Utrecht University

Performance of Specific Source and Common Source Bayes Factors for Trace-Reference Problems

Emilie Jessica Boudens

Utrecht University

Performance of Specific Source and Common Source Bayes Factors for Trace-Reference Problems

Master's Thesis

To fulfill the requirements for the degree of
Master of Science in Mathematical Sciences
at Utrecht University under the daily supervision of
Dr. Peter Vergeer (Netherlands Forensic Institute (NFI))
and
Dr. Cristian Spitoni (Department of Mathematics, Utrecht University)
with second reader Dr. Ivan Kryven (Department of Mathematics, Utrecht University)

Utrecht University

Netherlands Forensic Institute
Ministry of Justice and Security

Emilie Jessica Boudens (5510775)

December 2, 2021

Acknowledgments

I am very grateful for everybody who has supported me through this process. In particular, the following people:
I'd mostly like to thank my daily supervisor Peter Vergeer for letting me execute one of his ideas and guiding me through the process. For helping me improve my programming skills and making me feel like a part of their team. In addition, I am also grateful to Marjan Sjerps, Ivo Alberink and Leen van der Ham for making me feel welcome in their team at the NFI. Furthermore I am thankful for Cristian Spitoni, for giving me very detailed feedback and being critical on the mathematics I used and assumptions I have made. I also want to mention my colleagues at El Mundo, for being a good distraction and making my weekends fun and long. Lastly, I want to thank my parents and my partner for always supporting me and helping me believe in myself when I got a bit lost or stressed and for pushing me to work as hard as I can.

Thank you, everyone!

Abstract

In the forensic science community, a distinction between the specific source and common source scenarios arose a couple years ago. In the forensic scientific literature, the specific source scenario is used when the question is whether a trace is from a specific source or not. This is considered to be a trace-reference problem. The common source scenario is used to answer the question whether two traces share an unknown common source or not. This is considered to be a trace-trace problem. The specific source scenario and the common source scenario also imply different modelling. In principle, it is possible to apply the modelling assumptions of the common source scenario not only to a trace-trace comparison but also to a trace-reference comparison. Because the specific source is not assumed to be part of the background population, we lose information about the parameters that describe the background population. What happens if we use the common source Bayes Factors to update the prior odds of the specific source hypotheses to posterior odds? Do they give us more information? We used simulated data for two models to compare the common source and specific source Bayes Factor systems and see which model provided more information for the trace-reference problem.

Contents

1 Introduction

When a crime has been committed, different types of evidence may be found at the crime scene. Examples of these types are biometric evidence (such as DNA or finger prints), pattern evidence (such as handwriting) or trace evidence (such as glass fragments or clothing fibres) of which we examine the elemental compositions [21]. An important question in a crime scene investigation is the identification of the source of certain evidence. The source can be a person or a specific object, such as a gun or a window. The term 'identification' in this context means that all possible sources except for one that we are interested in, are to be excluded [29]. Unfortunately, actual identification is usually not possible due to randomness in measurements, degradation of traces or the fact that multiple sources might have the same forensic characteristics. This is one reason why forensic scientists make use of *likelihood ratios*. A likelihood ratio is a ratio of two probabilities: the probabilities of the observed evidence conditioned on two mutually exclusive hypotheses for how the evidence was generated. These hypotheses are often referred to as the prosecution hypothesis (H_p) and the defence hypothesis (H_d). A court room is more interested in probabilities, while forensic scientists work with likelihoods. In other words: a detective or court is interested in questions of the type: "what is the probability that the person of interest was at the crime scene?", but a forensic scientist is concerned with the question: "what is the probability of the evidence, given that the suspect was (or alternatively, was not) at the crime scene?". To summarize: the forensic scientists are *not* expected to deliver probabilities, but can only make a statement about likelihood ratios, giving the *value of evidence* [13]. A visualization of this can be found in Figure 1.

Definition 1.1 (Likelihood Ratio)**.** A likelihood ratio (or LR) is given by the following expression

$$LR(E) := \frac{\mathbb{P}(E \mid H_p, I)}{\mathbb{P}(E \mid H_d, I)}, \tag{1}$$

where $\mathbb{P}$ denotes a probability measure, E denotes the considered forensic evidence (modeled as an event in the probability space) and I the relevant background information that is the same for both hypotheses. This notation will be used throughout my whole thesis.

The background information is different for all criminal cases, but generally consists of police findings, witness testimonies, etc. We will not specifically model I, but it is considered to be known. The likelihood ratio is a measure of the evidential value of the evidence E and can be used to update prior odds of the hypotheses to posterior odds (see theorem 2.1).

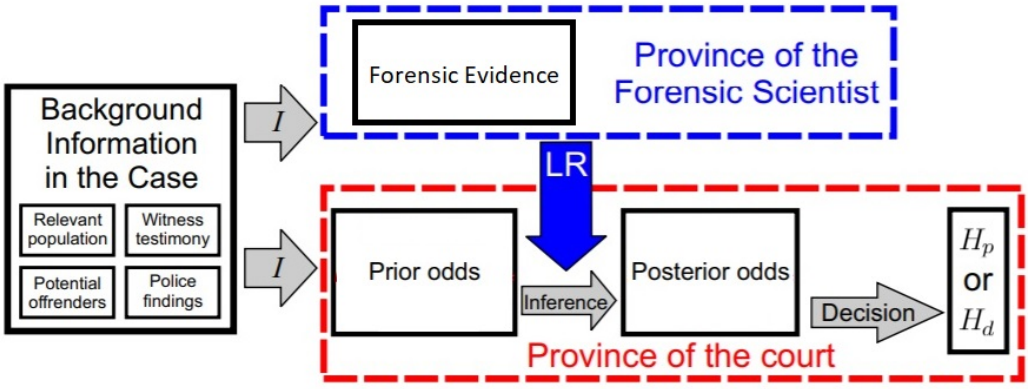


Figure 1: Elements of the decision process. (Adapted from [33])

1.1 Identification of Common Source

A common problem in forensic science is to determine whether two crimes are related. Consider, for example, that two different girls were attacked by an unknown person in the same area within the same week and that hair was found on their clothes, which didn't originate from their body. A logical question to ask is if they were attacked by the same person. The hairs will then be examined to see if they belonged to the same person. This is an example of the common source identification problem since the question of interest is whether or not the two girls were attacked 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 both originate from the same unknown source.

H_d: The two sets of unknown source evidence originate from two different unknown sources.

We adapt the notation of Ommen [29] to denote the evidence set for the common source problem as $E = \{e_{u_1}, e_{u_2}, e_a\}$. These elements are defined as:

- e_{u_1}: measurements performed on the evidence that originates from the first unknown source;

- e_{u_2}: measurements performed on evidence that originates from the second unknown source;

- e_a: measurements performed on the population of alternative sources, often referred to as the background population;

1.2 Identification of Specific Source

While the common source problem is often helpful to solve a case, for the court it is ultimately more interesting to determine whether a suspect can be linked to the evidence found at the crime scene. Following the previous example, suppose a person has been arrested for attacking (one of) the girls. His hair will be examined to compare it to the hairs found on the victim. This is an example of the specific source identification problem, since the question of interest is whether or not the specified suspect left his hair on the victim. In the common source problem, the person who left the hair is not identified and is treated as unknown. In the specific source problem, the person is identified as the suspect or person of interest, and is treated as known. The forensic hypotheses for the specific source problem are typically stated as follows:

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

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

The complete evidence set for the specific source problem is given by $E = \{e_s, e_u, e_a\}$, where the elements are defined as:

- e_s: measurements performed on evidence that originates from the known specific source;
- e_u: measurements performed on evidence that originates from the unknown source;
- e_a: measurements performed on the population of alternative sources.

1.3 Research Questions

This distinction between specific source and common source scenario arose only a couple years prior to this thesis [29]. It was brought up because the specific source is a known source to which we want to compare trace evidence and this approach results in a different interpretation of the results. Before the specific source system was presented, the most commonly used hypotheses in forensic science were the same source / different source hypotheses

(which are equivalent to the hypotheses in the common source scenario). A discussion arose along with the specific source scenario: which scenario is best to evaluate the value of evidence? The specific source question is often of most interest in the courtroom, because this likelihood ratio gives information about how the specific source might be related to the crime. The common source question might be more important during the investigating phase of a crime, because it may relate two crime scenes or places to each other. This is what has been published primarily in the literature [27, 30]. An important motivation for this research is the difference between the common source and specific source model specifications. In the literature it is assumed that the specific source is a known source, which is not exchangeable and the parameters of the probability distribution of this source are considered independent from the parameters that describe the background population [30]. This is not the case when we consider the common source scenario: we will assume that the source of the reference evidence is known, but now it's part of the relevant background population. Depending on the relevance of the background population, there is potentially a lot of information buried in the background population, which we do not consider to be as important in the specific source problem as in the common source problem. This motivated us to find a way to see if we can still use this information to update the prior probabilities of the specific scenario to the posterior probabilities (see Theorem 2.1). This leads to the following questions we wish to research and hopefully answer:

Q_1: Can we use the Common Source Likelihood Ratio models to update the prior probabilities of the Specific Source hypotheses to posterior probabilities?

Q_2: Does the Common Source Likelihood Ratio, in some instances, have more value than the Specific Source Likelihood Ratio when updating the prior probabilities of the Specific Source hypotheses?

These questions have not been considered in the literature, because the distinction between these two models itself is still very new, so they have not been compared in this way yet. We want to provide a way for the forensic science community to use common source statistical models to address specific source questions.

1.4 Thesis Outline

Before we address these questions, in Chapter 2 we introduce mathematical methods and concepts needed in the thesis. Then we will consider two different models. First we will define a generative[1] Beta-Binomial model for discrete evidence, of which we will describe the results in Chapter 3. We will compare common source and specific source Bayes Factors from [36] to assess their value in updating the specific source prior probabilities. In Chapter 4 we will define a generative hierarchical normal-normal model for continuous evidence following the model given by Ommen et al. in [31]. We use Markov Chain Monte Carlo sampling to be able to approximate the Bayes Factors and then use ECE-plots to determine their performance. Chapter 5 provides a discussion and comparison of the results of the two models.

We try to keep the models as simple and general as possible, because general models are easy to interpret. And secondly because we are only interested in comparing the performance of the likelihood ratios in the two scenarios and not specifically in how these likelihood ratios are determined, because we already know the fundamental differences between the models. For example, we will use scalar evidence instead of using more-dimensional evidence. Only one feature of the traces will be measured, while in reality we could look at more than one.

[1] A generative model describes the distribution of the data itself.

2 Concepts and Mathematical Methods Used

In this chapter we will introduce the mathematical background used for defining our models, validating them and assessing the performance of the different likelihood ratios.

2.1 Bayes Factors

Defining the probability $\mathbb{P}$ in definition 1.1 can be done in two ways within the forensic science community. The frequentist paradigm computes equation (1) directly, given some choice for the parameter values. The Bayesian paradigm assigns a probability $\mathbb{P}$ to the evidence, based on the prior belief that we have about the parameter values. In the Bayesian paradigm, we do not compute likelihood ratios but Bayes Factors, because we work with unknown parameters [23]. The Bayesian method of quantifying the value of evidence centers around the odds form of Bayes' Theorem to convert prior odds of the two competing hypotheses to posterior odds via the likelihood ratio.

Theorem 2.1 (Odds form of Bayes' Theorem)**.**

$$\underbrace{\frac{\mathbb{P}(H_p \mid E)}{\mathbb{P}(H_d \mid E)}}_{\text{Posterior Odds}} = \underbrace{\frac{\mathbb{P}(E \mid H_p)}{\mathbb{P}(E \mid H_d)}}_{\text{Likelihood Ratio}} \times \underbrace{\frac{\mathbb{P}(H_p)}{\mathbb{P}(H_d)}}_{\text{Prior Odds}} \tag{2}$$

As we will continue to work in the Bayesian paradigm and we need to consider parameters with unknown value, the likelihood ratio will be replaced by the Bayes Factor. Throughout this research, the measurements on the evidence will follow parametric models. Some of these parameters might be unknown, which means we have to integrate them out of the expressions to determine the probabilities. This results in the formal definition of the Bayes Factor:

Definition 2.2 (Bayes Factor)**.** A Bayes Factor (or BF) is given by the following expression

$$BF(E) := \frac{\int_{\theta_p} f(E \mid H_p, \theta_p, I)\, f(\theta_p \mid H_p, I) d\theta_p}{\int_{\theta_d} f(E \mid H_d, \theta_d, I)\, f(\theta_d \mid H_d, I) d\theta_d}, \tag{3}$$

where f denotes the likelihood function, I is again the relevant background information that is the same for both hypotheses and θ_p and θ_d denote the model parameter spaces under H_p and H_d respectively, which may be the same parameters depending on the model.

We can view the Bayes Factor as the relative weighted likelihoods of observing the evidence under each hypothesis with respect to the corresponding prior measures on the parameters. When this probative value of evidence is presented to a judge or jury, they can use this value to update their personal prior belief about the two different hypotheses. A Bayes Factor less than one indicates that the evidence supports H_d over H_p, while a Bayes Factor greater than one indicates that the evidence supports H_p over H_d. A Bayes Factor that is equal to one means that the evidence cannot make a distinction between the two hypotheses. A Bayes Factor does *not* mean that one hypothesis 'has a higher probability' than the other. This is a common mistake made in practice, which leads us to think about how evidence should be interpreted and presented to the court or outside world. It should be clear that the interpretation of evidence takes place within a framework of circumstances. We make assumptions when generating a model and the more of these assumptions we make, the less complex the mathematics, the more tractable the solutions and the less difficult it is to determine our Bayes Factor. But at the same time we move further away from reality and we make the domain of applicability smaller which each assumption. It is therefore necessary that all these assumptions are made clear.

2.2 Markov Chain Monte Carlo Methods

Computing the integrals in the Bayes Factor may be computationally difficult or infeasible to do analytically. They can be simplified by partitioning the evidence and therefore reducing the amount of evidence that needs to be evaluated in the likelihood. A popular technique in forensic science to numerically approximate the Bayes Factor is Monte Carlo integration [24]. In order to integrate quantities of the form

$$m(x) = \int_{\theta} f(x \mid \theta) g(\theta) d\theta, \tag{4}$$

where f is the likelihood function for the data x indexed by parameters θ and g is the density of θ. Kass and Raftery discuss three different integration techniques [20]. We will use the *Arithmetic Mean Estimate* of $m(x)$, which is defined as follows: if n is the Monte Carlo sample size, the estimate is given by

$$\hat{m}(x) = \frac{1}{n} \sum_{i=1}^{n} f(x \mid \theta^{(i)}), \tag{5}$$

where $(\theta^{(1)}, \ldots, \theta^{(n)})$ is an independent sample drawn from $g(\theta)$. We will simplify the integrals in the Bayes Factors to derive a distribution for the parameters θ posterior on a subset $e \subset E$ of the evidence. This will give us the posterior density $g(\theta \mid e)$, which might be difficult to determine analytically. The typical solution is to draw independent samples from the probability distribution, then repeat this process many times to approximate the desired quantity. This is referred to as Monte Carlo sampling. For high dimensional problems however, this method is infeasible. When this problem arises, a common solution in forensic science is to use *Markov Chain Monte Carlo (MCMC)* methods [7, 17]. When using MCMC methods to infer a posterior distribution, samples are drawn from the desired probability distribution by constructing a Markov Chain, where the next sample that is drawn is dependent upon the previously drawn sample (while Monte Carlo sampling draws independent samples). The idea is that the chain will find its equilibrium on the desired distribution. An important foundation for this method is that the posterior distribution is proportional to the prior belief and the likelihood:

$$g(\theta \mid e) \propto \mathbb{P}(e \mid \theta) g(\theta).$$

So the bottom line is, that given a set of observations and a prior belief, MCMC can be used to compute a sample of the posterior distribution. The basic procedure is defined as follows:

1. Select an initial set of values for the parameters of which we wish to know the posterior distribution;

2. Randomly assign new values to the parameters based on the current state;

3. Check if the new random values agree with the observations to a certain extent (based on the chosen method). If they do agree, accept the values as the new current state. If they do not, reject the values and return to the previous state;

4. Repeat steps 2 and 3 for the specified number of iterations.

The output of our MCMC sampling are parameter values, which can be used to calculate the Bayes Factor. Widely used methods are Gibbs sampling, the Metropolis-Hasting algorithm and Hamiltonian Monte Carlo. We used PyMC3 [35] in Python to perform our MCMC sampling, which implements the *No U-Turn Sampler* (NUTS) [18] by default for continuous distributions.

2.3 Uncertainties

An important field of study for interpreting results of research is information theory. It also introduces a concept that is important to clarify: *uncertainty*. Uncertainty can often be divided into two categories [12]. It can arise when there are things we are unsure of simply due to lack of knowledge. This means that we can reduce our uncertainty by gathering more information on the subject. This type of uncertainty is what we call *epistemic uncertainty* and is often regarding quantities that have a fixed value in reality. A forensic example can be the uncertainty about which hypothesis is true. We can update our knowledge about this by gathering more evidence. Another example are the unknown parameter values of probability distributions, which we will encounter when calculating Bayes factors. Alternatively, we have *aleatory uncertainty*, which arises due to random variability and can therefore not be reduced by gathering more information. Examples are measurement errors or taking a random sample from a population. Aleatory uncertainty is present in almost all data in the field of statistics. Furthermore, since the whole purpose of statistics is to learn from data, there must also be epistemic uncertainty in all statistical problems. The uncertainty in the data themselves is both aleatory and epistemic, because there are always unknown parameters about which we want to learn more. The terms randomness and uncertainty have also been used for aleatory variability and epistemic uncertainty. However, these terms are commonly used in generic ways and as a result, they are often mixed up when used. In the Bayesian paradigm, epistemic and aleatory uncertainty are treated in a similar way: probability distributions are assigned to the relevant variables, based on the prior belief of what their values are. To clarify the previous concepts and their difference I would like to introduce an example (adapted from [28]):

> **Example 1.** If we toss an ordinary coin, the probability that it will land on heads is 0.5. Suppose that we also have a bag of poker chips, and we know that some are red and some are green, but we have no idea how many of each color or how many chips there are in total. Now we we want to know what the probability of pulling a red chip is, when we pull one chip out of the bag. Our state of knowledge on the composition of colors in the bag is complete ignorance. There may as well be the same amount of red and green chips. Based on a Bernoulli distribution that is common for these types of problems, if one chip is to be pulled out of the bag the probability that it will be red is 0.5, given that we do not know how many chips there are of each color. Now surely we can agree that

we have little epistemic uncertainty about the coin toss and a lot of epistemic uncertainty about the bag of poker chips. The uncertainty about the coin toss is purely aleatory, since the probability is caused by irreducible randomness, whereas there is clearly epistemic uncertainty about the precise content of the bag of chips. It becomes more interesting when we consider a sequence of tosses of that coin, and a sequence of chips drawn from the bag. Our uncertainty about the coin tosses is still purely aleatory, no matter how many times the coin is tossed, our probability of heads is still 0.5. On the other hand, as more chips are drawn from the bag, our epistemic uncertainty about its composition reduces, and my probability for the next chip being red changes according to the chips I have now seen. The epistemic uncertainty is about the parameter of the proportion of chips in the bag that is red. In the coin tossing there is no epistemic uncertainty and no unknown parameter to learn about.

2.4 Scoring Rules

A way to measure these uncertainties is by using the entropy (Section 2.5). The entropy is based on the concept of scoring rules. Since a major purpose of statistical analysis is to make forecasts for the future and provide suitable measures of the uncertainty associated with them, the so called *scoring rules* can be used to evaluate the qualities of these probabilistic forecasts. The forecasts are compared to the true outcome such that the forecaster will receive a higher score (or reward) for an accurate prediction of a probability measure [9] and he is motivated to give significant statements.

Suppose we have a random variable X with a set of outcomes $\mathfrak{X}$ and family $\mathcal{P}$ of distributions over X. Our task is to choose a distribution $Q \in \mathcal{P}$ to represent our belief of what P (the true distribution of X) may be. After we learn that the true value of X equals a certain value $x \in \mathfrak{X}$, we will receive a score $S(Q,x)$ based on how accurate our prediction is. This score can be interpreted as a reward for choosing the right distribution or of course alternatively as a penalty for choosing the wrong distribution. The goal of the predictor should always be to maximize the expected score.

Definition 2.3 (Expected Score)**.** The expected score for believing a distribution is Q when in truth it is P is denoted by [15]:

$$S(Q:P) = \mathbb{E}_P\, S(Q,x) = \begin{cases} \sum_{x\in\mathfrak{X}} S(Q,x)P(x) & \text{for discrete } X; \quad (6) \\ \int_{\mathfrak{X}} S(Q,x)P(x)dx & \text{for continuous } X. \quad (7) \end{cases}$$

We will equation (6) to assess the performance of Bayes Factors, since our set of outcomes consists of the hypotheses H_p and H_d, which are discrete variables.

Definition 2.4 (Scoring Rule)**.** A scoring rule is a real-valued function $S : \mathcal{P} \times \mathfrak{X} \to \mathbb{R}$. We say it is a *proper scoring rule* if truthfulness maximizes the expected score

$$S(P : P) \geq S(Q : P) \quad \forall\, P, Q \in \mathcal{P}. \tag{8}$$

We say the scoring rule is *strictly proper* when truthfulness uniquely maximizes the expected score

$$S(P : P) > S(Q : P) \quad \forall\, P, Q \in \mathcal{P}. \tag{9}$$

Clearly we acquire better forecasts when we work with (strictly) proper scoring rules, since we want the scoring rule to encourage honesty. If you would get a higher reward for a wrong prediction, this would not encourage you to try to find a good prediction.

Example 2. Examples of strictly proper scoring rules are

- the Quadratic Score:
$$S(P,x) = 2P(x) - \sum_{x \in \mathfrak{X}} P(x)^2.$$
- the Logarithmic Score [16]:
$$S(P,x) = \log\, P(x) \in (-\infty, 0].$$

  Let's check that this is strictly proper by looking at the difference of the expected score:

$$S(P : P) - S(Q : P) = \sum_{x \in \mathfrak{X}} P(x)(\log P(x) - \log Q(x)) = \sum_{x \in \mathfrak{X}} P(x) \log \left(\frac{P(x)}{Q(x)} \right).$$

  This is known as the Kullback-Leibler divergence which is proven with Gibb's inequality to always be non-negative and only equals zero when $Q = P$ almost everywhere.

Some early applications of the mathematical theory of scoring rules were to meteorology [4] and subjective Bayesianism [16, 11]. Later, applications in statistical inference are described by Dawid and Musio [10]. They state that at a theoretical level, any proper scoring rule can

be used as a foundational basis for the theory of subjective probability. At an applied level a proper scoring can be used to compare and improve probability forecasts, and, in a parametric setting, as an alternative tool for inference. Gneiting and Raftery [15] describe a probabilistic model for weather forecasting and used this to show that if there are no parameters being estimated, then the likelihood ratio (as in equation (2)) is equal to the difference in logarithmic score. This adds to the wide variety of applications that scoring rules have in statistical analysis. One application is the entropy, which we will use to define a performance measure for our Bayes Factors.

2.5 Entropy

Entropy is a concept that is widely used in different fields of science. In mathematics it can be seen as a measure of uncertainty of a random variable and is defined as follows by Cover [8]:

Definition 2.5 (Entropy)**.** The entropy $H_{(p)}(X)$ of a discrete random variable X with probability mass function $p(x)$ and sample space $\mathfrak{X}$ is given by

$$H_{(p)}(X) := -\sum_{x\in\mathfrak{X}} p(x)\log_2(p(x)) \tag{10}$$

With equation (6) we see that the entropy is the negative of the expected score for the logarithmic scoring rule. The convention is used that $0\log 0 = 0$ to ensure that adding events of zero probability does not change the entropy. Note that a higher entropy means more uncertainty, so a lower entropy indicates that we have more information about the random variable. This is a very widely used scoring rule in information theory and Bayesian inference. A property of the logarithmic score that is great for forensic applications, is that a likelihood ratio of 0 yields an infinite penalty, when H_p is in fact true. Other motivations to use the logarithmic scoring rule can be found in [5].

We can adapt this to a forensic context in the same way as Ramos [33] by taking the sample space $H = \{H_p, H_d\}$ and adding background information I. Once the evidence is gathered and measurements are performed, we can determine the likelihood ratio. This value may or may not reduce the uncertainty about the hypotheses variables and can consequently also increase or reduce the entropy. This can be seen through the Cross-Entropy, which we define here for discrete evidence E:

Definition 2.6 (Cross Entropy)**.**

$$H_{Q\|P}(H \mid E) := -\sum_{i\in\{p,d\}} Q(H_i) \sum_{e\in E} q(e \mid H_i) \log_2 P(H_i \mid e), \tag{11}$$

where P denotes the probabilities obtained with the forensic model and Q denotes the reference (or reality) probabilities.

The cross entropy can be decomposed into the sum of the posterior entropy of the reference distribution (Q) and the Kullback-Leibler (KL) divergence [33]. The posterior entropy of the reference distribution measures the uncertainty about the hypotheses if distribution Q is used for computing posterior probabilities. The Kullback-Leibler divergence is a measure of distance between two probability distributions P and Q [8]. Since it is always non-negative, its value can only increases the entropy and therefore defines a measure of inefficiency (or information loss) of assuming a distribution P when the real distribution is Q.
When the cross entropy cannot be calculated directly (when due to unknown parameters or lack of knowledge of distributions we cannot compute $q(e \mid H_i)$), we may use an empirical approximation yielding the Empirical Cross Entropy (ECE).

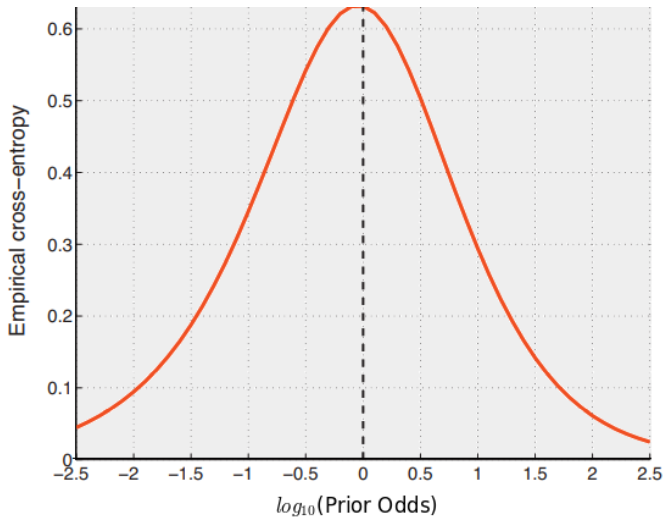


Figure 2: An example of the Empirical Cross Entropy function of a system of Bayes Factors plotted against the $\log_{10}$ of the prior odds.

Definition 2.7 (Empirical Cross Entropy)**.**

$$ECE := \hat{H}_{Q\|P} = -\sum_{i\in\{p,d\}} \frac{P(H_i)}{N_i} \sum_{e\in E} \log_2 P(H_i \mid e), \tag{12}$$

where N_i denotes the total number of observations where H_i is true.

We can rewrite the expression for the posterior probabilities $P(H_i \mid e)$, by using $P(H_p \mid e) = 1 - P(H_d \mid e)$ and equation (2):

$$P(H_p \mid E) = \frac{BF \times \frac{P(H_p)}{P(H_d)}}{1 + BF \times \frac{P(H_p)}{P(H_d)}}, \tag{13}$$

$$P(H_d \mid E) = \frac{1}{1 + BF \times \frac{P(H_p)}{P(H_d)}} = 1 - P(H_p \mid E). \tag{14}$$

So plugging these expressions into equation (12), we see that the empirical cross entropy is prior-dependent. Hence, it is not possible in general for a forensic scientist to compute one certain value of the empirical cross entropy for a given particular case, because the prior probabilities are not always given to them. However, we can compute and present it for a range of prior probabilities, without assuming a certain value for $P(H_p)$. An example of what this would look like can be found in figure 2. Rewriting equation (12) and adding background information I to place it in forensic context, gives us the final expression we will use for the Empirical Cross Entropy:

$$\begin{aligned} ECE = &\frac{P(H_p \mid I)}{N_p} \sum_{i:H_p \text{ is true}} \log_2 \left(1 + \frac{1}{BF_i \times \frac{P(H_p|I)}{P(H_d|I)}}\right) + \\ &\frac{P(H_d \mid I)}{N_d} \sum_{j:H_d \text{ is true}} \log_2 \left(1 + BF_j \times \frac{P(H_p \mid I)}{P(H_d \mid I)}\right). \end{aligned} \tag{15}$$

The ECE is a strictly proper scoring rule. For Bayes Factor systems that have no misleading evidence, the Bayes Factors will equal 0 when H_d is true and go to infinity when H_p is true. This results in a value of 0 for the ECE. Hence, the ECE penalizes misleading Bayes Factors with a penalty that is proportionate to the level of support for the wrong hypothesis. A lower ECE therefore means less uncertainty and a 'better' Bayes Factor system.

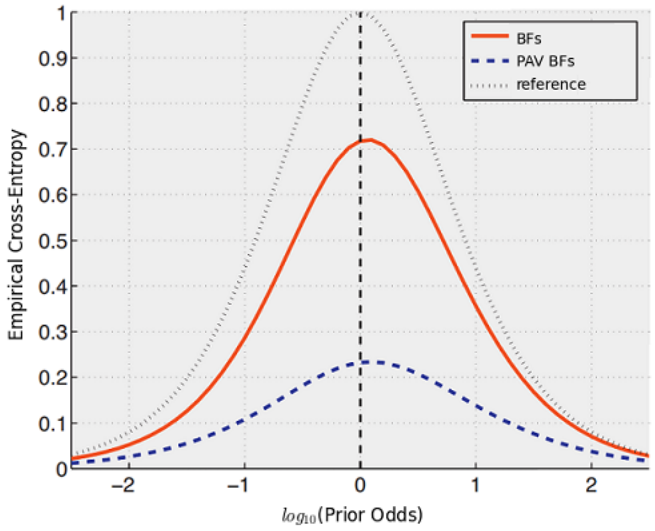


Figure 3: An example of an ECE plot, showing the PAV transformed likelihood ratios and a reference system where all likelihood ratios are equal to 1.

2.6 Pool Adjacent Violators Algorithm

When defining a model, it is important to check if it is an accurate well-calibrated model. We define the *accuracy* of a Bayes Factor system as the extent in which a Bayes Factor supports the true hypothesis. This accuracy can be decomposed into a *calibration* term and a *discriminating power* term [5, 34]. A set of likelihood ratios is well calibrated if *the likelihood ratio of the likelihood ratio is the likelihood ratio* [37] or when the likelihood ratio is exactly as is determined by the data. The discriminating power is the performance property representing the capability of the model to distinguish amongst forensic comparisons where the different hypotheses are true [26]. Calibration can be visualized by applying the *Pool Adjacent Violators Algorithm* (PAV) to the set of Bayes Factors. The algorithm is briefly described in [5] as follows:

1. Sort all Bayes Factors from lowest to highest;
2. Assign a posterior probability of one to all Bayes Factors for which H_p is true and of zero to all Bayes Factors for which H_d is true. This sequence will be used as input for the PAV algorithm;
3. The PAV algorithm pools adjacent Bayes Factors for which monotonicity is violated and then replaces all values in the pooled region by the mean over that region;

4. Reverse the second step by recovering the log-Bayes-Factors from the posterior probability following Bayes' rule, where the prior odds are the proportion of ones in the one-zero sequence of step 2;

5. Undo the sorting, such that the log-Bayes-Factor value correspond to the original input.

The output values of this algorithm are our *PAV Bayes Factors*. These are optimally calibrated, which means that all of the loss in accuracy that is measured by the ECE is due to loss in discriminating power. We can measure the calibration of our Bayes Factors by plotting their ECE and the ECE of the PAV Bayes Factors in one plot and seeing if they are close together. An example of an ECE plot can be found in figure 3. The accuracy is shown by the solid curve: the lower the curve, the more accurate your Bayes Factors are. The discriminating power is shown by the dashed curve: the lower the curve, the better the discriminating power. The calibration is represented by the difference between the solid and the dashed curve: if they are close, the system is well calibrated.

3 Model for Categorical Evidence

For categorical evidence the number of events of the random variable is finite. When a trace and reference sample are compared for this evidence type, they can *match* (the events are identical) or not (they are different). Examples of categorical evidence are blood type and DNA-profiles. In order to evaluate the evidential value of a match of this evidence with another DNA profile (for example that of a suspect), we need to weigh how probable the profiles are under the hypothesis that the suspect left the evidence (H_p) against how probable the profiles are under the hypothesis that someone else left the evidence (H_d). This match will depend on finding or not finding a certain characteristic of the profile, which we will call γ. Assuming that this characteristic is always detected correctly, this probability under H_p is 1 and under H_d it is equal to the proportion of people in the relevant population that have the same characteristic profile. We assume that we know that DNA profiles with this characteristic occur with frequency f_γ in our background population. When $f_\gamma \approx 0$, we speak of a *rare type match problem* [6].

3.1 Model Specification

We will use the notation that is provided in Sections 1.1 and 1.2 to specify the model used to calculate the Bayes Factors. We will adopt the model of van Dorp et. al. [36], where they check for a match between the DNA profiles in the available discrete evidence set E. In our case this match will denote finding the characteristic γ in the DNA profile.

3.1.1 Common Source

In the common source scenario, the prosecution hypothesis implies that e_{u_1} and e_{u_2} originate from the same source and are therefore the same DNA profile with probability 1. This is because there is no within-source variation, whereas under the defence model e_{u_1} and e_{u_2} are independent. We let e_{u_1} be the reference sample with which we compare the rest of the samples and check for the characteristic γ. We assume that the background population evidence e_a consists of n_a different sources. Let y_{a_i} denote the random variable corresponding to (not) finding the characteristic γ in the DNA profile of the i-th source in the background material e_a, for $i = 1, 2, \ldots, n_a$. In addition, y_{u_1} denotes the random variable corresponding to (not) finding the characteristic γ in the DNA profile of the first unknown source evidence e_{u_1} and y_{u_2} denotes the random variable corresponding to (not) finding the characteristic γ in the DNA profile of the unknown source evidence e_{u_2}. We have that all y_{a_i} are independently

identically distributed for each $i = 1,2,\ldots,n_a$, so

$$y_{a_i} \overset{iid}{\sim} G(\cdot \mid \theta), \quad \text{for} \quad i = 1,2,\ldots,n_a, \tag{16}$$

where G denotes the probability distribution of the matching of sources indexed by parameters θ. We know that under H_p, e_{u_1} and e_{u_2} are generated by the same source (hence, the DNA profiles) came from the same person, so we have

$$y_{u_1} \sim G(\cdot \mid \theta), \tag{17}$$

and $\mathbb{P}(y_{u_1} = y_{u_2}) = 1$. Under H_d, we know that e_{u_1} and e_{u_2} are generated independently, hence

$$y_{u_1} \sim G(\cdot \mid \theta) \quad \text{and} \quad y_{u_2} \sim G(\cdot \mid \theta) \quad \text{independently.} \tag{18}$$

3.1.2 Specific Source

For the specific source problem, the prosecution hypothesis implies that e_u and e_s originate from the same (specific) source and are therefore the same DNA profile with probability 1. Again, this is because there is no within-source variation, whereas under the defence model e_u and e_s are independent. Because e_s is generated from a known specific source, there is no randomness in its discrete evidence model and e_s remains fixed. Analogously to the common source scenario, let y_{a_i} denote the random variable corresponding to the matching of the evidence from the i-th source in the background material e_a, for $i = 1,2,\ldots,n_a$ and let y_u denote the random variable corresponding to the matching of the unknown source evidence e_u. We have that all y_{a_i} are independently identically distributed for each $i = 1,2,\ldots,n_a$, so

$$y_{a_i} \overset{iid}{\sim} G(\cdot \mid \theta), \quad \text{for} \quad i = 1,2,\ldots,n_a, \tag{16}$$

where G denotes the probability distribution of the matching of sources (other than the specific source) indexed by parameters θ. Under H_p, we have that $y_u = y_s$ with probability 1, because the unknown evidence is assumed to originate from the specific source. Finally under H_d, we have

$$y_u \sim G(\cdot \mid \theta). \tag{19}$$

3.1.3 Bayes Factor Calculation

We can view the random variables y_{a_i} corresponding to the background material as a result of a sequence of n_a Bernoulli trials with probability of success f_γ, where success corresponds to observing the characteristic γ and failure corresponds to not observing characteristic γ. Therefore, it is a logical next step to let distribution G in equations (16) - (19) be a Bernoulli distribution with parameter $\theta = f_\gamma$. This is equivalent to having the total number of matches in the background material being represented by a binomial model with parameters n_a and f_γ. We define a random variable X to be the number of matches in the background material. Then

$$X \sim \text{Bin}(n_a, f_\gamma). \tag{20}$$

We let $s_a = \sum_{i=1}^{n_a} y_{a_i}$ denote the total of *observed* matches in the background population. Note that the rare type match problem is defined by getting $s_a = 0$, because the frequency f_γ may be so low that the characteristic only appears in 1 out of 10 populations or even less [6]. Due to its conjugacy[2] with the binomial distribution, it's a convenient choice to let the prior distribution of f_γ be the Beta distribution [3]. This results in the following prior and posterior distributions for f_γ, where the posterior is known because of the conjugacy:

$$f_\gamma \sim \text{Beta}(\alpha, \beta), \quad \alpha, \beta > 0. \tag{21}$$

$$f_\gamma \mid e_a \sim \text{Beta}(\alpha + s_a,\ \beta + n_a - s_a). \tag{22}$$

Van Dorp et al. calculated the common source and specific source Bayes Factors according to this model [36]. We have added this calculation in appendix A for completeness. The Bayes Factors are given by

$$BF_{CS}(E) = \frac{\alpha + \beta + n_a + 1}{\alpha + s_a + 1}; \tag{A4}$$

$$BF_{SS}(E) = \frac{\alpha + \beta + n_a}{\alpha + s_a}. \tag{A6}$$

[2] If the posterior distribution of a variable is in the same probability distribution family as the prior probability distribution, the prior and posterior are called conjugate distributions, and the prior is called a conjugate prior for the likelihood function. A conjugate prior gives a closed-form expression for the posterior, which eliminates the necessity of numerical integration. It is therefore often algebraically convenient to work with conjugate priors.

3.2 Bayes Factor Comparison

Since we have discrete evidence, by definition we can use the cross entropy to compare the performance of both Bayes Factor systems in updating the prior odds of the specific source hypotheses. Recall from definition 2.6 that the cross entropy is given by

$$\begin{aligned} H_{Q\|P}(H \mid E) = &- Q(H_p) \sum_{e \in E} q(e \mid H_p) \log_2 P(H_p \mid e) \\ &- Q(H_d) \sum_{e \in E} q(e \mid H_d) \log_2 P(H_d \mid e). \end{aligned}$$

This equation can be evaluated for a range of prior probabilities, without choosing a specific value for $Q(H_p)$. As we let the prior probability be treated as a parameter, we have that the entropies are equal for different distributions ($H_{(p)}(X) = H_{(q)}(X)$ from definition 2.5) since they both take values in the interval $[0,1]$ and we consider all possible values. This translates into the definition of the cross entropy, which means we can choose $Q(H_i) = P(H_i)$ for $i \in \{p,d\}$ as the prior probability of the hypotheses [32]. We add background information[3] I and rewrite the above expression using equations (13) and (14) to get:

$$\begin{aligned} H_{Q\|P}(H \mid E) = &- P(H_p \mid I) \sum_{e \in E} q(e \mid H_p, I) \log_2 \left(\frac{BF \times \frac{P(H_p|I)}{P(H_d|I)}}{1 + BF \times \frac{P(H_p|I)}{P(H_d|I)}} \right) \\ &- P(H_d \mid I) \sum_{e \in E} q(e \mid H_d, I) \log_2 \left(\frac{1}{1 + BF \times \frac{P(H_p|I)}{P(H_d|I)}} \right). \end{aligned} \tag{23}$$

Note that the Bayes Factors are only dependent on parameters α, β, n_a and s_a and not specifically on the DNA profiles that have been compared.

Before we can evaluate this expression, we need to determine the probabilities $q(e \mid H_p, I)$ and $q(e \mid H_d, I)$, which denote our reference probabilities. We will set this in a specific source scenario, since we want to assess the value (influence) that the different Bayes Factors have on it. In this model, the evidence over which we sum are the number of matches in the background population (s_a), the DNA profile of the trace sample (y_u) and the DNA profile of the reference sample (y_s). So to calculate the cross entropy, we have to consider all possibilities of these values. From the Bernoulli distribution, we set the random variables y_u and y_s equal to 1 if

[3] We do not model this background information I, but here it denotes our belief about the hypotheses (due to police information etc.) and the reasoning that is given below for the values of the probability q.

the DNA profile has the characteristic and equal to 0 if the DNA profile does not contain γ. We will choose the convention that $q(y_s = 1) = 1$, so that the specific source always has the characteristic γ with probability 1. This is equivalent to what has been done in [30] and [36], but it is not immediately clear to the reader why it has been chosen. We do not completely agree with this choice. The only reason to set the probability of having γ equal to 1, is when the trace found on the crime scene also has the characteristic γ and we found the specific source based on this. But in our model, the DNA profile of the trace can also not contain γ. If we do not choose to follow the literature, we would have $q(y_s = 1) = f_\gamma$ and this would add exactly one factor f_γ in front of expression (23) for both scenarios, which means that it would not influence our analysis of the cross entropy values at all. And since the specific source is a fixed source (not random), we choose to follow the literature and set $q(y_s = 1) = 1$. Furthermore, we let y_s be the reference profile and y_u the trace profile.

Firstly, we consider $q(e \mid H_p, I)$. Note that under H_p we do not consider the specific source to be a part of the background population, which means that the specific source DNA profile and s_a are independent. But the profile of the trace and the reference profile are considered dependent given H_p, since under H_p we assume that they come from the same source (or equivalently, they are donated by the same person). So then the event where the DNA profile of the trace does not contain γ occurs with probability 0.

$$\begin{aligned}
\sum_{e\in E} q(e \mid H_p, I) &= \sum_{s_a} q(s_a, \text{trace and suspect have } \gamma \mid H_p) + q(s_a, \text{suspect has } \gamma \text{ and trace does not} \mid H_p) \\
&= \sum_{s_a=0}^{n_a} q(X = s_a, y_s = 1, y_u = 1 \mid H_p,\, I) + q(X = s_a, y_s = 1, y_u = 0 \mid H_p,\, I) \\
&= \sum_{s_a=0}^{n_a} q(X = s_a)\; q(y_s = 1, y_u = 1 \mid H_p,\, I) + 0 \\
&\overset{(*)}{=} \sum_{s_a=0}^{n_a} q(X = s_a)\; q(y_u = 1 \mid y_s = 1, H_p)\; q(y_s = 1) \\
&\overset{(**)}{=} \sum_{s_a=0}^{n_a} \binom{n_a}{s_a} f_\gamma^{s_a} (1 - f_\gamma)^{(n_a - s_a)} \times 1 \times 1,
\end{aligned} \tag{24}$$

where we applied the rule of conditional probability to get (*) and the distribution in equation (20) and the assumption that $y_s = 1$ with probability 1 to get (**).

Secondly, we consider $q(e \mid H_d, I)$. Under H_d, the reference profile, the trace profile and s_a are all independent of each other. We know that measuring characteristic γ is a Bernoulli trial with parameter f_γ, hence $q(y_u = 1) = f_\gamma$ and $q(y_u = 0) = 1 - f_\gamma$. To calculate the cross entropy we need to consider both values for y_u. However, a non-match $(y_u = 0)$ results in a cross entropy term[4] of 0, which does not influence our analysis and therefore, we do not take it into account when we calculate the reference probabilities. Recall that by assumption, the event where the suspect does not have γ occurs with probability 0. So then the probability $q(e \mid H_d, I)$ can be written as follows:

$$\begin{aligned}
q(e \mid H_d, I) &= \sum_{s_a} q(s_a, \text{trace and suspect have } \gamma \mid H_d,\, I) \\
&= \sum_{s_a=0}^{n_a} q(X = s_a, y_s = 1, y_u = 1 \mid H_d,\, I) \\
&\overset{(*)}{=} \sum_{s_a=0}^{n_a} q(X = s_a)\; q(y_s = 1)\; q(y_u = 1) \\
&\overset{(**)}{=} \sum_{s_a=0}^{n_a} \binom{n_a}{s_a} f_\gamma^{s_a} (1 - f_\gamma)^{(n_a - s_a)} \times 1 \times f_\gamma \\
&= \sum_{s_a=0}^{n_a} \binom{n_a}{s_a} f_\gamma^{(s_a+1)} (1 - f_\gamma)^{(n_a - s_a)}, \qquad (25)
\end{aligned}$$

where we used the distribution of X from (20) to get (**) and the fact that y_u and y_s are independent under H_d to get (*). We now have an expression for each term in equation (23) and we use algorithm 1 to compute the cross entropy numerically and plot its values for a range of prior odds. The calculations in this algorithm are easy, so for each set of values it only took a couple seconds to calculate the cross entropy. In this algorithm, we used $O(H_p) = \frac{\mathbb{P}(H_p)}{\mathbb{P}(H_d)}$.

We provide a short explanation for the values that we used in the algorithm:

- Prior probability values 0 and 1 for $\mathbb{P}(H_p)$ cannot be included, since this either gives prior odds of 0, which is not in the domain of the logarithm, or it gives $\mathbb{P}(H_d) = 0$ and of course we cannot divide by 0. We chose a domain of $[-6, 6]$ for the log odds since this showed a slightly flattened curve for the cross entropy on both ends of the domain.

[4] A non-match shows us that H_d must be true, since the two profiles cannot come from the same person. This gives us a Bayes Factor of 0 or $q(H_p \mid E) = 0$, which results in a logarithm of minus infinity. As with the definition of Entropy, the convention is used that $0 \log 0 = 0$. Additionally we have $q(H_d \mid E) = 1$ and $\log_{10}(q(H_d \mid E)) = 0$.

Algorithm 1: Algorithm for calculating the specific source and common source cross entropy

Input: $n_a = 1000$;

Output: Specific source cross entropy ($-CE_{SS}$) and common source cross entropy ($-CE_{CS}$) ;

for *Scenario* $S \in \{$*Common Source, Specific Source*$\}$ **do**

  for $\log_{10}(O(H_p)) \in \{-6, -5, 99, -5.98, \ldots, 5.98, 5.99, 6\}$ **do**

    $P(H_p) \leftarrow (10^{\log_{10}(O(H_p))})/((10^{\log_{10}(O(H_p))} + 1)$;

    $P(H_d) \leftarrow 1 - P(H_p)$;

    for $f_\gamma \in \{0.0001, 0.2, 0.4, 0.5\}$ **do**

      $q(e \mid H_p) \leftarrow \binom{n_a}{s_a} f_\gamma^{s_a} (1 - f_\gamma)^{(n_a - s_a)}$;

      $q(e \mid H_d) \leftarrow \binom{n_a}{s_a} f_\gamma^{(s_a+1)} (1 - f_\gamma)^{(n_a - s_a)}$;

      for $(\alpha, \beta) \in \{(0.001, 0.001), (0.01, 0.01), (0.01, 0.02), (0.5, 0.5), (2, 8), (8, 2), (5, 15), (10, 10), (100, 0.01), (1, 10.000)\}$ **do**

        if $S =$ *Common Source* **then**

          $BF_S \leftarrow (\alpha + \beta + n_a + 1)/(\alpha + s_a + 1)$;

        else

          $BF_S \leftarrow (\alpha + \beta + n_a)/(\alpha + s_a)$;

        end

        $P(H_p \mid E) \leftarrow (BF_S \times O(H_p))/(1 + BF_S \times O(H_p))$;

        $P(H_d \mid E) \leftarrow 1 - P(H_p \mid E)$;

        $CE_S \leftarrow \sum_{s_a=0}^{n_a} \Big(P(H_p) \times q(e \mid H_p) \times \log_2(P(H_p \mid E) + P(H_d) \times q(e \mid H_d) \times \log_2(P(H_d \mid E)\Big)$

        return $-CE_S$

      end

    end

  end

end

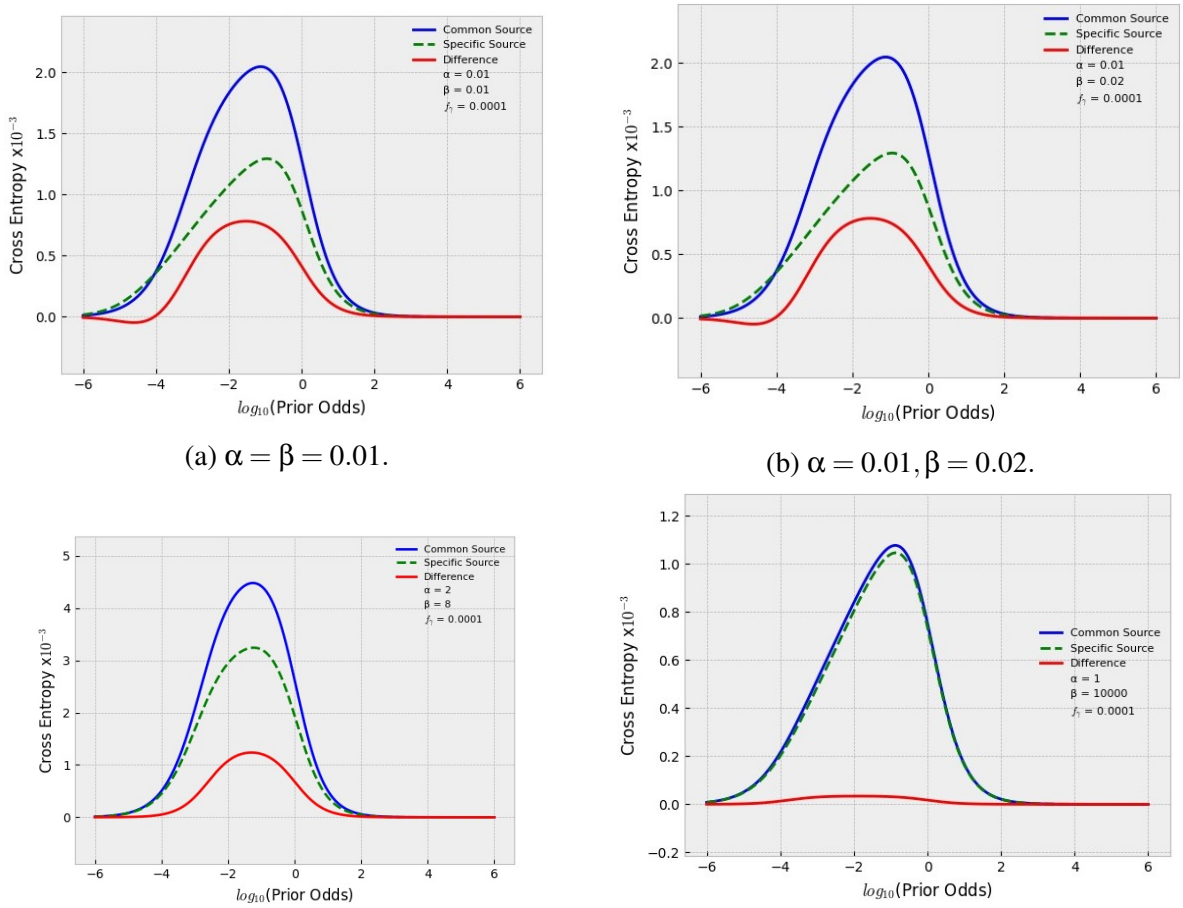
- We use different values for f_γ to compare the different scenarios. The one we are most interested in is the low value 0.0001, because it represents a rare characteristic and therefore has more evidential value. If a characteristic is chosen that a large portion of the population has, then this does not give us a lot of information about whether a person is guilty or not.

- First we choose α and β close to 0, to have only a small impact on the Bayes Factors compared to the sample size and to let the value of s_a provide a big influence. Jeffrey's prior [19] of $\alpha = \beta = 0.5$ provides an uninformative prior. Values of α and β are varied more to use different expected values of f_γ and to see the influence of the prior parameters on our results.

- Lastly, we sum over all possible values of s_a, since this is how our evidence and cross entropy are defined.

Recall that the cross entropy measures the uncertainty and inefficiency of using a distribution P when the true distribution is Q. We want to assess the performance of the Bayes Factors in updating the prior probabilities of the specific source hypotheses, as noted in research question Q2. Therefore, we define the *better Bayes Factor-system* to be the one with the lowest cross entropy. In figure 4 the common source and specific source cross entropies are plotted for $f_\gamma = 0.0001$ and different values of prior parameters α and β. Since we have a population of size 1000, this defines a rare type match problem, where $s_a = 0$. We immediately see that the common source Bayes Factor has a higher entropy in general for all parameters.
The only exceptions are figures 4a and 4b on the left side of the x-axis. This indicates that for lower prior odds, where H_d has a higher probability, the common source Bayes Factor-system performs better than the specific source Bayes Factor-system. We think the reason for this might be the following: the prior probability for H_d is very large in this area, but if we find the characteristic γ in the DNA profile of the trace, that provides strong evidential value for H_p. This means the scoring rule (or entropy) will give a penalty to the Bayes Factors since the prior probability for H_d was so high. A high prior probability for H_d means that the H_d cross entropy term is a much larger factor than the H_p cross entropy term when calculating the cross entropy and the penalty increases when the Bayes Factor is larger. For the rare type match problem with small values of α and β and a large value for $P(H_d)$, the common source Bayes Factor is smaller than the specific source Bayes Factor, which results in a higher penalty for the specific source Bayes Factors. We added a logarithmic plot of these left ends in figure 5 to double check if the specific source entropy dips under the common source cross

entropy on the far left end, but this is not the case in our domain. And we see that for larger values of α, the common source entropy is larger for all prior odds. Large values of β cause the specific source entropy to increase significantly as can be seen in figure 4d. A large ratio α/β results into convergence of the difference to 0. Additional plots are added in appendix B for $f_\gamma = 0.0001$. They show that for higher values of α and β the entropy gets much larger relatively to each other and the two lines seem to get closer as well. However, the difference between the two entropies seems to have the same peak and spread when we compare the difference plots.



(a) $\alpha = \beta = 0.01$.

(b) $\alpha = 0.01, \beta = 0.02$.

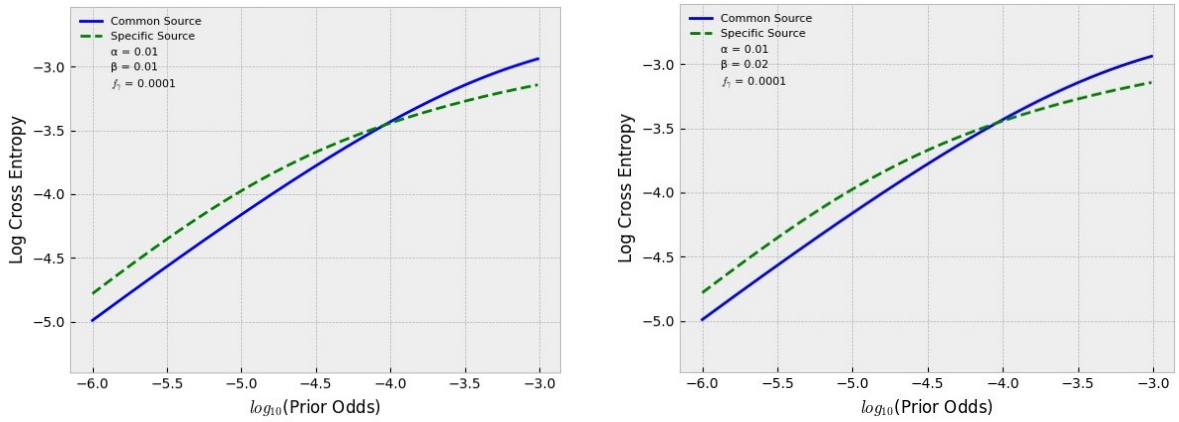
(c) $\alpha = 2$, $\beta = 8$.

(d) $\alpha = 1, \beta = 10000$.

Figure 4: Common source and specific source cross entropy (and their difference) plotted for a range of $\log_{10}$(Prior Odds) with frequency $f_\gamma = 0.0001$ and different prior parameter values.

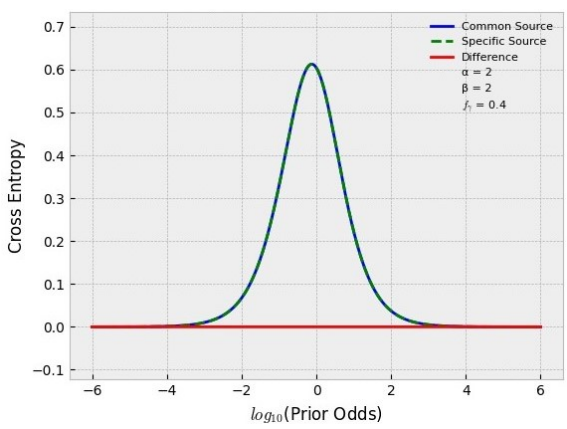
When we use higher frequencies of the characteristic, we see that the difference between the common source and specific source entropies is always really close to 0. We have added two examples in figure 6. For all values of α and β that we tried, all plots looked like this: the values of the y-axis differ, but the difference is always close to 0.

We conclude that it is possible to use the common source Bayes Factor to update the specific source hypotheses. The entropies are close together for a low frequency and high values of α and β. The difference in Bayes Factors for $s_a = 0$ is inflated when we use small values of α and β and therefore the entropies are also very different in value. Furthermore, our study shows that in most scenarios (that we looked at) the specific source Bayes Factor-system performs better than the common source Bayes Factor-system. This may be because there is only one level of uncertainty in this particular model.

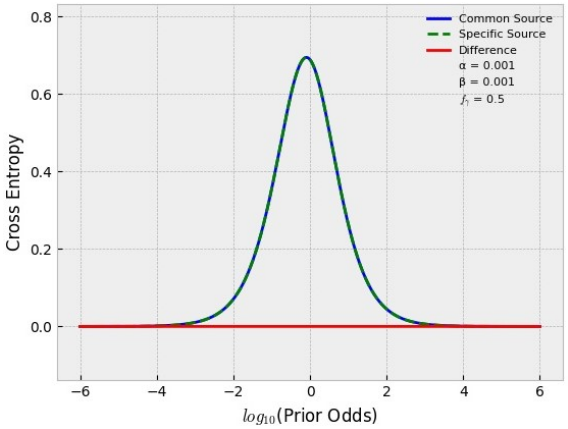


(a) $\alpha = \beta = 0.01$. (b) $\alpha = 0.01, \beta = 0.02$.

Figure 5: Left end plots of the common source and specific source log-cross entropies with frequency $f_\gamma = 0.0001$.



(a) $\alpha = \beta = 2$.



(b) $\alpha = \beta = 0.001$.

Figure 6: Common source and specific source entropy (and their difference) plotted for a range of $\log_{10}$(Prior Odds) with frequency $f_\gamma = 0.4$ (a) and $f_\gamma = 0.5$ (b).

4 Two-level Model for Continuous Evidence

In this chapter we will define a model for using continuous evidence to compare the Bayes Factor systems. After the specification of the model we will approximate the Bayes Factors using Markov Chain Monte Carlo sampling and then compare the results of the common source system to the results of the specific source system. Some sensitivity analysis has been made to clarify the model and check our assumptions and parameter choices. These will be described in the last section of this chapter.

4.1 Model Specification

We will follow the model of Ommen, Saunders and Neumann [31] who used evidence consisting of measurements on the elemental compositions of glass fragments. These fragments originate from a window which is considered to be the (unknown) source of the glass fragments. We will use this as an exemplary scenario for our two-level model.
We will define a general normal-normal model for the evidence sets and impose specific values on the parameters when we do our simulation study. The big difference between our model and the model in [31] is that we choose the variances to be unknown, while Ommen et al. use an estimation given by Aitken and Lucy [1] and therefore have a specific value. This means we have more unknown parameters and therefore more epistemic uncertainty. Therefore, it is even more important for us to take the background population into account as evidence with this model than the previous, because we need to estimate the posterior distributions of these variances.
We will generate evidence again in a specific source scenario. This evidence will then be used to define the two different models. The biggest difference between the models is that for the common source model we assume that all windows from which the glass fragments originate, are part of the background population. This includes the reference window.

4.1.1 Specific Source

The specific source evidence set is given by $E = \{e_s, e_u, e_a\}$, where e_a represents the measurements performed on the alternative window population, e_s denotes the measurements performed on the glass fragments from the specific source and e_u represents the measurements performed on the glass fragments with unknown source.

Firstly, we consider the specific source evidence set e_s, which is constructed as follows: one random sample of glass fragments is taken from the specific source window and m_s measurements are performed on it to get the column vector[5]

$$e_s = \vec{x}_s = (x_{s1}, \ldots, x_{sm_s})^T.$$

These vector elements are distributed as

$$x_{sj} \overset{iid}{\sim} \mu_s + \varepsilon_s \quad \text{for} \quad j \in \{1, \ldots, m_s\} \quad \text{with} \quad \varepsilon_s \sim N(0, \sigma_s^2). \tag{26}$$

Here σ_s^2 represents the variance due to measurement errors, which is the same for each measurement, hence for each $j \in \{1, \ldots, m_s\}$ and μ_s denotes the mean of the considered specific source. We will impose a prior normal distribution on μ_s and a prior uniform distribution on the standard deviation σ_s. Since all the measurements are independent and identically distributed Gaussian variables, the vector $\vec{x_u}$ is multivariate normally distributed as follows:

$$\vec{x_s} \sim N_{m_s}(\vec{\mu_s}, \Sigma_s) \quad \text{with} \quad \vec{\mu_s} = (\mu_s, \ldots, \mu_s)^T. \tag{27}$$

All measurements are made independently from each other, which means the covariance between elements x_{si} and x_{sj} is zero for all $i \neq j$ and we have

$$\Sigma_s = \begin{pmatrix} \sigma_s^2 & 0 & \cdots & 0 \\ 0 & \sigma_s^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_s^2 \end{pmatrix} = \sigma_s^2 I_{m_s},$$

where I_{m_s} is the $m_s \times m_s$-identity matrix.

Secondly, we consider the background evidence set, e_a, which is constructed as follows: n_a independent sources are randomly sampled from the alternative window population, then one sample a_i is taken from each window $i \in \{1, \ldots, n_a\}$ on which m_a measurements are per-

[5] We use the arrow notation $\vec{x}$ to denote a vector.

formed. Each sample will be represented by a column vector, so our evidence set e_a becomes:

$$e_a = (\vec{x}_{a_1}, \ldots, \vec{x}_{a_{n_a}}) = \begin{pmatrix} x_{a_1 1} & x_{a_2 1} & \cdots & x_{a_{n_a} 1} \\ x_{a_1 2} & x_{a_2 2} & \cdots & x_{a_{n_a} 2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{a_1 m_a} & x_{a_2 m_a} & \cdots & x_{a_{n_a} m_a} \end{pmatrix}.$$

We let the vector elements be independent and identically distributed given the source a_i:

$$x_{a_i j} \overset{iid}{\sim} \mu_{a_i} + \varepsilon_a \quad \text{for} \quad i \in \{1, \ldots, n_a\} \quad \text{and} \quad j \in \{1, \ldots, m_a\} \quad \text{with} \quad \varepsilon_a \sim N(0, \sigma_a^2). \quad (28)$$

The variance σ_a^2 represents the measurement errors for the background glass fragments and μ_{a_i} is the unknown mean of the elemental composition of window a_i and ε_a is the unknown deviation of the mean which has the same distribution for each window and each measurement (hence, does not depend on i or j). We assume that μ_{a_i} and μ_{a_j} are independent for all $i \neq j$ and that μ_{a_i} and σ_a^2 are independent from each other (for all $i \in \{1, \ldots, n_a\}$). When we consider glass fragments as part of a larger population, we will add an extra level of distributions to the hierarchical model. Hence, we let the mean μ_{a_i} follow a Normal distribution:

$$\mu_{a_i} \overset{iid}{\sim} \mu_M + \varepsilon_M \quad \text{for} \quad i \in \{1, \ldots, n_a\} \quad \text{with} \quad \varepsilon_M \sim N(0, \sigma_M^2). \quad (29)$$

Here μ_M denotes the mean of the grand mean population and σ_M^2 denotes the between-source variation. We will impose a prior normal distribution on μ_M and a prior uniform distribution on σ_a and σ_M.

Lastly, we consider e_u which consists of one random sample taken from the trace material on which we have performed m_u measurements to get the column vector

$$e_u = \vec{x}_u = (x_{u1}, \ldots, x_{um_u})^T.$$

Under H_p, the unknown source glass fragments and glass fragments taken from the specific source originate from the same window and therefore they follow the same distribution. Hence, the distribution of each vector element x_{uj} given that H_p is true, is given by

$$x_{uj} \mid H_p \overset{iid}{\sim} \mu_s + \varepsilon_s \quad \text{for} \quad j \in \{1, \ldots, m_s\} \quad \text{with} \quad \varepsilon_s \sim N(0, \sigma_s^2). \quad (30)$$

Under H_d, the unknown source glass fragments originate from another window than the specific source window, which means that the unknown source will be a source a_u different from the specific source, but still part of the population of alternative sources. So given that H_d is true, $\vec{x_u}$ follows the same model as $\vec{x_{a_u}}$ and we get the same distribution

$$x_{uj} \mid H_d \overset{iid}{\sim} \mu_{a_u} + \varepsilon_a \quad \text{for} \quad j \in \{1,\ldots,m_u\} \quad \text{with} \quad \varepsilon_a \sim N(0,\sigma_a^2), \tag{31}$$

where μ_{a_u} follows the distribution given in equation (29). The unknown parameters of the specific source scenario are then given by $\theta_{ss} = \{\mu_s,\, \sigma_s,\, \mu_M,\, \sigma_M,\, \sigma_a\}$.

4.1.2 Common Source

We use the same evidence sets as in the specific source model, because the evidence is generated in the specific source scenario. In other words, the evidence is the same, but the models are different.
Note that the model for e_a is exactly the same as in the specific source model, so we have the following distributions for the vector elements of $e_a = (\vec{x}_{a_1},\ldots,\vec{x}_{a_{n_a}})$:

$$x_{a_i j} \overset{iid}{\sim} N(\mu_{a_i},\sigma_a^2), \quad \text{for} \quad i \in \{1,\ldots,n_a\} \quad \text{and} \quad j \in \{1,\ldots,m_a\}, \tag{28}$$

$$\mu_{a_i} \sim N(\mu_M,\sigma_M^2), \quad \text{for} \quad i \in \{1,\ldots,n_a\}. \tag{29}$$

We will impose a prior normal distribution on μ_M and a prior uniform distribution on standard deviations σ_M and σ_a. The remainder of the evidence sets are defined as:

$$e_s = \vec{x}_s = (x_{s1},\ldots,x_{sm_u})^T \qquad \text{and} \qquad e_u = \vec{x}_u = (x_{u1},\ldots,x_{um_u})^T.$$

The difference with the specific source model is that we consider both evidence sets to originate from an unknown source. We choose e_s to be the reference evidence to which we compare the trace e_u. In the specific source model our reference window was unique and measurements on the other windows did not contain any information about the specific source. In the common source model the reference window is not unique and measurements in e_a contain information about e_s. In the common source model, we assume that all windows, from which the glass fragment originate, including the reference window, are part of the background population. Since we chose e_s as the reference trace, we assume that $\vec{x_s}$ comes from a fixed (unknown) source. The mean of its distribution is μ_s as in the specific source model, however we now assume that the source is part of the background population and μ_s

is distributed in the same way as μ_{a_i} in equation (29). Under H_p the glass fragments e_s and e_u came from the same window and therefore follow the same distribution. Hence, we have

$$x_{sj} \overset{iid}{\sim} \mu_s + \varepsilon_a \quad \text{for} \quad j \in \{1,\ldots,m_u\} \quad \text{with} \quad \varepsilon_a \sim N(0,\sigma_a^2), \tag{32}$$

$$x_{uj} \mid H_p \overset{iid}{\sim} \mu_s + \varepsilon_a \quad \text{for} \quad j \in \{1,\ldots,m_u\} \quad \text{with} \quad \varepsilon_a \sim N(0,\sigma_a^2), \tag{33}$$

$$\mu_s \sim N(\mu_M,\sigma_M^2). \tag{34}$$

Under H_d the second sample of glass fragments (e_u) does not originate from the same window as the first sample (e_s), so we can say it originates from another window a_u, which is also part of the background population. This results into the following distributions:

$$x_{uj} \mid H_d \overset{iid}{\sim} \mu_{a_u} + \varepsilon_a \quad \text{for} \quad j \in \{1,\ldots,m_u\} \quad \text{with} \quad \varepsilon_a \sim N(0,\sigma_a^2), \tag{35}$$

$$\mu_{a_u} \sim N(\mu_M,\sigma_M^2). \tag{29}$$

The unknown hyperparameters of the common source scenario are then given by $\theta_{cs} = \{\mu_M,\sigma_M,\sigma_a\}$.

An overview of the two models can be found in table 1.

4.2 Bayes Factor Approximation

Our first attempt at calculating the Bayes Factors was to integrate out all uncertainty about the hyperparameters by using prior distributions on them. We tried a conjugate model, where we impose a Normal-Inverse-Gamma distribution on the pairs (μ_s,σ_s) and (μ_M,σ_M) and an Inverse-Gamma distribution on σ_a. This was inspired by the Normal-Inverse-Chi-squared distribution[6] used in [2]. To find a closed expression for the Bayes Factors we would have to solve an integral over each hyperparameter for H_p and also for H_d, which (for the specific source scenario) adds up to 10 integrals over three likelihoods and three prior distributions. It is infeasible to do this analytically, which is why we decided to use Markov Chain Monte Carlo methods and the Arithmetic Mean Estimate as described in section 2.2.

[6]The Inverse-Gamma and Inverse-Chi-squared distribution are closely related as follows: if a random variable $X \sim Inv-\chi^2(k,s^2)$ then equivalently $X \sim Inv-\Gamma\left(\frac{k}{2},\frac{ks^2}{2}\right)$.

Table 1: Two-Level model for continous evidence.

Evidence	Specific Source	Common Source
e_s	$x_{sj} \sim N(\mu_s, \sigma_s^2)$	$x_{sj} \sim N(\mu_s, \sigma_a^2)$ $\mu_s \sim N(\mu_M, \sigma_M^2)$
e_a	$x_{a_i j} \sim N(\mu_{a_i}, \sigma_a^2)$ $\mu_{a_i} \sim N(\mu_M, \sigma_M^2)$	$x_{a_i j} \sim N(\mu_{a_i}, \sigma_a^2)$ $\mu_{a_i} \sim N(\mu_M, \sigma_M^2)$
e_u	$x_{uj} \mid H_p \sim N(\mu_s, \sigma_s^2)$ $x_{uj} \mid H_d \sim N(\mu_{a_u}, \sigma_a^2)$ $\mu_{a_u} \sim N(\mu_M, \sigma_M^2)$	$x_{uj} \mid H_p \sim N(\mu_s, \sigma_a^2)$ $\mu_s \sim N(\mu_M, \sigma_M^2)$ $x_{uj} \mid H_d \sim N(\mu_{a_u}, \sigma_a^2)$ $\mu_{a_u} \sim N(\mu_M, \sigma_M^2)$

4.2.1 Underlying Truth

We use *PyMC3* in Python [35] to approximate the Bayes Factors[7]. Before we can apply Monte Carlo sampling, we need to specify the underlying truth and generate evidence with these *true* values of the parameters. This evidence will be used as the input for the PyMC3 sampling. Similar to the reference probabilities q in section 3.2, generated evidence is represented by a specific source scenario. Since all location parameters can be expressed by their difference to μ_M, it is convenient to set $\mu_M = 0$. Secondly, we take $\sigma_M = 10$ and $\sigma_a = \sigma_s = 1$, because we want the measurements on the traces to be significantly closer to each other than the measurements on the grand mean population. In other words, the grand mean population must have a wider distribution than the trace distribution in order to get Bayes Factors with significant values[8]. We choose σ_a and σ_s to be equal, because they represent measurement errors, which is a property of the measuring machine and not of the windows. The variance σ_M^2 does not denote a measurement error but the variance of the grand mean population. We want to consider a specific source that has common characteristics as well as a specific source that has rare characteristics and then compare the specific source Bayes Factor to the common source Bayes Factor in both scenarios. This is also done by Neumann and Ausdemore in [27]. The common characteristics will be represented by choosing $\mu_s = \mu_M$, because the density of the normal distribution is the highest in the interval around μ_M. Therefore, choosing μ_s equal to μ_M means that these characteristics have a high probability of occurring more in the population. The rare characteristics are represented by setting $\mu_s = \mu_M + 2 \times \sigma_M$, since this translates into choosing a specific source in the tail of the normal distribution of the background population, of which the characteristics occur with low probability. We will only describe the full process for calculating the Bayes Factors for the scenario where $\mu_s = 0$, because the process for $\mu_s = 20$ is exactly the same.
Let us denote $\vec{x_{up}}$ as the true value of $\vec{x_u}$ when H_p is true and $\vec{x_{ud}}$ as the true value of $\vec{x_u}$ when H_d is true. These vectors are generated by sampling from normal distributions of which the parameters are equal to the true values.

[7] We used Python version 3.9 and PyMC3 version 3.11.3.
[8] With significant values we mean Bayes Factors with 'higher' orders, hence $\log_{10}(BFs) \gg 0$

So for $\mu_s = 0$, we get:

$$\mu_{a_u} \sim N(0, 10),$$

$$\vec{x_{up}} \sim N_{m_u}(\vec{0}, I_{m_u}) = N_{m_u}\left(\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}\right),$$

$$\vec{x_{ud}} \sim N_{m_u}(\vec{\mu_{a_u}}, I_{m_u}) = N_{m_u}\left(\begin{pmatrix} \mu_{a_u} \\ \mu_{a_u} \\ \vdots \\ \mu_{a_u} \end{pmatrix}, \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}\right)$$

for the true distribution of $\vec{x_{up}}$ and $\vec{x_{ud}}$. The true values of $\vec{x_{a_i}}$ and $\vec{x_s}$ are generated in the same way and are the same for both hypotheses.

4.2.2 Markov Chain Monte Carlo

For our simulation study, we set the number of measurements equal to 10, hence $m_a = m_s = m_u = 10$, and the number of sources in the background population (n_a) equal to 100. We chose this because we want the number of measurements on the background population (in this case 1000) to be large compared to the trace/reference measurements to get a lot of information about the background population. However, due to lack of time we did not have the opportunity to calculate 2000 Bayes Factors per model based on evidence consisting of 100 or 1000 measurements for example. Before we can start the MCMC sampling, we need to impose prior distributions on our hyperparameters. We choose the same prior distribution in both scenarios for each parameter that is in θ_{CS} as well as in θ_{SS}. Note that all distributions in the previous sections are one-dimensional normal distributions for the vector elements of measurement vectors $\vec{x_s}$, $\vec{x_u}$ and $\vec{x_{a_i}}$ for $i \in \{1, \ldots, 100\}$. This was to keep interpretation of results simple. Equivalently, the measurement vectors have a 10-dimensional multivariate normal distribution with 10-dimensional mean vectors and a 10×10 covariance matrix. The multivariate normal distribution is typically used as a conjugate prior for the mean of a multivariate normal distribution [14]. Therefore, for the hyperparameters μ_M and μ_s we choose

prior distributions

$$\mu_M \sim N(0, 40^2), \tag{36}$$

$$\mu_s \sim N(0, 60^2), \tag{37}$$

such that $\vec{\mu_M}$ and $\vec{\mu_s}$ are multivariate normally distributed. The prior distribution for μ_s only holds for the specific source model, since in the common source model it's drawn according to equation (34). The mean value of 0 is chosen since we have $\mu_M = 0$ as the underlying truth. The variances are chosen to denote an uninformative prior (wide distribution) and such that both underlying truth scenarios of μ_s can be contained in the prior of μ_s. For the variance hyperparameters σ_a, σ_M and σ_s we choose an uninformative uniform distribution, since it is a very simple distribution to interpret and we want the model to stay as general as possible. We chose the following prior distributions:

$$\sigma_s \sim U[0, 20], \tag{38}$$

$$\sigma_M \sim U[0, 40], \tag{39}$$

$$\sigma_a \sim U[0, 20]. \tag{40}$$

The prior distribution for σ_s only holds for the specific source model, since it does not occur in the common source model. These upper and lower values are chosen quite arbitrarily, but it should contain the underlying values of the parameter and since σ_M was chosen to be larger, we take a larger prior interval as well. Note that all these distributions are chosen independently of the hypotheses and scenarios. After running the MCMC sampler, we can check in the trace plots whether most of the probability mass of the estimated prior distributions of the μ's and σ's are within prior range.

Secondly, we want to rewrite the Bayes Factors. Recall that the distributions of e_a and e_s are independent of the hypotheses. In appendix C we showed that the Bayes Factors can be rewritten to

$$BF_{SS}(E) = \frac{\int_{\theta_{ss}} f(e_u \mid \theta_{ss}, H_p, I) f(\theta_{ss} \mid e_a, e_s, I) d\theta_{ss}}{\int_{\theta_{ss}} f(e_u \mid \theta_{ss}, H_d, I) f(\theta_{ss} \mid e_a, e_s, I) d\theta_{ss}} = \frac{BF_{SS,1}}{BF_{SS,2}}, \tag{C1}$$

$$BF_{CS}(E) = \frac{\int_{\theta_{cs}} f(e_u \mid \theta_{cs}, H_p, I) f(\theta_{cs} \mid e_a, e_s, I) d\theta_{cs}}{\int_{\theta_{cs}} f(e_u \mid \theta_{cs}, H_d, I) f(\theta_{cs} \mid e_a, e_s, I) d\theta_{cs}} = \frac{BF_{CS,1}}{BF_{CS,2}}. \tag{C2}$$

To approximate these Bayes Factors we want to generate a Markov Chain Monte Carlo sample of the unknown parameters, posterior on e_a and e_s. We generated a posterior sample of 4000 values using the PyMC3 package in Python, which uses the No U-Turn Sampler by default [18]. We denote the MCMC samples as follows:

$$f_{MC}(\theta_{ss} \mid e_a,\, e_s,\, I) = \left\{\mu_s^{(ss)}, \sigma_s^{(ss)}, \mu_M^{(ss)}, \sigma_M^{(ss)}, \sigma_a^{(ss)}\right\}, \tag{41}$$

$$f_{MC}(\theta_{cs} \mid e_a,\, e_s,\, I) = \left\{\mu_s^{(cs)}, \mu_M^{(cs)}, \sigma_M^{(cs)}, \sigma_a^{(cs)}\right\}. \tag{42}$$

Each element in these samples is a 4000-dimensional vector containing sampled posterior values of the parameter of interest. We added the initial trace plots of the specific source and common source model in Figures 7 and 8. The left side of these trace plots denotes an estimation of the posterior distribution of the parameters of the model and from this we can check if the model is well calibrated or not. So primarily, we expect the peak of this distribution to lie around the true value that we chose. The right side of the trace plots shows the Markov Chain steps that the sampling method has traveled through.

In the trace plots of this sampling method, we see the true value is in the interval for each parameter, but the peak of each distribution is slightly deviated from this value. This can be due to randomness or because we only performed 10 measurements on the glass fragments and this might not be enough to get a good estimation. In Figure 9 we added two more distribution plots taken from the common source trace plots: for Figure 9a we used 100 measurements to generate the evidence instead of 10 and for Figure 9b we used a different random seed[9]. Compared to Figure 8, we see that in the extra plots, the peak of the distribution for μ_M can be found on the other side of value 1 and the estimation for σ_a is better, while the estimations for σ_M and μ_s are slightly worse. This makes us more confident about our assumption that the deviation occurs due to randomness and not due to an error or a wrong assumption in our model. Although we do believe that the estimation is accurate because the peak lies in a small interval around the true value and the wideness of the distributions has decreased significantly compared to our input. We hope that this indicates a good estimation of the posterior parameter values as well. The parameters μ_{a_i} are nuisance parameters in both scenarios, because they are not of immediate interest since they're not the parameters that we need to estimate, but they must be accounted for in the analysis of the parameters in θ_{cs} and θ_{ss}. The trace plot for the parameter denoted as `mu_a` contains a trace plot for μ_{a_i} for each $i \in \{1, \ldots, 100\}$.

[9] By setting a random seed value, you are able to reproduce randomly sampled data when you run the code again.

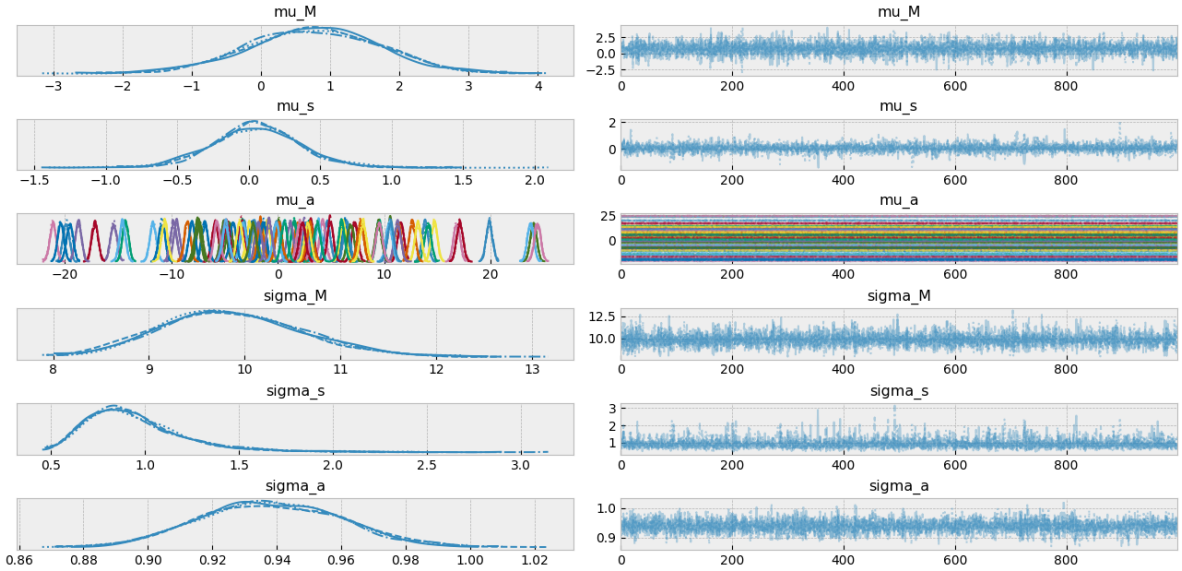


Figure 7: Trace plot for the specific source model Markov Chain Monte Carlo sampler (with true value of $\mu_s = 0$).

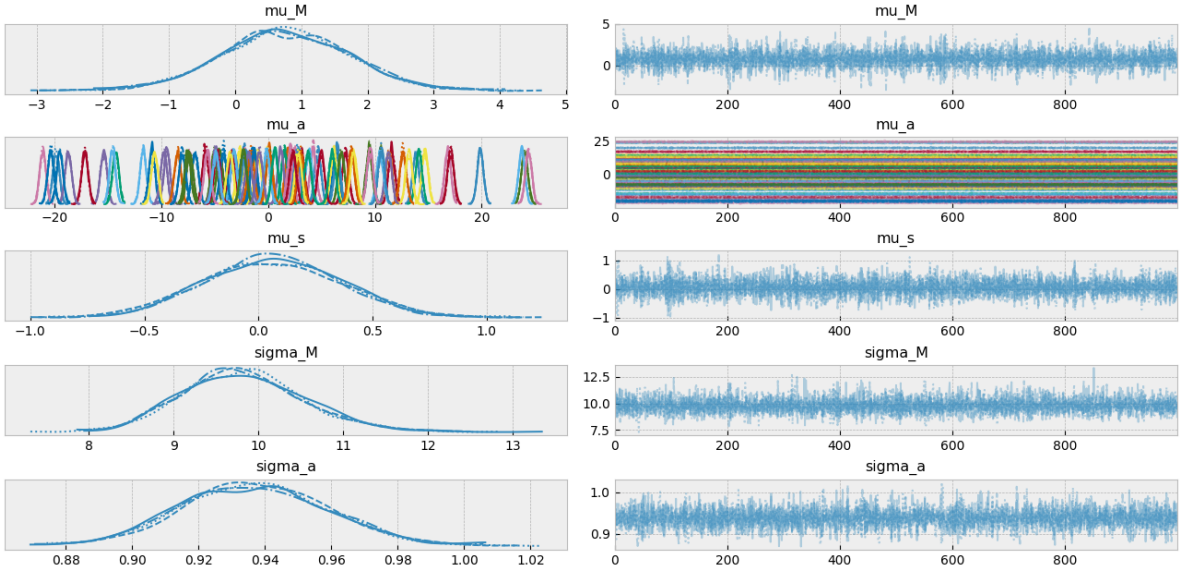
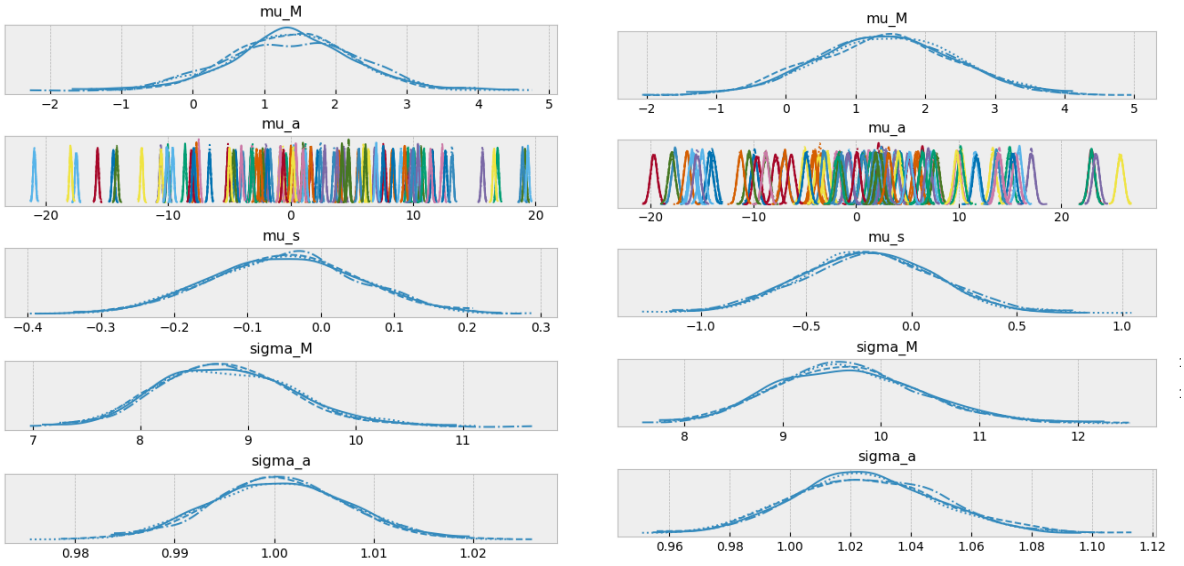


Figure 8: Trace plot for the common source model Markov Chain Monte Carlo sampler (with true value of $\mu_s = 0$).



(a) $m_a = m_s = 100$.

(b) Different random seed ($m_a = m_s = 10$).

Figure 9: Extra trace plots for the common source model Markov Chain Monte Carlo sampler (with true value of $\mu_s = 0$).

Using the Arithmetic Mean Estimate (or sample mean), we can now approximate the numerator and denominator of the Bayes Factors as

$$BF_{SS,1} \approx \frac{1}{4000} \sum_{i=1}^{4000} f(e_u \mid \theta_{ss}^{(i)}, H_p), \tag{43}$$

$$BF_{SS,2} \approx \frac{1}{4000} \sum_{i=1}^{4000} f(e_u \mid \theta_{ss}^{(i)}, H_d), \tag{44}$$

$$BF_{CS,1} \approx \frac{1}{4000} \sum_{i=1}^{4000} f(e_u \mid \theta_{cs}^{(i)}, H_p), \tag{45}$$

$$BF_{CS,2} \approx \frac{1}{4000} \sum_{i=1}^{4000} f(e_u \mid \theta_{cs}^{(i)}, H_d), \tag{46}$$

where $\theta_{ss}^{(i)}$ is drawn from the posterior sample $f_{MC}(\theta_{ss} \mid e_a, e_s, I)$, $\theta_{cs}^{(i)}$ is drawn from the posterior sample $f_{MC}(\theta_{cs} \mid e_a, e_s, I)$ and we omitted I for clarity. With one posterior sample (of 4000 values per parameter) we can compute 1 Bayes Factor. We will calculate 1000 common source and 1000 specific source Bayes Factors for when H_p is true and also for when H_d is true. We think this is enough to get a good sample of possible values. So for each MCMC sample we calculate two Bayes Factors: one where we use $\overrightarrow{x_{up}}$ as input to calculate

the likelihoods in equations (43) - (46) (H_p is true) and one with $\vec{x_{ud}}$ as input (H_d is true), since the distribution of $\vec{x_u}$ is different given each hypothesis (see equations (30) and (31)). Let us assume that H_p is true[10] and we want to calculate the four likelihoods that appear in the approximation of the Bayes Factors. Conditioned on H_p, we know the distribution of e_u in our models and we can draw parameter values from the MCMC samples to calculate the likelihood. We only use the MCMC samples for μ_s and σ_a since the others are not relevant for these distributions:

$$f(e_u \mid \theta_{ss}^{(i)}, H_p, I) = \phi_{10}(\vec{x_{up}} \mid \mu_s^{(ss)}, \sigma_a^{(ss)}) = \prod_{j=1}^{10} \phi(x_{up,j} \mid \ \mu_s^{(ss)}, \sigma_a^{(ss)}); \tag{47}$$

$$f(e_u \mid \theta_{cs}^{(i)}, H_p, I) = \phi_{10}(\vec{x_{up}} \mid \mu_s^{(cs)}, \sigma_a^{(cs)}) = \prod_{j=1}^{10} \phi(x_{up,j} \mid \ \mu_s^{(cs)}, \sigma_a^{(cs)}). \tag{48}$$

Here ϕ is the Gaussian density. We can take the product of the density of the vector elements of $\vec{x_{up}}$, because they are independent of each other.
Conditioned on H_d it is slightly more computationally intensive, because the distribution of $\vec{x_u}$ depends on μ_{a_u} in both models (see equations (31) and (35)), which is not an element of the MCMC samples. Consequently, we need to solve the following integrals[11]

$$\begin{aligned} f(e_u \mid \theta_{ss}^{(i)}, H_d, I) &= f(\vec{x_u} \mid \mu_M^{(ss)}, \sigma_M^{(ss)}, \sigma_a^{(ss)}, H_d) \\ &= \int_{\mu_{a_u}} \phi_{10}(\vec{x_u} \mid \mu_{a_u}, \sigma_a^{(ss)}) \ \phi(\mu_{a_u} \mid \mu_M^{(ss)}, \sigma_M^{(ss)}) d\mu_{a_u} \end{aligned} \tag{49}$$

$$\begin{aligned} f(e_u \mid \theta_{cs}^{(i)}, H_d, I) &= f(\vec{x_u} \mid \mu_M^{(cs)}, \sigma_M^{(cs)}, \sigma_a^{(cs)}, H_d) \\ &= \int_{\mu_{a_u}} \phi_{10}(\vec{x_u} \mid \mu_{a_u}, \sigma_a^{(cs)}) \ \phi(\mu_{a_u} \mid \mu_M^{(cs)}, \sigma_M^{(cs)}) d\mu_{a_u} \end{aligned} \tag{50}$$

[10] We will calculate the likelihoods for when H_p is true, since the scenario where H_d is true is exactly the same, but we use $\vec{x_{ud}}$ instead of $\vec{x_{up}}$ as input.

[11] If a random variable X has a distribution indexed with an unknown parameter θ, by the law of total probability, we can write the probability as $\mathbb{P}(X) = \int_\theta \mathbb{P}(X \mid \theta)\mathbb{P}(\theta) d\theta$.

The computation has been done in appendix D which yields the result in equation (D1), where we use $\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_{u,j}$ to get:

$$f(e_u \mid \theta_{ss}^{(i)}, H_d, I) = \exp\left(-\frac{\sum_{j=1}^{10}(\bar{x}-x_{u,j})^2}{2\left(\sigma_a^{(ss)}\right)^2}\right) \times \frac{1}{\sqrt{10\left(2\pi\left(\sigma_a^{(ss)}\right)^2\right)^9}} \times \phi\left(\bar{x} \mid \mu_M^{(ss)}, \left(\sigma_M^{(ss)}\right)^2 + \frac{\left(\sigma_a^{(ss)}\right)^2}{10}\right); \tag{51}$$

$$f(e_u \mid \theta_{cs}^{(i)}, H_d, I) = \exp\left(-\frac{\sum_{j=1}^{10}(\bar{x}-x_{u,j})^2}{2\left(\sigma_a^{(cs)}\right)^2}\right) \times \frac{1}{\sqrt{10\left(2\pi\left(\sigma_a^{(cs)}\right)^2\right)^9}} \times \phi\left(\bar{x} \mid \mu_M^{(cs)}, \left(\sigma_M^{(cs)}\right)^2 + \frac{\left(\sigma_a^{(cs)}\right)^2}{10}\right); \tag{52}$$

We now have all *ingredients* to calculate 1000 Bayes Factors for when H_p is true and 1000 Bayes Factors when H_d is true. This is done with algorithm 2. Generating the evidence, computing the likelihoods and computing a Bayes Factor only took a couple seconds in this algorithm. The MCMC sampling, however, took about 1 minute per sample, which means that it took about 67 hours to approximate 2000 Bayes Factors for the common source model and 2000 Bayes Factors for the specific source model. Of course this is far from ideal, but we need a new MCMC sample for each Bayes Factor because we cannot fix this posterior sample when it is chosen randomly.

Algorithm 2: Algorithm for calculating the specific source and common source Bayes Factors for the two-level model.

Input: $\mu_s = \mu_M = 0,\ \sigma_M = 10,\ \sigma_a = \sigma_s = 1.$;

Output: $BF_{SS},\ BF_{CS}$;

for $j \in \{1, 2, \ldots, 1000\}$ **do**

  Generate underlying specific source truth;

  return $\overrightarrow{x_{up}},\ \overrightarrow{x_{ud}}$;

  for $y \in \{\overrightarrow{x_{up}},\ \overrightarrow{x_{ud}}\}$ **do**

    for *Scenario S = Specific Source* **do**

      Generate posterior MCMC sample $f_{MC}(\theta_{ss} \mid e_a,\ e_s,\ I)$;

      Compute $f(e_u \mid H_p, \theta_{ss}^{(i)}, I)$ and $f(e_u \mid H_d, \theta_{ss}^{(i)}, I)$ using y as the observed value;

      Compute $BF_{SS} = BF_{SS,1}/BF_{SS,2}$.

    end

    for *Scenario S = Common Source* **do**

      Generate posterior MCMC sample $f_{MC}(\theta_{cs} \mid e_a,\ e_s,\ I)$;

      Compute $f(e_u \mid H_p, \theta_{cs}^{(i)}, I)$ and $f(e_u \mid H_d, \theta_{cs}^{(i)}, I)$ using y as the observed value;

      Compute $BF_{CS} = BF_{CS,1}/BF_{CS,2}$.

    end

    return $BF_{SS},\ BF_{CS}$

  end

end

4.3 Bayes Factor Comparison

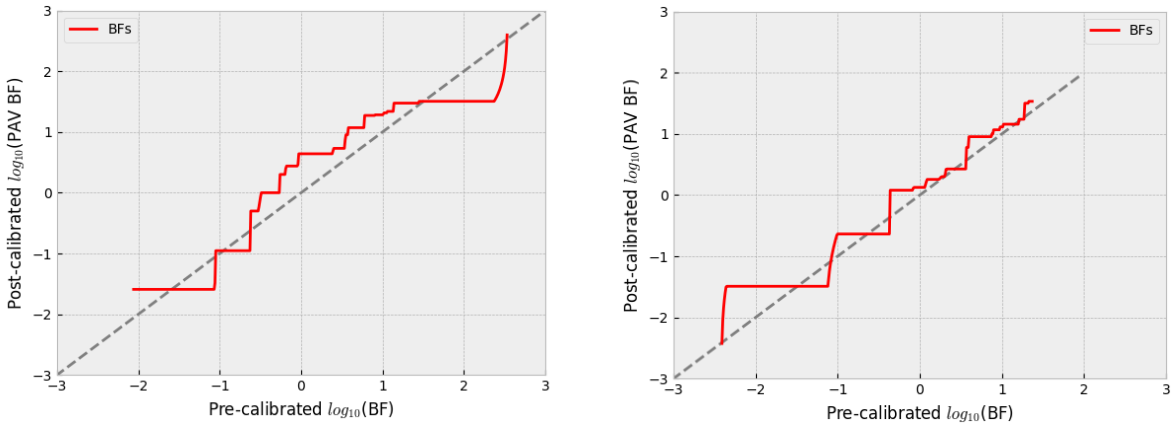
Since we calculated all the Bayes Factors, we can now compare the two scenarios like we did for the previous model. But since we have some unknown parameters we cannot compute the Cross Entropy. Recall the definition of the Empirical Cross Entropy from section 2.5 to approximate the Cross Entropy:

$$\begin{aligned} ECE = &\frac{P(H_p \mid I)}{1000} \sum_{i:H_p \text{ is true}} \log_2 \left(1 + \frac{1}{BF_i \times \frac{P(H_p|I)}{P(H_d|I)}} \right) + \\ &\frac{P(H_d \mid I)}{1000} \sum_{j:H_d \text{ is true}} \log_2 \left(1 + BF_j \times \frac{P(H_p \mid I)}{P(H_d \mid I)} \right). \end{aligned} \tag{15}$$

We can compute this for a range of prior odds.

4.3.1 Specific Source with General Characteristics

Firstly, we want to check the calibration of the Bayes Factors. The PAV method (as explained in Section 2.6) is a visual method for calibration for which we can make a PAV-plot, where we plot the log-Bayes Factors against the log of the Bayes Factors after they underwent the PAV transformation. The PAV transformed Bayes Factors are perfectly calibrated, so if the Bayes Factors are already well-calibrated before undergoing the PAV transformation, the difference between these two should be small. Therefore, if the PAV-plot is very close to the line $y = x$, the Bayes Factors indicate good calibration, whereas a large area between the two lines indicates bad calibration [39, 38]. The wideness of the steps in the plot is just the wideness of the bins used to make the PAV transform. In Figure 10 we see that the area between the two lines is the biggest in the specific source scenario, which means that the common source model has better calibration. The calibration of the specific source model is not as good as we were hoping. We expect that this is the case because we only used 10 measurements and σ_s is only determined by e_s. So we also made PAV plots for the model where we used $m_a = m_s = 100$, which can be found in 11. We see that the PAV Bayes Factors of the specific source model at least alternate above and below the line $y = x$, instead of being mostly above this line. This indicates an improvement in calibration. However, in Figure 10 both red lines seem to move around the line $y = x$ closely enough for us to conclude that the model is calibrated well enough for us to continue our analysis.



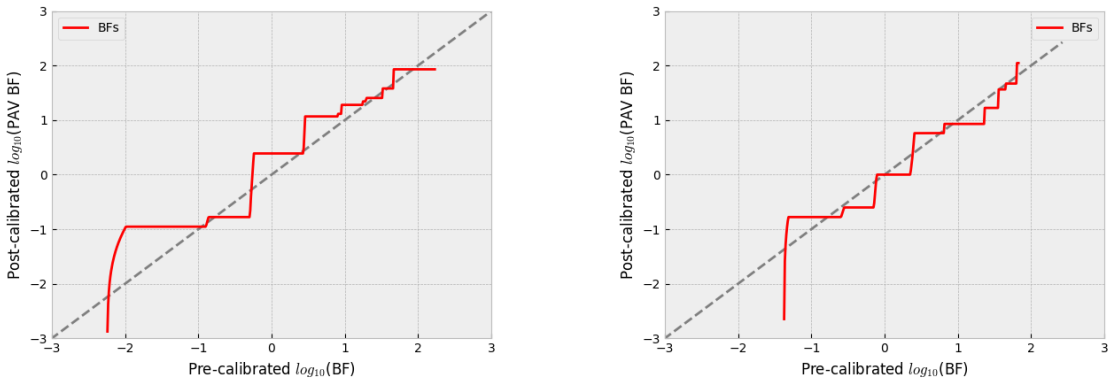
(a) Specific Source

(b) Common Source

Figure 10: PAV transforms of the specific source and common source Bayes Factors (with true value of $\mu_s = 0$).

Now let us consider the ECE plots in Figure 12. Compared to the reference plot, our Bayes Factor curve is very low. From Section 2.6 we know that a low Bayes Factor curve defines an accurate system. The PAV Bayes Factor curve is also very low, which shows us that the system also has a high discriminating power. The distance between the dashed and solid curves is shown more clearly in Figures 13a and 13b. The curves are very close together which indicates a well calibrated system like we concluded from the PAV-plots. The common source curves lie closer together than the specific source curves, which also confirms that the common source system is better calibrated than the specific source system, even if the difference is small. All of these features of our model are looking positive.

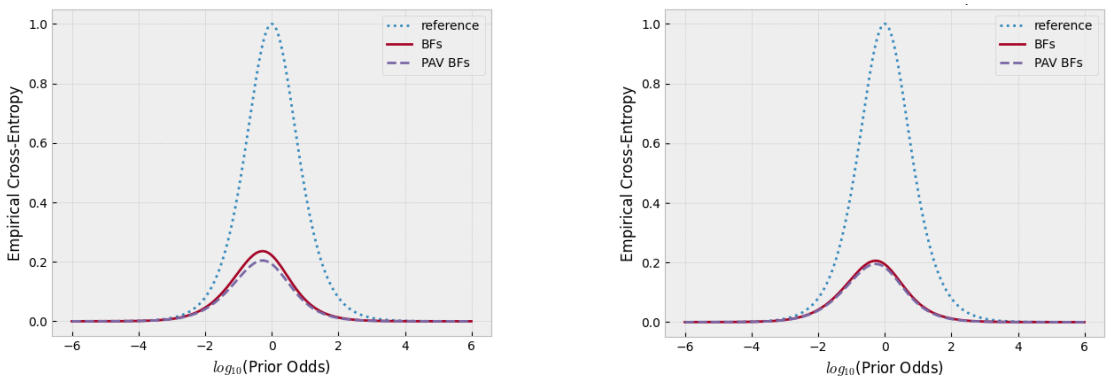
The difference between the Empirical Cross Entropy of the specific source and common source scenario is shown in Figure 13c. We see that the common source Bayes Factors perform better overall. However, the difference between the two entropies seems to be of the same magnitude as the difference in calibration. So this might indicate that the performance of the specific source model is worse because its calibration is worse. To check this, in Figure 14 we added the ECE plots and ECE comparison for the Bayes Factors we calculated with more measurements. We see that the distance between the dashed and solid curves in Figures 14a and 14b is very similar (about 0.01), which indicates that both models are equally well calibrated. In Figure 14c we see that the common source model still performs better in this case. We also see that the dashed curves are lower for the common source model, which means that after perfect calibration the common source model also performs better



(a) Specific Source (b) Common Source

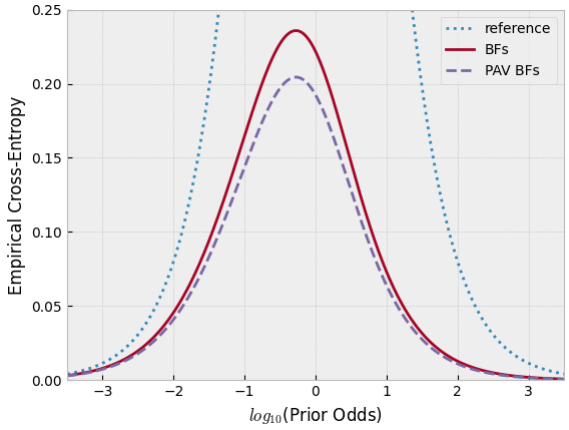
Figure 11: Extra PAV transforms of the specific source and common source Bayes Factors with $m_a = m_s = 100$ (and true value of $\mu_s = 0$).

than the specific source model. Our explanation for this is the one we gave in Section 1.3: the common source model provides more information on the parameters μ_s, σ_s and σ_a via the background population than the specific source hypotheses because we assume that the reference source is part of this population and we do not assume this in the specific source scenario. Furthermore, the common source scenario performs better in this model, but not necessarily in the DNA model in Chapter 3 and we expect that this is due to the difference in amount of uncertainty we have in the models. This two-level model has more uncertainties and randomness and therefore, the background information that we take into account with the common source model provides more information.

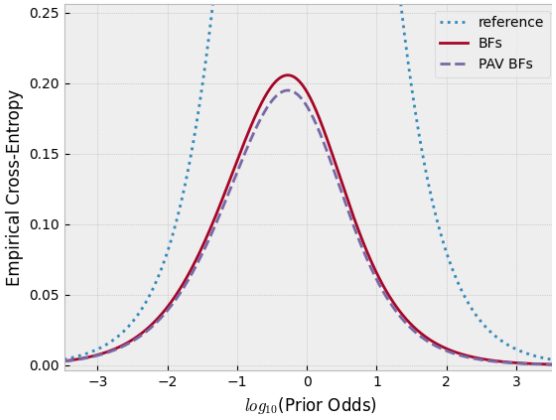


(a) Specific Source (b) Common Source

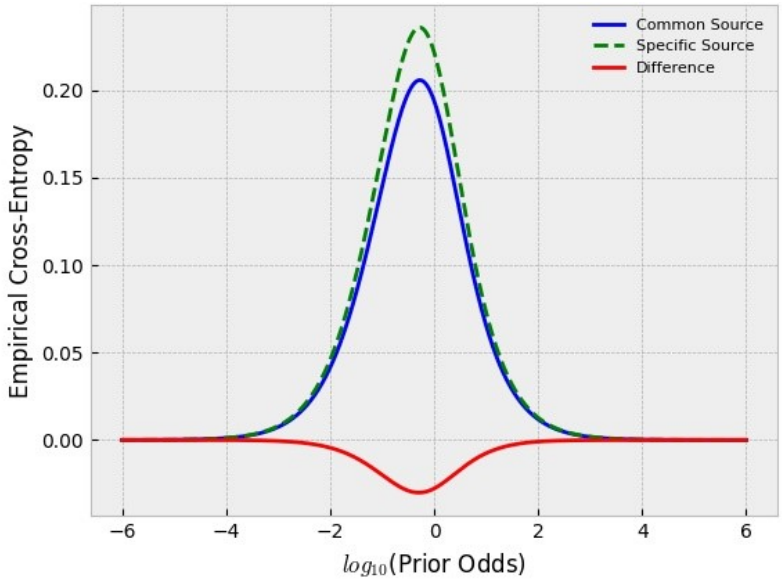
Figure 12: ECE plots for the specific source and common source Bayes Factors (with true value of $\mu_s = 0$).



(a) Specific Source

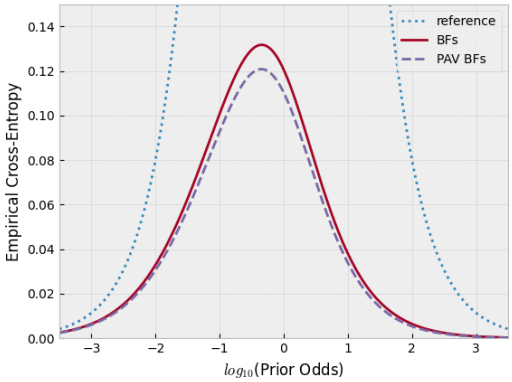


(b) Common Source

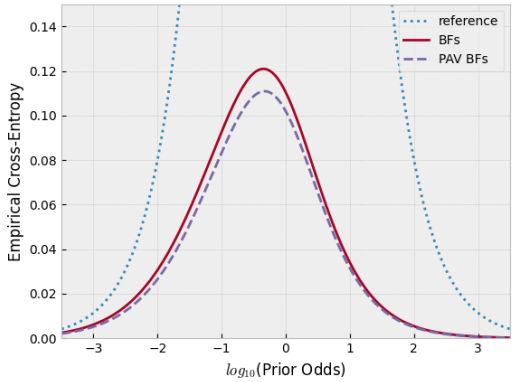


(c) ECE Comparison

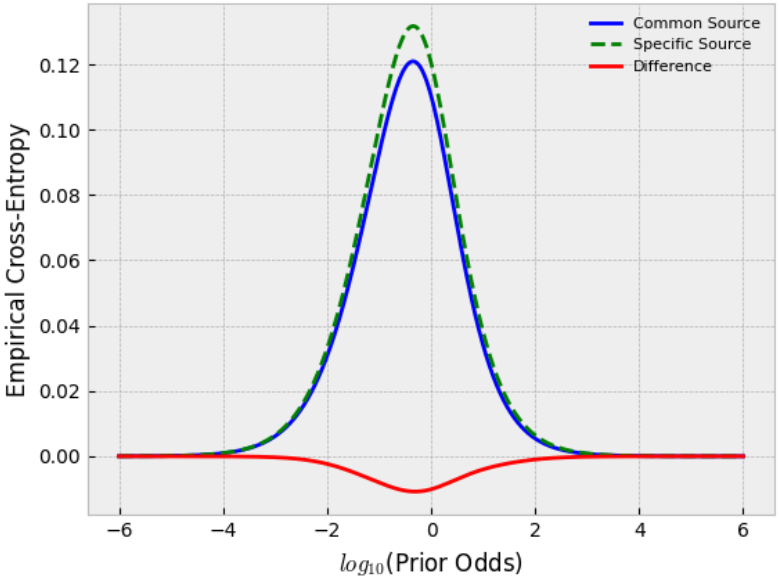
Figure 13: Zoomed ECE plots for the specific source and common source Bayes Factors and their comparison (with true value of $\mu_s = 0$).



(a) Specific Source

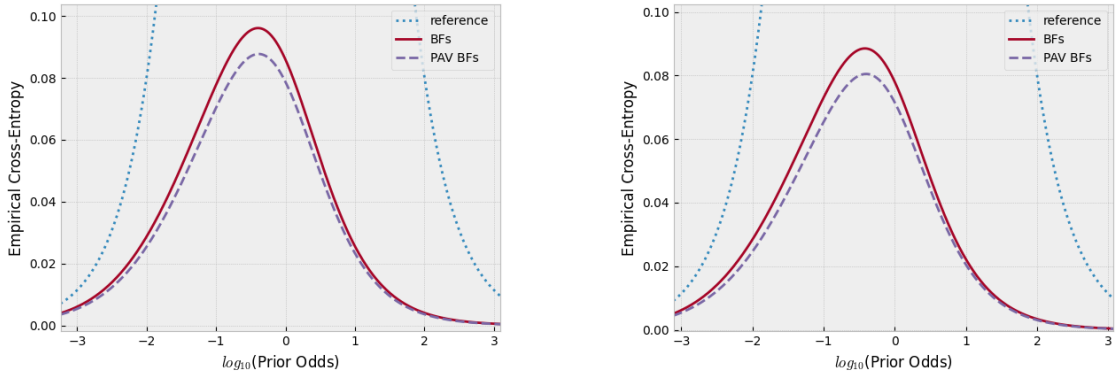


(b) Common Source



(c) ECE Comparison

Figure 14: ECE plots for the specific source and common source Bayes Factors and their comparison (with $m_a = m_s = 100$).



(a) Specific Source

(b) Common Source

Figure 15: Zoomed ECE plots for the specific source and common source Bayes Factors (with true value of $\mu_s = 20$).

4.3.2 Specific Source with Rare Characteristics

In this section we will look at the results where we used true value of $\mu_s = \mu_M + 2 \times \sigma_M = 20$ to generate evidence. The priors that we used will remain the same and the Bayes Factor calculation as well.

As expected the ECE plot curves are lower than in the general characteristic scenario, which means the discriminating power and accuracy are higher for rare characteristics (see Figure 15). This can also be seen in the Empirical Cross Entropy comparison plot in Figure 16, these curves are lower than the Empirical Cross Entropy plot for the model where $\mu_s = 0$. We see that the difference is also smaller and the common source scenario still performs better than the specific source scenario. Additional plots can be found in appendix F. In the trace plots (Figures 22 and 23) we see that the intervals of the estimated prior distributions of the parameters still contain the true value and the means are actually closer to this value than in the previous case. From Figure 25 we conclude that these Bayes Factors are well-calibrated because the lines move around the line $y = x$ nicely and the same can be concluded from Figure 15: the difference between the solid and dashed curves is small.

This model has small differences compared to the previous model, but they are both well-calibrated and show that the common source scenario provides more information.

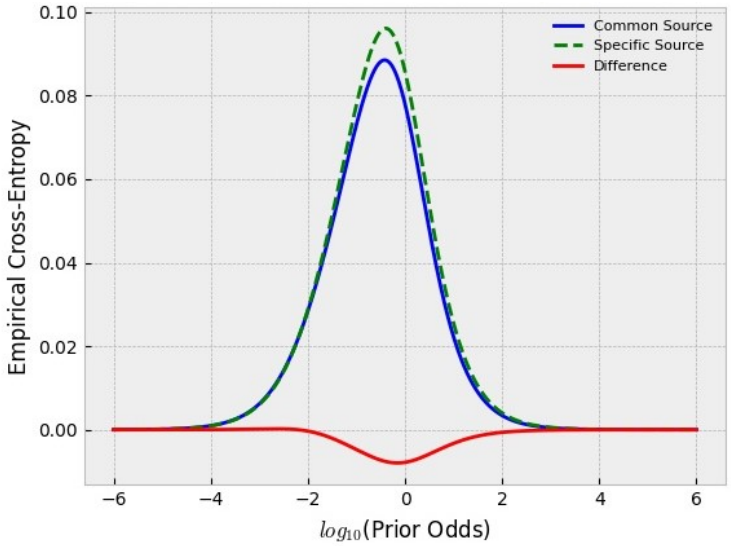


Figure 16: Common source and specific source Empirical Cross Entropy (and their difference) plotted for a range of $\log_{10}$(Prior Odds) (with true value of $\mu_s = 20$).

4.3.3 Sensitivity Analysis

Since we chose the parameters of the prior distributions in equations (36) - (40) quite arbitrarily, we also did the Markov Chain Monte Carlo sampling with wider priors to see if there would be a significant difference in the posterior distribution samples for our unknown parameters. The prior distributions that we used are:

$$\mu_M \sim N(0, 100^2), \tag{53}$$
$$\mu_s \sim N(0, 200^2), \tag{54}$$
$$\sigma_s \sim U[0, 100], \tag{55}$$
$$\sigma_M \sim U[0, 200], \tag{56}$$
$$\sigma_a \sim U[0, 100]. \tag{57}$$

In the trace plots (Figures 17 and 18) we see that the posterior distributions of the MCMC sample are similar to the posterior distributions that we got for the 'smaller' priors in the sense that they fall in the same intervals. The same result is obtained when we use true value $\mu_s = 20$ (Figures 26 and 27 in appendix F). The posterior distributions fall in the same intervals as when we used the smaller priors. Therefore, we conclude that the choice of parameters for the prior do not have a large influence on the results of the MCMC sampling method, but that the distributions are mostly determined by the values drawn for e_a, e_s and e_u.

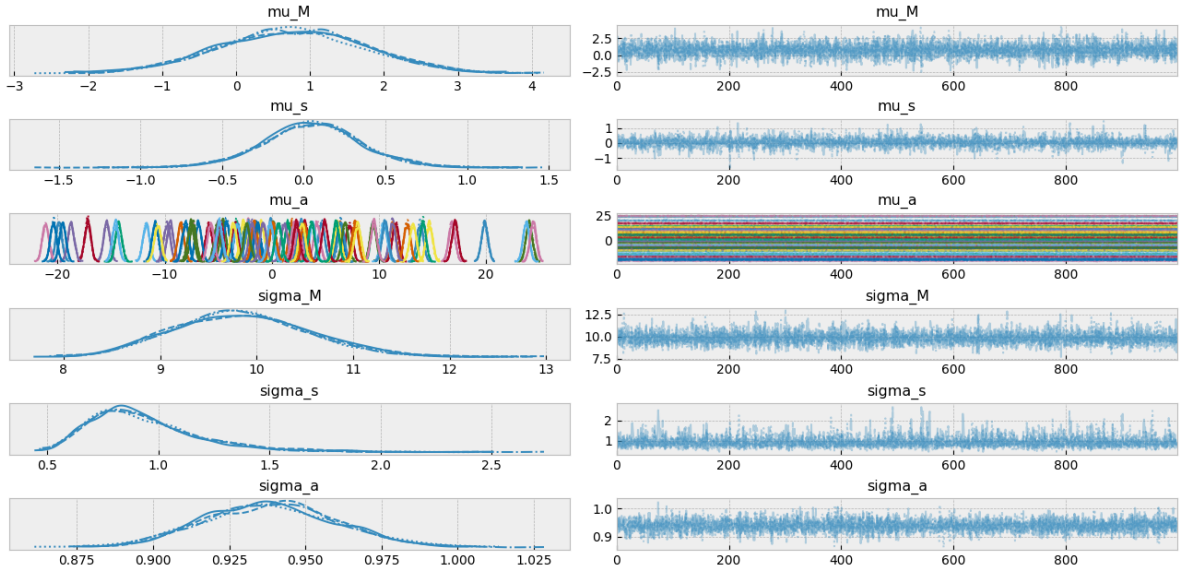


Figure 17: Trace plot for the specific source model Markov Chain Monte Carlo sampler with wider priors (with true value of $\mu_s = 0$).

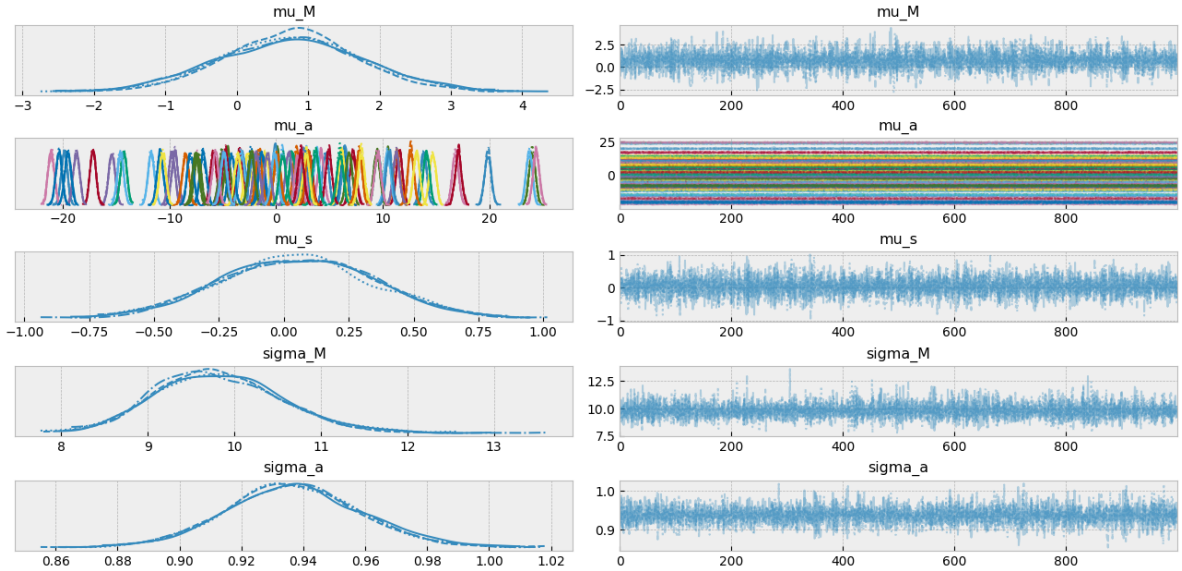
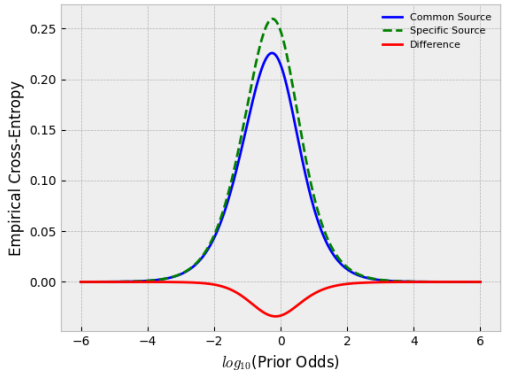
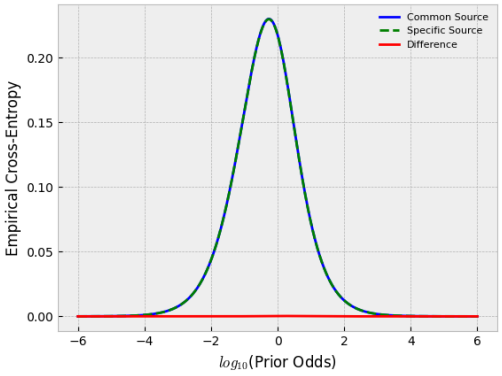


Figure 18: Trace plot for the common source model Markov Chain Monte Carlo sampler with wider priors (with true value of $\mu_s = 0$).



(a) Analysis 1: $\mu_s \sim N(\mu_M, \sigma_M^2)$.



(b) Analysis 2: $\sigma_s = \sigma_a$.

Figure 19: Common source and specific source Empirical Cross Entropy (and their difference) plotted for a range of $\log_{10}$(Prior Odds) for the extra analyses.

We also tried to isolate which difference in the models causes the specific source scenario to perform worse than the common source scenario. There are two things that are different between the two models: in the common source scenario μ_s is normally distributed with parameters μ_M and σ_M, whereas in the specific source scenario we impose a prior distribution on it; and we used σ_a^2 as the variance for the distribution of $\vec{x_s}$ in the common source model, whereas we used σ_s^2 as the variance in the specific source model. We calculated the Bayes Factors again twice, but for each time we changed something in the specific source model to make the models more similar: for the Markov Chain Monte Carlo method (1) we let $\mu_s \sim N(\mu_M, \sigma_M)$ or (2) we use σ_a instead of σ_s in the distribution of x_s.
After we generated the MCMC sample[12] and we calculated all the Bayes Factors, we can compute the Empirical Cross Entropy. The validation plots for these models can be found in appendix G. Both models are well-calibrated and have high discriminating power. The most interesting plot is the comparison of the ECE between both models, these can be found in Figure 19. Here we can see that for the second analysis the difference between the empirical cross entropies is very close to 0. In Figure 20 we zoomed in on this plot, to actually see that the specific source system even performs slightly better in this case. For the first analysis however, the common source system has lower entropy and therefore performs better. This shows us that the choice to use σ_s instead of σ_a in the specific source model largely explains why the specific source scenario performs worse.

[12] We used true value of $\mu_s = 0$.

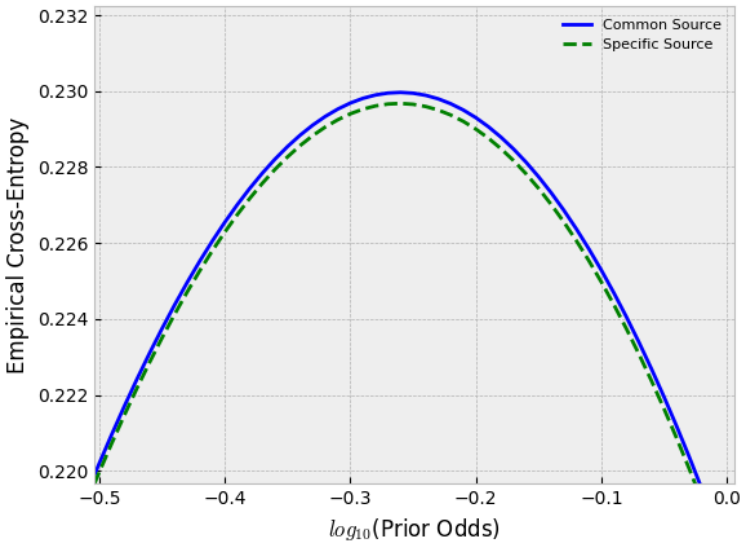


Figure 20: Zoomed plot for analysis 2.

5 Discussion

In this thesis, we defined two models to assess the performance of common source and specific source Bayes Factor systems and their value in updating the prior odds of the specific source hypotheses through Bayes' theorem.
The first model was an adaptation of the Beta-Binomial model described by Van Dorp, Leegwater, Alberink and Jongbloed [36] applied to discrete evidence. We saw that the specific source scenario performed better in general, but for higher values of prior parameters α and β the entropies were very close in value. This led us to believe that it is possible to use the common source Bayes Factor to update the prior odds of the specific source hypotheses. However, between these models this Bayes Factor does not have more value or give us more information than the specific source Bayes Factor. This might be the case because there is only one level of uncertainty in the model and the evidence based on the background population does not give us as much information.
The second model was an adaptation of the two-level Normal model described by Ommen, Neumann and Saunders [31] defined for continuous evidence. We used a Markov Chain Monte Carlo sampler to approximate the Bayes Factors and assessed the performance of the Bayes Factor systems using the Empirical Cross Entropy. We saw that the common source Bayes Factor system performed better overall. The common source model is better calibrated than the specific source model, however neither of them were perfectly calibrated. In addition we also saw that the ECE of the Bayes Factors after we applied the PAV transformation, still showed that the common source system performed better. So we expect that a model for continuous evidence that is calibrated better, will still tell us that the specific source Bayes Factor system provides less information. We believe the common source model provides more information on the parameters μ_s, σ_s and σ_a, because the background population is more important in this model and therefore we have more information and the entropy is lower. Furthermore, we concluded that the specific source Bayes Factor system performed worse, mainly because we have less information about the variance for the specific source measurements in the specific source model (σ_s^2) than in the common source model (σ_a^2). This is probably due to the fact that we only had 10 measurements, which might not be enough for the model to estimate this parameter properly.

Our advice for the forensic science community is to consider using the common source model when working on a specific source problem. Or more specifically, consider whether the measurement error can be described by a parametric model common to for specific source and the background sources. This advice is based on our two-level models, which had more uncertainty than the Beta-Binomial model and this resulted in the common source model performing better. If this advice is followed, common source models may outperform specific source models. The calibration plots showed us that with simulated data, it's strongly recommended to use more measurements, such that the model will be calibrated better. However, this is often not possible due to lack of time. If this is the case, the common source model should be used since it's better calibrated overall.

5.1 Future Work

Our recommendation for future work is primarily that the models should be validated with real data, such that we can see how the entropies will relate to each other in a real forensic case. Secondly, we need to research how the models perform when we use different priors for the MCMC sampler and with different values for the underlying truth. Mostly to see how sensitive this research is to priors and to be able to draw a substantiated conclusion from it. Another way to improve the research is to find a way to draw MCMC samples faster, to improve the numerical performance. I was very new to this method, so the way I implemented it in Python might not have been the most efficient one. We chose the Markov Chain Monte Carlo method, because this is what is common in the forensic community. It might be beneficial to use a Randomized Markov Chain Quasi-Monte Carlo method instead, since it may approximate the true distribution better than the regular MCMC [22] and it has a faster convergence rate of $O(n^{-2})$ as opposed to the $O(n^{-1})$ Monte Carlo rate [25].

Bibliography

[1] Colin GG Aitken and David Lucy. Evaluation of trace evidence in the form of multivariate data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 53(1):109–122, 2004.

[2] Ivo Alberink, Annabel Bolck, and Sonja Menges. Posterior likelihood ratios for evaluation of forensic trace evidence given a two-level model on the data. *Journal of Applied Statistics*, 40(12):2579–2600, 2013.

[3] Charles H Brenner. Fundamental problem of forensic mathematics—the evidential value of a rare haplotype. *Forensic Science International: Genetics*, 4(5):281–291, 2010.

[4] Glenn W Brier et al. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

[5] Niko Brümmer and Johan Du Preez. Application-independent evaluation of speaker detection. *Computer Speech & Language*, 20(2-3):230–275, 2006.

[6] Giulia Cereda. Bayesian approach to lr assessment in case of rare type match. *Statistica Neerlandica*, 71(2):141–164, 2017.

[7] Siddhartha Chib and Bradley P Carlin. On mcmc sampling in hierarchical longitudinal models. *Statistics and Computing*, 9(1):17–26, 1999.

[8] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[9] A.P. Dawid. Probality forecasting. *Kotz, S., Johnson, N.L., Read, C.B. (eds.) Encyclopedia of statistical sciences*, 7:210–218, 1986.

[10] Herbert Dawid, Philipp Harting, and Michael Neugart. Economic convergence: Policy implications from a heterogeneous agent model. *Journal of Economic Dynamics and Control*, 44:54–80, 2014.

[11] B De Finetti. *Theory of probability*, volume 1 & 2. John Wiley & Sons, New York, 1975.

[12] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112, 2009.

[13] Ian W Evett. Bayesian inference and forensic science: problems and perspectives. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 36(2/3):99–105, 1987.

[14] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2003 (second edition).

[15] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

[16] Irving John Good. Rational decisions. In *Breakthroughs in statistics*, pages 365–377. Springer, 1992.

[17] Cong Han and Bradley P Carlin. Mcmc methods for computing bayes factors: a comparative review. *Biometrika*, 82(4):711–732, 2000.

[18] Matthew D Hoffman, Andrew Gelman, et al. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

[19] Harold Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461, 1946.

[20] Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

[21] Terrence F Kiely. *Forensic evidence: science and the criminal law*. CRC Press, 2005.

[22] Pierre L'Ecuyer, Christian Lécot, and Bruno Tuffin. A randomized quasi-monte carlo simulation method for markov chains. *Operations research*, 56(4):958–975, 2008.

[23] Dennis V Lindley. A problem in forensic science. *Biometrika*, 64(2):207–213, 1977.

[24] David Lunn, Christopher Jackson, Nicky Best, Andrew Thomas, and David Spiegelhalter. The bugs book. *A Practical Introduction to Bayesian Analysis, Chapman Hall, London*, 2013.

[25] Pierre L'Ecuyer, David Munger, Christian Lécot, and Bruno Tuffin. Sorting methods and convergence rates for array-rqmc: Some empirical comparisons. *Mathematics and Computers in Simulation*, 143:191–201, 2018.

[26] Didier Meuwly, Daniel Ramos, and Rudolf Haraksim. A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation. *Forensic science international*, 276:142–153, 2017.

[27] Cedric Neumann and Madeline Ausdemore. Defence against the modern arts: the curse of statistics—part ii:'score-based likelihood ratios'. *Law, Probability and Risk*, 19(1):21–42, 2020.

[28] Tony O'Hagan. Dicing with the unknown. *Significance*, 1(3):132–133, 2004.

[29] Danica M Ommen. *Approximate statistical solutions to the forensic identification of source problem*. South Dakota State University, 2017.

[30] Danica M Ommen and Christopher P Saunders. Building a unified statistical framework for the forensic identification of source problems. *Law, Probability and Risk*, 17(2):179–197, 2018.

[31] Danica M Ommen, Christopher P Saunders, and Cedric Neumann. The characterization of Monte Carlo errors for the quantification of the value of forensic evidence. *Journal of Statistical Computation and Simulation*, 87(8):1608–1643, 2017.

[32] Daniel Ramos. Phd: Forensic evaluation of the evidence using automatic speaker recognition systems. 2007.

[33] Daniel Ramos and Joaquin Gonzalez-Rodriguez. Cross-entropy analysis of the information in forensic speaker recognition. In *Odyssey 2008: The Speaker and Language Recognition Workshop*. International Speech Communication Association, 2008.

[34] Daniel Ramos and Joaquin Gonzalez-Rodriguez. Reliable support: measuring calibration of likelihood ratios. *Forensic science international*, 230(1-3):156–169, 2013.

[35] John Salvatier, Thomas V Wiecki, and Christopher Fonnesbeck. Probabilistic programming in python using pymc3. *PeerJ Computer Science*, 2:e55, 2016.

[36] IN Van Dorp, AJ Leegwater, I Alberink, and G Jongbloed. Value of evidence in the rare type match problem: common source versus specific source. *Law, Probability and Risk*, 19(1):85–98, 2020.

[37] David A Van Leeuwen and Niko Brümmer. The distribution of calibrated likelihood-ratios in speaker recognition. *arXiv preprint arXiv:1304.1199*, 2013.

[38] Peter Vergeer, Ivo Alberink, Marjan Sjerps, and Rolf Ypma. Why calibrating lr-systems is best practice. a reaction to "the evaluation of evidence for microspectrophotometry data using functional data analysis", in fsi 305. *Forensic Science International*, 314:110388, 2020.

[39] Grzegorz Zadora, Agnieszka Martyna, Daniel Ramos, and Colin Aitken. *Statistical analysis in forensic science: evidential value of multivariate physicochemical data*. John Wiley & Sons, 2013.

Appendices

A Rewriting the Bayes Factor Expressions for the Beta-Binomial Model

Recall that a general Bayes Factor is of the form

$$BF(E) := \frac{\int_{\theta_p} f(E \mid H_p,\, \theta_p,\, I)\, f(\theta_p \mid H_p,\, I) d\theta_p}{\int_{\theta_d} f(E \mid H_d,\, \theta_d,\, I)\, f(\theta_d \mid H_d, I) d\theta_d}. \tag{3}$$

In [31], Ommen et Al. derive different expressions for the specific source and common source Bayes Factors, by rewriting the expressions in the integral. To do this we assume that the parameters of the models are the same for both hypotheses, so $\theta = \theta_p = \theta_d$ and $f(\theta \mid H_p) = f(\theta \mid H_d) = f(\theta)$. Also note that the alternative evidence e_a is independent from the hypotheses, hence $f(e_a \mid H_p, \theta, I) = f(e_a \mid H_d, \theta, I) = f(e_a \mid \theta, I)$. So we get as the common source Bayes Factor:

$$BF_{CS}(E) = \frac{\int_{\theta} f(e_{u_1}, e_{u_2} \mid H_p, \theta, I) f(e_a \mid \theta, I)\, f(\theta \mid I) d\theta}{\int_{\theta} f(e_{u_1}, e_{u_2} \mid H_d, \theta, I) f(e_a \mid \theta, I)\, f(\theta \mid I) d\theta}. \tag{A1}$$

and as the specific source Bayes Factor:

$$BF_{SS}(E) = \frac{\int_{\theta} f(e_s, e_u \mid H_p, \theta, I) f(e_a \mid \theta, I)\, f(\theta \mid I) d\theta}{\int_{\theta} f(e_s, e_u \mid H_d, \theta, I) f(e_a \mid \theta, I)\, f(\theta \mid I) d\theta}, \tag{A2}$$

Bayes' rule $\mathbb{P}(A \mid B) = \mathbb{P}(B \mid A)\mathbb{P}(A)/\mathbb{P}(B)$ and the rule of conditional probability $\mathbb{P}(A, B) = \mathbb{P}(A \mid B)\mathbb{P}(B)$ will be used to rewrite these Bayes Factors as has been done in [36].

A.1 Common Source Bayes Factor

With Bayes' rule we see that

$$f(e_a \mid \theta, I) = \frac{f(\theta \mid e_a, I) f(e_a \mid I)}{f(\theta \mid I)}.$$

We can substitute this into the above expression for the common source Bayes Factor and get

$$\begin{aligned}
BF_{CS}(E) &= \frac{\int_\theta f(e_{u_1}, e_{u_2} \mid H_p, \theta, I) \frac{f(\theta|e_a,I) f(e_a|I)}{f(\theta|I)} f(\theta \mid I) d\theta}{\int_\theta f(e_{u_1}, e_{u_2} \mid H_d, \theta, I) \frac{f(\theta|e_a,I) f(e_a|I)}{f(\theta|I)} f(\theta \mid I) d\theta} \\
&= \frac{f(e_a \mid I) \int_\theta f(e_{u_1}, e_{u_2} \mid H_p, \theta, I) f(\theta \mid e_a, I) d\theta}{f(e_a \mid I) \int_\theta f(e_{u_1}, e_{u_2} \mid H_d, \theta, I) f(\theta \mid e_a, I) d\theta} \\
&= \frac{\int_\theta f(e_{u_1}, e_{u_2} \mid H_p, \theta, I) f(\theta \mid e_a, I) d\theta}{\int_\theta f(e_{u_1}, e_{u_2} \mid H_d, \theta, I) f(\theta \mid e_a, I) d\theta}.
\end{aligned}$$

Implementing that e_{u_1} and e_{u_2} are independent under H_d, we have

$$\begin{aligned}
BF_{CS}(E) &= \frac{\int_\theta f(e_{u_1}, e_{u_2} \mid H_p, \theta, I) f(\theta \mid e_a, I) d\theta}{\int_\theta f(e_{u_1} \mid H_d, \theta, I) f(e_{u_2} \mid H_d, \theta, I) f(\theta \mid e_a, I) d\theta} \\
&= \frac{\int_\theta f(e_{u_2} \mid e_{u_1}, H_p, \theta, I) f(e_{u_1} \mid H_p, \theta, I) f(\theta \mid e_a, I) d\theta}{\int_\theta f(e_{u_1} \mid H_d, \theta, I) f(e_{u_2} \mid H_d, \theta, I) f(\theta \mid e_a, I) d\theta} \\
&= \frac{\int_\theta g(y_{u_1} \mid \theta, I) \pi(\theta \mid e_a, I) d\theta}{\int_\theta g(y_{u_1} \mid \theta, I) g(y_{u_2} \mid \theta, I) \pi(\theta \mid e_a, I) d\theta} \qquad \text{(A3)} \\
&= \frac{BF_{CS,1}}{BF_{CS,2}}
\end{aligned}$$

Expression (A3) is obtained with the model specified in section 3.1.1. Here π is the distribution of θ, g is the density corresponding to distribution G in equations (16) - (19) and recall that under H_p, the two unknown source evidences are equal with probability 1. We can now calculate the numerator and the denominator by applying the Beta-Binomial model defined

in section 3.1.3.

$$\begin{aligned} BF_{CS,1} &= \int_{\theta} g(y_{u_1} \mid \theta, I)\pi(\theta \mid e_a, I)d\theta \\ &= \int_{\theta} \theta \frac{\Gamma(\alpha+\beta+n_a)}{\Gamma(\alpha+s_a)\Gamma(\beta+n_a-s_a)}\theta^{\alpha+s_a-1}(1-\theta)^{\beta+n_a-s_a-1}d\theta \\ &= \frac{\alpha+s_a}{\alpha+\beta+n_a} \end{aligned}$$

$$\begin{aligned} BF_{CS,2} &= \int_{\theta} g(y_{u_1} \mid \theta, I)g(y_{u_2} \mid \theta, I)\pi(\theta \mid e_a, I)d\theta \\ &= \int_{\theta} \theta^2 \frac{\Gamma(\alpha+\beta+n_a)}{\Gamma(\alpha+s_a)\Gamma(+n_a-s_a)}\theta^{\alpha+s_a-1}(1-\theta)^{\beta+n_a-s_a-1}d\theta \\ &= \frac{(\alpha+s_a)(\alpha+s_a+1)}{(\alpha+\beta+n_a)(\alpha+\beta+n_a+1)} \end{aligned}$$

This results in a common source Bayes Factor of the form

$$BF_{CS}(E) = \frac{\alpha+\beta+n_a+1}{\alpha+s_a+1} \tag{A4}$$

A.2 Specific Source Bayes Factor

Note that e_s does not depend on any parameter and therefore under H_p, e_u is independent of θ as well. We also note that e_s is independent from the hypotheses, so $f(e_s \mid H_p, I) = f(e_s \mid H_d, I) = f(e_s \mid I)$. This results in a specific source Bayes Factor of the form:

$$
\begin{aligned}
BF_{SS}(E) &= \frac{f(e_u \mid e_s, H_p, I) f(e_s \mid H_p, I) \int_\theta f(e_a \mid \theta, I)\, f(\theta \mid I) d\theta}{f(e_s \mid H_d, I) \int_\theta f(e_u \mid H_d, \theta, I) f(e_a \mid \theta, I)\, f(\theta \mid I) d\theta} \\
&= \frac{f(e_u \mid e_s, H_p, I) \int_\theta f(e_a \mid \theta, I)\, f(\theta \mid I) d\theta}{\int_\theta f(e_u \mid H_d, \theta, I) f(e_a \mid \theta, I)\, f(\theta \mid I) d\theta} \times \frac{f(e_a \mid I)}{f(e_a \mid I)} \\
&= f(e_u \mid e_s, H_p, I) \times \frac{\int_\theta f(e_a \mid \theta, I) f(\theta \mid I) d\theta}{f(e_a \mid I)} \times \frac{f(e_a \mid I)}{\int_\theta f(e_u \mid \theta, H_d, I) f(e_a \mid \theta, I) f(\theta \mid I) d\theta} \\
&= f(e_u \mid e_s, H_p, I) \times \frac{f(e_a \mid I)}{f(e_a \mid I)} \times \frac{f(e_a \mid I)}{\int_\theta f(e_u \mid \theta, H_d, I) \frac{f(\theta \mid e_a, I) f(e_a \mid I)}{f(\theta \mid I)} f(\theta \mid I) d\theta} \\
&= f(e_u \mid e_s, H_p, I) \times \frac{f(e_a \mid I)}{f(e_a \mid I) \int_\theta f(e_u \mid \theta, H_d, I) f(\theta \mid e_a, I) d\theta} \\
&= \frac{f(e_u \mid e_s, H_p, I)}{\int_\theta f(e_u \mid \theta, H_d, I) f(\theta \mid e_a, I) d\theta} \\
&= \frac{1}{\int_\theta g(y_u \mid \theta) \pi(\theta \mid e_a, I) d\theta} \\
&= \frac{1}{BF_{SS,2}},
\end{aligned}
\tag{A5}
$$

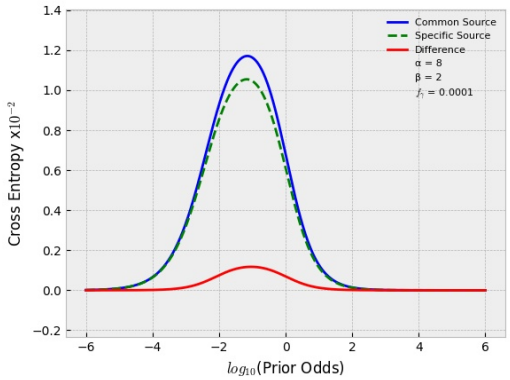
where we used Bayes' rule and that $f(e_a) = \int_\theta f(e_a \mid \theta) f(\theta) d\theta$, by basic Bayesian probability rules. Expression (A5) is obtained with the model specified in section 3.1.2. Here π is the distribution of θ, g is the density corresponding to distribution G in equations (16) - (19) and recall that under H_p, the unknown and specific source evidences are equal with probability 1. We can now calculate the denominator by applying the Beta-Binomial model defined in section 3.1.3.

$$
\begin{aligned}
BF_{SS,2} &= \int_\theta g(y_u \mid \theta) \pi(\theta \mid e_a, I) d\theta \\
&= \int_\theta \theta \frac{\Gamma(\alpha + \beta + n_a)}{\Gamma(\alpha + s_a)\Gamma(\beta + n_a - s_a)} \theta^{\alpha + s_a - 1} (1-\theta)^{\beta + n_a - s_a - 1} d\theta \\
&= \frac{\alpha + s_a}{\alpha + \beta + n_a}
\end{aligned}
$$

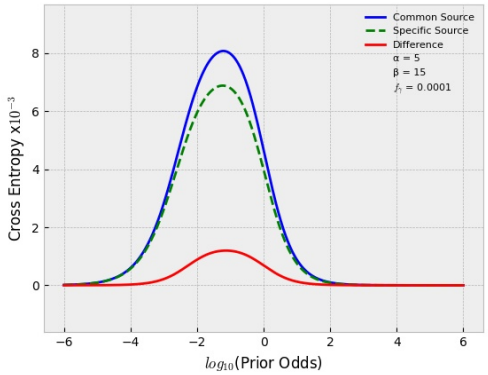
This results in a specific source Bayes Factor of the form

$$
BF_{SS}(E) = \frac{\alpha + \beta + n_a}{\alpha + s_a} \tag{A6}
$$

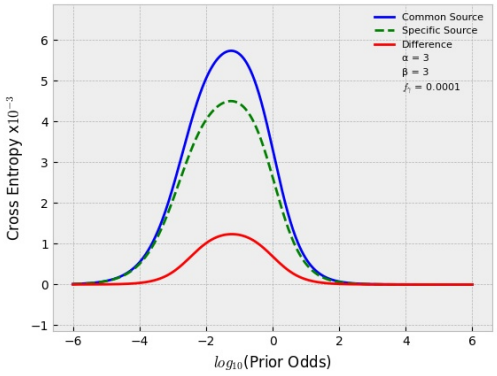
B Additional Cross Entropy Plots for the Beta-Binomial Model



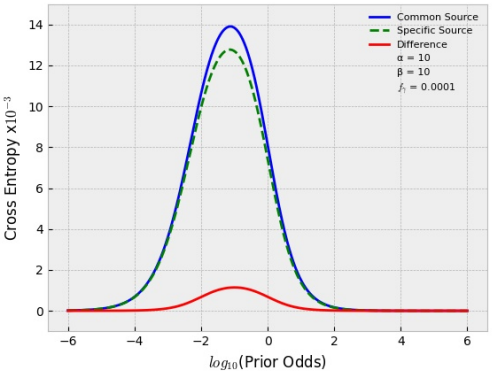
(a) $\alpha = 8, \beta = 2$.



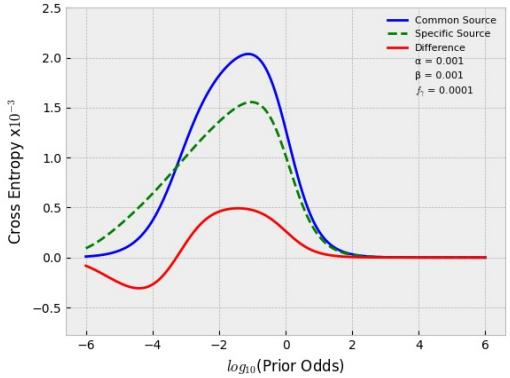
(b) $\alpha = 5, \beta = 15$.



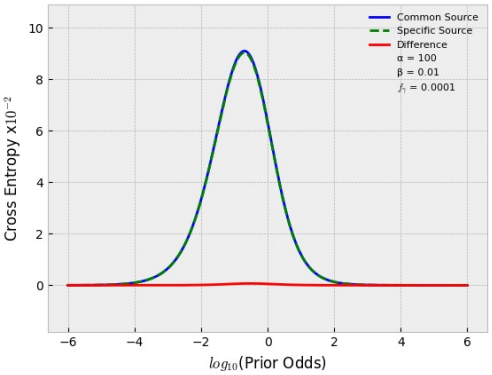
(c) $\alpha = \beta = 3$.



(d) $\alpha = \beta = 10$.



(e) $\alpha = \beta = 0.001$.



(f) $\alpha = 100, \beta = 0.01$.

Figure 21: Common source and specific source cross entropy (and their difference) plotted for a range of $\log_{10}$(Prior Odds) with frequency $f_\gamma = 0.0001$ and different prior parameter values.

C Rewriting the Bayes Factor Expressions for the two-level Model

Recall that with $\theta_{cs} = \{\mu_M,\, \sigma_M,\, \sigma_a\}$ the common source Bayes factor is given by:

$$\begin{aligned} BF_{cs}(E) &= \frac{\int_{\theta_{cs}} f(E \mid \theta_{cs},\, H_p,\, I) f(\theta_{cs} \mid H_p,\, I) d\theta_{cs}}{\int_{\theta_{cs}} f(E \mid \theta_{cs},\, H_d,\, I) f(\theta_{cs} \mid H_d,\, I) d\theta_{cs}} \\ &= \frac{\int_{\theta_{cs}} f(e_u \mid \theta_{cs},\, H_p,\, I) f(e_a,\, e_s \mid \theta_{cs},\, I) f(\theta_{cs} \mid I) d\theta_{cs}}{\int_{\theta_{cs}} f(e_u \mid \theta_{cs},\, H_d,\, I) f(e_a,\, e_s \mid \theta_{cs},\, I) f(\theta_{cs} \mid I) d\theta_{cs}}. \end{aligned}$$

and that with $\theta_{ss} = \{\mu_s,\, \sigma_s,\, \mu_M,\, \sigma_M,\, \sigma_a\}$ the specific source Bayes factor is given by:

$$\begin{aligned} BF_{ss}(E) &= \frac{\int_{\theta_{ss}} f(E \mid \theta_{ss},\, H_p,\, I) f(\theta_{ss} \mid H_p,\, I) d\theta_{ss}}{\int_{\theta_{ss}} f(E \mid \theta_{ss},\, H_d,\, I) f(\theta_{ss} \mid H_d,\, I) d\theta_{ss}} \\ &= \frac{\int_{\theta_{ss}} f(e_u \mid \theta_{ss},\, H_p,\, I) f(e_a,\, e_s \mid \theta_{ss},\, I) f(\theta_{ss} \mid I) d\theta_{ss}}{\int_{\theta_{ss}} f(e_u \mid \theta_{ss},\, H_d,\, I) f(e_a,\, e_s \mid \theta_{ss},\, I) f(\theta_{ss} \mid I) d\theta_{ss}}. \end{aligned}$$

Here we used that the likelihoods of e_a, e_s and e_{u_1} are independent of the hypotheses. Remembering Bayes' rule we can rewrite the likelihood of e_a and e_s as

$$f(e_a,\, e_s \mid \theta_{ss},\, I) = \frac{f(\theta_{ss} \mid e_a,\, e_s,\, I) f(e_a,\, e_s \mid I)}{f(\theta_{ss} \mid I)}.$$

When we plug this into the expression for the Bayes factor and we denote the numerator and denominator of the specific source Bayes Factor as $BF_{ss,1}$ and $BF_{ss,2}$ respectively, we get:

$$\begin{aligned} BF_{SS}(E) &= \frac{\int_{\theta_{ss}} f(e_u \mid \theta_{ss},\, H_p,\, I) \frac{f(\theta_{ss} \mid e_a,\, e_s,\, I) f(e_a,\, e_s \mid I)}{f(\theta_{ss} \mid I)} f(\theta_{ss} \mid I) d\theta_{ss}}{\int_{\theta_{ss}} f(e_u \mid \theta_{ss},\, H_d,\, I) \frac{f(\theta_{ss} \mid e_a,\, e_s,\, I) f(e_a,\, e_s \mid I)}{f(\theta_{ss} \mid I)} f(\theta_{ss} \mid I) d\theta_{ss}} \\ &= \frac{f(e_a,\, e_s \mid I) \int_{\theta_{ss}} f(e_u \mid \theta_{ss},\, H_p,\, I) f(\theta_{ss} \mid e_a,\, e_s,\, I) d\theta_{ss}}{f(e_a,\, e_s \mid I) \int_{\theta_{ss}} f(e_u \mid \theta_{ss},\, H_d,\, I) f(\theta_{ss} \mid e_a,\, e_s,\, I) d\theta_{ss}} \\ &= \frac{\int_{\theta_{ss}} f(e_u \mid \theta_{ss},\, H_p,\, I) f(\theta_{ss} \mid e_a,\, e_s,\, I) d\theta_{ss}}{\int_{\theta_{ss}} f(e_u \mid \theta_{ss},\, H_d,\, I) f(\theta_{ss} \mid e_a,\, e_s,\, I) d\theta_{ss}} \\ &= \frac{BF_{SS,1}}{BF_{SS,2}}. \end{aligned} \tag{C1}$$

In exactly the same way, we can rewrite the likelihood of e_a and e_s in the common source

scenario and we get

$$BF_{CS}(E) = \frac{\int_{\theta_{cs}} f(e_u \mid \theta_{cs},\, H_p,\, I) f(\theta_{cs} \mid e_a,\, e_s,\, I) d\theta_{cs}}{\int_{\theta_{cs}} f(e_u \mid \theta_{cs},\, H_d,\, I) f(\theta_{cs} \mid e_a,\, e_s,\, I) d\theta_{cs}} = \frac{BF_{cs,1}}{BF_{cs,2}}. \tag{C2}$$

D Determining $f(e_u \mid H_d, \theta, I)$ for the two-level model

We know that the distribution of x_u under H_d is $N(\mu_{a_u}, \sigma_a^2)$ and the mean follows the distribution $N(\mu_M, \sigma_M^2)$, of which the parameters are sampled with MCMC, but we do not denote them with superscript (cs) or (ss) here, because the calculation might become unclear and messy. To determine

$$f(e_u \mid \theta, H_d) = \phi(\vec{x_u} \mid \mu_M, \sigma_M^2, \sigma_a^2, H_d),$$

we need to solve the integral

$$\phi(\vec{x_u} \mid \mu_M, \sigma_M^2, \sigma_a^2, H_d) = \int_{\mu_{a_u}} \phi(\vec{x_u} \mid \mu_{a_u}, \sigma_a^2)\phi(\mu_{a_u} \mid \mu_M, \sigma_M^2) d\mu_{a_u}. \tag{58}$$

The probability density function of a normal distribution with mean μ and standard deviation σ is given by $\phi(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$. Recall that $\vec{x_u}$ is a vector of 10 measurements with covariance matrix Σ_a with only σ_a^2 on the diagonal. We solve equation (58) as follows:

$$\begin{aligned}
&\int_{\mu_{a_u}} \phi(\vec{x_u} \mid \mu_{a_u}, \sigma_a^2)\phi(\mu_{a_u} \mid \mu_M, \sigma_M^2) d\mu_{a_u} \\
&\quad= \int_{\mu_{a_u}} \frac{1}{\sqrt{(2\pi\sigma_a^2)^{10}}} \exp\left(-\frac{1}{2}\frac{\sum_{j=1}^{10}(x_{u,j} - \mu_{a_u})^2}{\sigma_a^2}\right) \phi(\mu_{a_u} \mid \mu_M, \sigma_M^2) d\mu_{a_u} \\
&\quad= \frac{1}{\sqrt{(2\pi\sigma_a^2)^{10}}} \int_{\mu_{a_u}} \exp\left(-\frac{1}{2}\frac{\sum_{j=1}^{10}(x_{u,j} - \bar{x} + \bar{x} - \mu_{a_u})^2}{\sigma_a^2}\right) \phi(\mu_{a_u} \mid \mu_M, \sigma_M^2) d\mu_{a_u} \\
(*) &= \frac{1}{\sqrt{(2\pi\sigma_a^2)^{10}}} \int_{\mu_{a_u}} \exp\left(-\frac{\sum_{j=1}^{10}(\bar{x} - \mu_{a_u})^2}{2\sigma_a^2}\right) \exp\left(\frac{-\sum_{j=1}^{10}(\bar{x} - x_{u,j})^2}{2\sigma_a^2}\right) \phi(\mu_{a_u} \mid \mu_M, \sigma_M^2) d\mu_{a_u} \\
&\quad= \exp\left(\frac{-\sum_{j=1}^{10}(\bar{x} - x_{u,j})^2}{2\sigma_a^2}\right) \frac{1}{\sqrt{(2\pi\sigma_a^2)^9}} \int_{\mu_{a_u}} \frac{1}{\sqrt{2\pi}\sigma_a} \exp\left(-\frac{10(\bar{x} - \mu_{a_u})^2}{2\sigma_a^2}\right) \phi(\mu_{a_u} \mid \mu_M, \sigma_M^2) d\mu_{a_u} \\
&\quad= \exp\left(\frac{-\sum_{j=1}^{10}(\bar{x} - x_{u,j})^2}{2\sigma_a^2}\right) \frac{1}{\sqrt{(2\pi\sigma_a^2)^9}} \int_{\mu_{a_u}} \frac{1}{\sqrt{10}} \phi(\bar{x} \mid \mu_{a_u}, \sigma_a^2/10)\phi(\mu_{a_u} \mid \mu_M, \sigma_M^2) d\mu_{a_u} \\
(**) &= \exp\left(-\frac{\sum_{j=1}^{10}(\bar{x} - x_{u,j})^2}{2\sigma_a^2}\right) \frac{1}{\sqrt{10(2\pi\sigma_a^2)^9}} \int_{\mu_{a_u}} \phi(\bar{x} \mid \mu_{a_u}, \sigma_a^2/10)\phi(\mu_{a_u} \mid \mu_M, \sigma_M^2) d\mu_{a_u}, \\
&\quad= \exp\left(-\frac{\sum_{j=1}^{10}(\bar{x} - x_{u,j})^2}{2\sigma_a^2}\right) \frac{1}{\sqrt{10(2\pi\sigma_a^2)^9}} \phi\left(\bar{x} \mid \mu_M, \sigma_M^2 + \frac{\sigma_a^2}{10}\right),
\end{aligned} \tag{D1}$$

where ϕ is the Gaussian density, we let $\bar{x} = \frac{1}{10}\sum_{j=1}^{10} x_{u,j}$ and consequently $\bar{x} \sim N(\mu_{a_u}, \sigma_a^2/10)$. To clarify (*), we let $a = \mu_{a_u} - \bar{x}$ and $b_j = \bar{x} - x_{u,j}$. Then the following hold:

$$\sum_{j=1}^{10}(x_{u,j} - \bar{x} + \bar{x} - \mu_{a_u})^2 = \sum_{j=1}^{10}(a+b_j)^2 = \sum_{j=1}^{10} a^2 + \sum_{j=1}^{10} b_j^2 + \sum_{j=1}^{10} 2ab_j$$

$$\sum_{j=1}^{10} 2ab_j = 2a\sum_{j=1}^{10} b_j = 2a\sum_{j=1}^{10} \bar{x} - x_{u,j} = 2a(10\bar{x} - \sum_{j=1}^{10} x_{u,j}) = 0$$

$$\sum_{j=1}^{10}(x_{u,j} - \bar{x} + \bar{x} - \mu_{a_u})^2 = \sum_{j=1}^{10} a^2 + \sum_{j=1}^{10} b_j^2 = \sum_{j=1}^{10}(\mu_{a_u} - \bar{x})^2 + \sum_{j=1}^{10}(\bar{x} - x_{u,j})^2$$

With the proof in appendix E, we know that the integral in (**) is equal to the density of a random variable with variance equal to the sum of the two variances, so in this case it's equal to: $\phi(\bar{x} \mid \mu_M, \sigma_M^2 + \sigma_a^2/10)$.

E Integral of two Gaussian probability densities

Suppose $x(\mu,\sigma_1^2)$ and $y \sim N(x,\sigma_2^2)$, and we want to calculate the following integral:

$$\frac{1}{\sqrt{2\pi}\sigma_1}\frac{1}{\sqrt{2\pi}\sigma_2}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma_1^2}} e^{-\frac{1}{2}\frac{(y-x)^2}{\sigma_2^2}}\,dx.$$

We rewrite it and complete the square to get the following:

$$\begin{aligned}
\frac{1}{\sqrt{2\pi}\sigma_1}\frac{1}{\sqrt{2\pi}\sigma_2}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma_1^2}} e^{-\frac{1}{2}\frac{(y-x)^2}{\sigma_2^2}}\,dx &= \frac{1}{2\pi\sigma_1\sigma_2}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{(x-\mu)^2}{\sigma_1^2}+\frac{(y-x)^2}{\sigma_2^2}\right)}dx \\
&= \frac{1}{2\pi\sigma_1\sigma_2}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{\sigma_2^2(x-\mu)^2+\sigma_1^2(y-x)^2}{\sigma_1^2\sigma_2^2}\right)}dx \\
(***) &= \frac{1}{2\pi\sigma_1\sigma_2}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{z}{\sigma_1^2\sigma_2^2}\right)}dx.
\end{aligned}$$

Let us rewrite the numerator of the exponent:

$$\begin{aligned}
z &= \sigma_2^2(x-\mu)^2+\sigma_1^2(y-x)^2 \\
&= \sigma_2^2(x^2+\mu^2-2\mu x)+\sigma_1^2(y^2+x^2-2yx) \\
&= x^2(\sigma_1^2+\sigma_2^2)-2x(\sigma_2^2\mu+\sigma_1^2 y)+\sigma_1^2y^2+\sigma_2^2\mu^2 \\
&= (\sigma_1^2+\sigma_2^2)\left[x^2-2\frac{\sigma_2^2\mu+\sigma_1^2 y}{\sigma_1^2+\sigma_2^2}x\right]+\sigma_1^2y^2+\sigma_2^2\mu^2 \\
&= (\sigma_1^2+\sigma_2^2)\left[\left(x-\frac{\sigma_2^2\mu+\sigma_1^2 y}{\sigma_1^2+\sigma_2^2}\right)^2-\frac{(\sigma_2^2\mu+\sigma_1^2 y)^2}{(\sigma_1^2+\sigma_2^2)^2}\right]+\sigma_1^2y^2+\sigma_2^2\mu^2 \\
&= (\sigma_1^2+\sigma_2^2)\left(x-\frac{\sigma_2^2\mu+\sigma_1^2 y}{\sigma_1^2+\sigma_2^2}\right)^2+\left[\sigma_1^2y^2+\sigma_2^2\mu^2-\frac{(\sigma_2^2\mu+\sigma_1^2 y)^2}{\sigma_1^2+\sigma_2^2}\right] \\
&= (\sigma_1^2+\sigma_2^2)\left(x-\frac{\sigma_2^2\mu+\sigma_1^2 y}{\sigma_1^2+\sigma_2^2}\right)^2+\left[\frac{(\sigma_1^2y^2+\sigma_2^2\mu^2)(\sigma_1^2+\sigma_2^2)}{\sigma_1^2+\sigma_2^2}-\frac{(\sigma_2^2\mu+\sigma_1^2 y)^2}{\sigma_1^2+\sigma_2^2}\right] \\
&= (\sigma_1^2+\sigma_2^2)\left(x-\frac{\sigma_2^2\mu+\sigma_1^2 y}{\sigma_1^2+\sigma_2^2}\right)^2+\left[\frac{\sigma_1^2\sigma_2^2(\mu^2+y^2-2\mu y)}{\sigma_1^2+\sigma_2^2}\right] \\
&= (\sigma_1^2+\sigma_2^2)\left(x-\frac{\sigma_2^2\mu+\sigma_1^2 y}{\sigma_1^2+\sigma_2^2}\right)^2+\left[\frac{\sigma_1^2\sigma_2^2(y-\mu)^2}{\sigma_1^2+\sigma_2^2}\right].
\end{aligned}$$

Plugging it back into the integral, we get:

$$
\begin{aligned}
(***) &= \frac{1}{2\pi\sigma_1\sigma_2}\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2}\left(\frac{(\sigma_1^2+\sigma_2^2)}{\sigma_1^2\sigma_2^2}\left(x-\frac{\sigma_2^2\mu+\sigma_1^2 y}{\sigma_1^2+\sigma_2^2}\right)^2+\frac{(y-\mu)^2}{\sigma_1^2+\sigma_2^2}\right)\right)dx \\
&= \frac{1}{2\pi\sigma_1\sigma_2}\exp\left(-\frac{1}{2}\frac{(y-\mu)^2}{\sigma_1^2+\sigma_2^2}\right)\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2}\left(\frac{(\sigma_1^2+\sigma_2^2)}{\sigma_1^2\sigma_2^2}\left(x-\frac{\sigma_2^2\mu+\sigma_1^2 y}{\sigma_1^2+\sigma_2^2}\right)^2\right)\right)dx \\
(****) &= \frac{1}{2\pi\sigma_1\sigma_2}\exp\left(-\frac{1}{2}\frac{(y-\mu)^2}{\sigma_1^2+\sigma_2^2}\right)\int_{-\infty}^{\infty}e^{-\frac{1}{2}\frac{(x-\nu)^2}{\tau^2}}dx,
\end{aligned}
$$

where

$$\nu = \frac{\sigma_2^2\mu+\sigma_1^2 y}{\sigma_1^2+\sigma_2^2} \quad \text{and} \quad \tau^2 = \frac{\sigma_1^2\sigma_2^2}{(\sigma_1^2+\sigma_2^2)}.$$

Since all densities integrate to 1, we get

$$\int_{-\infty}^{\infty}e^{-\frac{1}{2}\frac{(x-\nu)^2}{\tau^2}} = \sqrt{2\pi}\tau = \sqrt{2\pi}\frac{\sigma_1\sigma_2}{\sqrt{\sigma_1^2+\sigma_2^2}}$$

and therefore

$$(****) = \frac{1}{2\pi\sigma_1\sigma_2}\exp\left(-\frac{1}{2}\frac{(y-\mu)^2}{\sigma_1^2+\sigma_2^2}\right)\sqrt{2\pi}\frac{\sigma_1\sigma_2}{\sqrt{\sigma_1^2+\sigma_2^2}} = \frac{1}{\sqrt{2\pi}\sqrt{\sigma_1^2+\sigma_2^2}}\exp\left(-\frac{1}{2}\frac{(y-\mu)^2}{\sigma_1^2+\sigma_2^2}\right),$$

which is exactly the probability density function of a Gaussian random variable with mean μ and variance $\sigma_1^2+\sigma_2^2$.

F Additional Plots for the two-level Model

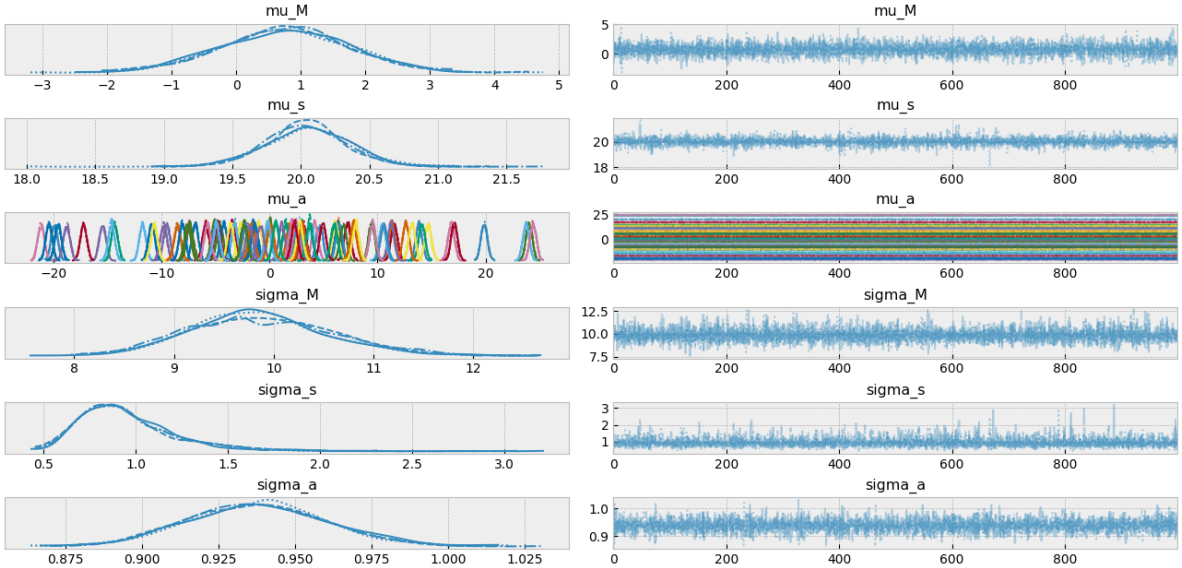


Figure 22: Trace plot for the specific source model Markov Chain Monte Carlo sampler (with true value of $\mu_s = 20$).

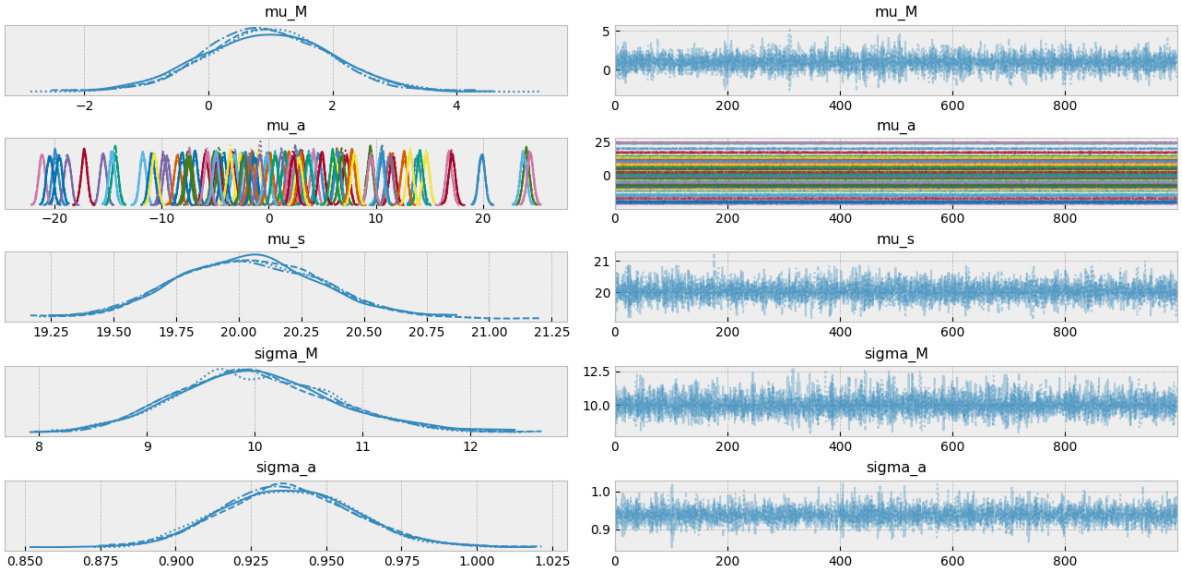
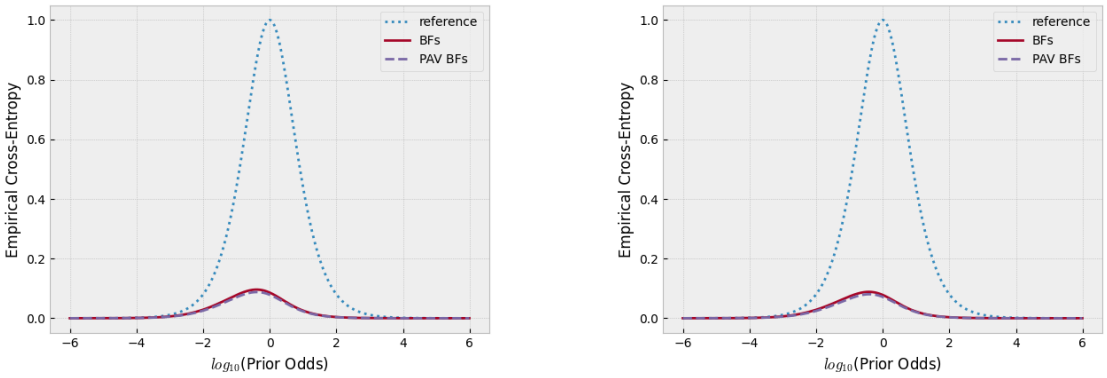
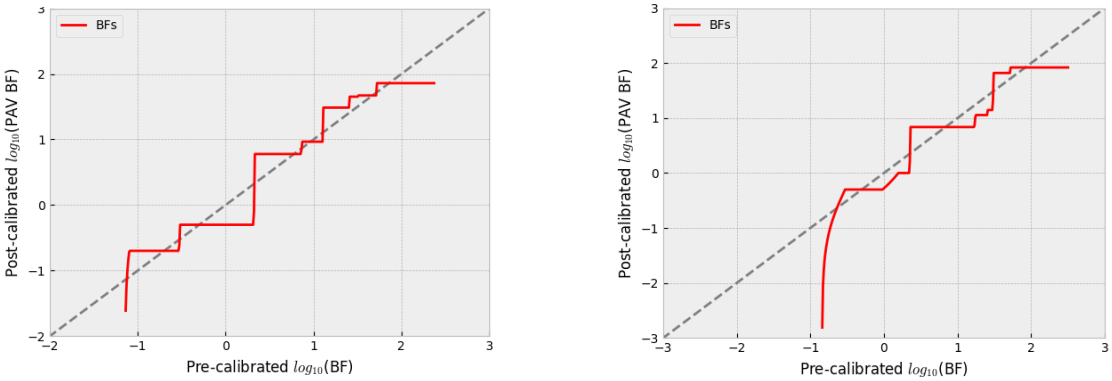


Figure 23: Trace plot for the common source model Markov Chain Monte Carlo sampler (with true value of $\mu_s = 20$).



(a) Specific Source (b) Common Source

Figure 24: ECE plots for the specific source and common source Bayes Factors (with true value of $\mu_s = 20$).



(a) Specific Source (b) Common Source

Figure 25: PAV transforms of the specific source and common source Bayes Factors (with true value of $\mu_s = 20$).

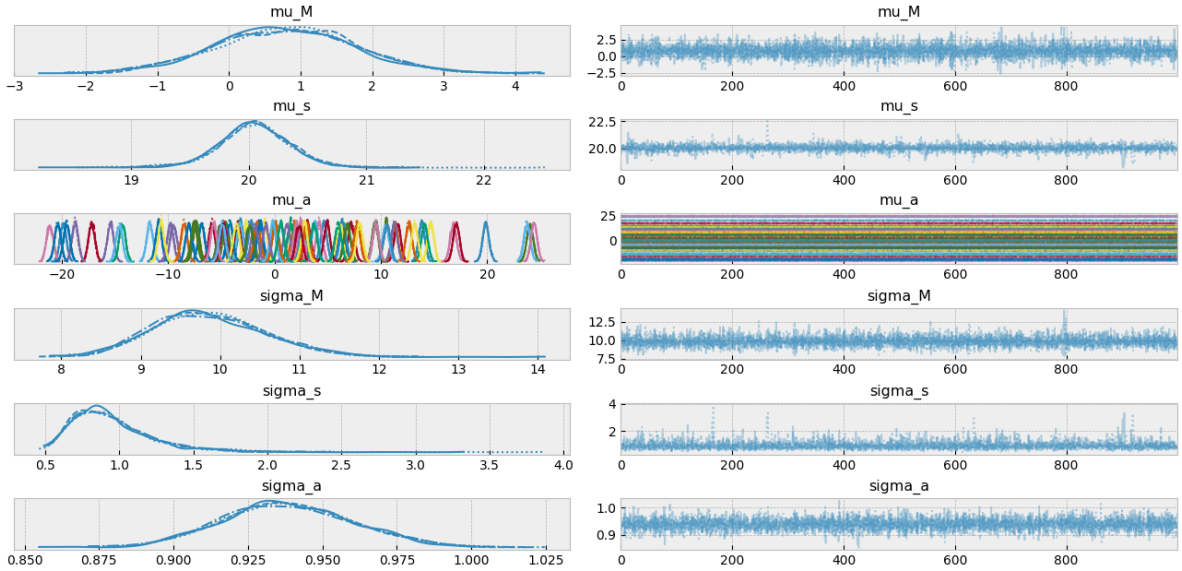


Figure 26: Trace plot for the specific source model Markov Chain Monte Carlo sampler with wider priors (with true value of $\mu_s = 20$).

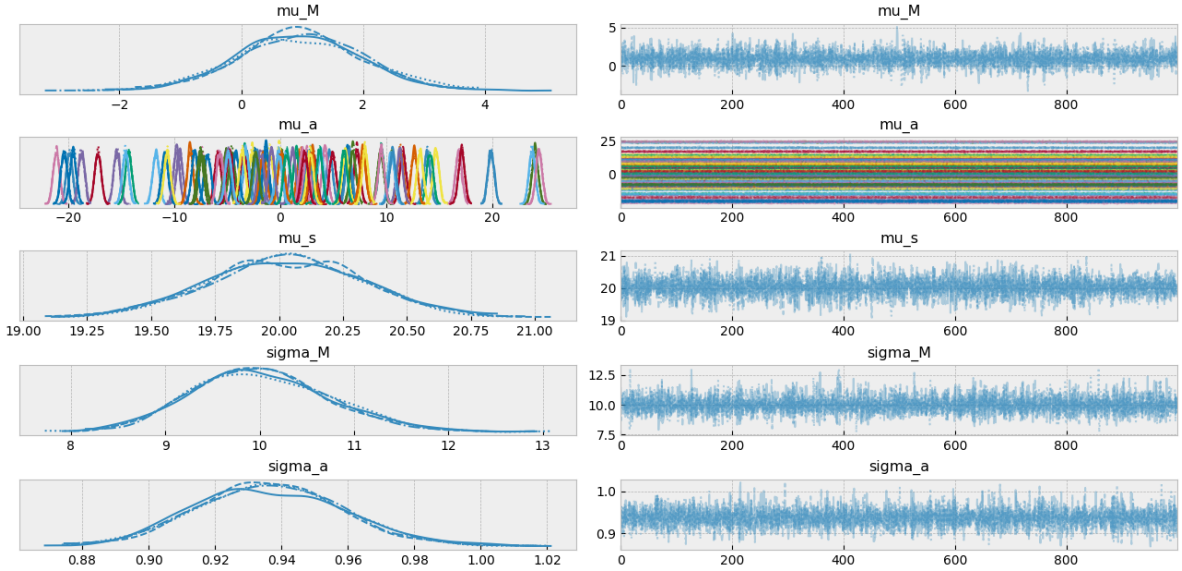
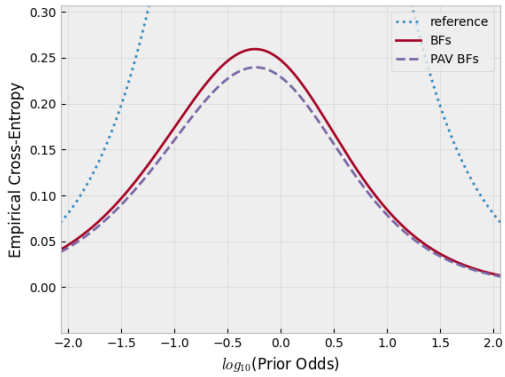
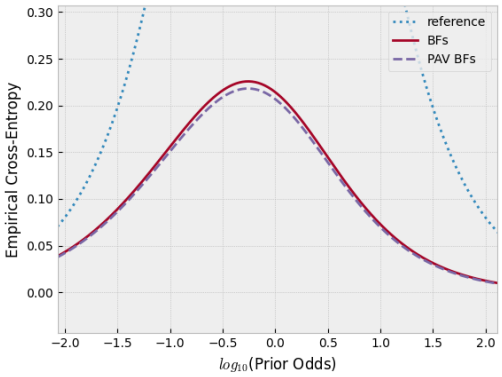


Figure 27: Trace plot for the common source model Markov Chain Monte Carlo sampler with wider priors (with true value of $\mu_s = 20$).

G Validation Plots for the Sensitivity Analysis of the two-level Model

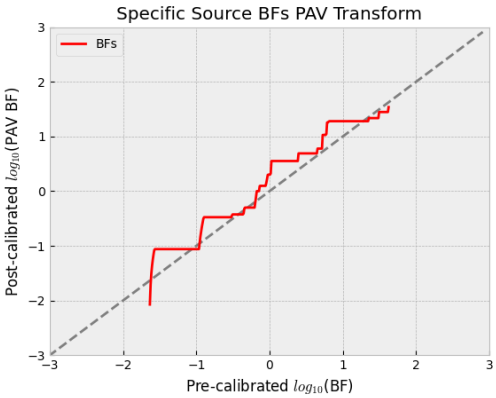


(a) Specific Source

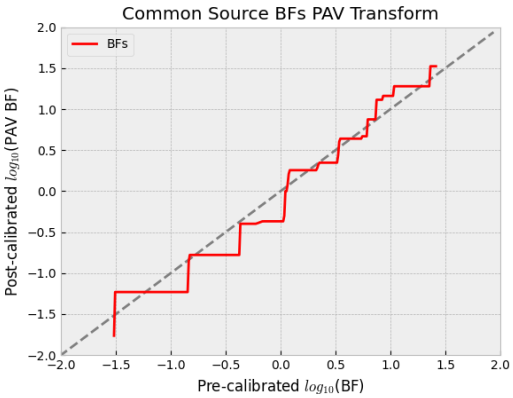


(b) Common Source

Figure 28: ECE plots for the specific source and common source Bayes Factors for extra analysis 1, where $\mu_s \sim N(\mu_M, \sigma^2_M)$.

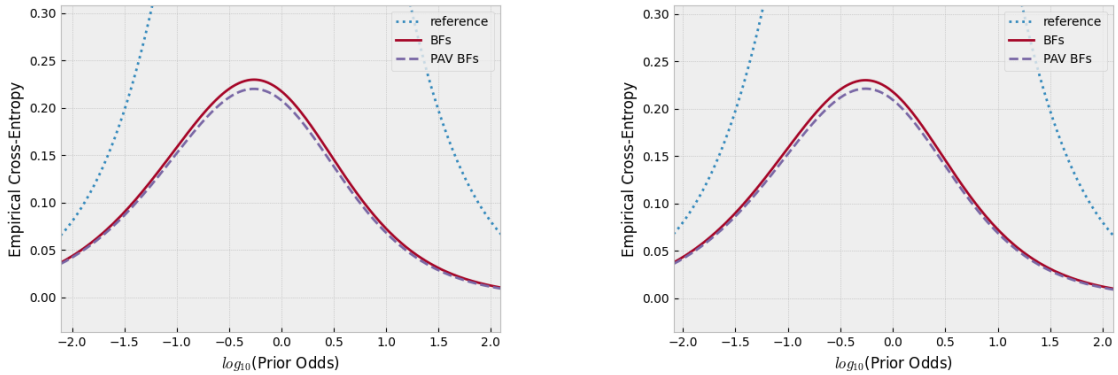


(a) Specific Source



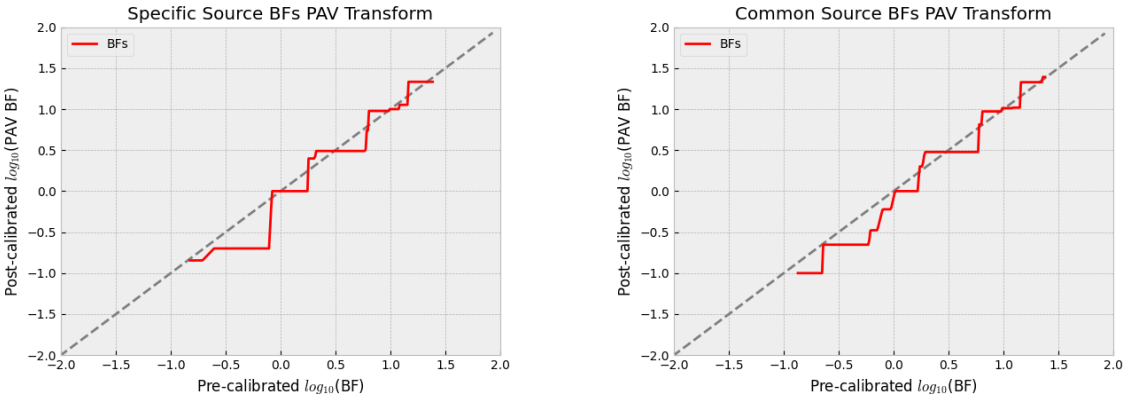
(b) Common Source

Figure 29: PAV transforms of the specific source and common source Bayes Factorsfor extra analysis 1, where $\mu_s \sim N(\mu_M, \sigma^2_M)$.



(a) Specific Source (b) Common Source

Figure 30: ECE plots for the specific source and common source Bayes Factors for extra analysis 2, where $\sigma_s = \sigma_a$.



(a) Specific Source (b) Common Source

Figure 31: PAV transforms of the specific source and common source Bayes Factors for extra analysis 2, where $\sigma_s = \sigma_a$.