} is the same, but the *exchangeability* models for e_{u_1} and e_{u_2} are different under the two different sampling models. Under the prosecution model, e_{u_1} and e_{u_2} are conditionally independent given the common source, and e_{u_1} and e_{u_2} are unconditionally independent under the defense model.

Again, the goal of specifying the sampling models is to indicate the exchangeability assumptions for the evidence. To specify the full Bayesian model for the evidence, a statement about the class of parametric models is needed [59]. To facilitate the use of Bayesian model selection methods, the classes of distributions will be constrained such that the indexing parameter is an element of a finite dimensional vector space, Θ_a, which is a subset of Euclidean space. Also, the indexing sets for these classes of distributions is assumed to be chosen such that the likelihoods are identifiable. The following notation will be used for the parametric models under M_a. The true sampling distribution for e_a, P_{a_0}, is assumed to be in a class of distributions indexed by the set of parameters Θ_a given by $\mathcal{P}_a = \{P_{\theta_a} : \theta_a \in \Theta_a\}$. The likelihood functions for the common source evidence will be denoted by $f(e_a|\theta_a)$, $f(e_{u_1}|\theta_a)$, and $f(e_{u_2}|\theta_a)$. Please see Section 3.4.3 for the details of the evidence and Section 3.4.1 for the construction of these likelihood functions. Finally, to complete the full Bayesian model a prior belief structure on the parameter space needs to be specified [59]. Here, I will use the standard Bayesian abuse of notation and let $\Pi(\theta_a)$ denote the proper prior probability measure on the parameter space Θ_a, with corresponding prior density function, $\pi(\theta_a)$, when it exists.

3.2   Identification of Specific Source

In contrast to the common source identification problem, it is often of interest to determine whether a suspect can be linked to evidence found at the scene of the crime. Consider that one bank has been robbed and one note has been left at the scene. After the bank robbery note is collected, the investigators find some handwritten documents at the suspect's place of residence. These documents serve as the suspect's template, which is compared to the bank robbery note to determine if the suspect wrote the bank robbery note. This is an example of the specific source identification problem

since the question of interest is whether or not the specified suspect wrote the bank robbery note. In the common source problem, the person who wrote the note is not identified and is treated as unknown. In the specific source problem, the person is identified as the suspect, and is treated as known.

The forensic hypotheses for the specific source problem are typically stated as follows:

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.

For the bank robbery, the specific source hypotheses are typically stated as "The suspect is the writer of the bank robbery note" versus "The suspect is not the writer of the bank robbery note."

For the specific source problem, the evidence consists of those materials recovered which originate from an unknown source, denoted e_u, those control materials which originate from a known, fixed specific source, denoted e_s, and those control materials which originate from the population of alternative sources, denoted e_a, and sometimes referred to as the background population. The entire set of evidence will be denoted $e = \{e_u, e_s, e_a\}$. In this case, the bank robbery note serves as e_u, the documents written by the suspect serve as e_s, and just as in the common source problem, the London Letters can serve as e_a [63].

Analogous to the common source problem, the specific source problem is suited to the methods of model selection. A typical example of the specific source sampling models are as follows:

M_s: e_s is a sample generated by randomly selecting n_s samples from the population

of known specific source samples

M_a: e_a is a sample generated by first randomly selecting n_a sources from the population of alternative sources, and then randomly selecting n_i samples from within the ith source for $i = 1, 2, \ldots, n_a$

M_p: e_u is a sample generated by randomly selecting n_u samples from the population of known specific source samples

M_d: e_u is a sample generated by first randomly selecting a new source from the population of alternative sources, and then randomly selecting n_u samples from within that source

For the specific source problem, the prosecution hypothesis implies that e_a and e_s have been generated according to model M_a and M_s, respectively, and that e_u has been generated according to model M_p. In contrast, the defense hypothesis implies that e_a and e_s have been generated according to model M_a and M_s, respectively, but that e_u has been generated according to model M_d. Similar to the common source problem, the model selection problem is then a selection between M_p and M_d for the unknown source evidence. In this scenario, the *distributional* models which generated the unknown source evidence are different under the different sampling models, and the *exchangeability* models are different, too. Under the prosecution model, e_u and e_s are conditionally independent given the parameters for the specific source population and both are unconditionally independent of e_a. However, under the defense model e_u and e_a are conditionally independent given the parameters for the alternative source population and both are unconditionally independent of e_s.

Similar to the common source problem, the sampling models only provide information about the exchangeability of the evidence so the parametric models need to be specified as well. The following notation will be used for the parametric models under M_s. The true sampling distribution for e_s, P_{s_0}, is assumed to be in a class of distributions

such that the indexing parameter is an element of a finite dimensional vector space $\Theta_s \subseteq \mathbb{R}^k$ given by $\mathcal{P}_s = \{P_{\theta_s} : \theta_s \in \Theta_s\}$. The following notation will be used for the parametric models under M_a. The true sampling distribution for e_a, P_{a_0}, is assumed to be in a class of distributions such that the indexing parameter is an element of a finite dimensional vector space $\Theta_a \subseteq \mathbb{R}^k$ given by $\mathcal{P}_a = \{P_{\theta_a} : \theta_a \in \Theta_a\}$. Both $\mathcal{P}_s$ and $\mathcal{P}_a$ should be defined so that the likelihoods are identifiable.

The likelihood function for the evidence from the specific source will be denoted $f(e_s|\theta_s)$ and the likelihood for the alternative source population will be denoted $f(e_a|\theta_a)$. The likelihood function for the trace evidence will depend upon the model under consideration. Under the prosecution model, the likelihood function for the trace evidence is denoted $f(e_u|\theta_s)$ and under the defense model it will be denoted $f(e_u|\theta_a)$. Please see Section 3.4.2 for the details of the evidence and Section 3.4.1 for the construction of these likelihood functions. Finally, to complete the full Bayesian model, a prior belief structure on the parameter spaces needs to be specified [59]. Again, the standard Bayesian abuse of notation will be used and let $\Pi(\theta_s)$ denote the proper prior probability measure on the parameter space Θ_s, with corresponding prior density function, $\pi(\theta_s)$, when it exists. Let $\Pi(\theta_a)$ denote the proper prior probability measure on the parameter space Θ_a, with corresponding prior density function, $\pi(\theta_a)$, when it exists.

3.3 Quantifying the Value of Evidence

In an institutional decision process, it is typically the task of the legal teams to determine the statement of the forensic hypotheses. From these hypotheses the sampling models will follow readily according to the previous sections. It is then the task of the forensic expert to determine the parametric models and the prior distributions.

Once the expert has specified the full Bayesian model for the evidence under each of the hypotheses, the next task is to provide a quantification of the value of evidence. Again, "Bayes Factor" is the term used by the statistics community when referring to the value of evidence, whereas "likelihood ratio" is the term typically used in the forensic science community [38]. However, the Bayes Factor (BF) and likelihood ratio (LR) are two distinct quantities in the author's opinion. In particular, when the values of the parameters for each of the sampling models are known with certainty, the value of evidence takes the form of the LR. The likelihood ratio is given by

$$V_{LR}(\theta_0; e_u) = \frac{f(e_u|\theta_0, M_p)}{f(e_u|\theta_0, M_d)} \tag{3.2}$$

where e_u denotes the unknown evidence, θ_0 represents the true value of the parameters, when they are known, and M_p and M_d denote the prosecution and defense models, respectively. Indeed, the likelihood structure, denoted f, and the values of the parameters, denoted θ_0, for these models need to be known with complete certainty to compute the likelihood ratio [10, 59]. When there is no uncertainty concerning the LR, it *can* formally be used as the value of evidence in Equation 3.1 to update the prior odds and arrive at the exact posterior odds. The particular forms of the LR for each of the identification of source problems is provided in the following sections.

The likelihood ratio is the statistic at the forefront of the Likelihood paradigm for evidence interpretation [61]. The particular forms of the LR for each of the identification of source problems is provided in Sections 3.3.1 and 3.3.2. It should be noted that in the Frequentist or Likelihood paradigms, the LR naturally exists for any well-defined population. Under these paradigms, θ_0 corresponds to the true value of the parameter for the statistical experiment which describes the methods for sampling from the population. Under the Bayesian paradigm, the concept of a "true" value

of the parameter is not well-defined (it is actually a highly debated concept among subjectivists in forensic science whether a "true" value of the parameter even exists [9, 71]). In the Bayesian paradigm, the belief in the existence of the true parameter value is implied (almost-surely) by Doob's Consistency Theorem (Theorem 2.10) relative to the specified prior distribution [75]. As a result, a Bayesian believes that a likelihood ratio (analogous to the Frequentist likelihood ratio) is guaranteed to eventually exist. Please see Section 5 for further details. Under the three considered statistical paradigms, it is often impossible to quantify the likelihood ratio in most practical applications since there is some uncertainty concerning the value of θ_0.

When there is uncertainty concerning the values of the parameters, other strategies need to be used to quantify the value of evidence. Consider the likelihood ratio function to be defined as

$$V_{LR}(\theta; e_u) = \frac{f(e_u|\theta, M_p)}{f(e_u|\theta, M_d)} \tag{3.3}$$

which is a function of the unknown parameter, θ. From this notation, it is clear that the likelihood ratio defined above represents a single value of the likelihood ratio function for the specified value of θ. There are many ad-hoc solutions to the forensic identification of source problem that involve the likelihood ratio function [39, 64, 27]. The most common ad-hoc solution is to take some estimate of the unknown parameter and substitute it into the likelihood ratio function [39]. This results in what is commonly referred to as the 'plug-in' LR for forensic evidence. In the best-case scenario, these ad-hoc methods are asymptotic approximations of the value of evidence, and therefore, can *not* be used formally in the odds form of Bayes Theorem above to arrive at the exact posterior odds. The precise forms of the likelihood ratio functions for the common source and specific source problems are provided in Sections 3.3.1 and 3.3.2.

As an alternative to the ad-hoc methods described above, one of the commonly used

formal strategies for incorporating the uncertainty into the value of evidence is to construct a Bayes Factor. The Bayes Factor takes the form

$$V_{BF}(e) = \frac{\int f(e|\theta, M_p) \, d\Pi(\theta|M_p)}{\int f(e|\theta, M_d) \, d\Pi(\theta|M_d)}. \tag{3.4}$$

which is the ratio of the *marginal* likelihood of observing the entire set of evidence under the prosecution model to the *marginal* likelihood of observing the entire set of evidence under the defense model. In contrast to the ad-hoc methods described above, the Bayes Factor *can* be used formally in the odds form of Bayes Theorem above to arrive at the exact posterior odds for a given set of prior beliefs [67]. The Bayes Factor is the statistic that is typically of interest in the Bayesian paradigm for hypothesis testing, model selection, and decision analysis [10, 59]. The particular forms of the Bayes Factor for each of the identification of source problems is given in the Sections 3.3.1 and 3.3.2. In contrast to the likelihood ratio, the Bayes Factor can be quantified in many practical applications, although it is a difficult and very computationally intensive process (see Ommen et al. [55] for details of a naive method of computing Bayes Factors and their corresponding standard errors). When the values of the parameters are known, then the Bayes Factor (constructed using an appropriate degenerate prior on θ) and the likelihood ratio are equal, and both are equivalent to the value of evidence in Equation 3.1.

3.3.1   Common Source

The likelihood ratio function, as defined in Equation 3.3, for the common source identification problem is given by

$$LR_{cs}(\theta_a; e_u) = \frac{f(e_{u_1}, e_{u_2}|\theta_a, M_p)}{f(e_{u_1}|\theta_a, M_d)f(e_{u_2}|\theta_a, M_d)} \tag{3.5}$$

which reflects that there are two sets of unknown evidence; under the prosecution model they are two sets from the same randomly selected source whose distribution is indexed by the parameter θ_a, in contrast to the defense model under which they are two independent sets from two different randomly selected sources whose distributions are indexed by the same parameter θ_a. For details of the likelihood structures for the unknown source evidence, see Section 3.4.3. The "true" common source likelihood ratio, as defined in Equation 3.2, takes the form

$$LR_{cs}(\theta_{a_0}; e_u) = \frac{f(e_{u_1}, e_{u_2} | \theta_{a_0}, M_p)}{f(e_{u_1} | \theta_{a_0}, M_d) f(e_{u_2} | \theta_{a_0}, M_d)} \tag{3.6}$$

where the only difference between the true likelihood ratio and the likelihood ratio function is that the true likelihood ratio represents a single point of the likelihood ratio function corresponding to the value of θ_{a_0} for θ_a. Under the Frequentist paradigm, θ_{a_0} corresponds to the true value of the parameter. Under the Bayesian paradigm, θ_{a_0} corresponds to the fixed value of the parameter suggested by Doob's Consistency Theorem [75].

The Bayes Factor, following the definition provided in Equation 3.4, for the common source identification problem considers the ratio of the marginal density of the evidence under the prosecution model to the marginal density of the evidence under the defense model. Under the assumption that the marginal distribution of e_a will be the same under both the prosecution and defense models, the commons source Bayes Factor can be given as the ratio of the posterior predictive distributions of the unknown source evidence given the alternative source population evidence and the prosecution model, to the defense model, respectively. This form of the common source Bayes Factor is given in Equation 3.7 below.

$$BF_{cs1}(e) = \frac{\int f(e_{u_1}, e_{u_2} | \theta_a, M_p) \, d\Pi(\theta_a | e_a)}{\int f(e_{u_1} | \theta_a, M_d) \, f(e_{u_2} | \theta_a, M_d) \, d\Pi(\theta_a | e_a)} \tag{3.7}$$

The derivation of Equation 3.7 is reproduced from Ommen et al. [55] below.

Derivation (3.7): The derivation starts from Equation 3.4 (step 3.8a). It is assumed that the prior choice has been made such that the integrals expressed exist and are finite. Then, using the likelihood structure presented in Equation 3.13 and the definition of the common source evidence in Section 3.4.3, the likelihoods are factored (step 3.8b). Next, we make the assumption that the marginal density for the alternative source population evidence e_a is the same for both of the models, i.e. $f(e_a|M_p) = f(e_a|M_d)$. Therefore, multiplying by $f(e_a|M_d)/f(e_a|M_p) = 1$ in step 3.8b doesn't change the value of the equation. The rest of the derivation (steps 3.8c and 3.8d) follows from standard definitions in Bayesian analysis, and step 3.8e from the fact that the prior belief of θ_a is the same under both models. Therefore, the posterior belief of θ_a given e_a is also the same under both models.

$$BF_{cs1}(e) = \frac{\int f(e|\theta_a, M_p)\, d\Pi(\theta_a|M_p)}{\int f(e|\theta_a, H_d)\, d\Pi(\theta_a|M_d)} \tag{3.8a}$$

$$= \frac{\int f(e_{u_1}, e_{u_2}|\theta_a, M_p) f(e_a|\theta_a, M_p)\, d\Pi(\theta_a|M_p)}{\int f(e_{u_1}, e_{u_2}|\theta_a, M_d) f(e_a|\theta_a, M_d)\, d\Pi(\theta_a|M_d)} \times \frac{f(e_a|M_d)}{f(e_a|M_p)} \tag{3.8b}$$

$$= \frac{\int f(e_{u_1}, e_{u_2}|\theta_a, M_p)\, \dfrac{f(e_a|\theta_a, M_p)\, d\Pi(\theta_a|M_p)}{f(e_a|M_p)}}{\int f(e_{u_1}|\theta_a, M_d) f(e_{u_2}|\theta_a, M_d)\, \dfrac{f(e_a|\theta_a, M_d)\, d\Pi(\theta_a|M_d)}{f(e_a|M_d)}} \tag{3.8c}$$

$$= \frac{\int f(e_{u_1}, e_{u_2}|\theta_a, M_p)\, d\Pi(\theta_a|e_a, M_p)}{\int f(e_{u_1}|\theta_a, M_d) f(e_{u_2}|\theta_a, M_d)\, d\Pi(\theta_a|e_a, M_d)} \tag{3.8d}$$

$$= \frac{\int f(e_{u_1}, e_{u_2}|\theta_a, H_p)\, d\Pi(\theta_a|e_a)}{\int f(e_{u_1}|\theta_a, H_d)\, f(e_{u_2}|\theta_a, H_d)\, d\Pi(\theta_a|e_a)} \tag{3.8e}$$

$\square$

3.3.2 Specific Source

The likelihood ratio function, as defined in Equation 3.3, for the specific source identification problem is given by

$$LR_{ss}(\theta; e_u) = \frac{f(e_u|\theta_s, M_p)}{f(e_u|\theta_a, M_d)} \tag{3.9}$$

which reflects that under the prosecution model the unknown evidence is a random sample from the specific source population whose distribution is indexed by the parameter θ_s, and that under the defense model the unknown evidence is a random sample from a randomly selected source in the alternative source populations whose distribution is indexed by the parameter θ_a. The "true" specific source likelihood ratio, as defined in Equation 3.2, represents a single point of the likelihood ratio function corresponding to a value of $\theta_0 = \{\theta_{s_0}, \theta_{a_0}\}$ for $\theta = \{\theta_s, \theta_a\}$, and is given below.

$$LR_{ss}(\theta_0; e_u) = \frac{f(e_u|\theta_{s_0}, M_p)}{f(e_u|\theta_{a_0}, M_d)} \tag{3.10}$$

The specific source Bayes Factor, as defined in Equation 3.4, does not require any assumptions on the prior distribution for the parameter. Analogous to the common source, the specific source Bayes Factor is the ratio of the marginal density of the evidence given the prosecution model to the marginal density of the evidence given the defense model. Under the assumption that the prior distribution of θ_s is statistically independent of the prior distribution for θ_a, the specific source Bayes Factor can be given as the ratio of posterior predictive distribution for the unknown source evidence given the specific source evidence to the posterior predictive distribution for the unknown source evidence given the alternative source population evidence. This

form of the specific source Bayes Factor is given in Equation 3.11 below.

$$BF_{ss1}(e) = \frac{\int f(e_u|\theta_s, M_p) \; d\Pi(\theta_s|e_s)}{\int f(e_u|\theta_a, M_d) \; d\Pi(\theta_a|e_a)} \tag{3.11}$$

The derivation of Equation 3.11 is reproduced from Ommen et al. [55] below.

Derivation (3.11): The derivation starts from Equation 3.4 (step 3.12a). Again, it is assumed that the prior choice has been made such that the integrals expressed exist and are finite. Next, the likelihood structure presented in Equation 3.13 and the definition of the specific source evidence in Section 3.4.2 is used to factor the likelihood into three different pieces, one for each of the evidence datasets (steps 3.12b and 3.12c). Then, using the assumption that the prior for θ_s is independent of the prior for θ_a, the marginal likelihoods in the numerator and denominator of the value of evidence can be split into two different integrals, one for each of the parameters θ_s and θ_a (steps 3.12d and 3.12e). The rest of the derivation (steps 3.12f, 3.12g, and 3.12h) follows from standard definitions in Bayesian analysis.

$$BF_{ss1}(e) = \frac{\int f(e|\theta, M_p) \; d\Pi(\theta|M_p)}{\int f(e|\theta, M_d) \; d\Pi(\theta|M_d)} \tag{3.12a}$$

$$= \frac{\int f(e_u|\theta, M_p)f(e_s|\theta, M_p)f(e_a|\theta, M_p) \; d\Pi(\theta|M_p)}{\int f(e_u|\theta, M_d)f(e_s|\theta, M_d)f(e_a|\theta, M_d) \; d\Pi(\theta|M_d)} \tag{3.12b}$$

$$= \frac{\int f(e_u|\theta_s)f(e_s|\theta_s)f(e_a|\theta_a) \; d\Pi(\theta)}{\int f(e_u|\theta_a)f(e_s|\theta_s)f(e_a|\theta_a) \; d\Pi(\theta)} \tag{3.12c}$$

$$= \frac{\int f(e_u|\theta_s)f(e_s|\theta_s)d\Pi(\theta_s) \; \int f(e_a|\theta_a)d\Pi(\theta_a)}{\int f(e_s|\theta_s)d\Pi(\theta_s) \; \int f(e_a|\theta_a)f(e_u|\theta_a)d\Pi(\theta_a)} \tag{3.12d}$$

$$= \frac{\int f(e_u|\theta_s)f(e_s|\theta_s) \; d\Pi(\theta_s)}{\int f(e_s|\theta_s) \; d\Pi(\theta_s)} \times \frac{\int f(e_a|\theta_a) \; d\Pi(\theta_a)}{\int f(e_a|\theta_a)f(e_u|\theta_a) \; d\Pi(\theta_a)}$$
$$\tag{3.12e}$$

$$= \frac{\int f(e_u|\theta_s)\dfrac{f(e_s|\theta_s) \; d\Pi(\theta_s)}{\int f(e_s|\theta_s) \; d\Pi(\theta_s)}}{\int f(e_u|\theta_a)\dfrac{f(e_a|\theta_a) \; d\Pi(\theta_a)}{\int f(e_a|\theta_a) \; d\Pi(\theta_a)}} \tag{3.12f}$$

$$= \frac{\int f(e_u|\theta_s) \dfrac{f(e_s|\theta_s) \ d\Pi(\theta_s)}{f(e_s)}}{\int f(e_u|\theta_a) \dfrac{f(e_a|\theta_a) \ d\Pi(\theta_a)}{f(e_a)}} \tag{3.12g}$$

$$= \frac{\int f(e_u|\theta_s) \ d\Pi(\theta_s|e_s)}{\int f(e_u|\theta_a) \ d\Pi(\theta_a|e_a)} \tag{3.12h}$$

$$\square$$

3.4   Technical Details

3.4.1   The Likelihood Function for the Evidence

Let $e = \{e_1, e_2, \ldots, e_r\}$ denote the forensic evidence (a collection of r independent numerical datasets) used to calculate the value of evidence. In the forensic identification of source problems considered in this dissertation $r = 3$. Let the likelihood function for e_i (for $i = 1, 2, \ldots, r$) be denoted as $f(\cdot|\theta_{ij}, H_j)$ under each hypothesis (for $j = p, d$). We use the following short-hand notation to represent the joint likelihood function for the entire set of forensic evidence e

$$f(e|\theta_j, H_j) = \prod_{i=1}^{r} f(e_i|\theta_{ij}, H_j), \tag{3.13}$$

where $\theta_j = \{\theta_{1j}, \theta_{2j}, \ldots, \theta_{rj}\}$ denotes the collection of parameters for all of the likelihood functions under hypothesis H_j, and H_j denotes either the prosecution or defense hypothesis. It will typically be the case that θ_p and θ_d will be the same, so for this case, we drop the subscript on θ to denote the common set of parameters under both hypotheses.

3.4.2 Specific Source Evidence

The specific source evidence will be the collection of three independent datasets, $e = \{e_s, e_a, e_u\}$, where e_s is the dataset containing the numerical measurements on the specific source evidence, e_a is the dataset containing the numerical measurements on samples from the alternative source population, and e_u is the dataset containing the numerical measurements on the unknown source evidence. For this development, only balanced designs are considered (i.e. there are no missing values for the measurements and the number of measurements per each sample is constant), but the generalization to unbalanced designs is straight-forward.

First, consider the dataset e_s which consists of n_s random samples from the specific source, each of which contains m_s different measurements. Then e_s has the matrix structure given in Equation 3.14 below where each column corresponds to a sample and each row corresponds to a measurement.

$$e_s = \begin{pmatrix} y_{s,11} & y_{s,21} & \cdots & y_{s,n_s1} \\ y_{s,12} & y_{s,22} & \cdots & y_{s,n_s2} \\ \vdots & \vdots & \ddots & \vdots \\ y_{s,1m_s} & y_{s,2m_s} & \cdots & y_{s,n_sm_s} \end{pmatrix} \equiv \begin{pmatrix} \mathbf{y}_{s1} & \mathbf{y}_{s2} & \cdots & \mathbf{y}_{sn_s} \end{pmatrix} \tag{3.14}$$

It will be beneficial to consider each column of the dataset as a random vector, independently and identically distributed to each of the other columns of the dataset (this assumption will be approximately met for a majority of forensic evidence types). Denote each column of the dataset as $\mathbf{y}_{si}$ which represents the m_s-dimensional vector of measurements made on the ith sample from e_s (for $i = 1, 2, \ldots, n_s$). Then,

$$\mathbf{y}_{si} \overset{iid}{\sim} F_s(\cdot|\theta_s)$$

where F_s is the distribution function indexed by the parameter θ_s corresponding to the probability measure P_{θ_s} under model M_s. If we denote the probability density function corresponding to the distribution F_s by f_s, then following the notational conventions in the previous subsection, we can define the likelihood function for e_s by

$$f(e_s|\theta_s, H_p) = f(e_s|\theta_s, H_d) = \prod_{i=1}^{n_s} f_s(\mathbf{y}_{si}|\theta_s) \tag{3.15}$$

since the observations in e_s are independent and identically distributed under model M_s (the sampling model implied by both H_p and H_d for the specific source evidence). Since this likelihood structure is the same under both the prosecution and defense hypotheses the notational dependence on H_p or H_d is typically dropped, $f(e_s|\theta_s)$.

Next, consider the dataset e_a which has a hierarchical sampling structure in which n_a sources are randomly sampled from the alternative source population and then from each of the sample sources, n_i samples are collected. From each of these sources, m_a measurements are recorded. Then e_a has the block matrix structure given in Equation 3.16 below where each block corresponds to a sampled source, each column within a block corresponds to a sample from that source, and each row within a block corresponds to a measurement on a sample.

$$e_a = \left(\; \mathbf{Y}_{a_1} \; \middle| \; \mathbf{Y}_{a_2} \; \middle| \cdots \middle| \; \mathbf{Y}_{a_{n_a}} \; \right)$$

and for $i = 1, 2, \ldots, n_a$

$$\mathbf{Y}_{a_i} = \begin{pmatrix} y_{a_i,11} & y_{a_i,21} & \cdots & y_{a_i,n_i1} \\ y_{a_i,12} & y_{a_i,22} & \cdots & y_{a_i,n_i2} \\ \vdots & \vdots & \ddots & \vdots \\ y_{a_i,1m_a} & y_{a_i,2m_a} & \cdots & y_{a_i,n_im_a} \end{pmatrix} \equiv \begin{pmatrix} \mathbf{y}_{i1} & \mathbf{y}_{i2} & \cdots & \mathbf{y}_{in_i} \end{pmatrix} \tag{3.16}$$

Again, it will be beneficial to consider each column of $\mathbf{Y}_{a_i}$ as a random vector, inde-

pendently and identically distributed to each of the other columns of the $\mathbf{Y}_{a_i}$ (again, this assumption will be approximately met for a majority of forensic evidence types). Denote each column of $\mathbf{Y}_{a_i}$ as $\mathbf{y}_{ij}$ which represents the m_a-dimensional vector of measurements made on the jth sample from the ith source in e_a (for $i = 1, 2, \ldots, n_a$ and $j = 1, 2, \ldots, n_i$). Then,

$$A_i \overset{iid}{\sim} G(\cdot|\theta_a) \quad \text{and} \quad \mathbf{y}_{ij}|a_i \overset{iid}{\sim} F_a(\cdot|a_i, \theta_a)$$

where A_i is the random variable used to denote the source being sampled from the alternative source population, and F_a and G are both probability distribution indexed by the parameters θ_a under both the prosecution and defense hypotheses. The distribution function F_a corresponds to the probability measure P_{θ_a} under model M_a. Therefore, the probability density function for $\mathbf{y}_{ij}$ can be defined by

$$f_a(\mathbf{y}_{ij}|\theta_a) = \int f_a(\mathbf{y}_{ij}|a_i, \theta_a)g(a_i|\theta_a) \, da_i$$

where g is the probability density function for a_i corresponding to the distribution G and f_a is the probability density function corresponding to F_a. However, the likelihood function for e_a is commonly intractable since the $\mathbf{y}_{ij}$ are not independent. Therefore, if we define the joint probability density function of the $\mathbf{y}_{ij}$ for a fixed i to be

$$\begin{aligned} f_{a_i}(\mathbf{y}_{i1}, \mathbf{y}_{i2}, \ldots, \mathbf{y}_{in_i}|\theta_a) &= \int \left[\prod_{j=1}^{n_i} f_a(\mathbf{y}_{ij}|a_i, \theta_a) \right] g(a_i|\theta_a) \, da_i \qquad (3.17) \\ &\equiv f(\mathbf{Y}_{a_i}|\theta_a), \end{aligned}$$

then we do in fact have independence due to the random sampling of sources from the alternative source population and we can define the likelihood function for e_a to

be

$$f(e_a|\theta_a, H_p) = f(e_a|\theta_a, H_d) = \prod_{i=1}^{n_a} f_{a_i}(\mathbf{y}_{i1}, \mathbf{y}_{i2}, \ldots, \mathbf{y}_{in_i}|\theta_a).$$

In general, there is no simplification of the likelihood structure presented in Equation 3.17. However, there is a simplification presented by Miller [49] under the assumption of normality which was discussed in Section 2.1. Again, since this likelihood structure is the same under both the prosecution and defense hypotheses the notational dependence on H_p or H_d is typically dropped, $f(e_a|\theta_a)$.

Finally, consider the dataset e_u which consists of n_u random samples from the unknown source, each of which contains m_u different measurements. Then e_u has the matrix structure given in Equation 3.18 below where each column corresponding to a sample and each row corresponds to a measurement.

$$e_u = \begin{pmatrix} y_{u,11} & y_{u,21} & \cdots & y_{u,n_u 1} \\ y_{u,12} & y_{u,22} & \cdots & y_{u,n_u 2} \\ \vdots & \vdots & \ddots & \vdots \\ y_{u,1m_u} & y_{u,2m_u} & \cdots & y_{u,n_u m_u} \end{pmatrix} \equiv \begin{pmatrix} \mathbf{y}_{u1} & \mathbf{y}_{u2} & \cdots & \mathbf{y}_{un_u} \end{pmatrix} \tag{3.18}$$

It will be beneficial to consider each column of the dataset as a random vector. Denote each column of the dataset as $\mathbf{y}_{ui}$ which represents the m_u-dimensional vector of measurements made on the ith sample from e_u (for $i = 1, 2, \ldots, n_u$). Then under H_p, e_u is an additional independent random sample from the specific source population, so the likelihood structure is the same as e_s. Conversely, under H_d, e_u is a random sample from an additional randomly selected source from the alternative source population, so the likelihood structure is the same as e_a. Therefore, the likelihood functions for e_u can be defined by

$$f(e_u|\theta_s, H_p) \equiv f(e_u|\theta_s, M_p) = \prod_{i=1}^{n_u} f_s(\mathbf{y}_{ui}|\theta_s)$$

and

$$f(e_u|\theta_a, H_d) \equiv f(e_u|\theta_a, M_d) = f_{a_u}(\mathbf{y}_{u1}, \mathbf{y}_{u2}, \ldots, \mathbf{y}_{un_u}|\theta_a).$$

3.4.3   Common Source Evidence

The common source evidence will be the collection of three independent datasets, $e = \{e_a, e_{u_1}, e_{u_2}\}$, where e_a is the dataset containing the numerical measurements on samples from the alternative source population, e_{u_1} is the dataset containing the numerical measurements taken from the first set of unknown source evidence, and e_{u_2} is the dataset containing the numerical measurements taken from the second set of unknown source evidence. Again, only balanced designs will be considered (i.e. there are no missing values for the measurements and the number of measurements per each sample is constant), but the generalization to unbalanced designs is straightforward.

The dataset e_a has the exact same matrix structure and likelihood function as described for e_a in the previous subsection, and e_{u_1} and e_{u_2} have the same matrix structure as e_u in the previous subsection. Now, under H_p, e_{u_1} and e_{u_2} are supposed to be conditionally independent samples drawn from the same randomly selected source in the alternative source population. Therefore, the joint likelihood function for e_{u_1} and e_{u_2} under M_p is defined as

$$
\begin{aligned}
f(e_{u_1}, e_{u_2}|\theta_a, H_p) &= f_{a_u}(\mathbf{y}_{u_11}, \ldots, \mathbf{y}_{u_1n_{u_1}}, \mathbf{y}_{u_21}, \ldots, \mathbf{y}_{u_2n_{u_2}}|\theta_a) \\
&\equiv f(e_{u_1}, e_{u_2}|\theta_a, M_p) \equiv f(e_u|\theta_a, M_p).
\end{aligned}
$$

Conversely, under H_d, e_{u_1} and e_{u_2} are supposed to be independent samples drawn from the two different randomly selected sources in the alternative source population.

Therefore, the joint likelihood function for e_{u_1} and e_{u_2} under M_d is defined as

$$
\begin{aligned}
f(e_{u_1}, e_{u_2} | \theta_a, H_d) &= f_{a_{u_1}}(\mathbf{y}_{u_1 1}, \ldots, \mathbf{y}_{u_1 n_{u_1}} | \theta_a) f_{a_{u_2}}(\mathbf{y}_{u_2 1}, \ldots, \mathbf{y}_{u_2 n_{u_2}} | \theta_a) \\
&\equiv f(e_{u_1} | \theta_a, H_d) f(e_{u_2} | \theta_a, H_d) \\
&\equiv f(e_{u_1} | \theta_a, M_d) f(e_{u_2} | \theta_a, M_d) \equiv f(e_u | \theta_a, M_d).
\end{aligned}
$$

CHAPTER 4

Methods for Computing Bayes Factors

Disclaimer: This chapter is based largely on Ommen et al. [55].

Recent developments in forensic science have led to a proliferation of methods for quantifying the probative value of evidence by constructing a Bayes Factor that allows a decision-maker to select between the prosecution and defense models. Unfortunately, the analytical form of a Bayes Factor is often computationally intractable. A typical approach in statistics uses Monte Carlo integration to numerically approximate the marginal likelihoods composing the Bayes Factor. This chapter focuses on developing a generally applicable method for characterizing the numerical error associated with Monte Carlo integration techniques used in constructing the Bayes Factor. The derivation of an asymptotic Monte Carlo standard error for the Bayes Factor will be presented and its applicability to quantifying the value of evidence will be explored using a simulation-based example involving a benchmark dataset. The simulation will also explore the effect of prior choice on the Bayes Factor approximations and corresponding Monte Carlo standard errors.

4.1   Monte Carlo Integration

This subsection will explain the details of using Monte Carlo integration methods to approximate the Bayes Factor. The computational complexity of the Mixture method is sufficiently prohibitive that it will not be used in the simulation study. Therefore,

the details of its implementation will not be included in this section, but can be found in Newton and Raftery [53]. The following compact notation will be used to denote a set of sampled parameter values throughout the subsection: $\{\theta^{(i)}\} = \theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(n)}$.

Recall from Equation 3.11 that the Bayes Factor for the specific source identification problem is given by

$$BF_{ss} = \frac{\int f(e_u|\theta_s, M_p)\, d\Pi(\theta_s|e_s)}{\int f(e_u|\theta_a, M_d)\, d\Pi(\theta_a|e_a)}.$$

This Bayes Factor can be approximated using the Arithmetic Mean method of Monte Carlo integration by

$$\widehat{BF}_{ss}^{(1)} = \frac{n_p^{-1} \sum_{i=1}^{n_p} f(e_u|\theta_s^{(i)}, M_p)}{n_d^{-1} \sum_{j=1}^{n_d} f(e_u|\theta_a^{(j)}, M_d)} \tag{4.1}$$

where $\{\theta_s^{(i)}\}$ is an independent sample of size n_p drawn from $\pi(\theta_s|e_s)$ and the $\{\theta_a^{(j)}\}$ is an independent sample of size n_d drawn from $\pi(\theta_a|e_a)$. The Bayes Factor approximation for BF_{ss} by method of Harmonic Mean Monte Carlo integration is given by

$$\widehat{BF}_{ss}^{(2)} = \frac{\left[n_p^{-1} \sum_{i=1}^{n_p} f(e_u|\theta_s^{(i)}, M_p)^{-1}\right]^{-1}}{\left[n_d^{-1} \sum_{j=1}^{n_d} f(e_u|\theta_a^{(j)}, M_d)^{-1}\right]^{-1}} \tag{4.2}$$

where $\{\theta_s^{(i)}\}$ is an independent sample of size n_p drawn from the importance sampling function $\pi(\theta_s|e_s, e_u, M_p) = f(e_u|\theta_s, M_p)\pi(\theta_s|e_s)/f(e_u|e_s, M_p)$ and $\{\theta_a^{(j)}\}$ is an independent sample of size n_d drawn from the importance sampling function $\pi(\theta_a|e_a, e_u, M_d) = f(e_u|\theta_a, M_d)\pi(\theta_a|e_a)/f(e_u|e_a, M_d)$.

Recall from Equation 3.7 that the Bayes Factor for the common source identifica-

tion problem can be written as

$$BF_{cs} = \frac{\int f(e_{u_1}, e_{u_2}|\theta_a, M_p) \ d\Pi(\theta_a|e_a)}{\int f(e_{u_1}|\theta_a, M_d) \ f(e_{u_2}|\theta_a, M_d) \ d\Pi(\theta_a|e_a)}.$$

This Bayes Factor can be approximated using the Arithmetic Mean method of Monte Carlo integration by

$$\widehat{BF}_{cs}^{(1)} = \frac{n_p^{-1} \sum_{i=1}^{n_p} f(e_{u_1}, e_{u_2}|\theta_a^{(i)}, M_p)}{n_d^{-1} \sum_{j=1}^{n_d} f(e_{u_1}|\theta_a^{(j)}, M_d) \ f(e_{u_2}|\theta_a^{(j)}, M_d)} \tag{4.3}$$

where $\{\theta_a^{(i)}\}$ is an independent sample of size n_p drawn from $\pi(\theta_a|e_a)$ and $\{\theta_a^{(j)}\}$ is an independent sample of size n_d drawn from $\pi(\theta_a|e_a)$. The Bayes Factor approximation for BF_{cs} by method of Harmonic Mean Monte Carlo integration is given by

$$\widehat{BF}_{cs}^{(2)} = \frac{\left[n_p^{-1} \sum_{i=1}^{n_p} f(e_{u_1}, e_{u_2}|\theta_a^{(i)}, M_p)^{-1} \right]^{-1}}{\left[n_d^{-1} \sum_{j=1}^{n_d} \left(f(e_{u_1}|\theta_a^{(j)}, M_d) \ f(e_{u_2}|\theta_a^{(j)}, M_d) \right)^{-1} \right]^{-1}} \tag{4.4}$$

where $\{\theta_a^{(i)}\}$ is an independent sample of size n_p drawn from the importance sampling function $\pi(\theta_a|e_a, e_{u_1}, e_{u_2}, M_p) = f(e_{u_1}, e_{u_2}|\theta_a, M_p)\pi(\theta_a|e_a)/f(e_{u_1}, e_{u_2}|e_a, M_p)$ and $\{\theta_a^{(j)}\}$ is an independent sample of size n_d drawn from the importance sampling function $\pi(\theta_a|e_a, e_{u_1}, e_{u_2}, M_d) = f(e_{u_1}, e_{u_2}|\theta_a, M_d)\pi(\theta_a|e_a)/f(e_{u_1}, e_{u_2}|e_a, M_d)$.

If any of the distributions from which parameter values need to drawn at random do not have closed for solutions, Gibbs sampling algorithms will be needed in order to sample from those distributions [70]. However, the resulting sample of $\theta^{(i)}$'s from the Gibbs sampling algorithm will only be approximately independent and will require a sample size correction to account for this dependence.

4.2 Monte Carlo Standard Errors

Suppose that the value of evidence for the identification of source problem is given by

$$V_{BF} = \frac{\int f(e|\theta, H_p) \, d\Pi(\theta|H_p)}{\int f(e|\theta, H_d) \, d\Pi(\theta|H_d)} \equiv \frac{f_p(e)}{f_d(e)}.$$

Also, suppose that $f_p(e)$ and $f_d(e)$ are estimated independently using Monte Carlo integration methods by $\hat{f}_p(e)$ and $\hat{f}_d(e)$, respectively, where the corresponding approximate MCSEs are denoted ϵ_p and ϵ_d, respectively. Therefore, let

$$\hat{V}_{BF} = \frac{\hat{f}_p(e)}{\hat{f}_d(e)}$$

be the approximation of the value of evidence via Monte Carlo integration in both the numerator and denominator. Assume that if a Gibbs sampler [70] is used to sample the parameter values for the Monte Carlo integration, that the Gibbs sampler has reached a stable state and is producing an independent sample of parameter values for θ. By Lemmas 2.14, 2.15, and 2.16, the Monte Carlo integration estimates converge to the actual value of the integral,

$$\hat{f}_p(e) \xrightarrow{as} f_p(e), \quad \text{as} \quad n_p \to \infty \tag{4.5a}$$

and

$$\hat{f}_d(e) \xrightarrow{as} f_d(e), \quad \text{as} \quad n_d \to \infty, \tag{4.5b}$$

where n_p and n_d are the respective number of Monte Carlo samples used to compute $\hat{f}_p(e)$ and $\hat{f}_d(e)$ [36, 65]. However, for the Harmonic Mean estimate, Neal suggests that the convergence is very slow and unstable since "the number of points required for this estimate to get close to the right answer will often be greater than the number of atoms in the observable universe" [52]. A major advantage of the Mixture estimate

over the Harmonic Mean estimate is that it has a more stable convergence as a function of n_p or n_d [44]. A disadvantage of the Arithmetic Mean estimate is that, typically, most of the values of $\theta^{(i)}$ will be small, causing a large approximate MCSE, and very slow convergence to the actual value of the integral [44].

Now, let σ_p^2 and σ_d^2 be defined such that

$$\sqrt{n_p}\,[\hat{f}_p(e) - f_p(e)] \rightsquigarrow N(0, \sigma_p^2), \quad \text{as} \quad n_p \to \infty \tag{4.6a}$$

and

$$\sqrt{n_d}\,[\hat{f}_d(e) - f_d(e)] \rightsquigarrow N(0, \sigma_d^2), \quad \text{as} \quad n_d \to \infty \tag{4.6b}$$

according to the Central Limit Theorem [65] (see Geweke [36] for the proofs of these results). Equations 4.5a, 4.5b, 4.6a, and 4.6b will be used in the following derivation to show the asymptotic normality (according to the definition presented in Serfling [65]) of the difference between the approximated value of evidence, denoted $\hat{V}_{BF}$, via Monte Carlo integration and the actual value of evidence.

$$
\begin{aligned}
\hat{V}_{BF} - V_{BF} &= \frac{\hat{f}_p(e)}{\hat{f}_d(e)} - \frac{f_p(e)}{f_d(e)} \\
&= \frac{\hat{f}_p(e)}{\hat{f}_d(e)} - \frac{f_p(e)}{\hat{f}_d(e)} + \frac{f_p(e)}{\hat{f}_d(e)} - \frac{f_p(e)}{f_d(e)} \\
&= \frac{\hat{f}_p(e) - f_p(e)}{\hat{f}_d(e)} + \frac{f_p(e)}{f_d(e)}\left(\frac{f_d(e) - \hat{f}_d(e)}{\hat{f}_d(e)}\right) \\
&= \frac{\sqrt{n_p}\,[\hat{f}_p(e) - f_p(e)]}{\sqrt{n_p}\,\hat{f}_d(e)} + \frac{f_p(e)}{f_d(e)}\left(\frac{\sqrt{n_d}\,[f_d(e) - \hat{f}_d(e)]}{\sqrt{n_d}\,\hat{f}_d(e)}\right)
\end{aligned}
$$

This suggests that $\hat{V}_{BF} - V_{BF}$ is asymptotically normal,

$$\mathcal{AN}\left(0, \frac{\sigma_p^2/n_p}{[f_d(e)]^2} + \frac{[f_p(e)]^2\,\sigma_d^2/n_d}{[f_d(e)]^4}\right).$$

This implies that the appropriate asymptotic MCSE for the approximate Bayes Factor will take the form

$$\epsilon_V = \frac{\sqrt{[\epsilon_p \ \hat{f}_d(e)]^2 + [\epsilon_d \ \hat{f}_p(e)]^2}}{[\hat{f}_d(e)]^2} \tag{4.7}$$

since ϵ_p and ϵ_d are estimators of $\sigma_p/\sqrt{n_p}$ and $\sigma_d/\sqrt{n_d}$, respectively, and $\hat{f}_p(e)$ and $\hat{f}_d(e)$ are the approximations of $f_p(e)$ and $f_d(e)$, respectively.

The effective sample size of a Gibbs sampler represents the equivalent number of independent samples needed to achieve the MCSE corresponding to the dependent sample obtained and is defined in terms of the auto-correlation function values for a sample [35]. Let ρ_k denote the sample auto-correlation function value at lag k and let n denote the Monte Carlo sample size. Then the effective sample size, denoted n^*, is defined by [35] as

$$n^* = \frac{n}{1 + 2\sum_{k=1}^{N} \rho_k},$$

for N sufficiently large. Using the effective sample size to update Equation 4.7, define the corrected MCSE to be

$$\epsilon_V^* = \frac{\sqrt{[\epsilon_p^* \ \hat{f}_d(e)]^2 + [\epsilon_d^* \ \hat{f}_p(e)]^2}}{[\hat{f}_d(e)]^2}, \tag{4.8}$$

where $\epsilon_p^* = \epsilon_p\sqrt{n_p/n_p^*}$ and $\epsilon_d^* = \epsilon_d\sqrt{n_d/n_d^*}$ are the corrected MCSE for $\hat{f}_p(e)$ and $\hat{f}_d(e)$, respectively. It should be noted that the auto-correlation function values at each lag which are smaller than the upper-confidence bound (the threshold for acceptable thinning) are set to zero. This will ensure that the effective sample size is bounded above by the Monte Carlo sample size, $1 \leq n^* \leq n$, achieving the upper bound only when the sample is determined to be independent by having no auto-correlation.

4.3 Simulation Study

A simulation study was designed using a dataset of measurements on glass fragments with the purpose of characterizing the asymptotic MCSE for the Bayes Factors using both the Arithmetic Mean and the Harmonic Mean techniques and studying the effect of prior choice on the outcome of the numerical procedure. Due to the computationally intensive nature of the iterative scheme needed to compute the Mixture technique, this method was not implemented for the simulation study. The simulation will be divided into six distinct scenarios which will be described in the following subsection. For each of the scenarios, two different values of evidence are calculated, given by Equation 3.7 and Equation 3.11, using each of the two different Monte Carlo integration techniques, given by Equation 2.13 and Equation 2.17 (four approximate values of evidence in total given by Equations 4.1-4.4) along with their corresponding asymptotic MCSE, given by Equation 2.14 and Equation 2.18. All four of these values of evidence were calculated using two different groups of windows to suggest the values of the hyperparameters for the prior distributions, described in the subsection to follow. Each one of the values of evidence for a given scenario is computed 30 different times using a Monte Carlo sample size of 1000 for n_p and n_d. The asymptotic MCSE for a given method is measured using the average of the 30 MCSE values for each corresponding Bayes Factor which is compared to the empirical standard error for the Bayes Factor as measured by the standard deviation in the 30 estimated Bayes Factors. The algorithm for the simulation study is given in Algorithm 2.

4.3.1 Application to Forensic Evidence

The dataset used in the simulation consists of measurements made on three groups of glass fragments from 62 different window panes. This dataset was collected by Dr.

Algorithm 2: Bayes Factor Simulation Study

---

for *each scenario $S \in \{Hp,\ CNM,\ 1stQ,\ Med,\ 3rdQ,\ Max\}$* **do**

    Choose the evidence e from window group 1 according to scenario S;

    for *each window group $W \in \{2,\ 3\}$* **do**

        Choose group W to suggest the prior hyperparameters;

        for *each MC integration method $M \in \{Arithmetic,\ Harmonic\}$ Mean* **do**

            Set $n_p = 1000$ and $n_d = 1000$;

            for $i = 1$ **to** 30 **do**

                for *each definition of V_{BF} in Equation $\{3.7,\ 3.11\}$* **do**

                    Compute $\hat{V}_{BF}^{(i)}$ using method M;

                    Compute the corresponding $\epsilon_V^{(i)}$;

                end

            end

            for *each definition of V_{BF} in Equation $\{3.7,\ 3.11\}$* **do**

                Compute the empirical standard error by $sd(\hat{V}_{BF}^{(1)}, \ldots, \hat{V}_{BF}^{(30)})$;

                Compute the asymptotic standard error by $mean(\epsilon_V^{(1)}, \ldots, \epsilon_V^{(30)})$;

            end

        end

    end

end
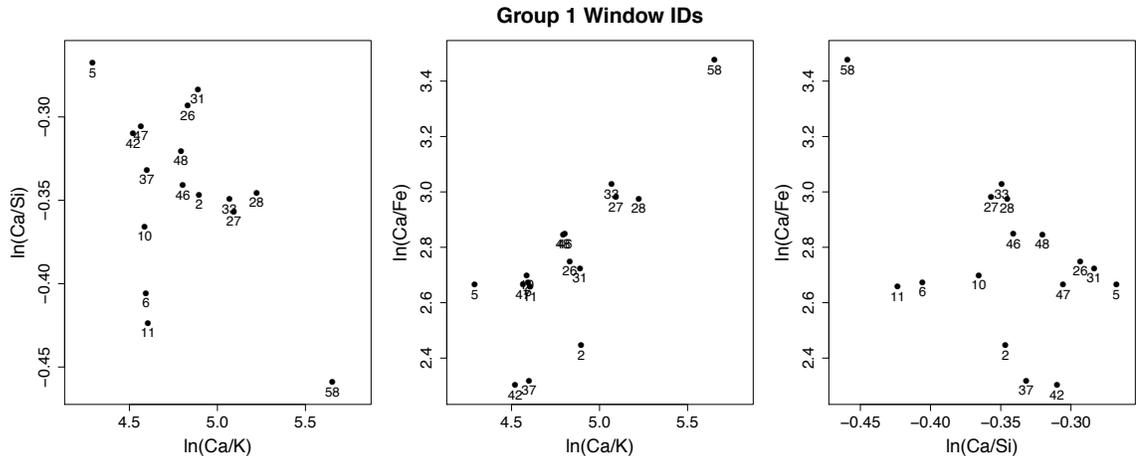
Figure 4.1: Pairwise plots of the overall mean elemental concentrations for each window in the first group of the FBI glass data along with the window identification number.

JoAnn Buscaglia of the Federal Bureau of Investigation Laboratory Division, analyzed in Aitken and Lucy, and publicly available online through the Journal of the Royal Statistical Society [1]. There are 16 windows in the first group, 16 in the second group, and 30 in the third group from which 5 fragments per window are measured (310 total fragments). The trace elemental composition of four different elements (which are thought to be the most discriminative [1]), calcium(Ca), potassium(K), silicon(Si), and iron(Fe), are measured for each of the fragments. As described in Aitken and Lucy [1], the values used for the analysis are the natural logarithms of the three ratios, Ca/K, Ca/Si, and Ca/Fe. The first group of windows was chosen for the evidence datasets in the simulation. Figure 4.1 shows the mean of the fragments for each window in the first group along with each window's identification number.

For the simulation, each window in the first group is used as the specific source for each of six different choices for the trace evidence window. The six different windows for the source of the trace evidence were chosen by first computing the Euclidean distance matrix containing all the pairwise distances between the window means (plotted in Figure 4.1). Then the distances for each window in turn were ordered from least to greatest. Under the prosecution hypothesis, denoted scenario 'Hp', the window itself was chosen as the trace evidence window. The next closest window to each specific source window was chosen for the closest non-match scenario, denoted 'CNM', the first quartile window was chosen for the '1stQ' scenario, the median window was chosen for the 'Med' scenario, the third quartile window was chosen for the '3rdQ' scenario, and finally the window farthest away from the specific source window was chosen for the 'Max' scenario. For all six of the scenarios, the remaining windows serve as the sources in the alternative source population. Figure 4.2 shows the values of the measurements on each of the fragments for each of the windows in group 1, plotted as the gray dots. The mean value of the measurements for each of the windows is plotted in black to give an idea of the variation of the fragments within each window. The
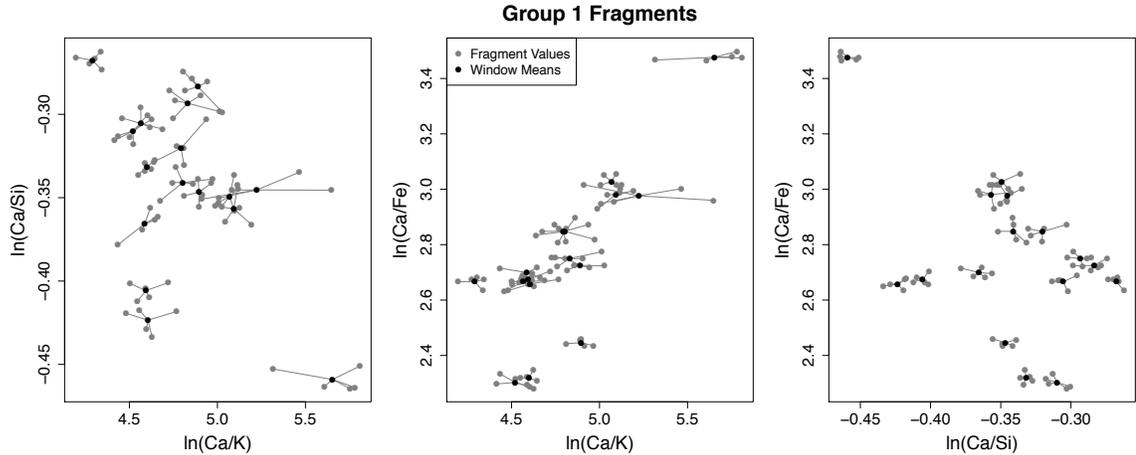
Figure 4.2: Pairwise plots of the elemental concentrations for each fragment in the first group of FBI glass windows along with the mean elemental concentration for each associated window.

Table 4.1: Test Results for Departures from Normality for Group 1 of the FBI Glass Data

Method	Component	Test Statistic	p-Value
Shapiro-Wilk	1	0.9617	0.01709
Shapiro-Wilk	2	0.9952	0.9928
Shapiro-Wilk	3	0.9906	0.8300
Fisher's Method		8.525635	0.7979345

fragments belonging to the window corresponding to the plotted mean are connected to that mean point with a line.

When the three-dimensional vectors of the measurements for the elemental compositions in the first group was tested for multivariate normality, no significant departures from a multivariate Normal distribution were observed. First, the data was mean-centered and then rotated using principal component analysis [33]. The resulting principal component scores were tested individually for departures from univariate normality using the Shapiro-Wilk test, see Saunders [62] for details. Finally, the p-values were combined using Fisher's method [32]. The results of the tests are given in Table 4.1.

Therefore, we will assume the following models for the evidence datasets. First,

define the model for $e_s = (\mathbf{y}_{s1}, \mathbf{y}_{s2}, \mathbf{y}_{s3})$ under the specific source problem, where $\mathbf{y}_{si}$ is the three-dimensional (column) vector of elemental compositions described above, as

$$\mathbf{y}_{si} \stackrel{iid}{\sim} \mathcal{N}_3(\mu_s, \Sigma_s) \tag{4.9a}$$

where $N_3(\mu_s, \Sigma_s)$ denotes a three-dimensional multivariate normal distribution with mean μ_s, a three-dimensional (column) vector, and variance Σ_s, a 3×3 positive definite covariance matrix. The following priors are given for μ_s and Σ_s:

$$\mu_s \sim \mathcal{N}_3(\mu_\pi, \Sigma_b) \quad \text{and} \quad \Sigma_s \sim \mathcal{W}_3^{-1}(\Sigma_e, \nu_e) \tag{4.9b}$$

where μ_π is a three-dimensional vector, Σ_b and Σ_e are 3×3 positive definite covariance matrices, and the number of degrees of freedom ν_e is a scalar which must be at least as large as the dimension. The multivariate normal distribution is typically used as the conjugate prior for the mean of a multivariate normal distribution [35]. The inverse Wishart distribution is typically used as the conjugate prior for covariance matrices of a multivariate normal distribution [35].

Next, for both the specific source and common source problems, define the alternative source population evidence as $e_a = (Y_{a_1}, Y_{a_2}, \ldots, Y_{a_{16}})$, where $Y_{a_i} = (\mathbf{y}_{i1}, \mathbf{y}_{i2}, \ldots, \mathbf{y}_{i5})$ is the (column) vector of measurements from a single source, and $\mathbf{y}_{ij}$ is the three-dimensional (column) vector of elemental compositions for the jth fragment from the ith window. Define the hierarchical random effects model for e_a, under both the specific source and common source problems to be

$$\mathbf{y}_{ij} = \mu_a + a_i + w_{ij} \tag{4.10a}$$

with

$$a_i \stackrel{iid}{\sim} \mathcal{N}_3(0, \Sigma_a) \quad \text{and} \quad w_{ij} \stackrel{iid}{\sim} \mathcal{N}_3(0, \Sigma_w) \tag{4.10b}$$

where μ_a is the three-dimensional grand mean vector of the elemental compositions, a_i is the deviation from the grand mean vector for the ith window in e_a, and w_{ij} is the deviation for the jth fragment from the mean vector for the ith window. In Section 2.1, a result presented by Miller [49] shows that

$$\mathbf{y}_{ci} \overset{iid}{\sim} \mathcal{N}_{15}(\mu_c, \Sigma_c) \tag{4.10c}$$

where μ_c is a block vector whose elements are functions of μ_a and Σ_c is a block matrix whose elements are functions of Σ_a and Σ_w. The priors for μ_a, Σ_a, and Σ_w are as follows:

$$\mu_a \sim \mathcal{N}_3(\mu_\pi, K\Sigma_b), \quad \Sigma_a \sim \mathcal{W}_3^{-1}(\Sigma_b, \nu_b), \quad \text{and} \quad \Sigma_w \sim \mathcal{W}_3^{-1}(\Sigma_e, \nu_e) \tag{4.11}$$

where μ_π is a three-dimensional vector, Σ_b and Σ_e are 3×3 positive definite covariance matrices, the number of degrees of freedom ν_b and ν_e are scalars which must be at least as large as the dimension, and K is a scalar.

Now, under the prosecution hypothesis for the specific source problem, $e_u = (\mathbf{y}_{u1}, \mathbf{y}_{u2})$ follows the same model as e_s. Conversely, under the defense hypothesis for the specific source problem, e_u follows the same model as e_a. For the common source problem, e_{u1} and e_{u2} both follow the same model as e_a under both hypotheses. However, under the prosecution hypothesis they are considered to be dependent samples from the same source in the alternative source population model, whereas under the defense hypothesis they are considered to be independent samples from two different sources in the alternative source population model. See Section 3.4.3 and Section 3.4.2 for further details on the models.

For the simulation study, the values for the prior hyperparameters have been determined from the second or third group of windows. Let the second and third group

of the data have the same structure as the alternative source population given in Section 3.4, where $\mathbf{y}_{kij}$ denotes the jth fragment from the ith source in the kth group (either the second or third group of windows) for $i = 1, \ldots, n_{g_k}$, $j = 1, \ldots, n_f$, and where n_{g_k} is the number of windows in the kth group of windows $(k = 2, 3)$ and n_f is the number of fragments from each window. Then

$$\bar{\mathbf{y}}_{ki} = \frac{1}{n_f} \sum_{j=1}^{n_f} \mathbf{y}_{kij} \tag{4.12a}$$

$$\bar{\mathbf{y}}_k = \frac{1}{n_{g_k}} \sum_{i=1}^{n_{g_k}} \bar{\mathbf{y}}_{ki} \tag{4.12b}$$

$$S_{kiw} = \frac{1}{n_f - 1} \sum_{j=1}^{n_f} (\mathbf{y}_{kij} - \bar{\mathbf{y}}_{ki})(\mathbf{y}_{kij} - \bar{\mathbf{y}}_{ki})^T \quad \text{for } i = 1, \ldots, n_{g_k} \tag{4.12c}$$

$$S_{kb} = \frac{1}{n_{g_k} - 1} \sum_{i=1}^{n_{g_k}} (\bar{\mathbf{y}}_{ki} - \bar{\mathbf{y}}_k)(\bar{\mathbf{y}}_{ki} - \bar{\mathbf{y}}_k)^T \tag{4.12d}$$

Then, μ_π is a three-dimensional vector representing the grand mean of the elemental compositions for the windows, Σ_b is the estimated 3×3 between group covariance matrix, and Σ_e is the estimated 3×3 within group covariance matrix as described in Aitken and Lucy [1] and given below.

$$\mu_{\pi_k} = \bar{\mathbf{y}}_k \tag{4.13a}$$

$$\Sigma_{e_k} = \frac{1}{n_{g_k}} \sum_{i=1}^{n_{g_k}} S_{kiw} \tag{4.13b}$$

$$\Sigma_{b_k} = S_{kb} - \frac{1}{n_f} \Sigma_{e_k} \tag{4.13c}$$

The number of degrees of freedom for the inverse Wishart distributions ν_b and ν_e are scalars that take the values $3, 9, 27$, and $K = 10$. The degrees of freedom are chosen to begin at the smallest possible value 3 and increase exponentially (i.e. 3^1, 3^2, 3^3, ...). For simplicity of the simulation, we chose to set $\nu_b = \nu_e$, however, this may not always be the case in practice. (One situation in which it may be appropriate to
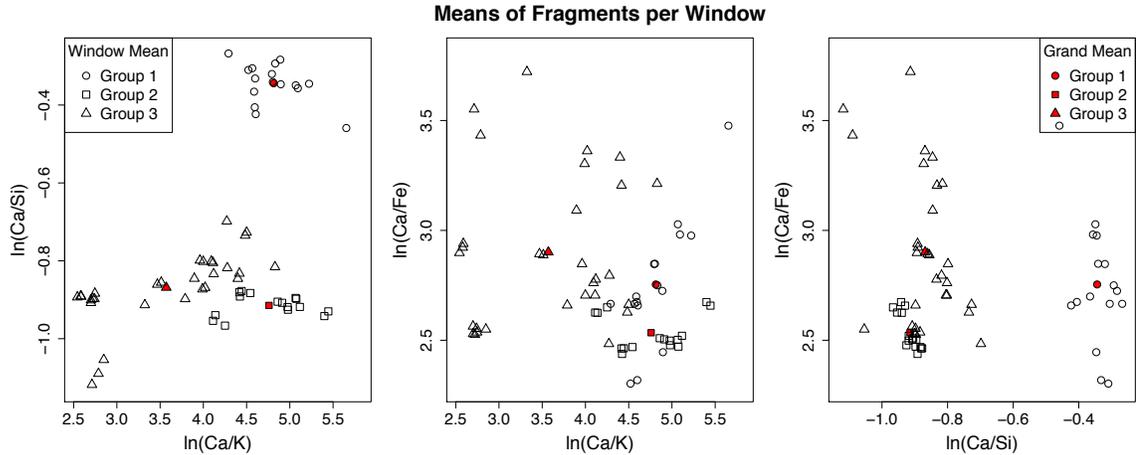
Figure 4.3: Pairwise plots of the mean elemental concentrations for each window in the FBI glass data and the grand mean elemental concentrations for each group in all three groups of windows.

use $\nu_b = \nu_e$ is when exploring a new forensic modality, such as copper wire [21], in which there is little prior information concerning the degrees of freedom. The degrees of freedom control the amount of information contained in the prior distribution about the covariance matrices.) The value of K was chosen to make the prior for μ_a less informative (or more spread out). It should be noted that the priors chosen are relatively restrictive priors, not non-informative or flat priors. These priors have the effect of shifting the grand mean of the first group towards the grand mean of whichever group is being used for the prior (either group 2 or 3). The grand means are plotted in Figure 4.3, which also provides a rough visualization of the between-window covariance.

The first and second groups have similar between-window covariances, so, when the second group is used as the prior there is not much effect on this variable. However, the third group has a much larger between-window covariance effectively widening the distribution for the first group when it is used in the prior. Figure 4.4 provides a visual representation of the within-window covariance for each of the groups. The third group has the smallest within-window covariance, the first group is in the middle,

and the second group has the largest within-window covariance. Therefore, when the second and third groups are used for the priors, the within-window covariance structure widens and narrows, respectively.

4.3.2 Diagnostic Summary

After running the simulation, the output from the Gibbs samplers via the 'MCM-Cglmm' function [40] in R was analyzed to see if the assumption of independent samples for the values of θ was valid. Within the 'MCMCglmm' function, a burn-in period of 1000 samples was specified to ensure that the algorithm reached a steady state and a thinning interval of 2 was used to facilitate the assumption of independence of the samples for the values of θ. The independence of the sample was measured by plotting the auto-correlation function for each sub-simulation to see whether the values fell within the recommended threshold [35]. It was determined that for the specific source sub-simulation when the prosecution model is true that the auto-correlation function values typically fell within the specified threshold value for all lag values between one and twenty. (Note that we chose to stop the lag values at 20 because theoretically the maximum lag is equal to the number of parameter values for θ, due to the Markov property, which is 9 for θ_s and 15 for θ_a [35].) This suggests that the chosen burn-in period and thinning interval values are appropriate for this example. A typical acceptable auto-correlation function plot can be seen in Figure 4.5(a).

However, for the specific source sub-simulation when the defense model is true and for both models in the common source sub-simulation, the auto-correlation function values did not fall within the desired tolerance for many of the lag values between one and twenty. A typical unacceptable auto-correlation function plot can be seen in Figure 4.5(b). This suggests that the thinning interval chosen was not appropriate for this example and should have been increased to something larger than 2 (perhaps
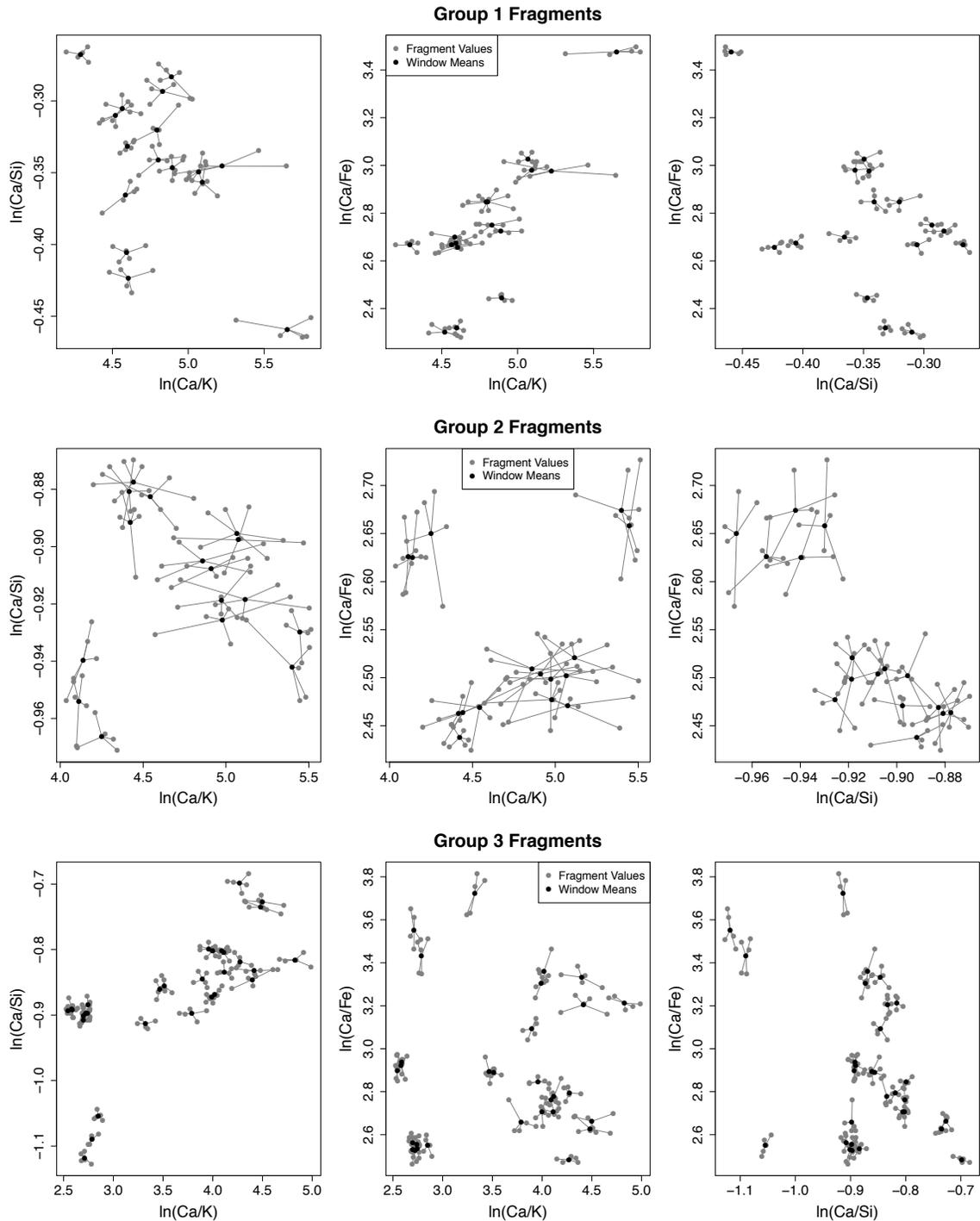
Figure 4.4: Pairwise plots of the elemental concentrations for each fragment in each window in the FBI glass data along with the mean elemental concentrations for each window in all three groups. (Note that each subplot has different scaling for the three variables.)

Figure 4.5: Typical auto-correlation function plots to determine whether the chosen thinning intervals for the Gibbs samplers are appropriate.
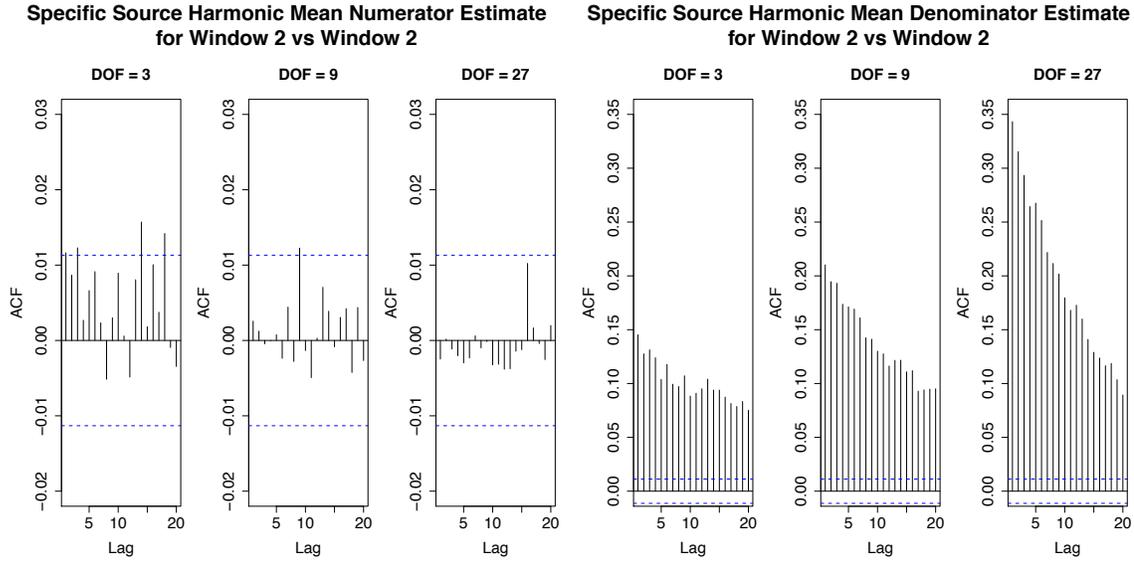
(a) A plot showing acceptable thinning inter- (b) A plot showing unacceptable thinning in-
vals.   tervals.

20 or even 50).

Another diagnostic test performed considered the trace plots for the values used in the sum-portion of the Monte Carlo integration (for the Arithmetic Mean we considered $log(f(x|\theta^{(i)}))$, and for the Harmonic Mean we considered $log(1/f(x|\theta^{(i)})))$. In general, the trace plots looked fairly well-mixed. A typical acceptable trace plot can be seen in Figure 4.6(a) distinguished by the apparent randomness between consecutively simulated values. There were some abnormal trace plots which is unsurprising given the large number of abnormal auto-correlation function plots. A typical abnormal plot is given in Figure 4.6(b) distinguished by the trending feature of consecutively simulated values. Due to the computational intensity of this simulation, the simulations were not re-run using a higher thinning interval (as suggested by the auto-correlation function plots) since many of the trace plots looked acceptable.

Figure 4.6: Typical trace plots to determine whether the chosen thinning intervals for the Gibbs samplers are appropriate.
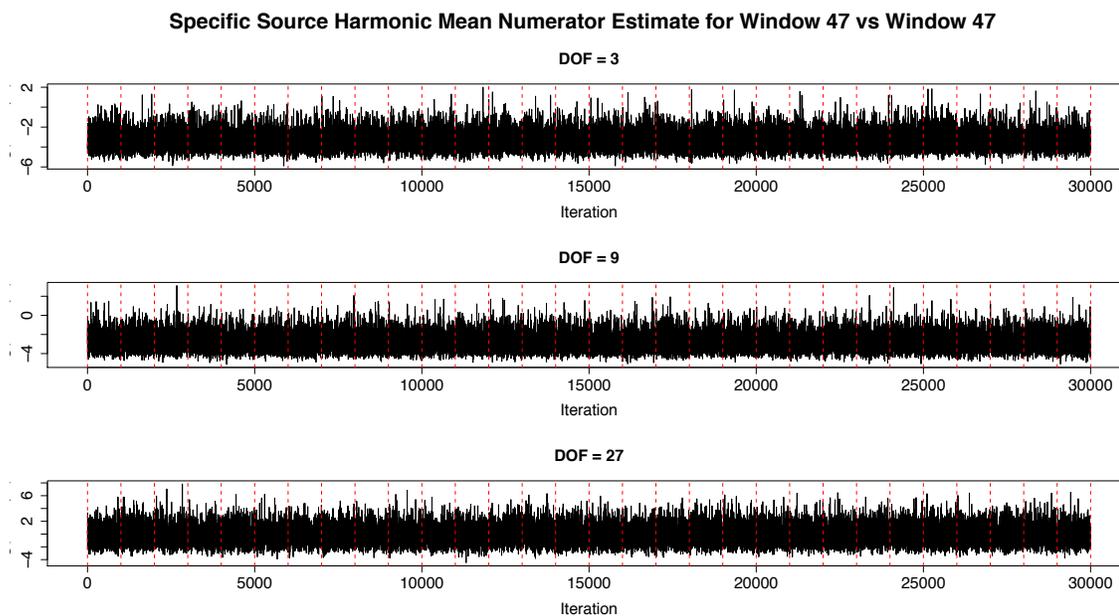
(a) Trace plot showing acceptable thinning intervals.
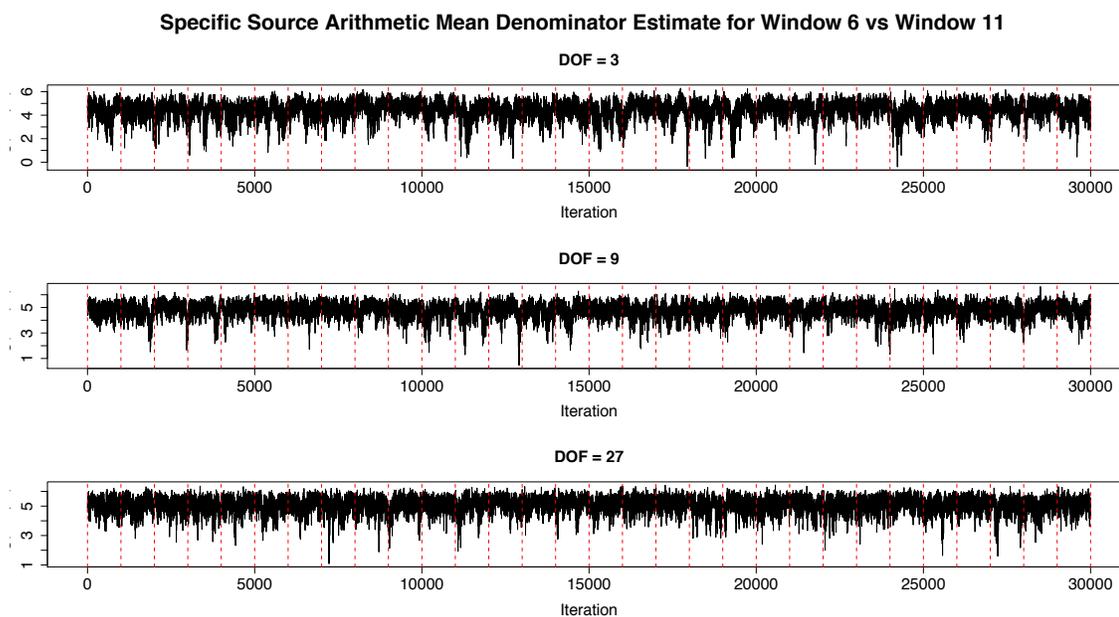
(b) Trace plot showing unacceptable thinning intervals.

Figure 4.7: Boxplots for the relative difference between the empirical standard error in the specific source Bayes Factors and the asymptotic MCSEs comparing the Arithmetic and Harmonic Mean methods.

Figure 4.8: Boxplots for the relative difference between the empirical standard error of the common source Bayes Factors and the asymptotic MCSEs comparing the Arithmetic and Harmonic Mean methods.

Table 4.2: Corrected Monte Carlo Standard Errors

DOF	Scenario	$sd(\hat{V}_{ss})$	$\bar{\epsilon}_V$	$\bar{\epsilon}_V^*$	$\frac{sd(\hat{V}_{ss})-\bar{\epsilon}_V}{sd(\hat{V}_{ss})}$	$\frac{sd(\hat{V}_{ss})-\bar{\epsilon}_V^*}{sd(\hat{V}_{ss})}$
3	Hp	6.611725e-02	1.830668e-02	5.166504e-02	0.7231179	0.2185845
3	CNM	9.289674e-03	5.290273e-03	7.733278e-03	0.4305211	0.1675404
3	1stQ	5.187942e-02	1.741958e-02	4.206387e-02	0.6642294	0.1891992
3	Med	2.220020e-04	2.373557e-04	2.865845e-04	-0.0691604	-0.2909096
3	3rdQ	1.635434e-08	4.818058e-09	4.819576e-09	0.7053957	0.7053029
3	Max	7.536286e-18	1.383777e-18	1.386807e-18	0.8163848	0.8159826
9	Hp	1.879532e-02	8.300683e-03	1.666579e-02	0.5583643	0.1133012
9	CNM	5.774035e-03	4.149662e-03	5.676135e-03	0.2813238	0.0169551
9	1stQ	2.274740e-02	9.529237e-03	2.345237e-02	0.5810845	-0.0309915
9	Med	1.450562e-04	9.329485e-05	1.056527e-04	0.3568364	0.2716427
9	3rdQ	6.104715e-12	1.115829e-12	1.116634e-12	0.8172185	0.8170866
9	Max	2.295706e-35	4.248643e-36	4.343302e-36	0.8149309	0.8108076
27	Hp	1.700196e-03	1.006384e-03	1.458469e-03	0.4080778	0.1421756
27	CNM	6.562763e-04	5.970101e-04	7.819605e-04	0.0903068	-0.1915111
27	1stQ	2.096175e-03	1.174497e-03	2.277635e-03	0.4396950	-0.0865675
27	Med	2.952653e-06	1.522242e-06	1.631392e-06	0.4844494	0.4474828
27	3rdQ	4.060306e-23	7.522703e-24	7.526123e-24	0.8147257	0.8146415
27	Max	4.72576e-112	8.84673e-113		0.8127980	

4.3.3   Simulation Results for Standard Errors

Before placing any trust in the asymptotic MCSE, it should be determined whether these asymptotic MCSEs accurately reflect the empirical standard error. So, the boxplots in Figure 4.7 and Figure 4.8 show the relative error between the empirical standard error and the asymptotic MCSE of the Bayes Factor. Ideally, the boxplots should have medians around zero with small ranges. However, this is not the case. For both the specific source and common source sub-simulations, the error associated with the Bayes Factor tends to be underestimated. This may be a byproduct of the issues concerning the thinning interval for the Gibbs sampler. Because the auto-correlation function values indicated that the samples generated by the Gibbs sampler were not independent, it may have been more appropriate to use the effective sample size from the Gibbs sampler instead of the Monte Carlo sample size to calculate the asymptotic MCSE.

Consider, for example, the specific source sub-simulation for the second window only and using the second group of windows to suggest the priors. The corrected MCSEs

(given by Equation 4.8) were calculated for each of the 30 estimated Bayes Factors in each of the six scenarios. The corrected MCSEs are compared to the empirical standard errors and the uncorrected MCSEs (given by Equation 4.7), and the results are provided in the Table 4.2. It should be noted that the machine precision used to compute the values in Table 4.2 was 2.220446×10^{-16}, and so the values smaller than this quantity are numerically zero. In addition, the auto-correlation function values at each lag which are smaller than the upper-confidence bound (the threshold for acceptable thinning) are set to zero. This will ensure that the effective sample size is bounded above by the Monte Carlo sample size.

The corrected MCSEs were calculated for all of the Arithmetic Mean Bayes Factors (for all of the windows) in each of the six scenarios, and the results are compared to the uncorrected MCSEs in Figure 4.9. See Figure 4.10 for the corresponding results of the corrected MCSEs for all of the Harmonic Mean Bayes Factors. In general, the corrected MCSEs have the effect of increasing the asymptotic standard error (so that the corrected MCSE is not underestimating the empirical standard error as often as the uncorrected MCSE). However, for the scenarios where the trace and control samples are very far apart resulting in a very small Bayes Factor (typically the '3rdQ' and 'Max' scenarios), the standard errors are so small that the sample size correction has little effect on the standard error. In certain cases, the corrected standard errors for the numerator and denominator were numerically zero causing the asymptotic MCSE for the Bayes Factor to not exist (these entries have been omitted from Table 4.2 and from the construction of the boxplots in Figure 4.9) and Figure 4.10.
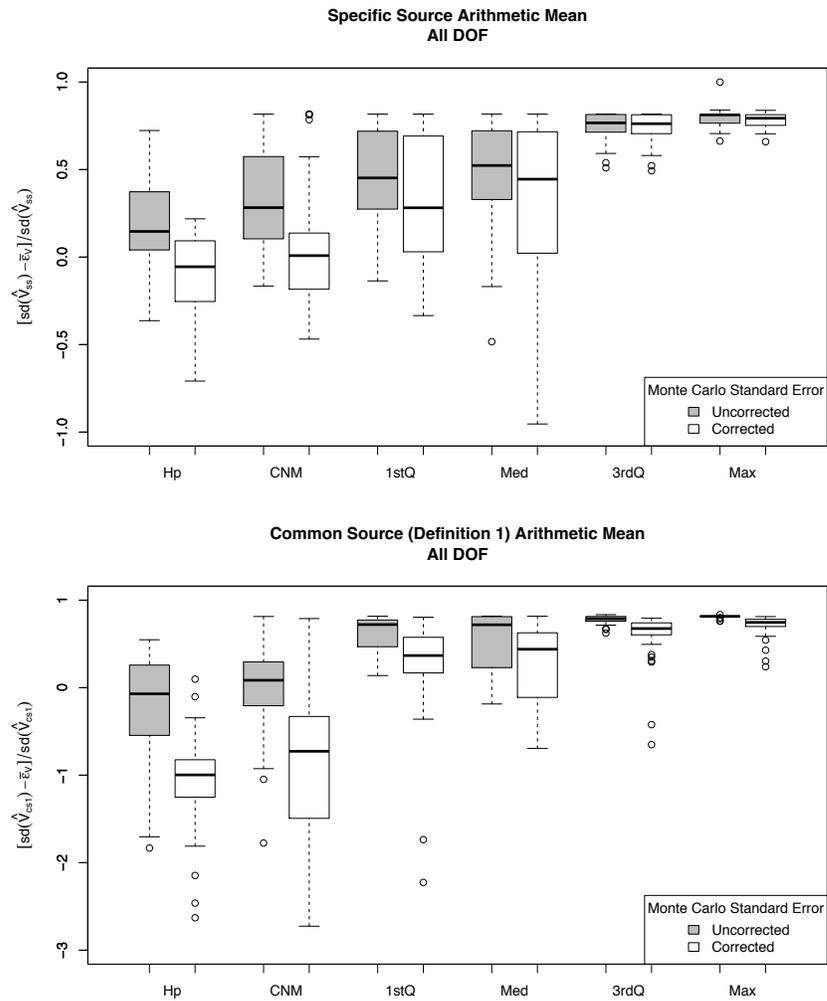
Figure 4.9: Boxplots for the relative difference between the empirical standard error in the Bayes Factors and the asymptotic MCSEs comparing the uncorrected and the corrected MCSE.
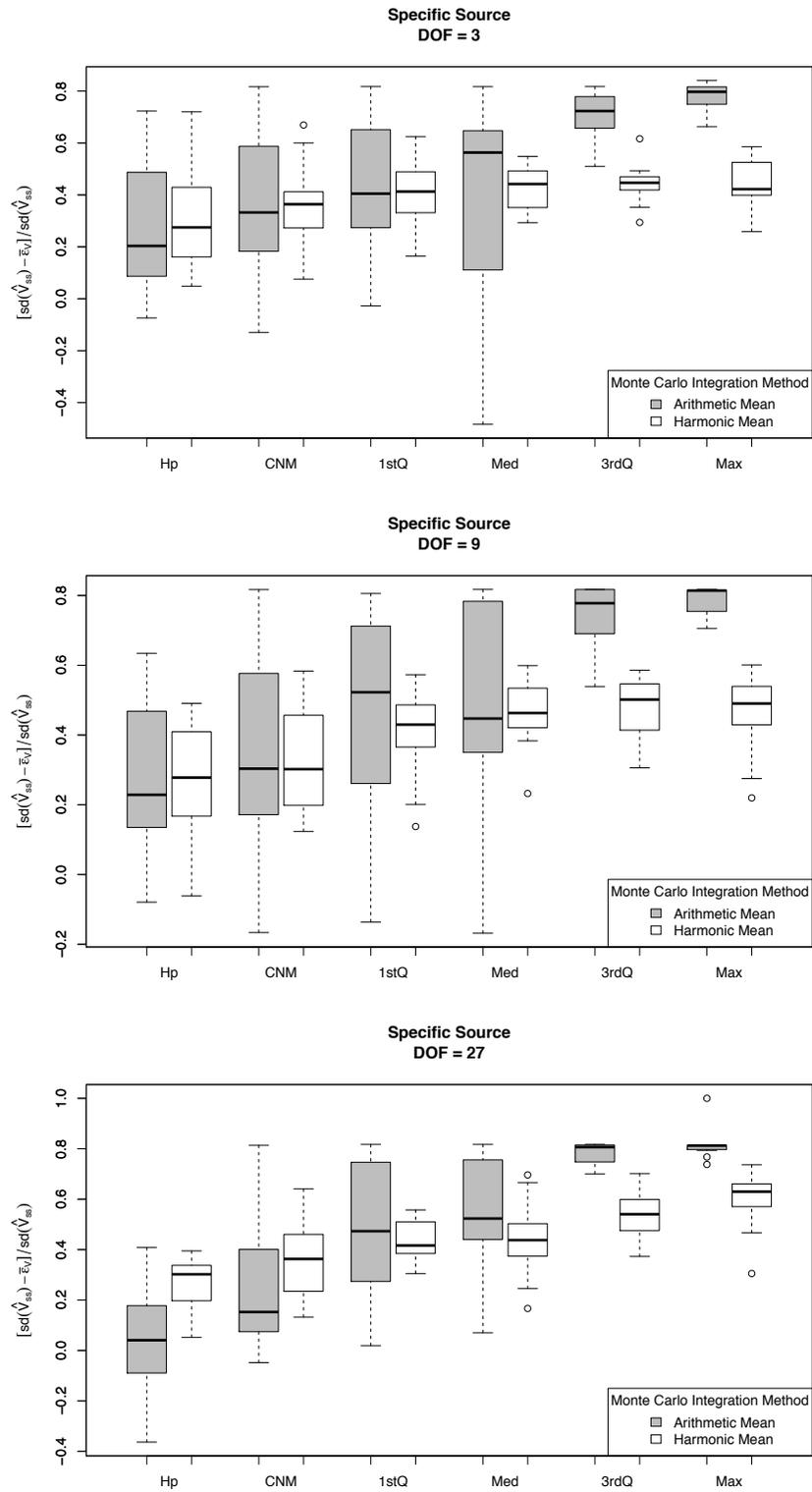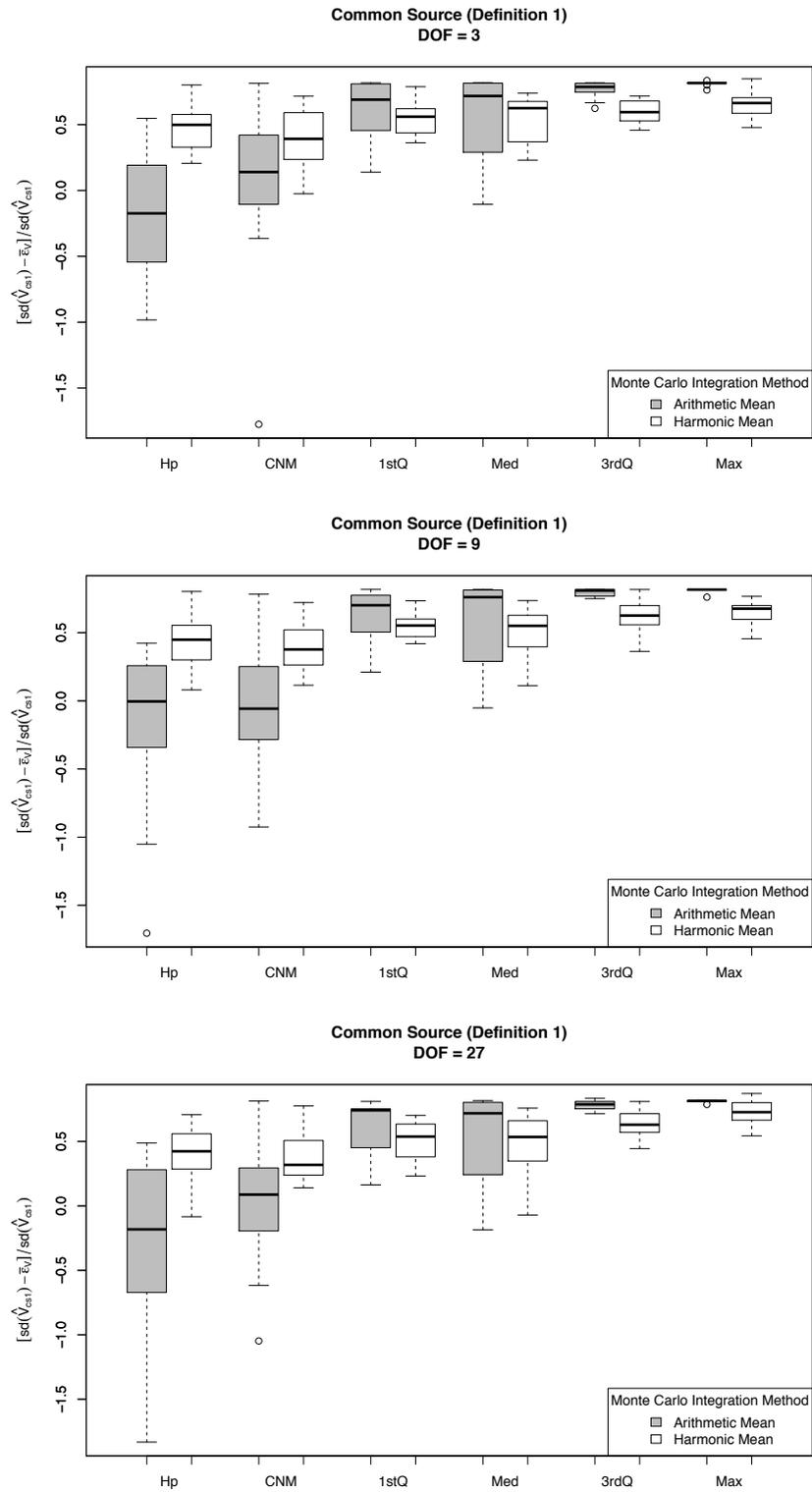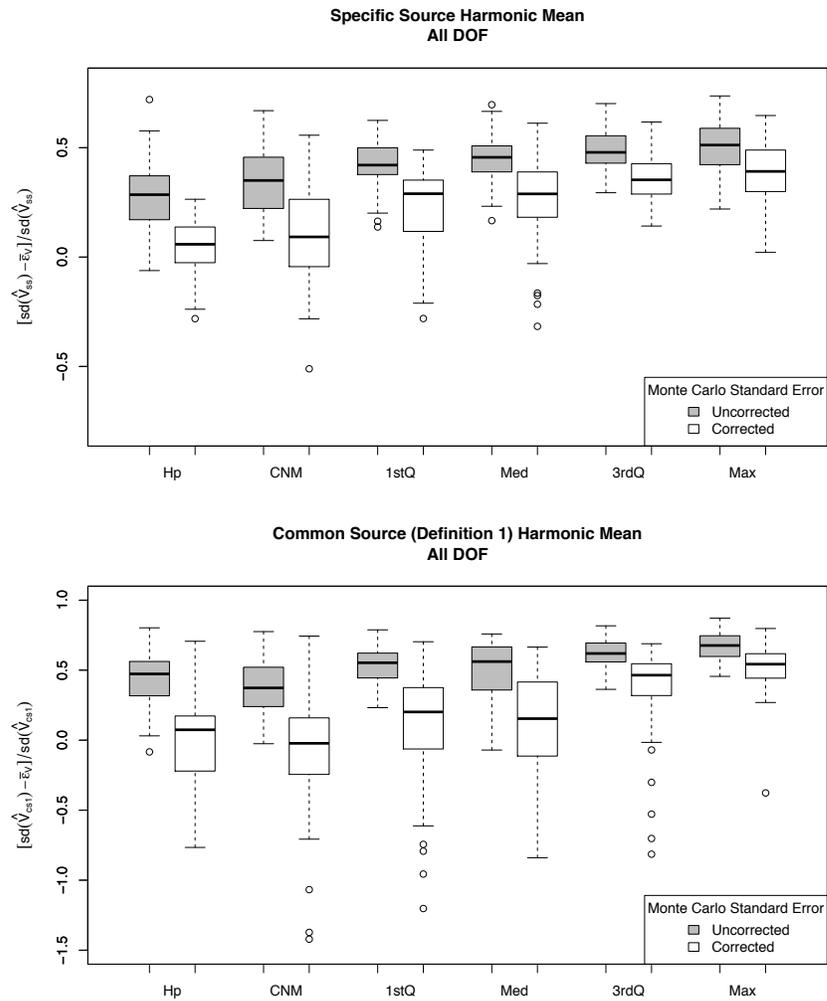
Figure 4.10: Boxplots for the relative difference between the empirical standard error in the second definition of the common source Bayes Factors and the asymptotic MCSE comparing the uncorrected and the corrected MCSE.

4.3.4 Sensitivity Analysis for Bayes Factors

A rough sensitivity analysis was performed using the results of the simulation study to determine which, if any, of the results for each of the methods were affected by changes in the prior choice and the inexactness of the numerical approximation methods used. Following Hepler et al. [42], the rates of misleading evidence (RME) and the rates of disagreement between experts were analyzed to detect possible differences in the Bayes Factors as a result of the choice of prior and of using approximation techniques. The results of this analysis are summarized in Table 4.3a and Table 4.3b.

In these tables, the 30 Bayes Factors for each method of approximating the value of evidence are summarized using the average, and then the results for all six prior choices were combined (two different groups of windows to suggest the prior hyper-parameters each with three different choices of degrees of freedom). The rates of disagreement were calculated by first determining for each of the 16 windows in each scenario ('Hp', 'CNM', '1stQ', 'Med', '3rdQ', and 'Max') if all of the Bayes Factor averages calculated for each of the six prior choices (all 6 Bayes Factor averages per scenario, total) fell into the same category for magnitude of support (Supports H_p, Inconclusive, or Supports H_d). If so, the window was labeled with its category for magnitude of support. If any one of the six Bayes Factor averages provided a different magnitude of support than the others, then the window was labeled as 'Disagree'-ing. The proportion of windows with labels in each category are summarized in Table 4.3a and Table 4.3b. For all six scenarios, the 'Disagree' category is a measurement of the sensitivity of the method to the choice of prior and numerical approximation technique. Methods with a high rate of disagreement are more sensitive to the prior choice and numerical approximation techniques, whereas, methods with a low rate of disagreement are less sensitive. For the 'Hp' scenarios, the RME gives the proportion of Bayes Factor averages which are less than 100 (the proportion that do not 'Sup-

Table 4.3: Rates of misleading evidence and rates of disagreement for the Bayes Factors between all prior combinations.

(a) Specific Source Bayes Factors

	Scenario	RME	Supports H_p $\bar{V} \geq 100$	Inconclusive $0.01 < \bar{V} < 100$	Supports H_d $\bar{V} \leq 0.01$	Disagree
	Hp	0.8229	0	0.0625	0	0.9375
	CNM	0.7500	0	0.1875	0.1250	0.6875
Arithmetic Mean	1stQ	0.2292	0	0	0.4375	0.5625
	Med	0.0729	0	0	0.6250	0.3750
	3rdQ	0	0	0	1	0
	Max	0	0	0	1	0
	Hp	0.7917	0	0.0625	0	0.9375
	CNM	0.8125	0	0.1875	0	0.8125
Harmonic Mean	1stQ	0.4271	0	0.0625	0	0.9375
	Med	0.3542	0	0	0.1250	0.8750
	3rdQ	0.2083	0	0	0.1875	0.8125
	Max	0.3021	0	0	0.1875	0.8125

(b) Common Source Bayes Factors

	Scenario	RME	Supports H_p $\bar{V} \geq 100$	Inconclusive $0.01 < \bar{V} < 100$	Supports H_d $\bar{V} \leq 0.01$	Disagree
	Hp	0.1250	0.7500	0	0	0.2500
	CNM	0.8125	0.1250	0.3750	0.1875	0.3125
Arithmetic Mean	1stQ	0	0	0	1	0
	Med	0.0833	0	0.0625	0.8750	0.0625
	3rdQ	0	0	0	1	0
	Max	0	0	0	1	0
	Hp	0.1146	0.7500	0	0	0.2500
	CNM	0.8125	0.1250	0.3750	0.1875	0.3125
Harmonic Mean	1stQ	0	0	0	1	0
	Med	0.0833	0	0.0625	0.8750	0.0625
	3rdQ	0	0	0	1	0
	Max	0	0	0	1	0

ports H_p'). For the remaining five scenarios, the RME give the proportion of Bayes Factor averages which are greater than 0.01 (the proportion that do not 'Supports H_d').

The results show comparable rates of misleading evidence and rates of disagreement between the Arithmetic Mean and Harmonic Mean methods. The specific source methods studied in the simulation have very high RME, as well as high rates of disagreement. This would indicate that the specific source values of evidence are highly sensitive to the choice of prior and to the numerical approximation technique. In

contrast, the common source methods tend to have lower RME and rates of disagreement. It should be noted that the highest RME are for the 'CNM' scenario under the common source methods. This is expected since the trace evidence will look very similar to the specific source evidence, but not actually originate from the specific source.

Due to the unexpectedly high rates of disagreement and rates of misleading evidence (especially for the specific source methods under the 'Hp' scenario), the tables were divided into two subsets based on the group of windows used to suggest the priors. These results are given in Tables 4.4a, 4.4b, and 4.4c. They have been reorganized to better compare results between the second and third group being used to suggest the prior hyperparameters. For the specific source using the Arithmetic mean method, the RME and rates of disagreement are higher for the second group than for the third group. The same is generally true for the specific source Harmonic mean method as well. For the common source methods, the tables for the Arithmetic mean and Harmonic mean methods were the same. Because they were all the same, only one table for the common source methods is presented in Table 4.4c. However, the group used to suggest the prior hyperparameters does not seem to make a difference for the common source methods as it does for the specific source methods.

Tables 4.4a and 4.4b indicate that the priors suggested by group 2 are causing the issues with the high RME seen in Table 4.3a. Further inspection of the evidence under these circumstances reveals the potential cause. Figure 4.11 shows the pairwise scatter plots of the evidence the specific source problem when the prosecution hypothesis is true along with a contour plot of the posterior predictive distribution. This figure reveals that under the specific source problem when the prosecution hypothesis is true, the prior suggested by the second group of windows for the specific source parameters, given by Equation 4.9b, is not a very good match for the specific source evidence.

Table 4.4: Rates of misleading evidence and rates of disagreement for the Bayes Factors within each prior group.

(a) Specific Source Bayes Factors using the Arithmetic Mean method

Prior	Scenario	RME	Supports H_p $\bar{V} \geq 100$	Inconclusive $0.01 < \bar{V} < 100$	Supports H_d $\bar{V} \leq 0.01$	Disagree
Group 2	Hp	0.9583	0	0.1875	0	0.8125
	CNM	0.6458	0	0.2500	0.1250	0.6250
	1stQ	0.1875	0	0.0625	0.6875	0.2500
	Med	0.0625	0	0	0.8125	0.1875
	3rdQ	0	0	0	1	0
	Max	0	0	0	1	0
Group 3	Hp	0.6875	0	0.3750	0	0.6250
	CNM	0.8542	0	0.5000	0.1250	0.3750
	1stQ	0.2708	0	0	0.5625	0.4375
	Med	0.0833	0	0	0.8125	0.1875
	3rdQ	0	0	0	1	0
	Max	0	0	0	1	0

(b) Specific Source Bayes Factors using the Harmonic Mean method

Prior	Scenario	RME	Supports H_p $\bar{V} \geq 100$	Inconclusive $0.01 < \bar{V} < 100$	Supports H_d $\bar{V} \leq 0.01$	Disagree
Group 2	Hp	0.9583	0	0.2500	0	0.7500
	CNM	0.6875	0	0.2500	0.0625	0.6875
	1stQ	0.2500	0	0.0625	0.6250	0.3125
	Med	0.1875	0	0	0.6875	0.3125
	3rdQ	0.0417	0	0	0.8750	0.1250
	Max	0.0625	0	0	0.8125	0.1875
Group 3	Hp	0.6250	0.0625	0.3750	0	0.5625
	CNM	0.9375	0	0.5625	0	0.4375
	1stQ	0.6042	0	0.1875	0.0625	0.7500
	Med	0.5208	0	0.0625	0.1250	0.8125
	3rdQ	0.3750	0	0	0.1875	0.8125
	Max	0.5412	0	0.0625	0.1875	0.7500

(c) Common Source Bayes Factors

Prior	Scenario	RME	Supports H_p $\bar{V} \geq 100$	Inconclusive $0.01 < \bar{V} < 100$	Supports H_d $\bar{V} \leq 0.01$	Disagree
Group 2	Hp	0.0417	0.9375	0	0	0.0625
	CNM	0.8125	0.4375	0.3750	0.1875	0
	1stQ	0	0	0	1	0
	Med	0.0833	0	0.0625	0.8750	0.0625
	3rdQ	0	0	0	1	0
	Max	0	0	0	1	0
Group 3	Hp	0.2083	0.7500	0.1875	0	0.0625
	CNM	0.8125	0.1250	0.6250	0.1875	0.0625
	1stQ	0	0	0	1	0
	Med	0.0833	0	0.0625	0.8750	0.0625
	3rdQ	0	0	0	1	0
	Max	0	0	0	1	0
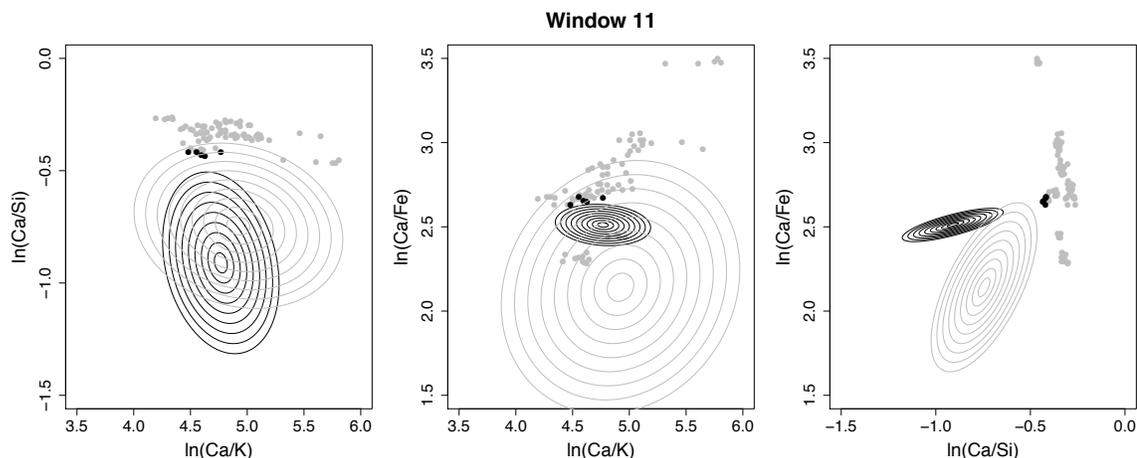
Figure 4.11: Pairwise scatter plots of the specific source evidence under the prosecution hypothesis, along with a contour plot of the posterior predictive distribution for the specific source evidence, is plotted in black. The alternative source population evidence and corresponding contour plot of the posterior predictive distribution is plotted in gray.

Even when the prior is updated with the evidence, the posterior distribution for the specific source evidence is still quite mismatched with the data. The same is true for the alternative source population evidence, but to a much lesser degree. This is due to the fact that there are many more samples in the alternative source population evidence compared with the specific source evidence. This is effectively decreasing the numerator of the Bayes Factor, even though it should be large since the prosecution hypothesis is true.

Since the previous analysis is based on the point values for the approximate Bayes Factors and does not take the approximate MCSEs into account, the results can often be misleading. In addition, it is impossible to differentiate whether the rates of misleading evidence and the rates of disagreement are due to the effect of the prior choice or due to the numerical approximation technique. The following analysis using the MCSEs will consider the sensitivity of the Bayes Factors to only the prior choice given the current level of MCSE.

Consider comparing, for each window under each scenario and approximation tech-

nique, the interval for the approximate Bayes Factors defined below for each of three different prior degrees of freedom:

$$[\bar{V} - 2\bar{\epsilon}_V^*, \bar{V} + 2\bar{\epsilon}_V^*]$$

where $\bar{V}$ represents the average of the 30 Bayes Factor approximations and $\bar{\epsilon}_V^*$ represents the square root of the mean of the squared corrected MCSEs for the 30 corresponding Bayes Factors. If the lower endpoint of the interval is negative, it is truncated to zero since Bayes Factors take only non-negative values. If the corrected MCSE does not exist, the uncorrected MCSE (which is most often zero) will be used. If all three of these intervals overlap each other, then that window is said to 'agree' for all prior choices. If any one of the intervals does not overlap another interval, that window is said to 'disagree' for all prior choices. See Figure 4.12 for an illustration of the different types of overlapping and non-overlapping intervals typically encountered during the simulation. If a large proportion of the windows under each scenario and approximation technique 'disagree', that method is sensitive to choice of the prior. It should be noted that if $\bar{\epsilon}_V^*$ is numerically zero for more than one comparable Bayes Factor (which is good in the sense that your estimate of the Bayes Factors do not contain any significant amount of numerical error), then the interval above will become essentially a point value. In this situation, this method of sensitivity analysis will not be the ideal method since it is unlikely that several point values will be exactly the same (meaning the 'intervals' will not overlap, even if the three point values are very similar). The proportion of windows that 'disagree' for all prior choices considered under each of the approximation scenarios is presented in Table 4.5. Table 4.5 is for the Arithmetic mean and Harmonic mean methods and for the group 2 priors only.

It should be noted that the levels of disagreement in Table 4.5 are a direct reflection

Potential Overlap Scenarios
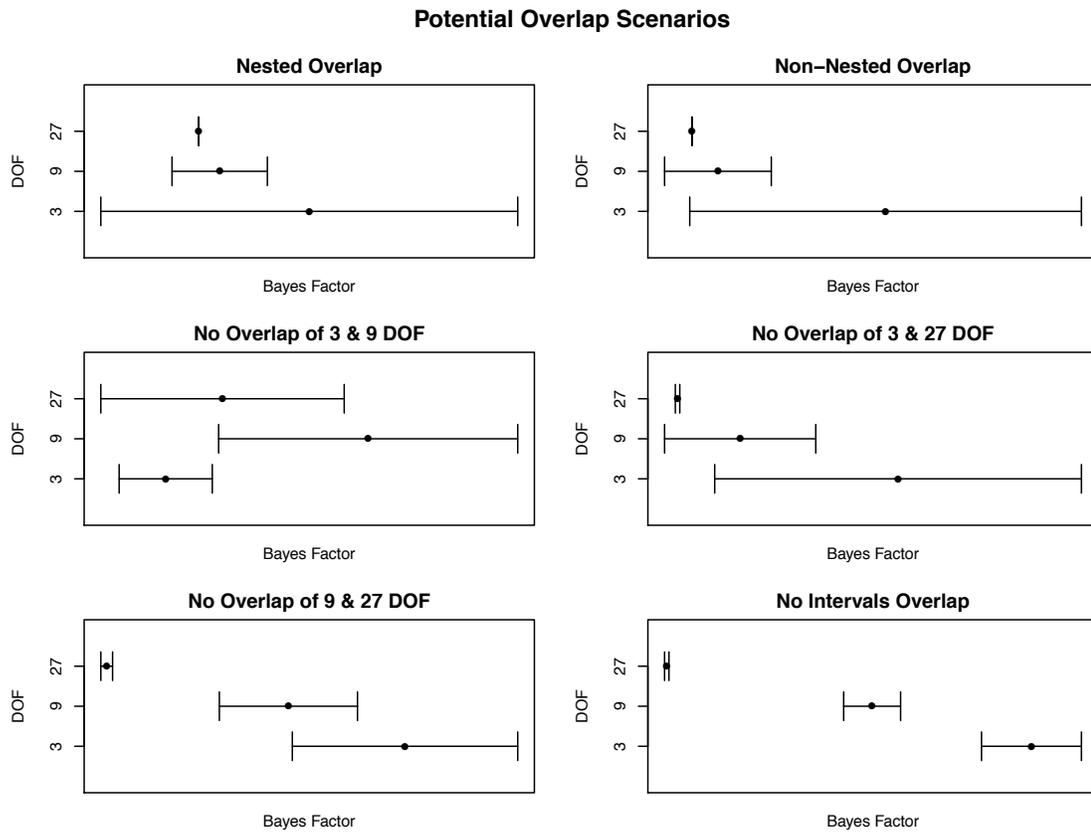
Figure 4.12: Potential overlap scenarios of intervals for Bayes Factors based on the corrected MCSE. (These exemplars have been chosen from the simulation results for Bayes Factors approximated using the Arithmetic Mean method under the specific source identification problem.)

Table 4.5: Sensitivity of the Bayes Factor to prior choice based on rate of disagreement of the Bayes Factor intervals using the MCSE.

Scenario:		Hp	CNM	1stQ	Med	3rdQ	Max
Arithmetic Mean	SS	1.0000	0.8750	0.4375	0.4375	0.0000	0.0000
	CS	0.5625	0.3125	0.0000	0.2500	0.0625	0.4375
Harmonic Mean	SS	0.9375	0.9375	1.0000	0.9375	0.8125	0.3125
	CS	0.3750	0.2500	0.0000	0.1875	0.0000	0.0000

of the choice of prior, and do not affect the validity of the Monte Carlo standard errors in this analysis. The specific source methods studied in the simulation have very high rates of disagreement. This would indicate that the specific source values of evidence are highly sensitive to the choice of prior. This effect of prior choice on the specific source Bayes Factors is particularly problematic in regards to using the Bayes Factors in the judicial system. The specific source methods are of the most interest in criminal trials because the specific source is fixed (typically related to the defendant). As a result, it might be beneficial to use methods of quantifying the value of evidence that do not depend on prior information for the nuisance parameters, or that use standard reference priors [11] for the evidence type of interest. When the Bayes Factor is particularly sensitive to prior choice, a common approach is to choose a non-informative prior. However, in most situations we have encountered, the non-informative priors are improper, leading to issues in the resulting Bayes Factor [59]. In contrast, the common source methods are less sensitive to the prior choice. The higher rates of disagreement for the common source Arithmetic mean methods under the extreme scenarios ('3rdQ' and 'Max') are a byproduct of the MCSE being numerically zero, so these cases are not concerning. Although some might find it appealing to use the common source methods in criminal trials since it appears to be less sensitive to prior choice, the common source value of evidence is not particularly applicable in this setting since it does not consider the source related to the defendant to be fixed.

4.4   Discussion

In the forensic science community, Bayesian methods have been proposed as the ideal method for quantifying the probative value of evidence. Due to these recommendations, many forensic statisticians have been focusing on formalizing methods of

characterizing the different types of uncertainties and errors associated with computing the value of evidence in practice. Using asymptotic properties of the Monte Carlo integration estimates for the numerator and denominator of the Bayes Factor, a single MCSE for the Bayes Factor was developed. This chapter proposes using the Bayes Factor MCSE to characterize the amount of numerical imprecision associated with using Monte Carlo integration methods to approximate the Bayes Factor. This article provides a way to measure the computational reliability of the reported Bayes Factors, not only for forensic science, but in general. (It is important to note that the MCSE does not characterize the accuracy or computational validity of the Bayes Factors.)

For the simulations studied, the diagnostic tools suggested that the thinning interval needs to be increased. When the thinning interval is not large enough, the calculation of the standard error should take into account the effective sample size instead of the Monte Carlo sample size. Without any sample size correction, the MCSE tends to underestimate the standard error for the Bayes Factor, giving a false sense of numerical precision. However, using the effective sample size will correct this standard error to reflect the added imprecision associated with using a Gibbs sampling algorithm. For most of the situations studied, the corrected MCSEs were closer to the empirical standard error in the Bayes Factors than the uncorrected MCSEs.

For the limited number of prior hyperparameters considered in the simulation study, the choice of prior seems to have little effect on the approximation of the Bayes Factors for the common source methods. However, for the specific source methods, the prior choice had a significant effect on the rates of misleading evidence for the Bayes Factors using the same evidence. This is problematic in the judicial system since it may lead a judge or jury to arrive at the wrong conclusion regarding the guilt/innocence of the defendant. Additionally, for the specific source methods, the prior choice had a

significant effect on level of disagreement of the Bayes Factors for the same evidence. This is also problematic in the judicial system since different forensic scientists could present different values for the Bayes Factor. The common source methods produced a lower rate of misleading evidence than the specific source methods, but the common source methods are less applicable in criminal trials since they do not assume that the specific source is fixed. Consequently, it may be beneficial to study methods of quantifying the value of evidence which do not depend on prior information for the nuisance parameters, or that use standard reference priors.

CHAPTER 5

The Bayes Factor & the Likelihood Ratio

While the Bayes Factor and the likelihood ratio are two different statistics used for quantifying the value of evidence under different statistical paradigms (see Section 3.3 for details), the two have many practical relationships. In this section, standard Bayesian analysis tools, including Doob's Consistency Theorem, are used to relate the Bayes Factor and likelihood ratio. In particular, Doob's Consistency Theorem will be used to show that under suitable conditions, a Bayesian believes the Bayes Factor will converge almost-surely to the likelihood ratio. This result is important because it reveals that Bayesians believe there is a fixed value of the likelihood ratio function analogous to the Frequentist true likelihood ratio. This means that a Bayesian believes that the Frequentist results, like the Bernstein-von Mises Theorem, will hold almost-surely with respect to their prior belief.

5.1   Common Source

A direct relationship between the Bayes Factor and likelihood ratio is given in the equation below, which provides an equivalent non-standard expression for the Bayes Factor. While the first expression for the Bayes Factor applies generally to all Bayesian model selection and hypothesis testing problems, this non-standard form of the Bayes Factor is mainly applicable in statistical pattern recognition problems, and of course, the forensic identification of source problems. This alternative expres-

sion of the Bayes Factor considers the expected value of the common source likelihood ratio function with respect to the posterior distribution for θ_a given the observation of the entire set of evidence under the defense model.

$$
\begin{aligned}
BF_{cs2}(e) &= \int \frac{f(e_{u_1}, e_{u_2}|\theta_a, M_p)}{f(e_{u_1}, e_{u_2}|\theta_a, M_d)} \ d\Pi(\theta_a|e_{u_1}, e_{u_2}, e_a, M_d) \\
&= \int LR_{cs}(\theta_a; e_{u_1}, e_{u_2}) \ d\Pi(\theta_a|e_{u_1}, e_{u_2}, e_a, M_d) \quad (5.1)
\end{aligned}
$$

The derivation of Equation 5.1 is given below. This derivation is a summary of the derivation given in Ommen et al. [55].

Derivation (5.1): This derivation essentially consists of multiplication by convenient factors of one and applications of standard Bayesian analysis tools. The derivation starts from the first definition of the common source Bayes Factor given by Equation 3.7 above. Again, it is assumed that the class of parametric distributions, $\mathcal{P}_a$, and the prior distributions, Π, have been chosen such that the integrals expressed exist and are finite.

$$
\begin{aligned}
BF_{cs2}(e) &= \frac{\int f(e_{u_1}, e_{u_2}|\theta_a, M_p) \ d\Pi(\theta_a|e_a)}{\int f(e_{u_1}|\theta_a, M_d) \ f(e_{u_2}|\theta_a, M_d) \ d\Pi(\theta_a|e_a)} & (5.2a) \\
&= \frac{\int f(e_{u_1}, e_{u_2}|\theta_a, M_p) \ d\Pi(\theta_a|e_a)}{f(e_{u_1}, e_{u_2}|e_a, M_d)} & (5.2b) \\
&= \int \frac{f(e_{u_1}, e_{u_2}|\theta_a, M_p)}{f(e_{u_1}, e_{u_2}|e_a, M_d)} \ d\Pi(\theta_a|e_a) & (5.2c) \\
&= \int \frac{f(e_{u_1}, e_{u_2}|\theta_a, M_p)}{f(e_{u_1}, e_{u_2}|e_a, M_d)} \times \frac{f(e_{u_1}, e_{u_2}|\theta_a, M_d)}{f(e_{u_1}, e_{u_2}|\theta_a, M_d)} \ d\Pi(\theta_a|e_a) & (5.2d) \\
&= \int \frac{f(e_{u_1}, e_{u_2}|\theta_a, M_p)}{f(e_{u_1}, e_{u_2}|\theta_a, M_d)} \times \frac{f(e_{u_1}, e_{u_2}|\theta_a, M_d) \ d\Pi(\theta_a|e_a)}{f(e_{u_1}, e_{u_2}|e_a, M_d)} & (5.2e) \\
&= \int \frac{f(e_{u_1}, e_{u_2}|\theta_a, M_p)}{f(e_{u_1}, e_{u_2}|\theta_a, M_d)} \ d\Pi(\theta_a|e_a, e_{u_1}, e_{u_2}, M_d) & (5.2f) \\
&= \int LR_{cs}(\theta_a; e_{u_1}, e_{u_2}) \ d\Pi(\theta_a|e, M_d) & (5.2g)
\end{aligned}
$$

$\square$

It should be noted that the reciprocal of the Bayes Factor, $1/BF_{cs2}(e)$ (where $BF_{cs2}(e)$ is defined as in Equation 5.1), is equivalent to the expected value of the reciprocal of the common source likelihood ratio function with respect to the posterior distribution for the entire set of evidence under the prosecution model, given below.

$$BF_{cs3}(e) = \left[\int \frac{1}{LR_{cs}(\theta_a; e_{u_1}, e_{u_2})} \, d\Pi(\theta_a | e_{u_1}, e_{u_2}, e_a, M_p) \right]^{-1} \qquad (5.3)$$

The derivation of Equation 5.3 is provided below for reference.

Derivation (5.3): In a similar manner to Derivation 5.1, the reciprocal of the common source Bayes Factor from Equation 3.7 is equal to the expected value of the reciprocal of the common source likelihood ratio with respect to the posterior distribution given the entire set of evidence has been generated according to the defense model.

$$\frac{1}{BF_{cs1}(e)} = \int \frac{f(e_{u_1}, e_{u_2} | \theta_a, M_d)}{f(e_{u_1}, e_{u_2} | e_a, M_p)} \times \frac{f(e_{u_1}, e_{u_2} | \theta_a, M_p)}{f(e_{u_1}, e_{u_2} | \theta_a, M_p)} \, d\Pi(\theta_a | e_a) \quad (5.4a)$$

$$= \int \frac{1}{LR_{cs}(\theta_a; e_{u_1}, e_{u_2})} \, d\Pi(\theta_a | e, M_p) \qquad (5.4b)$$

Therefore,

$$BF_{cs3}(e) = \left[\int \frac{1}{LR_{cs}(\theta_a; e_{u_1}, e_{u_2})} \, d\Pi(\theta_a | e, M_p) \right]^{-1} .$$

$\square$

There are no strong assumptions necessary for the alternative form of the common source Bayes Factor to hold.

Another relationship between the Bayes Factor and likelihood ratio is provided as a consequence of the Doob's Consistency Theorem [75], which is included for clarity in Section 2.3. It can be shown that the BF and the LR for the common source identification problem are asymptotically equivalent (as n_a tends to infinity) under a

variety of prior distributions for the nuisance parameters in the Bayes Factor. The result is formalized in the theorem to follow.

Theorem 5.1 (Common Source Bayes Factor Consistency)**:**

Given a fixed observation of e_{u_1} and e_{u_2}, suppose that $f(e_{u_1}, e_{u_2}|\theta_a, M_p)$ is bounded random variable with respect to $\Pi(\theta_a)$ and that $f(e_{u_1}, e_{u_2}|\theta_a, M_d)$ is a random variable with respect to $\Pi(\theta_a)$ which is not degenerate at 0. Let the assumptions of Doob's Consistency Theorem be satisfied. Then for every prior probability measure $\Pi(\theta_a)$ on Θ_a, the sequence of Bayes Factors, $BF_{cs}(e_n)$, converges almost surely to the likelihood ratio, $LR_{cs}(\theta_a; e_{u_1}, e_{u_2})$, as $n_a \to \infty$ for Π-almost every θ_a and for $P_{\theta_a}^\infty$-almost every e_a^∞.

Proof: For this proof, we will use the standard abuse of notation and let $\Pi_{n_a}(\theta_a'|e_{a,n_a})$ denote the cumulative distribution function corresponding to the posterior measure on the parameter space given the observation e_{a,n_a}. We will also let $\delta_{\theta_a}(\theta_a')$ denote the cumulative distribution function corresponding to the probability measure degenerate at θ_a. By Doob's Consistency Theorem (Theorem 2.10), for Π_a-almost every θ_a and for $P_{\theta_a}^\infty$-almost every $e_{a,\infty}$, then as $n_a \to \infty$

$$\Pi_{n_a}(\theta_a'|e_{a,n_a}) \to \delta_{\theta_a}(\theta_a') \tag{5.5}$$

for all continuity points θ_a' of δ_{θ_a} where δ_{θ_a} is the probability measure degenerate at θ_a.

Let $\mathscr{D}$ be the class of all Cadlag functions [76] and let $g : \mathscr{D} \mapsto \mathbb{R}$ be a continuous map such that $g(D) = \int f dD$ for $D \in \mathscr{D}$ and bounded, continuous function f. Then Equation 5.5 and the Continuous Mapping Theorem, imply that

$$g(\Pi_{n_a}(\theta_a'|e_{a,n_a})) \xrightarrow{as} g(\delta_{\theta_a}(\theta_a'))$$

for all continuity points θ'_a of δ_{θ_a}, for Π_a-almost every θ_a, and for $P^\infty_{\theta_a}$-almost every $e_{a,\infty}$. Using alternative notation and letting $h = p$ or $h = d$, this means that as $n_a \to \infty$

$$\int f(e_{u_1}, e_{u_2}|\theta'_a, M_h)d\Pi_{n_a}(\theta'_a|e_{a,n_a}) \xrightarrow{as} \int f(e_{u_1}, e_{u_2}|\theta'_a, M_h)d\delta_{\theta_a}(\theta'_a).$$

Therefore, for both $h = p$ and $h = d$, as $n_a \to \infty$

$$f(e_{u_1}, e_{u_2}|e_{a,n_a}, M_h) \xrightarrow{as} f(e_{u_1}, e_{u_2}|\theta_a, M_h)$$

for Π_a-almost every θ_a and for $P^\infty_{\theta_a}$-almost every $e_{a,\infty}$.

Now, because the non-negative inverse function is continuous, then by the Continuous Mapping Theorem

$$\frac{1}{f(e_{u_1}, e_{u_2}|e_{a,n_a}, M_d)} \xrightarrow{as} \frac{1}{f(e_{u_1}, e_{u_2}|\theta_a, M_d)}.$$

Finally, by Slutsky's Theorem, as $n_a \to \infty$,

$$BF_{cs}(e_n) = \frac{f(e_{u_1}, e_{u_2}|e_{a,n_a}, M_p)}{f(e_{u_1}, e_{u_2}|e_{a,n_a}, M_d)} \xrightarrow{as} \frac{f(e_{u_1}, e_{u_2}|\theta_a, M_p)}{f(e_{u_1}, e_{u_2}|\theta_a, M_d)} = LR_{cs}(\theta_a; e_{u_1}, e_{u_2})$$

for Π_a-almost every θ_a and for $P^\infty_{\theta_a}$-almost every $e_{a,\infty}$. ∎

5.2   Specific Source

An alternative form of the specific source Bayes Factor is given in an analogous manner to the common source. Under the assumption that the prior distribution for θ_s is statistically independent of the prior distribution for θ_a, the specific source Bayes

Factor is given by the expected value of the likelihood ratio with respect to the joint posterior distribution for the parameters given the entire set of evidence under the defense model.

$$
\begin{aligned}
BF_{ss2}(e) &= \int \frac{f(e_u|\theta_s, M_p)}{f(e_u|\theta_a, M_d)} \, d\Pi(\theta_s|e_s) \, d\Pi(\theta_a|e_u, e_a, M_d) \\
&= \int LR_{ss}(\theta; e_u) \, d\Pi(\theta|e_u, e_s, e_a, M_d) \quad\quad (5.6)
\end{aligned}
$$

The full derivation of Equation 5.6 can be found below.

> **Derivation (5.6):** Similar to the common source derivation above, this derivation also consists of multiplication by a convenient factor of one and applications of standard Bayesian analysis tools. The derivation begins from Equation 3.4 for the Bayes Factor specifically for the specific source identification problem. Again, it is assumed that the classes of parametric distributions, $\mathcal{P}_s$ and $\mathcal{P}_a$, and the prior distributions, Π, have been chosen such that the integrals expressed exist and are finite.

$$
\begin{aligned}
BF_{ss2}(e) &= \frac{\int f(e|\theta, M_p) \, d\Pi(\theta|M_p)}{\int f(e|\theta, M_d) \, d\Pi(\theta|M_d)} & (5.7\text{a}) \\
&= \frac{\int f(e_u|\theta_s)f(e_s|\theta_s)f(e_a|\theta_a) \, d\Pi(\theta_s, \theta_a)}{f(e_u, e_s, e_a|M_d)} & (5.7\text{b}) \\
&= \int \frac{f(e_u|\theta_s)f(e_s|\theta_s)f(e_a|\theta_a)}{f(e_u, e_s, e_a|M_d)} \, d\Pi(\theta_s, \theta_a) & (5.7\text{c}) \\
&= \int \frac{f(e_u|\theta_s)f(e_s|\theta_s)f(e_a|\theta_a)}{f(e_u, e_s, e_a|M_d)} \times \frac{f(e_u|\theta_a)}{f(e_u|\theta_a)} \, d\Pi(\theta_s, \theta_a) & (5.7\text{d}) \\
&= \int \frac{f(e_u|\theta_s)}{f(e_u|\theta_a)} \times \frac{f(e_u|\theta_a)f(e_s|\theta_s)f(e_a|\theta_a)}{f(e_u, e_s, e_a|M_d)} \, d\Pi(\theta_s, \theta_a) & (5.7\text{e}) \\
&= \int \frac{f(e_u|\theta_s)}{f(e_u|\theta_a)} \times \frac{f(e_u, e_s, e_a|\theta_s, \theta_a, M_d)}{f(e_u, e_s, e_a|M_d)} \, d\Pi(\theta_s, \theta_a) & (5.7\text{f}) \\
&= \int \frac{f(e_u|\theta_s)}{f(e_u|\theta_a)} \, d\Pi(\theta_s, \theta_a|e_u, e_s, e_a, M_d) & (5.7\text{g}) \\
&= \int LR_{ss}(\theta_s, \theta_a; e_u) \, d\Pi(\theta_s, \theta_a|e_u, e_s, e_a, M_d) & (5.7\text{h}) \\
&= \int LR_{ss}(\theta; e_u) \, d\Pi(\theta|e, M_d) & (5.7\text{i})
\end{aligned}
$$

$\square$

Similarly, the reciprocal of $BF_{ss2}(e)$ (the alternative form of the specific source Bayes Factor given in Equation 5.6) is equivalent to the expected value of the reciprocal of the likelihood ratio with respect to the joint posterior distribution for the parameters given the entire set of evidence under the prosecution model, as given in Equation 5.8 below.

$$BF_{ss3}(e) = \left[\int \frac{1}{LR_{ss}(\theta; e_u)} \, d\Pi(\theta | e_u, e_s, e_a, M_p) \right]^{-1} \tag{5.8}$$

A derivation of Equation 5.8 is provided below. Again, it should be noted that in order for this form of the specific source Bayes Factor to hold, there is a strong assumption that the prior distribution for θ_s is statistically independent of the prior distribution for θ_a.

Derivation (5.8): Similar to the common source derivation above, the reciprocal of the specific source Bayes Factor is equal to the expected value of the reciprocal of the specific source likelihood ratio with respect to the posterior distribution given the entire set of evidence has been generated according to the defense model.

$$\frac{1}{BF_{ss1}(e)} = \int \frac{f(e_u | \theta_a) f(e_s | \theta_s) f(e_a | \theta_a)}{f(e_u, e_s, e_a | M_p)} \times \frac{f(e_u | \theta_s)}{f(e_u | \theta_s)} \, d\Pi(\theta_s, \theta_a) \tag{5.9a}$$

$$= \int \frac{1}{LR_{ss}(\theta_s, \theta_a; e_u)} \, d\Pi(\theta_s, \theta_a | e_u, e_s, e_a, M_p) \tag{5.9b}$$

Therefore,

$$BF_{ss3}(e) = \left[\int \frac{1}{LR_{ss}(\theta; e_u)} \, d\Pi(\theta | e, M_p) \right]^{-1}.$$

$\square$

As a consequence of the Doob's Consistency Theorem [75], it can be shown that the Bayes Factor and the likelihood ratio for the specific source identification problem are asymptotically equivalent as the number of samples from the specific source (n_a) and the number of sources in the alternative source population (n_a) tend to infinity.

The result is formalized in the theorem to follow.

Theorem 5.2 (Specific Source Bayes Factor Consistency)**:**

Given a fixed observation of e_u, suppose that $f(e_u|\theta_s)$ is bounded random variable with respect to $\Pi(\theta_s)$ and that $f(e_u|\theta_a)$ is a random variable with respect to $\Pi(\theta_a)$ which is not degenerate at 0. Let the assumptions of Doob's Consistency Theorem be satisfied. Then for every joint prior probability measure $\Pi(\theta)$ on Θ, the sequence of Bayes Factors, $BF_{ss}(e_n)$, converges almost surely to the likelihood ratio, $LR_{ss}(\theta; e_u)$, as $n = n_s = n_a \to \infty$ for Π-almost every θ and for P_θ^∞-almost every e_∞.

Proof: For this proof, we will use the standard abuse of notation and let $\Pi_{n_a}(\theta_a'|e_{a,n_a})$ denote the cumulative distribution function corresponding to the posterior measure on Θ_a the observation e_{a,n_a}, and let $\Pi_{n_s}(\theta_s'|e_{s,n_s})$ denote the cumulative distribution function corresponding to the posterior measure on Θ_s given the observation e_{s,n_s}. We will also let $\delta_{\theta_a}(\theta_a')$ denote the cumulative distribution function corresponding to the probability measure degenerate at θ_a, and $\delta_{\theta_s}(\theta_s')$ denote the cumulative distribution function corresponding to the probability measure degenerate at θ_s. By Doob's Consistency Theorem (Theorem 2.10), for Π_a-almost every θ_a and for $P_{\theta_a}^\infty$-almost every $e_{a,\infty}$, then as $n_a \to \infty$

$$\Pi_{n_a}(\theta_a'|e_{a,n_a}) \to \delta_{\theta_a}(\theta_a') \tag{5.10}$$

for all continuity points θ_a' of δ_{θ_a} where δ_{θ_a} is the probability measure degenerate at θ_a. Similarly, for Π_s-almost every θ_s and for $P_{\theta_s}^\infty$-almost every $e_{s,\infty}$, then as $n_s \to \infty$

$$\Pi_{n_s}(\theta_s'|e_{s,n_s}) \to \delta_{\theta_s}(\theta_s') \tag{5.11}$$

for all continuity points θ_s' of δ_{θ_s} where δ_{θ_s} is the probability measure degenerate at θ_s.

Let $\mathscr{D}$ and $g : \mathscr{D} \mapsto \mathbb{R}$ be defined as above for the proof of Theorem 5.1. Then Equations 5.10 and 5.11 and the Continuous Mapping Theorem, imply that

$$g(\Pi_{n_a}(\theta'_a|e_{a,n_a})) \xrightarrow{as} g(\delta_{\theta_a}(\theta'_a))$$

for all continuity points θ'_a of δ_{θ_a}, for Π_a-almost every θ_a, and for $P_{\theta_a}^\infty$-almost every $e_{a,\infty}$, and

$$g(\Pi_{n_s}(\theta'_s|e_{s,n_s})) \xrightarrow{as} g(\delta_{\theta_s}(\theta'_s))$$

for all continuity points θ'_s of δ_{θ_s}, for Π_s-almost every θ_s, and for $P_{\theta_s}^\infty$-almost every $e_{s,\infty}$.

Using alternative notation, this means that as $n_a \to \infty$

$$\int f(e_u|\theta'_a)d\Pi_{n_a}(\theta'_a|e_{a,n_a}) \xrightarrow{as} \int f(e_u|\theta'_a)d\delta_{\theta_a}(\theta'_a),$$

and as $n_s \to \infty$

$$\int f(e_u|\theta'_s)d\Pi_{n_s}(\theta'_s|e_{s,n_s}) \xrightarrow{as} \int f(e_u|\theta'_s)d\delta_{\theta_s}(\theta'_s).$$

Therefore, for Π_a-almost every θ_a and for $P_{\theta_a}^\infty$-almost every $e_{a,\infty}$, as $n_a \to \infty$

$$f(e_u|e_{a,n_a}) \xrightarrow{as} f(e_u|\theta_a),$$

and for Π_s-almost every θ_s and for $P_{\theta_s}^\infty$-almost every $e_{s,\infty}$, as $n_s \to \infty$

$$f(e_u|e_{s,n_s}) \xrightarrow{as} f(e_u|\theta_s).$$

Now, because the non-negative inverse function is continuous, then by the Continuous

Mapping Theorem

$$\frac{1}{f(e_u|e_{a,n_a})} \xrightarrow{as} \frac{1}{f(e_u|\theta_a)}.$$

Finally, by Slutsky's Theorem, as $n \equiv n_a = n_s \to \infty$,

$$BF_{ss}(e_n) = \frac{f(e_u|e_{s,n_s})}{f(e_u|e_{a,n_a})} \xrightarrow{as} \frac{f(e_u|\theta_s)}{f(e_u|\theta_a)} = LR_{ss}(\theta_s, \theta_a; e_u)$$

for for Π_s-almost every θ_s, for $P_{\theta_s}^{\infty}$-almost every $e_{s,\infty}$, for Π_a-almost every θ_a, and for $P_{\theta_a}^{\infty}$-almost every $e_{a,\infty}$. ∎

It should be noted that for the statement of this theorem, it is assumed that n_s and n_a are equal. However, this assumption can be generalized to include values for n_s and n_a that are proportional, although the proof becomes more complicated.

5.3   Application Example

The purpose of this example is to demonstrate the methods of computing both the common source and specific source Bayes Factors using the alternative expressions given by Equation 5.1 and Equation 5.6. The Monte Carlo integration methods for computing these values of evidence are provided in Section 5.3.3. The resulting values of evidence will be compared with those computed using the first expression given by Equation 3.7 and Equation 3.11. For both of these examples, two different datasets will be used to examine any sample size effects on the resulting values of evidence, the glass dataset described in Section 4.3 and a simulated glass dataset with larger samples sizes than the former.

5.3.1   Observed Glass Data

The dataset used for this example is the same glass data provided in Section 4.3. Similarly to Section 4.3, the data from the first group of glass fragments, used for the evidence data, is modeled with a normal distribution. The data used for the evidence in the specific source identification example under the scenario that the prosecution hypothesis is true is given by

e_u: Two fragments from window number 10;

e_s: Three different fragments from window number 10;

e_a: Five fragments from each of 14 windows in the first group (excluding windows number 10 and 48).

The data used for the evidence in the specific source identification example under the scenario that the defense hypothesis is true is given by

e_u: Two fragments from window number 10;

e_s: Five fragments from window number 48;

e_a: Five fragments from each of 14 windows in the first group (excluding windows number 10 and 48).

The data used for the evidence in the common source identification example under the scenario that the prosecution hypothesis is true is given by

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

e_{u_2}: Three different fragments from window number 10;

e_a: Five fragments from each of 14 windows in the first group (excluding windows number 10 and 48).

The data used for the evidence in the specific source identification example under the scenario that the defense hypothesis is true is given by

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

e_{u_2}: Five fragments from window number 48;

e_a: Five fragments from each of 14 windows in the first group (excluding windows number 10 and 48).

The datasets described above for this application example are for illustrative purposes only and do not necessarily reflect realistic data commonly encountered in casework. In the scenarios under which the data is chosen according the prosecution hypothesis being true, it is expected that the Bayes Factors will be larger than one. In contrast, for the scenarios under which the data is chosen according the the defense hypothesis being true, it is expected that the Bayes Factors will be less than one. Following Section 4.3, the measurements from the second and third groups of windows are combined and used to suggest the prior hyperparameters. Again, a normal distribution is used as the prior distribution for the mean parameters and an inverse Wishart distribution is used as the prior distribution for the covariance parameters, see Equations 4.9–4.13. The number of degrees of freedom for the inverse Wishart distributions was chosen to be 27 for this example and the value of K was chosen to be 10, similar to Section 4.3.

5.3.2   Simulated Glass Data

Notice that the glass data described above has relatively small sample sizes, only a maximum of five fragments per source (including the specific source) and only 14 sources in the alternative source population. Because these samples sizes are so small, it was desired to simulate some glass data with larger sample sizes to compare the

behavior of the computed values of evidence. The simulated glass data will have five fragments in each of the unknown source evidence sets, 25 fragments in the specific source evidence, and 50 sources in the alternative source population with ten fragments within each source.

To simulate this data, the first group of the glass data was used to suggest the true parameter values for the multivariate normal distributions. The measurements on the fragments from the window with identification number 10, where $\mathbf{y}_{si}^{(10)}$ is the three-dimensional (column) vector of elemental compositions for the ith fragment from window number 10, were used to determine the parameter values, $\mu_{s_0}^{(10)}$ and $\Sigma_{s_0}^{(10)}$, given as the sample estimates below.

$$\mu_{s_0}^{(10)} = \frac{1}{n_s^{(10)}} \sum_{i=1}^{n_s^{(10)}} \mathbf{y}_{si}^{(10)} = \begin{pmatrix} 4.586980 & -0.365652 & 2.699196 \end{pmatrix}^T$$

$$\begin{aligned} \Sigma_{s_0}^{(10)} &= \frac{1}{n_s^{(10)} - 1} \sum_{i=1}^{n_s^{(10)}} \left(\mathbf{y}_{si}^{(10)} - \mu_{s_0}^{(10)} \right) \left(\mathbf{y}_{si}^{(10)} - \mu_{s_0}^{(10)} \right)^T \\ &= \begin{pmatrix} 8.378358 \times 10^{-03} & 6.880485 \times 10^{-04} & -6.512652 \times 10^{-04} \\ 6.880485 \times 10^{-04} & 7.254092 \times 10^{-05} & -5.674613 \times 10^{-05} \\ -6.512652 \times 10^{-04} & -5.674613 \times 10^{-05} & 2.673494 \times 10^{-04} \end{pmatrix} \end{aligned}$$

The measurements on the fragments from the window with identification number 48 were used to determine the parameter values, $\mu_{s_0}^{(48)}$ and $\Sigma_{s_0}^{(48)}$ in a similar manner to the above sample estimates.

$$\mu_{s_0}^{(48)} = \begin{pmatrix} 4.792984 & -0.320326 & 2.846948 \end{pmatrix}^T$$

$$\Sigma_{s_0}^{(10)} = \begin{pmatrix} 1.138632 \times 10^{-02} & 8.849329 \times 10^{-04} & 1.260511 \times 10^{-03} \\ 8.849329 \times 10^{-04} & 1.165956 \times 10^{-04} & 1.137223 \times 10^{-04} \\ 1.260511 \times 10^{-03} & 1.137223 \times 10^{-04} & 5.526833 \times 10^{-04} \end{pmatrix}$$

Finally, the measurements on the fragments from the remaining 14 windows were used to determine the parameter values, μ_{a_0}, Σ_{b_0}, and Σ_{e_0} in a manner similar to Equation 4.13 (the only difference is that Equations 4.13 uses either group 2 or group 3 of the glass data, and here group 1 is used).

$$\mu_{a_0} = \begin{pmatrix} 4.830995 & -0.344223 & 2.751252 \end{pmatrix}^T$$

$$\Sigma_{b_0} = \begin{pmatrix} 0.118334236 & -0.009927254 & 0.087147020 \\ -0.009927254 & 0.002945243 & -0.008224465 \\ 0.087147020 & -0.008224465 & 0.094283548 \end{pmatrix}$$

$$\Sigma_{e_0} = \begin{pmatrix} 1.653828 \times 10^{-02} & 8.373814 \times 10^{-05} & 3.564552 \times 10^{-04} \\ 8.373814 \times 10^{-05} & 5.010762 \times 10^{-05} & -1.088538 \times 10^{-05} \\ 3.564552 \times 10^{-04} & -1.088538 \times 10^{-05} & 5.169330 \times 10^{-04} \end{pmatrix}$$

First of all, for both the common source and specific source identification problems and both of the scenarios under H_p and H_d being assumed true, the parameter values μ_{a_0}, Σ_{b_0}, and Σ_{e_0} are used for generating the alternative source population evidence, e_a, according to

$$\mathbf{y}_{ci} \stackrel{iid}{\sim} \mathcal{N}_{15}(\mu_{c_0}, \Sigma_{c_0}),$$

for $i = 1, 2, \ldots, n_a$, where μ_{c_0} depends on μ_{a_0} and Σ_{c_0} depends on Σ_{b_0} and Σ_{e_0} according to Theorem 2.2. For additional details, see Equation 4.10. Now, for the specific source problem under the H_p scenario, e_u and e_s are generated according to

$$\mathbf{y}_{ui} \stackrel{iid}{\sim} \mathcal{N}_3(\mu_{s_0}^{(10)}, \Sigma_{s_0}^{(10)}) \quad \text{and} \quad \mathbf{y}_{sj} \stackrel{iid}{\sim} \mathcal{N}_3(\mu_{s_0}^{(10)}, \Sigma_{s_0}^{(10)}),$$

for $i = 1, \ldots, 5$ and $j = 1, \ldots, 25$. Under the H_d scenario, e_u is generated using the same parameter values while e_s is generated according to

$$\mathbf{y}_{sj} \overset{iid}{\sim} \mathcal{N}_3(\mu_{s_0}^{(48)}, \Sigma_{s_0}^{(48)}),$$

for $j = 1, \ldots, 25$. Next, for the common source problem under the H_p scenario, e_{u_1} and e_{u_2} are generated according to

$$\mathbf{y}_{u_1 i} \overset{iid}{\sim} \mathcal{N}_3(\mu_{s_0}^{(10)}, \Sigma_{s_0}^{(10)}) \qquad \text{and} \qquad \mathbf{y}_{u_2 i} \overset{iid}{\sim} \mathcal{N}_3(\mu_{s_0}^{(10)}, \Sigma_{s_0}^{(10)}),$$

for $i = 1, \ldots, 5$. Under the H_d scenario, e_{u_1} is generated using the same parameter values while e_{u_2} is generated according to

$$\mathbf{y}_{u_2 i} \overset{iid}{\sim} \mathcal{N}_3(\mu_{s_0}^{(48)}, \Sigma_{s_0}^{(48)}),$$

for $i = 1, \ldots, 5$.

5.3.3   Monte Carlo Integration for Non-Standard BF

Recall from Equation 5.1 that the Bayes Factor for the common source identification problem can be written as

$$BF_{cs2} = \int \frac{f(e_{u_1}, e_{u_2} | \theta_a, M_p)}{f(e_{u_1}, e_{u_2} | \theta_a, M_d)} \, d\Pi(\theta_a | e_{u_1}, e_{u_2}, e_a, M_d).$$

This Bayes Factor can be approximated using the Arithmetic Mean method of Monte Carlo integration by

$$\widehat{BF}_{cs}^{(1)} = n^{-1} \sum_{i=1}^{n} \frac{f(e_{u_1}, e_{u_2} | \theta_a^{(i)}, M_p)}{f(e_{u_1}, e_{u_2} | \theta_a^{(i)}, M_d)} \tag{5.12}$$

where $\{\theta_a^{(i)}\}$ is an independent sample of size n drawn from $\pi(\theta_a|e_{u_1}, e_{u_2}, e_a, M_d)$. The Bayes Factor approximation for BF_{cs2} by method of Harmonic Mean Monte Carlo integration is given by

$$\widehat{BF}_{cs}^{(2)} = \left[n^{-1} \sum_{i=1}^{n} \frac{f(e_{u_1}, e_{u_2}|\theta_a^{(i)}, M_d)}{f(e_{u_1}, e_{u_2}|\theta_a^{(i)}, M_p)} \right]^{-1} \tag{5.13}$$

where $\{\theta_a^{(i)}\}$ is an independent sample of size n drawn from the importance sampling function $\pi(\theta_a|e_{u_1}, e_{u_2}, e_a, M_p) = f(e_{u_1}, e_{u_2}|\theta_a, M_p)\pi(\theta_a|e_a)/f(e_{u_1}, e_{u_2}|e_a, M_p)$. Note that this posterior distribution is different from the posterior distribution for Equation 5.12 since the given model for the evidence is different. The derivation that this Harmonic Mean estimate of the common source Bayes Factor is a proper Monte Carlo integration estimate that converges to BF_{cs2} is given below.

Derivation (5.13): It will be shown that $\widehat{BF}_{cs}^{(2)}$ converges almost-surely to BF_{cs2}. In the context above, the importance sampling function is given by

$$I(\theta_a) \equiv \pi(\theta_a|e_{u_1}, e_{u_2}, e_a, M_p) = \frac{f(e_{u_1}, e_{u_2}|\theta_a, M_p)\pi(\theta_a|e_a)}{f(e_{u_1}, e_{u_2}|e_a, M_p)}.$$

Therefore, by the SLLN,

$$n^{-1} \sum_{i=1}^{n} \frac{f(e_{u_1}, e_{u_2}|\theta_a^{(i)}, M_d)}{f(e_{u_1}, e_{u_2}|\theta_a^{(i)}, M_p)} \xrightarrow{as} E_I \left[\frac{f(e_{u_1}, e_{u_2}|\theta_a, M_d)}{f(e_{u_1}, e_{u_2}|\theta_a, M_p)} \right],$$

as the Monte Carlo sample size, n, tends to infinity, where

$$
\begin{aligned}
E_I \left[\frac{f(e_{u_1}, e_{u_2}|\theta_a, M_d)}{f(e_{u_1}, e_{u_2}|\theta_a, M_p)} \right] &= \int \frac{f(e_{u_1}, e_{u_2}|\theta_a, M_d)}{f(e_{u_1}, e_{u_2}|\theta_a, M_p)} \, d\Pi(\theta_a|e_{u_1}, e_{u_2}, e_a, M_p) \\
&= \int \frac{f(e_{u_1}, e_{u_2}|\theta_a, M_d)}{f(e_{u_1}, e_{u_2}|\theta_a, M_p)} \frac{f(e_{u_1}, e_{u_2}|\theta_a, M_p)}{f(e_{u_1}, e_{u_2}|e_a, M_p)} \, d\Pi(\theta_a|e_a) \\
&= \int \frac{f(e_{u_1}, e_{u_2}|\theta_a, M_d)}{f(e_{u_1}, e_{u_2}|e_a, M_p)} \, d\Pi(\theta_a|e_a) \\
&= \frac{\int f(e_{u_1}, e_{u_2}|\theta_a, M_d) \, d\Pi(\theta_a|e_a)}{f(e_{u_1}, e_{u_2}|e_a, M_p)} \\
&= \frac{\int f(e_{u_1}, e_{u_2}|\theta_a, M_d) \, d\Pi(\theta_a|e_a)}{\int f(e_{u_1}, e_{u_2}|\theta_a, M_p) \, d\Pi(\theta_a|e_a)}
\end{aligned}
$$

$$= \frac{1}{BF_{cs1}(e)} = \frac{1}{BF_{cs2}(e)}.$$

Finally, by Slutsky's Theorem, as $n \to \infty$

$$\widehat{BF}_{cs}^{(2)} = \left[n^{-1} \sum_{i=1}^{n} \frac{f(e_{u_1}, e_{u_2} | \theta_a^{(i)}, M_d)}{f(e_{u_1}, e_{u_2} | \theta_a^{(i)}, M_p)} \right]^{-1} \xrightarrow{as} BF_{cs2}.$$

$\square$

Note that the Harmonic Mean estimate of $BF_{cs2}(e)$ given by Equation 5.13 is equivalent to the Arithmetic Mean estimate of $BF_{cs3}(e)$ from Equation 5.3. Also, that the Arithmetic Mean estimate of $BF_{cs2}(e)$ given by Equation 5.12 is equivalent to the Harmonic Mean estimate of $BF_{cs3}(e)$ from Equation 5.3. These derivations are rather straight-forward, and so they will be omitted from this section.

Now, recall from Equation 5.6 that the specific source Bayes Factor is given by

$$BF_{ss2}(e) = \int \frac{f(e_u | \theta_s, M_p)}{f(e_u | \theta_a, M_d)} \, d\Pi(\theta | e_u, e_s, e_a, M_d).$$

The specific source Bayes Factor can be approximated using the Arithmetic Mean method of Monte Carlo integration by

$$\widehat{BF}_{ss}^{(1)} = n^{-1} \sum_{i=1}^{n} \frac{f(e_u | \theta_s^{(i)}, M_p)}{f(e_u | \theta_a^{(i)}, M_d)} \tag{5.14}$$

where $\{\theta_s^{(i)}\}$ is an independent sample of size n drawn from $\pi(\theta_s | e_s)$ and $\{\theta_a^{(i)}\}$ is an independent sample of size n drawn from $\pi(\theta_a | e_u, e_a, M_d)$ (or you can think of this as a single joint sample $\{\theta^{(i)}\} = \{(\theta_s^{(i)}, \theta_a^{(i)})\}$ of size n from $\pi(\theta | e_u, e_s, e_a, M_d)$). The Bayes Factor approximation for BF_{cs2} by method of Harmonic Mean Monte Carlo

integration is given by

$$\widehat{BF}_{ss}^{(2)} = \left[n^{-1} \sum_{i=1}^{n} \frac{f(e_u|\theta_a^{(i)}, M_d)}{f(e_u|\theta_s^{(i)}, M_p)} \right]^{-1} \tag{5.15}$$

where $\{\theta_s^{(i)}\}$ is an independent sample of size n drawn from $\pi(\theta_s|e_u, e_s, M_p)$ and $\{\theta_a^{(i)}\}$ is an independent sample of size n drawn from $\pi(\theta_a|e_a)$ (or you can think of this as a single joint sample $\{\theta^{(i)}\} = \{(\theta_s^{(i)}, \theta_a^{(i)})\}$ of size n from the importance sampling function $I(\theta) \equiv \pi(\theta|e_u, e_s, e_a, M_p))$. Note, similarly to the common source Bayes Factor, that this posterior distribution is different from the posterior distribution for Equation 5.14 since the given evidence model is different. The derivation that this Harmonic Mean estimate of the specific source Bayes Factor is a proper Monte Carlo integration estimate that converges to BF_{ss2} is given below.

Derivation (5.15): It will be shown that $\widehat{BF}_{ss}^{(2)}$ converges almost-surely to BF_{ss2}. In the context above, the importance sampling function is given by

$$I(\theta) \equiv \pi(\theta_a|e_u, e_s, e_a, M_p) = \pi(\theta_s|e_u, e_s, M_p) \, \pi(\theta_a|e_a).$$

Therefore, by the SLLN, as $n \to \infty$ (where n is the Monte Carlo sample size)

$$n^{-1} \sum_{i=1}^{n} \frac{f(e_u|\theta_a^{(i)}, M_d)}{f(e_u|\theta_s^{(i)}, M_p)} \xrightarrow{as} E_I \left[\frac{f(e_u|\theta_a, M_d)}{f(e_u|\theta_s, M_p)} \right],$$

where

$$\begin{aligned}
E_I \left[\frac{f(e_u|\theta_a, M_d)}{f(e_u|\theta_s, M_p)} \right] &= \int \frac{f(e_u|\theta_a, M_d)}{f(e_u|\theta_s, M_p)} \, d\Pi(\theta_a|e_u, e_s, e_a, M_p) \\
&= \int \int \frac{f(e_u|\theta_a, M_d)}{f(e_u|\theta_s, M_p)} \, d\Pi(\theta_s|e_u, e_s, M_p) \, d\Pi(\theta_a|e_a) \\
&= \int f(e_u|\theta_a, M_d) \, d\Pi(\theta_a|e_a) \int \frac{1}{f(e_u|\theta_s, M_p)} \, d\Pi(\theta_s|e_u, e_s, M_p) \\
&= f(e_u|e_a, M_d) \int \frac{1}{f(e_u|\theta_s, M_p)} \frac{f(e_u|\theta_s, M_p)}{f(e_u|e_s, M_p)} d\Pi(\theta_s|e_s) \\
&= f(e_u|e_a, M_d) \int \frac{1}{f(e_u|e_s, M_p)} d\Pi(\theta_s|e_s)
\end{aligned}$$

$$= \frac{f(e_u|e_a, M_d)}{f(e_u|e_s, M_p)} \int d\Pi(\theta_s|e_s)$$

$$= \frac{f(e_u|e_a, M_d)}{f(e_u|e_s, M_p)} = \frac{1}{BF_{ss1}(e)} = \frac{1}{BF_{ss2}(e)}.$$

Finally, by Slutsky's Theorem, as $n \to \infty$

$$\widehat{BF}_{ss}^{(2)} = \left[n^{-1} \sum_{i=1}^{n} \frac{f(e_u|\theta_a^{(i)}, M_d)}{f(e_u|\theta_s^{(i)}, M_p)} \right]^{-1} \xrightarrow{as} BF_{ss2}.$$

$\square$

Again, it should be noted that the Harmonic Mean estimate of $BF_{ss2}(e)$ given by Equation 5.15 is equivalent to the Arithmetic Mean estimate of $BF_{ss3}(e)$ from Equation 5.8. Also, that the Arithmetic Mean estimate of $BF_{ss2}(e)$ given by Equation 5.14 is equivalent to the Harmonic Mean estimate of $BF_{ss3}(e)$ from Equation 5.8. These derivations are also omitted from this section since they are trivial extensions of the results above.

5.3.4   Application Results

Based on the results of Section 4.3, the Bayes Factor was computed using Monte Carlo integration techniques with a Monte Carlo sample size of $100,000$, a burn-in period of 2000, and a thinning interval of 15. Details of the Monte Carlo approximation of both the common source and specific source Bayes Factors by both the Arithmetic and Harmonic mean methods are provided in Section 5.3.3. The auto-correlation and trace plots were examined for each of the computed Bayes Factors. A typical auto-correlation function plot and a typical trace plot are provided in Figure 5.1. These plots indicated that the burn-in and thinning intervals were appropriately chosen to ensure that the samples from the Gibbs sampling algorithm, as implemented by the 'MCMCglmm' function in R [40], were approximately independent. The auto-correlation values fall within the appropriate tolerance for all lag-values considered

Figure 5.1: Typical diagnostic plots to determine whether the chosen thinning intervals for the Gibbs samplers are appropriate.
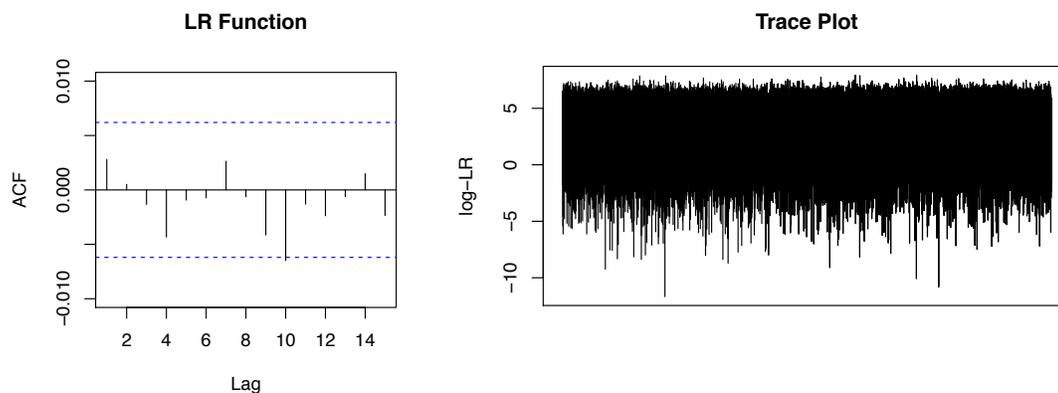
(a) Auto-correlation function plot showing acceptable thinning intervals.

(b) Trace plot showing acceptable thinning intervals.

and the trace plot of the log-likelihood ratio function values used in the sum-portion of the Monte Carlo integration method are randomly positioned without any apparent systematic pattern. Therefore, the Monte Carlo standard errors can be computed without any sample size corrections needed.

Using the methods described above, the Bayes Factors were computed for both the common source and specific source identification problems under two different scenarios each. The first scenario uses data created under the prosecution hypothesis and the second scenario uses data created under the defense hypothesis. Also, each of the Bayes Factors are computed using the alternative expression, where the Bayes Factor is given by the expected value of the likelihood ratio function with respect to the posterior distribution for the parameters given the entire set of evidence under the defense model, using both the Arithmetic and Harmonic Mean methods of Monte Carlo integration. The results are provided in Table 5.1 for the observed glass data and in Table 5.2 for the simulated glass data. As expected, all of the Bayes Factors for the evidence under the H_p scenario are above one, and all of the Bayes Factors for

Table 5.1: Computed Bayes Factors (and corresponding MCSE) for the observed glass data using the alternative expressions given in Equation 5.1 and Equation 5.6.

(a) Specific Source

	Scenario	
	H_p	H_d
Arithmetic	101.2017	8.0131×10^{-4}
Mean	(0.50845)	(2.3209×10^{-4})
Harmonic	98.98313	1.5477×10^{-3}
Mean	(2.02394)	(4.7023×10^{-4})

(b) Common Source

	Scenario	
	H_p	H_d
Arithmetic	256.4003	4.8869×10^{-9}
Mean	(0.22683)	(9.6985×10^{-10})
Harmonic	256.6640	3.7440×10^{-9}
Mean	(0.22553)	(2.0478×10^{-9})

Table 5.2: Computed Bayes Factors (and corresponding MCSE) for the simulated glass data using the alternative expressions given in Equation 5.1 and Equation 5.6.

(a) Specific Source

	Scenario	
	H_p	H_d
Arithmetic	4809.033	5.1356×10^{-10}
Mean	(13.55218)	(1.1316×10^{-10})
Harmonic	4800.352	2.5095×10^{-9}
Mean	(14.91021)	(2.2921×10^{-9})

(b) Common Source

	Scenario	
	H_p	H_d
Arithmetic	694.8544	5.5534×10^{-35}
Mean	(0.3558765)	(2.4121×10^{-35})
Harmonic	695.0588	6.0625×10^{-34}
Mean	(0.3532926)	(3.2642×10^{-34})

the evidence under the H_d scenario are less than one. These results are compared to the corresponding Bayes Factors computed using the original expression in Table 5.3 for the observed glass data and in Table 5.4 for the simulated glass data.

It is clear that the corresponding values of the Bayes Factor are very comparable between the alternative and original expressions. This is expected since the two expressions are equivalent. However, it should be noted that the Bayes Factors computed using the alternative expression are expected to be more stable since the computation involves only a single Monte Carlo integration instead of two, as is the case for

Table 5.3: Computed Bayes Factors (and corresponding MCSE) for the observed glass data using the original expressions given in Equation 3.7 and Equation 3.11.

(a) Specific Source

	Scenario	
	H_p	H_d
Arithmetic	101.7335	5.20015×10^{-4}
Mean	(0.49266)	(1.0420×10^{-4})
Harmonic	101.3096	2.16392×10^{-3}
Mean	(1.82651)	(7.07263×10^{-4})

(b) Common Source

	Scenario	
	H_p	H_d
Arithmetic	256.6531	4.21411×10^{-9}
Mean	(0.64141)	$(5.86046 \times 10^{-10})$
Harmonic	256.5720	1.41288×10^{-8}
Mean	(0.67984)	(2.0639×10^{-9})

Table 5.4:   Computed Bayes Factors (and corresponding MCSE) for the simulated glass data using the original expressions given in Equation 3.7 and Equation 3.11.

(a) Specific Source

	Scenario	
	H_p	H_d
Arithmetic	4805.955	4.2106×10^{-10}
Mean	(13.02153)	(1.2798×10^{-10})
Harmonic	4802.957	3.0628×10^{-8}
Mean	(14.46239)	(1.7698×10^{-9})

(b) Common Source

	Scenario	
	H_p	H_d
Arithmetic	695.0156	1.5637×10^{-33}
Mean	(1.613102)	(1.5165×10^{-33})
Harmonic	694.084	4.6234×10^{-34}
Mean	(1.649854)	(2.9533×10^{-34})

Table 5.5:   Computed likelihood ratio values for the simulated glass example.

	Scenario	
	H_p	H_d
Specific Source	2449715	1.708589×10^{-50}
Common Source	482.056	5.31445×10^{-48}

the original expression of the Bayes Factor. Also, in this case, the computation is additionally reduced by simplification of computing the Monte Carlo standard error. Refer to Section 4.2 for details of the Monte Carlo standard error for Bayes Factors approximated by Monte Carlo integration from the original expressions given by Equation 3.7 and Equation 3.11.

It should also be noted that the Bayes Factors computed using the simulated glass data, which has larger sample sizes than the observed glass data, tend to be closer to the values of the likelihood ratios computed using the true likelihood structures from the simulation. The values of these likelihood ratios are given in Table 5.5 for reference. This limited example is compatible with Theorem 5.1 and Theorem 5.2, that the Bayes Factor is consistent towards the likelihood ratio as the sample size increases.

5.4   Discussion

In Chapter 4, it was determined that the Monte Carlo integration estimates of the Bayes Factors (especially for the specific source) given in Equation 3.7 and Equation 3.11 are computationally complex and require additional computational effort to compute the corresponding Monte Carlo standard errors. In this section, two non-standard forms for the Bayes Factor were presented. The first is given by

$$V_{BF_2}(e) = \int V_{LR}(\theta; e_u) \ d\Pi(\theta | e, M_d), \tag{5.16}$$

and the second is given by

$$V_{BF_3}(e) = \left[\int V_{LR}(\theta; e_u)^{-1} \ d\Pi(\theta | e, M_d) \right]^{-1}. \tag{5.17}$$

While the alternative forms of the Bayes Factor given by Equation 5.16 (Equation 5.1 and Equation 5.6) do not eliminate the need for Gibbs sampling techniques, they may still be preferred computationally since there is only a single integral to approximate (instead of two) and because it is easier to characterize the corresponding numerical standard error. Please see Section 5.3.3 for further details of the numerical approximation methods for these alternative Bayes Factor forms.

Additionally, the Monte Carlo integration estimates of the Bayes Factors (especially for the specific source) given in Equation 3.7 and Equation 3.11 are sensitive to the choice of prior distributions for the parameters. The non-standard forms of the Bayes Factors given in Equation 5.1 and Equation 5.6 will aid in the development of various approximations to the value of evidence which are not heavily influenced by the prior choice.

CHAPTER 6

Bernstein-von Mises & Laplace Approximations

6.1 Bernstein-von Mises Approximation

In this section, an asymptotic approximation of the Bayes Factor is introduced as a consequence of the Bernstein-von Mises Theorem. Please refer to Section 2.3 for details of the theorem and for a discussion of its assumptions. The Bernstein-von Mises Theorem gives a set of conditions under which the posterior distribution for the parameter is approximately normal for large sample sizes. This result suggests a natural approximation to the Bayes Factor, the Bernstein-von Mises (BVM) approximation, which is produced by replacing the posterior distribution in the Bayes Factor with the Normal distributions suggested by the Bernstein-von Mises Theorem. The general form of the BVM approximation is given by

$$\hat{V}_{BVM}(e) = \int V_{LR}(\theta; e_u) \ d\Phi(\theta|\hat{\mu}_d, \hat{\Sigma}_d) \tag{6.1}$$

where $\Phi\left(x; \mu, \Sigma\right)$ represents the distribution function for a multivariate normal random variable X with mean μ and covariance matrix Σ, and with corresponding estimates $\hat{\mu}_d$ and $\hat{\Sigma}_d$, respectively, where the estimates are computing from the evidence, e, under the defense model. It should be noted that $\Phi\left(x; \mu, \Sigma\right)$ is the distribution

function corresponding to the measure $\mathcal{N}(\mu, \Sigma)$ on the sample space for X.

The forms of the BVM approximations for the common source and specific source identification problems corresponding to the alternative forms of the Bayes Factor given in Equation 5.1 and Equation 5.6, respectively, are given in Section 6.1.1 and Section 6.1.2 below.

6.1.1 Common Source

Recall from Section 5.1 that the alternative form for the common source Bayes Factor is given by

$$BF_{cs2}(e) = \int LR_{cs}(\theta_a; e_{u_1}, e_{u_2}) \, d\Pi(\theta_a | e_{u_1}, e_{u_2}, e_a, M_d).$$

By the Bernstein-von Mises Theorem, the posterior distribution $\Pi(\theta_a | e_{u_1}, e_{u_2}, e_a, M_d)$ converges in probability under the total variation norm to $\Phi(\theta_a; \hat{\theta}_a^d, \mathcal{I}_d(\hat{\theta}_a^d)^{-1})$ where $\hat{\theta}_a^d$ is the MLE for θ_a given the observation of the entire set of evidence under the defense model, and $\mathcal{I}_d(\hat{\theta}_a^d)^{-1}$ is the inverse of the observed Fisher's information matrix for the observation of the entire set of evidence under the defense model evaluated at $\hat{\theta}_a^d$. Therefore, the BVM approximation of BF_{cs2} is given by

$$BVM_{cs2}(e) = \int LR_{cs}(\theta_a; e_{u_1}, e_{u_2}) \, d\Phi\left(\theta_a; \hat{\theta}_a^d, \mathcal{I}_d(\hat{\theta}_a^d)^{-1}\right). \tag{6.2}$$

In the following theorem, it will be shown that as the number of sources in the alternative source population (n_a) increases to infinity, then the difference between the common source BVM approximation and Bayes Factor converges to zero in probability. To facilitate notation for this theorem, let e_n denote the sequence of evidence sets as n_a increases. For clarity, $e_n = \{e_{u_1}, e_{u_2}, e_a : n_a = n\}$ for $n = 1, 2, \ldots, \infty$. It should be noted that the number of samples for the unknown source evidence n_{u_1} and n_{u_2} are constant, as well as the number of samples within each of the sources from the

alternative source population. The only sample size that is changing is the number of sources in the alternative source population, n_a.

Theorem 6.1:

Suppose that $\hat{\theta}_a^d$, the MLE for the entire set of evidence under the defense model as defined above, is an asymptotically efficient estimator of θ_{a_0}. Let the assumptions of the Bernstein-von Mises Theorem hold, and let the likelihood ratio function $LR_{cs}(\theta_a; e_{u_1}, e_{u_2})$ be bounded in a neighborhood of $\hat{\theta}_a^d$. Then as $n_a \to \infty$

$$\left| BVM_{cs}(e_n) - BF_{cs}(e_n) \right| \xrightarrow{P_{\theta_{a_0}}^n} 0.$$

Proof: Consider the difference in the common source Bayes Factor and BVM approximation.

$$
\begin{aligned}
& \left| BVM_{cs}(e_n) - BF_{cs}(e_n) \right| \\
&= \left| \int LR_{cs}(\theta_a; e_{u_1}, e_{u_2}) \, d\Phi(\theta_a; \hat{\theta}_a^d, \mathcal{I}_d(\hat{\theta}_a^d)^{-1}) \right. \\
&\qquad \left. - \int LR_{cs}(\theta_a; e_{u_1}, e_{u_2}) \, d\Pi(\theta_a | e_n, M_d) \right| \\
&= \left| \int LR_{cs}(\theta_a; e_{u_1}, e_{u_2}) \, d\left[\Phi(\theta_a; \hat{\theta}_a^d, \mathcal{I}_d(\hat{\theta}_a^d)^{-1}) - \Pi(\theta_a | e_n, M_d) \right] \right| \\
&\equiv \left| \int LR_{cs}(\theta_a) \, d\left[\Phi_n(\theta_a) - \Pi_n(\theta_a) \right] \right|
\end{aligned}
$$

First, note that $\Phi_n(\theta_a) - \Pi_n(\theta_a)$ is a signed measure. For any signed measure λ it follows that $|\lambda(A)| \leq ||\lambda||_{TV}(A)$ where $||\lambda||_{TV}$ is the total variation norm [5]. The total variation norm with respect to a measure λ is defined as $||\lambda||_{TV} = \int |d\lambda|$ [73]. Therefore,

$$\left| \int LR_{cs}(\theta_a) \, d\left[\Phi_n(\theta_a) - \Pi_n(\theta_a) \right] \right| \leq \int LR_{cs}(\theta_a) \, d\left| \left[\Phi_n(\theta_a) - \Pi_n(\theta_a) \right] \right|$$

By the assumption that $LR_{cs}(\theta_a)$ is bounded in a neighborhood of $\hat{\theta}_a^d$, then suppose

that $LR_{cs}(\theta_a) \leq C$ for some real number $C > 0$. This implies that, when θ_a is in a neighborhood of $\hat{\theta}_a^d$,

$$
\begin{aligned}
\left| \int LR_{cs}(\theta_a) \, d\left[\Phi_n(\theta_a) - \Pi_n(\theta_a)\right] \right| &\leq \int LR_{cs}(\theta_a) \, d\left| \left[\Phi_n(\theta_a) - \Pi_n(\theta_a)\right] \right| \\
&\leq \int C \, d\left| \left[\Phi_n(\theta_a) - \Pi_n(\theta_a)\right] \right| \\
&\leq C \int d\left| \left[\Phi_n(\theta_a) - \Pi_n(\theta_a)\right] \right| \\
&\leq \left\| \Phi_n(\theta_a) - \Pi_n(\theta_a) \right\|_{TV}.
\end{aligned}
$$

By the Bernstein-von Mises Theorem (Theorem 2.11), as $n_a \to \infty$, then when θ_a is in a neighborhood of $\hat{\theta}_a^d$

$$
\left\| \Phi_n(\theta_a) - \Pi_n(\theta_a) \right\|_{TV} \xrightarrow{P_{\theta_{a_0}}^n} 0.
$$

By the law of total probability, then for any $\varepsilon > 0$, there exists a $\zeta > 0$ such that

$$
\begin{aligned}
P_{\theta_{a_0}}^n &\left(\left\| \Phi_n(\theta_a) - \Pi_n(\theta_{a_0}) \right\|_{TV} > \varepsilon \right) \\
&= P_{\theta_{a_0}}^n \left(\|\Phi_n(\theta_a) - \Pi_n(\theta_a)\|_{TV} > \varepsilon \;\middle|\; \theta_a \in N_\zeta(\hat{\theta}_a^d) \right) \Pi_n(\theta_a \in N_\zeta(\hat{\theta}_a^d)) \\
&\quad + P_{\theta_{a_0}}^n \left(\|\Phi_n(\theta_a) - \Pi_n(\theta_a)\|_{TV} > \varepsilon \;\middle|\; \theta_a \notin N_\zeta(\hat{\theta}_a^d) \right) \Pi_n(\theta_a \notin N_\zeta(\hat{\theta}_a^d)).
\end{aligned}
$$

From the result above $P_{\theta_{a_0}}^n \left(\|\Phi_n(\theta_a) - \Pi_n(\theta_a)\|_{TV} > \varepsilon \;\middle|\; \theta_a \in N_\zeta(\hat{\theta}_a^d) \right) \to 0$. By the Bernstein-von Mises Theorem (Theorem 2.11), then $\Pi_n(\theta_a \in N_\zeta(\hat{\theta}_a^d)) \to 1$ and $\Pi_n(\theta_a \notin N_\zeta(\hat{\theta}_a^d)) \to 0$. In conclusion,

$$
\left| BVM_{cs}(e_n) - BF_{cs}(e_n) \right| \xrightarrow{P_{\theta_{a_0}}^n} 0
$$

by Slutsky's Theorem. ∎

6.1.2 Specific Source

Recall from Section 5.2 that the alternative form of the specific source Bayes Factor is given by

$$BF_{ss2}(e) = \int LR_{ss}(\theta_s, \theta_a; e_u) \, d\Pi(\theta_s, \theta_a | e_u, e_s, e_a, M_d)$$

where the posterior distribution can be given by

$$\Pi(\theta_s, \theta_a | e_u, e_s, e_a, M_d) = \Pi(\theta_s | e_s) \, \Pi(\theta_a | e_u, e_a, M_d)$$

by the assumption that the prior for θ_s is statistically independent of the prior for θ_a. By the Bernstein-von Mises Theorem, the posterior distributions $\Pi(\theta_s, \theta_a | e_u, e_s, e_a, M_d)$ converges in probability under the total variation norm to $\Phi(\theta; \hat{\theta}_d, \mathcal{I}_d(\hat{\theta}_d)^{-1})$. Also, by the Bernstein-von Mises Theorem the posterior distribution $\Pi(\theta_s | e_s)$ and $\Pi(\theta_a | e_u, e_a, M_d)$ converge in probability under the total variation norm to $\Phi(\theta_s; \hat{\theta}_s, \mathcal{I}_s(\hat{\theta}_s)^{-1})$ and $\Phi(\theta_a; \hat{\theta}_a^*, \mathcal{I}_a^*(\hat{\theta}_a^*)^{-1})$, respectively. Therefore, the BVM approximation of BF_{ss2} is given by

$$
\begin{aligned}
BVM_{ss2}(e) &= \int \frac{f(e_u | \theta_s, M_p)}{f(e_u | \theta_a, M_d)} \, d\Phi\left(\theta_s; \hat{\theta}_s, \mathcal{I}_s(\hat{\theta}_s)^{-1}\right) \Phi\left(\theta_a; \hat{\theta}_a^*, \mathcal{I}_a^*(\hat{\theta}_a^*)^{-1}\right) \\
&= \int LR_{ss}(\theta; e_u) \, d\Phi\left(\theta; \hat{\theta}_d, \mathcal{I}_d(\hat{\theta}_d)^{-1}\right)
\end{aligned}
\tag{6.3}
$$

where $\hat{\theta}_s$ is the MLE for θ_s given the observation of e_s (under the defense model), $\hat{\theta}_a^*$ is the MLE for θ_a given the observation of e_u and e_a under the defense model, and $\hat{\theta}_d$ is the MLE for θ given the observation for the entire set of evidence e under the defense model; also, $\mathcal{I}_s(\hat{\theta}_s)^{-1}$ is the inverse of the observed Fisher's information matrix for the observation of e_s (under the defense model), $\mathcal{I}_a^*(\hat{\theta}_a^*)^{-1}$ is the inverse of the observed Fisher's information matrix for the observation of e_u and e_a under the defense model, and $\mathcal{I}_d(\hat{\theta}_d)^{-1}$ is the inverse of the observed Fisher's information matrix

for the observation of the entire set of evidence e under the defense model.

In the following theorem, it will be shown that as the number of sources in the alternative source population (n_a) and the number of samples in the specific source population (n_s) increase to infinity, then the difference between the specific source BVM approximation and Bayes Factor converges to zero in probability. To facilitate notation for this theorem, let e_n denote the sequence of evidence sets as n_a and n_s increase. For clarity, $e_n = \{e_u, e_s, e_a : n_s = n_a = n\}$ for $n = 1, 2, \ldots, \infty$. It should be noted that the number of samples for the unknown source evidence n_u is constant, as well as the number of samples within each of the sources from the alternative source population. For this theorem, n_s and n_a are constrained to be equal (this is the same constraint used for Theorem 5.2).

> **Theorem 6.2:**
> *Suppose that $\hat{\theta}_d$, the MLE for the entire set of evidence under the defense model as defined above, is an asymptotically efficient estimator of θ_0. Let the assumptions of the Bernstein-von Mises Theorem hold, and let the likelihood ratio function $LR_{ss}(\theta; e_u)$ be bounded in a neighborhood of $\hat{\theta}_d$. Then as $n_a \to \infty$ and $n_s \to \infty$*
>
> $$\left| BVM_{ss}(e_n) - BF_{ss}(e_n) \right| \xrightarrow{P_{\theta_0}^n} 0.$$

Proof: Consider the difference in the specific source Bayes Factor and BVM approximation. In a similar manner to the common source proof, suppose that $LR_{ss}(\theta) \le C$ for some real number $C > 0$, then

$$
\begin{aligned}
\left| BVM_{ss}(e_n) - BF_{ss}(e_n) \right| &= \left| \int LR_{ss}(\theta) \, d\left[\Phi_n(\theta) - \Pi_n(\theta) \right] \right| \\
&\le \int LR_{ss}(\theta) \left| d\left[\Phi_n(\theta) - \Pi_n(\theta) \right] \right| \\
&\le \left\| \Phi_n(\theta) - \Pi_n(\theta) \right\|_{TV}.
\end{aligned}
$$

By the Bernstein-von Mises Theorem, as $n_a \to \infty$ and $n_s \to \infty$, then when θ is in a neighborhood of $\hat{\theta}_d$

$$\left|\left|\Phi_n(\theta) - \Pi_n(\theta)\right|\right|_{TV} \xrightarrow{P_{\theta_0}^n} 0.$$

Analogously to the common source proof, by the law of total probability, the Bernstein-von Mises Theorem, and Slutsky's Theorem, then as $n_a \to \infty$ and $n_s \to \infty$

$$\left|BVM_{ss}(e_n) - BF_{ss}(e_n)\right| \xrightarrow{P_{\theta_0}^n} 0.$$

∎

6.1.3   Discussion

The advantage of this method is that the Bernstein-von Mises (BVM) approximation mitigates the effect of the prior choice on the resulting quantification of the value of evidence since it replaces the posterior distribution with a normal approximation. While the effect of prior choice has been alleviated, there is still a large computational component to the BVM approximation since Monte Carlo integration may be necessary in order to estimate the integrals. See Section 6.2.1 for details of the computational methods needed for the BVM approximation. However, one of the computational advantages of this methods is that there is no longer a need to perform Gibbs sampling during the Monte Carlo integration because it is straight-forward to simulate values from a multivariate normal distribution.

6.2   Application Example

The dataset used for this example is the data from the first group of glass windows described in Section 5.3. For this example, the Bernstein-von Mises approximation will

be computed for both the common source and specific source identification problems under the two different scenarios, H_p and H_d, using both the observed and simulated glass datasets described in Section 5.3. The methods of computing the BVM approximations for this example by means of two different Monte Carlo integration methods is provided in Section 6.2.1. Additionally, for each BVM approximation computed, the required Fisher's information matrices will be computed using three different methods, described in the sections to follow.

6.2.1 Methods for Computing the BVM Approximation

Recall from Equation 6.2 that the BVM approximation for the common source identification problem can be written as

$$BVM_{cs} = \int \frac{f(e_{u_1}, e_{u_2} | \theta_a, M_p)}{f(e_{u_1}, e_{u_2} | \theta_a, M_d)} \, d\Phi(\theta_a; \hat{\theta}_a^d, \mathcal{I}_d(\hat{\theta}_a^d)^{-1}).$$

This BVM approximation can be approximated using the Arithmetic Mean method of Monte Carlo integration by

$$\widehat{BVM}_{cs}^{(1)} = n^{-1} \sum_{i=1}^{n} \frac{f(e_{u_1}, e_{u_2} | \theta_a^{(i)}, M_p)}{f(e_{u_1}, e_{u_2} | \theta_a^{(i)}, M_d)} \tag{6.4}$$

where $\{\theta_a^{(i)}\}$ is an independent sample of size n drawn from $\phi(\theta_a; \hat{\theta}_a^d, \mathcal{I}_d(\hat{\theta}_a^d)^{-1})$ where $\hat{\theta}_a^d$ is the MLE corresponding to the objective function $\ell_d(\theta_a) = \ln\left[f(e_{u_1}, e_{u_2}, e_a | \theta_a, M_d)\right]$ and $\mathcal{I}_d(\theta_a)$ is the corresponding Fisher's information matrix. The BVM approximation by method of Harmonic Mean Monte Carlo integration is given by

$$\widehat{BVM}_{cs}^{(2)} = \left[n^{-1} \sum_{i=1}^{n} \frac{f(e_{u_1}, e_{u_2} | \theta_a^{(i)}, M_d)}{f(e_{u_1}, e_{u_2} | \theta_a^{(i)}, M_p)} \right]^{-1} \tag{6.5}$$

where $\{\theta_a^{(i)}\}$ is an independent sample of size n drawn from the importance sampling function $\phi(\theta_a; \hat{\theta}_a^p, \mathcal{I}_p(\hat{\theta}_a^p)^{-1})$ where $\hat{\theta}_a^p$ is the MLE corresponding to the objective function $\ell_p(\theta_a) = \ln\left[f(e_{u_1}, e_{u_2}, e_a | \theta_a, M_p)\right]$ and $\mathcal{I}_p(\theta_a)$ is the corresponding Fisher's information matrix. Note that this sampling distribution is different from the sampling distribution for Equation 6.4 since the given model for the evidence is different.

Now, recall from Equation 6.3 that the specific source BVM approximation is given by

$$BVM_{ss}(e) = \int LR_{ss}(\theta; e_u) \; d\Phi\left(\theta; \hat{\theta}_d, \mathcal{I}_d(\hat{\theta}_d)^{-1}\right)$$

where $\theta = (\theta_s, \theta_a)$ is the joint parameter, $\hat{\theta}_d$ is the MLE corresponding to the objective function $\ell_d(\theta) = \ln\left[f(e_u, e_s, e_a | \theta, M_d)\right]$, and $\mathcal{I}_d(\theta)$ is the corresponding Fisher's information matrix.

The specific source Bayes Factor can be approximated using the Arithmetic Mean method of Monte Carlo integration by

$$\widehat{BVM}_{ss}^{(1)} = n^{-1} \sum_{i=1}^{n} \frac{f(e_u | \theta_s^{(i)}, M_p)}{f(e_u | \theta_a^{(i)}, M_d)} \tag{6.6}$$

where $\{\theta_s^{(i)}\}$ is an independent sample of size n drawn from $\phi(\theta_s; \hat{\theta}_s, \mathcal{I}_s(\hat{\theta}_s)^{-1})$ where $\hat{\theta}_s$ is the MLE corresponding to the objective function $\ell_s(\theta_s) = \ln\left[f(e_s | \theta_s)\right]$ and $\mathcal{I}_s(\theta_s)$ is the corresponding Fisher's information matrix; also, $\{\theta_a^{(i)}\}$ is an independent sample of size n drawn from $\phi(\theta_a; \hat{\theta}_a^*, \mathcal{I}_a^*(\hat{\theta}_a^*)^{-1})$ where $\hat{\theta}_a^*$ is the MLE corresponding to the objective function $\ell_a^*(\theta_a) = \ln\left[f(e_u, e_a | \theta_a)\right]$ and $\mathcal{I}_a^*(\theta_a)$ is the corresponding Fisher's information matrix. Alternatively, instead of considering two independent samples, one for $\{\theta_s^{(i)}\}$ and one for $\{\theta_a^{(i)}\}$, the BVM approximation can be computed using a single joint sample $\{\theta^{(i)}\} = \{(\theta_s^{(i)}, \theta_a^{(i)})\}$ of size n from $\phi(\theta; \hat{\theta}_d, \mathcal{I}_d(\hat{\theta}_d)^{-1})$. However, in this case, the form of the Fisher's information matrix becomes more complicated to determine.

The BVM approximation by method of Harmonic Mean Monte Carlo integration is given by

$$\widehat{BVM}_{ss}^{(2)} = \left[n^{-1} \sum_{i=1}^{n} \frac{f(e_u|\theta_a^{(i)}, M_d)}{f(e_u|\theta_s^{(i)}, M_p)} \right]^{-1} \tag{6.7}$$

where $\{\theta_s^{(i)}\}$ is an independent sample of size n drawn from $\phi(\theta_s; \hat{\theta}_s^*, \mathcal{I}_s^*(\hat{\theta}_s^*)^{-1})$ where $\hat{\theta}_s^*$ is the MLE corresponding to the objective function $\ell_s^*(\theta_s) = \ln\left[f(e_u, e_s|\theta_s)\right]$ and $\mathcal{I}_s^*(\theta_s)$ is the corresponding Fisher's information matrix; also, $\{\theta_a^{(i)}\}$ is an independent sample of size n drawn from $\phi(\theta_a; \hat{\theta}_a, \mathcal{I}_a(\hat{\theta}_a)^{-1})$ where $\hat{\theta}_a$ is the MLE corresponding to the objective function $\ell_a(\theta_a) = \ln\left[f(e_a|\theta_a)\right]$ and $\mathcal{I}_a(\theta_a)$ is the corresponding Fisher's information matrix. Similar to the above approximation, instead of considering two independent samples, the BVM approximation can be computed using a single joint sample $\{\theta^{(i)}\} = \{(\theta_s^{(i)}, \theta_a^{(i)})\}$ of size n from $\phi(\theta; \hat{\theta}_p, \mathcal{I}_p(\hat{\theta}_p)^{-1})$ where $\hat{\theta}_p$ is the MLE corresponding to the objective function $\ell_p(\theta) = \ln\left[f(e_u, e_s, e_a|\theta, M_p)\right]$, and $\mathcal{I}_p(\theta)$ is the corresponding Fisher's information matrix. Note, similarly to the common source BVM approximation, that this sampling distribution is different from the sampling distribution for Equation 6.6 since the given evidence model is different.

It should be noted that in order to sample from the multivariate normal distributions, then the maximum likelihood estimates and the Fisher's information matrices need to be computed first. There are a couple of ways in which the Fisher's information matrices can be computed. Depending on the numerical optimization method used to compute the maximum-likelihood estimates, the Fisher's information matrix may be estimated as a necessary step. If this is not the case, then the following methods may be considered for computing the Fisher's information matrices: the first method is by exact calculation, the second is by parametric bootstrap, and the third is by jackknife resampling [23].

6.2.2 Exact Fisher's Information Matrices

For the first method, the following definition provides the exact form of the Fisher's information matrix for the evidence from the specific source under the assumption of normality.

Definition 6.1: Consider the specific source evidence $e_s = \{\mathbf{y}_{si}\}_{i=1}^{n_s}$ where n_s is the number of samples in the specific source population evidence, $\mathbf{y}_{si}$ is an m-dimensional column vector, and $\mathbf{y}_{si} \overset{iid}{\sim} \mathcal{N}_m(\mu_s, \Sigma_s)$ for $i = 1, 2, \ldots, n_s$ where

$$
\mu_s(\beta) = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{pmatrix} \, , \quad \Sigma_s(\eta) = \begin{pmatrix} \eta_{11} & \eta_{12} & \cdots & \eta_{1m} \\ \eta_{12} & \eta_{22} & \cdots & \eta_{2m} \\ \vdots & \ddots & \ddots & \vdots \\ \eta_{1m} & \cdots & \eta_{m-1,m} & \eta_{mm} \end{pmatrix} \, ,
$$

$$
\text{and } \Sigma_s^{-1}(\sigma) = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{12} & \sigma_{22} & \cdots & \sigma_{2m} \\ \vdots & \ddots & \ddots & \vdots \\ \sigma_{1m} & \cdots & \sigma_{m-1,m} & \sigma_{mm} \end{pmatrix} .
$$

Notice that there are $N_m = \binom{m}{2} + m$ unique elements in $\Sigma_s(\eta)$ and $\Sigma_s^{-1}(\sigma)$ since they are both symmetric. Consider the following vectorization of the unique elements in the matrix $\Sigma_s(\eta)$:

$$
\boldsymbol{\Sigma_s} = \begin{pmatrix} \eta_{11} & \cdots & \eta_{1m} & \eta_{22} & \cdots & \eta_{2m} & \cdots & \eta_{m-1,m-1} & \eta_{m-1,m} & \eta_{mm} \end{pmatrix}^T
$$

which is an N_m-dimensional column vector. Let $\theta_s = (\mu_s^T \ \boldsymbol{\Sigma_s}^T)^T$ be the $(m + N_m)$-dimensional column vector of unique parameter values. Then the *Fisher's information matrix for the specific source evidence* is an $(m + N_m) \times (m + N_m)$ symmetric matrix given by

$$
\mathcal{I}_s(\theta_s) = \begin{bmatrix} n_s \Sigma_s^{-1} & 0 \\ 0 & n_s I(\eta) \end{bmatrix} \tag{6.8}
$$

where $I(\eta)$ is an $N_m \times N_m$ matrix with elements

$$
I_{ijkl}(\eta) = \begin{cases}
\frac{1}{2}\sigma_{ik}^2 & : i = j, \ k = l \\
\sigma_{ik}\sigma_{il} & : i = j, \ k \neq l \\
\sigma_{ik}\sigma_{jk} & : i \neq j, \ k = l \\
\sigma_{il}\sigma_{jk} + \sigma_{ik}\sigma_{jl} & : i \neq j, \ k \neq l
\end{cases}
$$

where the $ijkl$-element is in the row corresponding to the position of η_{ij} in $\boldsymbol{\Sigma_s}$ and the column corresponding to the position of η_{kl} in $\boldsymbol{\Sigma_s}$.

When the evidence follows a multivariate normal distribution, the observed Fisher's information matrix $\mathcal{I}_s(\hat{\theta}_s)$ needed for the specific source BVM approximation using the Arithmetic Mean method of Monte Carlo integration can be computed by replacing θ_s with $\hat{\theta}_s$ in Definition 6.1 for $\mathcal{I}_s(\theta_s)$. When the evidence follows a multivariate normal distribution, the observed Fisher's information matrix $\mathcal{I}_s^*(\hat{\theta}_s^*)$ needed for the specific source BVM approximation using the Harmonic Mean method of Monte Carlo integration can be computed by replacing θ_s with $\hat{\theta}_s^*$ in Definition 6.1 for $\mathcal{I}_s(\theta_s)$.

The exact form of the Fisher's information matrix for the alternative source population evidence under the assumption of normality is given in the following definition.

Definition 6.2: Consider the alternative source population evidence $e_a = \{\mathbf{y}_{ij}\}_{i=1 \ j=1}^{n_a \ n_w}$ where n_w is the number of samples in the ith source, n_a is the number of sources from the alternative source population, and $\mathbf{y}_{ij}$ is an m-dimensional column vector. In addition, $\mathbf{y}_{ij} = \mu_a + a_i + w_{ij}$ with $a_i \overset{iid}{\sim} \mathcal{N}_m(0, \Sigma_b)$ and $w_{ij} \overset{iid}{\sim} \mathcal{N}_m(0, \Sigma_w)$ where

$$
\mu_a(\alpha) = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{pmatrix}, \quad
\Sigma_b(\beta) = \begin{pmatrix}
\beta_{11} & \beta_{12} & \cdots & \beta_{1m} \\
\beta_{12} & \beta_{22} & \cdots & \beta_{2m} \\
\vdots & \ddots & \ddots & \vdots \\
\beta_{1m} & \cdots & \beta_{m-1,m} & \beta_{mm}
\end{pmatrix},
$$

$$\text{and } \Sigma_w(\eta) = \begin{pmatrix} \eta_{11} & \eta_{12} & \cdots & \eta_{1m} \\ \eta_{12} & \eta_{22} & \cdots & \eta_{2m} \\ \vdots & \ddots & \ddots & \vdots \\ \eta_{1m} & \cdots & \eta_{m-1,m} & \eta_{mm} \end{pmatrix}.$$

Notice that there are $N_m = \binom{m}{2} + m$ unique elements in $\Sigma_b(\beta)$ and $\Sigma_w(\eta)$ since they are both symmetric. Consider the following vectorization of the unique elements in the matrix $\Sigma_b(\beta)$:

$$\boldsymbol{\Sigma_b} = \begin{pmatrix} \beta_{11} & \cdots & \beta_{1m} & \beta_{22} & \cdots & \beta_{2m} & \cdots & \beta_{m-1,m-1} & \beta_{m-1,m} & \beta_{mm} \end{pmatrix}^T$$

which is an N_m-dimensional column vector. Similarly, consider the following vectorization of the unique elements in the matrix $\Sigma_w(\eta)$:

$$\boldsymbol{\Sigma_w} = \begin{pmatrix} \eta_{11} & \cdots & \eta_{1m} & \eta_{22} & \cdots & \eta_{2m} & \cdots & \eta_{m-1,m-1} & \eta_{m-1,m} & \eta_{mm} \end{pmatrix}^T$$

which is an N_m-dimensional column vector. Let $\theta_a = (\mu_s^T \ \boldsymbol{\Sigma_b}^T \ \boldsymbol{\Sigma_w}^T)^T$ be the $(m + 2N_m)$-dimensional column vector of unique parameter values. Therefore, $\mathbf{Y}_{ai} = \begin{pmatrix} \mathbf{y}_{i1}^T & \mathbf{y}_{i2}^T & \cdots & \mathbf{y}_{in_w}^T \end{pmatrix}^T \overset{iid}{\sim} \mathcal{N}_{mn_w}(\mu_c, \Sigma_c)$ where

$$\mu_c(\alpha) = \begin{pmatrix} \mu_a \\ \mu_a \\ \vdots \\ \mu_a \end{pmatrix}, \quad \Sigma_c(\beta, \eta) = \begin{pmatrix} \Sigma_b + \Sigma_w & \Sigma_b & \cdots & \Sigma_b \\ \Sigma_b & \Sigma_b + \Sigma_w & \ddots & \vdots \\ \vdots & \ddots & \ddots & \Sigma_b \\ \Sigma_b & \cdots & \Sigma_b & \Sigma_b + \Sigma_w \end{pmatrix}.$$

Define the $mn_w \times mn_w$ matrix

$$\Sigma_c^{-1}(d, s) = \begin{pmatrix} D & S & \cdots & S \\ S & D & \ddots & \vdots \\ \vdots & \ddots & \ddots & S \\ S & \cdots & S & D \end{pmatrix}$$

where D is the $m \times m$ matrix

$$D(d) = \begin{pmatrix} d_{11} & d_{12} & \cdots & d_{1m} \\ d_{12} & d_{22} & \cdots & d_{2m} \\ \vdots & \ddots & \ddots & \vdots \\ d_{1m} & \cdots & d_{m-1,m} & d_{mm} \end{pmatrix}$$

and S is the $m \times m$ matrix

$$S(s) = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1m} \\ s_{12} & s_{22} & \cdots & s_{2m} \\ \vdots & \ddots & \ddots & \vdots \\ s_{1m} & \cdots & s_{m-1,m} & s_{mm} \end{pmatrix}.$$

Then the *Fisher's information matrix for the alternative source population evidence* is the $(m + 2N_m) \times (m + 2N_m)$ matrix given by

$$\mathcal{I}_a(\theta_a) = \begin{bmatrix} n_a I(\alpha) & 0 & 0 \\ 0 & n_a I(\beta) & 0 \\ 0 & 0 & n_a I(\eta) \end{bmatrix} \qquad (6.9)$$

with

$$I(\alpha) = n_w D + n_w(n_w - 1)S$$

$$I(\beta) = n_w^2 \left[I(D) + (n_w - 1)I(DS) + (n_w - 1)^2 I(S) \right]$$

$$I(\eta) = n_w I(D) + n_w(n_w - 1)I(S)$$

where $I(D)$ is an $N_m \times N_m$ matrix with elements

$$I_{ijkl}(D) = \begin{cases} \frac{1}{2}d_{ik}^2 & : i = j, \ k = l \\ d_{ik}d_{il} & : i = j, \ k \neq l \\ d_{ik}d_{jk} & : i \neq j, \ k = l \\ d_{il}d_{jk} + d_{ik}d_{jl} & : i \neq j, \ k \neq l \end{cases},$$

$I(DS)$ is an $N_m \times N_m$ matrix with elements

$$
I_{ijkl}(DS) = \begin{cases}
d_{ik}s_{ik} & : i = j, \ k = l \\
d_{ik}s_{il} + d_{il}s_{ik} & : i = j, \ k \neq l \\
d_{ik}s_{jk} + d_{jk}s_{ik} & : i \neq j, \ k = l \\
d_{ik}s_{jl} + d_{il}s_{jk} + d_{jk}s_{il} + d_{jl}s_{ik} & : i \neq j, \ k \neq l
\end{cases},
$$

and $I(S)$ is an $N_m \times N_m$ matrix with elements

$$
I_{ijkl}(S) = \begin{cases}
\frac{1}{2}s_{ik}^2 & : i = j, \ k = l \\
s_{ik}s_{il} & : i = j, \ k \neq l \\
s_{ik}s_{jk} & : i \neq j, \ k = l \\
s_{il}s_{jk} + s_{ik}s_{jl} & : i \neq j, \ k \neq l
\end{cases}.
$$

When using $I(D)$, $I(DS)$, and $I(S)$ to compute $I(\beta)$ the $ijkl$-element for all three matrices is in the row corresponding to the position of β_{ij} in $\boldsymbol{\Sigma_b}$ and the column corresponding to the position of β_{kl} in $\boldsymbol{\Sigma_b}$. When using $I(D)$ and $I(S)$ to compute $I(\eta)$ the $ijkl$-element for both matrices is in the row corresponding to the position of η_{ij} in $\boldsymbol{\Sigma_w}$ and the column corresponding to the position of η_{kl} in $\boldsymbol{\Sigma_w}$.

When the evidence follows a multivariate normal distribution, the observed Fisher's information matrix $\mathcal{I}_a^*(\hat{\theta}_a^*)$ needed for the specific source BVM approximation using the Arithmetic Mean method of Monte Carlo integration can be computed by replacing θ_a with $\hat{\theta}_a^*$ in Definition 6.2 for $\mathcal{I}_a(\theta_a)$. The observed Fisher's information matrix $\mathcal{I}_a(\hat{\theta}_a)$ needed for the specific source BVM approximation using the Harmonic Mean method of Monte Carlo integration can be computed by replacing θ_a with $\hat{\theta}_a$ in Definition 6.2 for $\mathcal{I}_a(\theta_a)$ when the evidence follows a multivariate normal distribution. When the evidence follows a multivariate normal distribution, the Fisher's information matrix $\mathcal{I}_d(\hat{\theta}_a^d)$ needed for the common source BVM approximation using the Arithmetic Mean method of Monte Carlo integration can be computed by replacing θ_a with $\hat{\theta}_a^d$ in Definition 6.2 for $\mathcal{I}_a(\theta_a)$. The Fisher's information matrix $\mathcal{I}_p(\hat{\theta}_a^p)$ needed for the common source BVM approximation using the Harmonic Mean method of

Monte Carlo integration can be computed by replacing θ_a with $\hat{\theta}_a^p$ in Definition 6.2 for $\mathcal{I}_a(\theta_a)$ when the evidence follows a multivariate normal distribution.

The implementation of these matrices in R, which was aided by the efforts of Dr. Cedric Neumann and Ms. Allison Lempola in relation to the NIJ grant mentioned in the Acknowledgements, provides some computational challenges. In rare cases, the resulting matrices may no longer be symmetric, positive definite, or numerically invertible. In these rare instances, the matrices are forced to have the desired behavior using the functions 'forceSymmetric' and 'nearPD' from the 'Matrix' package [7]. While I admit that these methods are far from ideal, finding solutions to these computational issues are beyond the scope of this dissertation.

6.2.3   Bootstrap Method for Fisher's Information Matrices

For the second method of computing the matrices using a parametric bootstrap, the computational issues mentioned above are avoided. However, this method is more computationally intensive and requires significantly more time to perform. The algorithm for computing the inverse observed Fisher's information matrix associated with the specific source evidence under the assumption of normality using this method is given in Algorithm 3. Algorithm 3 should be used to replace the exact computation of the inverse Fisher's information matrix when it is defined by Definition 6.1. The algorithm for computing the inverse observed Fisher's information matrix associated with the alternative source population evidence under the assumption of normality using this method is given in Algorithm 4. Algorithm 4 should be used to replace the exact computation of the inverse Fisher's information matrix when it is defined by Definition 6.2.

Algorithm 3: Bootstrap for Specific Source Fisher's Information Matrix

Input: $\hat{\mu}_s$ is the MLE for the mean parameter;

$\hat{\Sigma}_s$ is the MLE for the covariance parameter;

n_s is the number of samples in e_s used to compute the MLEs;

m is the sample dimension;

Initialize an empty matrix M of size $N \times (\binom{m}{2} + 2m)$;

for $i = 1, 2, \ldots, N$ **do**

    Simulate a sample e_s^* of size n_s from $\mathcal{N}_m(\hat{\mu}_s, \hat{\Sigma}_s)$;

    Compute $\hat{\mu}_s^*$ the MLE for the mean from e_s^*;

    Compute $\hat{\Sigma}_s^*$ the MLE for the covariance from e_s^*;

    Store the vectorization of the unique elements in $\hat{\mu}_s^*$ and $\hat{\Sigma}_s^*$ into row i of M

end

Compute the covariance matrix of M and store in $invFIM$;

Output: $invFIM$ is the inverse observed Fisher's Information Matrix $\mathcal{I}_s(\hat{\theta}_s)$

Algorithm 4: Bootstrap for Alternative Source Fisher's Information Matrix

Input: $\hat{\mu}_a$ is the MLE for the mean parameter;

$\hat{\Sigma}_b$ is the MLE for the between-source covariance parameter;

$\hat{\Sigma}_w$ is the MLE for the within-source covariance parameter;

n_w is the number of sources in e_a used to compute the MLEs;

n_w is the number of samples in each source of e_a;

m is the sample dimension;

Initialize an empty matrix M of size $N \times (2\binom{m}{2} + 3m)$;

for $i = 1, 2, \ldots, N$ **do**

    Initialize an empty dataset e_a^* of the same size as e_a;

    Simulate a sample a_i of size n_a from $\mathcal{N}_m(\hat{\mu}_a, \hat{\Sigma}_b)$;

    for $j = 1, 2, \ldots, n_a$ **do**

        Simulate a sample of size n_w from $\mathcal{N}_m(a_{ij}, \hat{\Sigma}_w)$;

        Store this sample in e_a^*

    end

    Compute $\hat{\mu}_a^*$ the MLE for the mean from e_a^*;

    Compute $\hat{\Sigma}_b^*$ the MLE for the between-source covariance from e_a^*;

    Compute $\hat{\Sigma}_w^*$ the MLE for the within-source covariance from e_a^*;

    Store vectorization of unique elements in $\hat{\mu}_a^*$, $\hat{\Sigma}_b^*$, and $\hat{\Sigma}_w^*$ into row i of M

end

Compute the covariance matrix of M and store in $invFIM$;

Output: $invFIM$ is the inverse observed Fisher's Information Matrix $\mathcal{I}_a(\hat{\theta}_a)$

6.2.4   Jackknife Method for Fisher's Information Matrices

Since both of the previous algorithms require the assumption of normality for the evidence in order to compute the exact Fisher's information matrices, a jackknife algorithm is utilized in order to avoid any distributional assumptions for the data. As long as the MLEs are consistent, then according to the Linearization of M-estimators Theorem 2.13, the MLEs have an asymptotic variance equal to the corresponding inverse Fisher's information matrix [75]. Since the jackknife estimate for the variance of a parameter estimated via the jackknife method is known [23], it will be used to estimate the inverse Fisher's information matrix. The following algorithms for computing the Fisher's information matrices using the jackknife algorithm utilize this theoretical result. The algorithm for computing the inverse observed Fisher's information matrix associated with the specific source evidence using this method is given in Algorithm 5. The algorithm for computing the inverse observed Fisher's information matrix associated with the alternative source population evidence is given in Algorithm 6.

Algorithm 5: Jackknife for Specific Source Fisher's Information Matrix

---

Input: $e_s = \{y_1, y_2, \ldots, y_N\}$;

Initialize an empty matrix M of size $N \times (\binom{m}{2} + 2m)$;

for $i = 1, 2, \ldots, N$ **do**

> Let $e_s^{(i)}$ be the set obtained by removing y_i from e_s;
> Compute $\hat{\mu}_s^{(i)}$ the MLE for the mean from $e_s^{(i)}$;
> Compute $\hat{\Sigma}_s^{(i)}$ the MLE for the covariance from $e_s^{(i)}$;
> Store the vectorization of unique elements in $\hat{\mu}_s^{(i)}$ and $\hat{\Sigma}_s^{(i)}$ into row i of M;

end

Compute the sample covariance of M and store in C;

$invFIM = (n-1)^2/n\ C$;

Output: $invFIM$ is the inverse observed Fisher's Information Matrix $\mathcal{I}_s(\hat{\theta}_s)$

Algorithm 6: Jacknife for Alternative Source Fisher's Information Matrix

---

Input: $e_a = \{\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_N\}$;

Initialize an empty matrix M of size $N \times (2\binom{m}{2} + 3m)$;

for $i = 1, 2, \ldots, N$ **do**

> Let $e_a^{(i)}$ be the set obtained by removing $\mathbf{Y}_i$ from e_a;
>
> Compute $\hat{\mu}_a^{(i)}$ the MLE for the mean from $e_a^{(i)}$;
>
> Compute $\hat{\Sigma}_b^{(i)}$ the MLE for the covariance from $e_a^{(i)}$;
>
> Compute $\hat{\Sigma}_w^{(i)}$ the MLE for the covariance from $e_a^{(i)}$;
>
> Store vectorization of unique elements in $\hat{\mu}_a^{(i)}$, $\hat{\Sigma}_b^{(i)}$, and $\hat{\Sigma}_w^{(i)}$ into row i of M

end

Compute the sample covariance of M and store in C;

$invFIM = (n-1)^2/n \; C$;

Output: $invFIM$ is the inverse observed Fisher's Information Matrix $\mathcal{I}_a(\hat{\theta}_a)$

6.2.5 Application Results

The Bernstein-von Mises approximations were computed using the R software and a Monte Carlo sample size of $100,000$ iterations. Using the methods described above, the Bernstein-von Mises approximations were computed for both the common source and specific source identification problems under the two different scenarios H_p and H_d, for both the observed and simulated glass datasets as described in Section 5.3. The results for the simulated glass data in Table 6.1. Since the methods performed so poorly on the observed glass data, the results are not reported here.

One of the greatest computational issues with this method, especially for the observed glass data and generally for data with small sample sizes, is the creation of likelihood ratios for the Monte Carlo integration (or reciprocal likelihood ratios for the Harmonic Mean method) which are undefined (in this case undefined can be a likelihood ratio value of $0/0$ or ∞). In the implementation of these methods, any undefined likelihood ratio values are discarded from the Monte Carlo integration and the number discarded is recorded. If this number is too high, resulting in a much smaller Monte Carlo sample size than desired, the method produces a computational error and aborts.

Table 6.1: Computed Bernstein-von Mises approximations (and corresponding MCSE) for the simulated glass data.

(a) Specific Source

Arithmetic Mean	Scenario	
	H_p	H_d
Exact FIM	710123.6 (11723.36)	2.196×10^{-17} (8.659×10^{-18})
Bootstrap FIM	1.13×10^{19} (1.14×10^{19})	3.331×10^{-17} (1.483×10^{-17})
Jackknife FIM	748425.2 (105338.1)	7.615×10^{-20} (2.473×10^{-20})

Harmonic Mean	Scenario	
	H_p	H_d
Exact FIM	2.14×10^{-280} (2.13×10^{-280})	0 (DNE)
Bootstrap FIM	1.56×10^{-261} (1.56×10^{-261})	2.56×10^{-294} (2.56×10^{-294})
Jackknife FIM	0 (DNE)	0 (DNE)

(b) Common Source

Arithmetic Mean	Scenario	
	H_p	H_d
Exact FIM	402.0035 (37.41627)	1.823×10^{-39} (4.966×10^{-40})
Bootstrap FIM	365.8552 (0.555391)	1.921×10^{-39} (8.156×10^{-40})
Jackknife FIM	351.9083 (0.4018931)	4.872×10^{-38} (2.023×10^{-38})

Harmonic Mean	Scenario	
	H_p	H_d
Exact FIM	332.7279 (0.296837)	4.286×10^{-60} (3.564×10^{-60})
Bootstrap FIM	1.25×10^{-38} (1.25×10^{-38})	9.426×10^{-60} (6.626×10^{-60})
Jackknife FIM	323.8939 (0.2832954)	2.022×10^{-64} (1.568×10^{-64})

This was the case with many attempts to compute the BVM approximation for the observed glass data. In general, the Harmonic Mean methods tend to have much higher numbers of discarded values than the Arithmetic Mean methods. These high numbers of discarded values are thought to be the reason why the values of evidence for the specific source BVM approximations computed using the Harmonic Mean method are all numerically zero with corresponding Monte Carlo standard errors which do not exist (labelled DNE in the table).

6.3   Laplace's Approximation

Another numerical approximation technique for the Bayes Factor which can be more computationally efficient than Monte Carlo integration, but which still requires the use of prior distributions for the parameters, is the Laplace's approximation method. The Laplace's approximation for a posterior expectation is given in the following asymptotic approximation result from Robert [59].

Theorem 6.3:

Let $L(\theta) = \ln[f(x|\theta)\pi(\theta)]$ and $L^*(\theta) = \ln[g(\theta)f(x|\theta)\pi(\theta)]$, assuming that f, π, and g are positive. Let $\hat{\theta}$ and $\hat{\theta}^*$ denote the parameter values maximizing L and L^*, respectively. Define $I^L(x)$ and $I^{L^*}(x)$ to be the $(p \times p)$ matrices with (i, j) elements, respectively,

$$I_{ij}^L(x) = -\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}L(\theta)\right]_{\theta=\hat{\theta}}, \qquad I_{ij}^{L^*}(x) = -\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}L^*(\theta)\right]_{\theta=\hat{\theta}^*}.$$

Then, under suitable conditions,

$$E_{\pi(\theta|x)}\big[g(\theta)\big] \approx \left(\frac{\left|I^{L^*}(x)\right|}{\left|I^L(x)\right|}\right)^{-\frac{1}{2}} \frac{g(\hat{\theta}^*)f(x|\hat{\theta}^*)\pi(\hat{\theta}^*)}{f(x|\hat{\theta})\pi(\hat{\theta})}.$$

The Laplace approximation of the Bayes Factor is a straight-forward applications of this result to the first alternative form of the Bayes Factor. The general form of the Laplace approximation is given by

$$\hat{V}_{LP}(e) = \left(\frac{\left|I^{L^*}(x)\right|}{\left|I^L(x)\right|}\right)^{-\frac{1}{2}} \frac{V_{LR}(\hat{\theta}^*; e_u)f(e|\hat{\theta}^*, M_d)\pi(\hat{\theta}^*)}{f(e|\hat{\theta}, M_d)\pi(\hat{\theta})}. \qquad (6.10)$$

It should be noted that there is no practical difference between the Laplace approximation of the first forms for the Bayes Factor given in Equations 3.7 and 3.11 and the Laplace approximation of the alternative forms of the Bayes Factor given in Equations 5.1 and 5.6. The application of the this result to the common source and specific source Bayes Factors are given in Sections 6.3.1 and 6.3.2.

6.3.1 Common Source

The Laplace approximation of the common source Bayes Factor given by Equation 3.7 is a straight-forward application of Theorem 6.3 in both the numerator and denominator of the Bayes Factor. Let $\tilde{\theta}_a$ denote the parameter value maximizing $L(\theta_a) = \ln[f(e_a|\theta_a)\pi(\theta_a)]$ with corresponding Fisher's information matrix $I^L(e_a)$, $\tilde{\theta}_a^p$

denote the parameter value maximizing $L^p(\theta_a) = \ln[f(e_{u_1}, e_{u_2}|\theta_a, M_p)f(e_a|\theta_a)\pi(\theta_a)]$ with corresponding Fisher's information matrix $I^{L^p}(e_a)$, and $\tilde{\theta}_a^d$ denote the parameter value maximizing $L^d(\theta_a) = \ln[f(e_{u_1}, e_{u_2}|\theta_a, M_d)f(e_a|\theta_a)\pi(\theta_a)]$ with corresponding Fisher's information matrix $I^{L^d}(e_a)$. Then the Laplace's approximation of the common source Bayes Factor is given by

$$
\begin{aligned}
LP_{cs}(e) &= \frac{\left(\dfrac{\left|I^{L^p}(e_a)\right|}{\left|I^L(e_a)\right|}\right)^{-\frac{1}{2}} \dfrac{f(e_{u_1}, e_{u_2}|\tilde{\theta}_a^p, M_p)f(e_a|\tilde{\theta}_a^p)\pi(\tilde{\theta}_a^p)}{f(e_a|\tilde{\theta}_a)\pi(\tilde{\theta}_a)}}{\left(\dfrac{\left|I^{L^d}(e_a)\right|}{\left|I^L(e_a)\right|}\right)^{-\frac{1}{2}} \dfrac{f(e_{u_1}, e_{u_2}|\tilde{\theta}_a^d, M_d)f(e_a|\tilde{\theta}_a^d)\pi(\tilde{\theta}_a^d)}{f(e_a|\tilde{\theta}_a)\pi(\tilde{\theta}_a)}} \\[2ex]
&= \left(\frac{\left|I^{L^p}(e_a)\right|}{\left|I^{L^d}(e_a)\right|}\right)^{-\frac{1}{2}} \frac{f(e_{u_1}, e_{u_2}|\tilde{\theta}_a^p, M_p)f(e_a|\tilde{\theta}_a^p)\pi(\tilde{\theta}_a^p)}{f(e_{u_1}, e_{u_2}|\tilde{\theta}_a^d, M_d)f(e_a|\tilde{\theta}_a^d)\pi(\tilde{\theta}_a^d)}. \quad (6.11)
\end{aligned}
$$

If the assumption that both $\tilde{\theta}_a^p$ and $\tilde{\theta}_a^d$ are consistent towards θ_{a_0} can reasonably be met, then it can be shown that the Laplace's approximation of the common source Bayes Factor will converge in probability to the likelihood ratio.

6.3.2 Specific Source

The Laplace approximation of the specific source Bayes Factor given by Equation 3.11 is a straight-forward application of Theorem 6.3 in both the numerator and denominator of the Bayes Factor. Let $\hat{\theta}_s$ denote the parameter value maximizing $L_s(\theta_s) = \ln[f(e_s|\theta_s)\pi(\theta_s)]$ with corresponding Fisher's information matrix $I^{L_s}(e_s)$, $\hat{\theta}_s^*$ denote the parameter value maximizing $L_s^*(\theta_s) = \ln[f(e_u|\theta_s)f(e_s|\theta_s)\pi(\theta_s)]$ with corresponding Fisher's information matrix $I^{L_s^*}(e_s)$, $\hat{\theta}_a$ denote the parameter value maximizing $L_a(\theta_a) = \ln[f(e_a|\theta_a)\pi(\theta_a)]$ with corresponding Fisher's information matrix $I^{L_a}(e_a)$, and $\hat{\theta}_a^*$ denote the parameter value maximizing $L_a^*(\theta_a) = \ln[f(e_u|\theta_a)f(e_a|\theta_a)\pi(\theta_a)]$ with corresponding Fisher's information matrix $I^{L_a^*}(e_a)$. Then the Laplace's approx-

imation of the specific source Bayes Factor is given by

$$
\begin{aligned}
LP_{ss}(e) &= \frac{\left(\dfrac{\left|I^{L_s^*}(e_s)\right|}{\left|I^{L_s}(e_s)\right|}\right)^{-\frac{1}{2}} \dfrac{f(e_u|\tilde{\theta}_s^*)f(e_s|\tilde{\theta}_s^*)\pi(\tilde{\theta}_s^*)}{f(e_s|\tilde{\theta}_s)\pi(\tilde{\theta}_s)}}{\left(\dfrac{\left|I^{L_a^*}(e_a)\right|}{\left|I^{L_a}(e_a)\right|}\right)^{-\frac{1}{2}} \dfrac{f(e_u|\tilde{\theta}_a^*)f(e_a|\tilde{\theta}_a^*)\pi(\tilde{\theta}_a^*)}{f(e_a|\tilde{\theta}_a)\pi(\tilde{\theta}_a)}} \\
&= \left(\frac{\left|I^{L_s^*}(e_s)\right|}{\left|I^{L_s}(e_s)\right|}\right)^{-\frac{1}{2}} \left(\frac{\left|I^{L_a^*}(e_a)\right|}{\left|I^{L_a}(e_a)\right|}\right)^{\frac{1}{2}} \frac{f(e_u|\tilde{\theta}_s^*)f(e_s|\tilde{\theta}_s^*)f(e_a|\tilde{\theta}_a)\pi(\tilde{\theta}_s^*)\pi(\tilde{\theta}_a)}{f(e_u|\tilde{\theta}_a^*)f(e_s|\tilde{\theta}_s)f(e_a|\tilde{\theta}_a^*)\pi(\tilde{\theta}_s)\pi(\tilde{\theta}_a^*)}. \quad (6.12)
\end{aligned}
$$

If the assumption that both $\tilde{\theta}_s$ and $\tilde{\theta}_s^*$ are consistent towards θ_{s_0} can reasonably be met, and if the assumption that both $\tilde{\theta}_a$ and $\tilde{\theta}_a^*$ are consistent towards θ_{a_0} can reasonably be met, then it can be shown that the Laplace's approximation of the specific source Bayes Factor will converge in probability to the likelihood ratio.

6.3.3   Discussion

The Bernstein-von Mises approximation involved the maximum likelihood estimates of the parameters, whereas the Laplace's approximation involves the maximum a-posteriori (MAP) estimates. In this sense, the Laplace's approximation for the Bayes Factor is indirectly influenced by the prior choice. The Laplace's approximation is directly influenced by the prior choice through the expression of the estimate including the prior densities. However, unlike the BVM approximation, the Laplace's approximation of the Bayes Factor doesn't require any Monte Carlo integration. The Fisher's information matrices necessary for the Laplace's approximation can be computed in a similar fashion to the methods for the Fisher's information matrices for the BVM approximation. In addition, Fisher's information matrices related to the prior distributions need to be computed for the Laplace approximation.

CHAPTER 7

Neyman-Pearson Approximation

Up to this point, the focus of the approximation techniques has been on the Bayes Factor. Recall from Section 5 that the likelihood ratio (LR) is the limiting form of the Bayes Factor. In this section, an approximation of the LR is presented which I call the Neyman-Pearson approximation since it is similar to a generalization of the Neyman-Pearson test statistic for non-nested models [69]. It should be noted that this approximation is different from the "plug-in" LR. Where the plug-in LR only considers the likelihood of the unknown source evidence, the Neyman-Pearson (NP) approximation properly considers the likelihood of the entire set of evidence [20]. The NP approximation numerator is the maximum likelihood of observing the entire set of evidence under the prosecution model, and the denominator is the maximum likelihood of observing the entire set of evidence under the defense model. The optimization in the numerator of the NP approximation is performed independently of the optimization in the denominator. The general form of the NP approximation is given by

$$V_{NP}(e) = \frac{\max\limits_{\theta_p \in \Theta_p} f(e|\theta_p, M_p)}{\max\limits_{\theta_d \in \Theta_d} f(e|\theta_d, M_d)} \equiv \frac{f(e|\hat{\theta}_p, M_p)}{f(e|\hat{\theta}_d, M_d)} \qquad (7.1)$$

where Θ_p and Θ_d are the indexing sets for the classes of distributions implied under the prosecution and defense models, respectively. This means that Θ_p and Θ_d are subsets of $\mathbb{R}^k$ such that $\mathcal{P}_p = \{P_{\theta_p} : \theta_p \in \Theta_p\}$ is the class of distributions for the evidence implied by the prosecution proposition and $\mathcal{P}_d = \{P_{\theta_d} : \theta_d \in \Theta_d\}$ is the

class of distributions for the evidence implied by the defense proposition. Please refer back to Chapter 3 for further details. In the remainder of this section, it will be shown that the NP approximation is an asymptotic approximation of the likelihood ratio under certain conditions, including consistency of all the maximum likelihood estimates. This result will be considered separately for the common source and specific source problems in Sections 7.1 and 7.2. In order to facilitate the final result, two intermediary results will be provided under each identification of source problem which involve Theorem 2.12 and Theorem 2.13.

7.1 Common Source

First consider the following maximum likelihood estimates (MLEs) related to the common source likelihood functions: $\hat{\theta}_a$ is the MLE corresponding to the alternative source population evidence e_a, $\hat{\theta}_a^p$ is the MLE corresponding to the entire set of evidence under the prosecution model M_p, and $\hat{\theta}_a^d$ is the MLE corresponding to the entire set of evidence under the defense model M_d. Using these MLEs, the common source Neyman-Pearson approximation is given by

$$NP_{cs}(e) = \frac{f(e_{u_1}, e_{u_2} | \hat{\theta}_a^p, M_p) f(e_a | \hat{\theta}_a^p)}{f(e_{u_1}, e_{u_2} | \hat{\theta}_a^d, M_d) f(e_a | \hat{\theta}_a^d)}. \tag{7.2}$$

The first lemma is necessary to show that if $\hat{\theta}_a$ is consistent for θ_{a_0} under $P_{\theta_{a_0}}$ probability, then both $\hat{\theta}_a^p$ and $\hat{\theta}_a^d$ are also consistent. In order to facilitate the understanding of the lemma, first consider the fixed function $m(\theta_a) = E\left[\ln[f(\mathbf{Y}_{a_i} | \theta_a)]\right]$, where $\mathbf{Y}_{a_i}$ is defined as in Equation 3.17, with a corresponding maximizing value of θ_{a_0}. Next, consider the sequence of functions $m_n(\theta_a) = n_a^{-1} \sum_{i=1}^{n_a} \ln[f(\mathbf{Y}_{a_i} | \theta_a)]$ and note that, by definition, the maximizing value of $m_n(\theta_a)$ is $\hat{\theta}_a$. Also by definition, $\hat{\theta}_a^p$ and $\hat{\theta}_a^d$ are the maximizing values of the sequences of functions $m_n^p(\theta_a) =$

$m_n(\theta_a) + n_a^{-1} \ln[f(e_u|\theta_a, M_p)]$ and $m_n^d(\theta_a) = m_n(\theta_a) + n_a^{-1} \ln[f(e_u|\theta_a, M_d)]$, respectively. The following lemma will show that both $\hat{\theta}_a^p$ and $\hat{\theta}_a^d$ are near maximizers of $m_n(\theta_a)$.

Lemma 7.1:

Let the following assumptions of Theorem 2.12 hold where $m(\theta_a)$ and $m_n(\theta_a)$ are defined as above:

$$\text{Assumption 1:} \quad \sup_{\theta_a : d(\theta_a, \theta_{a_0}) \geq \epsilon} m(\theta_a) \leq m(\theta_{a_0}), \quad \forall \epsilon > 0;$$

$$\text{Assumption 2:} \quad \sup_{\theta_a \in \Theta_a} \left| m_n(\theta_a) - m(\theta_a) \right| \xrightarrow{P_{\theta_{a_0}}} 0, \quad n_a \to \infty;$$

$$\text{Assumption 3:} \quad m_n(\hat{\theta}_a) \geq m_n(\theta_{a_0}) - o_{P_{\theta_{a_0}}}(1), \quad n_a \to \infty.$$

Suppose that $f(e_u|\theta_a, M_p)$ and $f(e_u|\theta_a, M_d)$ are both bounded in $P_{\theta_{a_0}}$-probability in a neighborhood of θ_{a_0}, then $\hat{\theta}_a$, $\hat{\theta}_a^p$, and $\hat{\theta}_a^d$ are consistent for θ_{a_0} as $n_a \to \infty$.

Proof: Consistency of $\hat{\theta}_a$ follows directly from Theorem 2.12. To complete the proof of the lemma, we need to verify that Results 1-3 below hold for $\hat{\theta}_a^h$ for $h = p$ and $h = d$ where $m_n^p(\theta_a)$ and $m_n^d(\theta_a)$ are defined as above.

$$\text{Result 1:} \quad \sup_{\theta_a : d(\theta_a, \theta_{a_0}) \geq \epsilon} m(\theta_a) \leq m(\theta_{a_0}), \quad \forall \epsilon > 0$$

$$\text{Result 2:} \quad \sup_{\theta_a \in \Theta_a} \left| m_n^h(\theta_a) - m(\theta_a) \right| \xrightarrow{P_{\theta_{a_0}}} 0$$

$$\text{Result 3:} \quad m_n^h(\hat{\theta}_a^h) \geq m_n^h(\theta_{a_0}) - o_{P_{\theta_{a_0}}}(1).$$

Result 1: This is exactly Assumption 1.

Result 2: Now, by the Triangle Inequality we have that

$$\sup_{\theta_a \in \Theta_a} \left| m_n^h(\theta_a) - m(\theta_a) \right| = \sup_{\theta_a \in \Theta_a} \left| m_n^h(\theta_a) - m_n(\theta_a) + m_n(\theta_a) - m(\theta_a) \right|$$

$$\leq \sup_{\theta_a \in \Theta_a} \left| m_n^h(\theta_a) - m_n(\theta_a) \right| + \sup_{\theta_a \in \Theta_a} \left| m_n(\theta_a) - m(\theta_a) \right|$$

Since $\ln[f(e_u|\theta_a, M_h)]$ is not a function of n_a, then as $n_a \to \infty$,

$$m_n^h(\theta_a) - m_n(\theta_a) = \frac{1}{n_a} \ln[f(e_u|\theta_a, M_h)] \xrightarrow{P_{\theta_{a_0}}} 0$$

provided that $\ln[f(e_u|\theta_a, M_h)]$ is bounded in probability in a neighborhood of θ_{a_0}. This implies that, as $n_a \to \infty$,

$$\sup_{\theta_a \in \Theta_a} \left| m_n^h(\theta_a) - m_n(\theta_a) \right| \xrightarrow{P_{\theta_{a_0}}} 0.$$

Also, by Assumption 2, then as $n_a \to \infty$,

$$\sup_{\theta_a \in \Theta_a} \left| m_n(\theta_a) - m(\theta_a) \right| \xrightarrow{P_{\theta_{a_0}}} 0.$$

Therefore, by Slutsky's Theorem we obtain, as $n_a \to \infty$,

$$\sup_{\theta_a \in \Theta_a} \left| m_n^h(\theta_a) - m(\theta_a) \right| \xrightarrow{P_{\theta_{a_0}}} 0.$$

Result 3: By Assumption 3 and by definition of $m_n^h(\theta_a) = m_n(\theta_a) + n_a^{-1} \ln[f(e_u|\theta_a, M_h)]$, we have

$$
\begin{aligned}
m_n(\hat{\theta}_a) &\geq m_n(\theta_{a_0}) - o_{P_{\theta_{a_0}}}(1) \\
m_n(\hat{\theta}_a) + o_{P_{\theta_{a_0}}}(1) &\geq m_n(\theta_{a_0}) \\
m_n^h(\hat{\theta}_a) - \frac{1}{n_a} \ln f(e_u|\hat{\theta}_a, M_h) + o_{P_{\theta_{a_0}}}(1) &\geq m_n(\theta_{a_0}) \\
m_n^h(\hat{\theta}_a) - \frac{1}{n_a} \ln f(e_u|\hat{\theta}_a, M_h) + o_{P_{\theta_{a_0}}}(1) &\geq m_n^h(\theta_{a_0}) - \frac{1}{n_a} \ln f(e_u|\theta_{a_0}, M_h).
\end{aligned}
$$

Now, by assumption $\ln[f(e_u|\hat{\theta}_a, M_h)]$ is a continuous function of θ_a and by assumption

$\hat{\theta}_a - \theta_{a_0} \xrightarrow{P_{\theta_{a_0}}} 0$. Therefore, by the Continuous Mapping Theorem,

$$\left[\ln[f(e_u|\theta_{a_0}, M_h)] - \ln[f(e_u|\hat{\theta}_a, M_h)]\right] \xrightarrow{P_{\theta_{a_0}}} 0, \quad \text{as } n_a \to \infty.$$

Therefore, by Slutsky's Theorem,

$$\frac{1}{n_a}\left[\ln[f(e_u|\theta_{a_0}, M_h)] - \ln[f(e_u|\hat{\theta}_a, M_h)]\right] \xrightarrow{P_{\theta_{a_0}}} 0, \quad \text{as } n_a \to \infty.$$

Now, this implies that, as $n_a \to \infty$,

$$m_n^h(\hat{\theta}_a) + \left[\frac{1}{n_a}\ln f(e_u|\theta_{a_0}, M_h) - \frac{1}{n_a}\ln f(e_u|\hat{\theta}_a, M_h)\right] + o_{P_{\theta_{a_0}}}(1) \geq m_n^h(\theta_{a_0})$$
$$m_n^h(\hat{\theta}_a) + o_{P_{\theta_{a_0}}}(1) + o_{P_{\theta_{a_0}}}(1) \geq m_n^h(\theta_{a_0})$$
$$m_n^h(\hat{\theta}_a) + o_{P_{\theta_{a_0}}}(1) \geq m_n^h(\theta_{a_0})$$

since $o_{P_{\theta_{a_0}}}(1) + o_{P_{\theta_{a_0}}}(1) = o_{P_{\theta_{a_0}}}(1)$ by Slutsky's Theorem. Next, we have defined $\hat{\theta}_a^h$ to be the maximizer of the function m_n^h, and therefore, $m_n^h(\hat{\theta}_a^h) \geq m_n^h(\hat{\theta}_a)$.

$$m_n^h(\hat{\theta}_a) + o_{P_{\theta_{a_0}}}(1) \geq m_n^h(\theta_{a_0})$$
$$m_n^h(\hat{\theta}_a^h) + o_{P_{\theta_{a_0}}}(1) \geq m_n^h(\hat{\theta}_a) + o_{P_{\theta_{a_0}}}(1) \geq m_n^h(\theta_{a_0})$$
$$m_n^h(\hat{\theta}_a^h) + o_{P_{\theta_{a_0}}}(1) \geq m_n^h(\theta_{a_0})$$

Now that we have Results 1-3, then by Theorem 2.12, $\hat{\theta}_a^h \xrightarrow{P_{\theta_{a_0}}} \theta_{a_0}$ as $n_a \to \infty$ (i.e. $\hat{\theta}_a^h$ is consistent for both $h = p$ and $h = d$). ∎

Unfortunately, in order to show that the NP approximation converges to the LR, it is not sufficient that all of the MLEs are consistent for θ_{a_0}. It is first necessary to show that the difference between $\hat{\theta}_a^p$ and $\hat{\theta}_a$, and the difference between $\hat{\theta}_a^d$ and $\hat{\theta}_a$, converges to zero in probability at a sufficiently fast rate. This is the result of the

following lemma. In the following sections, the element-wise differential operator ∇ is used.

First, consider the following measurable function $m(\theta_a) = E\,[\,\ln[f(\mathbf{Y}_{a_i}|\theta_a)]\,]$ as given in Lemma 7.1 with corresponding maximizer θ_{a_0}. By definition, $m(\theta_a)$ is differentiable $P_{\theta_{a_0}}$–almost everywhere, with $n_y \times 1$ vector of first partial derivatives $\Psi_{\theta_a} = \nabla m(\theta_a)$ with elements given by

$$\psi_q = E\left[\frac{\partial \ln[f(\mathbf{Y}_{a_i}|\theta_a)]}{\partial \theta_{aq}}\right], \text{ for } q = 1, \ldots, n_y.$$

Furthermore, suppose that $m(\theta_a)$ admits a second-order Taylor expansion at a point of maximum θ_{a_0} with an $n_y \times n_y$ nonsingular, symmetric second derivative matrix $W_{\theta_a} = \nabla\nabla^T m(\theta_a)$ with elements given by

$$w_{q,r} = E\left[\frac{\partial^2 \ln[f(\mathbf{Y}_{a_i}|\theta_a)]}{\partial \theta_{aq}\partial \theta_{ar}}\right], \text{ for } q, r = 1, \ldots, n_y.$$

Also, consider the sequence of functions $m_n(\theta_a) = n_a^{-1}\sum_{i=1}^{n_a}\ln[f(\mathbf{Y}_{a_i}|\theta_a)]$ as given in Lemma 7.1. Note that by definition, the maximizer of $m_n(\theta_a)$ is $\hat{\theta}_a$.

Next, consider the following quantities related to $\hat{\theta}_a^h$ where $h = p$ or $h = d$. First, consider the following measurable function $m^h(\theta_a) = m(\theta_a) + n_a^{-1}E[\,\ln[f(e_u|\theta_a, M_h)]\,]$ with respect to $P_{\theta_{a_0}}$. Also, suppose that $m^h(\theta_a)$ possesses an $n_y \times 1$ vector of first partial derivatives $\Psi_{\theta_a}^h = \nabla m^h(\theta_a) = \Psi_{\theta_a} + \dfrac{1}{n_a}\nabla E[\,\ln[f(e_u|\theta_a, M_h)]\,]$ with elements given by

$$\psi_q^h = \psi_q + \frac{1}{n_a}E\left[\frac{\partial \ln[f(e_u|\theta_a, M_h)]}{\partial \theta_{aq}}\right], \text{ for } q = 1, \ldots, n_y.$$

Furthermore, suppose that $m^h(\theta_a)$ admits a second-order Taylor expansion at a point of maximum θ_{a_0} with an $n_y \times n_y$ nonsingular, symmetric second derivative matrix

$$W_{\theta_a}^h = \nabla\nabla^T m^h(\theta_a) = W_{\theta_a} + \frac{1}{n_a}\nabla\nabla^T \ln[f(e_u|\theta_a, M_h)] \text{ with elements given by}$$

$$w_{q,r}^h = w_{q,r} + \frac{1}{n_a}E\left[\frac{\partial^2 \ln[f(e_u|\theta_a, M_h)]}{\partial\theta_{aq}\partial\theta_{ar}}\right], \text{ for } q, r = 1, \ldots, n_y.$$

Finally, consider the sequence of functions $m_n^h(\theta_a) = m_n(\theta_a) + n_a^{-1}\ln[f(e_u|\theta_a, M_h)]$ as given in the introduction and proof of Lemma 7.1. Note that $\hat{\theta}_a^h$ is the maximizing value of $m_n^h(\theta_a)$. The following lemma will show that the difference between $\hat{\theta}_a^p$ and $\hat{\theta}_a$, and the difference between $\hat{\theta}_a^d$ and $\hat{\theta}_a$, converges to zero in probability at a sufficiently fast rate.

Lemma 7.2:

Let the assumptions of Theorem 2.13 hold for $h = p$ or $h = d$, in particular the following:

Assumption 1:
$$|m(\theta_1) - m(\theta_2)| \leq \Psi_{\theta_{a0}}||\theta_1 - \theta_2||, \quad \forall\ \theta_1,\ \theta_2 \text{ in a neighborhood of } \theta_{a0};$$

Assumption 2:
$$|m^h(\theta_1) - m^h(\theta_2)| \leq \Psi_{\theta_{a0}}^h||\theta_1 - \theta_2||, \quad \forall\ \theta_1,\ \theta_2 \text{ in a neighborhood of } \theta_{a0};$$

Assumption 3: $\ m_n(\hat{\theta}_a) \geq \sup_{\theta_a \in \Theta_a} m_n(\theta_a) - o_{P_{\theta_{a0}}}(n_a^{-1}), \quad n_a \to \infty;$

Assumption 4: $\ m_n^h(\hat{\theta}_a) \geq \sup_{\theta_a \in \Theta_a} m_n^h(\theta_a) - o_{P_{\theta_{a0}}}(n_a^{-1}), \quad n_a \to \infty;$

where m, m_n, Ψ_{θ_a}, W_{θ_a}, m^h, m_n^h, $\Psi_{\theta_a}^h$, and $W_{\theta_a}^h$ are defined as above. If $\hat{\theta}_a$ and $\hat{\theta}_a^h$ are both consistent, then as $n_a \to \infty$

$$\sqrt{n_a}(\hat{\theta}_a^h - \hat{\theta}_a) \xrightarrow{P_{\theta_{a0}}} 0.$$

Proof: By Theorem 2.13 for $\hat{\theta}_a$, we have the following result, as $n_a \to \infty$:

$$\sqrt{n_a}\left(\hat{\theta}_a - \theta_{a0}\right) = -W_{\theta_{a0}}^{-1}\frac{1}{\sqrt{n_a}}\sum_{i=1}^{n_a}\Psi_{\theta_{a0}} + o_{P_{\theta_{a0}}}(1)$$

$$\sqrt{n_a}\,W_{\theta_{a0}}\left(\hat{\theta}_a - \theta_{a0}\right) \equiv -G_{n_a}\Psi_{\theta_{a0}} + o_{P_{\theta_{a0}}}(1).$$

By Theorem 2.13 for $\hat{\theta}_a^h$, we have the following result, as $n_a \to \infty$:

$$\sqrt{n_a}\left(\hat{\theta}_a^h - \theta_{a_0}\right) = -W_{\theta_{a_0}}^{h}{}^{-1} \frac{1}{\sqrt{n_a}} \sum_{i=1}^{n_a} \Psi_{\theta_{a_0}}^h + o_{P_{\theta_{a_0}}}(1)$$

$$\sqrt{n_a}\, W_{\theta_{a_0}}^h\left(\hat{\theta}_a^h - \theta_{a_0}\right) = -\frac{1}{\sqrt{n_a}} \sum_{i=1}^{n_a} \left[\Psi_{\theta_{a_0}} + \frac{1}{n_a}\nabla \ln[f(e_u|\theta_a, M_h)]\right] + o_{P_{\theta_{a_0}}}(1)$$

$$= -\frac{1}{\sqrt{n_a}} \sum_{i=1}^{n_a} \Psi_{\theta_{a_0}} + \frac{1}{\sqrt{n_a}}\nabla \ln[f(e_u|\theta_a, M_h)] + o_{P_{\theta_{a_0}}}(1)$$

$$\equiv -G_{n_a}\Psi_{\theta_{a_0}} + \frac{1}{\sqrt{n_a}}\nabla \ln[f(e_u|\theta_a, M_h)] + o_{P_{\theta_{a_0}}}(1)$$

Since $\nabla \ln[f(e_u|\theta_a, M_h)]$ is not dependent on n_a, then $\frac{1}{\sqrt{n_a}}\nabla \ln[f(e_u|\theta_a, M_h)] = O(n_a^{-\frac{1}{2}})$ as $n_a \to \infty$. By Slutsky's Theorem, we have, as $n_a \to \infty$,

$$\sqrt{n_a}\, W_{\theta_{a_0}}^h\left(\hat{\theta}_a^h - \theta_{a_0}\right) = -G_{n_a}\Psi_{\theta_{a_0}} + O(n_a^{-\frac{1}{2}}) + o_{P_{\theta_{a_0}}}(1)$$

$$= -G_{n_a}\Psi_{\theta_{a_0}} + o_{P_{\theta_{a_0}}}(1)$$

Substituting $\sqrt{n_a}\, W_{\theta_{a_0}}\left(\hat{\theta}_a - \theta_{a_0}\right) = -G_{n_a}\Psi_{\theta_{a_0}} + o_{P_{\theta_{a_0}}}(1)$ into the display above, we get

$$\sqrt{n_a}\, W_{\theta_{a_0}}^h\left(\hat{\theta}_a^h - \theta_{a_0}\right) = \sqrt{n_a}\, W_{\theta_{a_0}}\left(\hat{\theta}_a - \theta_{a_0}\right) + o_{P_{\theta_{a_0}}}(1)$$

Therefore, as $n_a \to \infty$,

$$\sqrt{n_a}\left[W_{\theta_{a_0}}^h(\hat{\theta}_a^h - \theta_{a_0}) - W_{\theta_{a_0}}(\hat{\theta}_a - \theta_{a_0})\right] = o_{P_{\theta_{a_0}}}(1)$$

$$W_{\theta_{a_0}}^h(\hat{\theta}_a^h - \theta_{a_0}) - W_{\theta_{a_0}}(\hat{\theta}_a - \theta_{a_0}) = o_{P_{\theta_{a_0}}}\left(n_a^{-\frac{1}{2}}\right)$$

$$\left[W_{\theta_{a_0}} + \frac{1}{n_a}\nabla\nabla^T \ln[f(e_u|\theta_a, M_h)]\right](\hat{\theta}_a^h - \theta_{a_0}) - W_{\theta_{a_0}}(\hat{\theta}_a - \theta_{a_0}) = o_{P_{\theta_{a_0}}}\left(n_a^{-\frac{1}{2}}\right)$$

$$W_{\theta_{a_0}}(\hat{\theta}_a^h - \theta_{a_0}) + \frac{1}{n_a}\nabla\nabla^T \ln[f(e_u|\theta_a, M_h)](\hat{\theta}_a^h - \theta_{a_0}) - W_{\theta_{a_0}}(\hat{\theta}_a - \theta_{a_0}) = o_{P_{\theta_{a_0}}}\left(n_a^{-\frac{1}{2}}\right)$$

$$W_{\theta_{a_0}}\left[(\hat{\theta}_a^h - \theta_{a_0}) - (\hat{\theta}_a - \theta_{a_0})\right] + \frac{1}{n_a}\nabla\nabla^T \ln[f(e_u|\theta_a, M_h)](\hat{\theta}_a^h - \theta_{a_0}) = o_{P_{\theta_{a_0}}}\left(n_a^{-\frac{1}{2}}\right)$$

$$W_{\theta_{a_0}}(\hat{\theta}_a^h - \hat{\theta}_a) + \frac{1}{n_a}\nabla\nabla^T \ln[f(e_u|\theta_a, M_h)](\hat{\theta}_a^h - \theta_{a_0}) = o_{P_{\theta_{a_0}}}\left(n_a^{-\frac{1}{2}}\right)$$

Since $\nabla\nabla^T \ln[f(e_u|\theta_a, M_h)]$ is not dependent on n_a then $\frac{1}{n_a}\nabla\nabla^T \ln[f(e_u|\theta_a, M_h)] = O(n_a^{-1})$ as $n_a \to \infty$. Since $\hat{\theta}_a^h$ is consistent, then $(\hat{\theta}_a^h - \theta_{a0}) = o_{P_{\theta_{a0}}}(1)$ as $n_a \to \infty$. Thus, by Slutksy's Theorem, as $n_a \to \infty$,

$$W_{\theta_{a0}}\left(\hat{\theta}_a^h - \hat{\theta}_a\right) + \frac{1}{n_a}\nabla\nabla^T \ln f(e_u|\theta_a, M_h)\left(\hat{\theta}_a^h - \theta_{a0}\right) = o_{P_{\theta_{a0}}}\left(n_a^{-\frac{1}{2}}\right)$$

$$W_{\theta_{a0}}\left(\hat{\theta}_a^h - \hat{\theta}_a\right) + O(n_a^{-1})o_{P_{\theta_{a0}}}(1) = o_{P_{\theta_{a0}}}\left(n_a^{-\frac{1}{2}}\right)$$

$$W_{\theta_{a0}}\left(\hat{\theta}_a^h - \hat{\theta}_a\right) + o_{P_{\theta_{a0}}}(n_a^{-1}) = o_{P_{\theta_{a0}}}\left(n_a^{-\frac{1}{2}}\right)$$

$$W_{\theta_{a0}}\left(\hat{\theta}_a^h - \hat{\theta}_a\right) = o_{P_{\theta_{a0}}}\left(n_a^{-\frac{1}{2}}\right) - o_{P_{\theta_{a0}}}\left(n_a^{-1}\right)$$

$$W_{\theta_{a0}}\left(\hat{\theta}_a^h - \hat{\theta}_a\right) = o_{P_{\theta_{a0}}}\left(n_a^{-\frac{1}{2}}\right)$$

$$\sqrt{n_a}\, W_{\theta_{a0}}\left(\hat{\theta}_a^h - \hat{\theta}_a\right) = o_{P_{\theta_{a0}}}(1)$$

Finally, because matrix inverses are continuous functions, then by the Continuous Mapping Theorem and Slutsky's Theorem

$$\sqrt{n_a}\left(\hat{\theta}_a^h - \hat{\theta}_a\right) = W_{\theta_{a0}}^{-1}o_{P_{\theta_{a0}}}(1)$$

$$\sqrt{n_a}\left(\hat{\theta}_a^h - \hat{\theta}_a\right) = o_{P_{\theta_{a0}}}(1), \quad \text{as } n_a \to \infty.$$

$\blacksquare$

Now that the necessary preliminary results are in place, the following theorem provides conditions under which the common source Neyman-Pearson approximation converges in probability to the true likelihood ratio as the number of sources in the alternative source population increases without bound.

Theorem 7.3:
Let the assumptions of Lemma 7.1 and Lemma 7.2 be satisfied. Then as $n_a \to \infty$,

$$NP_{cs}(e_n) \xrightarrow{P_{\theta_{a0}}} LR_{cs}(\theta_{a0}; e_{u_1}, e_{u_2}).$$

Proof: Consider the Neyman-Pearson approximation for the common source problem

$$
\begin{aligned}
NP_{cs}(e_n) &= \frac{f(e_{u_1}, e_{u_2}|\hat{\theta}_a^p, M_p)f(e_a|\hat{\theta}_a^p)}{f(e_{u_1}, e_{u_2}|\hat{\theta}_a^d, M_d)f(e_a|\hat{\theta}_a^d)} \\
&= \frac{f(e_{u_1}, e_{u_2}|\hat{\theta}_a^p, M_p)}{f(e_{u_1}, e_{u_2}|\hat{\theta}_a^d, M_d)} \times \frac{f(e_a|\hat{\theta}_a^p)}{f(e_a|\hat{\theta}_a^d)} \\
&\equiv U \times A
\end{aligned}
$$

The goal of the proof is to show that $A \xrightarrow{P_{\theta_{a_0}}} 1$ and $U \xrightarrow{P_{\theta_{a_0}}} LR_{cs}(\theta_{a_0}; e_{u_1}, e_{u_2})$ as $n_a \to \infty$. First, consider

$$
A = \frac{f(e_a|\hat{\theta}_a^p)}{f(e_a|\hat{\theta}_a^d)} = \frac{\prod_{i=1}^{n_a} f(\mathbf{Y}_{a_i}|\hat{\theta}_a^p)}{\prod_{i=1}^{n_a} f(\mathbf{Y}_{a_i}|\hat{\theta}_a^d)},
$$

then we have

$$
\begin{aligned}
\ln(A) &= \ln\left(\frac{\prod_{i=1}^{n_a} f(\mathbf{Y}_{a_i}|\hat{\theta}_a^p)}{\prod_{i=1}^{n_a} f(\mathbf{Y}_{a_i}|\hat{\theta}_a^d)}\right) \\
&= \sum_{i=1}^{n_a} \ln f(\mathbf{Y}_{a_i}|\hat{\theta}_a^p) - \sum_{i=1}^{n_a} \ln f(\mathbf{Y}_{a_i}|\hat{\theta}_a^d) \\
&\equiv \ell_A(\hat{\theta}_a^p) - \ell_A(\hat{\theta}_a^d).
\end{aligned}
$$

Now, let $\tilde{\theta}_a^p$ be on the line between $\hat{\theta}_a$ and $\hat{\theta}_a^p$. Let ∇ be the derivative operator (i.e. ∇f is the vector of first partial derivatives, $\nabla\nabla^T f$ is the matrix of second partial derivatives, and so on). Using the second-order Multivariate Taylor's Expansion (Wade [77] p. 421) with the Mean Value Theorem, then

$$
\ell_A(\hat{\theta}_a^p) = \ell_A(\hat{\theta}_a) + (\hat{\theta}_a^p - \hat{\theta}_a)^T \nabla \ell_A(\hat{\theta}_a) + \frac{1}{2}(\hat{\theta}_a^p - \hat{\theta}_a)^T \nabla\nabla^T \ell_A(\tilde{\theta}_a^p)(\hat{\theta}_a^p - \hat{\theta}_a).
$$

Similarly, let $\tilde{\theta}_a^d$ be on the line between $\hat{\theta}_a$ and $\hat{\theta}_a^d$. Using the second-order Multivariate

Taylor's Expansion (Wade [77] p. 421) with the Mean Value Theorem, then

$$\ell_A(\hat{\theta}_a^d) = \ell_A(\hat{\theta}_a) + (\hat{\theta}_a^d - \hat{\theta}_a)^T \nabla \ell_A(\hat{\theta}_a) + \frac{1}{2}(\hat{\theta}_a^d - \hat{\theta}_a)^T \nabla \nabla^T \ell_A(\tilde{\theta}_a^d)(\hat{\theta}_a^d - \hat{\theta}_a).$$

Therefore,

$$
\begin{aligned}
\ln(A) &= \ell_A(\hat{\theta}_a^p) - \ell_A(\hat{\theta}_a^d) \\
&= \ell_A(\hat{\theta}_a) + (\hat{\theta}_a^p - \hat{\theta}_a)^T \nabla \ell_A(\hat{\theta}_a) + \frac{1}{2}(\hat{\theta}_a^p - \hat{\theta}_a)^T \nabla \nabla^T \ell_A(\tilde{\theta}_a^p)(\hat{\theta}_a^p - \hat{\theta}_a) \\
&\quad - \ell_A(\hat{\theta}_a) - (\hat{\theta}_a^d - \hat{\theta}_a)^T \nabla \ell_A(\hat{\theta}_a) - \frac{1}{2}(\hat{\theta}_a^d - \hat{\theta}_a)^T \nabla \nabla^T \ell_A(\tilde{\theta}_a^d)(\hat{\theta}_a^d - \hat{\theta}_a) \\
&= \left[(\hat{\theta}_a^p - \hat{\theta}_a)^T - (\hat{\theta}_a^d - \hat{\theta}_a)^T \right] \nabla \ell_A(\hat{\theta}_a) + \frac{1}{2}(\hat{\theta}_a^p - \hat{\theta}_a)^T \nabla \nabla^T \ell_A(\tilde{\theta}_a^p)(\hat{\theta}_a^p - \hat{\theta}_a) \\
&\quad - \frac{1}{2}(\hat{\theta}_a^d - \hat{\theta}_a)^T \nabla \nabla^T \ell_A(\tilde{\theta}_a^d)(\hat{\theta}_a^d - \hat{\theta}_a).
\end{aligned}
$$

Consider $\nabla \ell_A(\hat{\theta}_a)$ which is the derivative of the log-likelihood function evaluated at the maximum likelihood estimate $\hat{\theta}_a$. By definition of $\hat{\theta}_a$ in the introduction to Lemmas 7.1 and 7.2, and by properties of MLE's, then $\nabla \ell_A(\hat{\theta}_a) = 0$. Thus,

$$\ln(A) = \frac{1}{2}(\hat{\theta}_a^p - \hat{\theta}_a)^T \nabla \nabla^T \ell_A(\tilde{\theta}_a^p)(\hat{\theta}_a^p - \hat{\theta}_a) - \frac{1}{2}(\hat{\theta}_a^d - \hat{\theta}_a)^T \nabla \nabla^T \ell_A(\tilde{\theta}_a^d)(\hat{\theta}_a^d - \hat{\theta}_a).$$

Next, for both $h = p$ and $h = d$, consider

$$\nabla \nabla^T \ell_A(\tilde{\theta}_a^h) = \nabla \nabla^T \left[\sum_{i=1}^{n_a} \ln[f(\mathbf{Y}_{a_i}|\tilde{\theta}_a^h)] \right] = \sum_{i=1}^{n_a} \nabla \nabla^T \ln[f(\mathbf{Y}_{a_i}|\tilde{\theta}_a^h)].$$

Because $\hat{\theta}_a^h$ and $\hat{\theta}_a$ are consistent by Lemma 7.1 and $\tilde{\theta}_a^h$ is on the line between the two, then $\tilde{\theta}_a^h$ will be in a neighborhood of θ_{a_0}. By the Strong Law of Large Numbers, then as $n_a \to \infty$

$$\frac{1}{n_a} \sum_{i=1}^{n_a} \nabla \nabla^T \ln[f(\mathbf{Y}_{a_i}|\tilde{\theta}_a^h)] \xrightarrow{as} \nabla \nabla^T E\left[\ln[f(\mathbf{Y}_{a_i}|\theta_a)] \right]$$

where the convergence of the matrix is element-wise. By assumption, $E\left[\ln[f(\mathbf{Y}_{a_i}|\theta_a)]\right]$ is finite. Therefore, $\nabla\nabla^T E\left[\ln[f(\mathbf{Y}_{a_i}|\theta_a)]\right]$ is also finite. Thus, as $n_a \to \infty$

$$\frac{1}{n_a}\nabla\nabla^T \ell_A(\tilde{\theta}_a^h) = O_{P_{\theta_{a_0}}}(1).$$

Also, by Lemma 7.2, as $n_a \to \infty$

$$\sqrt{n_a}\left(\hat{\theta}_a^h - \hat{\theta}_a\right) = o_{P_{\theta_{a_0}}}(1).$$

So, by Slutsky's Theorem, as $n_a \to \infty$

$$
\begin{aligned}
\ln(A) &= \frac{1}{2}(\hat{\theta}_a^p - \hat{\theta}_a)^T \nabla\nabla^T \ell_A(\tilde{\theta}_a^p)(\hat{\theta}_a^p - \hat{\theta}_a) - \frac{1}{2}(\hat{\theta}_a^d - \hat{\theta}_a)^T D^2 \ell_A(\tilde{\theta}_a^d)(\hat{\theta}_a^d - \hat{\theta}_a) \\
&= \frac{1}{2}\left[\sqrt{n_a}(\hat{\theta}_a^p - \hat{\theta}_a)^T\right]\left[\frac{1}{n_a}\nabla\nabla^T \ell_A(\tilde{\theta}_a^p)\right]\left[\sqrt{n_a}(\hat{\theta}_a^p - \hat{\theta}_a)\right] \\
&\quad - \frac{1}{2}\left[\sqrt{n_a}(\hat{\theta}_a^d - \hat{\theta}_a)^T\right]\left[\frac{1}{n_a}\nabla\nabla^T \ell_A(\tilde{\theta}_a^d)\right]\left[\sqrt{n_a}(\hat{\theta}_a^d - \hat{\theta}_a)\right] \\
&= \frac{1}{2}o_{P_{\theta_{a_0}}}(1)\, O_{P_{\theta_{a_0}}}(1)\, o_{P_{\theta_{a_0}}}(1) - \frac{1}{2}o_{P_{\theta_{a_0}}}(1)\, O_{P_{\theta_{a_0}}}(1)\, o_{P_{\theta_{a_0}}}(1) \\
&= o_{P_{\theta_{a_0}}}(1) + o_{P_{\theta_{a_0}}}(1) \\
&= o_{P_{\theta_{a_0}}}(1).
\end{aligned}
$$

This implies that $\ln(A) \xrightarrow{P_{\theta_{a_0}}} 0$, as $n_a \to \infty$. Since the exponential function is continuous, then by the Continuous Mapping Theorem, as $n_a \to \infty$

$$A = e^{\ln(A)} \xrightarrow{P_{\theta_{a_0}}} e^0 = 1.$$

To complete the proof, we need to show that, as $n_a \to \infty$

$$U = \frac{f(e_{u_1}, e_{u_2}|\hat{\theta}_a^p, M_p)}{f(e_{u_1}, e_{u_2}|\hat{\theta}_a^d, M_d)} \xrightarrow{P_{\theta_{a_0}}} LR_{cs}(\theta_{a_0}; e_{u_1}, e_{u_2}) = \frac{f(e_{u_1}, e_{u_2}|\theta_{a_0}, M_p)}{f(e_{u_1}, e_{u_2}|\theta_{a_0}, M_d)}.$$

By Lemma 7.1, we have that $\hat{\theta}_a^p \xrightarrow{P_{\theta_{a_0}}} \theta_{a_0}$ and $\hat{\theta}_a^d \xrightarrow{P_{\theta_{a_0}}} \theta_{a_0}$, as $n_a \to \infty$. By the Continuous Mapping Theorem, as $n_a \to \infty$

$$f(e_{u_1}, e_{u_2} | \hat{\theta}_a^p, M_p) \xrightarrow{P_{\theta_{a_0}}} f(e_{u_1}, e_{u_2} | \theta_{a_0}, M_p)$$

and

$$f(e_{u_1}, e_{u_2} | \hat{\theta}_a^d, M_d) \xrightarrow{P_{\theta_{a_0}}} f(e_{u_1}, e_{u_2} | \theta_{a_0}, M_d).$$

Finally, use Slutsky's Theorem to obtain the result. In conclusion, as $n_a \to \infty$

$$NP_{cs}(e_n) \xrightarrow{P_{\theta_{a_0}}} LR_{cs}(\theta_{a_0}; e_{u_1}, e_{u_2}).$$

∎

7.2 Specific Source

Next, the specific source Neyman-Pearson approximation is given by

$$NP_{ss}(e) = \frac{f(e_u | \hat{\theta}_s^*, M_p) f(e_s | \hat{\theta}_s^*) f(e_a | \hat{\theta}_a)}{f(e_u | \hat{\theta}_a^*, M_d) f(e_s | \hat{\theta}_s) f(e_a | \hat{\theta}_a^*)} \tag{7.3}$$

where $\hat{\theta}_s^*$ is the MLE for θ_s corresponding to e_u and e_s under the prosecution model, $\hat{\theta}_a$ is the MLE for θ_a corresponding to e_a under the prosecution model, $\hat{\theta}_a^*$ is the MLE for θ_a corresponding to e_u and e_a under the defense model, and $\hat{\theta}_s$ is the MLE for θ_s corresponding e_s under the defense model.

Similar to the common source, the first lemma is necessary to show that if $\hat{\theta}_a$ is consistent for θ_{a_0} under $P_{\theta_{a_0}}$ probability, then $\hat{\theta}_a^*$ is also consistent, and if $\hat{\theta}_s$ is consistent for θ_{s_0} under $P_{\theta_{s_0}}$ probability, then $\hat{\theta}_s^*$ is also consistent. In order to facilitate the

understanding of the lemma, for either $k = a$ or $k = s$, consider the fixed function $m(\theta_k) = E\left[\ln[f(\mathbf{y}_{k_i}|\theta_k)]\right]$, where $\mathbf{y}_{k_i} = \mathbf{Y}_{a_i}$ for $k = a$ is defined as in Equation 3.17 with a corresponding maximizing value of θ_{a_0}, and where $\mathbf{y}_{s_i}$ for $k = s$ is defined as in Equation 3.15 with a corresponding maximizing value of θ_{s_0}. Next, consider the sequence of functions $m_n(\theta_k) = n_k^{-1} \sum_{i=1}^{n_k} \ln[f(\mathbf{y}_{k_i}|\theta_k)]$ and note that, by definition, the maximizing value of $m_n(\theta_k)$ is $\hat{\theta}_k$. Also by definition, $\hat{\theta}_k^*$ is the maximizing value of the sequence of functions $m_n^*(\theta_k) = m_n(\theta_k) + n_k^{-1} \ln[f(e_u|\theta_k)]$. The following lemma will show that $\hat{\theta}_a^*$ and $\hat{\theta}_s^*$ are near maximizers of $m_n(\theta_a)$ and $m_n(\theta_s)$, respectively.

Lemma 7.4:

Let $k = a$ or $k = s$. Let the following assumptions of Theorem 2.12 hold where $m(\theta_k)$ and $m_n(\theta_k)$ are defined as above:

Assumption 1: $\displaystyle\sup_{\theta_k : d(\theta_k, \theta_{k_0}) \geq \epsilon} m(\theta_k) \leq m(\theta_{k_0}), \quad \forall \epsilon > 0;$

Assumption 2: $\displaystyle\sup_{\theta_k \in \Theta_k} \left| m_n(\theta_k) - m(\theta_k) \right| \xrightarrow{P_{\theta_{k_0}}} 0;$

Assumption 3: $m_n(\hat{\theta}_k) \geq m_n(\theta_{k_0}) - o_{P_{\theta_{k_0}}}(1).$

Suppose that $f(e_u|\theta_k)$ is bounded in $P_{\theta_{k_0}}$-probability in a neighborhood of θ_{k_0}, then $\hat{\theta}_k$ and $\hat{\theta}_k^*$ are consistent.

Proof: Consistency of $\hat{\theta}_k$ follows directly from Theorem 2.12. To complete the proof of the lemma, we need to verify that Results 1-3 below hold for $\hat{\theta}_a^k$ for $k = a$ and $k = s$ where $m_n^*(\theta_k)$ is defined as above.

Result 1: $\displaystyle\sup_{\theta_k : d(\theta_k, \theta_{k_0}) \geq \epsilon} m(\theta_k) \leq m(\theta_{k_0}), \quad \forall \epsilon > 0;$

Result 2: $\displaystyle\sup_{\theta_k \in \Theta_k} \left| m_n^*(\theta_k) - m(\theta_k) \right| \xrightarrow{P_{\theta_{k_0}}} 0;$

Result 3: $m_n^*(\hat{\theta}_k^*) \geq m_n^*(\theta_{k_0}) - o_{P_{\theta_{k_0}}}(1).$

Result 1: This is exactly Assumption 1.

Result 2: This proof is analogous to the corresponding proof for Lemma 7.1. Now, by the Triangle Inequality, by Assumption 2, and by Slutsky's Theorem, as $n_k \to \infty$

$$\sup_{\theta_k \in \Theta_k} \left| m_n^*(\theta_k) - m(\theta_k) \right| \xrightarrow{P_{\theta_{k_0}}} 0.$$

Result 3: This proof is analogous to the corresponding proof for Lemma 7.1. By Assumption 3, then as $n_k \to \infty$

$$
\begin{aligned}
m_n(\hat{\theta}_k) &\geq m_n(\theta_{k_0}) - o_{P_{\theta_{k_0}}}(1) \\
m_n^*(\hat{\theta}_k) - \frac{1}{n_k} \ln f(e_u | \hat{\theta}_k) + o_{P_{\theta_{k_0}}}(1) &\geq m_n^*(\theta_{k_0}) - \frac{1}{n_k} \ln f(e_u | \theta_{k_0}) \\
m_n^*(\hat{\theta}_k^*) + o_{P_{\theta_{k_0}}}(1) \geq m_n^*(\hat{\theta}_k) + o_{P_{\theta_{k_0}}}(1) &\geq m_n^*(\theta_{k_0}) \\
m_n^*(\hat{\theta}_k^*) + o_{P_{\theta_{k_0}}}(1) &\geq m_n^*(\theta_{k_0}).
\end{aligned}
$$

Now that we have Results 1-3, then by Theorem 2.12 $\hat{\theta}_k^* \xrightarrow{P_{\theta_{k_0}}} \theta_{k_0}$, as $n_k \to \infty$ (i.e. $\hat{\theta}_k^*$ is consistent). ∎

Similar to the common source problem, it is first necessary to show that the difference between $\hat{\theta}_a^*$ and $\hat{\theta}_a$, and the difference between $\hat{\theta}_s^*$ and $\hat{\theta}_s$, converges to zero in probability at a sufficiently fast rate. This is the result of the following lemma.

First, consider the following measurable function for $k = a$ or $k = s$, $m(\theta_k) = E\left[\ln[f(\mathbf{y}_{k_i} | \theta_k)] \right]$ as given in Lemma 7.4 which is differentiable $P_{\theta_{k_0}}-$almost everywhere by definition, with $n_y \times 1$ vector of first partial derivatives $\Psi_{\theta_k} = \nabla m(\theta_k)$ with elements given by

$$\psi_q = E\left[\frac{\partial \ln[f(\mathbf{y}_{k_i} | \theta_k)]}{\partial \theta_{kq}} \right], \text{ for } q = 1, \dots, n_y.$$

Furthermore, suppose that $m(\theta_k)$ admits a second-order Taylor expansion at a point

of maximum θ_{k_0} with an $n_y \times n_y$ nonsingular, symmetric second derivative matrix $W_{\theta_k} = \nabla\nabla^T m(\theta_k)$ with elements given by

$$w_{q,r} = E\left[\frac{\partial^2 \ln[f(\mathbf{y}_{k_i}|\theta_k)]}{\partial\theta_{kq}\partial\theta_{kr}}\right], \text{ for } q, r = 1, \ldots, n_y.$$

Also, consider the sequence of functions $m_n(\theta_k) = n_k^{-1} \sum_{i=1}^{n_k} \ln[f(\mathbf{y}_{k_i}|\theta_k)]$ as given in Lemma 7.4. By definition, $\hat{\theta}_k$ is the maximizer of $m_n(\theta_k)$.

Next, consider the following quantities related to $\hat{\theta}_k^*$ where $k = a$ or $k = s$. First, consider the following measurable function $m^*(\theta_k) = m(\theta_k) + n_k^{-1}E[\ln[f(e_u|\theta_k)]]$ with respect to the measure $P_{\theta_{k_0}}$, and suppose that it possesses an $n_y \times 1$ vector of first partial derivatives $\Psi_{\theta_k}^* = \nabla m^*(\theta_k) = \Psi_{\theta_k} + \frac{1}{n_k}\nabla E[\ln[f(e_u|\theta_k)]]$ with elements given by

$$\psi_q^h = \psi_q + \frac{1}{n_k}E\left[\frac{\partial \ln[f(e_u|\theta_k)]}{\partial\theta_{kq}}\right], \text{ for } q = 1, \ldots, n_y.$$

Furthermore, suppose that $m^*(\theta_k)$ admits a second-order Taylor expansion at a point of maximum θ_{k_0} with an $n_y \times n_y$ nonsingular, symmetric second derivative matrix $W_{\theta_k}^* = \nabla\nabla^T m^*(\theta_k) = W_{\theta_k} + \frac{1}{n_k}\nabla\nabla^T \ln[f(e_u|\theta_k)]$ with elements given by

$$w_{q,r}^h = w_{q,r} + \frac{1}{n_k}E\left[\frac{\partial^2 \ln[f(e_u|\theta_k)]}{\partial\theta_{kq}\partial\theta_{kr}}\right], \text{ for } q, r = 1, \ldots, n_y.$$

Finally, consider the sequence of functions $m_n^*(\theta_k) = m_n(\theta_k) + n_k^{-1}\ln[f(e_u|\theta_k)]$ as given in the introduction and proof of Lemma 7.4. By definition, then $\hat{\theta}_k^*$ is the maximizer of $m_n^*(\theta_k)$. The following lemma will show that the difference between $\hat{\theta}_a^*$ and $\hat{\theta}_a$, and the difference between $\hat{\theta}_s^*$ and $\hat{\theta}_s$, converges to zero in probability at a sufficiently fast rate.

Lemma 7.5:
Let the assumptions of Theorem 2.13 hold for $k = a$ or $k = s$, in particular the following:

Assumption 1:

$|m(\theta_{k_1}) - m(\theta_{k_2})| \leq \Psi_{\theta_{k_0}} ||\theta_{k_1} - \theta_{k_2}|| \quad \forall \, \theta_{k_1}, \, \theta_{k_2}$ *in a neighborhood of* θ_{k_0};

Assumption 2:

$|m^*(\theta_{k_1}) - m^*(\theta_{k_2})| \leq \Psi^*_{\theta_{k_0}} ||\theta_{k_1} - \theta_{k_k}|| \; \forall \, \theta_{k_1}, \, \theta_{k_2}$ *in a neighborhood of* θ_{k_0};

Assumption 3: $m_n(\hat{\theta}_k) \geq \sup_{\theta_k \in \Theta_k} m_n(\theta_k) - o_{P_{\theta_{k_0}}}(n_k^{-1}), \quad n_k \to \infty$;

Assumption 4: $m_n^*(\hat{\theta}_k) \geq \sup_{\theta_k \in \Theta_k} m_n^*(\theta_k) - o_{P_{\theta_{k_0}}}(n_k^{-1}), \quad n_k \to \infty$;

where $m, \, m_n, \, \Psi_{\theta_k}, \, W_{\theta_k}, \, m^*, \, m_n^*, \, \Psi^*_{\theta_k},$ *and* $W^*_{\theta_k}$ *are defined as above.*
If $\hat{\theta}_k$ *and* $\hat{\theta}_k^*$ *are both consistent, then as* $n_k \to \infty$

$$\sqrt{n_k}(\hat{\theta}_k^* - \hat{\theta}_k) \xrightarrow{P_{\theta_{k_0}}} 0.$$

Proof: This theorem is proved in an analogous manner to the common source Lemma 7.2.
By Theorem 2.13 for $\hat{\theta}_k$, then as $n_k \to \infty$:

$$\sqrt{n_k} \, W_{\theta_{k_0}} \left(\hat{\theta}_k - \theta_{k_0} \right) \equiv -G_{n_k} \Psi_{\theta_{k_0}} + o_{P_{\theta_{k_0}}}(1).$$

By Theorem 2.13 for $\hat{\theta}_k^*$ and Slutsky's Theorem, then as $n_k \to \infty$:

$$\sqrt{n_k} \, W^*_{\theta_{k_0}} \left(\hat{\theta}_k^* - \theta_{k_0} \right) = -G_{n_k} \Psi_{\theta_{k_0}} + o_{P_{\theta_{k_0}}}(1).$$

Substituting $\sqrt{n_k} \, W_{\theta_{k_0}} \left(\hat{\theta}_k - \theta_{k_0} \right) = -G_{n_k} \Psi_{\theta_{k_0}} + o_{P_{\theta_{k_0}}}(1)$ into the display above, as $n_k \to \infty$

$$\sqrt{n_k} \left[W^*_{\theta_{k_0}}(\hat{\theta}_k^* - \theta_{k_0}) - W_{\theta_{k_0}}(\hat{\theta}_k - \theta_{k_0}) \right] = o_{P_{\theta_{k_0}}}(1)$$

$$W_{\theta_{k_0}}(\hat{\theta}_k^* - \hat{\theta}_k) + \frac{1}{n_k} \nabla \nabla^T \ln[f(e_u|\theta_k)](\hat{\theta}_k^* - \theta_{k_0}) = o_{P_{\theta_{k_0}}}\left(n_k^{-\frac{1}{2}}\right)$$

$$W_{\theta_{k_0}}(\hat{\theta}_k^* - \hat{\theta}_k) + O(n_k^{-1}) o_{P_{\theta_{k_0}}}(1) = o_{P_{\theta_{k_0}}}\left(n_k^{-\frac{1}{2}}\right)$$

$$W_{\theta_{k_0}}(\hat{\theta}_k^* - \hat{\theta}_k) = o_{P_{\theta_{k_0}}}\left(n_k^{-\frac{1}{2}}\right) - o_{P_{\theta_{k_0}}}\left(n_k^{-1}\right)$$

$$\sqrt{n_k} \, W_{\theta_{k_0}}(\hat{\theta}_k^* - \hat{\theta}_k) = o_{P_{\theta_{k_0}}}(1).$$

Finally, because matrix inverses are continuous functions, then by the Continuous Mapping Theorem and Slutsky's Theorem

$$\sqrt{n_k}(\hat{\theta}_k^* - \hat{\theta}_k) = W_{\theta_{k_0}}^{-1} o_{P_{\theta_{k_0}}}(1)$$

$$\sqrt{n_k}(\hat{\theta}_k^* - \hat{\theta}_k) = o_{P_{\theta_{k_0}}}(1), \quad \text{as } n_k \to \infty.$$

■

Theorem 7.6:

Let the assumptions of Lemma 7.4 and Lemma 7.5 be satisfied. Then as $n_s \to \infty$ and $n_a \to \infty$ at the same rate,

$$NP_{ss}(e_n) \xrightarrow{P_{\theta_0}} LR_{ss}(\theta_0; e_u).$$

Proof: Consider the Neyman-Pearson Likelihood Ratio for the specific source problem

$$
\begin{aligned}
NP_{ss}(e_n) &= \frac{f(e_s|\hat{\theta}_s^*)f(e_u|\hat{\theta}_s^*)f(e_a|\hat{\theta}_a)}{f(e_s|\hat{\theta}_s)f(e_u|\hat{\theta}_a^*)f(e_a|\hat{\theta}_a^*)} \\
&= \frac{f(e_u|\hat{\theta}_s^*)}{f(e_u|\hat{\theta}_a^*)} \times \frac{f(e_s|\hat{\theta}_s^*)}{f(e_s|\hat{\theta}_s)} \times \frac{f(e_a|\hat{\theta}_a)}{f(e_a|\hat{\theta}_a^*)} \\
&\equiv U \times S \times A
\end{aligned}
$$

Our goal is to show that $U \xrightarrow{P_{\theta_0}} LR_{ss}(\theta_0; e_u)$ as $n_s \to \infty$ and $n_a \to \infty$ at the same rate, $S \xrightarrow{P_{\theta_{s_0}}} 1$ as $n_s \to \infty$, and $A \xrightarrow{P_{\theta_{a_0}}} 1$ as $n_a \to \infty$.

The proof that $A \xrightarrow{P_{\theta_{a_0}}} 1$ is analogous to that for Theorem 7.3. First, consider

$$
\begin{aligned}
\ln(A) &= \sum_{i=1}^{n_a} \ln[f(Y_{ai}|\hat{\theta}_a)] - \sum_{i=1}^{n_a} \ln[f(Y_{ai}|\hat{\theta}_a^*)] \\
&\equiv \ell_A(\hat{\theta}_a) - \ell_A(\hat{\theta}_a^*).
\end{aligned}
$$

Using the second-order Multivariate Taylor's Expansion (Wade [77] p. 421) with the

Mean Value Theorem, then

$$\ln(A) = -\frac{1}{2}(\hat{\theta}_a^* - \hat{\theta}_a)^T \nabla\nabla^T \ell_A(\tilde{\theta}_a)(\hat{\theta}_a^* - \hat{\theta}_a)$$

where $\tilde{\theta}_a$ be on the line between $\hat{\theta}_a$ and $\hat{\theta}_a^*$. Next, consider

$$\nabla\nabla^T \ell_A(\tilde{\theta}_a) = \nabla\nabla^T \left[\sum_{i=1}^{n_a} \ln[f(Y_{ai}|\tilde{\theta}_a)]\right] = \sum_{i=1}^{n_a} \nabla\nabla^T \ln[f(Y_{ai}|\tilde{\theta}_a)].$$

Because $\hat{\theta}_a^*$ and $\hat{\theta}_a$ are consistent by Lemma 7.4 and by the Strong Law of Large Numbers, then as $n_a \to \infty$

$$\frac{1}{n_a} \nabla\nabla^T \ell_A(\tilde{\theta}_a) = O_{P_{\theta_{a0}}}(1).$$

Also, by Lemma 7.5, as $n_a \to \infty$

$$\sqrt{n_a}\left(\hat{\theta}_a^* - \hat{\theta}_a\right) = o_{P_{\theta_{a0}}}(1).$$

So, by Slutsky's Theorem, as $n_a \to \infty$

$$\begin{aligned}
\ln(A) &= -\frac{1}{2}(\hat{\theta}_a^* - \hat{\theta}_a)^T \nabla\nabla^T \ell_A(\tilde{\theta}_a)(\hat{\theta}_a^* - \hat{\theta}_a) \\
&= o_{P_{\theta_{a0}}}(1).
\end{aligned}$$

This implies that $\ln(A) \xrightarrow{P_{\theta_{a0}}} 0$ as $n_a \to \infty$, and by the Continuous Mapping Theorem

$$A \xrightarrow{P_{\theta_{a0}}} 1.$$

Using a similar proof as above, then as $n_s \to \infty$

$$S = \frac{f(e_s|\hat{\theta}_s^*)}{f(e_s|\hat{\theta}_s)} = \frac{\prod_{i=1}^{n_s} f(\mathbf{y}_{s_i}|\hat{\theta}_s^*)}{\prod_{i=1}^{n_s} f(\mathbf{y}_{s_i}|\hat{\theta}_s)} \xrightarrow{P_{\theta_{s0}}} 1.$$

To complete the proof, we need to show that, as $n_a \to \infty$ and as $n_s \to \infty$ at the same rate,

$$U = \frac{f(e_u|\hat{\theta}_s^*)}{f(e_u|\hat{\theta}_a^*)} \xrightarrow{P_{\theta_0}} LR_{ss}(\theta_0; e_u) = \frac{f(e_u|\theta_{s_0})}{f(e_u|\theta_{a_0})}.$$

By Lemma 7.4, we have that $\hat{\theta}_s^* \xrightarrow{P_{\theta_{s_0}}} \theta_{s_0}$ as $n_s \to \infty$ and $\hat{\theta}_a^* \xrightarrow{P_{\theta_{a_0}}} \theta_{a_0}$ as $n_a \to \infty$. By the Continuous Mapping Theorem

$$f(e_u|\hat{\theta}_s^*) \xrightarrow{P_{\theta_{s_0}}} f(e_u|\theta_{s_0})$$

and

$$\frac{1}{f(e_u|\hat{\theta}_a^*)} \xrightarrow{P_{\theta_{a_0}}} \frac{1}{f(e_u|\theta_{a_0})}.$$

Finally, by Slutsky's Theorem

$$NP_{ss}(e_n) \xrightarrow{P_{\theta_0}} LR_{ss}(\theta_0; e_u),$$

as $n_s \to \infty$ and $n_a \to \infty$ at the same rate.

$\blacksquare$

7.3   Application Example

The dataset used for this example is the exact same glass data described in Section 5.3. Four different Neyman-Pearson approximations are computed in the section, two for the specific source identification problem and two for the common source identification problem. For each identification of source problem, the Neyman-Pearson approximations are computed for two different scenarios; one scenario in which the evidence is chosen such that the prosecution hypothesis is true, and the second scenario in which

Table 7.1: Computed Neyman-Pearson approximations for the glass example.

(a) Observed Data

	Scenario	
	H_p	H_d
Specific Source	3.7117×10^{-23}	7.0100×10^{-2}
Common Source	68.45877	4.4915×10^{-11}

(b) Simulated Data

	Scenario	
	H_p	H_d
Specific Source	$1,355,357$	8.4163×10^{-9}
Common Source	340.3552	5.9925×10^{-37}

Computed likelihood ratio values for the glass simulated example.

	Scenario	
	H_p	H_d
Specific Source	2449715	1.708589×10^{-50}
Common Source	482.056	5.31445×10^{-48}

the evidence is chosen such that the defense hypothesis is true. Since the data is modeled using a multivariate normal distribution, see Section 4.3 for details of the models, the maximum likelihood estimates needed to compute the Neyman-Pearson approximation can be computed using the 'lme' function from the 'nlme' package in R [57]. The results of the example are provided in Table 7.1. For comparison, Table 5.5 is reproduced which shows the value of the true likelihood ratio for the simulated glass data under each of the scenarios and identification of source problems.

For the small number of samples in the specific source evidence (only three fragments under the H_p scenario and five fragments under the H_d scenario), the specific source Neyman-Pearson approximation results in misleading evidence (evidence pointing to the incorrect hypothesis) under the H_p scenario. This result is an example of Lindley's paradox in which two methods for evaluating the exact same evidence lead to two opposing conclusions. This indicates that when n_s, the number of samples in the specific source evidence e_s, is small (smaller than 10 samples) the Neyman-Pearson approximation provides an unstable approximation of the value of evidence. However, as indicated by the results presented in Section 7.1 and Section 7.2, this issue is

resolved when the sample sizes are larger. This is indicated by the results for the simulated glass data, since the samples sizes are larger than the observed glass data. It can be seen that the values of the Neyman-Pearson approximation for the simulated data are very close to the true value of the likelihood ratio, and do not result in Lindley's paradox. See the following Section 7.4 for further details.

7.4   Discussion

One of the most important consequences of the Neyman-Pearson approximation is that it can be used to replace the Bayes Factor in Equation 3.1 to update the prior odds and arrive at the approximate posterior odds. This result is formalized in the following corollary.

Corollary 7.6.1:
Suppose that the likelihood ratio function is bounded in P_{θ_0}-probability in a neighborhood of θ_0. The Neyman-Pearson approximation is an approximate value of evidence for the forensic identification of source problems:

$$\frac{\mathbb{P}\left(H_p|e, I\right)}{\mathbb{P}\left(H_d|e, I\right)} = \left[V_{NP}(e_n) \times \frac{\mathbb{P}\left(H_p|I\right)}{\mathbb{P}\left(H_d|I\right)}\right] + o_{P_{\theta_0}}(1).$$

Proof: For the first case, consider either $\mathbb{P}\left(H_p|I\right) \in \{0, 1\}$. Therefore, the prior odds are either zero or infinite, and so the result is trivial.

For the second case, consider Equation 3.1 for $\mathbb{P}\left(H_p|I\right) \in (0, 1)$ and $\mathbb{P}\left(H_d|I\right) \in (0, 1)$

$$
\begin{aligned}
\frac{\mathbb{P}\left(H_p|e, I\right)}{\mathbb{P}\left(H_d|e, I\right)} &= V_{BF}(e_n) \times \frac{\mathbb{P}\left(H_p|I\right)}{\mathbb{P}\left(H_d|I\right)} \\
&= \left[(V_{BF}(e_n) - V_{NP}(e_n)) + V_{NP}(e_n)\right] \times \frac{\mathbb{P}\left(H_p|I\right)}{\mathbb{P}\left(H_d|I\right)} \\
&= \left[(V_{BF}(e_n) - V_{NP}(e_n)) \times \frac{\mathbb{P}\left(H_p|I\right)}{\mathbb{P}\left(H_d|I\right)}\right] + \left[V_{NP}(e_n) \times \frac{\mathbb{P}\left(H_p|I\right)}{\mathbb{P}\left(H_d|I\right)}\right]
\end{aligned}
$$

Now, from Theorem 7.3 and Theorem 7.6, then $V_{NP}(e_n) \xrightarrow{P_{\theta_0}} V_{LR}(\theta_0; e_u)$ as $n \to \infty$. If it can be shown that $V_{BF}(e_n) \xrightarrow{P_{\theta_0}} V_{LR}(\theta_0; e_u)$ as $n \to \infty$, then the proof is complete. Recall from Equation 5.16 that the Bayes Factor is given by

$$V_{BF}(e_n) = \int V_{LR}(\theta; e_u) \, d\Pi(\theta|e_n, M_d),$$

where $||\Pi(\theta|e_n, M_d) - \Phi(\theta|\hat{\theta}_d, \frac{1}{n}\mathcal{I}_d(\hat{\theta}_d)^{-1})||_{TV} \xrightarrow{P_{\theta_0}} 0$ as $n \to \infty$ by the Bernstein-von Mises Theorem (Theorem 2.11) and where $\hat{\theta}_d$ is the MLE for θ and $\mathcal{I}_d(\hat{\theta}_d)^{-1}$ is the inverse of the observed Fisher's information matrix corresponding to the entire set of evidence under the defense model. Under the assumptions of Theorem 7.3 and Theorem 7.6, then $\hat{\theta}_d$ is consistent towards θ_0. Additionally, $\frac{1}{n}\mathcal{I}_d(\hat{\theta}_d)^{-1} \xrightarrow{P_{\theta_0}} 0$ as $n \to \infty$ provided that $\mathcal{I}_d(\hat{\theta}_d)^{-1}$ is bounded in P_{θ_0}-probability in a neighborhood of θ_0. Therefore, this implies that $||\Phi(\theta|\hat{\theta}_d, \frac{1}{n}\mathcal{I}_d(\hat{\theta}_d)^{-1}) - \delta_{\theta_0}(\theta)||_{TV} \xrightarrow{P_{\theta_0}} 0$ as $n \to \infty$ where $\delta_{\theta_0}(\theta)$ is the measure degenerate at θ_0. Now, consider

$$
\begin{aligned}
&\left|\left|\Pi(\theta|e_n, M_d) - \delta_{\theta_0}(\theta)\right|\right|_{TV} \\
&= \left|\left|\Pi(\theta|e_n, M_d) - \Phi(\theta|\hat{\theta}_d, n^{-1}\mathcal{I}_d(\hat{\theta}_d)^{-1}) + \Phi(\theta|\hat{\theta}_d, n^{-1}\mathcal{I}_d(\hat{\theta}_d)^{-1}) - \delta_{\theta_0}(\theta)\right|\right|_{TV} \\
&\leq \left|\left|\Pi(\theta|e_n, M_d) - \Phi(\theta|\hat{\theta}_d, n^{-1}\mathcal{I}_d(\hat{\theta}_d)^{-1})\right|\right|_{TV} + \left|\left|\Phi(\theta|\hat{\theta}_d, n^{-1}\mathcal{I}_d(\hat{\theta}_d)^{-1}) - \delta_{\theta_0}(\theta)\right|\right|_{TV} \\
&= o_{P_{\theta_0}}(1) + o_{P_{\theta_0}}(1)
\end{aligned}
$$

Therefore, $\left|\left|\Pi(\theta|e_n, M_d) - \delta_{\theta_0}(\theta)\right|\right|_{TV} \xrightarrow{P_{\theta_0}} 0$ as $n \to \infty$.

Now, consider the difference between the Bayes Factor and the Likelihood Ratio

$$
\begin{aligned}
&\left|V_{BF}(e_n) - V_{LR}(\theta_0; e_u)\right| \\
&= \left|\int V_{LR}(\theta; e_u) \, d\Pi(\theta|e_n, M_d) - \int V_{LR}(\theta; e_u) \, d\delta_{\theta_0}(\theta)\right| \\
&= \left|\int V_{LR}(\theta; e_u) \, d\left[\Pi(\theta|e_n, M_d) - \delta_{\theta_0}(\theta)\right]\right|
\end{aligned}
$$

Similar to the proofs of Theorem 6.1 and Theorem 6.2, by properties of signed measures and the total variation norm,

$$\left| \int V_{LR}(\theta; e_u) \, d\left[\Pi(\theta|e_n, M_d) - \delta_{\theta_0}(\theta)\right] \right| \leq \int V_{LR}(\theta; e_u) \, d\left| \left[\Pi(\theta|e_n, M_d) - \delta_{\theta_0}(\theta)\right] \right|.$$

By the assumption that $V_{LR}(\theta; e_u)$ is bounded in a neighborhood of θ_0, then suppose that $V_{LR}(\theta; e_u) \leq C$ for some real number $C > 0$. This implies that, when θ is in a neighborhood of θ_0,

$$
\begin{aligned}
\left| \int V_{LR}(\theta; e_u) \, d\left[\Pi(\theta|e_n, M_d) - \delta_{\theta_0}(\theta)\right] \right| &\leq \int V_{LR}(\theta; e_u) \, d\left| \left[\Pi(\theta|e_n, M_d) - \delta_{\theta_0}(\theta)\right] \right| \\
&\leq \int C \, d\left| \left[\Pi(\theta|e_n, M_d) - \delta_{\theta_0}(\theta)\right] \right| \\
&\leq C \int d\left| \left[\Pi(\theta|e_n, M_d) - \delta_{\theta_0}(\theta)\right] \right| \\
&\leq \left\| \Pi(\theta|e_n, M_d) - \delta_{\theta_0}(\theta) \right\|_{TV}.
\end{aligned}
$$

By the law of total probability and the previous result, as $n \to \infty$

$$\left| V_{BF}(e_n) - V_{LR}(\theta_0; e_u) \right| \xrightarrow{P_{\theta_0}} 0.$$

Finally, this implies (by Slutsky's Lemma) that as $n \to \infty$

$$
\begin{aligned}
&(V_{BF}(e_n) - V_{NP}(e_n)) \times \frac{\mathbb{P}(H_p|I)}{\mathbb{P}(H_d|I)} \\
=\ & (V_{BF}(e_n) - V_{LR}(\theta_0; e_u) + V_{LR}(\theta_0; e_u) - V_{NP}(e_n)) \times \frac{\mathbb{P}(H_p|I)}{\mathbb{P}(H_d|I)} \\
=\ & \left[(V_{BF}(e_n) - V_{LR}(\theta_0; e_u)) \times \frac{\mathbb{P}(H_p|I)}{\mathbb{P}(H_d|I)} \right] + \left[(V_{LR}(\theta_0; e_u) - V_{NP}(e_n)) \times \frac{\mathbb{P}(H_p|I)}{\mathbb{P}(H_d|I)} \right] \\
=\ & \left[o_{P_{\theta_0}}(1) \times \frac{\mathbb{P}(H_p|I)}{\mathbb{P}(H_d|I)} \right] + \left[o_{P_{\theta_0}}(1) \times \frac{\mathbb{P}(H_p|I)}{\mathbb{P}(H_d|I)} \right] \\
=\ & o_{P_{\theta_0}}(1).
\end{aligned}
$$

In conclusion,

$$
\begin{aligned}
\frac{\mathbb{P}\left(H_p|e,I\right)}{\mathbb{P}\left(H_d|e,I\right)} &= \left[(V_{BF}(e_n) - V_{NP}(e_n)) \times \frac{\mathbb{P}\left(H_p|I\right)}{\mathbb{P}\left(H_d|I\right)}\right] + \left[V_{NP}(e_n) \times \frac{\mathbb{P}\left(H_p|I\right)}{\mathbb{P}\left(H_d|I\right)}\right] \\
&= o_{P_{\theta_0}}(1) + \left[V_{NP}(e_n) \times \frac{\mathbb{P}\left(H_p|I\right)}{\mathbb{P}\left(H_d|I\right)}\right].
\end{aligned}
$$

∎

The Neyman-Pearson approximation only requires the computation of the maximum likelihood estimates for the parameters, in contrast to the Bernstein-von Mises or Laplace's approximations which require both the parameter estimates as well as computation of Fisher's information matrices. Also, the Neyman-Pearson requires no Monte Carlo integration techniques since there are no integrals to compute, as in the Bernstein-von Mises approximation. However, the Neyman-Pearson approximation will require an optimization routine to find the maximum likelihood estimates for many classes of parametric distributions. Finally, the Neyman-Pearson approximation does not have any dependence on a choice of prior distribution, like the Bayes Factor (and potentially the Laplace's approximation).

7.5 Simulation Study

To further investigate the relationship between the Neyman-Pearson approximation and the Bayes Factor, a simulation study was designed with the goal of showing that the Neyman-Pearson approximation is similar to a Bayes Factor computed using a non-informative prior. Theoretically, a Bayes Factor computed using a (improper) non-informative prior is not well-defined [10]. Computationally, several issues arise when attempting to compute a Bayes Factor using a non-informative prior [10, 59]. However, the Neyman-Pearson approximation is well-defined for the forensic identifi-

cation of source problems and computationally viable when samples sizes for e_s and e_a are not too small.

7.5.1   Details of Simulation Methodology

For this simulation study, the first group of the glass data was used to suggest the true parameter values for the generation of evidence sets in the manner described in Section 5.3.2. Once the evidence has been generated for the simulation study, it is assumed that the evidence follows the multivariate normal distribution originally formulated in Section 4.3. These distributions will be used to compute both the Neyman-Pearson approximations and the Bayes Factors. In order to complete the full Bayesian model needed to compute the Bayes Factors, the specified true parameter values are used as the values of the hyperparameters for the prior distributions.

$$\mu_s \sim \mathcal{N}_3(\mu_{s_0}, K\Sigma_{s_0}) \quad \text{and} \quad \Sigma_s \sim \mathcal{W}_3^{-1}(\Sigma_{s_0}, \nu_s) \tag{7.4a}$$

$$\mu_a \sim \mathcal{N}_3(\mu_{a_0}, K\Sigma_{b_0}), \quad \Sigma_a \sim \mathcal{W}_3^{-1}(\Sigma_{b_0}, \nu_b), \quad \text{and} \quad \Sigma_w \sim \mathcal{W}_3^{-1}(\Sigma_{e_0}, \nu_e) \tag{7.4b}$$

Under the H_p scenario $\mu_{s_0} = \mu_{s_0}^{(10)}$ and $\Sigma_{s_0} = \Sigma_{s_0}^{(10)}$, while under the H_d scenario $\mu_{s_0} = \mu_{s_0}^{(48)}$ and $\Sigma_{s_0} = \Sigma_{s_0}^{(48)}$. In addition, the Bayes Factors for the simulation study were computed using varying degrees of certainty in the prior distribution, as designated by the scaling variable, K, for the covariance hyperparameters of the multivariate normal prior for the mean parameters and by the degrees of freedom hyperparameters, ν, of the inverse Wishart priors for the covariance parameters. See the following Algorithm 7 for further details of the simulation study.

Algorithm 7: Neyman-Pearson/Bayes Factor Simulation Study

Input: Compute the following true parameters values (as described above):
   $\mu_{s_0}^{(10)}, \Sigma_{s_0}^{(10)}, \mu_{s_0}^{(48)}, \Sigma_{s_0}^{(48)}, \mu_{a_0}, \Sigma_{b_0}, \Sigma_{e_0}$;

for *each identification of source problem $ID \in \{CS,\ SS\}$* **do**

    for *each scenario $S \in \{H_p,\ H_d\}$* **do**

        Generate unknown source evidence according to S for ID problem;

        Compute the Likelihood Ratio using Equation 3.6 or Equation 3.10;

        for $i = 1, 2, \ldots, N$ **do**

            Generate control evidence according to S for ID problem;

            Compute the NP approximation using Equation 7.2 or Equation 7.3;

            for $d \in \{3^1, 3^2, 3^4, 3^6, 3^8, 3^{10}, 3^{12}, 3^{14}, 3^{16}\}$ **do**

                Compute $\widehat{BF}(e)$ using Equation 5.12 or Equation 5.14

                    where the prior distributions are given in Equation 7.4

                    with $\nu_s = \nu_b = \nu_e = d$ and $K = 100/d$;

            end

            Store all Bayes Factors and NP approximations for comparison;

        end

   Store all Likelihood Ratios for reference;

    end

end

7.5.2   Simulation Results

The results of the simulation study are shown in the following plots. For the scenarios considered in this small simulation study, it can be seen in Figures 7.1-7.4 that as the degrees of freedom used for the prior distributions to compute the Bayes Factor increase, the Bayes Factors converge to the value of the likelihood ratio, given by the red line. These figures also show that as the degrees of freedom decrease towards zero, the Bayes Factor approaches the value of the Neyman-Pearson approximation. These results support the conjecture that the Neyman-Pearson approximation behaves like a Bayes Factor computed using uninformative prior distributions.

For the scenarios explored in the simulation study, the Neyman-Pearson approximation tends to be larger (often much larger) than the value of the likelihood ratio. This relationship may not hold for all cases, as is indicated by the common source problem

Figure 7.1: The simulation results for the specific source problem under the H_p scenario represented by the log-10 Neyman-Pearson approximation (NP) and the various log-10 Bayes Factors computed using differing degrees of certainty in the prior distribution.
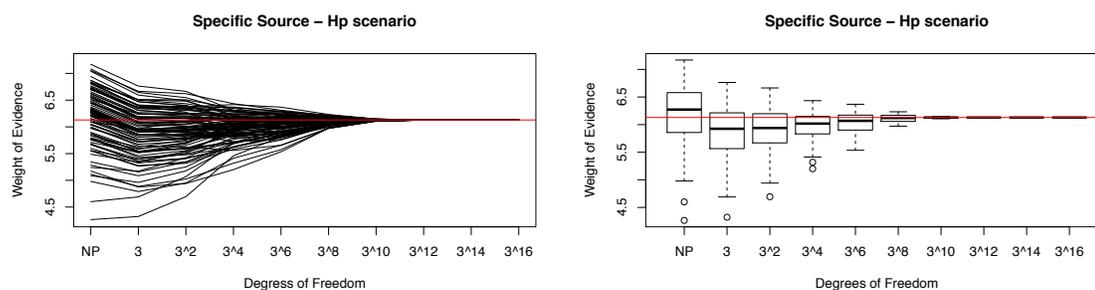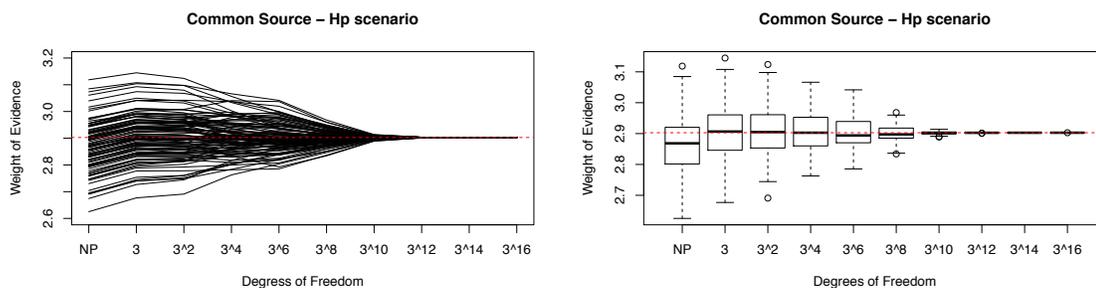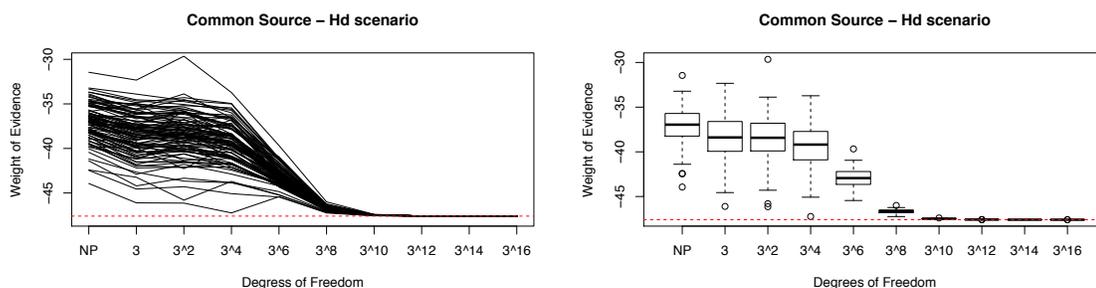

(a) Each solid black line represents the resulting values for the weight of evidence computed using the same evidence in a single iteration of the simulation. The true log-10 likelihood ratio is represented by the solid red line.

(b) Boxplots corresponding to the simulation results for Figure(a). Again, the true log-10 likelihood ratio is represented by the solid red line.

Figure 7.2: The simulation results for the specific source problem under the H_d scenario represented by the log-10 Neyman-Pearson approximation (NP) and the various log-10 Bayes Factors computed using differing degrees of certainty in the prior distribution.
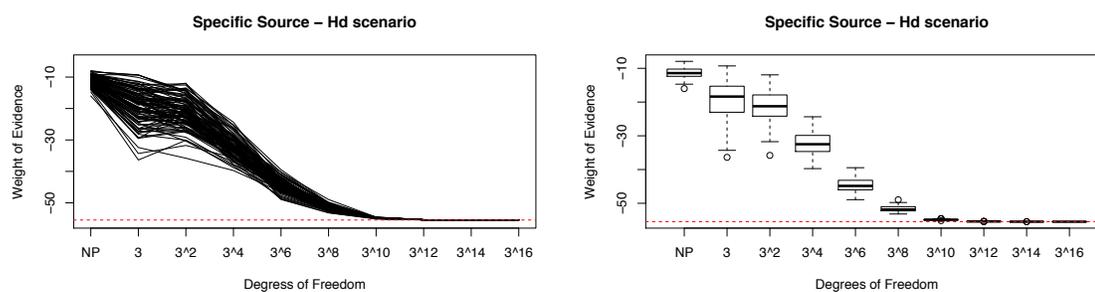
(a) Each solid black line represents the resulting values for the weight of evidence computed using the same evidence in a single iteration of the simulation. The true log-10 likelihood ratio is represented by the solid red line.

(b) Boxplots corresponding to the simulation results for Figure(a). Again, the true log-10 likelihood ratio is represented by the solid red line.

Figure 7.3: The simulation results for the common source problem under the H_p scenario represented by the log-10 Neyman-Pearson approximation (NP) and the various log-10 Bayes Factors computed using differing degrees of certainty in the prior distribution.

(a) Each solid black line represents the resulting values for the weight of evidence computed using the same evidence in a single iteration of the simulation. The true log-10 likelihood ratio is represented by the solid red line.

(b) Boxplots corresponding to the simulation results for Figure(a). Again, the true log-10 likelihood ratio is represented by the solid red line.

Figure 7.4: The simulation results for the common source problem under the H_d scenario represented by the log-10 Neyman-Pearson approximation (NP) and the various log-10 Bayes Factors computed using differing degrees of certainty in the prior distribution.

(a) Each solid black line represents the resulting values for the weight of evidence computed using the same evidence in a single iteration of the simulation. The true log-10 likelihood ratio is represented by the solid red line.

(b) Boxplots corresponding to the simulation results for Figure(a). Again, the true log-10 likelihood ratio is represented by the solid red line.

under the H_p scenario, in which the Neyman-Pearson tended to be smaller than the values of the likelihood ratio. Since the Neyman-Pearson approximation is behaving similarly to a Bayes Factor computed using an uninformative prior, the direction of the relationship between the Neyman-Pearson approximation and the likelihood ratio is difficult (possibly impossible) to predict in practical applications.

7.5.3   Simulation Conclusions

The results of this simulation study provide support for the conjecture that the Neyman-Pearson approximation behaves like a Bayes Factor computed using uninformative prior distributions. This method of computing the Bayes Factor is referred to as a generalized Bayes Rule [10]. In Berger [10], it is shown that decisions based on generalized Bayes Rules are often statistically admissible with respect to a particular loss function. Based on these results, it is believed that decisions based on the Neyman-Pearson approximation, with respect to the "0-1" loss function, will be statistically admissible rules for the forensic identification of source problems.

CHAPTER 8

Interval Quantifications

Disclaimer: This chapter is based in part on Ommen et al. [54].

Many legal and scientific scholars in the various forensic science disciplines agree that the value of evidence should be reported as a likelihood ratio or a Bayes Factor [24]. Recent analytical developments paired with modern statistical computational tools have led to the proliferation of ad-hoc techniques for quantifying these probative values of forensic evidence. Therefore, quantifications for the value of evidence are subjected to many sources of variability and uncertainty. There is currently a debate on how to characterize the reliability of the value of evidence [9, 13, 18, 50, 54, 71, 72, 74, 68]. Some authors have proposed associating a confidence interval or credible interval with the value of evidence assigned to a collection of forensic evidence. In this chapter, a method of providing reasonable credible intervals for the likelihood ratio is introduced.

8.1   Introduction

First, as discussed by Taroni et al. [71], the Bayes Factor already incorporates the uncertainty associated with the unknown parameters into the final assessment of the evidence. Therefore, it is clearly redundant to include an interval estimate for the Bayes Factor and intervals should not be used to characterize this type of uncertainty associated with the Bayes Factor. Hence, only interval quantifications for the

likelihood ratio will be considered in this chapter.

Some researchers have suggested using various credible or confidence intervals for the likelihood ratio [6, 8, 19]. Some examples of credible intervals include the highest posterior density, equal tails, and approximate normal posterior interval constructions presented in Buckleton and Curran [14, 19]. It is unclear whether the entire interval, although informative, can be used in the Bayesian paradigm to make a logical and coherent decision. Although no stance regarding interval quantifications for the value of evidence is taken in this dissertation, a potential solution for those who wish to quantify uncertainty regarding a likelihood ratio with a credible interval in an appropriate manner is presented. The two non-standard forms for the Bayes Factor provided in Section 5 facilitate the Bayesian asymptotic theory leading to the derivation of three approximately equal credible intervals for the likelihood ratio.

In Ommen et al. [54], one of the arguments against presenting intervals for the likelihood ratio is that the resulting quantification tends to misrepresent the value of evidence. For this result, Jensen's inequality is used to show that the posterior mean of the likelihood ratio (given e_s and e_a) is at least as large as the Bayes Factor. The result is reproduced for clarity below.

$$
\begin{aligned}
\int LR_{ss}(\theta; e_u) d\Pi(\theta | e_s, e_a) &= \int \int \frac{f(e_u | \theta_s)}{f(e_u | \theta_a)} d\Pi(\theta_a | e_a) d\Pi(\theta_s | e_s) \\
&= \int f(e_u | \theta_s) d\Pi(\theta_s | e_s) \int \frac{1}{f(e_u | \theta_a)} d\Pi(\theta_a | e_a) \\
&\geq \int f(e_u | \theta_s) d\Pi(\theta_s | e_s) \frac{1}{\int f(e_u | \theta_a) d\Pi(\theta_a | e_a)} \quad (8.1) \\
&= BF_{ss}(e)
\end{aligned}
$$

It should be noted that in order for this inequality to hold, there is a strong assumption that the prior distribution for θ_s is statistically independent of the prior distribution for θ_a. Notice that, in this case, the expectation of the likelihood ratio is taken

with respect to the joint posterior distribution of the parameters given e_s and e_a. In contrast, if the posterior distribution of the parameters given the entire set of evidence under the defense model is used when taking the expectation of the likelihood ratio, the result is Equation 5.6 for the specific source problem and Equation 5.1 for the common source problem. This means that there is a direct equality between the expected value of the likelihood ratio and the Bayes Factor. Therefore, the credible interval for the likelihood ratio based on this posterior distribution for the parameters given the entire set of evidence under the defense model will *not* misrepresent the value of evidence like the result above.

8.2 Credible Intervals for the Likelihood Ratio

As a culmination of the results presented thus far relating the Bayes Factor to the likelihood ratio, the Bayes Factor can be used as the center-point to construct an approximate $1 - \alpha$ credible intervals for the likelihood ratio. The Jensen's inequality result, given by Equation 8.1, provides motivation for a posterior distribution of the parameters which results in the posterior mean of the likelihood ratio being unbiased towards the Bayes Factor. This led to the derivation of the alternative form of the value of evidence given by Equations 5.1 and 5.6. Using the posterior distribution suggested by these Bayes Factors, a posterior distribution for the likelihood ratio is derived. The Bernstein-von Mises Theorem implies that a scaled version of this posterior distribution is shown to be asymptotically normal. This result is formalized in Theorem 8.1 below.

First, some notation will be defined to facilitate the result. Let λ denote the likelihood ratio function defined in Equation 3.3 for a fixed observation of the unknown source evidence. The notation here is generalized so that the results in this section will

apply generally to both the common source and specific source identification problem. Since the Bernstein-von Mises Theorem uses the rescaled parameter $\sqrt{n}(\theta - \theta_0)$, then consider the rescaled version of the likelihood ratio function

$$\Lambda_n(\theta) = \sqrt{n}[\lambda(\theta) - \lambda(\hat{\theta}_n^d)]$$

where $n = n_s = n_a$ and $\hat{\theta}_n^d$ is a consistent maximum likelihood estimate of θ under the defense model M_d ($\hat{\theta}_n^d \xrightarrow{P_{\theta_0}} \theta_0$ as $n \to \infty$). Next, let $\lambda'(\hat{\theta}_n^d)$ denote the vector of first partial derivatives of $\lambda(\theta)$ with respect to θ and $\mathcal{I}_{\hat{\theta}_n^d}^{-1}$ denote the inverse observed Fisher's information matrix evaluated at $\hat{\theta}_n^d$. For the following set of results, the metric $||\cdot||_{TV}$ denotes the total variation norm (see van der Vaart [75] or van de Geer [73]), and $N_\epsilon(x)$ denotes an ϵ-neighborhood of the point x. For two measures P and Q on $(\mathcal{X}, \mathcal{A})$, the total variation distance between P and Q is given by $||P - Q||_{TV} = \sup_{A \in \mathcal{A}} |P(A) - Q(A)|$. Finally, let e_n denote the entire set of evidence where the unknown source evidence always has a fixed number of samples (n_u for specific source, and n_{u_1} and n_{u_2} for the common source) and where the control samples have a varying size n (recall n_a is the number of sources in the alternative source population evidence and n_s is the number of samples in the specific source evidence).

Theorem 8.1 (Asymptotic Posterior Distribution for the LR)**:**
Suppose that $\theta \in N_\varepsilon(\theta_0)$ for any $\varepsilon > 0$ and let $\Pi(\Lambda_n(\theta)|e_n, M_d)$ denote the posterior distribution of $\Lambda_n(\theta)$ given the entire set of evidence e_n under M_d. If the assumptions of the Bernstein-von Mises Theorem hold, then as $n \to \infty$

$$\left|\left|\Pi(\Lambda_n(\theta)|e_n, M_d) - \Phi(\Lambda_n(\theta); 0, \gamma_n^d)\right|\right|_{TV} \xrightarrow{P_{\theta_0}^n} 0$$

where the convergence is with respect to the total variation norm and

$$\gamma_n^d = \lambda'(\hat{\theta}_n^d)^T \; \mathcal{I}_{\hat{\theta}_n^d}^{-1} \; \lambda'(\hat{\theta}_n^d).$$

Proof: Consider the following Taylor's expansion for either the specific source or commons source likelihood ratio function:

$$\lambda(\theta) = \lambda(\hat{\theta}_n^d) + (\theta - \hat{\theta}_n^d)^T \lambda'(\hat{\theta}_n^d) + \frac{1}{2}(\theta - \hat{\theta}_n^d)^T \lambda''(\tilde{\theta}_n^d)(\theta - \hat{\theta}_n^d)$$

where $\hat{\theta}_n^d$ is the maximum likelihood estimate for the parameter θ under the defense model for the evidence, $\lambda'(\theta)$ denotes the vector of first partial derivatives of the likelihood ratio function, $\lambda''(\theta)$ denotes the matrix of second partial derivatives of the likelihood ratio function, and $\tilde{\theta}_n^d$ is between θ and $\hat{\theta}_n^d$. Note that for this derivation, we will assume that $\hat{\theta}_n^d \xrightarrow{P_{\theta_0}} \theta_0$ as $n \to \infty$ where θ_0 is the true parameter value for θ (in the frequentist sense). Next, consider the corresponding Taylor's expansion of $\Lambda_n(\theta)$:

$$\Lambda_n(\theta) = \sqrt{n}\left[\lambda(\theta) - \lambda(\hat{\theta}_n^d)\right] = \sqrt{n}(\theta - \hat{\theta}_n^d)^T \lambda'(\hat{\theta}_n^d) + \frac{\sqrt{n}}{2}(\theta - \hat{\theta}_n^d)^T \lambda''(\tilde{\theta}_n^d)(\theta - \hat{\theta}_n^d)$$

By the Bernstein-von Mises Theorem, we have that as $n \to \infty$

$$\left\|\Pi(\sqrt{n}(\theta - \hat{\theta}_n^d)|e_n, M_d) - \Phi(\sqrt{n}(\theta - \hat{\theta}_n^d); 0, I_{\hat{\theta}_n^d}^{-1})\right\|_{TV} \xrightarrow{P_{\theta_0}^n} 0,$$

and therefore,

$$\sqrt{n}(\theta - \hat{\theta}_n^d) = O_{P_{\theta_0}^n}(1).$$

By the assumptions of the Bernstein-von Mises Theorem, then $\lambda''(\theta)$ is bounded in probability for any θ in a neighborhood of $\hat{\theta}_n^d$. This implies that as $n \to \infty$

$$\frac{\lambda''(\tilde{\theta}_n^d)}{\sqrt{n}} = o_{P_{\theta_0}^n}(1).$$

By Slutsky's Lemma, as $n \to \infty$

$$
\begin{aligned}
\frac{\sqrt{n}}{2}(\theta - \hat{\theta}_n^d)\lambda''(\tilde{\theta}_n^d)(\theta - \hat{\theta}_n^d)^T &= \frac{1}{2}\left[\sqrt{n}(\theta - \hat{\theta}_n^d)^T\right]\left[\frac{\lambda''(\tilde{\theta}_n^d)}{\sqrt{n}}\right]\left[\sqrt{n}(\theta - \hat{\theta}_n^d)\right] \\
&= O_{P_{\theta_0}^n}(1) \; o_{P_{\theta_0}^n}(1) \; O_{P_{\theta_0}^n}(1) \\
&= o_{P_{\theta_0}^n}(1).
\end{aligned}
$$

This means that the limiting distribution of $\Lambda_n(\theta)$ is determined by the limiting distribution of the scaled parameter.

$$
\Lambda_n(\theta) = \sqrt{n}(\theta - \hat{\theta}_n^d)^T \lambda'(\hat{\theta}_n^d) + o_{P_{\theta_0}^n}(1), \quad n \to \infty
$$

By Bernstein-von Mises Theorem and by Slutsky's Lemma, then as $n \to \infty$

$$
\left|\left|\Pi(\Lambda_n(\theta)|e_n, M_d) - \Phi(\Lambda_n(\theta); 0, \gamma_n^d)\right|\right|_{TV} \xrightarrow{P_{\theta_0}^n} 0
$$

where the convergence is with respect to the total variation norm and $\gamma_n^d = \lambda'(\hat{\theta}_n^d)^T \mathcal{I}_{\hat{\theta}_n^d}^{-1} \lambda'(\hat{\theta}_n^d)$. ∎

Because the posterior distribution for the likelihood ratio is asymptotically normal, we can construct an approximate $1 - \alpha$ credible region for the likelihood ratio based on this posterior distribution. This result is formalized in the theorem below.

Corollary 8.1.1 (Approximate $1 - \alpha$ Credible Interval for the LR)**:**
Let let the assumptions of the Bernstein-von Mises Theorem hold. Then for $0 < \alpha < 1$, as $n = n_s = n_a \to \infty$

$$
\Pi(\lambda(\theta) \in I_n | e_n, M_d) \to 1 - \alpha,
$$

where I_n is the interval such that $I_n = \lambda(\hat{\theta}_n^d) \pm \Phi^{-1}(1 - \frac{\alpha}{2}) \sqrt{\gamma_n^d/n}$ and

Φ^{-1} *represents the standard normal quantile function. Therefore, the existence of an approximate* $1 - \alpha$ *credible interval for the likelihood ratio is guaranteed.*

Proof: Consider

$$
\begin{aligned}
\Pi(\lambda(\theta) \in I_n | e_n, M_d) \;=\; & \Pi(\lambda(\theta) \in I_n) | e_n, M_d, \theta \in N_\varepsilon(\theta_0)) \; \Pi(\theta \in N_\varepsilon(\theta_0) | e_n, M_d) \\
+\; & \Pi(\lambda(\theta) \in I_n) | e_n, M_d, \theta \notin N_\varepsilon(\theta_0)) \; \Pi(\theta \notin N_\varepsilon(\theta_0) | e_n, M_d)
\end{aligned}
$$

By the equivalent unscaled result for Theorem 8.1 and the Extended Continuous Mapping Theorem [76], there exists an $\varepsilon_1 > 0$ such that, as $n \to \infty$

$$
\Pi(\lambda(\theta) \in I_n) | e_n, M_d, \theta \in N_{\varepsilon_1}(\theta_0)) \xrightarrow{P_{\theta_0}^n} 1 - \alpha.
$$

By the unscaled version of Bernstein-von Mises Theorem given in Equation 2.12, there exists $\varepsilon_2 > 0$ such that, as $n \to \infty$

$$
\Pi(\theta \in N_{\varepsilon_2}(\theta_0) | e_n, M_d) \xrightarrow{P_{\theta_0}^n} 1
$$

and

$$
\Pi(\theta \notin N_{\varepsilon_2}(\theta_0) | e_n, M_d) \xrightarrow{P_{\theta_0}^n} 0.
$$

Letting $\varepsilon = \min\{\varepsilon_1, \varepsilon_2\}$, then as $n \to \infty$ and by Slutsky's Lemma [75], we have that

$$
\Pi(\lambda(\theta) \in I_n) | e_n, M_d) \xrightarrow{P_{\theta_0}^n} 1 - \alpha.
$$

This proves the existence of a $1 - \alpha$ credible interval for the likelihood ratio that covers the true likelihood ratio which is based on the posterior distribution for the likelihood ratio given the entire set of evidence under the defense model. ∎

As a consequence of the result above, there are three types of common intervals that will be considered for the likelihood ratio, the approximate normal-based interval, the highest posterior density interval, and the equal tails interval [14, 19]. As a result of Theorems 5.1 and 5.2, these intervals for the likelihood ratio will cover the the Bayes Factor with $1 - \alpha$ coverage (for sufficiently large n), and cover the true likelihood ratio with approximate $1 - \alpha$ coverage. The following theorem formalizes this result.

Corollary 8.1.2 (Unbiased Credible Intervals for the LR)**:**

Let the assumptions of the Bernstein-von Mises Theorem hold. Then the following approximate $1 - \alpha$ credible intervals for the likelihood ratio, λ_θ, are all approximately equivalent:

Approximate Normal-Based: $V(e_n) \pm \Phi^{-1}(1 - \frac{\alpha}{2})\, \sigma_n$

> *where σ_n denotes the posterior standard deviation for the likelihood ratio and Φ^{-1} represents the standard normal quantile function;*

Equal Tails: $\left(\Pi_{\lambda|e_n}^{-1}(\frac{\alpha}{2}), \Pi_{\lambda|e_n}^{-1}(1 - \frac{\alpha}{2}) \right)$

> *where $\Pi_{\lambda|e_n}^{-1}$ represents the quantile function of the posterior distribution for the likelihood ratio, $\Pi(\lambda_\theta|e_n, M_d)$;*

Highest Posterior Density: $C_n = \{\lambda_\theta : \pi(\lambda_\theta|e_n, M_d) \geq c\}$

> *where c is chosen such that*

$$\int_{C_n} \pi(\lambda_\theta|e_n, M_d)\, d\lambda_\theta = 1 - \alpha$$

and $\pi(\lambda_\theta|e_n, M_d)$ is the density corresponding to $\Pi(\lambda_\theta|e_n, M_d)$.

Proof: From Theorem 5.1 and Theorem 5.2, we get that the Bayes Factor is the posterior mean of the likelihood ratio. From Theorem 8.1 we get that the posterior

distribution of the likelihood ratio will be approximately symmetric for sufficiently large n. From Theorem 8.1.1 we get that the approximate coverage of these intervals is $1 - \alpha$. Therefore, since the normal distribution is symmetric and unimodal, each of the three credible intervals will be approximately centered on the Bayes Factor with the exact same coverage probability of $1 - \alpha$. ∎

8.3   Simulation Study

A simulation example has been designed to investigate how the three different credible intervals for the likelihood ratio (approximate normal, equal tails, and highest posterior density) behave as the number of samples from the specific source population (n_s) and the number of sources in the alternative source population (n_a) increase. For this simulation, let N be a sufficiently large constant denoting the number of samples to increase n_s and n_a to. In Section 8.2, there is a constraint on the evidence sets that the initial number of samples from the specific source and the initial number of sources in the alternative source population are equal $(n_s = n_a)$. Therefore, this simulation study will use the simulated glass data for the evidence with $N = 100$ and with five unknown source fragments and 5 fragments within each source in e_a.

8.3.1   Details of Simulation Methodology

For a given sample size $(n_s = n_a = n = 1, 2, \ldots, N)$, the simulation will draw multiple samples of the likelihood ratio from its posterior distribution given the entire set of evidence generated under M_d using a method similar to the parametric bootstrap technique [23]. For both the common source and specific source problem, two different sets of prior distributions for the parameters are used, resulting in two different posterior distributions. The first set of prior distributions are centered on the true

values of the parameters used for simulating the glass data, refer to Section 5.3.2 for details. All of the degrees of freedom hyperparameters are set to three and the value of K is ten. For the second set of prior distributions, the priors suggested using groups 2 and 3 of the observed glass data given in Section 4.3.1 are used. These posterior likelihood ratio samples will be used to compute the three different credible intervals defined in Corollary 8.1.2. The algorithms for the simulating the posterior LR samples for both the specific source and common source problems are given in Algorithms 8 and 9 below. The initial evidence sets for the simulation are composed of the simulated glass data described in Section 5.3.2.

Algorithm 8: Specific Source Posterior Likelihood Ratio Simulation

Input: Evidence set $e^{(n)} = \{e_u, e_s^{(n)}, e_a^{(n)}\}$ with sample size n

for $m = 1, 2, ..., M$ **do**

  Draw θ_s^* from $\Pi(\theta_s | e_s^{(n)}, M_d)$
  Draw θ_a^* from $\Pi(\theta_a | e_u, e_a^{(n)}, M_d)$
  Compute $\lambda_{\theta,m}^{(n)} = f(e_u | \theta_s^*) / f(e_u | \theta_a^*)$

end

Output: $\{\lambda_{\theta,m}^{(n)} : m = 1, 2, \ldots, M\}$ a sample from $\Pi(\lambda_\theta | e^{(n)}, M_d)$

Algorithm 9: Common Source Posterior Likelihood Ratio Simulation

Input: Initial evidence set $e^{(n)} = \{e_{u_1}, e_{u_2}, e_a^{(n)}\}$ with sample size n

for $m = 1, 2, ..., M$ **do**

  Draw θ_a^* from $\Pi(\theta_a | e_{u_1}, e_{u_2}, e_a^{(n)}, M_d)$
  Compute $\lambda_{\theta,m}^{(n)} = f(e_{u_1}, e_{u_2} | \theta_a^*, M_p) / f(e_{u_1}, e_{u_2} | \theta_a^*, M_d)$

end

Output: $\{\lambda_{\theta,m}^{(n)} : m = 1, 2, \ldots, M\}$ a sample from $\Pi(\lambda_\theta | e^{(n)}, M_d)$

Once the posterior LR samples have been simulated, the credible intervals for the likelihood ratio are computed from these samples. As a result of Equations 5.1 and 5.6, the Bayes Factors, which serve as the centers of the approximate normal-based credible intervals, are computed using the sample mean of the posterior LR sample for a given sample size n. And, as a result of Theorem 8.1.1, the posterior standard deviations for the approximate normal-based credible interval are computed using

the sample standard deviation of the posterior LR samples for a given sample size n. Next, the equal tails credible intervals are computed using the sample quantile function for the posterior LR samples. Finally, the highest posterior density credible intervals are computed using the implementation in the 'HDInterval' package in R [48].

8.3.2   Simulation Results

The results of the simulation are provided in the Figures 8.1-8.8. Figure 8.1 shows the behavior of the credible intervals for the specific source likelihood ratio under the H_p scenario as the sample size increases for both the number of samples in the specific source evidence and the number of sources in the alternative source population evidence. As Figure 8.1 shows, under the first prior choice, the credible intervals for the likelihood ratio contain the true value of the likelihood ratio for the simulated data, plotted with the red dotted line. However, this is not the case under the second prior choice (which was shown to be mismatched to the specific source data in Section 4.3.4). In Figure 8.1(b), due to the mismatched prior choice, the credible intervals for the likelihood ratio point in the incorrect direction (all the interval centers are less than one instead of greater than one). Also, as the sample size increases, the widths of the intervals tends to decrease for the first choice of prior distributions. In contrast, as the sample sizes increases using the second choice of prior distributions, the interval widths tend to increase. This result is shown in Figure 8.3.

Figure 8.2 shows the behavior of the credible intervals for the specific source likelihood ratio under the H_d scenario as the sample size increases. As Figure 8.2 shows, under the first prior choice, the centers of the credible intervals for the likelihood ratio are very close to the value of the likelihood ratio for the simulated data, plotted with the red dotted line. Again, this is not the case under the second prior choice. It should be

Figure 8.1: Credible intervals for the specific source likelihood ratio under the H_p scenario. The center of the intervals is plotted with a ∘ and the endpoints of the intervals are plotted with a −. The true value of the likelihood ratio is plotted with a red dotted line.
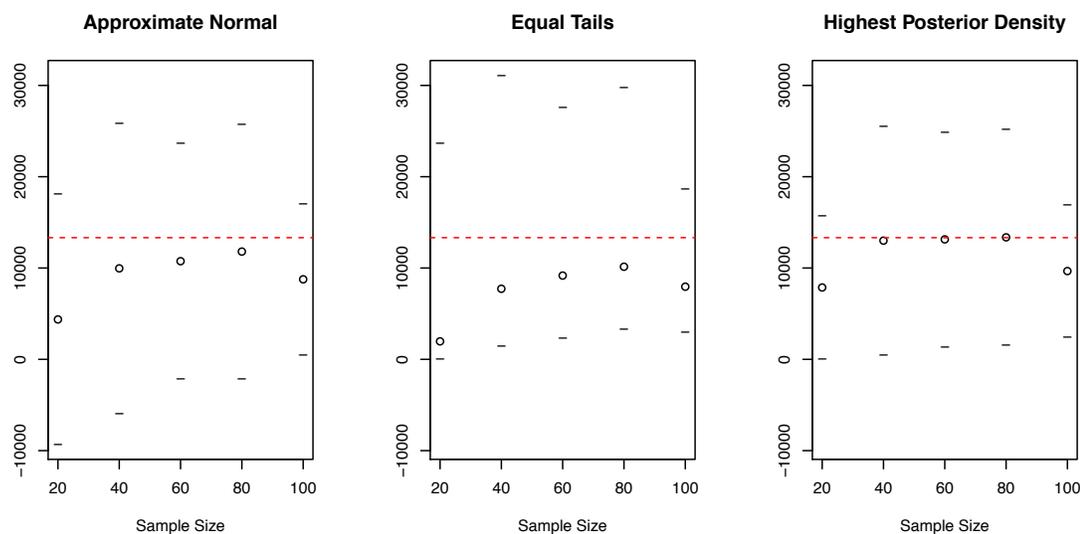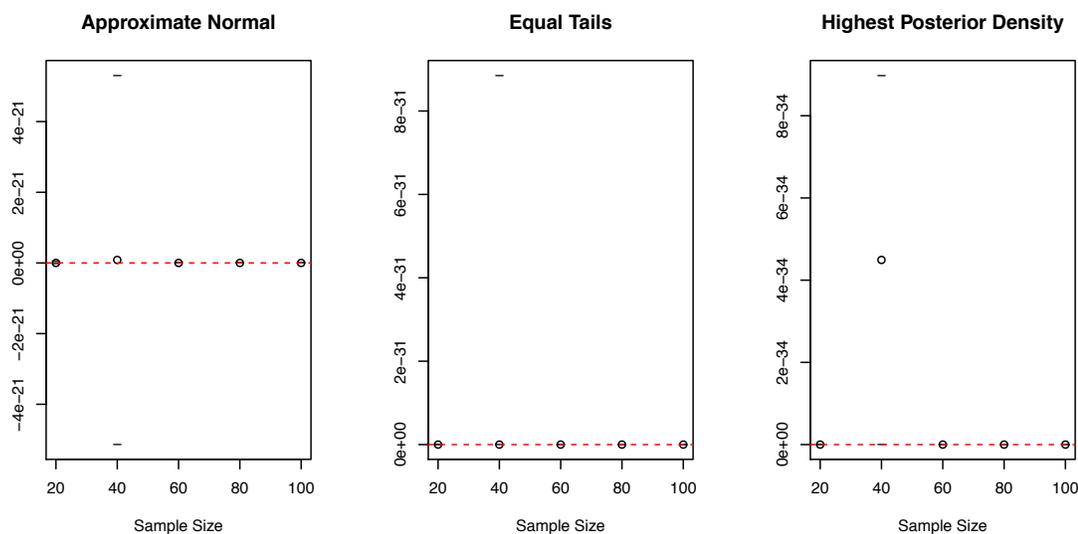

(a) Credible intervals constructed using prior distributions centered on the true parameter values.
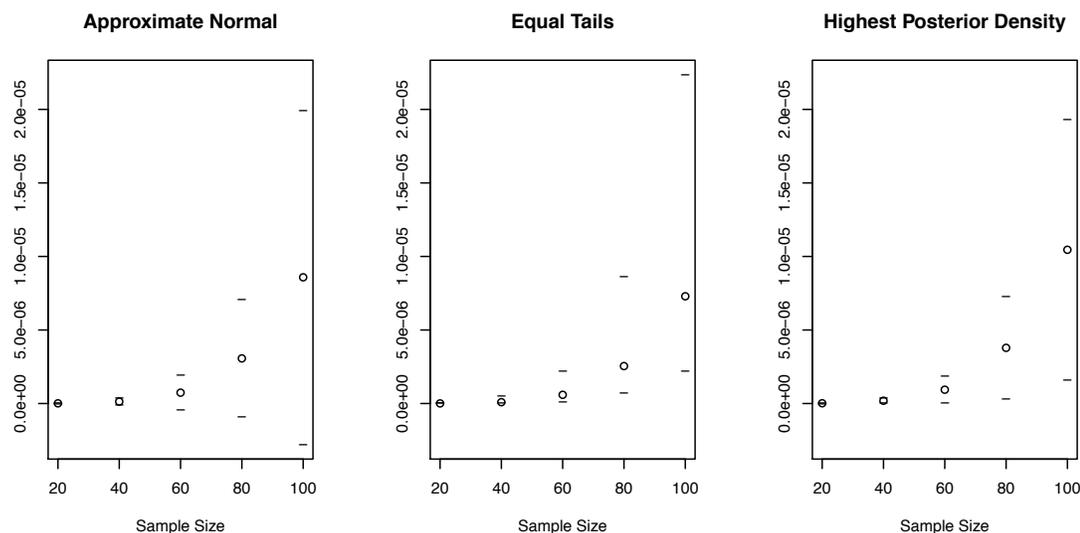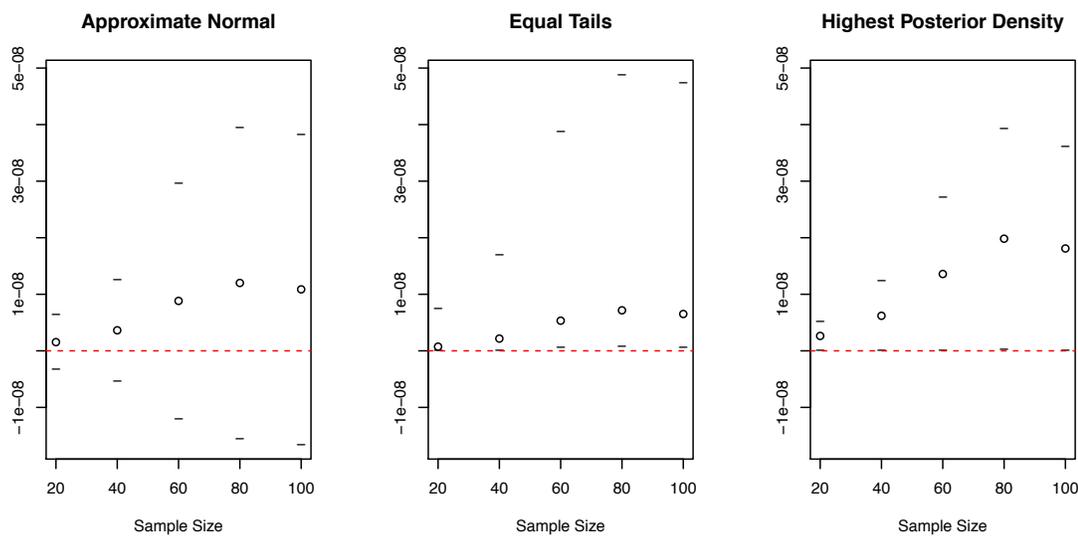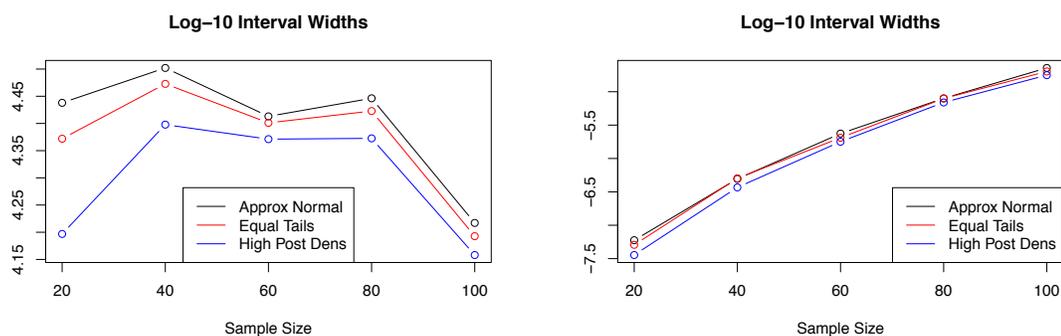

(b) Credible intervals constructed using prior distributions suggested by groups 2 and 3 of the observed glass data.

Figure 8.2: Credible intervals for the specific source likelihood ratio under the H_d scenario. The center of the intervals is plotted with a ∘ and the endpoints of the intervals are plotted with a −. The true value of the likelihood ratio is plotted with a red dotted line.

(a) Credible intervals constructed using prior distributions centered on the true parameter values.
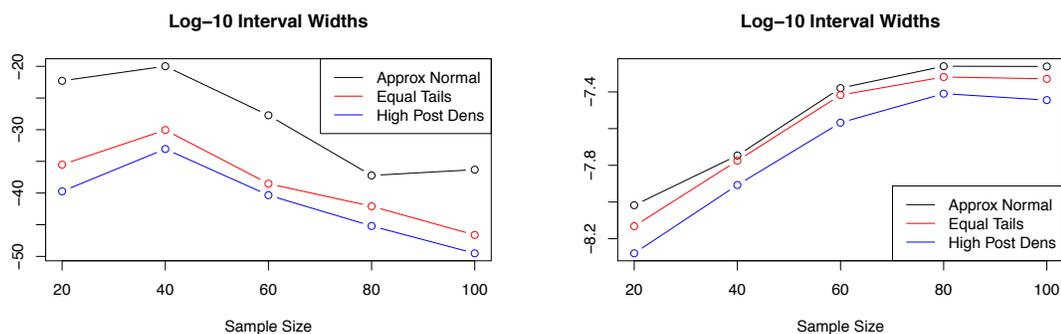

(b) Credible intervals constructed using prior distributions suggested by groups 2 and 3 of the observed glass data.

Figure 8.3: The log-10 widths of each of the intervals for the specific source problem under the H_p scenario.

(a) Credible intervals constructed using prior distributions centered on the true parameter values.

(b) Credible intervals constructed using prior distributions suggested by groups 2 and 3 of the observed glass data.

Figure 8.4: The log-10 widths of each of the intervals for the specific source problem under the H_d scenario.
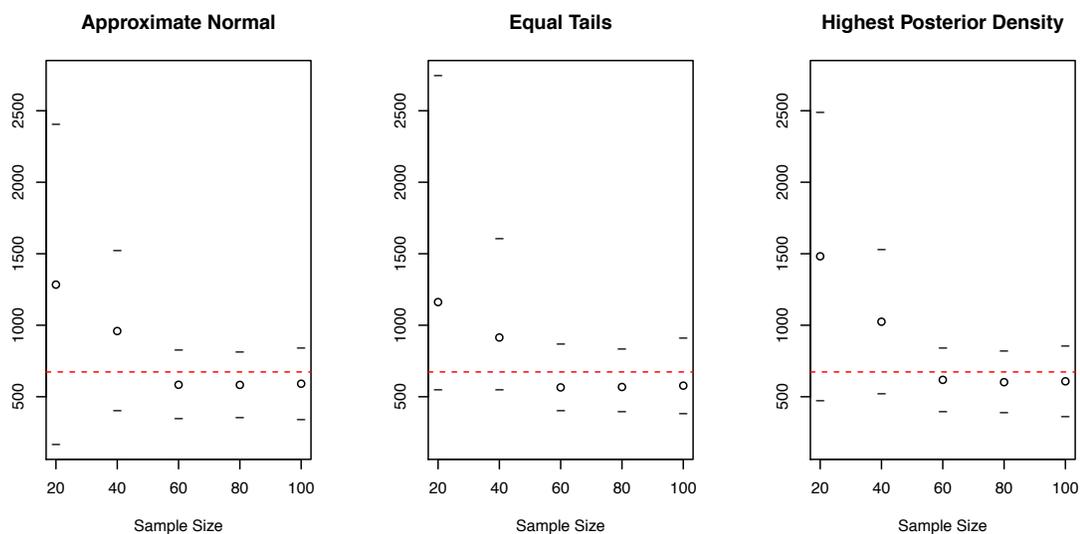

(a) Credible intervals constructed using prior distributions centered on the true parameter values.

(b) Credible intervals constructed using prior distributions suggested by groups 2 and 3 of the observed glass data.

noted that the y-axes in Figure 8.2(a) are not plotted on the same scale. Therefore, Figure 8.4 can be referenced to better compare the widths of the intervals. It can be seen in Figure 8.4 that under the first prior choice, the widths of the intervals tend to decrease, which is the desired behavior. When the prior distributions are mismatched to the observed evidence, as is the case for the second choice of prior, the intervals tend to become more spread out.

Figure 8.5 shows the behavior of the credible intervals for the common source likelihood ratio under the H_p scenario as the sample size increases. Like the common source problem, the credible intervals for the first choice of prior distributions contains the true values of the likelihood ratio for the simulated evidence. However, the credible intervals for the mismatched, second choice of prior distributions do not contain the true value of the likelihood ratio. Figure 8.7 shows that the widths of the intervals drop off quickly, and then stabilize. Also, the credible intervals tend to be shorter for the first prior choice compared to the credible intervals for the second prior choice.

Figure 8.6 shows the behavior of the credible intervals for the common source likelihood ratio under the H_d scenario as the sample size increases. It is difficult to determine from the figure, but none of the credible intervals for the second choice of prior contain the true value of the likelihood ratio for the simulated evidence. Figure 8.8 shows that the widths of the intervals drop off quickly, and then increases significantly. The odd behavior of these intervals indicates that maybe the number of posterior likelihood ratio samples should be increased, or that the simulated evidence for the second unknown source is potentially very rare. This might cause the computational algorithms to suffer from issues concerning ratios of very small probabilities which fluctuate considerably.

Figure 8.5: Credible intervals for the common source likelihood ratio under the H_p scenario. The center of the intervals is plotted with a $\circ$ and the endpoints of the intervals are plotted with a $-$. The true value of the likelihood ratio is plotted with a red dotted line.
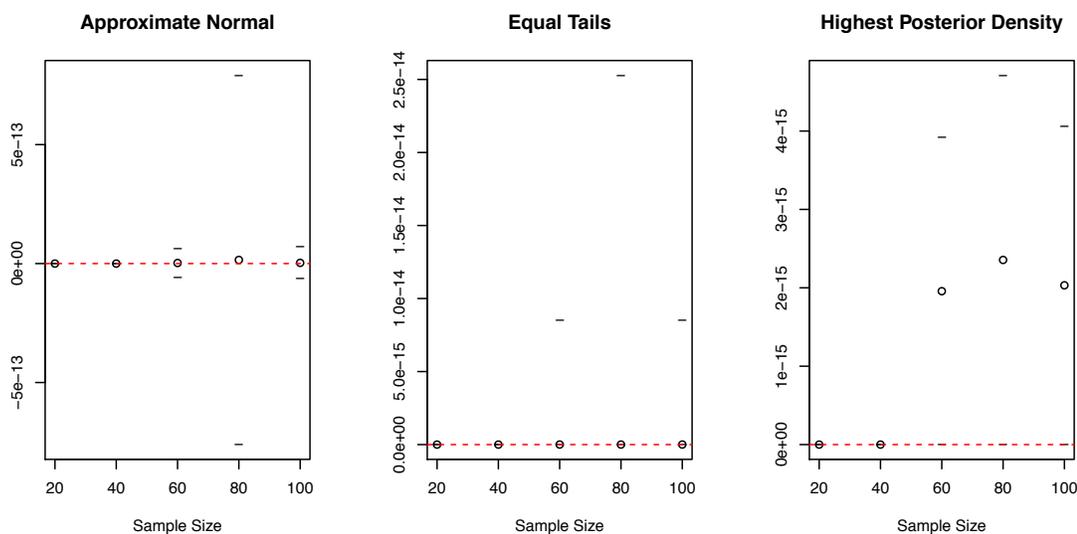

(a) Credible intervals constructed using prior distributions centered on the true parameter values.
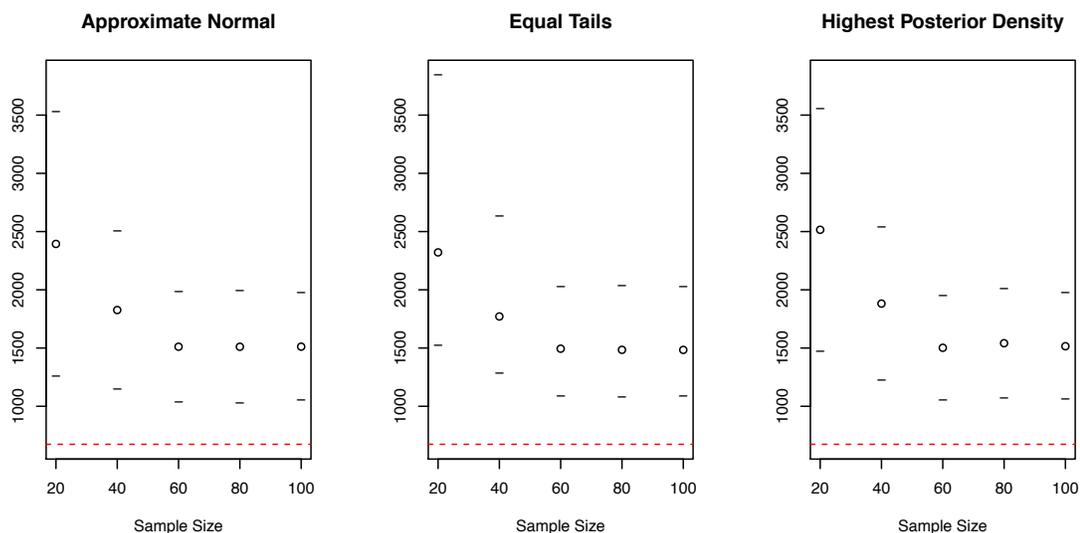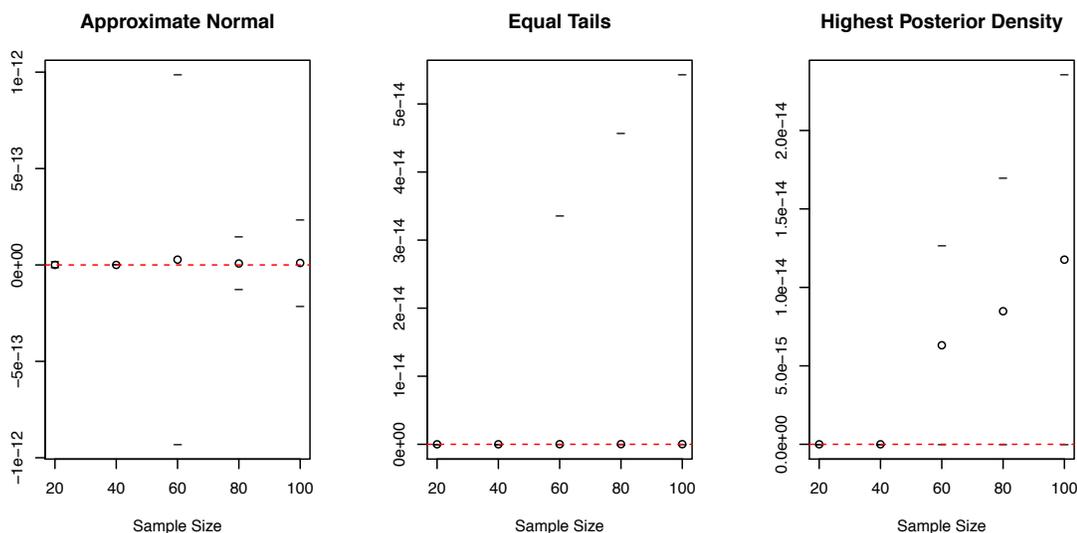

(b) Credible intervals constructed using prior distributions suggested by groups 2 and 3 of the observed glass data.

Figure 8.6: Credible intervals for the common source likelihood ratio under the H_d scenario. The center of the intervals is plotted with a $\circ$ and the endpoints of the intervals are plotted with a $-$. The true value of the likelihood ratio is plotted with a red dotted line.

(a) Credible intervals constructed using prior distributions centered on the true parameter values.

(b) Credible intervals constructed using prior distributions suggested by groups 2 and 3 of the observed glass data.

Figure 8.7: The log-10 widths of each of the intervals for the common source problem under the H_p scenario.
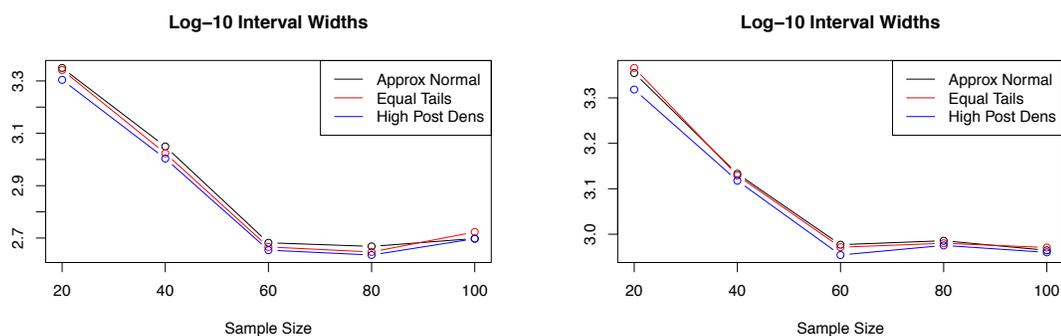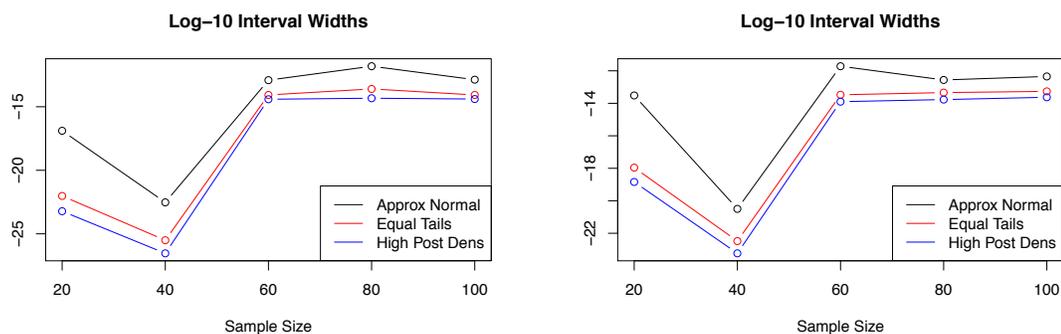

(a) Credible intervals constructed using prior distributions centered on the true parameter values.

(b) Credible intervals constructed using prior distributions suggested by groups 2 and 3 of the observed glass data.

Figure 8.8: The log-10 widths of each of the intervals for the common source problem under the H_d scenario.

(a) Credible intervals constructed using prior distributions centered on the true parameter values.

(b) Credible intervals constructed using prior distributions suggested by groups 2 and 3 of the observed glass data.

8.4 Discussion

It should be noted that as a result of Corollary 7.6.1, that the Neyman-Pearson approximation may replace the Bayes Factor for an approximately equivalent credible interval for the likelihood ratio. Also, similar to the Bayes Factor, these credible intervals for the likelihood ratio are dependent on the prior choice. In fact, if the prior distributions are changed, then the credible intervals for the likelihood ratio will change accordingly. If a chosen prior is particularly mismatched to the observed evidence, then the resulting credible intervals for the likelihood ratio may not contain the true value of the likelihood ratio. Regardless, these intervals may play a significant role in the determination of proper sample sizes for the forensic identification of source problems. This is a topic of particular interest to the forensic science community that will require future research.

CHAPTER 9

Conclusion

9.1   Concluding Remarks

Due to the intractability of computing the marginal likelihoods which compose the numerator and denominator of the Bayes Factor, Monte Carlo integration methods were implemented for computing the Bayes Factor. Since two separate approximations are used for computing the numerator and denominator, it was necessary to develop methods for quantifying the numerical standard error associated with the computation of the entire Bayes Factor by Monte Carlo methods. The results of this preliminary study clearly demonstrate that the prior choice for the nuisance parameters is one of the main causes of differences between the Bayes Factor and the likelihood ratio. This research motivated the development of the Bernstein-von Mises and Neyman-Pearson approximations.

The central result of this dissertation is the formal development of a generalization of the Neyman-Pearson likelihood ratio test statistic which serves as an approximation to the Bayes Factor. To the best of my knowledge, the results in Chapter 7 are the first results demonstrating that a non-Bayesian statistic can be used as an approximation to the subjective Bayes Factor in all aspects of the forensic identification of source problems. These results are applicable to both the common source and specific source identification problems, and for a large class of reasonable prior distributions for the indexing parameters on the class of probability models for the

evidence. The Neyman-Pearson approximation can replace the Bayes Factor in the Bayesian decision-making process and in all intermediate steps, including reasonable interval quantifications for the likelihood ratio. One of the main advantages of the Neyman-Pearson approximation is that this method avoids the complications associated with choosing a subjective prior distribution for the nuisance parameters.

In the process of developing the Neyman-Pearson approximation, it was necessary to consider a normal approximation to the posterior distribution for the nuisance parameters. This lead to an approximation of the Bayes Factor based on the Bernstein-von Mises Theorem which has similar asymptotic properties as the Neyman-Pearson approximation. While this Bernstein-von Mises approximation of the Bayes Factor also avoids any subjective determination of prior distributions, it possesses similar computational issues as the Bayes Factor. Therefore, methods for the computational implementation of the Bernstein-von Mises approximation were developed using Monte Carlo integration in Chapter 6. To avoid some of these computational issues, the Laplace approximation of the Bayes Factor can be used. However, the Laplace approximation still requires subjective prior determinations for the nuisance parameters.

One of the preliminary results necessary to study the asymptotic properties of the value of evidence develops a relationship between the subjective Bayes Factor and its limiting form, the likelihood ratio. In addition, a new expression for the common source and specific source Bayes Factors was developed. This new expression takes the form of the expected value of the corresponding likelihood ratio function with respect to the posterior distribution for the entire set of evidence under the defense model. These new non-standard expressions for the Bayes Factor facilitate the derivation of a posterior distribution of the likelihood ratio. This posterior distribution allows for a number of interval quantifications of the value of evidence that have the property

that the center of the interval is the Bayes Factor. These intervals avoid the bias properties typical of most interval-based value of evidence approaches.

9.2   Future Research

As a consequence of the results in this dissertation, I would like to focus my future research efforts on demonstrating that when the Neyman-Pearson approximation is used as a surrogate for the Bayes Factor, the resulting decision will be a statistically admissible decision rule for selecting between the prosecution and defense models for the forensic identification of source problems.

Conjecture:
For the forensic identification of source problems, choosing the model with the highest approximate posterior probability, where the approximate posterior odds are determined by multiplying the Neyman-Pearson approximation of the Bayes Factor by the prior odds, is the (generalized) Bayes Rule with respect to the "0-1" loss function.

It should be noted that a (generalized) Bayes Rule is a decision rule which minimizes the posterior expected loss with respect to the "0-1" loss functions and the (improper) prior distribution [10]. Several theorems from Berger [10] provide conditions under which (generalized) Bayes Rules are statistically admissible. It is expected that decisions based on the conjecture stated above will be statistically admissible. Although the Bernstein-von Mises approximation has asymptotic properties similar to the Neyman-Pearson approximation, it is not expected that using the Bernstein-von Mises approximation will lead to a statistically admissible decision rule.

As a natural extension of the results related to interval quantifications for the value of evidence, I would like to use a simulation algorithm to determine how many control samples are needed, relative to a given choice of categorical/verbal scale used in

the presentation of evidence [24], to guarantee that the reported value of evidence is reasonably close to the actual likelihood ratio. The algorithm will use posterior predictive distributions for the evidence in a manner similar to a parametric bootstrap to generate several samples from the posterior distribution for the likelihood ratio. The lengths of the resulting credible intervals for the likelihood ratio, in conjunction with the categorical scale, will be used as a metric to determine an appropriate sample size that will guarantee that the reported value of evidence is reasonable. The development of these methods will allow modern Bayesian methods for designing clinical trials to be used in sample size determinations for the forensic identification of source problems [12]. This research will be conducted with Dr. Saunders and Dr. Neumann to fulfill the requirements for the final phase of the National Institute of Justice grant listed in the acknowledgements.

Finally, the results of the simulation study for computing Bayes Factors revealed that the Bayes Factors are particularly sensitive to the choice of prior distribution for the nuisance parameters, especially for the specific source identification problem. Therefore, it would be beneficial to perform a systematic study of reference priors for various types of common forensic evidence. In this context, a reference prior is a prior distribution that would be used in all applications of a similar nature to mitigate any differences in reported values of evidence as a consequence of different practitioners choosing different prior distributions. These reference priors could then be managed by the various agencies for practical standards in forensic science, such as the National Institute for Standards and Technology in the United States.

REFERENCES

[1] 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 (2004), pp. 109–122.

[2] C. G. G. Aitken, P. Roberts, and G. Jackson, *Fundamentals of Probability and Statistical Evidence in Criminal Proceedings; Guidance for Judges, Lawyers, Forensic Scientists and Expert Witnesses*, Royal Statistical Society's Working Group on Statistics and the Law, London, UK, 1st ed., 2010.

[3] C. G. G. Aitken and F. Taroni, *Statistics and the Evaluation of Evidence for Forensic Scientists*, John Wiley and Sons, Ltd., West Sussex, UK, 2nd ed., 2004.

[4] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, John Wiley and Sons, Ltd., Hoboken, NJ, USA, 3rd ed., 2003.

[5] R. B. Ash and C. A. Doleans-Dade, *Probability and Measure Theory*, Harcourt/Academic Press, San Diego, CA, USA, 2000.

[6] D. J. Balding, *Weight-of-evidence for Forensic DNA Profiles*, Statistics in Practice, John Wiley and Sons, Ltd., Hoboken, NJ, USA, 2005.

[7] D. Bates and M. Maechler, *Matrix: Sparse and Dense Matrix Classes and Methods*, 2016, https://CRAN.R-project.org/package=Matrix. R package version 1.2-6.

[8] G. W. Beecham and B. S. Weir, *Confidence interval of the likelihood ratio associated with mixed stain DNA evidence*, Journal of Forensic Sciences, 56

(2011), pp. 166–171.

[9] C. E. Berger and K. Slooten, *The LR does not exist*, Science and Justice, 56 (2016), pp. 388–391.

[10] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*, Springer Series in Statistics, Springer, New York, NY, USA, 2nd ed., 1985.

[11] J. O. Berger, J. M. Bernardo, and D. Sun, *The formal definition of reference priors*, The Annals of Statistics, 37 (2009), pp. 905–938.

[12] S. M. Berry, B. P. Carlin, J. J. Lee, and P. Muller, *Bayesian Adaptive Methods for Clinical Trials*, CRC Biostatistics Series, CRC Press, Boca Raton, FL, USA, 2011.

[13] A. Biedermann, S. Bozza, F. Taroni, and C. G. G. Aitken, *Reframing the debate: A question of probability, not of likelihood ratio*, Science and Justice, 56 (2016), pp. 392–396.

[14] J. Buckleton, C. M. Triggs, and S. J. Walsh, eds., *Forensic DNA Evidence Interpretation*, CRC Press, Boca Raton, FL, USA, 2005.

[15] B. P. Carlin and S. Chib, *Bayesian model choice via Markov Chain Monte Carlo methods*, Journal of the Royal Statistical Society. Series B (Methodological), 57 (1995), pp. 473–484.

[16] K. P. S. Chan and C. G. G. Aitken, *Estimation of the Bayes' Factor in a forensic science problem*, Journal of Statistical Computation and Simulation, 33 (1989), pp. 249–264.

[17] R. Cook, I. W. Evett, G. Jackson, P. J. Jones, and J. A. Lambert, *A hierarchy of propositions: deciding which level to address in casework*, Science and Justice, 38 (1998), pp. 231–239.

[18] J. M. CURRAN, *Admitting to uncertainty in the LR*, Science and Justice, 56 (2016), pp. 380–382.

[19] J. M. CURRAN, J. S. BUCKLETON, C. M. TRIGGS, AND B. S. WEIR, *Assessing uncertainty in DNA evidence caused by sampling effects*, Science and Justice, 42 (2002), pp. 29–37.

[20] A. P. DAWID, *Forensic likelihood ratio: Statistical problems and pitfalls*, Science and Justice, 57 (2017), pp. 73–75.

[21] J. R. DETTMAN, A. A. CASSABAUM, C. P. SAUNDERS, D. L. SNYDER, AND J. BUSCAGLIA, *Forensic discrimination of copper wire using trace element concentrations*, Analytical Chemistry, 86 (2014), pp. 8176–8182.

[22] R. M. DUDLEY, *Real Analysis and Probability*, Mathematics Series, Wadsworth and Brooks/Cole Advanced Books and Software, Belmont, CA, USA, 1989.

[23] B. EFRON, *The Jackknife, the Bootstrap and Other Resampling Plans*, Society for Industrial and Applied Mathematics, 1982, https://doi.org/10.1137/1.9781611970319.

[24] EUROPEAN NETWORK OF FORENSIC SCIENCE INSTITUTES, *ENFSI Guideline for Evaluative Reporting in Forensic Science*, 2015.

[25] I. W. EVETT, *The interpretation of refractive index measurements*, Forensic Science, 9 (1977), pp. 209–217.

[26] I. W. EVETT, *The interpretation of refractive index measurements: II*, Forensic Science International, 12 (1978), pp. 27–47.

[27] I. W. EVETT, *A Bayesian approach to the problem of interpreting glass evidence in forensic science casework*, Journal of the Forensic Science Society, 26 (1986), pp. 3–18.

[28] I. W. Evett and J. A. Lambert, *The interpretation of refractive index measurements: III*, Forensic Science International, 20 (1982), pp. 237–245.

[29] I. W. Evett and J. A. Lambert, *The interpretation of refractive index measurements: IV*, Forensic Science International, 24 (1984), pp. 149–163.

[30] I. W. Evett and J. A. Lambert, *The interpretation of refractive index measurements: V*, Forensic Science International, 27 (1985), pp. 97–110.

[31] T. S. Ferguson, *Mathematical Statistics: A Decision Theoretic Approach*, Academic Press, New York, NY, USA, 1967.

[32] R. A. Fisher, *Statistical Methods for Research Workers*, Oliver and Boyd, Edinburgh, UK, 5th ed., 1934.

[33] B. Flury, *A First Course in Multivariate Statistics*, Springer Texts in Statistics, Springer-Verlag, New York, NY, USA, 1997.

[34] S. Fruhwirth-Schnatter, *Finite Mixture and Markov Switching Models*, Springer Series in Statistics, Springer, New York, NY, USA, 2006.

[35] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*, Texts in Statistical Science, CRC Press, Boca Raton, FL, USA, 3 ed., 2014.

[36] J. Geweke, *Bayesian inference in econometric models using Monte Carlo integration*, Econometric, 57 (1989), pp. 1317–1339.

[37] S. Ghosal and A. van der Vaart, *Fundamentals of Nonparametric Bayesian Inference*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge, UK, 2017.

[38] I. J. Good, *"Weight of evidence and the Bayesian likelihood ratio" in The Use of Statistics in Forensic Science*, CRC Press, Boca Raton, FL, USA, Oct. 1991.

[39] D. M. Grove, *The interpretation of forensic evidence using a likelihood ratio*, Biometrika, 67 (1980), pp. 243–264.

[40] J. D. Hadfield, *MCMC methods for multi–response generalised linear mixed models: The MCMCglmm R package*, Journal of Statistical Software, 33 (2010), pp. 1–22.

[41] C. Han and B. P. Carlin, *MCMC methods for computing Bayes factors: a comparative review*, Journal of the American Statistical Association, 96 (2001), pp. 1122–1132.

[42] A. Hepler, C. P. Saunders, L. Davis, and J. Buscaglia, *Score-based likelihood ratios for handwriting evidence*, Forensic Science International, 219 (2012), pp. 129–140.

[43] A. J. Izenman, *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*, Springer Texts in Statistics, Springer, New York, NY, USA, 2013.

[44] R. E. Kass and A. E. Raftery, *Bayes factors*, Journal of the American Statistical Association, 90 (1995), pp. 773–795.

[45] Q. Y. Kwan, *Inference of Identify of Source*, Ph.D. Dissertation in Criminology, University of California, Berkeley, 1977.

[46] D. V. Lindley, *A problem in forensic science*, Biometrika, 64 (1977), pp. 207–213.

[47] S. P. Lund and H. Iyer, *Likelihood ratio as weight of forensic evidence: A metrological perspective*, 2016, https://arxiv.org/abs/1608.07598v2.

[48] M. Meredith and J. Kruschke, *HDInterval: Highest (Posterior) Density Intervals*, 2016, https://CRAN.R-project.org/package=HDInterval. R package

version 0.1.3.

[49] J. J. MILLER, *Asymptotic properties of maximum likelihood estimates in the mixed model of the analysis of variance*, Annals of Statistics, 5 (1977), pp. 746–762.

[50] G. S. MORRISON AND E. ENZINGER, *What should a forensic practitioner's likelihood ratio be?*, Science and Justice, 56 (2016), pp. 374–379.

[51] NATIONAL RESEARCH COUNCIL COMMITTEE ON IDENTIFYING THE NEEDS OF THE FORENSIC SCIENCES COMMUNITY, *Strengthening Forensic Science in the United States: A Path Forward*, The National Academies Press, Washington, D.C., USA, 2009.

[52] R. NEAL, *The Harmonic Mean of the Likelihood: Worst Monte Carlo Method Ever*, 2008, http://radfordneal.wordpress.com/2008/08/17/the-harmonic-mean-of-the-likelihood-worst-monte-carlo-method-ever/ (accessed 15 May 2014).

[53] M. A. NEWTON AND A. E. RAFTERY, *Approximate Bayesian inference with the weighted likelihood bootstrap*, Journal of the Royal Statistical Society. Series B (Methodological), 56 (1994), pp. 3–48.

[54] D. M. OMMEN, C. P. SAUNDERS, AND C. NEUMANN, *An argument against presenting interval quantifications as a surrogate for the value of evidence*, Science and Justice, 56 (2016), pp. 383–387.

[55] D. M. OMMEN, C. P. SAUNDERS, AND C. NEUMANN, *The characterization of Monte Carlo errors for the quantification of the value of forensic evidence*, Journal of Statistical Computation and Simulation, 87 (2017), pp. 1608–1643.

[56] J. B. PARKER, *A statistical treatment of identification problems*, Journal of the Forensic Science Society, 6 (1966), pp. 33–39.

[57] J. Pinheiro, D. Bates, S. DebRoy, D. Sarkar, and R Core Team, *nlme: Linear and Nonlinear Mixed Effects Models*, 2016, http://CRAN.R-project.org/package=nlme. R package version 3.1-128.

[58] President's Council of Advisors on Science and Technology, *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*, Executive Office of the President of the United States, Sept. 2016.

[59] C. P. Robert, *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implements*, Springer Texts in Statistics, Springer, New York, NY, USA, 2001.

[60] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, Springer Texts in Statistics, Springer, New York, NY, USA, 2nd ed., 2004.

[61] R. Royall, *Statistical Evidence: A Likelihood Paradigm*, vol. 71, CRC Press, Boca Raton, FL, USA, 1997.

[62] C. P. Saunders, *Empirical Processes for Estimated Projections of Multivariate Normal Vectors with Applications to E.D.F. and Correlation Type Goodness of Fit Tests*, Ph.D. Dissertation in Statistics, University of Kentucky, 2006, http://uknowledge.uky.edu/gradschool_diss/464.

[63] C. P. Saunders, L. J. Davis, A. C. Lamas, J. J. Miller, and D. T. Gantz, *Construction and evaluation of classifiers for forensic document analysis*, The Annals of Applied Statistics, 5 (2011), pp. 381–399.

[64] A. Seheult, *On a problem in forensic science*, Biometrika, 65 (1978), pp. 646–648.

[65] R. J. Serfling, *Approximation Theorems of Mathematical Statistics*, Wiley Series in Probability and Statistics, John Wiley and Sons, Ltd., New York, NY,

USA, 1980.

[66] G. Shafer, *Lindley's paradox*, Journal of the American Statistical Association, 77 (1982), pp. 325–334.

[67] D. A. Shum, *The Evidential Foundations of Probabilistic Reasoning*, Northwestern University Press, Evanston, IL, USA, 2001.

[68] M. J. Sjerps, I. Alberink, A. Bolck, R. Stoel, P. Vergeer, and J. H. van Zanten, *Uncertainty and LR; to integrate or not to integrate, that's the question*, Law, Probability, and Risk, 15 (2016), pp. 23–29.

[69] C. L. Smith, *Model Selection for Nonnested Linear Mixed Models*, Ph.D. Dissertation in Biostatistics, University of North Carolina at Chapel Hill, 2015.

[70] M. A. Tanner, *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, Springer, New York, NY, USA, 3rd ed., 1996.

[71] 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 (2016), pp. 1–16.

[72] D. Taylor, T. Hicks, and C. Champod, *Using sensitivity analyses in Bayesian Networks to highlight the impact of data paucity and direct future analyses: a contribution to the debate on measuring and reporting the precision of likelihood ratios*, Science and Justice, 56 (2016), pp. 402–410.

[73] S. van de Geer, *Empirical Processes in M-Estimation*, Cambridge University Press, Cambridge, UK, 2000.

[74] A. van den Hout and I. Alberink, *Posterior distributions for likelihood ratios in forensic science*, Science and Justice, 56 (2016), pp. 397–401.

[75] A. W. van der Vaart, *Asymptotic Statistics*, Cambridge University Press, Cambridge, UK, 1998.

[76] A. W. van der Vaart and J. Wellner, *Weak Convergence and Empirical Processes*, Springer Series in Statistics, Springer, New York, NY, USA, 2000.

[77] W. R. Wade, *An Introduction to Analysis*, Pearson, Upper Saddle River, NJ, USA, 4th ed., 2010.

[78] L. Wasserman, *All of Statistics: A Concise Course in Statistical Inference*, Springer, New York, NY, USA, 2010.

[79] J. Wayman, *What is probability?*, in Technical Colloquium on Quantifying the Weight of Forensic Evidence, https://www.nist.gov/sites/default/files/documents/2016/12/07/04_wayman _what_is_probability_2016.pdf, https://www.nist.gov/news-events/events/2016/ 05/ibpc-technical-colloquium-quantifying-weight-forensic-evidence, 2016, National Institute of Standards and Technology.

CURRICULUM VITAE

Education

2017	*Ph.D. Candidate in Computational Science & Statistics* South Dakota State University
2014	*M.S. in Mathematics with emphasis in Statistics* South Dakota State University
2012	*B.S. in Mathematics* South Dakota State University

Professional Experience

starting 08/2017	*Assistant Professor of Statistics* Iowa State University
12/2014 – 08/2017	*Graduate Research Assistant* South Dakota State University
05/2015 – 08/2015	*Visiting Scientist* Oak Ridge Institute for Science and Education and Federal Bureau of Investigation Laboratory Division Counterterrorism and Forensic Science Research Unit
08/2012 – 12/2014	*Graduate Teaching Assistant* South Dakota State University

Publications

Danica M. Ommen, Christopher P. Saunders, and Cedric Neumann (2017) *"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, DOI: 10.1080/00949655.2017.1280036.

Danica M. Ommen, Christopher P. Saunders, and Cedric Neumann, *"An Argument Against Presenting Interval Quantifications as a Surrogate for the Value of Evidence"*, Science and Justice 56 (2016) 383-387.

Invited Presentations & Panels

02/2017	*"A South Dakotan's View on the Difference between the Bayes Factor and the Likelihood Ratio"* Danica Ommen Netherlands Forensic Institute; The Hague, The Netherlands
05/2016	*"New Approaches to the Quantification of Trace Evidence for Source Identification"* Danica Ommen, Christopher Saunders, JoAnn Buscaglia Technical Colloquium: Quantifying the Weight of Forensic Evidence International Biometric Performance Testing Conference 2016; National Institute of Standards and Technology; Gaithersburg, MA
05/2016	*Panel on the Use of Interval Quantifications for the Value of Forensic Evidence* Technical Colloquium: Quantifying the Weight of Forensic Evidence International Biometric Performance Testing Conference 2016; National Institute of Standards and Technology; Gaithersburg, MA
07/2015	*"Recent Developments in Approximate Solutions to Forensic Source Identification Problems"* Danica Ommen, Chris Saunders, and JoAnn Buscaglia Algorithms for Threat Detection Program Review; Defense Threat Reduction Agency National Science Foundation; Arlington, VA

Oral Presentations & Panels

08/2016	*"Information Criteria Approximations to the Value of Evidence for Forensic Identification of Source Problems"* Danica Ommen and Christopher Saunders Joint Statistical Meetings; American Statistical Association; Chicago, IL (Contributed Presentation)
09/2015	*"Convergence of Different Computationally Efficient Approximations of the Weight of the Forensic Evidence"* Danica Ommen, Doug Armstrong, Cedric Neumann, Chris Saunders European Academy of Forensic Sciences Conference; Prague, Czech Republic
08/2015	*"Convergence of Different Computationally Efficient Approximations of the Weight of the Forensic Evidence"* Danica Ommen, Chris Saunders, and Cedric Neumann Joint Statistical Meetings; American Statistical Association; Seattle, WA

01/2015   Facilitated the Handwriting Modality Panel discussions
Symposium on Improving Biometric and Forensic Technology:
The Future of Research Datasets
National Institute of Standards and Technology; Gaithersburg, MD

10/2014   *"The Common Source Value of Evidence in the Presence of*
Uncertainty about the Alternative Source Population"
Danica Ommen
Computational Science and Statistics Seminar;
South Dakota State University; Brookings, SD

Poster Presentations

06/2016   *"Information Criteria Approximations to the Value of*
Evidence for Forensic Identification of Source Problems"
Danica Ommen, Christopher Saunders
International Society for Bayesian Analysis 2016 World Meeting;
Cagliari, Sardinia, Italy

09/2015   *"The Interpretation and Presentation of Trace Element*
Analysis of High Purity Copper Evidence"
Chris Saunders, Danica Ommen*, Joshua Dettman, JoAnn Buscaglia
European Academy of Forensic Sciences Conference;
Prague, Czech Republic

08/2014   *"Computational and Statistical Aspects of the Forensic*
Identification Source Problem"
Danica Ommen, Chris Saunders, and Cedric Neumann
Joint Statistical Meetings; American Statistical Association;
Boston, MA

Co-authored Presentations

08/2016   *"On the Different Classes of Forensic Identification of Source*
Problems"
Danica Ommen, Christopher Saunders*, and Cedric Neumann,
Joint Statistical Meetings; American Statistical Association;
Chicago, IL (Topic Contributed Presentation)

08/2014   *"Computational and Statistical Aspects of the Forensic Identification*
of Source Problem: The Specific Source Problem"
Danica Ommen, Chris Saunders*, and Cedric Neumann
International Conference on Forensic Inference and Statistics;
Leiden, Netherlands (Contributed Presentation)

08/2014	*"Computational and Statistical Aspects of the Forensic Identification of Source Problem: Asymptotic Properties of the Bayes Factor"*
	Danica Ommen, Chris Saunders*, and Cedric Neumann
	University of Salzburg; Salzburg, Austria (Invited Presentation)
05/2014	*"Computational and Statistical Aspects of the Forensic Identification of Source Problem: A General Overview"*
	Danica Ommen, Chris Saunders*, and Cedric Neumann
	Algorithms for Threat Detection Program Review
	Defense Threat Reduction Agency; National Science Foundation
	Boulder, CO (Invited Presentation)

Professional Memberships

International Society for Bayesian Analysis
Institute of Mathematical Statistics
Golden Key National Honour Society

Awards & Honors

Stephen E. Fienberg CSAFE Young Investigator Travel Award - $1500