Forensic Science International: Genetics Supplement Series 7 (2019) 572–574

Contents lists available at ScienceDirect

Forensic Science International: Genetics Supplement Series

journal homepage: www.elsevier.com/locate/fsigss

Are reported likelihood ratios well calibrated?

Jan Hannig[a,c,*], Sarah Riman[b], Hari Iyer[a], Peter M. Vallone[b]

[a] Statistical Design, Analysis, and Modeling Group, ITL/NIST, United States
[b] Applied Genetics Group, National Institute of Standards and Technology, United States
[c] Department of Statistics and Operations Research, UNC-Chapel Hill, United States

ARTICLE INFO

Keywords:
Likelihood ratio
Calibration
Probabilistic genotyping
Generalized fiducial inference

ABSTRACT

In this work we introduce a new statistical methodology for empirically examining the validity of model-based Likelihood Ratio (LR) systems by applying a general statistical inference approach called generalized fiducial inference.

LR systems are gaining widespread acceptance in many forensic disciplines, especially in the interpretation of DNA evidence, in the form of probabilistic genotyping systems (PGS). These systems output a Bayes factor, commonly referred to as a likelihood ratio in forensic science applications. Methods for examining the validity of such systems is a topic of ongoing interest. In addition to summarizing existing approaches and developing our new approach, we illustrate the methods using the PROVEDIt dataset by examining LR values calculated with two PG software packages.

1. Background

The Likelihood (LR) framework is commonly used for quantifying the value of evidence in forensic DNA analysis. Advances in Probabilistic Genotyping have resulted in our ability to interpret complex DNA mixtures. Many probabilistic genotyping software systems, some open-source and some proprietary, are currently available. As a consequence, the same electropherogram information could result in different LR assessments depending on the software system used. A common attitude towards this multiplicity of LR values is that each LR assessment is valid as long as the assumptions underlying the parent models are reasonable approximations of reality. However, it is widely accepted that continuous models have greater power to discriminate between the so-called prosecution hypothesis,

H_P. The person of interest is a contributor to the mixture, and the defense hypothesis.

H_D. The person of interest is NOT a contributor to the mixture.

Under the circumstances, one can ask the following questions: (a) Are LR values from different systems close enough to each other to not have an impact in casework? (b) In situations where the differences are large enough to be impactful which LR should one use?

We propose to empirically assess an LR system, when an adequate number of ground-truth known samples is available, using two key metrics: (a) The power to discriminate between H_P and H_D, and (b) the degree to which the LR system is well-calibrated.

Discrimination power can be empirically examined using receiver-operating characteristic (ROC) plots and associated summary statistics such as the area under the ROC curve (AUC). The property of being well-calibrated is a bit more involved to assess and is the topic of this work. Among all available LR systems we need to identify those that are well-calibrated and have high discriminating power.

A model making probability assessments (such as probability of rain on a given day) is said to be well-calibrated if it actually rains on 100p% of the days that the model predicts p as the probability for rain. In contrast, a model making LR assessments is said to be well-calibrated if the value LR = r occurs r times as often under H_P as it does than under H_D. This property is captured in the well-known statement that LR of LR is LR (see, [1], Eq. 1.32, page 26, Section 1.8).

Many previous approaches for assessing LR calibration are based on assessing the calibration of posterior probabilities in controlled, ground-truth known conditions with varying prior probabilities using familiar methods for assessing calibration of probabilities. This is based on the property that LR is the multiplier that converts a prior probability into a posterior probability. See [2,3] in the references. Other methods that have been proposed to assess overall performance of LR systems involve the use of a single metric, e.g., empirical cross-entropy (see [4,5]). Our approach is based on a direct assessment of how well the property that LR of LR is LR is satisfied and does not require consideration of prior or posterior probabilities.

2. Our approach

Estimation of the LR of LR is difficult. We use a novel approach

* Corresponding author at: Department of Statistics and Operations Research, UNC-Chapel Hill, United States.
E-mail address: jan.hannig@unc.edu (J. Hannig).

https://doi.org/10.1016/j.fsigss.2019.10.094
Received 20 September 2019; Received in revised form 2 October 2019; Accepted 8 October 2019
Available online 09 October 2019
1875-1768/ Published by Elsevier B.V.

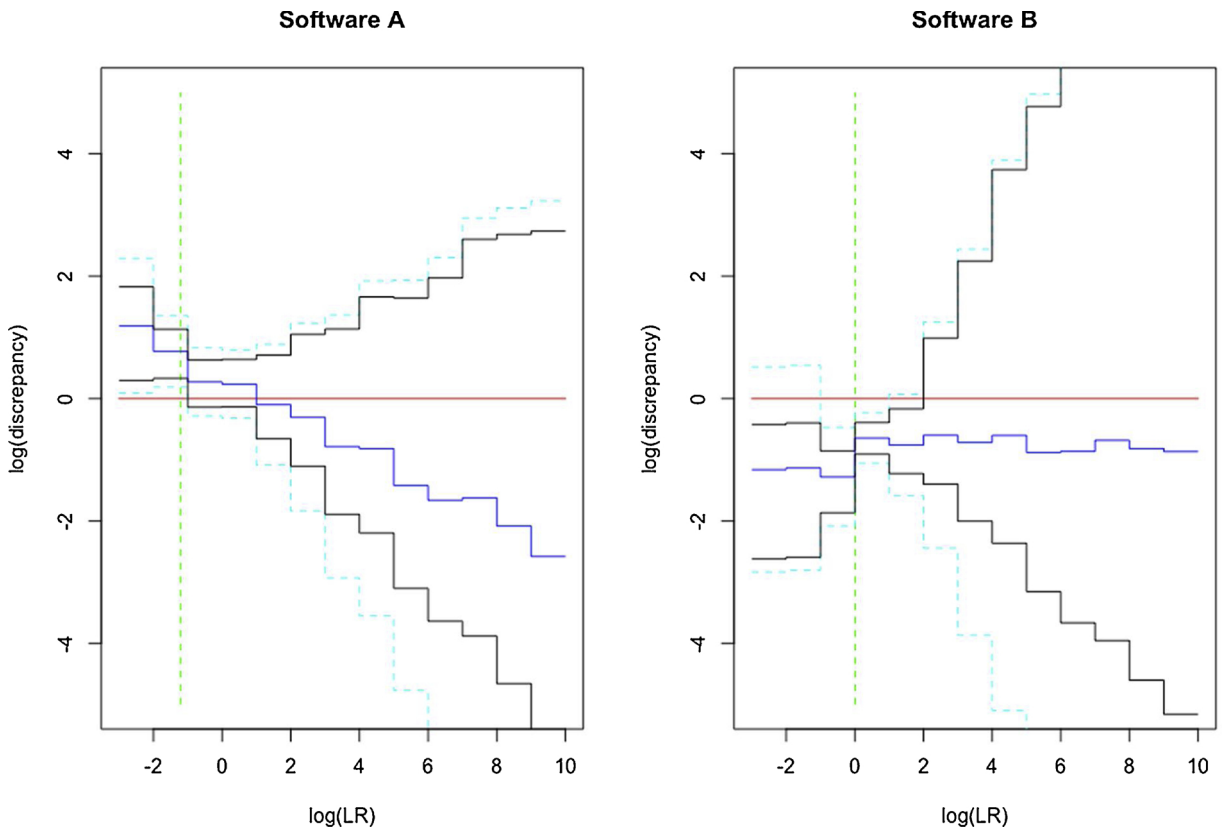

Fig. 1. (Left) Calibration plot for software system A. (Right) Calibration plot for software system B. The blue trace represents the Fiducial median estimate for the Logarithm of the calibration discrepancy. The black and cyan traces on either side of the blue trace are, respectively, the boundaries of a 95% pointwise and simultaneous fiducial confidence interval for the log discrepancy. The red horizontal line shows the log discrepancy curve for a perfectly calibrated LR system. The green vertical line marks the spot where GPD extrapolation begins. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

based on the integrated version of the equation LR of LR = LR. To account for sampling uncertainty in the empirical data used to make this assessment we use the 'Fiducial Approach' proposed by R. A. Fisher [6]. The method of Fiducial Inference was unpopular during the 20th century but has been re-discovered and highly developed during the past 20 years. See [7,8].

3. Key mathematical details

*Let $g(r)$ denote the probability density function (pdf) and $G(r)$ denote the corresponding cumulative distribution function (cdf) of the likelihood ratio LR under H_P. Likewise, let $f(r)$ denote the pdf and $F(r)$ the cdf of LR under H_D. We need to check whether or not $g(r) = r f(r),\ r \geq 0$. By integrating this equation over the interval (a, b) we observe that the following equation must hold

$$\log(G(b) - G(a)) - \log(bF(b) - aF(a) - \int_a^b F(r)dr) = 0,\ 0 < a < b < \infty. \quad (1)$$

We estimate $G(r)$ and $F(r)$ with fiducial distributions obtained from the ground-truth known empirical data. That is, we have a collection of LR values known to have come from H_P true cases and another collection of LR values known to have come from H_D true cases. We check the condition in Eq. (1) for a sequence of intervals (a, b).

The fiducial distributions of $G(r)$ and $F(r)$ allow us to form confidence intervals for the left-hand side of (1). The confidence interval can provide statistical evidence for or against calibration, e.g., if it does not include 0 everywhere in the interval (a, b), we conclude lack of calibration in that interval.

4. Illustrative example

We considered ground-truth known 2-Person and 3-Person mixtures and obtained LR values using two different Probgen systems (continuous models), say Software A and Software B. We pooled both sets (2 P and 3 P) for this illustration. *Since ground truth is known, we can examine LR values from H_P True cases and H_D True cases.* The quantity of our interest is the factor by which the stated LR value differs from what is supportable by empirical data. Because of lack of data for the 'right tail' of the H_D-true LR distribution, extrapolation based on a Generalized Pareto Distribution (GPD) is used. Fiducial confidence interval estimates are computed for logarithm of the LR Calibration Discrepancy shown on the left-hand side of (1). Results are shown in the following plots where the axes use a logarithmic (base 10) scale. The Fiducial median is used as the point estimator for the discrepancy and is shown in blue. 95% fiducial confidence interval (band) is shown in black. The red line (horizontal line at 0) corresponds to perfect calibration. A negative Log discrepancy means LR value overstated the weight of evidence in favor of H_P whereas a positive Log discrepancy means LR value overstated the weight of evidence in favor of H_D. Clearly calibration (or lack of it) cannot be demonstrated for large LR values due to lack of adequate information (Fig. 1).

In the illustrative examples above, the calibration plot suggests the following: Software A has a downward slope suggesting that as the reported LR values get increasingly larger than 1 they tend to increasingly overstate the weight of evidence in favor of H_P. In the case of Software B, the calibration plot suggests that it may be overstating the weight of evidence in favor of H_P by little less than a factor of 10.

5. Summary

In this presentation we have proposed an approach for directly assessing

calibration of LR systems using the fact that a well calibrated LR system should possess the property that LR of LR is LR. We illustrated the application of the method in some real examples using two of the available continuous LR systems. We discussed aggregate measures of performance only. Further examination of performance in subpopulations is of interest.

Acknowledgement

Jan Hannig's research was supported in part by the National Science Foundation under Grant No. DMS-1512945 and IIS-1633074 and DMS-1916115.

References

[1] D.M. Green, J.A. Swets, Signal Detection Theory and Psychophysics, John Wiley & Sons, 1966.
[2] G. Brier, Verification of forecasts expressed in terms of probability, Monthly Weather Rev. 78 (1950) 1–3.
[3] M.H. DeGroot, S.E. Fienberg, The comparison and evaluation of forecasters, Statistician. 32 (1983) 12–22.
[4] D. Van Leeuwen, N. Brümmer, An introduction to application-independent evaluation of speaker recognition systems, in: C. Müller (Ed.), Speaker Classification, Lecture Notes in Computer Science/Artificial Intelligence, vol. 4343, Springer, Heidelberg/Berlin, Germany; New York, NY, USA, 2007, pp. 330–353.
[5] D. Ramos, J. Franco-Pedroso, A. Lozano-Diez, J. Gonzalez-Rodriguez, Deconstructing cross-entropy for probabilistic binary classifiers, Entropy 20 (208) (2018) 1–20 Issue 3.
[6] R.A. Fisher, The fiducial argument in statistical inference, Ann. Eugen. 6 (1935) 391–398.
[7] J. Hannig, H. Iyer, R.C.S. Lai, T.C.M. Lee, Generalized fiducial inference: a review and new results, J. Am. Stat. Assoc. 111 (2016) 1346–1361.
[8] Y. Cui, J. Hannig, Estimation and testing of survival functions via generalized fiducial inference with censored data, with discussion and rejoinder by the authors, Biometrika 106 (2019) 501–518.