The Annals of Applied Statistics
2023, Vol. 17, No. 1, 378–402
https://doi.org/10.1214/22-AOAS1632

GENERALIZED FIDUCIAL FACTOR: AN ALTERNATIVE TO THE BAYES FACTOR FOR FORENSIC IDENTIFICATION OF SOURCE PROBLEMS

BY JONATHAN P. WILLIAMS[1,a], DANICA M. OMMEN[2,b] AND JAN HANNIG[3,c]

[1]*Department of Statistics, North Carolina State University,* [a]*jwilli27@ncsu.edu*
[2]*Department of Statistics, Iowa State University,* [b]*dmommen@iastate.edu*
[3]*Department of Statistics and Operations Research, UNC at Chapel Hill and National Institute of Standards and Technology,* [c]*jan.hannig@unc.edu*

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 comparing the marginal distribution of measurements of trace evidence under two competing propositions for whether or not the unknown source evidence originated from a specific source. 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 choice/specification has a strong influence on the value of the Bayes factor. To remedy this problem of prior specification and choice, we develop an alternative to the Bayes factor, within the framework of generalized fiducial inference, that we term a *generalized fiducial factor*. Furthermore, we demonstrate empirically, on synthetic and real Netherlands Forensic Institute casework data, deficiencies in Bayes factor and classical/frequentist likelihood ratio approaches.

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 (Lindley (1971), Savage (1961)) 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 a range of such prior probability distributions that are reasonable, depending on the experts' role in the courtroom. On one extreme there is the model representing the belief of the prosecution, and on the other extreme is the 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 Netherlands Forensic Institute (NFI) (van Es et al. (2017)).

Received January 2021; revised March 2022.
Key words and phrases. Bayes factor, generalized fiducial inference, likelihood ratio.

Several approaches for assigning value to forensic evidence have been explored, including the two-stage approach (Evett (1977), Parker (1966)), likelihood ratio (LR) with Bayesian treatment of parameter uncertainty (Aitken and Lucy (2004), Evett (1986), Lindley (1977)) or with maximum likelihood estimates (MLE) of parameters (Grove (1980), Ommen (2017)) as well as score-based approaches (Bolck et al. (2009), Egli, Champod and Margot (2006), Gonzalez-Rodriguez et al. (2006), Gonzalez-Rodriguez et al. (2005), Hepler et al. (2012), Neumann et al. (2007)). The Bayes factor (BF) approach is the most commonly recommended among European countries (Berger and Slooten (2016), Biedermann et al. (2016), ENFSI (2015), Taroni et al. (2016)), while a non-Bayesian approach is often recommended in the U.S. (Kafadar (2018), Swofford et al. (2018)). Recently, all of these methods have been scrutinized due to their lack of attention to the handling of uncertainty (Lund and Iyer (2017), Morrison (2016)). In this paper we contribute to the discussion regarding how to handle uncertainty when quantifying the value of evidence, and we focus on a similar question to the one proposed in Lund and Iyer (2017): "What do you really know vs. what are you claiming to know (using prior information)?"

The gist of the LR approaches is to compare the probability of observing the evidence under two competing 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 and 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 data sets 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 and Ausdemore (2020), Neumann, Hendricks and 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, Hendricks and Ausdemore (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 generalized fiducial inference (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 may be poorly calibrated to reflect the strength of evidence that they convey. Our empirical results demonstrate very transparently that the degree to which the BF could vary often may be 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 generalized fiducial factor (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 in our Supplementary Material (Williams, Ommen and Hannig (2023), also 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 have 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 the 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 in our Supplementary Material (Williams, Ommen and Hannig (2023), also 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 evidential 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 same 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 which would link the suspect to the scene of the crime.

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 and Saunders (2019)). In this formulation, material evidence, such as trace elements (i.e., the chemical composition of the chemical components that are useful for discrimination; see Dettman et al. (2014)) 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 evidence originated from the specific source.
H_d: The unknown source evidence originated from some other source in the alternative source population.

We confine the rest of our exposition to modeling evidence arising from trace elements of glass fragments. 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, also can be 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 have been documented in the alternative source database. This logical framework lends itself to describing the alternative source data by a random effects model,

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 relevant 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. The alternative sources of data provide a benchmark for comparison. In the sections that follow, we develop and evaluate statistical methodology to address this question. Further, we design a simulation study consistent with 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.

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

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

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

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

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 $y = G(U,\theta)$, assuming that $G^{-1}(\cdot,\theta)$ exists (thus leading to a unique solution $u = G^{-1}(y,\theta)$). In our applications it is the case that $G^{-1}(\cdot,\theta)$ exists. 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 noninformativie, weakly informative, etc.) priors are well accepted in the literature and, in practice, as both meaningful and useful inferential tools (Berger, Bernardo and Sun (2009), Bernardo (1979), Gelman et al. (2008), Jeffreys (1946), Martin and Walker (2019), Mukerjee and Reid (1999), Staicu and Reid (2008)). 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 log-transformed concentration of p elements and record the measurements as a column vector $y_{s,k} \in \mathbb{R}^p$ for $k \in \{1,\dots,m\}$. Then, assuming a multivariate Gaussian data-generating equation, as in Aitken and Lucy (2004) and Ommen, Saunders and Neumann (2017), for $k \in \{1,\dots,m\}$,

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

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

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

where $q_s(\mu_s, A|\{y_{s,k}\}) := f_s(\{y_{s,k}\}|\mu_s, A) \cdot J_s(\{y_{s,k}\}, (\mu_s, A))$ is the unnormalized GF density with normalizing constant c_s, $f_s(\cdot|\mu_s, A)$ is a multivariate Gaussian density and the Jacobian term $J_s(\{y_{s,k}\}, (\mu_s, A))$ is computed below.

For the construction of a GFF, similar to a BF, the identifiability of the matrix A in a data-generating equation of the form in (3) is superfluous because we are not interested in inference on the parameters of this model. The utility of defining (3) as the data-generating equation for multivariate Gaussian data is that it is unnecessary to add any constraints to the A matrix to guarantee that AA^T is symmetric and nonnegative definite (i.e., a covariance matrix) so that $Y_{s,k} \sim \mathrm{N}(\mu_s, AA^T)$ is well defined. Not needing to impose constraints on the A matrix makes derivation of the gradient $\nabla_{(\mu_s,A)}G$ uncomplicated, and it avoids dealing with problematic mixing conditions in Markov chain Monte Carlo (MCMC) computations (e.g., the same reason why it is common to perform MCMC on the transformation $\log(\sigma)$ rather than σ for some parameter $\sigma > 0$). Moreover, (3) is a well-studied choice of data-generating equation for Gaussian linear models in the GF literature (see, e.g., Hannig et al. (2016), Shi et al. (2021), Williams and Hannig (2019), Williams, Xie and Hannig (2019)), and so it is most prudent to construct and study a first ever GFF using this data-generating equation. Consequences on a GF distribution from constraining the matrix A (for concerns relating to the identifiability of A, etc.) are an area of active research (see, e.g., Murph, Hannig and Williams (2020)).

As in Shi et al. (2021), denote by w the vector of length mp obtained by stacking the vectors $y_{s,1}, \ldots, y_{s,m}$ on top of each other. Applying definition (2) gives $J_s(\{y_{s,k}\}, (\mu_s, A))$, 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

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

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,

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

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 log-transformed concentration of p elements for source (i.e., window pane) $i \in \{1, \ldots, n\}$, 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

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

where $q_a(\mu_a, B, C|\{t_i\}, \{y_{a,i,k}\}) = f_a(\{y_{a,i,k}\}|\mu_a, B, C, \{t_i\}) \cdot J_a(\{y_{a,i,k}\}, (\mu_a, B, C))$ is the unnormalized GF density with normalizing constant c_a and $f_a(\cdot|\mu_a, B, C, \{t_i\})$ is a multivariate Gaussian density. To compute the Jacobian term, 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 $J_a(\{y_{a,i,k}\}, (\mu_a, B, C))$, 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

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

where

$$W := \begin{pmatrix} 1_{m_1} \otimes t_1' \\ \vdots \\ 1_{m_n} \otimes t_n' \end{pmatrix} \quad \text{and} \quad \widetilde{Q} := \begin{pmatrix} v_{1,1}' \\ \vdots \\ v_{1,m_1}' \\ \vdots \\ v_{n,1}' \\ \vdots \\ v_{n,m_n}' \end{pmatrix} = \underbrace{\begin{pmatrix} (y_{a,1,1} - \mu_a - Bt_1)' \\ \vdots \\ (y_{a,1,m_1} - \mu_a - Bt_1)' \\ \vdots \\ (y_{a,n,1} - \mu_a - Bt_n)' \\ \vdots \\ (y_{a,n,m_n} - \mu_a - Bt_n)' \end{pmatrix}}_{=:Q} (C^{-1})'.$$

Thus,

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

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 term, which is the analogue for the prior in GFI, is a function of the data. To illustrate this distinction, 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 log-transformed 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/training data, 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, as expressed in Kass and Raftery (1995), is

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

In the case that improper prior densities are assigned for θ_s and θ_a, the BF in (7) is defined in the sense of an intrinsic BF using reference priors, as in Berger and Pericchi (1996), with $\{y_{s,k}\}$ and $\{y_{a,i,k}\}$ serving as training samples leading to proper posterior densities $\pi_s(\cdot|\{y_{s,k}\})$ and $\pi_a(\cdot|\{y_{a,i,k}\})$, respectively. Using reference priors is one strategy for dealing with the arbitrary normalizing constant that appears in a BF constructed from improper priors (see Berger et al. (2001) for a discussion on this problem). We define the GFF by analogy to equation (7), but note that the GFF has the advantage that the normalizing constant is determined by the data-generating equation, and is thus *not arbitrary*. It is demonstrated by Theorem 4 in Hannig et al. (2016) that the ratio of normalizing constants from GF densities are meaningful for model selection.

The GF densities $r_s(\cdot|\{y_{s,k}\})$ and $r_a(\cdot|\{y_{a,i,k}\})$ are proper density functions and share similar large-sample properties as posterior density functions in the sense of Bernstein-von Mises type theory. Hence, by analogy, we define

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

In the remaining sections of this paper, we demonstrate empirically that the defined GFF both has 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

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

The numerator is the expected value of $f_s(\{y_{u,l}\}|\mu_s, A)$ (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 $f_s(\{y_{u,l}\}|\mu_s, A)$ over a GF sample of the parameters μ_s and A. We thus construct a random walk Metropolis–Hastings 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 importance samples of the $\{T_i\}$ are needed to properly identify BB' and CC'. Accordingly, we construct the following point approximations for the unobserved random effects.

First, construct the statistics

$$\begin{aligned}
\widehat{\mu}_a &:= \frac{1}{N}\sum_{i=1}^{n}\sum_{k=1}^{m_i} y_{a,i,k},\\
\widehat{B}\widehat{B}' &:= \frac{\tau-2}{\tau}\cdot\frac{1}{n-1}\sum_{i=1}^{n}(\bar{y}_{a,i,\cdot}-\widehat{\mu}_a)(\bar{y}_{a,i,\cdot}-\widehat{\mu}_a)',\\
\widehat{C}\widehat{C}' &:= \frac{1}{N-1}\sum_{i=1}^{n}\sum_{k=1}^{m_i}(y_{a,i,k}-\bar{y}_{a,i,\cdot})(y_{a,i,k}-\bar{y}_{a,i,\cdot})',
\end{aligned} \tag{9}$$

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. Substituting these statistics into data-generating equation (5) yields the repeated observations regression model, $Y_{a,i,k} - \widehat{\mu}_a = \widehat{B}t_i + \widehat{C}V_{i,k}$, for $k \in \{1, \ldots, m_i\}$ and $i \in \{1, \ldots, n\}$, where $t_1, \ldots, t_n$ are the realized but unobserved values of the random effects coefficients that generated the data. Averaging over each measurement k and rescaling the systems of equations gives the Gauss–Markov model

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

with the resulting least squares solution $\widehat{t}_i := (\widehat{B}'(\widehat{C}\widehat{C}')^{-1}\widehat{B})^{-1}\widehat{B}'(\widehat{C}\widehat{C}')^{-1}(\bar{y}_{a,i,\cdot} - \widehat{\mu}_a)$ for every source $i \in \{1, \ldots, n\}$ in the alternative source data set.

Using $\{\widehat{t}_i\}$, we estimate the GFF as

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

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 (DiCiccio et al. (1997), Gelman and Meng (1998), Meng and Wong (1996)). Other popular approaches include importance, bridge and path sampling (Gelman and 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 supplied by the NFI, but the NFI was not further involved in this research. Currently, these data are not publicly available but 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 log-transformed 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 is considered. Further details of these data are documented in van Es et al. (2017).

In the context of our formulation, 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 glass fragment measurements 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.

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*

Element	Ti49	Sr88	K39	Zr90	Mn55	Ba137	Ce140	La139	**Pb208**	**Rb85**
st dev	0.00000	0.00001	0.00001	0.00001	0.00001	0.00002	0.00006	0.00007	**0.00012**	**0.00013**

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, Saunders and Neumann (2017) and Ommen and Saunders (2019) (see their *specific source* formulation). The LR is defined in Chapter 7.2 of Ommen (2017) and uses plug-in MLEs of the parameter values under each hypothesis. For reference, the exact details of the BF and LR 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 log-transformed 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 consists only of two measurements, consistent with a sure independence screening strategy (Fan and 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 (i.e., window panes) and glass fragment measurements in the alternative source data set (3×659 measurements in total). Table 1 presents the variance observed for the rescaled, log-transformed concentrations of each of the 10 elements from which we select elements "Pb208" and "Rb85."

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

$$\widehat{\mu}_s := \frac{1}{m}\sum_{k=1}^{m} y_{s,k},$$

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

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 3000 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 3000 simulated GFF, BF and LR values for H_d true are computed using 3000 nonassociated

pairs of unknown and specific source data sets (the specific sources are randomly selected from among the 320 for each of the 3000 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 3320 simulated events. The LR crashed for one of the 3000 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 2999 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 3000 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*. First, Figure 1 presents a box plot of the performance of the GFF, BF, and LR over the 3000 simulations under H_d and 320 simulations under H_p. The BF is evaluated for three different prior specifications: a prior that favors the prosecution hypothesis, denoted "BF_p prior." a prior that favors the defense hypothesis, denoted "BF_d prior," and an "oracle" prior centered on the parameter values used to generate the synthetic data, denoted "BF_o prior." Note that the parameter

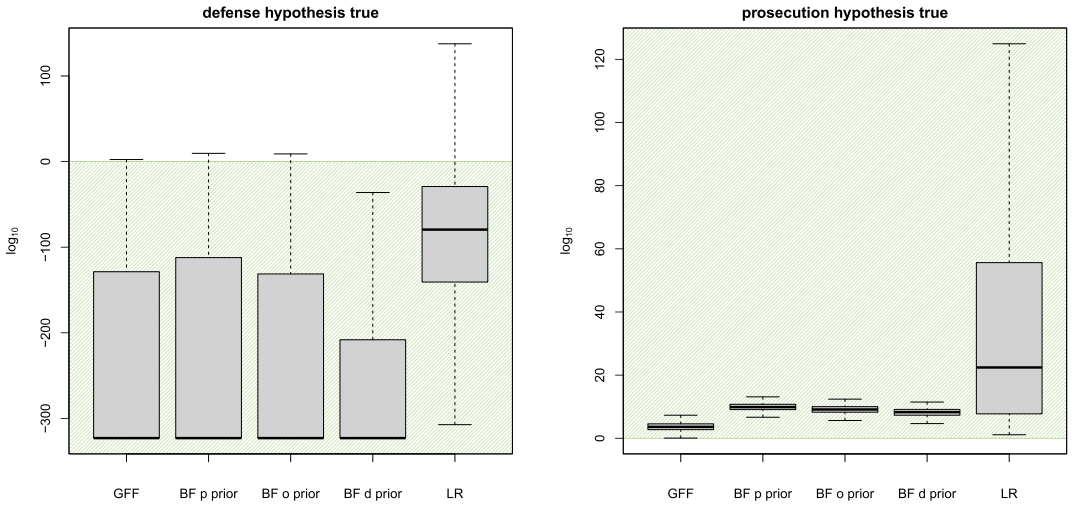

FIG. 1. *Box plots of the sampling distributions of the GFF, BF and LR over the* 3000 *simulations under* H_d (*left panel*) *and* 320 *simulations under* H_p (*right panel*). *For this synthetic "ideal sample size" simulation,* $m_u = 2$, $m = 150$, $n = 659$ *and* $m_i = 3$. *BF_p prior denotes the BF constructed from priors that favor* H_p, *whereas BF_d prior denotes the BF constructed from priors that favor* H_d. *The shaded regions in each panel correspond to values of the GFF, BF and LR that favor the true hypothesis. Outliers are omitted.*

values used to generate the synthetic data are defined from empirical analogues calculated from the real data; the precise details for each of these prior specifications is provided in the Appendix.

Figure 1 demonstrates that all five 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 vs. BF_d prior is a result of the unrealistically ideal sample sizes of this synthetic simulation design. The next simulation design illustrates this point.

Second, the fiducial distributions of the area under the receiver operating characteristic curve (AUC) for the GFF, BF and LR are displayed in Figure 2. The AUC measures the adequacy of each of the five 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 1. There is almost no overlap in the observed GFF, BF_p prior and BF_d prior values, respectively, for H_d true vs. 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 2. However, the LR values exhibit some overlap in tail values between H_d true vs. 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 three 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 three instances of an LR of 3 when H_p is true. As described in Hannig and Iyer (2022), a shorthand for this notion of calibration is the the expression "LR(LR) = LR."

We follow the method in Hannig and Iyer (2022) for estimating the calibration of the 3000 simulations under H_d and 320 simulations under H_p, for GFF, BF and LR; see Figure 3 for the estimated calibrations, and observe that the GFF is the best calibrated of the five 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

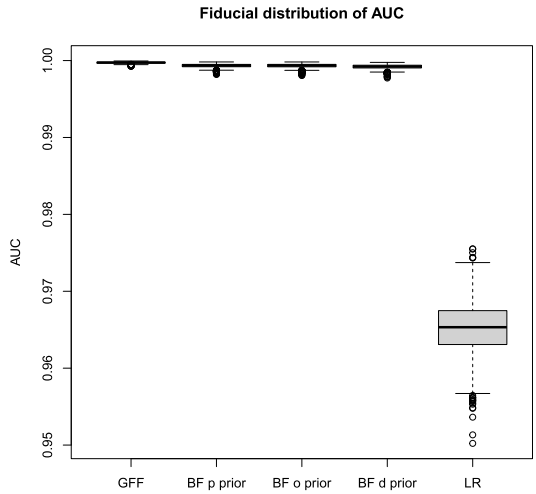

FIG. 2. *Fiducial distributions of the AUC for the GFF, BF and LR over the* 3000 *simulations under* H_d *and* 320 *simulations under* H_p. *For this "ideal sample size" simulation,* $m_u = 2$, $m = 150$, $n = 659$ *and* $m_i = 3$.

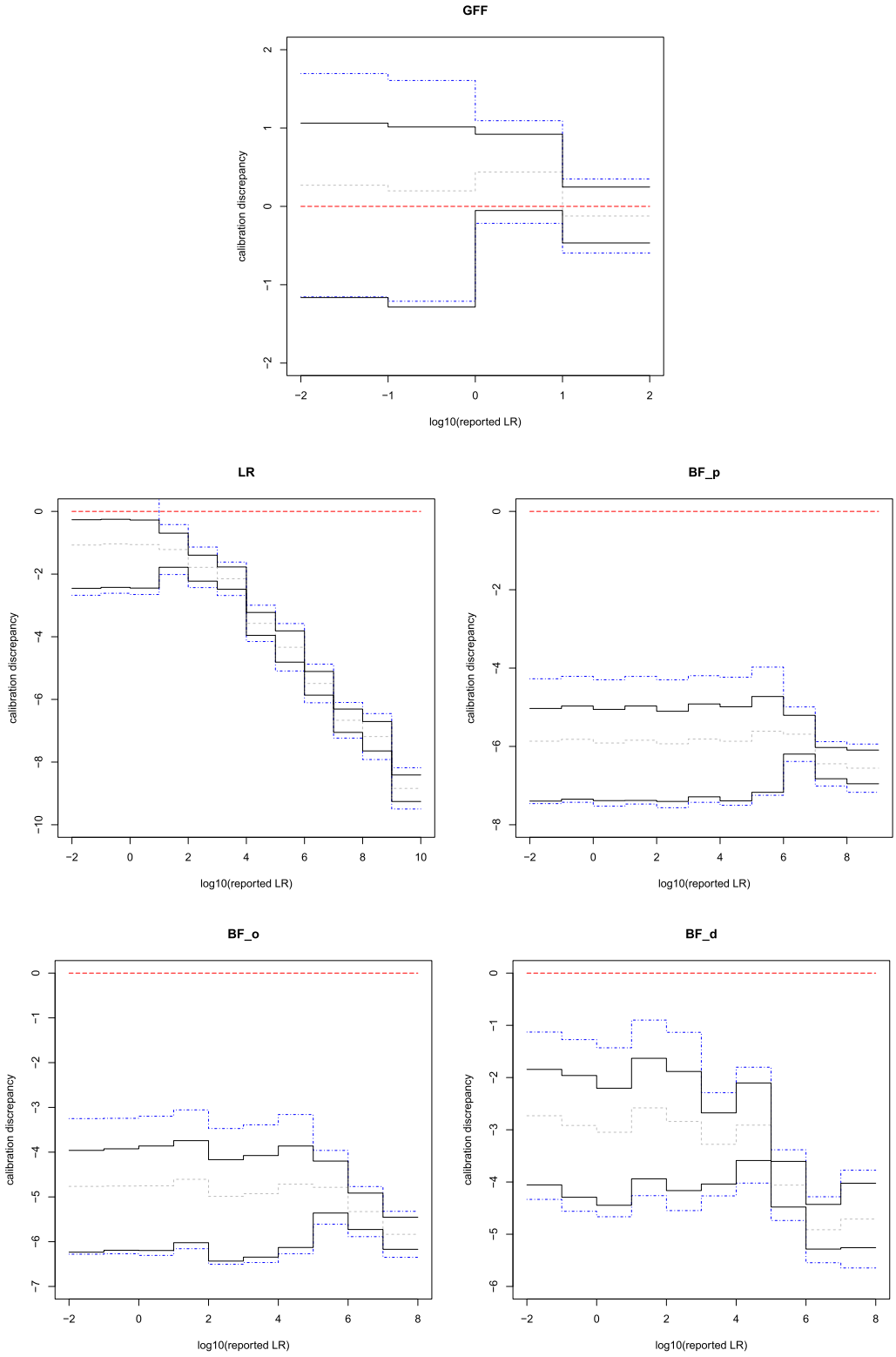

FIG. 3. *Calibration for the GFF, BF and LR over the* 3000 *simulations under* H_d *and* 320 *simulations under* H_p. *The horizontal dashed line at zero corresponds to perfect calibration* (*i.e.*, $LR(LR) = LR$). *The dotted grey line is the fiducial median log discrepancy. The solid black and dot-dashed lines are upper and lower* 0.95 *pointwise and simultaneous fiducial confidence intervals, respectively, for the log discrepancy. For this "ideal sample size" simulation,* $m_u = 2$, $m = 150$, $n = 659$, *and* $m_i = 3$.

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.

Note that the calibration plots are constructed in Hannig and Iyer (2022), and they apply to any likelihood ratios (i.e., BF, GFF and LR). The calibration discrepancy is based on the fiducial distributions of the empirical distribution functions (for further details on these, see Cui and Hannig (2019)) of the likelihood ratios. The fiducial median log discrepancy is the median of the fiducial samples of the calibration discrepancy. These fiducial samples are *not* to be confused with the fiducial distributions that underly the GFF.

We conclude this section by presenting an alternative calibration analysis described in Ramos and Gonzalez-Rodriguez (2008); see Figure 4. It is again observed that the GFF values

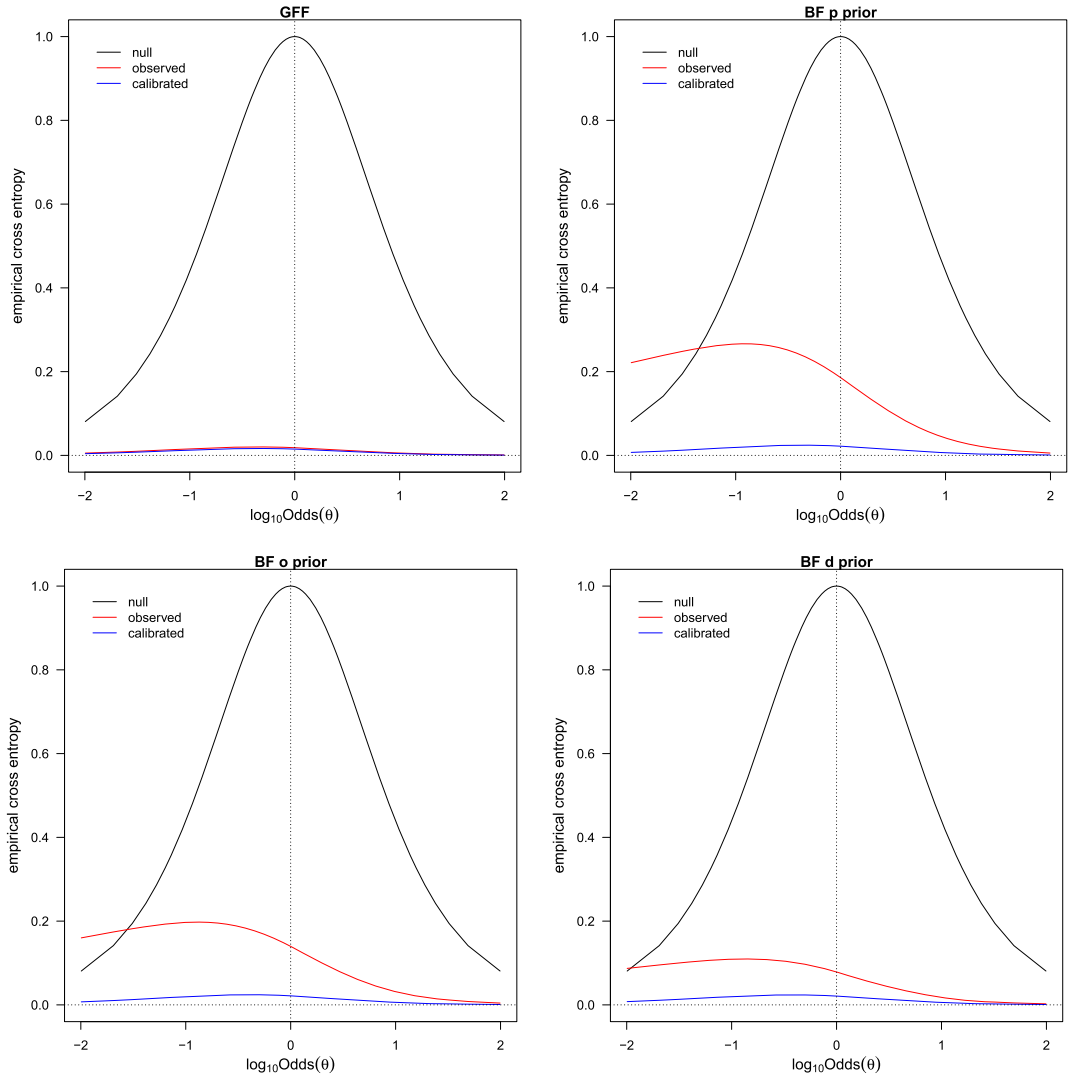

FIG. 4. *Empirical cross entropy for the GFF, BF and LR over the* 3000 *simulations under* H_d *and* 320 *simulations under* H_p. *This calibration diagnostic tool is proposed in Ramos and Gonzalez-Rodriguez* (2008). *Good calibration is exhibited when the red line is nested between the blue and black lines and as close as possible the blue. The code from Ramos and 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* $m_i = 3$.

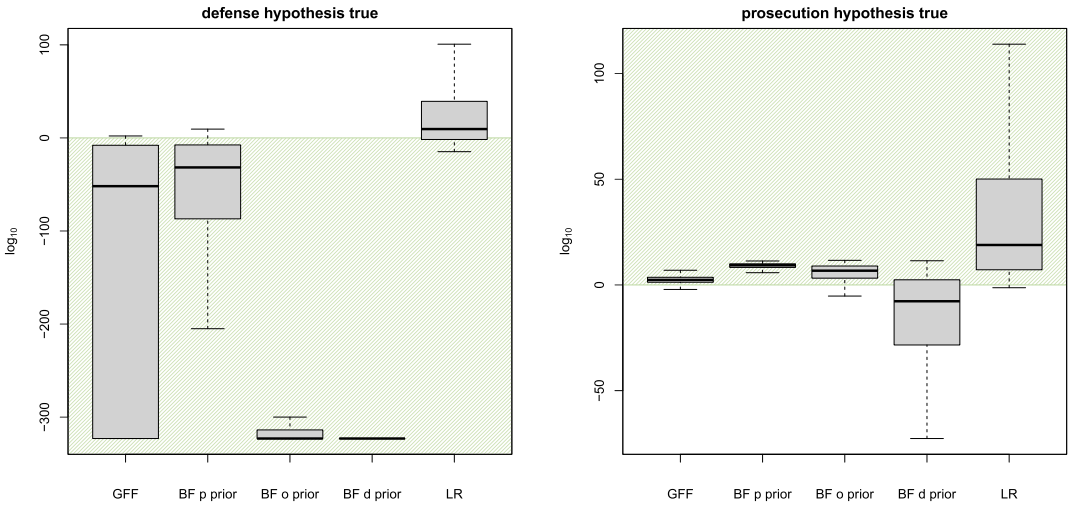

FIG. 5. *Box plots of the sampling distributions of the GFF, BF and LR over the* 3000 *simulations under* H_d (*left panel*) *and* 320 *simulations under* H_p (*right panel*). *For this synthetic "NFI casework data sample sizes" simulation,* $m_u = 2$, $m = 3$, $n = 659$ *and* $m_i = 3$. *BF_p prior 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 regions in each panel correspond to values of the GFF, BF and LR that favor the true hypothesis. Outliers are omitted.*

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 data sample sizes*. The sampling distributions are displayed in Figure 5. 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 $f_s(\{y_{s,k}\}|\widehat{\theta}_s)$ in the denominator.

The next feature to observe in Figure 5 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 and 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 5, 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 vs. defense) for or against the hypothesis, even to the point where the BF entirely favors the wrong hypothesis. Conversely, it is observed in Figure 5 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 5 are further supported by the discrimination and calibration analyses presented in Figures 6 and 7, respectively.

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 five methods is BF_d prior values, as displayed in Figure 5. There is a clear distinction between the distribution of BF_d prior values under H_d vs. H_p, even though both

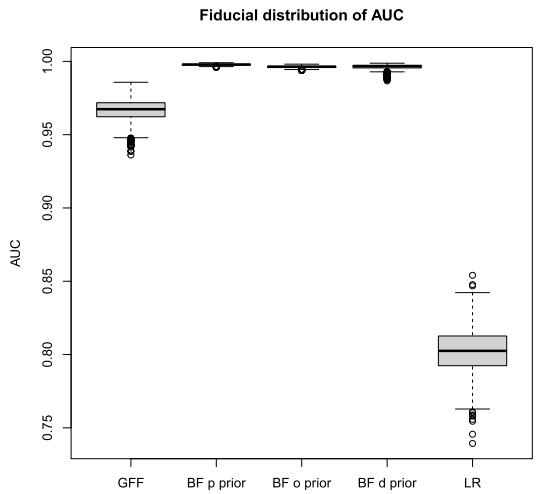

FIG. 6. *Fiducial distributions of the AUC for the GFF, BF and LR over the* 3000 *simulations under* H_d *and* 320 *simulations under* H_p. *For this synthetic "NFI casework data sample sizes" simulation,* $m_u = 2$, $m = 3$, $n = 659$ *and* $m_i = 3$.

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 6 (along with that for the other three methods as well). Notice that, in Figure 6 as well as Figure 9 for the real NFI data, the BF values exhibit improved discrimination over the GFF values due to the fact that the BF values are characterized by a larger magnitude of separation for the competing hypotheses. However, such a large magnitude of separation of these values that results in improved discrimination comes at the cost of deteriorated calibration, as demonstrated in Figures 7 and 10 (in fact, the calibration software crashed for the BF_d prior in this simulation design). 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 5.

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 5) and are effective at discriminating between H_d and H_p (i.e., Figure 6), there is still a danger that they are not calibrated to appropriately reflect the strength of evidence that their values suggest. Figure 7 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 7 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*. Once again, the resulting sampling distributions of the methods are presented as box plots in Figure 8. The fiducial distributions of the AUC to assess discrimination effectiveness between the hypotheses are presented in Figure 9, and the calibration analysis is displayed in Figure 10 and 11. Four instances out of the 3320 simulations corresponding to one extreme outlier window (of the 320 specific/unknown source windows) have been removed from the analysis in Figure 10; for comparison, the full 3320 simulations are shown 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

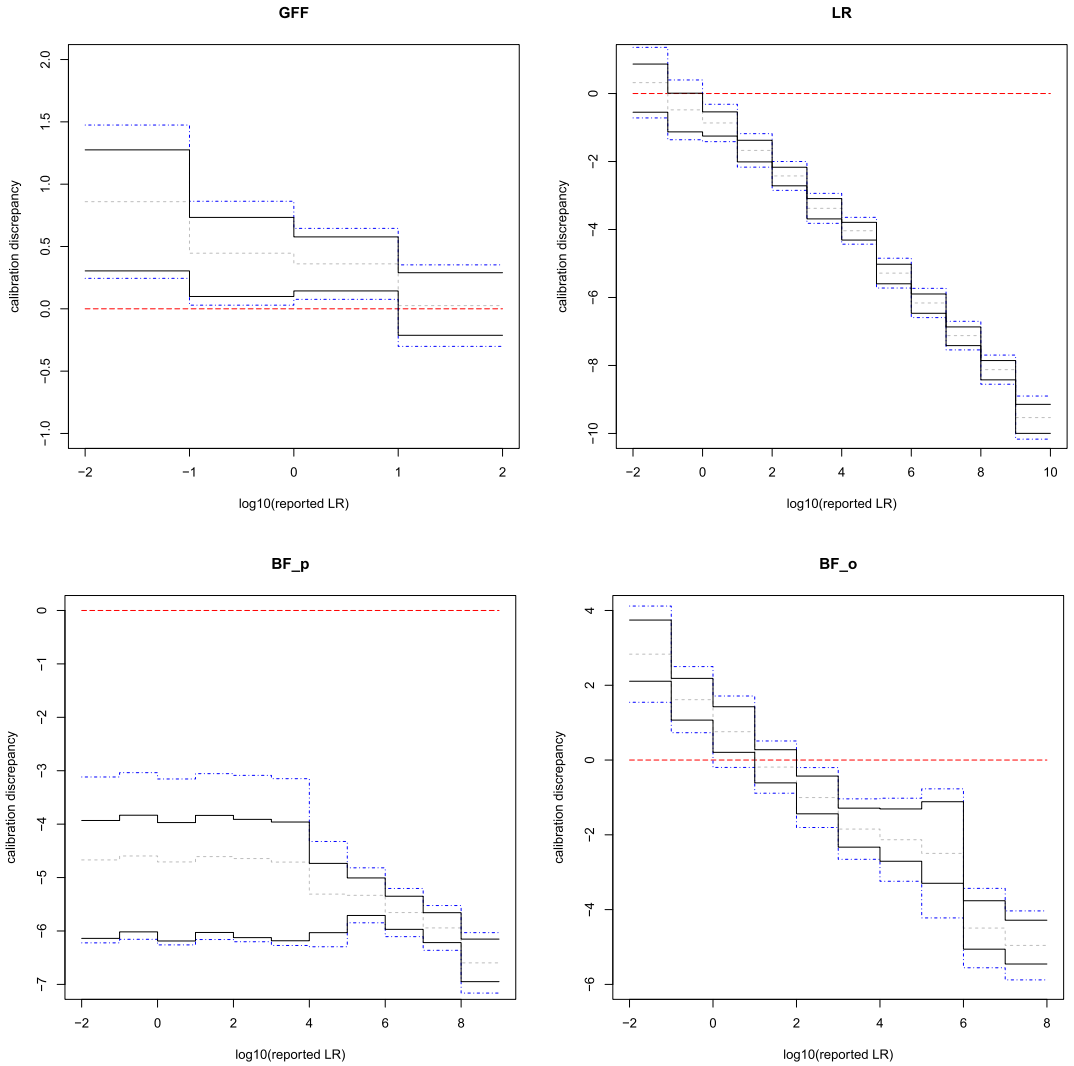

FIG. 7. *Calibration for the GFF, BF and LR over the* 3000 *simulations under* H_d *and* 320 *simulations under* H_p. *The horizontal dashed line at zero corresponds to perfect calibration* (*i.e.*, $LR(LR) = LR$). *The dotted grey line is the fiducial median log discrepancy. The solid black and dot-dashed lines are upper and lower* 0.95 *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* $m_i = 3$.

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. The GFF tends to less extreme values than it did for the synthetic data, most noticeably for the H_d true scenario, but nonetheless, the calibration of the values suggests that the GFF is relatively well calibrated (aside from the four simulations due to the one outlier window).

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.

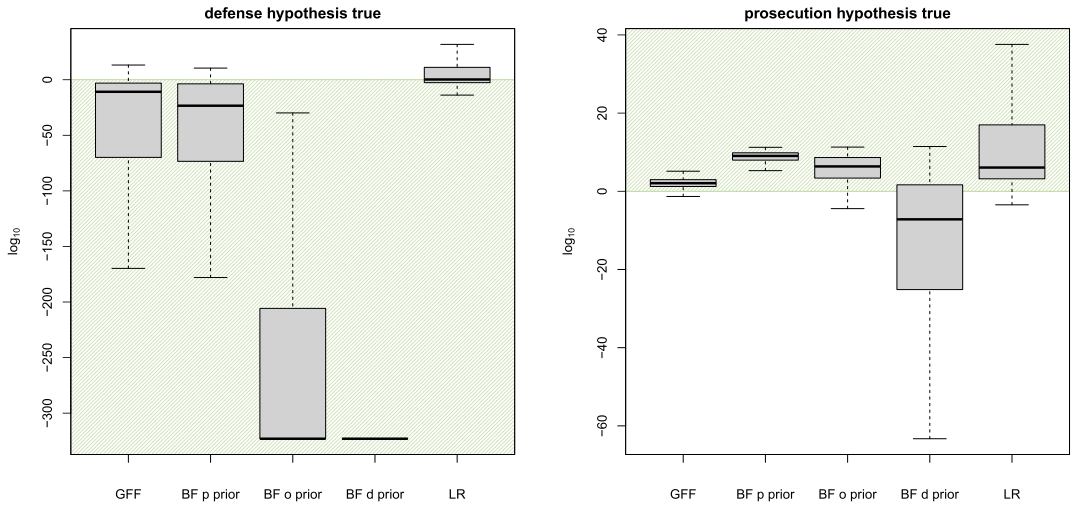

FIG. 8. *Box plots of the sampling distributions of the GFF, BF and LR over the* 3000 *simulations under* H_d (*left panel*) *and* 320 *simulations under* H_p (*right panel*). *For this "real NFI casework data" simulation,* $m_u = 2$, $m = 3$, $n = 659$ *and* $m_i = 3$. *BF_p prior denotes the BF constructed from priors that favor* H_p, *whereas BF_d prior denotes the BF constructed from priors that favor* H_d. *The shaded regions in each panel correspond to values of the GFF, BF and LR that favor the true hypothesis. Outliers are omitted.*

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 vs. 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,

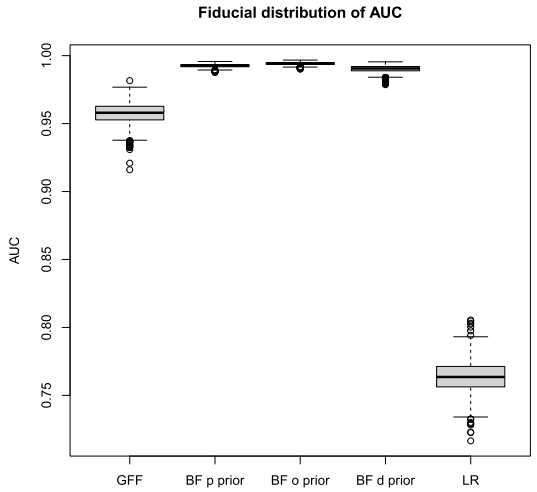

FIG. 9. *Fiducial distributions of the AUC for the GFF, BF and LR over the* 3000 *simulations under* H_d *and* 320 *simulations under* H_p. *For this "real NFI casework data" simulation,* $m_u = 2$, $m = 3$, $n = 659$ *and* $m_i = 3$.

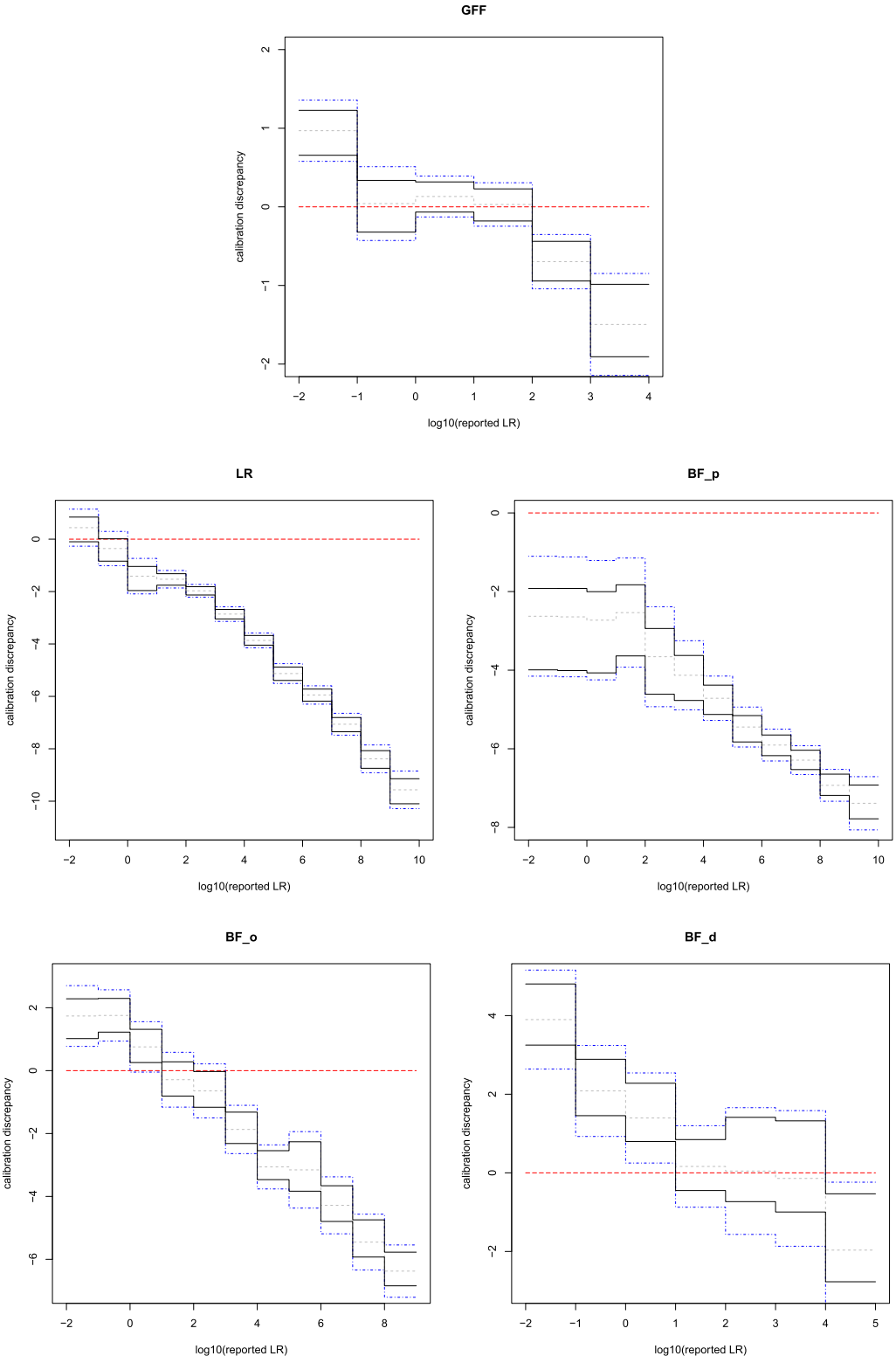

FIG. 10. *Calibration for the GFF, BF and LR over the* 3000 *simulations under* H_d *and* 320 *simulations under* H_p. *The horizontal dashed line at zero corresponds to perfect calibration* (*i.e.*, $LR(LR) = LR$). *The dotted grey line is the fiducial median log discrepancy. The solid black and dot-dashed lines are upper and lower* 0.95 *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* $m_i = 3$. *Four outliers have been removed to produce these plots.*

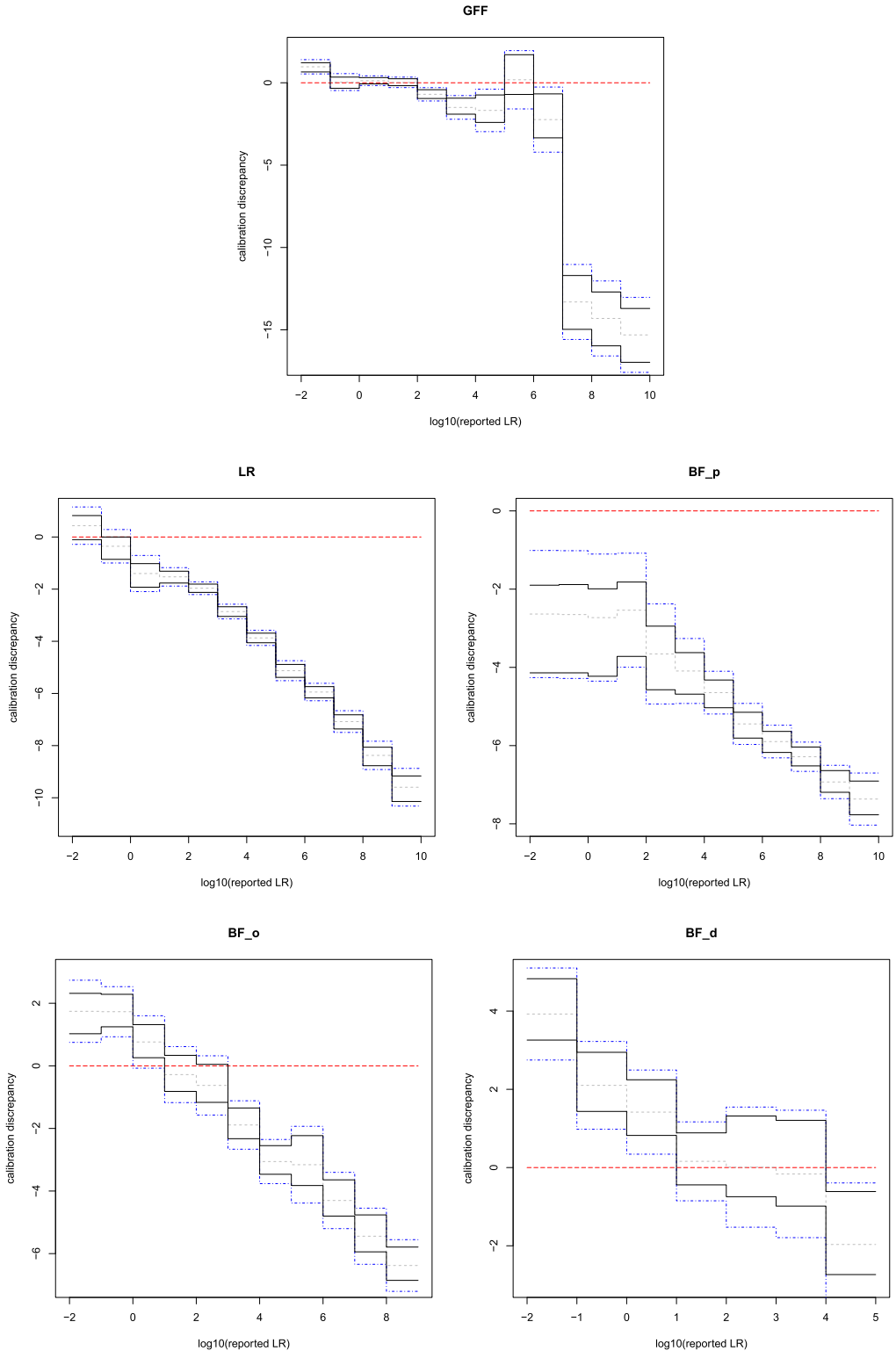

FIG. 11. *Calibration for the GFF, BF and LR over the* 3000 *simulations under* H_d *and* 320 *simulations under* H_p. *The horizontal dashed line at zero corresponds to perfect calibration* (*i.e.*, $LR(LR) = LR$). *The dotted grey line is the fiducial median log discrepancy. The solid black and dot-dashed lines are upper and lower* 0.95 *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* $m_i = 3$. *No outliers have been removed to produce these plots.*

and they were not calibrated to represent the strength of evidence they appeared to convey. 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 evaluated a GFF as an alternative methodological approach and tool that does not suffer from the demonstrated deficiencies in the BF and LR. Moreover, 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.

Lastly, while it is beyond the scope of our current investigation, an obvious next step is the development of a GFF (and fiducial factors, more broadly) from a formal decision theoretic perspective. Compelling ideas for fiducial decision theory were introduced in the seminal paper Taraldsen and Lindqvist (2013). Developments since then are contributed and summarized in the recent preprints (and references therein) Taraldsen and Lindqvist (2021) and Martin (2021), the latter within the inferential models framework (Martin and Liu (2016)).

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, Saunders and Neumann (2017), Ommen and Saunders (2019). Assuming the posterior distributions of θ_s and θ_a are independent, the BF from equation (7) is expressed as

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

where

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

is the posterior distribution of (θ_s,θ_a) under the defense hypothesis that the unknown source data are generated from the alternative source and is constructed after including the unknown source data as part of the alternative source dataset. 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, Saunders and Neumann (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 our simulation studies throughout Section 4, the prosecution, oracle and defense priors are specified as

$$\mu_\pi = \begin{cases} (1,1)' & \text{for BF\_p and BF\_d prior specification,} \\ (0.6322523, 0.6265417)' & \text{for BF\_o prior specification,} \end{cases}$$

$$\Sigma_b = \begin{cases} 1000 \cdot \begin{pmatrix} 0.13 & 0.03 \\ 0.03 & 0.13 \end{pmatrix} & \text{for BF\_p prior specification,} \\ \begin{pmatrix} 0.07615601 & 0.02221717 \\ 0.02221717 & 0.08497993 \end{pmatrix} & \text{for BF\_o prior specification,} \\ 0.1 \cdot \begin{pmatrix} 0.13 & 0.03 \\ 0.03 & 0.13 \end{pmatrix} & \text{for BF\_d prior specification,} \end{cases}$$

$$\Sigma_e = \begin{cases} \begin{pmatrix} 4.5 \cdot 10^{-4} & 5 \cdot 10^{-5} \\ 5 \cdot 10^{-5} & 4.5 \cdot 10^{-4} \end{pmatrix} & \text{for BF\_p and BF\_d prior specification,} \\ \begin{pmatrix} 8.954729 \cdot 10^{-4} & 4.636142 \cdot 10^{-5} \\ 4.636142 \cdot 10^{-5} & 5.302711 \cdot 10^{-4} \end{pmatrix} & \text{for BF\_o prior specification,} \end{cases}$$

$$\nu_e = \begin{cases} 5 & \text{for BF\_p prior specification,} \\ 27 & \text{for BF\_o prior specification,} \\ 100 & \text{for BF\_d prior specification,} \end{cases}$$

and $\nu_b = 27$ and $k = 10$ for all prior specifications. These BF_p and BF_d prior specifications are prescribed in Ommen, Saunders and Neumann (2017) (and the accompanying code). The BF_o prior specifications are set as $\mu_\pi = \widehat{\mu}_a$, $\Sigma_b = \widehat{B}\widehat{B}'$ and $\Sigma_e = \widehat{C}\widehat{C}'$, as in (9) using the alternative source data set. These values for $\widehat{\mu}_a$, $\widehat{B}$ and $\widehat{C}$ are used to generate the synthetic data in our simulation studies.

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

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

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

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

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 in our Supplementary Material (Williams, Ommen and Hannig (2023), also available at https://jonathanpw.github.io/research.html).

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

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

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 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 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.

SUPPLEMENTARY MATERIAL

Time capsuled code (DOI: 10.1214/22-AOAS1632SUPP; .zip). This Supplementary Material contains the code used to produce all of the results presented in this manuscript. The glass fragment data set that we investigate (van Es et al. (2017)) was kindly supplied by the NFI, but the NFI was not further involved in this research. Currently, these data are not publicly available, but are available on request by emailing p.zoon@nfi.nl.

REFERENCES

AITKEN, C. G. G. and LUCY, D. (2004). Evaluation of trace evidence in the form of multivariate data. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **53** 109–122. MR2037883 https://doi.org/10.1046/j.0035-9254.2003.05271.x

BERGER, C. E. H. and SLOOTEN, K. (2016). The LR does not exist. *Sci. Justice* **56** 388–391. https://doi.org/10.1016/j.scijus.2016.06.005

BERGER, J. O., BERNARDO, J. M. and SUN, D. (2009). The formal definition of reference priors. *Ann. Statist.* **37** 905–938. MR2502655 https://doi.org/10.1214/07-AOS587

BERGER, J. O. and PERICCHI, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *J. Amer. Statist. Assoc.* **91** 109–122. MR1394065 https://doi.org/10.2307/2291387

BERGER, J. O., PERICCHI, L. R., GHOSH, J., SAMANTA, T., DE SANTIS, F., BERGER, J. and PERICCHI, L. (2001). Objective Bayesian methods for model selection: Introduction and comparison. In *Lecture Notes—Monograph Series* 135–207.

BERNARDO, J.-M. (1979). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. Ser. B* **41** 113–147. MR0547240

BIEDERMANN, A., BOZZA, S., TARONI, F. and AITKEN, C. G. G. (2016). Reframing the debate: A question of probability, not of likelihood ratio. *Sci. Justice* **56** 392–396.

BOLCK, A., WEYERMANN, C., DUJOURDY, L., ESSEIVA, P. and VAN DEN BERG, J. (2009). Different likelihood ratio approaches to evaluate the strength of MDMA tablet comparisons. *Forensic Sci. Int.* **191** 42–51.

CUI, Y. and HANNIG, J. (2019). Nonparametric generalized fiducial inference for survival functions under censoring. *Biometrika* **106** 501–518. MR3992384 https://doi.org/10.1093/biomet/asz016

DETTMAN, J. R., CASSABAUM, A. A., SAUNDERS, C. P., SNYDER, D. L. and BUSCAGLIA, J. (2014). Forensic discrimination of copper wire using trace element concentrations. *Anal. Chem.* **86** 8176–8182. https://doi.org/10.1021/ac5013514.

DICICCIO, T. J., KASS, R. E., RAFTERY, A. and WASSERMAN, L. (1997). Computing Bayes factors by combining simulation and asymptotic approximations. *J. Amer. Statist. Assoc.* **92** 903–915. MR1482122 https://doi.org/10.2307/2965554

EGLI, N. M., CHAMPOD, C. and MARGOT, P. (2006). Evidence evaluation in fingerprint comparison and automated fingerprint identification systems—modeling between finger variability. *Forensic Sci. Int.* **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. *J. Forensic Sci.* **9** 209–217.

EVETT, I. W. (1986). A Bayesian approach to the problem of interpreting glass evidence in forensic science casework. *J. - Forensic Sci. Soc.* **26** 3–18.

FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 849–911. MR2530322 https://doi.org/10.1111/j.1467-9868.2008.00674.x

GELMAN, A. and MENG, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statist. Sci.* **13** 163–185. MR1647507 https://doi.org/10.1214/ss/1028905934

GELMAN, A., JAKULIN, A., PITTAU, M. G. and SU, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Ann. Appl. Stat.* **2** 1360–1383. MR2655663 https://doi.org/10.1214/08-AOAS191

GONZALEZ-RODRIGUEZ, J., DRYGAJLO, A., RAMOS-CASTRO, D., GARCIA-GOMAR, M. and ORTEGA-GARCIA, J. (2006). Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. *Comput. Speech Lang.* **20** 331–355.

GONZALEZ-RODRIGUEZ, J., FIERREZ-AGUILAR, J., RAMOS-CASTRO, D. and ORTEGA-GARCIA, J. (2005). Bayesian analysis of fingerprint, face and signature evidences with automatic biometric systems. *Forensic Sci. Int.* **155** 126–140.

GROVE, D. M. (1980). The interpretation of forensic evidence using a likelihood ratio. *Biometrika* **67** 243–246. MR0570530 https://doi.org/10.1093/biomet/67.1.243

HANNIG, J. and IYER, H. (2022). Testing for calibration discrepancy of reported likelihood ratios in forensic science. *J. Roy. Statist. Soc. Ser. A* **185** 267–301. MR4384306

HANNIG, J., IYER, H., LAI, R. C. S. and LEE, T. C. M. (2016). Generalized fiducial inference: A review and new results. *J. Amer. Statist. Assoc.* **111** 1346–1361. MR3561954 https://doi.org/10.1080/01621459.2016.1165102

HEPLER, A., SAUNDERS, C. P., DAVIS, L. and BUSCAGLIA, J. (2012). Score-based likelihood ratios for handwriting evidence. *Forensic Sci. Int.* **219** 129–140.

JEFFREYS, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. R. Soc. Lond. Ser. A, Math. Phys. Sci.* **186** 453–461. MR0017504 https://doi.org/10.1098/rspa.1946.0056

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

KASS, R. E. and RAFTERY, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90** 773–795. MR3363402 https://doi.org/10.1080/01621459.1995.10476572

LINDLEY, D. V. (1971). *Bayesian Statistics, a Review*. *Conference Board of the Mathematical Sciences Regional Conference Series in Applied Mathematics*, *No*. 2. SIAM, Philadelphia, PA. MR0329081

LINDLEY, D. V. (1977). A problem in forensic science. *Biometrika* **64** 207–213. MR0518935 https://doi.org/10.1093/biomet/64.2.207

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

LUND, S. P. and IYER, H. (2017). Likelihood ratio as weight of forensic evidence: A closer look. *J. Res. Natl. Inst. Stand. Technol.* **122** 1–32.

MARTIN, R. (2021). Inferential models and the decision-theoretic implications of the validity property. Preprint. Available at arXiv:2112.13247.

MARTIN, R. and LIU, C. (2016). *Inferential Models*: *Reasoning with Uncertainty*. *Monographs on Statistics and Applied Probability* **147**. CRC Press, Boca Raton, FL. MR3618727

MARTIN, R. and WALKER, S. G. (2019). Data-driven priors and their posterior concentration rates. *Electron. J. Stat.* **13** 3049–3081. MR4010592 https://doi.org/10.1214/19-ejs1600

MENG, X.-L. and WONG, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statist. Sinica* **6** 831–860. MR1422406

MORRISON, G. S. (2016). Special issue on measuring and reporting the precision of forensic likelihood ratios: Introduction to the debate. *Sci. Justice* **56** 371–373.

MUKERJEE, R. and REID, N. (1999). On a property of probability matching priors: Matching the alternative coverage probabilities. *Biometrika* **86** 333–340. MR1705343 https://doi.org/10.1093/biomet/86.2.333

MURPH, A., HANNIG, J. and WILLIAMS, J. (2020). Introduction to Generalized Fiducial Inference. In *CRC Press BFF Handbook*. To appear.

NEUMANN, C. and AUSDEMORE, M. A. (2020). Defence against the modern arts: the curse of statistics—Part II: 'Score-based likelihood ratios.' *Law Probab. Risk* **19** 21–42. https://doi.org/10.1093/lpr/mgaa006

NEUMANN, C., CHAMPOD, C., PUCH-SOLIS, R., EGLI, N. M., ANTHONIOZ, A. and BROMAGE-GRIFFITHS, A. (2007). Computation of likelihood ratios in fingerprint identification for configurations of any number of minutiae. *J. Forensic Sci.* **52** 54–64.

NEUMANN, C., HENDRICKS, J. and AUSDEMORE, M. A. (2020). Statistical support for conclusions in fingerprint examinations. In *Handbook of Forensic Statistics*, 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.

OMMEN, D. M. and SAUNDERS, C. P. (2019). Reconciling the bayes factor and likelihood ratio for two non-nested model selection problems. Preprint. Available at arXiv:1901.09798.

OMMEN, D. M., SAUNDERS, C. P. and NEUMANN, C. (2017). The characterization of Monte Carlo errors for the quantification of the value of forensic evidence. *J. Stat. Comput. Simul.* **87** 1608–1643. MR3625235 https://doi.org/10.1080/00949655.2017.1280036

PARKER, J. B. (1966). A statistical treatment of identification problems. *J. - Forensic Sci. Soc.* **6** 33–39.

PINHEIRO, J., BATES, D., DEBROY, S. and SARKAR, D. (2019). nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-140. https://CRAN.R-project.org/package=nlme.

RAMOS, D. and 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 *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I: Contributions to the Theory of Statistics* 575–586. Univ. California Press, Berkeley, CA. MR0133898

SHAFER, G. (1982). Lindley's paradox. *J. Amer. Statist. Assoc.* **77** 325–351. MR0664677

SHI, W. J., HANNIG, J., LAI, R. C. S. and LEE, T. C. M. (2021). Covariance estimation via fiducial inference. *Stat. Theory Relat. Fields* **5** 316–331. MR4335036 https://doi.org/10.1080/24754269.2021.1877950

STAICU, A.-M. and REID, N. M. (2008). On probability matching priors. *Canad. J. Statist.* **36** 613–622. MR2532255 https://doi.org/10.1002/cjs.5550360408

SWOFFORD, H., KOERNTNER, A., ZEMP, F., AUSDEMORE, M., LIU, A. and SALYARDS, M. (2018). A method for the statistical interpretation of friction ridge skin impression evidence: Method development and validation. *Forensic Sci. Int.* **287** 113–126.

TARALDSEN, G. and LINDQVIST, B. H. (2013). Fiducial theory and optimal inference. *Ann. Statist.* **41** 323–341. MR3059420 https://doi.org/10.1214/13-AOS1083

TARALDSEN, G. and LINDQVIST, B. (2021). Fiducial inference and decision theory. Preprint. Available at arXiv:2112.07060.

TARONI, F., BOZZA, S., BIEDERMANN, A. and AITKEN, C. G. G. (2016). Dismissal of the illusion of uncertainty in the assessment of a likelihood ratio. *Law Probab. Risk* **15** 1–16.

VAN ES, A., WIARDA, W., HORDIJK, M., ALBERINK, I. and VERGEER, P. (2017). Implementation and assessment of a likelihood ratio approach for the evaluation of la-icp-ms evidence in forensic glass analysis. *Sci. Justice* **57** 181–192.

WASSERSTEIN, R. L. and LAZAR, N. A. (2016). The ASA's statement on *p*-values: Context, process, and purpose [Editorial]. *Amer. Statist.* **70** 129–133. MR3511040 https://doi.org/10.1080/00031305.2016.1154108

WILLIAMS, J. P. and HANNIG, J. (2019). Nonpenalized variable selection in high-dimensional linear model settings via generalized fiducial inference. *Ann. Statist.* **47** 1723–1753. MR3911128 https://doi.org/10.1214/18-AOS1733

WILLIAMS, J. P., OMMEN, D. M. and HANNIG, J. (2023). Supplement to "Generalized fiducial factor: an alternative to the Bayes factor for forensic identification of source problems." https://doi.org/10.1214/22-AOAS1632SUPP

WILLIAMS, J. P., XIE, Y. and HANNIG, J. (2019). The EAS approach for graphical selection consistency in vector autoregression models. *Canad. J. Statist.* 1–28. To appear.

ZADORA, G., MARTYNA, A., RAMOS, D. and AITKEN, C. (2013). *Statistical Analysis in Forensic Science: Evidential Value of Multivariate Physicochemical Data*. Wiley, New York.