Reducing the degree of uncertainty within automatic speaker recognition systems using a Bayesian calibration model

Bruce Xiao Wang & Vincent Hughes
bruce.wang@alumni.york.ac.uk/ vincent.hughes@york..uk

1. Introduction

Forensic validation

-Validity: a measure of discrimination, i.e., how good or bad the system is at separating two samples
- Repeatability: intra-examiner reliability
- Reproducibility: inter-examiner reliability

Most forensic validation implicitly focuses on
- Overall performance of methods under casework conditions
- Discriminability (see Smith & Neal 2021) with different methods chosen, or decisions made
- Based low values for the validity metric used

However, the expert's primary concern should be to reduce uncertainty (Morrison & Enzinger 2016, Ramos et al 2021).

Source of uncertainty

- Subjective decision: varies from features used for analysis to statistical models, reference population used
- Variation in the disputed sample
- Variation in the known sample
- Sampling variability

Problems:

- Each case is unique in terms of case materials/conditions,

Questions:

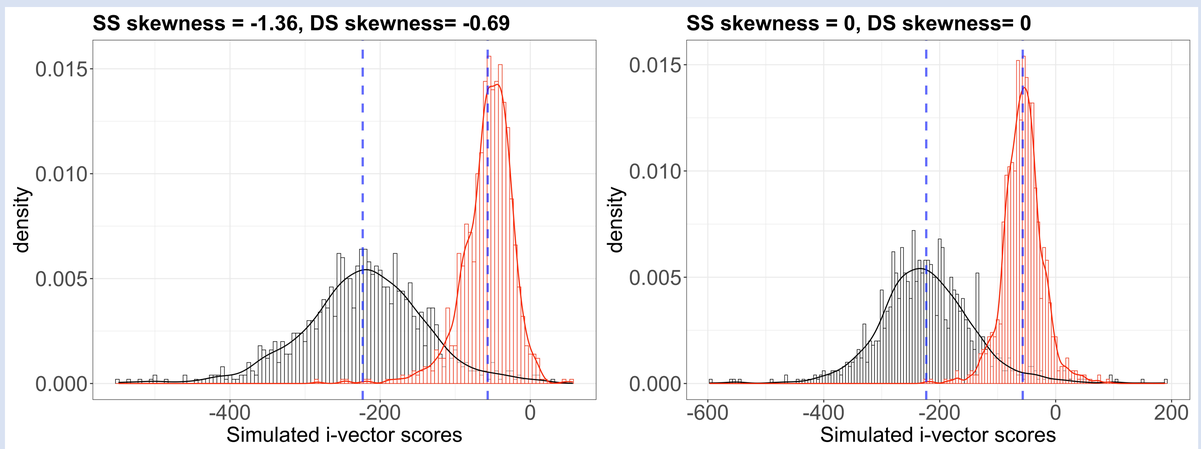
- Is reproducibility possible?
- What happens to the reliability caused by uncertainty on system performance?

2. Aim

- Use simulated score to demonstrate the variability in validation results as a function of sampling variability
- Demonstrate how Bayesian calibration (Brümmer & Swart 2014) may reduce such uncertainty.

3. Method

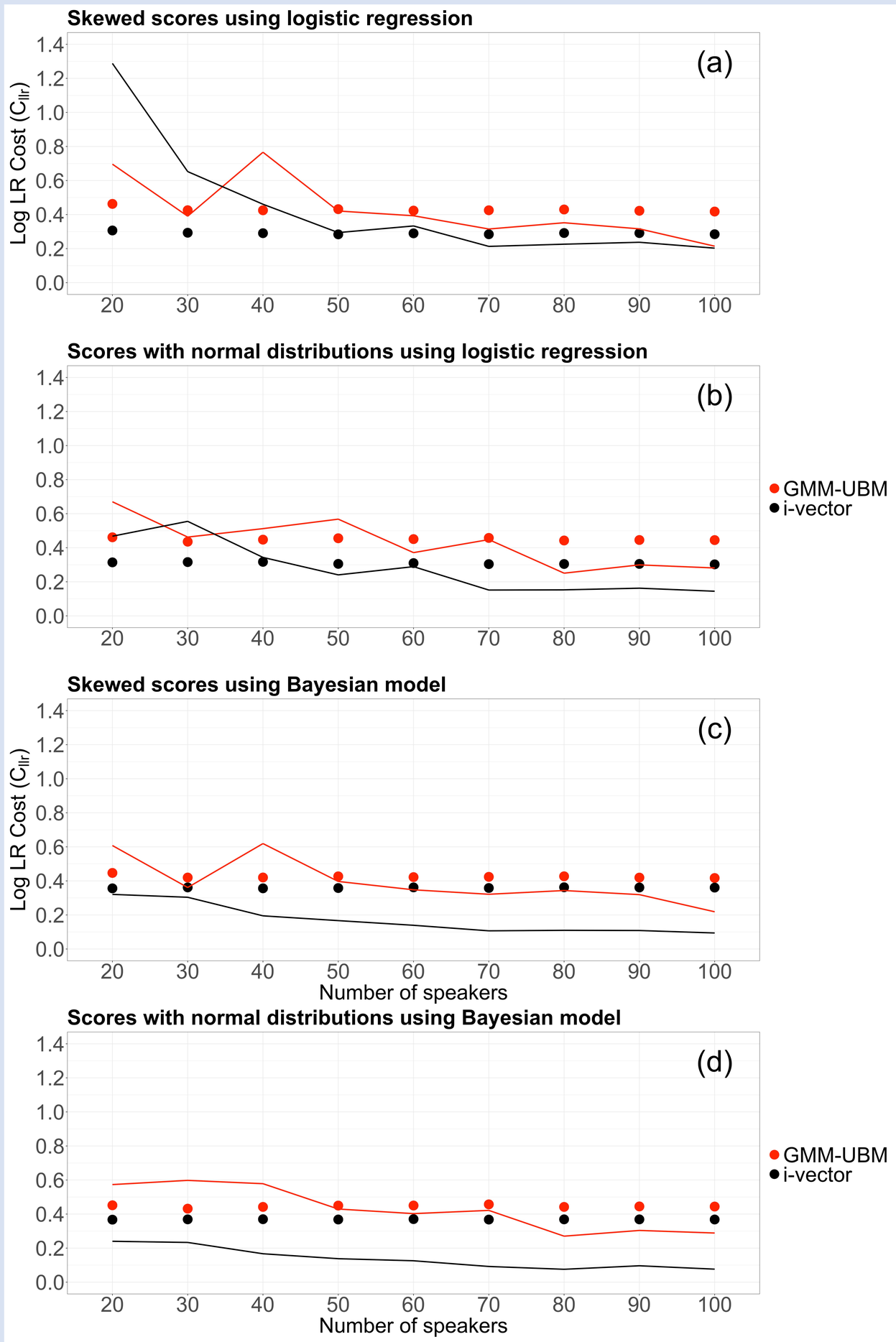
- Scores were simulated based on score distribution parameters obtained from (Enzinger & Morrison, 2016; Morrison & Pho, 2018).
- The skewness for both SS and DS scores were changed to 0 (i.e., normal distribution) while the kurtosis, mean and standard deviation were kept fixed.
- Calibration (i.e., logistic regression and Bayesian model) conducted 100 times for scores with normal and skewed distribution with different sample size respectively



Examples of simulated i-vector (top panel) and GMM-UBM (bottom panels) scores, sample size = 1000 in each of the SS and DS scores. Blue dashed lines indicate the mean.

4. Results

Validation in experiment



C_{llr} mean (dots) and range (lines) as a function of score skewness, sample size and calibration methods.

5. Take home message

- Recognise that forensic comparison is a process involving numerous decisions which introduce uncertainty via both systematic and random factors

- Be explicit about the decisions made at each stage of the process and the implications of such decisions for uncertainty in terms of the results LRs **and** overall method validity

- In the forensic context, it is not the case that the modern 'state-of-art' system which is capable of the best validity is necessarily the optimal choice.

- Rather, this choice is dependent on sample size, score skewness, and the choice of calibration method (likely amongst of considerations).

6. References

Akmeemana, A., Weis, P., Corzo, R., Ramos, D., Zoon, P., Trejos, T., ... & Almirall, J. (2021). Interpretation of chemical data from glass analysis for forensic purposes. *Journal of Chemometrics*, *35*(1), e3267.
Brümmer, N. and Swart, A. (2014) Bayesian calibration for forensic evidence reporting. *Proceedings of Interspeech*, Singapore, 14-18 September. pp. 388-392.
Morrison, G. S. and Enzinger, E. (2016) What should a forensic practitioner's likelihood ratio be? *Science and Justice* (Virtual Special Issue on measuring and reporting the precision of forensic likelihood ratios), 56: 374-379.
Morrison, G. S. and Poh, N. (2018) Avoiding overstating the strength of forensic evidence: Shrunk likelihood ratios/ Bayes factors. *Science and Justice*, 58: 200-218.
Smith, A. M. and Neal, T. M. S. (2021) The distinction between discriminability and reliability in forensic science. *Science and Justice*, 61: 319-331.

Read more by scanning the QR code