Generalized fiducial factor: an alternative to the Bayes factor for forensic identification of source problems

Jonathan P Williams[(1)], Danica M Ommen[(2)], Jan Hannig[(3)]

North Carolina State University[(1)]
Iowa State University[(2)]
University of North Carolina at Chapel Hill[(3)]

Abstract

One formulation of forensic identification of source problems is to determine the source of trace evidence, for instance, glass fragments found on a suspect for a crime. The current state of the science is to compute a Bayes factor (BF) comparing the marginal distribution of measurements of trace evidence under two competing propositions for whether or not the trace evidence on the suspect originated from the crime scene. The obvious problem with such an approach is the ability to tailor the prior distributions (placed on the features/parameters of the statistical model for the measurements of trace evidence) in favor of the defense or prosecution, which is further complicated by the fact that the typical number of measurements of trace evidence is typically sufficiently small that prior information (which arguably does not exist) has a strong influence on the value of the BF. To remedy this problem of prior specification and choice, we develop an alternative to the BF, within the framework of generalized fiducial inference (GFI), that we term a *generalized fiducial factor* (GFF). GFI is an alternative paradigm for statistical inference allows for a prior-free, probabilistic approach to forensic identification of source problems. Building on developments over the past decade on the theory of GFI, we define, construct, and investigate properties of this first ever GFF. We present empirical analyses, on synthetic data and on real Netherlands Forensic Institute (NFI) casework data, of the GFF performance. Furthermore, we demonstrate empirically, on the same synthetic and real casework data, deficiencies in the BF and likelihood ratio (LR) approaches.

Keywords: Bayes factor; generalized fiducial inference; likelihood ratio

Running title: Generalized fiducial factor for forensic identification of source problems

Corresponding Author: Jonathan P Williams, `jwilli27@ncsu.edu`

1 Introduction

The adversarial nature of the criminal courtroom is extraordinarily troublesome in the context of Bayesian prior specification and choice. In its purest form, subjectivist Bayesian theory (Savage 1961, Lindley 1972) only admits prior probability distributions that reflect genuine beliefs about unknown features of a posited statistical model. However, in the criminal courtroom setting there are inherently two sets of such prior probability distributions, one model representing the belief of the prosecution and one model representing the belief of the defense. Further, given the high stakes nature of the outcome of a criminal court proceeding it is not hard to imagine that the subjectivist Bayesian inference from the evidence provided could lead to an extreme answer favoring either the prosecution or the defense, depending on which prior distribution is assumed for the statistical model features/parameters.

Historically, the alternative to subjectivist Bayesian theory is to consider a class of *objective* prior distributions. The problem with this approach is how to define *objective* in this context, and how to determine if the *objective* prior tends to favor the prosecution or the defense. The critical question focuses on whether Bayesian methodology is actually appropriate for the criminal courtroom setting involving beliefs of expert witnesses (i.e., not only appropriate for each individual juror). As statisticians, we have a responsibility to assess whether the methodological assumptions are safe and reliable. To this end, we investigate a particular class of problems commonly referred to as forensic identification of source problems, and we motivate our work with a real data set of glass fragments that was gathered from 10 years of casework by the NFI (van Es et al. 2017).

Several approaches for assigning value to forensic evidence have been explored, including the Two-Stage approach (Parker 1966, Evett 1977), LRs with Bayesian treatment of parameter uncertainty (Lindley 1977, Evett 1986, Aitken & Lucy 2004) or with maximum likelihood estimates (MLE) of parameters (Grove 1980, Ommen 2017), as well as score-based approaches (Gonzalez-Rodriguez et al. 2005, Egli et al. 2006, Gonzalez-Rodriguez et al. 2006, Neumann et al. 2007, Bolck et al. 2009, Hepler et al. 2012). The BF approach is the most commonly recommended among European countries (ENFSI 2015, Taroni et al. 2016, Berger & Slooten 2016, Biedermann et al. 2016), while a non-Bayesian approach is often recommended in the US (Lund & Iyer 2017, Swofford et al. 2018, Kafadar 2018). More precisely, Lund & Iyer (2017) does not object to Bayesian approaches, but does object to providing a single numerical value as the answer (regardless of whether a Bayesian or non-Bayesian method is used). The question that Lund & Iyer (2017) focus on is "what do you really know" versus what you are claiming to know (using prior information).

The gist of the LR approaches is to compare the probability of observing the evidence under two competing (and collectively exhaustive) explanations for how the evidence was generated. The Two-Stage approach, as it is most commonly presented, relies on statistical significance testing to compare two pieces of evidence; first to determine whether the evidence can be considered a "match," and then to compare to other sources to determine how many others might also "match". This approach is not directly comparable to the recommended LR approaches (Shafer 1982), and will likely come under scrutiny due to the movement away from significance testing for applications with "high-stakes" decisions (Wasserstein & Lazar 2016). The score-based likelihood ratio (SLR) approaches evolved from difficulties with the LR approaches for high-dimensional pattern and impression evidence (such as fingerprints, footwear, firearms, and handwriting evidence). These SLR approaches rely on extensive training datasets consisting of pairwise comparison scores between evidential objects, and these scores can be created in

a variety of different ways (Hepler et al. 2012, Neumann et al. 2020, Neumann & Ausdemore 2020). Again, this approach is not directly comparable to the recommended LR approaches due to the focus on modeling pairwise comparison scores as opposed to the features of one single object (Neumann et al. 2020). Due to the expressed concerns with the Two-Stage and SLR approaches, we will not consider these in this article.

Our contributions are the following. First and most fundamentally, we develop methodology for a new solution to forensic identification of source problems based on the GFI approach (Hannig et al. 2016). It has been shown in the literature that GFI is asymptotically valid in the sense of Bernstein von-Mises type theory (again, see Hannig et al. (2016)). Second, we illustrate empirically via simulating the real NFI casework data that the BF can yield remarkably different answers when the priors reflect the prosecution instead of the defense hypotheses and vice versa, and that the BF values can never be calibrated to reflect the strength of evidence that they convey. Our empirical results demonstrate very transparently that the degree to which the BF varies is more than enough to change the narrative of presented forensic evidence in a courtroom to the extent that a jury decision could conceivably be contrived. Furthermore, an alternative LR statistic for this application is numerically unstable and poorly calibrated to these data.

GFI is a prior-free approach to estimating a posterior distribution which reflects the uncertainty associated with unknown model parameters. We use GFI to define and construct the first ever GFF, particularly for application to statistical inference for forensic identification of source problems. Moreover, we demonstrate in a real NFI data simulation that the GFF, which does not rely on prior specification, is able to provide meaningful, consistent, and well calibrated inference. We make our R code and documentation for implementing the GFF publicly available at `https://jonathanpw.github.io/software.html`. The GFF can loosely be interpreted by analogy to a BF for particular choices of objective, data-driven priors, but the approach is justified independently of such interpretation. However, the GFI, and by extension the GFF, approach has principled foundational roots in statistical theory. We provide a gentle introduction to GFI prior to our construction of the GFF.

The organization of the paper is as follows. Section 2 precisely defines and describes the context of forensic identification of source problems. The real data is described and references are provided in Section 3.5. Section 3 introduces the central notions for GFI, provides a brief overview of the established theory, and proceeds by deriving the necessary components for the GFF in the context of forensic identification of source for glass fragment data. Thereafter, the main empirical results of the paper are presented in Section 4. Finally, concluding remarks are provided in closing, and an appendix accounts specific details for the BF and LR. The R code, along with a bash workflow file for reproducing all of our results is available at `https://jonathanpw.github.io/research.html`.

2 Motivating application

The motivation for the development of methodology for a GFF is the adversarial courtroom setting in which subjectivist BFs become problematic. We focus our attention on the particular class of forensic identification of source problems. The basic premise for such problems is that there is a crime that occurred at a specified location, and some standard materials (e.g., blood, weapons, gunpowder, glass fragments, etc.) were found at the scene of the crime. Next, a suspect for the crime is identified and is found with these standard materials. For example, glass fragments might be found at both the crime scene and fixed to the clothes of the suspect. Perhaps the glass fragments are tiny, but nonetheless can be analyzed for chemical composition.

Then an important question involves assessing how likely it is that the glass fragments on the suspect originated from the window at the crime scene. Strong evidence suggesting this is the case makes a compelling story linking the suspect to the crime.

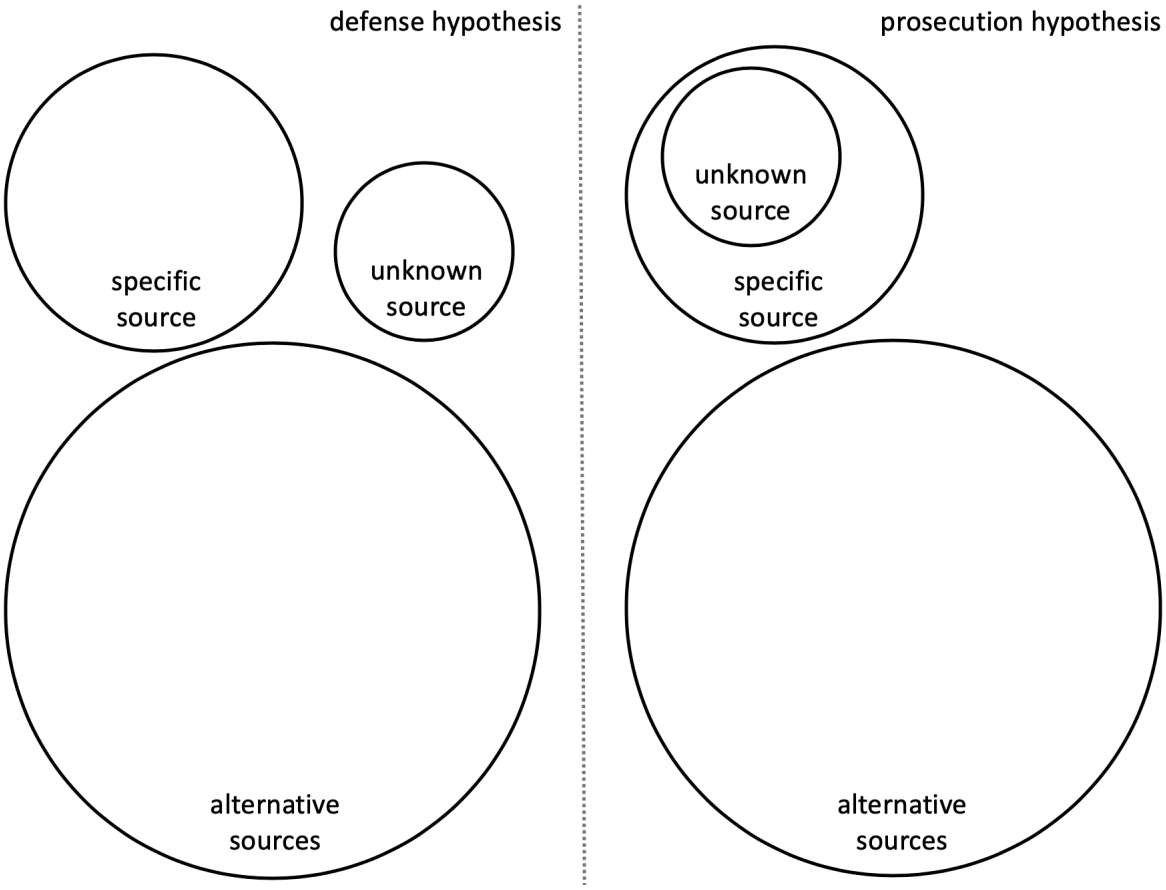

Figure 1: Graphical description of the likely relationship between the specific, unknown, and alternative sources.

Within the context of forensic identification of source problems, we consider the following framework for constructing the competing hypotheses, sometimes referred to as the *specific source* formulation (Ommen & Saunders 2019). In this formulation, material evidence such as trace elements of glass fragments found on a suspect are regarded as having been generated from either the *specific source* or some *alternative source*. In the case of glass evidence, the data gathered from the suspect is regarded as having been generated from an *unknown source* (either the specific source at the crime scene or an alternative source often characterized by a background database), and the competing hypotheses are

H_p : The unknown source is the specific source.

H_d : The unknown source is not the specific source.

We confine the rest of our exposition to modeling evidence arising from trace elements of glass fragments. Figure 1 gives a visual illustration of the likely relationship between the sources of data in the identification of source problem. The alternative sources characterize a large database of panes of glass found in windows and doors used to describe the variation of trace element compositions found between and within panes of glass. Glass fragments from a

pane found at a specific source such as a crime scene can be uniquely characterized based on the composition and variation of their trace elements. When glass fragments are discovered on a suspect for a crime (i.e., the unknown source data), an analyst can compare the composition and variation of its trace elements to that of glass found at the specific source (i.e., the crime scene) and that of all types of glass that has been documented in the alternative source database. This logical framework lends itself to modeling the alternative source data as a random effect, where the random effects component describes variation of trace elements between panes. In Section 3 we formulate the construction of these data generating models.

Unfortunately, forensic databases are not sufficiently exhaustive for it to be realistic to assume that all sources are represented in the alternative source data. Nonetheless, the meaningful question for the forensic identification of source problem remains whether the unknown source data are consistent with the specific source data. In the sections that follow we develop and evaluate statistical methodology to address this question. Further, we design a simulation study consistent with Figure 1 from the real NFI casework data to investigate and assess our methods.

3 Methodology

The motivation for GFI is to construct prior-free probabilistic inference on meaningful parameters in a data generating model. An overview of the ideas, common examples, and theoretical guarantees for GFI is presented in Hannig et al. (2016). The formal definition of a GF distribution begins with a data generating equation G for the realization of data Y, depending on some underlying pivotal quantity U and some unknown fixed parameters θ. That is,

$$Y = G(U, \theta),$$

where G is deterministic and U is a random variable whose distribution is completely known. The idea for GFI is to invert the function G to solve for the unknown parameters, and then switch the roles of θ and the observed data y to construct a distributional estimator for θ that inherits the uncertainty associated with U.

More precisely, consider the following inverse problem,

$$Q_y(u^\star) = \underset{\theta^\star}{\operatorname{argmin}} \|y - G(u^\star, \theta^\star)\|.$$

For $\epsilon > 0$, define the random variable $\theta^\star_\epsilon = Q_y(U^\star_\epsilon)$, where $U^\star_\epsilon$ has the same distribution as U truncated to the set

$$\mathcal{C}_\epsilon = \{U^\star_\epsilon : \|y - G(U^\star_\epsilon, \theta^\star_\epsilon)\| = \|y - G(U^\star_\epsilon, Q_y(U^\star_\epsilon))\| \leq \epsilon\}.$$

Then assuming that the random variables $\theta^\star_\epsilon$ converge in distribution as $\epsilon \to 0$, the GF distribution is defined as the limiting distribution $\theta^\star = \lim_{\epsilon \to 0} \theta^\star_\epsilon$. Notice that the fiducial distribution of $\theta^\star$ depends on the observed data y. The intuition for understanding this distribution is similar to that for approximate Bayesian computations (Beaumont et al. 2002).

Moreover, under certain conditions applicable to many practical settings (Hannig et al. 2016), the GF distribution can be computed as

$$r(\theta \mid y) = \frac{f(y \mid \theta) J(y, \theta)}{\int_\Theta f(y \mid \tilde{\theta}) J(y, \tilde{\theta}) \, d\tilde{\theta}}, \tag{1}$$

where $f(y \mid \theta)$ is the likelihood function, and

$$J(y,\theta) := D\Big(\nabla_\theta G(u,\theta)\Big|_{u=G^{-1}(y,\theta)}\Big), \tag{2}$$

with $D(A) := \sqrt{\det(A'A/n)}$, where n is the number of samples observed (dimension of y). The function $J(y,\theta)$ is a Jacobian-like quantity that results from inverting the data generating equation $y = G(U,\theta)$. Viewed from another perspective, (1) defines a posterior-like distribution for a class of data-driven, objective priors. A variety of classes of objective (or non-informativie, weakly informative, etc.) priors are well-accepted in the literature and in practice as both meaningful and useful inferential tools (Jeffreys 1946, Bernardo 1979, Mukerjee & Reid 1999, Gelman et al. 2008, Staicu & Reid 2008, Berger et al. 2009, Martin et al. 2019). In fact, any prior distribution that is constructed for any reason other than to reflect the true state of the prior knowledge is not properly Bayesian. In the following two subsections we use (1) to construct GF distributions for the forensic identification of source problems described in the previous section.

3.1 GF distribution of specific source data

For the glass fragments found at the specific source, let m denote the number of measurements of the concentration of p elements, and record the measurements as a column vector $y_{s,k} \in \mathbb{R}^p$ for $k \in \{1,\ldots,m\}$. Then, assuming a multivariate Gaussian data generating equation, for $k \in \{1,\ldots,m\}$,

$$Y_{s,k} = G\big(Z_k, (\mu_s, A)\big) = \mu_s + AZ_k, \tag{3}$$

where $Z_k \sim \mathrm{N}_p(0, I_p)$ and A is nonsingular. The GF distribution of (μ_s, A) then has the form,

$$r_s\big(\mu_s, A \mid \{y_{s,k}\}\big) = \frac{q_s\big(\mu_s, A \mid \{y_{s,k}\}\big)}{c_s},$$

where $q_s\big(\mu_s, A \mid \{y_{s,k}\}\big) := f_s\big(\{y_{s,k}\} \mid \mu_s, A\big) \cdot J_s\big(\{y_{s,k}\}, (\mu_s, A)\big)$ is the unnormalized GF density with normalizing constant c_s, $f_s(\cdot \mid \mu_s, A)$ is a multivariate Gaussian density, and the Jacobian $J_s\big(\{y_{s,k}\}, (\mu_s, A)\big)$ is computed as follows. As in Shi et al. (2017), denoting $w := (y'_{s,1}, \ldots, y'_{s,m})'$ and applying definition (2) gives $J_s\big(\{y_{s,k}\}, (\mu_s, A)\big)$, where

$$\nabla_{(\mu_s,A)} G = \begin{pmatrix} \frac{\partial w_1}{\partial (\mu_s)_1} & \cdots & \frac{\partial w_1}{\partial (\mu_s)_p} & \frac{\partial w_1}{\partial A_{11}} & \frac{\partial w_1}{\partial A_{12}} & \cdots & \frac{\partial w_1}{\partial A_{pp}} \\ \frac{\partial w_2}{\partial (\mu_s)_1} & \cdots & \frac{\partial w_2}{\partial (\mu_s)_p} & \frac{\partial w_2}{\partial A_{11}} & \frac{\partial w_2}{\partial A_{12}} & \cdots & \frac{\partial w_2}{\partial A_{pp}} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial w_{mp}}{\partial (\mu_s)_1} & \cdots & \frac{\partial w_{mp}}{\partial (\mu_s)_p} & \frac{\partial w_{mp}}{\partial A_{11}} & \frac{\partial w_{mp}}{\partial A_{12}} & \cdots & \frac{\partial w_{mp}}{\partial A_{pp}} \end{pmatrix} = \begin{pmatrix} I_p & I_p \otimes z'_1 \\ \vdots & \vdots \\ I_p & I_p \otimes z'_m \end{pmatrix}.$$

Rearranging rows of $\nabla_{(\mu_s,A)} G$ and denoting $\widetilde{U} := (z_1, \ldots, z_m)'$ simplifies the expression to,

$$\begin{aligned} J_s\big(\{y_{s,k}\}, (\mu_s, A)\big) &= \left| \begin{pmatrix} I_p \otimes 1'_m \\ I_p \otimes \widetilde{U}' \end{pmatrix} \begin{pmatrix} I_p \otimes 1_m & I_p \otimes \widetilde{U} \end{pmatrix} \right|^{\frac{1}{2}} m^{-\frac{p+p^2}{2}} \\ &= \left| \begin{pmatrix} I_p & 0 \\ 0 & I_p \otimes A^{-1} \end{pmatrix} \begin{pmatrix} mI_p & I_p \otimes 1'_m U \\ I_p \otimes U' 1_m & I_p \otimes U'U \end{pmatrix} \begin{pmatrix} I_p & 0 \\ 0 & I_p \otimes (A^{-1})' \end{pmatrix} \right|^{\frac{1}{2}} m^{-\frac{p+p^2}{2}}, \end{aligned}$$

where 1_m is an $m \times 1$ vector of ones, and $U := (y_{s,1} - \mu_s, \ldots, y_{s,m} - \mu_s)'$ so that $\widetilde{U} = U(A^{-1})'$. Thus,

$$q_s(\mu_s, A \mid \{y_{s,k}\}) = (2\pi)^{-\frac{mp}{2}} |AA'|^{-\frac{m+p}{2}} e^{-\frac{1}{2}\mathrm{tr}\left(S_s(AA')^{-1}\right)} \left| \begin{pmatrix} mI_p & I_p \otimes 1_m' U \\ I_p \otimes U' 1_m & I_p \otimes U'U \end{pmatrix} \right|^{\frac{1}{2}} m^{-\frac{p+p^2}{2}},$$

where

$$S_s := \sum_{k=1}^{m} (y_{s,k} - \mu_s)(y_{s,k} - \mu_s)'. \tag{4}$$

3.2 GF distribution of alternative source data

For the glass fragments available in the alternative sources, let m_i denote the number of measurements of the concentration of p elements for source $i \in \{1, \ldots, n\}$, where n is the total number of sources contained in the alternative source data. Record the p measurements as a column vector $y_{a,i,k} \in \mathbb{R}^p$ for $k \in \{1, \ldots, m_i\}$ and $i \in \{1, \ldots, n\}$. Then, consistent with the specific source setup in the previous section, we assume that the data from each source in the alternative source data set is generated from a multivariate Gaussian distribution (Zadora et al. 2013) with a unique mean vector $\mu_a + Bt_i$, where $\mu_a \in \mathbb{R}^p$ is a fixed effect, and $Bt_i \in \mathbb{R}^p$ is a draw from a multivariate T random effect with τ degrees of freedom and positive-definite covariance matrix BB' describing the variation in mean vectors over each source in the alternative source set. The heavy tails of the multivariate T distribution reflect the inherently large variation that is observed in element composition exhibited by different panes of glass, while the light tails of the multivariate Gaussian distribution reflect the relatively small variance in element composition found in a single pane of glass.

Accordingly, for $k \in \{1, \ldots, m_i\}$ and $i \in \{1, \ldots, n\}$,

$$Y_{a,i,k} = \mu_a + BT_i + CV_{i,k}, \tag{5}$$

where $V_{i,k} \sim \mathrm{N}_p(0, I_p)$, C is nonsingular, and $T_i \sim \mathrm{T}_\tau(0, I_p)$. Consequently, the GF distribution of (μ_a, B, C) can be expressed as,

$$\begin{aligned} r_a\big(\mu_a, B, C \mid \{y_{a,i,k}\}\big) &:= \frac{q_a\big(\mu_a, B, C \mid \{y_{a,i,k}\}\big)}{c_a} \\ &= \frac{1}{c_a} \int \cdots \int q_a\big(\mu_a, B, C, \{t_i\} \mid \{y_{a,i,k}\}\big)\ dt_1 \cdots dt_n \\ &= \frac{1}{c_a} \int \cdots \int q_a\big(\mu_a, B, C \mid \{t_i\}, \{y_{a,i,k}\}\big) f_{T_1}(t_1) \cdots f_{T_n}(t_n)\ dt_1 \cdots dt_n, \end{aligned}$$

where $q_a\big(\mu_a, B, C \mid \{t_i\}, \{y_{a,i,k}\}\big) = f_a\big(\{y_{a,i,k}\} \mid \mu_a, B, C, \{t_i\}\big) \cdot J_a\big(\{y_{a,i,k}\}, (\mu_a, B, C)\big)$ is the unnormalized GF density with normalizing constant c_a, and $f_a\big(\cdot \mid \mu_a, B, C, \{t_i\}\big)$ is a multivariate Gaussian density. To compute the Jacobian, as in the specific source derivation let $w := (y'_{a,1,1}, \ldots, y'_{a,1,m_1}, \ldots, y'_{a,n,1}, \ldots, y'_{a,n,m_n})'$, denote $N := \sum_{i=1}^{n} m_i$, and apply definition (2)

which gives $J_a\big(\{y_{a,i,k}\},(\mu_a,B,C)\big)$, where

$$\nabla_{(\mu_a,B,C)}G = \begin{pmatrix} I_p & I_p\otimes t_1' & I_p\otimes v_{1,1}' \\ \vdots & \vdots & \vdots \\ I_p & I_p\otimes t_1' & I_p\otimes v_{1,m_1}' \\ \vdots & \vdots & \vdots \\ I_p & I_p\otimes t_n' & I_p\otimes v_{n,1}' \\ \vdots & \vdots & \vdots \\ I_p & I_p\otimes t_n' & I_p\otimes v_{n,m_n}' \end{pmatrix}$$

Next, rearranging rows of $\nabla_{(\mu_a,B,C)}G$ gives,

$$\begin{aligned} J_a\big(\{y_{a,i,k}\},(\mu_a,B,C)\big) &= \left| \begin{pmatrix} I_p\otimes 1_N' \\ I_p\otimes W' \\ I_p\otimes \widetilde{Q}' \end{pmatrix} \begin{pmatrix} I_p\otimes 1_N & I_p\otimes W & I_p\otimes \widetilde{Q} \end{pmatrix} \right|^{\frac{1}{2}} N^{-\frac{p+2p^2}{2}} \\ &= \left| \begin{pmatrix} I_p & 0 & 0 \\ 0 & I_{p^2} & 0 \\ 0 & 0 & I_p\otimes (CC')^{-1} \end{pmatrix} \begin{pmatrix} NI_p & I_p\otimes 1_N'W & I_p\otimes 1_N'Q \\ I_p\otimes W'1_N & I_p\otimes W'W & I_p\otimes W'Q \\ I_p\otimes Q'1_N & I_p\otimes Q'W & I_p\otimes Q'Q \end{pmatrix} \right|^{\frac{1}{2}} N^{-\frac{p+2p^2}{2}}, \end{aligned}$$

where $W := \begin{pmatrix} 1_{m_1}\otimes t_1' \\ \vdots \\ 1_{m_n}\otimes t_n' \end{pmatrix}$, and

$$\widetilde{Q} := \begin{pmatrix} v_{1,1}' \\ \vdots \\ v_{1,m_1}' \\ \vdots \\ v_{n,1}' \\ \vdots \\ v_{n,m_n}' \end{pmatrix} = \underbrace{\begin{pmatrix} (y_{a,1,1}-\mu_a-Bt_1)' \\ \vdots \\ (y_{a,1,m_1}-\mu_a-Bt_1)' \\ \vdots \\ (y_{a,n,1}-\mu_a-Bt_n)' \\ \vdots \\ (y_{a,n,m_n}-\mu_a-Bt_n)' \end{pmatrix}}_{=:\,Q} (C^{-1})'.$$

Thus,

$$q_a(\mu_a,B,C\mid\{t_i\},\{y_{a,i,k}\}) = \frac{e^{-\frac{1}{2}\mathrm{tr}\left(S_a(CC')^{-1}\right)}}{(2\pi)^{\frac{pN}{2}}|CC'|^{\frac{N+p}{2}}N^{\frac{p+2p^2}{2}}} \left| \begin{pmatrix} NI_p & I_p\otimes 1_N'W & I_p\otimes 1_N'Q \\ I_p\otimes W'1_N & I_p\otimes W'W & I_p\otimes W'Q \\ I_p\otimes Q'1_N & I_p\otimes Q'W & I_p\otimes Q'Q \end{pmatrix} \right|^{\frac{1}{2}},$$

where

$$S_a := \sum_{i=1}^{n}\sum_{k=1}^{m_i}(y_{a,i,k}-\mu_a-Bt_i)(y_{a,i,k}-\mu_a-Bt_i)'. \tag{6}$$

3.3 Generalized fiducial factor

With the GF densities constructed for the specific source data in Section 3.1 and alternative source data in Section 3.2, it remains to construct the GFF from them. A key distinction between a BF and a GFF results from the fact that a prior distribution is necessarily independent

of the data while the Jacobian, which is the analogue for the prior in GFI, is a function of the data. To illustrate how this distinction breaks the construction of a BF, consider the data $y_{u,1}, \ldots, y_{u,m_u} \in \mathbb{R}^p$ from an unknown source, as described in Section 2 (i.e., m_u measurements of the concentration of p elements from glass fragments found on the suspect for a crime). Let M_s and M_a denote the specific and alternative source models, respectively, and for conciseness let $\theta_s := (\mu_s, A)$ corresponding to the parameters for the specific source model (described in section 3.1) and $\theta_a := (\mu_a, B, C)$ corresponding to the parameters for the alternative source model (described in section 3.2). Then the BF is defined as in Kass & Raftery (1995) as,

$$\text{BF} := \frac{\pi\big(\{y_{u,j}\} \mid M_s\big)}{\pi\big(\{y_{u,j}\} \mid M_a\big)} = \frac{\int \pi\big(\theta_s, \{y_{u,j}\} \mid M_s\big)\, d\theta_s}{\int \pi\big(\theta_a, \{y_{u,j}\} \mid M_a\big)\, d\theta_a} = \frac{\int f_s\big(\{y_{u,j}\} \mid \theta_s\big)\pi_s\big(\theta_s \mid \{y_{s,k}\}\big)\, d\theta_s}{\int f_a\big(\{y_{u,j}\} \mid \theta_a\big)\pi_a\big(\theta_a \mid \{y_{a,i,k}\}\big)\, d\theta_a}. \quad (7)$$

This last equality does not make sense in the GFI paradigm in the same way that it would not make sense for an improper prior.

The use of the conditional densities $\pi_s\big(\cdot \mid \{y_{s,k}\}\big)$ and $\pi_a\big(\cdot \mid \{y_{a,i,k}\}\big)$ requires them to be proper density functions (or at least integrable). Nonetheless, the GF densities $r_s\big(\cdot \mid \{y_{s,k}\}\big)$ and $r_a\big(\cdot \mid \{y_{a,i,k}\}\big)$ are proper density functions, and share similar large-sample properties in the sense of Bernstein von-Mises type theory. Hence, by analogy we define,

$$\text{GFF} := \frac{\int f_s\big(\{y_{u,j}\} \mid \theta_s\big) \cdot r_s\big(\theta_s \mid \{y_{s,k}\}\big)\, d\theta_s}{\int f_a\big(\{y_{u,j}\} \mid \theta_a\big) \cdot r_a\big(\theta_a \mid \{y_{a,i,k}\}\big)\, d\theta_a}, \quad (8)$$

and note the distinction from the BF. In the remaining sections of this paper, we demonstrate empirically that the defined GFF has both practical utility for the identification of source problem and overcomes limitations of the BF and LR approaches.

3.4 Remarks on computation

In this section we describe our approach to compute the GFF defined in (8) from actual data. Applying the derivations of the GF distributions from Sections 3.1 and 3.2 directly into (8) gives

$$\text{GFF} = \frac{\int\int f_s\big(\{y_{u,l}\} \mid \mu_s, A\big) \cdot r_s\big(\mu_s, A \mid \{y_{s,k}\}\big)\, d\mu_s\, dA}{E_{T_1,\ldots,T_{n+1}}\Big(\int\int\int f_a\big(\{y_{u,l}\} \mid \mu, B, C, t_{n+1}\big) \cdot \frac{1}{c_a} q_a\big(\mu_a, B, C \mid \{t_i\}, \{y_{a,i,k}\}\big)\, d\mu_a\, dB\, dC\Big)}.$$

The numerator is the expected value of $f_s\big(\{y_{u,l}\} \mid \mu_s, A\big)$ (i.e., the specific source likelihood evaluated for the unknown source data) with respect to the GF density for the specific source model. Accordingly, a natural estimate for this expected value is the average value of $f_s\big(\{y_{u,l}\} \mid \mu_s, A\big)$ over a GF sample of the parameters μ_s and A. We thus construct a random walk Metropolis-Hastings Markov chain Monte Carlo (MCMC) algorithm to estimate a GF sample of μ_s and A.

The denominator is computationally much more difficult to deal with due to the expectation over the random effect components $T_1, \ldots, T_{n+1}$. We have experimented with various strategies for importance sampling over all $T_1, \ldots, T_{n+1}$, but these samples result in very poor mixing within the MCMC algorithm to estimate the GF distribution of μ_a, B, and C. A prohibitively large number of importances samples of the $\{T_i\}$ are needed to properly identify BB' and CC'. Accordingly, we construct the following point estimators for $t_1, \ldots, t_n$.

First, construct the estimates,

$$\begin{aligned}
\widehat{\mu}_a &:= \frac{1}{N}\sum_{i=1}^{n}\sum_{k=1}^{m_i} y_{a,i,k} \\
\widehat{B}\widehat{B}' &:= \frac{1}{n-1}\sum_{i=1}^{n}\big(\bar{y}_{a,i,\cdot} - \widehat{\mu}_a\big)\big(\bar{y}_{a,i,\cdot} - \widehat{\mu}_a\big)' \\
\widehat{C}\widehat{C}' &:= \frac{1}{N-1}\sum_{i=1}^{n}\sum_{k=1}^{m_i}\big(y_{a,i,k} - \bar{y}_{a,i,\cdot}\big)\big(y_{a,i,k} - \bar{y}_{a,i,\cdot}\big)',
\end{aligned} \tag{9}$$

where $\bar{y}_{a,i,\cdot} := \frac{1}{m_i}\sum_{k=1}^{m_i}\bar{y}_{a,i,k}$ for each $i \in \{1,\ldots,n\}$ and $\widehat{B}$ and $\widehat{C}$ are triangular Cholesky decomposition factors. Then substituting these estimates into the data generating equation (5) yields the following repeated observations regression model in terms of unknown coefficients t_i,

$$Y_{a,i,k} - \widehat{\mu}_a = \widehat{B}t_i + \widehat{C}V_{i,k},$$

for $k \in \{1,\ldots,m_i\}$ and $i \in \{1,\ldots,n\}$. Averaging over each measurement k and rescaling the systems of equations gives the Gauss-Markov model,

$$\big(\widehat{C}\widehat{C}'\big)^{-\frac{1}{2}}\big(\bar{Y}_{a,i,\cdot} - \widehat{\mu}_a\big) = \big(\widehat{C}\widehat{C}'\big)^{-\frac{1}{2}}\widehat{B}t_i + \big(\widehat{C}\widehat{C}'\big)^{-\frac{1}{2}}\widehat{C}\left(\frac{1}{m_i}\sum_{k=1}^{m_i}V_{i,k}\right),$$

with the resulting least squares solution

$$\widehat{t_i} := \Big(\widehat{B}'\big(\widehat{C}\widehat{C}'\big)^{-1}\widehat{B}\Big)^{-1}\widehat{B}'\big(\widehat{C}\widehat{C}'\big)^{-1}\big(\bar{y}_{a,i,\cdot} - \widehat{\mu}_a\big),$$

for every source $i \in \{1,\ldots,n\}$ in the alternative source data set. Note that $\widehat{t_i}$ is a consistent estimator for t_i for large m_i.

Using the estimates $\{\widehat{t_i}\}$ we estimate the GFF as

$$\text{GFF} = \frac{\int\int f_s\big(\{y_{u,l}\} \mid \mu_s, A\big)\cdot r_s\big(\mu_s, A \mid \{y_{s,k}\}\big)\; d\mu_s\; dA}{\int\int\int E_{T_{n+1}}\Big(f_a\big(\{y_{u,l}\} \mid \mu, B, C, t_{n+1}\big)\Big)\cdot \frac{1}{c_a}q_a\big(\mu_a, B, C \mid \{\widehat{t_i}\}, \{y_{a,i,k}\}\big)\; d\mu_a\; dB\; dC},$$

where the expectation $E_{T_{n+1}}(\cdot)$ is estimated by evaluating the average of the integrand over some large number of importance samples of $T_{n+1} \sim \text{T}_5(0, I_p)$.

The computation of the ratio of marginal densities such as a BF or the GFF is a difficult endeavor and a well explored topic in the literature (Meng & Wong 1996, DiCiccio et al. 1997, Gelman & Meng 1998). Other popular approaches include importance, bridge, and path sampling (Gelman & Meng 1998), but we find that these methods nonetheless tend to require a fair amount of finesse and tailoring to a given data model. The remaining sections of this paper serve to evaluate the empirical performance of our proposed GFF, and to illustrate shortcomings in the BF and LR. The real data are described next

3.5 NFI casework data

The glass fragment data set that we investigate (van Es et al. 2017) was kindly received from the NFI, but the NFI was not further involved in this research. Currently, these data are available on request by emailing p.zoon@nfi.nl.

The data set consists of fragments from 979 unique windows from crime scenes spanning approximately 10 years of casework (van Es et al. 2017). Of the 979 sources, 659 are designated as training data and the remaining 320 as calibration data. Measurements of the concentration of 18 elements for three fragments are recorded for the glass corresponding to each crime scene window in the training data, for a total of 3×659 measurements. Measurements of the 18 elements for five fragments for each window in the calibration data are recorded, for a total of 5×320 measurements. As discussed in van Es et al. (2017), a meaningful subset of 10 of the 18 elements are considered. Further details of these data are documented in van Es et al. (2017).

In the context of our formulation (i.e., Figure 1), the training data corresponds to the alternative source data. We then separate the first three measurements of each source in the calibration data set to denote a set of specific source data (each set corresponding to one unique window as the specific source), and leave the remaining two measurements to comprise sets of unknown source data. Accordingly, we have 320 observed instances in which the unknown source is the specific source (i.e., the prosecution hypothesis, H_p), and 320×319 observed instances in which the unknown source is *not* the specific source (i.e., the defense hypothesis, H_d). We study our methods by simulating over these data and evaluating the performance of the GFF we construct, compared to the truth and compared to the BF and LR.

4 Empirical results

In the empirical analysis that follows, we first demonstrate that all three methods (GFF, BF, and LR) perform well on fully synthetic data simulated from the data generation equations (3) and (5) when there are many specific and unknown source data measurements available. Next, we illustrate the performance of all three methods in a similar simulation design but with only three data points in the specific source data sets, and two in the unknown source data sets. This second simulation design allows us to exhibit the behavior of the GFF, BF, and LR using data generation equations (3) and (5), but with sample sizes the same as in the real NFI data. Lastly, to assess performance using the real data we show the results of a simulation design that simply samples data sets from the real NFI data.

Preprocessing of the data is described next, followed by a summary of each of the three simulation designs. The results are presented and discussed in the subsections that remain. The implementation of the BF follows as described in Ommen et al. (2017) and Ommen & Saunders (2019) (see their *specific source* formulation). The LR is constructed from Chapter 7.2 of Ommen (2017). For reference, the exact details are presented in the appendix.

A limitation of the NFI casework data is that each specific source consists of only three measurements of the glass fragments making it difficult to obtain very reliable estimates of the specific source parameters, μ_s and A, regardless of the statistical framework (i.e., Bayesian, frequentist, or GF). Since each of the three measurements records the concentration of 10 elements (down from the original 18 as in van Es et al. (2017)), with so few measurements, this is in fact a relatively high-dimensional inference problem. Moreover, since the unknown source data only consists of two measurements, consistent with a sure independence screening strategy (Fan & Lv 2008), in our analysis we reduce the dimension of the measurements to reflect only the two elements (i.e., $p = 2$) that exhibit the largest variation (after being rescaled to have unit norm) over all sources and measurements in the alternative source data set (3×659 measurements in total). Table 1 presents the variance observed for each of the 10 elements, from which we select elements "Pb208" and "Rb85".

element	Ti49	Sr88	K39	Zr90	Mn55	Ba137	Ce140	La139	**Pb208**	**Rb85**
st dev	.00000	.00001	.00001	.00001	.00001	.00002	.00006	.00007	**.00012**	**.00013**

Table 1: Sample standard deviation (rounded to five decimal places) of each element over all 3×659 measurements in the NFI training data set. The data vector for each element was first rescaled to have unit Euclidean norm.

In the first simulation design we generate $n = 659$ alternative sources of data from (5) with $m_i = 3$ measurements for each source. The values of μ_a, B, and C used to generate the data are computed from the real NFI alternative source data via the equations in (9). Next, 320 specific source data sets are generated from (3), each with $m = 150$ measurements. Each of the 320 specific source data sets are generated from unique values of μ_s and A, each corresponding to a particular source of the 320 specific sources in the real NFI data set and computed as

$$\widehat{\mu}_s := \frac{1}{m} \sum_{k=1}^{m} y_{s,k}$$
$$\widehat{A}\widehat{A}' := \frac{1}{m-1} \sum_{k=1}^{m} \big(y_{s,k} - \widehat{\mu}_s\big)\big(y_{s,k} - \widehat{\mu}_s\big)'.$$

To simulate H_p true and H_d true events, respectively, we must generate additional data with unknown sources. For H_p true, an additional $m_u = 2$ measurements for each of the 320 specific sources are generated from (3) using the respective, previously computed values of μ_s and A. For H_d true, an additional 3,000 sets of $m_u = 2$ measurements are generated the same as the alternative sources of data. Accordingly, 320 simulated GFF, BF, and LR values for H_p true are computed using the 320 pairs of unknown and specific source data sets, and 3,000 simulated GFF, BF, and LR values for H_d true are computed using 3,000 non-associated pairs of unknown and specific source data sets (the specific sources are randomly selected from among the 320, for each of the 3,000 unknown sources).

While we could have generated only one data set of $n = 659$ alternative sources of data and one set of 320 specific sources of data, to account for variation in these sources a new set is generated for each of the 3,320 simulated events. The LR crashed for one of the 3,000 simulated H_d true events, so for comparison sake, we omit the data associated with this random number generator seed for all three simulation designs (i.e., all simulation designs have data for 2,999 data sets for H_d true). We describe this simulation design as having ideal sample sizes because $m = 150$ whereas $m = 3$ for the real NFI data. This difference has a particularly significant effect on the stability of the LR, as will be seen in the two simulation designs that follow. See the results in Section 4.1.

This second simulation design is the same as the first, with the modification being that the specific sources each contain only $m = 3$ measurements, as is the case for the real NFI data set. Thus, this simulation is designed to observe the performance of the GFF, BF, and LR on synthetic data that most closely resembles the real NFI data. See the results in Section 4.2.

The third simulation design uses the measurement values from real NFI data set. Recall from Section 3.5 that for each of the 320 specific sources (each containing $m = 3$ measurements) there are an associated two held out measurements. With these 320 sets of $m_u = 2$ measurements each serving as unknown sources, we are able to simulate 320 H_p true events and 320×319 H_d true events. For comparison with the first and second simulation designs, however, we only sample a random subset of 3,000 of the 320×319 H_d true events. See the results in Section 4.3.

4.1 Simulation 1: fully synthetic data with ideal sample sizes

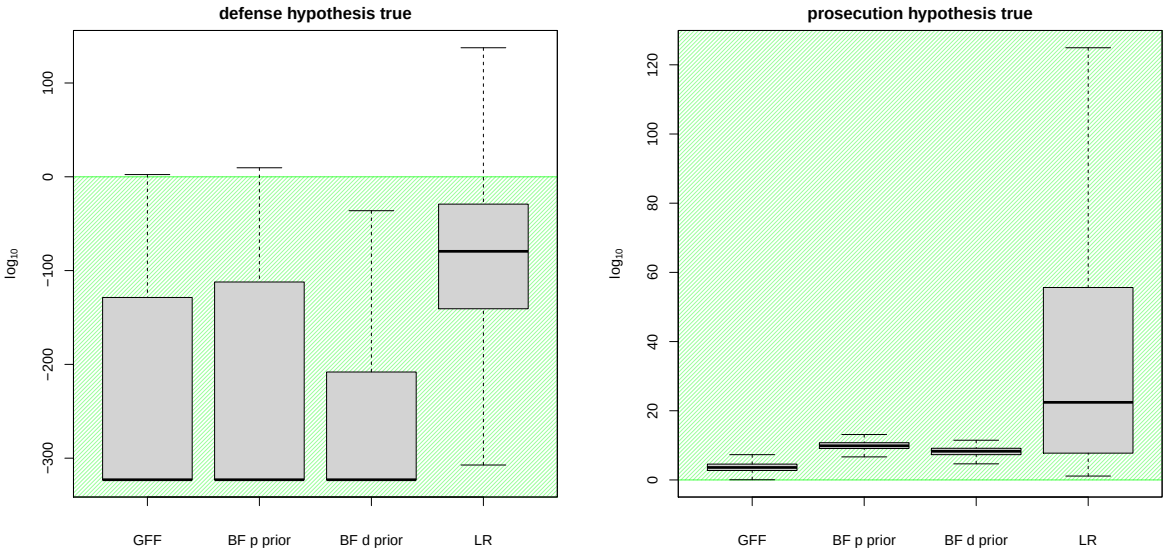

Figure 2: Box plots of the sampling distributions of the GFF, BF, and LR over the 3,000 simulations under H_d (left panel) and 320 simulations under H_p (right panel). For this synthetic 'ideal sample size' simulation, $m_u = 2$, $m = 150$, $n = 659$, and $m_i = 3$. BF p prior denotes the BF constructed from priors that favor H_p, whereas BF d prior denotes the BF constructed from priors that favor H_d. The shaded green regions in each panel correspond to values of the GFF, BF, and LR that favor the true hypothesis. Outliers are omitted.

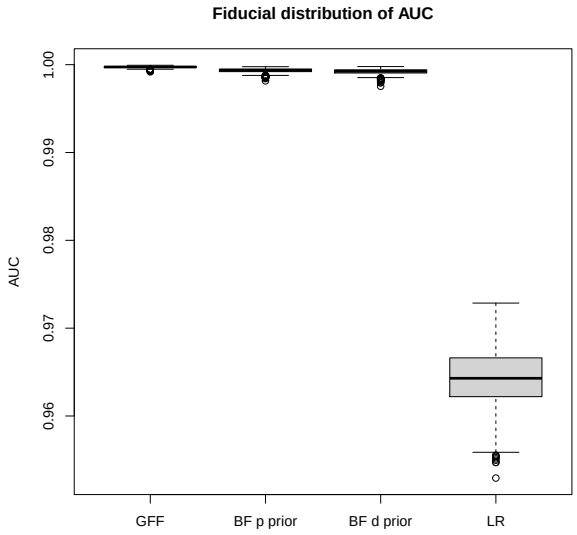

Figure 3: Fiducial distributions of the AUC for the GFF, BF, and LR over the 3,000 simulations under H_d and 320 simulations under H_p. For this 'ideal sample size' simulation, $m_u = 2$, $m = 150$, $n = 659$, and $m_i = 3$.

First, Figure 2 presents a box plot of the performance of the GFF, BF, and LR over the 3,000 simulations under H_d and 320 simulations under H_p. The BF is evaluated both with a prior that

favors the prosecution hypothesis, denoted 'BF p prior', and with a prior that favors the defense hypothesis, denoted 'BF d prior'. The construction of these competing priors is as described in the previous section, where the prior specification is discussed. Figure 2 demonstrates that all four methods perform as reasonably desired in this ideal size synthetic data simulation (i.e., their sampling distributions favor the true hypothesis in under either scenario). Note that the arguably inconsequential difference in the performance of the BF p prior versus BF d prior is a result of the unrealistically ideal sample sizes of this synthetic simulation design. The next simulation design illustrates this point.

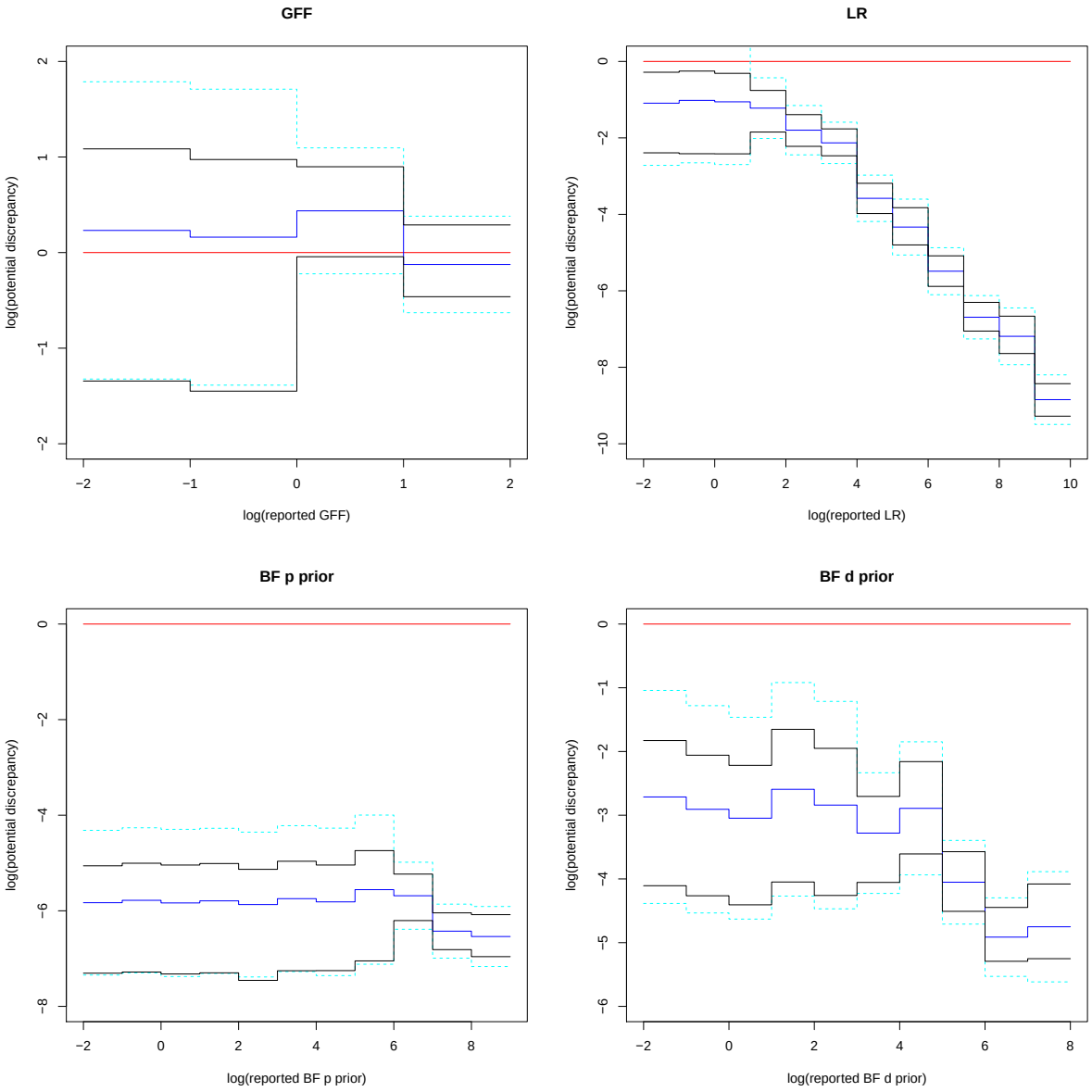

Figure 4: Calibration for the GFF, BF, and LR over the 3,000 simulations under H_d and 320 simulations under H_p. The horizontal red line at zero corresponds to perfect calibration (i.e., LR(LR) = LR). The blue line is the fiducial median log discrepancy. The black and cyan lines are upper and lower .95 point-wise and simultaneous fiducial confidence intervals, respectively, for the log discrepancy. For this 'ideal sample size' simulation, $m_u = 2$, $m = 150$, $n = 659$, and $m_i = 3$.

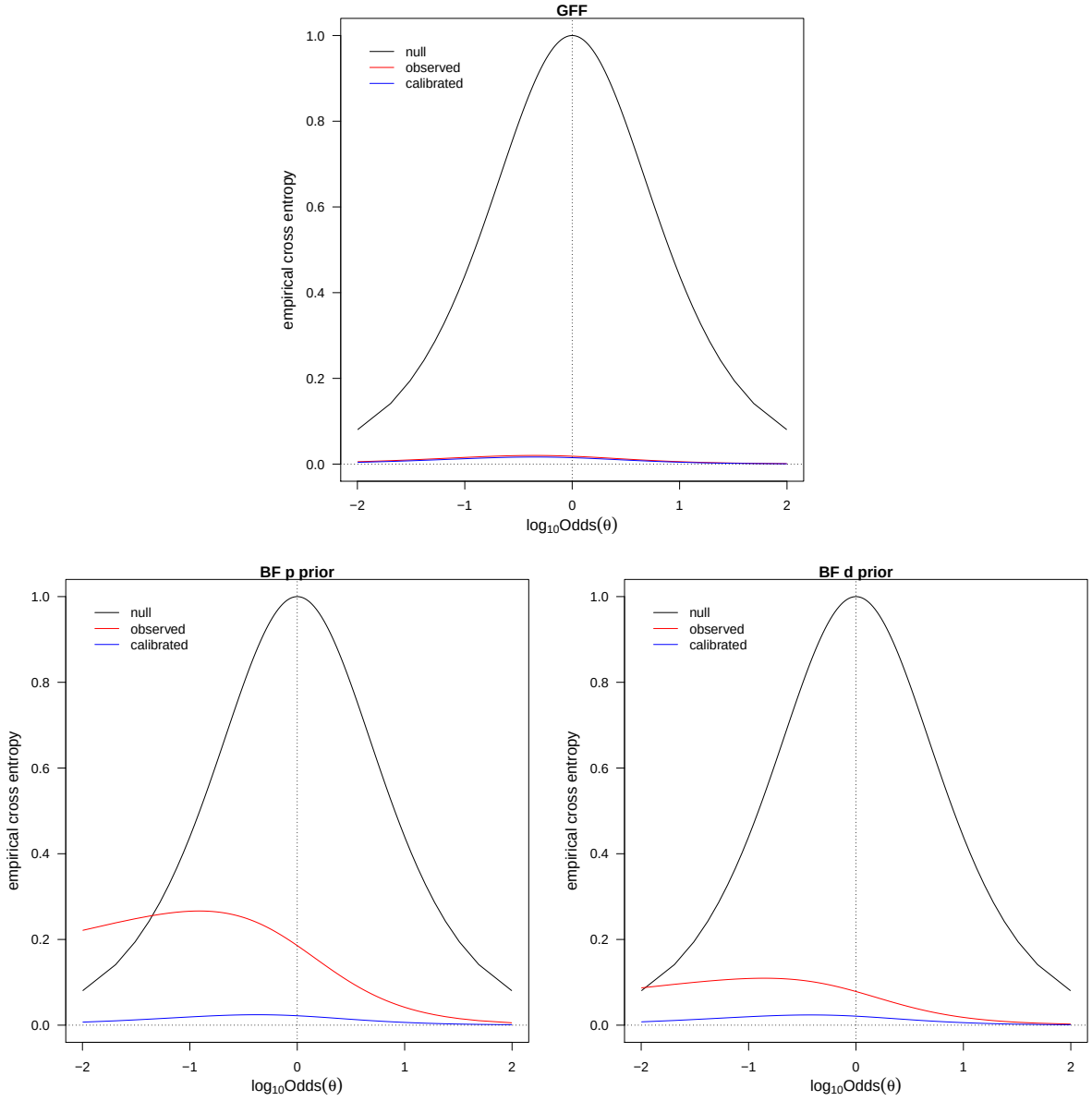

Figure 5: Empirical cross entropy for the GFF, BF, and LR over the 3,000 simulations under H_d and 320 simulations under H_p. This calibration diagnostic tool is proposed in Ramos & Gonzalez-Rodriguez (2008). Good calibration is exhibited when the red line is nested between the blue and black lines, and as close as possible the blue. The code from Ramos & Gonzalez-Rodriguez (2008) crashed for the LR, and for all subsequent simulation designs. For this 'ideal sample size' simulation, $m_u = 2$, $m = 150$, $n = 659$, and $m_i = 3$.

Second, the fiducial distributions of the area under the receiver operating characteristic curve (AUC) for the GFF, BF, and LR are displayed in Figure 3. The AUC measures the adequacy of each of the four methods for accurately discriminating between H_p and H_d, and the observed AUC values reflect an important feature observed in the distributions of the GFF, BF, and LR values in Figure 2. There is almost no overlap in the observed GFF, BF p prior, and BF d prior values, respectively, for H_d true versus H_p true, which means there exist an effective threshold for discriminating between these two hypotheses for each of these methods. Hence, the AUC values are clustered very close to the boundary at one in Figure 3. Conversely,

the LR values exhibit some overlap in tail values between H_d true versus H_p true, and so the LR AUC values reflect this loss of discriminating ability, though not a dramatic loss in this ideal sized simulation design.

Next, a meaningful notion for assessing the performance of ratio quantities such as the GFF, BF, and LR is to determine whether they are well calibrated to the values they exhibit. For example, an LR value of 3 has the interpretation that it is 3 times as likely to observe the evidence when H_p is true than when H_d is true. For this interpretation to be meaningful, for every instance that we observe an LR of 3 when H_d is true we should observe 3 instances of an LR of 3 when H_p is true. As described in Hannig et al. (2019), a shorthand for this notion of calibration is the the expression 'LR(LR) = LR'. We follow the method in Hannig et al. (2019) for estimating the calibration of the 3,000 simulations under H_d and 320 simulations under H_p, for GFF, BF, and LR. See Figure 4 for the estimated calibrations, and observe that the GFF is the best calibrated of the four methods. Note that these ratio quantities can yield very poorly calibrated values while still being effective at discriminating between hypotheses, as seen for the LR, BF p prior, and the BF d prior. The consequence of poor calibration is a misrepresentation, often an exaggeration of the strength of evidence supporting the respective hypotheses. In the context of forensic identification of source problems, such misrepresentation can lead to the false conclusion that the evidence in favor of a particular hypothesis is overwhelming, or beyond any doubt. Thus, the implication of a lack of calibration cannot be overstated.

We conclude this section by presenting an alternative calibration analysis described in Ramos & Gonzalez-Rodriguez (2008). See Figure 5. It is again observed that the GFF values are the best calibrated. Unfortunately, the code (Lucy 2013) for this calibration analysis only worked for the GFF, BF p prior, and BF d prior values in this ideal size synthetic data simulation, and so similar figures are not available for the two simulation designs that follow.

4.2 Simulation 2: fully synthetic data with NFI casework data sample sizes

The sampling distributions are displayed in Figure 6. A first observation is that the LR tends to favor H_p in both scenarios, and as noted for the previous simulation design this results from an unstable MLE of the specific source parameters with $m = 3$. Referring back to the LR construction in equation (13), the instability stems from the evaluation of $f_s\big(\{y_{s,k}\} \mid \widehat{\theta}_s\big)$ in the denominator.

The next feature to observe in Figure 6 is that the strength of evidence for H_d is characterized by the BF p prior an order of magnitude smaller than by the BF d prior, in the H_d true scenario. These prosecution and defense priors were constructed to reflect extreme beliefs, to demonstrate that any values between the BF p prior and BF d prior values can reasonably result from the prior specification. The H_p true scenario is even more problematic because the BF p prior and the BF d prior tend to favor opposite hypotheses. This consequence of subjectivist Bayesian prior choice for forensic identification of source problems, as illustrated in Figure 6, is exceedingly problematic because it demonstrates that the strength of evidence for or against a hypothesis is heavily influenced by the competing prior beliefs (prosecution versus defense) for or against the hypothesis, even to the point where the BF entirely favors the wrong hypothesis. Conversely, it is observed in Figure 6 that the GFF values tend to favor the true hypothesis in each scenario. Moreover, the GFF values do not suffer from the instability exhibited by the LR values. These important features illustrated in Figure 6 are further supported by the discrimination and calibration analyses presented in Figures 7 and 8, respectively.

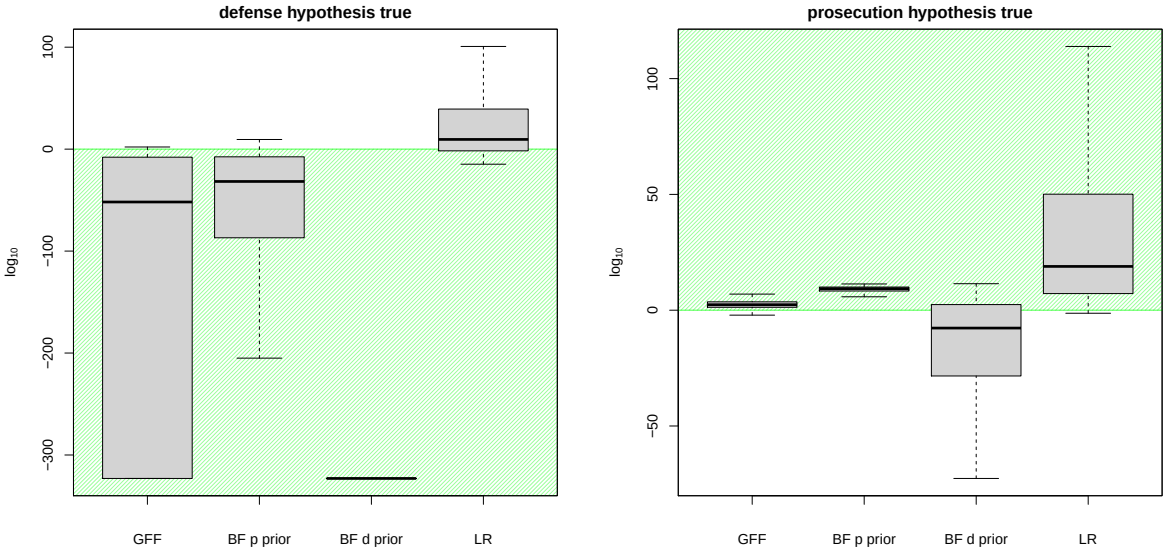

Figure 6: Box plots of the sampling distributions of the GFF, BF, and LR over the 3,000 simulations under H_d (left panel) and 320 simulations under H_p (right panel). For this synthetic 'NFI casework data sample sizes' simulation, $m_u = 2$, $m = 3$, $n = 659$, and $m_i = 3$. BF p prior denotes the BF constructed from priors that favor H_p, whereas BF d prior denotes the BF constructed from priors that favor H_d. The shaded green regions in each panel correspond to values of the GFF, BF, and LR that favor the true hypothesis. Outliers are omitted.

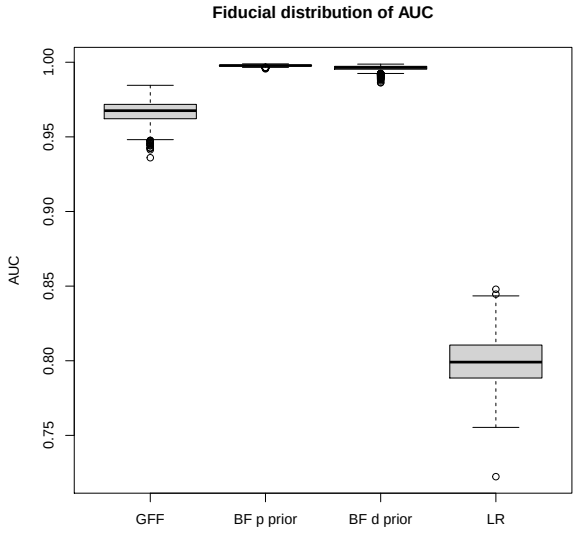

Figure 7: Fiducial distributions of the AUC for the GFF, BF, and LR over the 3,000 simulations under H_d and 320 simulations under H_p. For this synthetic 'NFI casework data sample sizes' simulation, $m_u = 2$, $m = 3$, $n = 659$, and $m_i = 3$.

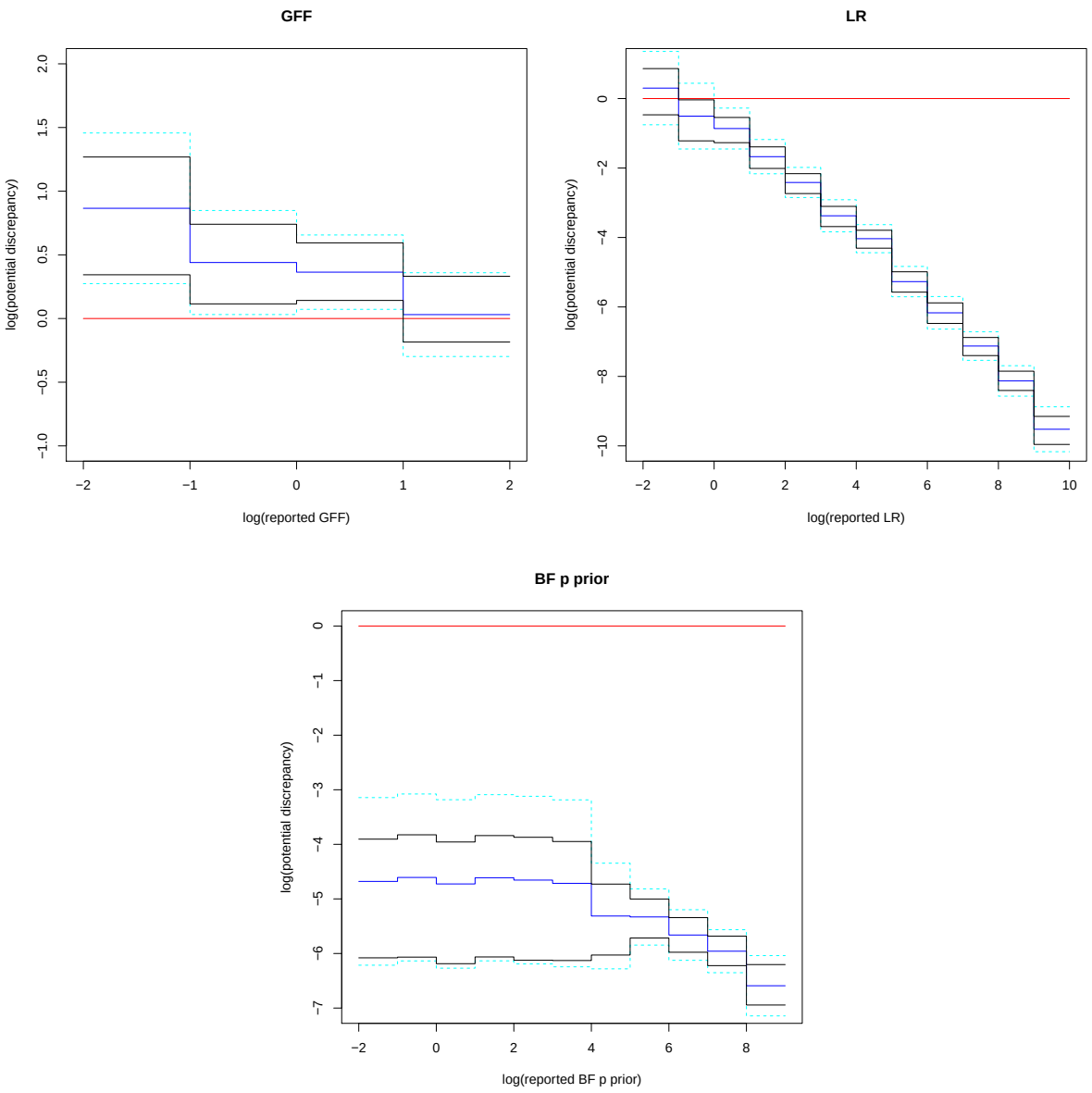

Figure 8: Calibration for the GFF, BF, and LR over the 3,000 simulations under H_d and 320 simulations under H_p. The horizontal red line at zero corresponds to perfect calibration (i.e., LR(LR) = LR). The blue line is the fiducial median log discrepancy. The black and cyan lines are upper and lower .95 pointwise and simultaneous fiducial confidence intervals, respectively, for the log discrepancy. For this 'NFI casework data sample sizes' simulation, $m_u = 2$, $m = 3$, $n = 659$, and $m_i = 3$.

As alluded to in the discussion for the previous simulation design, even if the values of the GFF, BF, or LR do not tend to be associated with the true hypothesis, it is still possible that these methods are effective at correctly discriminating between H_d and H_p. The most notable of these four methods is BF d prior values, as displayed in Figure 6. There is a clear distinction between the distribution of BF d prior values under H_d versus H_p, even though both distributions tend to exhibit values associated with H_d true. The distinction between the BF d prior values for the two hypotheses is characterized by the fiducial AUC distributions shown in Figure 7 (along with that for the other three methods, as well). As in the previous simulation

design, with AUC values very close to one, the GFF, BF p prior, and BF d prior are very effective at discriminating between H_d and H_p. The LR suffers in its ability to discriminate, due to the issues with numerical instability for sample sizes so small, as described at the beginning of this section and illustrated in Figure 6.

While in this 'actual sample size' simulation design, the GFF and BF p prior methods tend to exhibit values associated with the correct hypothesis (i.e., Figure 6) and are effective at discriminating between H_d and H_p (i.e., Figure 7), there is still a danger that they are not calibrated to appropriately reflect the strength of evidence that their values suggest. Figure 8 presents the calibration analysis for the GFF, BF p prior, and LR values. Note that the calibration for the BF d prior values is missing; the values are very poorly calibrated and so the calibration softwares crashed. Furthermore, Figure 8 suggests that the LR and BF p prior values are also poorly calibrated. The GFF values are much better, and in fact, reasonably well calibrated in light of the very small sample sizes that characterize this simulation design and the real NFI casework data.

4.3 Simulation 3: real NFI casework data

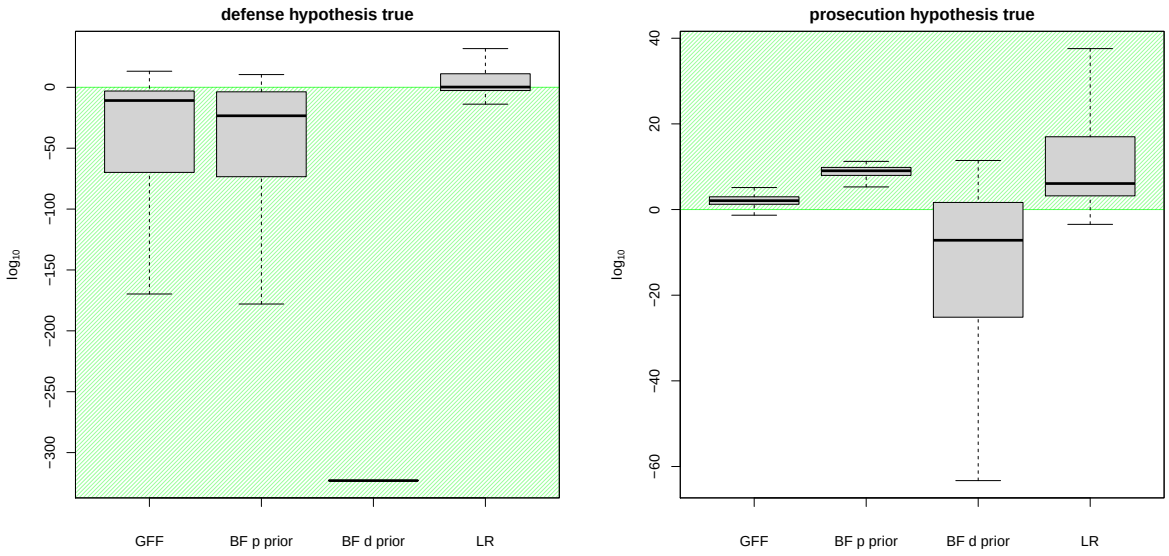

Figure 9: Box plots of the sampling distributions of the GFF, BF, and LR over the 3,000 simulations under H_d (left panel) and 320 simulations under H_p (right panel). For this 'real NFI casework data' simulation, $m_u = 2$, $m = 3$, $n = 659$, and $m_i = 3$. BF p prior denotes the BF constructed from priors that favor H_p, whereas BF d prior denotes the BF constructed from priors that favor H_d. The shaded green regions in each panel correspond to values of the GFF, BF, and LR that favor the true hypothesis. Outliers are omitted.

Once again, the resulting sampling distributions of the methods are presented as box plots in Figure 9. The fiducial distributions of the AUC to assess discrimination effectiveness between the hypotheses are presented in Figure 10, and the calibration analysis is displayed in Figure 11. What is most noteworthy about the results of this simulation design is that they are largely unchanged from those of the synthetic simulation design (with matching sample sizes). This suggests that the assumed data generating models are reasonable approximations to this real casework data, with respect to quantifying the evidence in favor of the competing hypotheses,

H_d and H_p. Likewise, the LR and BF exhibit the same deficiencies that they did with the synthetic data. However, the GFF tends to less extreme values than it did for the synthetic data, most noticeably for the H_d true scenario.

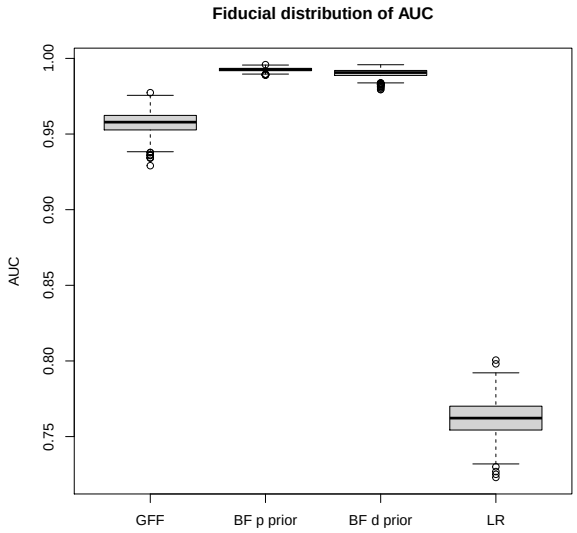

Figure 10: Fiducial distributions of the AUC for the GFF, BF, and LR over the 3,000 simulations under H_d and 320 simulations under H_p. For this 'real NFI casework data' simulation, $m_u = 2$, $m = 3$, $n = 659$, and $m_i = 3$.

The less extreme GFF values may indicate that the GFF is less robust to model misspecification than the BF or LR. Or it may indicate that a Gaussian random effect for the alternative source model is a better approximation for describing the NFI casework data (since the BF and LR were constructed on this assumption). Nonetheless, the calibration of the GFF values presented in Figure 11 suggests that the GFF is well calibrated for values below 10^7. The degradation in calibration for values greater than 10^7 seems to be driven by 3 very large values for GFF under H_d true in the simulation of 3,000 data sets. Under H_d true for the real NFI casework data, the unknown source data is neither (truly) associated with the specific nor alternative sources. Accordingly, it is possible that 1 in a 1,000 of the unknown source data looks very different from the alternative source data (and maybe not so different from the specific source data) in which case the denominator of the GFF is very small, so that the GFF value becomes excessively large. We also see a similar phenomenon occurring for LR and BF (depending on the prior). This would not happen for the simulated synthetic data simulations because under H_d the unknown source data is actually generated from the alternative source.

In forensic identification of source applications, and particularly for those that rely on such small sample sizes, it is very important that the inferential methods being used are appropriately calibrated to reflect the strength of evidence provided by the data. Accordingly, for small sample sizes ($m = 3$ and $m_u = 2$, in this case) practitioners should be very skeptical of any tool that conveys extreme confidence in favor of either of the competing hypothesis.

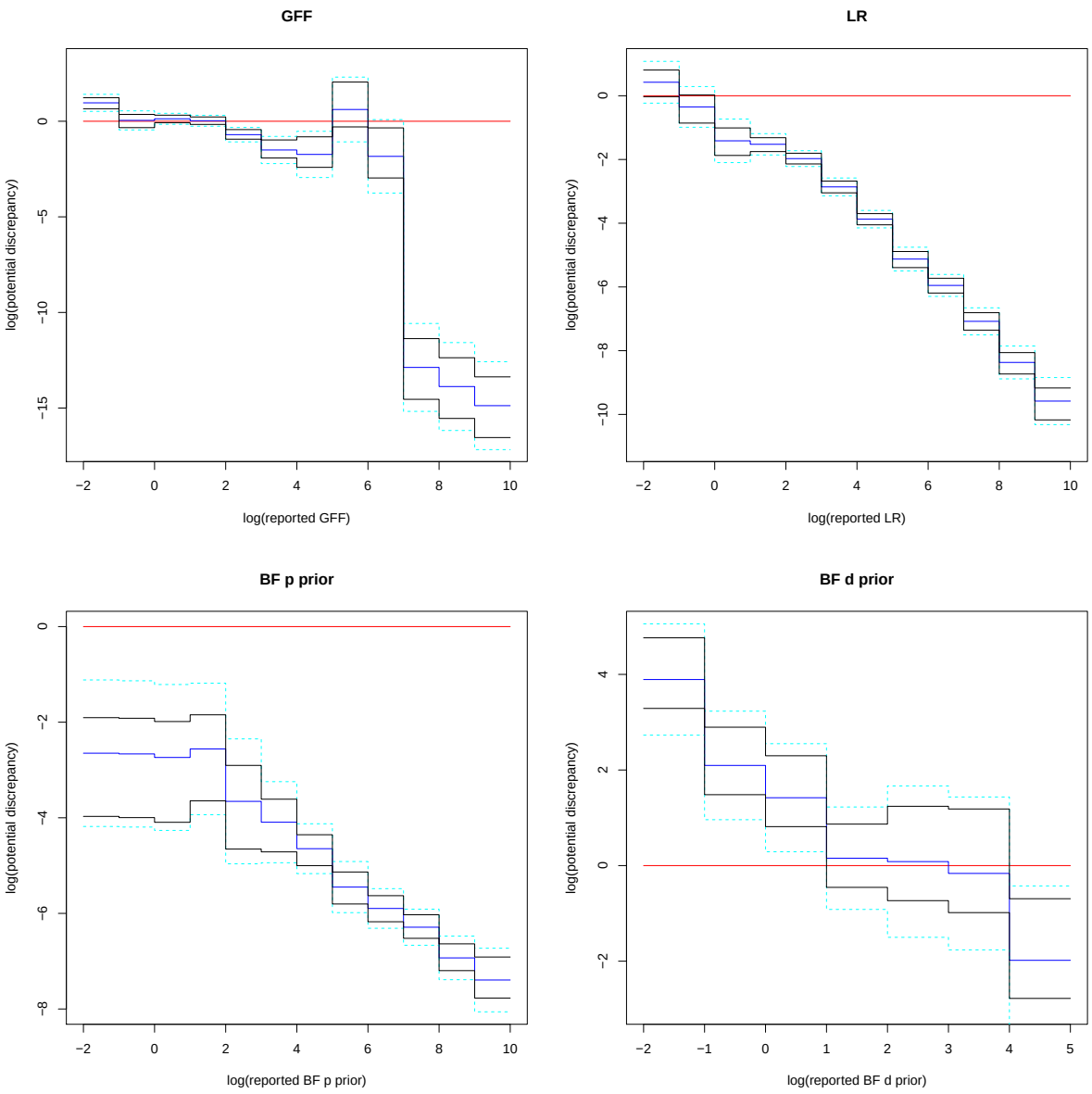

Figure 11: Calibration for the GFF, BF, and LR over the 3,000 simulations under H_d and 320 simulations under H_p. The horizontal red line at zero corresponds to perfect calibration (i.e., LR(LR) = LR). The blue line is the fiducial median log discrepancy. The black and cyan lines are upper and lower .95 pointwise and simultaneous fiducial confidence intervals, respectively, for the log discrepancy. For this 'real NFI casework data' simulation, $m_u = 2$, $m = 3$, $n = 659$, and $m_i = 3$.

5 Concluding remarks

The motivations for this research and the writing of this manuscript are multifaceted. The use of the BF or LR in the context of forensic identification of source applications is problematic. Given the high stakes nature of such applications in criminal justice systems around the world, the statistics community must take responsibility for both communicating the dangerous shortcomings of these methods that are in widespread use, and for developing new methods that

overcome such shortcomings.

In regards to the BF, the entire notion of "reasonableness" has no meaning in the context of subjectivist Bayesian prior specification/choice, especially in an adversarial scenario (e.g., prosecution versus defense). Furthermore, while we observed the BF to be effective at discriminating between H_d and H_p, the BF values were highly influenced by the choice of prior and they were not calibrated to represent the strength of evidence they conveyed. In regards to the LR, ratios of likelihood functions evaluated at MLEs computed from excessively small data sets are very unstable, the LR values fell short in their ability to discriminate between H_d and H_p, and they were poorly calibrated. We have provided evidence to demonstrate these assertions empirically, and on real casework data, and we have constructed and validated a GFF as an alternative methodological approach and tool that does not suffer from the demonstrated deficiencies in the BF and LR.

Lastly, there is an argument to be made that the shortcomings in the BF approach can be remedied via the construction of *objective* priors (however that is to be defined). To this point, in reference to equation (1), the GFF can be interpreted precisely as a BF arising from a particular choice of *objective*, data-driven priors.

References

Aitken, C. G. G. & Lucy, D. (2004), 'Evaluation of trace evidence in the form of multivariate data', *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **53**(1), 109–122.

Beaumont, M. A., Zhang, W. & Balding, D. J. (2002), 'Approximate bayesian computation in population genetics', *Genetics* **162**(4), 2025–2035.

Berger, C. E. & Slooten, K. (2016), 'The LR does not exist', *Science and Justice* **56**(5), 388–391.

Berger, J. O., Bernardo, J. M., Sun, D. et al. (2009), 'The formal definition of reference priors', *The Annals of Statistics* **37**(2), 905–938.

Bernardo, J. M. (1979), 'Reference posterior distributions for bayesian inference', *Journal of the Royal Statistical Society: Series B (Methodological)* **41**(2), 113–128.

Biedermann, A., Bozza, S., Taroni, F. & Aitken, C. G. G. (2016), 'Reframing the debate: A question of probability, not of likelihood ratio', *Science and Justice* **56**(5), 392–396.

Bolck, A., Weyermann, C., Dujourdy, L., Esseiva, P. & van den Berg, J. (2009), 'Different likelihood ratio approaches to evaluate the strength of MDMA tablet comparisons', *Forensic Science International* **191**(1), 42–51.

DiCiccio, T. J., Kass, R. E., Raftery, A. & Wasserman, L. (1997), 'Computing bayes factors by combining simulation and asymptotic approximations', *Journal of the American Statistical Association* **92**(439), 903–915.

Egli, N. M., Champod, C. & Margot, P. (2006), 'Evidence evaluation in fingerprint comparison and automated fingerprint identification systems - modeling between finger variability', *Forensic Science International* **176**, 189–195.

ENFSI (2015), 'Enfsi guideline for evaluative reporting in forensic science', http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf.

Evett, I. W. (1977), 'The interpretation of refractive index measurements', *Forensic Science* **9**, 209–217.

Evett, I. W. (1986), 'A Bayesian approach to the problem of interpreting glass evidence in forensic science casework', *Journal of the Forensic Science Society* **26**, 3–18.

Fan, J. & Lv, J. (2008), 'Sure independence screening for ultrahigh dimensional feature space', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**(5), 849–911.

Gelman, A., Jakulin, A., Pittau, M. G., Su, Y.-S. et al. (2008), 'A weakly informative default prior distribution for logistic and other regression models', *The annals of applied statistics* **2**(4), 1360–1383.

Gelman, A. & Meng, X.-L. (1998), 'Simulating normalizing constants: From importance sampling to bridge sampling to path sampling', *Statistical science* pp. 163–185.

Gonzalez-Rodriguez, J., Drygajlo, A., Ramos-Castro, D., Garcia-Gomar, M. & Ortega-Garcia, J. (2006), 'Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition', *Computer Speech and Language* **20**, 331–355.

Gonzalez-Rodriguez, J., Fierrez-Aguilar, J., Ramos-Castro, D. & Ortega-Garcia, J. (2005), 'Bayesian analysis of fingerprint, face and signature evidences with automatic biometric systems', *Forensic Science International* **155**, 126–140.

Grove, D. M. (1980), 'The interpretation of forensic evidence using a likelihood ratio', *Biometrika* **67**(1), 243–264.

Gupta, A., Martinez-Lopez, C., Curran, J. M. & Almirall, J. R. (2019), 'Multi-element comparisons of tapes evidence using dimensionality reduction for calculating likelihood ratios', *Forensic science international* **301**, 426–434.

Hannig, J., Iyer, H., Lai, R. C. & Lee, T. C. (2016), 'Generalized fiducial inference: A review and new results', *Journal of the American Statistical Association* **111**(515), 1346–1361.

Hannig, J., Riman, S., Iyer, H. & Vallone, P. M. (2019), 'Are reported likelihood ratios well calibrated?', *Forensic Science International: Genetics Supplement Series* **7**(1), 572–574.

Hepler, A., Saunders, C. P., Davis, L. & Buscaglia, J. (2012), 'Score-based likelihood ratios for handwriting evidence', *Forensic Science International* **219**(1), 129–140.

Jeffreys, H. (1946), '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.

Kafadar, K. (2018), 'The critical role of statistics in demonstrating the reliability of expert evidence', *Fordham Law Review* **86**(4), 1617–1637.

Kass, R. E. & Raftery, A. E. (1995), 'Bayes factors', *Journal of the American Statistical Association* **90**(430), 773–795.

Lindley, D. V. (1972), *Bayesian statistics, a review*, Vol. 2, SIAM.

Lindley, D. V. (1977), 'A problem in forensic science', *Biometrika* **64**(2), 207–213.

Lucy, D. (2013), *comparison: Multivariate likelihood ratio calculation and evaluation.* R package version 1.0-4.
URL: *https://CRAN.R-project.org/package=comparison*

Lund, S. P. & Iyer, H. (2017), 'Likelihood ratio as weight of forensic evidence: A closer look', *Journal of the Research of National Institute of Standards and Technology* **122**(27), 1–32.

Martin, R., Walker, S. G. et al. (2019), 'Data-driven priors and their posterior concentration rates', *Electronic Journal of Statistics* **13**(2), 3049–3081.

Meng, X.-L. & Wong, W. H. (1996), 'Simulating ratios of normalizing constants via a simple identity: a theoretical exploration', *Statistica Sinica* pp. 831–860.

Mukerjee, R. & Reid, N. (1999), 'On a property of probability matching priors: matching the alternative coverage probabilities', *Biometrika* **86**(2), 333–340.

Neumann, C. & Ausdemore, M. A. (2020), 'Defence Against the Modern Arts: the Curse of Statistics - Part II: "Score-based likelihood ratios"', *Law, Probability, and Risk* **https://doi.org/10.1093/lpr/mgaa006**.

Neumann, C., Champod, C., Puch-Solis, R., Egli, N. M., Anthonioz, A. & Bromage-Griffiths, A. (2007), 'Computation of likelihood ratios in fingerprint identification for configurations of any number of minutiae', *Journal of Forensic Sciences* **52**, 54–64.

Neumann, C., Hendricks, J. & Ausdemore, M. A. (2020), *"Statistical support for conclusions in fingerprint examinations" chapter in Handbook of Forensic Statistics*, In Press edn, CRC Press, Boca Raton, FL, USA.

Ommen, D. M. (2017), 'Approximate statistical solutions to the forensic identification of source problem', *Electronic Theses and Dissertations.* **1710**.
URL: *https://openprairie.sdstate.edu/etd/1710*

Ommen, D. M. & Saunders, C. P. (2019), 'Reconciling the bayes factor and likelihood ratio for two non-nested model selection problems', *arXiv preprint arXiv:1901.09798* .

Ommen, D. M., Saunders, C. P. & Neumann, C. (2017), 'The characterization of monte carlo errors for the quantification of the value of forensic evidence', *Journal of Statistical Computation and Simulation* **87**(8), 1608–1643.

Parker, J. B. (1966), 'A statistical treatment of identification problems', *Journal of the Forensic Science Society* **6**(1), 33–39.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. & R Core Team (2019), *nlme: Linear and Nonlinear Mixed Effects Models.* R package version 3.1-140.
URL: *https://CRAN.R-project.org/package=nlme*

Ramos, D. & Gonzalez-Rodriguez, J. (2008), Cross-entropy analysis of the information in forensic speaker recognition, *in* 'Odyssey 2008: The Speaker and Language Recognition Workshop', International Speech Communication Association.

Savage, L. J. (1961), The foundations of statistics reconsidered, *in* 'Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics', The Regents of the University of California.

Shafer, G. (1982), 'Lindley's paradox', *Journal of the American Statistical Association* **77**(378), 325–334.

Shi, W. J., Hannig, J., Lai, R. & Lee, T. (2017), 'Covariance estimation via fiducial inference', *arXiv preprint arXiv:1708.04929* .

Staicu, A.-M. & Reid, N. M. (2008), 'On probability matching priors', *Canadian Journal of Statistics* **36**(4), 613–622.

Swofford, H., Koerntner, A., Zemp, F., Ausdemore, M., Liu, A. & Salyards, M. (2018), 'A method for the statistical interpretation of friction ridge skin impression evidence: method development and validation', *Forensic Science International* **287**, 113–126.

Taroni, F., Bozza, S., Biedermann, A. & Aitken, C. G. G. (2016), 'Dismissal of the illusion of uncertainty in the assessment of a likelihood ratio', *Law, Probability, and Risk* **15**(1), 1–16.

van Es, A., Wiarda, W., Hordijk, M., Alberink, I. & Vergeer, P. (2017), 'Implementation and assessment of a likelihood ratio approach for the evaluation of la-icp-ms evidence in forensic glass analysis', *Science & Justice* **57**(3), 181–192.

Wasserstein, R. L. & Lazar, N. A. (2016), 'The ASA Statement on p-Values: Context, Process, and Purpose', *The American Statistician* **70**(2), 129–133.

Zadora, G., Martyna, A., Ramos, D. & Aitken, C. (2013), *Statistical analysis in forensic science: evidential value of multivariate physicochemical data*, John Wiley & Sons.

6 Appendix

In this section, the details of the BF and LR specification and computations are given. These details for the BF are as in Ommen et al. (2017), Ommen & Saunders (2019). Assuming the posterior distributions of θ_s and θ_a are independent, the BF from equation (7) is expressed as

$$\begin{aligned} \text{BF} &= \frac{\int\int f_s\big(\{y_{u,j}\} \mid \theta_s\big) \cdot \pi_s\big(\theta_s \mid \{y_{s,k}\}\big) \cdot \pi_a\big(\theta_a \mid \{y_{a,i,k}\}\big)\ d\theta_s\ d\theta_a}{\int\int f_a\big(\{y_{u,j}\} \mid \theta_a\big) \cdot \pi_s\big(\theta_s \mid \{y_{s,k}\}\big) \cdot \pi_a\big(\theta_a \mid \{y_{a,i,k}\}\big)\ d\theta_s\ d\theta_a} \\ &= \int\int \frac{f_s\big(\{y_{u,j}\} \mid \theta_s\big)}{f_a\big(\{y_{u,j}\} \mid \theta_a\big)} \cdot \pi_d\big(\theta_s, \theta_a \mid \{y_{s,k}\}, \{y_{a,i,k}, y_{u,j}\}\big)\ d\theta_s\ d\theta_a, \end{aligned} \tag{10}$$

where

$$\pi_d\big(\theta_s, \theta_a \mid \{y_{s,k}\}, \{y_{a,i,k}, y_{u,j}\}\big) := \frac{f_a\big(\{y_{u,j}\} \mid \theta_a\big) \cdot \pi_s\big(\theta_s \mid \{y_{s,k}\}\big) \cdot \pi_a\big(\theta_a \mid \{y_{a,i,k}\}\big)}{\int\int f_a\big(\{y_{u,j}\} \mid \theta_a\big) \cdot \pi_s\big(\theta_s \mid \{y_{s,k}\}\big) \cdot \pi_a\big(\theta_a \mid \{y_{a,i,k}\}\big)\ d\theta_s\ d\theta_a}$$

is the posterior distribution of (θ_s, θ_a) under the defense hypothesis that the unknown source data are generated from the alternative. Note that this is simply a method for computing the BF, and it does not favor one hypothesis over another.

The random effects term in (5) is assumed to follow a multivariate Gaussian distribution in Ommen et al. (2017), and they construct the following conjugate priors for the various

parameters.

$$\begin{aligned}
\mu_s &\sim \mathrm{N}_p(\mu_\pi, \Sigma_b) \\
AA' &\sim \text{inv-Wishart}_p(\Sigma_e, \nu_e) \\
\mu_a &\sim \mathrm{N}_p(\mu_\pi, k\Sigma_b) \\
BB' &\sim \text{inv-Wishart}_p(\Sigma_b, \nu_b) \\
CC' &\sim \text{inv-Wishart}_p(\Sigma_e, \nu_e),
\end{aligned} \tag{11}$$

where k is some scalar. Particularly with small samples sizes for the observed specific and unknown source data, even small variations in the data can lead to numerically unreliable BF values, especially due to the light tails of the Gaussian likelihood function. Accordingly, from these priors it follows that it is most consistent with a belief in the prosecution hypothesis to set as diffuse as possible the specific source priors so that the unknown source data is as consistent as possible with the specific source posterior distribution. This is done by choosing large components for Σ_b for the prior on μ_s and small degrees of freedom parameter ν_e for the prior on AA'. Conversely, it is most consistent with a belief in the defense hypothesis to choose small components for Σ_b and a large ν_e so as to make the unknown source data appear as distinct as possible from the specific source posterior distribution. In the simulation studies that follow, we construct priors from the extremes of both hypotheses in order to illustrate the excessive range in variation of the resulting BF values.

Recall from the computational expression of the BF in (10), the unknown source data is appended to the alternative source data. With the updated $\{y_{a,i,k}\} = \{y_{a,i,k}, y_{u,j}\}$ and denoting $m_{n+1} := m_u$, the conditional posteriors resulting from the priors in (11) are

$$\begin{aligned}
\mu_s \mid \{y_{s,k}\}, AA' &\sim \mathrm{N}_p\big(M^{-1}L, M^{-1}\big) \\
AA' \mid \{y_{s,k}\}, \mu_s &\sim \text{inv-Wishart}_p\big(S_s + \Sigma_e, \nu_e + m\big) \\
\mu_a \mid \{y_{a,i,k}\}, BB', CC' &\sim \mathrm{N}_p\big(Q^{-1}R, Q^{-1}\big) \\
CV_{i,k} \mid CC' &\sim \mathrm{N}_p(0, CC') \\
BB' \mid \{y_{a,i,k}\}, \{CV_{i,k}\}, \mu_a &\sim \text{inv-Wishart}_p\big(S_v + \Sigma_b, N + m_{n+1} + \nu_b\big) \\
BT_i \mid BB' &\sim \mathrm{N}_p(0, BB') \\
CC' \mid \{y_{a,i,k}\}, \{BT_i\}, \mu_a &\sim \text{inv-Wishart}_p\big(S_a + \Sigma_e, N + m_{n+1} + \nu_e\big),
\end{aligned} \tag{12}$$

where S_s is defined in (4), S_a is defined in (6) with an additional m_{n+1} terms corresponding to the $\{y_{u,j}\}$ components, and

$$\begin{aligned}
M &:= m(AA')^{-1} + \Sigma_b^{-1} \\
L &:= m(AA')^{-1}\bar{y}_{s,\cdot} + \Sigma_b^{-1}\mu_\pi \\
Q &:= (N + m_{n+1})\big(BB' + CC'\big)^{-1} + (k\Sigma_b)^{-1} \\
R &:= (N + m_{n+1})\big(BB' + CC'\big)^{-1}\bar{y}_{a,\cdot,\cdot} + (k\Sigma_b)^{-1}\mu_\pi \\
S_v &:= \sum_{i=1}^{n+1}\sum_{k=1}^{m_i}(y_{a,i,k} - \mu_a - CV_{i,k})(y_{a,i,k} - \mu_a - CV_{i,k})'.
\end{aligned}$$

To compute the joint posterior distribution of all the model parameters, we wrote a custom Gibbs sampler that iterates according to the updates enumerated in (12). This code is available at `https://jonathanpw.github.io/research.html`.

The LR is constructed from Chapter 7.2 of Ommen (2017) as,

$$LR := \frac{f_s\big(\{y_{u,j}\} \mid \widehat{\theta}_s^{\star}\big) \cdot f_s\big(\{y_{s,k}\} \mid \widehat{\theta}_s^{\star}\big) \cdot f_a\big(\{y_{a,i,k}\} \mid \widehat{\theta}_a\big)}{f_a\big(\{y_{u,j}\} \mid \widehat{\theta}_a^{\star}\big) \cdot f_s\big(\{y_{s,k}\} \mid \widehat{\theta}_s\big) \cdot f_a\big(\{y_{a,i,k}\} \mid \widehat{\theta}_a^{\star}\big)}, \tag{13}$$

where $\widehat{\theta}_s^{\star}$ is the MLE of the specific source parameters $\theta_s = \{\mu_s, AA'\}$ from the pooled data $\{y_{s,k}, y_{u,j}\}$ based on the the prosecution hypothesis, $\widehat{\theta}_s$ is the MLE of θ_s from the data $\{y_{s,k}\}$, $\widehat{\theta}_a^{\star}$ is the MLE of the alternative source parameters $\theta_a = \{\mu_a, BB', CC'\}$ from the pooled data $\{y_{a,i,k}, y_{u,j}\}$ based on the the defense hypothesis, and $\widehat{\theta}_a$ is the MLE of θ_a from the data $\{y_{a,i,k}\}$. The `lme` function from the `nlme` R package (Pinheiro et al. 2019) is used to compute the MLE for each of the parameters.