Document Title: Probabilistic Foundations for the use of the Logistic Regression Bayes Factor in Forensic Source Identification

Author(s): Dylan Borchert

Document Number: 308217

Date Received: December 2023

Award Number: N/A

Department of Justice
National Institute of Justice

SOUTH DAKOTA
STATE UNIVERSITY

Travel Support Report

Probabilistic foundations for the use of the logistic regression Bayes factor in forensic source identification

Travel Support Period: June 10 2023 – June 16 2023
Report for DOJ/NIJ Travel Support

Submitted: August 31, 2023

Submitted by:
Dylan Borchert*
dylan.borchert@jacks.sdstate.edu
(605) 688-6196

To:
Mr. Tom Wilson and Dr. Greg Dutton
U.S. Department of Justice
National Institute of Justice

*This is joint work with Peter Vergeer of the Netherlands Forensic Institute and Christopher P. Saunders of South Dakota State University.

Table of Contents

Acknowledgments

The presenter of this work was funded by the National Justice Institute in the form of a Travel Grant to attend the 11^{th} International Conference on Forensic Inference and Statistics (ICFIS) hosted in Lund, Sweden. This report is on the contents of a poster presentation given at ICFIS on June 13, 2023. This work was funded in part by the National Science Foundation through the National Research Traineeship site titled "Cyber-Physical-Social System for Understanding and Thwarting the Illicit Economy" under NSF award numbers: DGE 1828492, 1828302, 1828462, 1828288.

The materials presented here solely represent the views of the authors and do not represent the view of the US Department of Justice, the National Science Foundation, or the United States government.

Abstract

In comparison to likelihood ratios (LRs), Bayes factors (BFs) have the advantage that uncertainty in model parameter values is taken into account in a logical and coherent manner. In forensic literature, it is common to calculate BFs for generative models. It is also common to calculate LRs for discriminative models, for example using maximum likelihood (ML) estimates of logistic regression parameters. In this report, we present an approach to calculate BFs when using logistic regression as a model to discriminate between two classes. In logistic regression, the log of the LR between the two classes follows a functional form. We will focus on the case where this functional form is linear. This is equivalent to the log of the posterior odds of group membership following a linear model. We propose the calculation of the BF utilizing the posterior odds ratio, as well as using the LR function in the context of Ommen and Saunders, 2021. Using a database of simulated observations generated under two different models, we can obtain a posterior distribution for the parameters of the logistic regression, and use this distribution to obtain the posterior odds of group membership for a new observation with unknown membership. This posterior odds ratio can then be divided by the prior odds ratio to obtain the corresponding BF. An important note is that by constructing the database with a prespecified number of observations under each model, we are fixing the base rates. This removes the Bernoulli sampling process of the labels used to construct the likelihood function for the logistic regression, which will be discussed in the context of McLachlan, 2004. As a result, our discriminative model is an approximation to the latent generative models of the two classes. We study the convergence of the BF to the LR for two different BF calculations, and show that for large sample sizes they both converge. Also, we compare the calculated BFs of the two approaches to a reference BF, LR, and the plug-in estimate of the LR.

Introduction

In forensic source identification the forensic examiner is tasked with providing a value of evidence for some evidence, $\boldsymbol{E}$, that can allow a decision maker to make a logical and coherent decision about the source of $\boldsymbol{E}$, under two competing sampling models M_1 and M_2. One way of providing this value of evidence for a realization of evidence, $\boldsymbol{e}$, is through a likelihood ratio (LR) or a Bayes factor (BF), where $LR = \frac{f(\boldsymbol{e}|M_1)}{f(\boldsymbol{e}|M_2)}$. From a practical point of view, for the typically chosen priors, the BF is generally dampened (i.e. value of evidence closer to 1) compared to its (maximum likelihood) LR counterpart. This makes BFs in principle more suitable for automated value of evidence calculation than LRs. Logistic regression has been used to provide a calibrated LR and to fuse different evidential values into a single value of evidence [2]. In this work we discuss using logistic regression to obtain a formal BF and discuss the changes in the sampling model implied by the use of logistic regression, and study the convergence of the BF to the LR when increasing the amount of training data.

Consider two competing models for the generation of an observation, X,

$$\begin{aligned} M_1 &: X \sim F_1 \\ M_2 &: X \sim F_2 \end{aligned} \tag{1}$$

where distributions F_1 and F_2 have density functions $f_{\boldsymbol{\theta},1}$ and $f_{\boldsymbol{\theta},2}$ that are parameterized by $\boldsymbol{\theta}$. We also assume

$$\log\left(\frac{f_{\boldsymbol{\theta},1}(x)}{f_{\boldsymbol{\theta},2}(x)}\right) = \beta_0(\boldsymbol{\theta}) + \beta_1(\boldsymbol{\theta})x \equiv g(x;\boldsymbol{\theta}) \tag{2}$$

that is, the natural log of the density ratio follows a linear function.

Now suppose we have a database with observations $X_1, \ldots, X_{n_1} \sim F_1$ and $X_{n_1+1}, \ldots, X_{n_1+n_2} \sim F_2$. We can construct labels z_i, for $i = 1, \ldots, n_1 + n_2$ such that $z_i = 1$ if $X_i \sim F_1$ and $z_i = 0$ if $X_i \sim F_2$. Now denote the realization of the data set as $D^n = \{(x_i, z_i) : i = 1, \ldots, n_1 + n_2\}$. We control the number of samples under each model, and thus the rates at which we encounter observations under each model. We can define an *auxiliary probability model.* This model is not the true sampling model, but a substitute model used to make probabilistic statements about the data. The auxiliary probability model is

$$\begin{aligned} Z_i &\sim Bernoulli(\tau) \\ X_i|Z_i = 1 &\sim F_1 \\ X_i|Z_i = 0 &\sim F_2. \end{aligned} \tag{3}$$

Notice that we have fixed the priors rates to be $\frac{n_1}{n_1+n_2}$ and $\frac{n_2}{n_1+n_2}$ in the design of our experiment, but the likelihood of the auxiliary probability model depends on the parameter τ, which is the prior probability of encountering an observation under the first model. Treating τ as known we can proceed to use $\boldsymbol{\theta}$ to parameterize the distribution of X and Z under the auxiliary probability model. We have $Z_i|x_i \sim Bernoulli(\zeta(x_i;\boldsymbol{\theta}))$, where

$$\zeta(x_i;\boldsymbol{\theta}) = \mathbb{E}(Z_i|x_i) = \frac{\frac{f_{\boldsymbol{\theta},1}(x_i)\tau}{f_{\boldsymbol{\theta},2}(x_i)(1-\tau)}}{1 + \frac{f_{\boldsymbol{\theta},1}(x_i)\tau}{f_{\boldsymbol{\theta},2}(x_i)(1-\tau)}}, \tag{4}$$

using the logit link function we have

$$logit(\tau(x_i;\boldsymbol{\theta})) = \beta_0(\boldsymbol{\theta}) + \beta_1(\boldsymbol{\theta})x_i + \log\left(\frac{\tau}{1-\tau}\right). \tag{5}$$

Thus the log posterior odds follows the same functional form as the log LR, but shifted by the log prior odds. To recover $\beta_0(\boldsymbol{\theta}) + \beta_1(\boldsymbol{\theta})x_i$, we can correct the log odds using the prior rates induced by the design of the database

$$\beta_0(\boldsymbol{\theta}) + \beta_1(\boldsymbol{\theta})x_i \approx logit(\tau(x_i;\boldsymbol{\theta})) - \log\left(\frac{n_1}{n_2}\right). \tag{6}$$

Now our auxiliary likelihood is $f_{\boldsymbol{\theta}}(x,z) = f_{\boldsymbol{\theta}}(z|x)f_{\boldsymbol{\theta}}(x)$, but we do no not want to impose any distributional assumption on X other than the linear model assumption. Thus there is no clear way to express the mixture components in $f_{\boldsymbol{\theta}}(x)$, without putting stricter assumptions on the distribution of X. This mixture also contains the components needed for the BF of interest. An advantage of using logistic regression is we focus on $f_{\boldsymbol{\theta}}(z|x)$, because these two likelihoods share the same ML estimates [1], but may have different shapes. This difference in likelihoods will become important in the evaluation of the posterior distribution of $\boldsymbol{\theta}$, which is needed in the evaluation of the BF, where

$$\pi(\boldsymbol{\theta}|x,z) = \frac{f_{\boldsymbol{\theta}}(z|x)\pi(\boldsymbol{\theta}|x)}{\int f_{\boldsymbol{\theta}}(z|x)d\Pi(\boldsymbol{\theta}|x)}. \tag{7}$$

The posterior density $\pi(\boldsymbol{\theta}|x)$ requires the likelihood $f_{\boldsymbol{\theta}}(x)$, which as stated earlier is unavailable without further assumptions being placed on the distribution of X. However, we can calculate the posterior of the logistic regression coefficients as

$$\pi(\boldsymbol{\theta}|x,z) = \frac{f_{\boldsymbol{\theta}}(z|x)\pi(\boldsymbol{\theta})}{\int f_{\boldsymbol{\theta}}(z|x)d\Pi(\boldsymbol{\theta})} \tag{8}$$

utilizing only the likelihood $f_{\boldsymbol{\theta}}(z|x)$. We will discuss the implications of using this auxiliary probability model and dropping this extra term in the likelihood to obtain a formal BF.

Methods

Given a new observation x^* known to arise under one of the models with unknown label Z^*, one method to get the BF utilizes the posterior odds ratio, where the BF is the posterior odds divided by the prior odds. That is

$$BF_1 = \frac{\pi(Z^*=1|x^*,D^n)}{\pi(Z^*=0|x^*,D^n)}\bigg/\frac{\pi(Z^*=1)}{\pi(Z^*=0)} \tag{9}$$

To get the posterior odds we first need to posterior probability of group membership by using

$$\pi(Z^*=1|x^*,D^n) = \int \zeta(x^*;\boldsymbol{\theta})d\Pi(\boldsymbol{\theta}|x^*,D^n). \tag{10}$$

Now we can obtain the posterior distribution of $\boldsymbol{\theta}$ using the law of total probability

$$\begin{aligned}\pi(\boldsymbol{\theta}|x^*,D^n) &= \pi(\boldsymbol{\theta}|x^*,Z^*=1,D^n)\pi(Z^*=1|x^*,D^n)\\ &\quad+\pi(\boldsymbol{\theta}|x^*,Z^*=0,D^n)\pi(Z^*=0|x^*,D^n).\end{aligned} \tag{11}$$

To simplify the derivation of the BF consider

$$A = \pi(Z^* = 1|x^*, D^n), \tag{12a}$$

$$B = \int \tau(x^*; \boldsymbol{\theta}) d\Pi(\boldsymbol{\theta}|x^*, Z^* = 1, D^n), \tag{12b}$$

$$C = \int \tau(x^*; \boldsymbol{\theta}) d\Pi(\boldsymbol{\theta}|x^*, Z^* = 0, D^n). \tag{12c}$$

Combining (10), (11), and (12) we get $A = AB + (1 - A)C$, assuming that $\pi(Z^* = 0|x^*, D^n) = 1 - \pi(Z^* = 1|x^*, D^n)$. This assumption follows if we believe x^* could only be generated under one of the two models. Solving for A we have $A = \frac{C}{1-B+C}$, and finally $BF_1 = \frac{A}{1-A}/\frac{n_1}{n_2}$. We notice that this calculation requires the evaluation of two integrals, which are both able to be estimated using Monte Carlo integration, each requiring posterior samples of the parameters with x^* assumed to be generated under one of the models.

We also notice that we have access to the LR function

$$\lambda(\boldsymbol{\theta}) = \exp\{\beta_0(\boldsymbol{\theta}) + \beta_1(\boldsymbol{\theta})x^*\}. \tag{13}$$

Now similar to Ommen and Saunders we have another way to calculate the BF, where the BF is equal to the LR function integrated with respect to the posterior given the denominator model is true [3]. This is

$$BF_2 = \int \exp\{\beta_0(\boldsymbol{\theta}) + \beta_1(\boldsymbol{\theta})x^*\} d\Pi(\boldsymbol{\theta}|x^*, z^* = 0, D^n). \tag{14}$$

This form of the BF only requires the evaluation of one integral, which is also in the form to be estimated using Monte Carlo integration.

Simulation

To compare these two forms for the BF, a simulation study is used to compare with a ground truth model known to comply with the logistic regression assumptions. The sampling model used for the simulation study is

$$\begin{aligned} M_1 &: X \sim N(\mu_1, \sigma^2) \\ M_2 &: X \sim N(\mu_2, \sigma^2) \end{aligned} \tag{15}$$

where μ_1 and μ_2 are real numbers, and $\sigma^2 > 0$ is known. We have $\boldsymbol{\theta} = \{\mu_1, \mu_2, \sigma\}$. It is important to note that the log linear assumption is met as

$$\log\left(\frac{f_{\boldsymbol{\theta},1}(x)}{f_{\boldsymbol{\theta},2}(x)}\right) = \frac{\mu_2^2 - \mu_1^2}{2\sigma^2} + \frac{\mu_1 - \mu_2}{\sigma^2}x. \tag{16}$$

Assigning prior distributions to μ_1 and μ_2 as follows

$$\begin{aligned} \mu_1 &\sim N(\delta_1, \nu\sigma^2) \\ \mu_2 &\sim N(\delta_2, \nu\sigma^2) \end{aligned} \tag{17}$$

where δ_1 and δ_2 are real numbers and $\nu > 0$. Now with the assumption that the prior distributions of $\boldsymbol{\theta}$ does not change depending on which model generated x^*, we have the BF utilizing the true sampling model is

$$BF_{TRUE} = \frac{\pi(x^*|D^n, Z^* = 1)}{\pi(x^*|D^n, Z^* = 0)}, \tag{18}$$

which is a ratio of posterior predictive densities. We have that

$$\begin{aligned} x^*|(D^n, Z^* = 1) &\sim N(\eta_1, \gamma_1) \\ x^*|(D^n, Z^* = 0) &\sim N(\eta_2, \gamma_2) \end{aligned} \tag{19}$$

where

$$\eta_i = \left(\frac{1}{\nu\sigma^2} + \frac{n_i}{\sigma^2}\right)^{-1} \left(\frac{\delta_i}{\nu\sigma^2} + \frac{n_i \bar{x}_i}{\sigma^2}\right) \tag{20}$$

and

$$\gamma_i = \left(\frac{1}{\nu\sigma^2} + \frac{n_i}{\sigma^2}\right)^{-1} + \sigma^2. \tag{21}$$

Here $\bar{x}_i$ is the sample mean of the samples generated under model M_i for $i = 1, 2$.

The first simulation is used to study the convergence of the BF calculations to the LR empirically. This simulation is implemented according to Algorithm 1. We will use varying samples sizes fixed to be the same under both models. That is to study the behavior of the calculations as $n = n_1 = n_2$ increases.

Algorithm 1 Convergence Simulation

```
inputs: $\mu_1 = 1.5, \mu_2 = 0, \sigma = 1, \delta_1 = 1.5, \delta_2 = 0, \nu = 10$
for $x^* \in \{-1.5, 0, 0.75, 1.5, 3\}$ do
  for increasing $n$ do
    Repeat 30 times
    1. Generate $n$ samples from $N(\mu_1, \sigma^2)$ and $n$ samples from $N(\mu_2, \sigma^2)$.
    2. Draw 100,000 posterior samples of $\boldsymbol{\beta}(\boldsymbol{\theta})$ under $M_1$ and $M_2$ [4].
    3. Compute $BF_1,\ BF_2,$ and $LR$.
    4. Store $BF_1,\ BF_2,\ LR,$ and $n$.
  end for
end for
```

The second simulation, according to Algorithm 2, is used to compare the different value of evidence calculations. We use $n_1 = 500$ and $n_2 = 1000$ to compare the calculations.

Algorithm 2 Comparison Simulation

```
inputs: $\mu_1 = 1.5, \mu_2 = 0, \sigma = 1, \delta_1 = 1.5, \delta_2 = 0, \nu = 10$
for $x^* \in \{-1.5, 0, 0.75, 1.5, 3\}$ do
  Repeat 100 times
  1. Generate $n_1 = 500$ samples from $N(\mu_1, \sigma^2)$ and $n_2 = 1000$ samples from $N(\mu_2, \sigma^2)$.
  2. Draw 100,000 posterior samples of $\boldsymbol{\beta}(\boldsymbol{\theta})$ under $M_1$ and $M_2$ [4].
  3. Compute $BF_1$, $BF_2$, $BF_{TRUE}$, $LR$, and the plug-in estimate of the $LR$ and store.
end for
```

Results

The results of the simulations are displayed in Figure 1. The left panels shows the results of the simulation according to Algorithm 1. From top to bottom the value of x^* takes on the values

$-1.5, 0, 0.75, 1.5,$ and, 3. The dashed blue line is the true value of the LR utilizing the true parameter values. The quantiles of the BF_1 replicates given the sample size were estimated using a quantile regression with $\log_{10}(BF)$ modeled using a cubic B-spline basis expansion on sample size with 5 degrees of freedom. The sample size, n, is displayed on a log scale for better illustration. The shaded region is the interquartile range (IQR) with a median curve in the middle. The maximum and minimum lines are $1.5 * IQR$ added and subtracted from the third and first quartiles respectively, and any points lying outside of the maximum and minimum lines are plotted as red points. We do not display the results using BF_2 as the results are almost identical.

We see that as the number of background objects increases the distribution of the BF collapses around the true value of the LR. We see that at the point where $LR(x^*) = 1$ or analogously $\log_{10}(LR(x^*)) = 0$, at $x^* = 0.75$, the distribution of BF values is narrower in the small sample size range. We also see that at this point the median of the BF values is close to the true value of the LR even at small samples sizes. As we move x^* away from this point the BF values are more conservative, meaning the median of the BF_1 replicates is closer to 1 than the true value of the LR.

The right panels show the results of the simulation according to Algorithm 2. The value of x^* again changes from the top to bottom in the same fashion as in the left panels. Again, the blue line shows the true value of the LR. The box plots moving left to right show the distributions of BF_{TRUE}, BF_1, BF_2, and the plug-in estimate of the LR using logistic regression.

We see that the distribution of BF_1 and BF_2 are wider than the distribution of BF_{TRUE} likely due to the term being dropped from the likelihood of the auxiliary probability model. We see that BF_1 and BF_2 behave similarly and are centered on the true value of the LR. We again see that when $x^* = 0.75$ the median of the BF_1 and BF_2 replicates are around the true value of the LR. As we move x^* away from this point the median values are more conservative.

Discussion

The two methods to obtain a BF using logistic regression behave similarly. In both cases there is more variation than the BF that utilizes the true generative model, but the distribution of BF replicates are centered near the true value of the LR. Empirically these BF values converges to the LR, which is a desired property. This is promising, as we can utilize methods such as Bayesian neural networks or other methods that provide posterior probabilities of group membership to estimate functional forms of the log density ratio to obtain a formal BF.

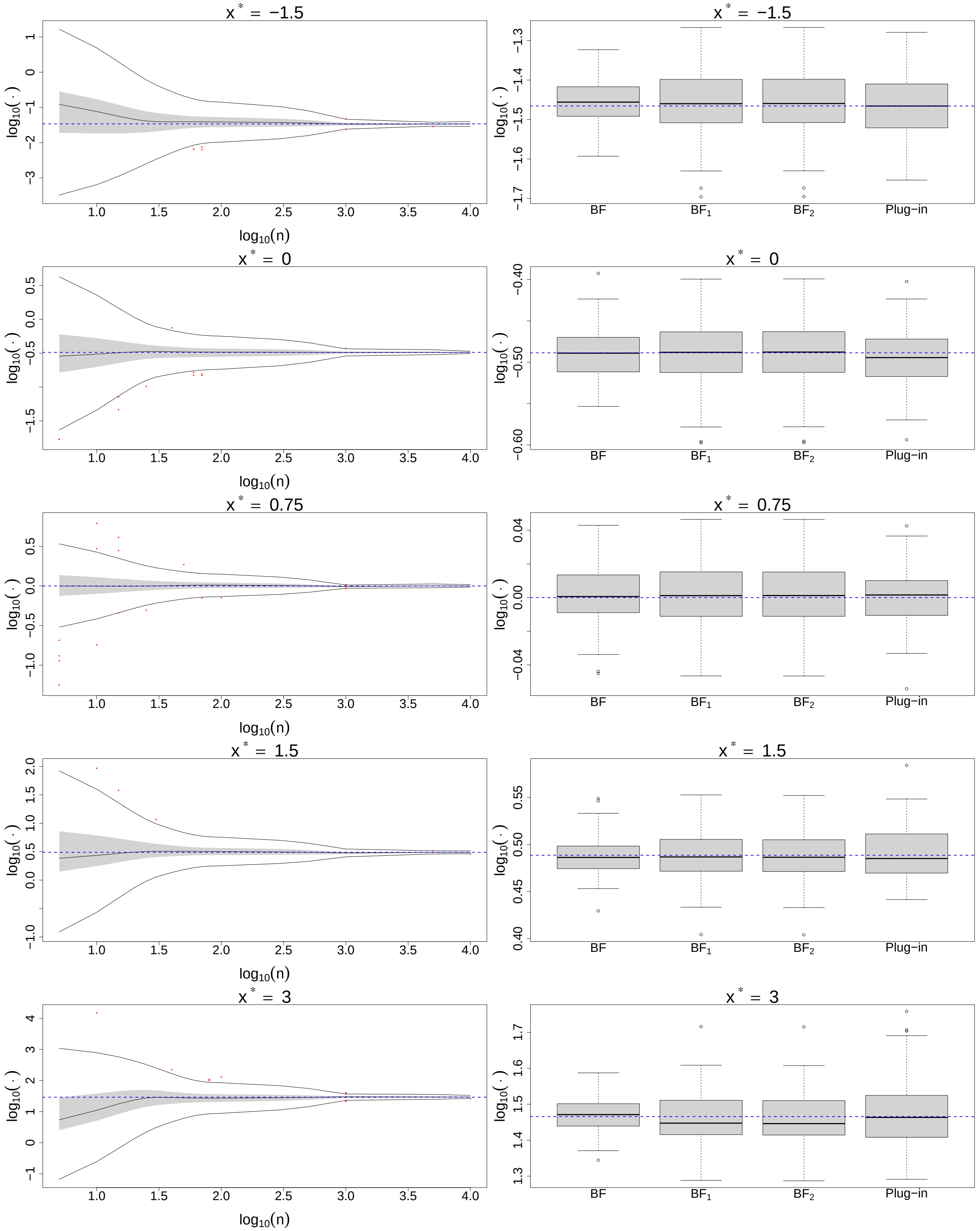

Figure 1: The left column shows the distribution of BF_1 as $n = n_1 = n_2$ increases. The right column shows the distribution of the different values of evidence when $n_1 = 500$ and $n_2 = 1000$. The blue dashed line is the true value of the LR.

References

[1] Geoffrey J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley & Sons, Hoboken, New Jersey, 2004.

[2] Geoffrey Stewart Morrison. Tutorial on logistic-regression calibration and fusion:converting a score to a likelihood ratio. *Australian Journal of Forensic Sciences*, 45(2):173–197, 2013.

[3] Danica M. Ommen and Christopher P. Saunders. A Problem in Forensic Science Highlighting the Differences between the Bayes Factor and Likelihood Ratio. *Statistical Science*, 36(3):344 – 359, 2021.

[4] Stan Development Team. RStan: the R interface to Stan, 2020. R package version 2.21.2.