Accuracy Assessment Methods for Likelihood-Ratio-based Evidence Evaluation

Daniel Ramos and Joaquin Gonzalez-Rodriguez
ATVS – Biometric Recognition Group
Universidad Autónoma de Madrid
http://atvs.ii.uam.es

Jose-Juan Lucena-Molina
Statistics Department, Criminalistics Service
Dirección General de la Policía y de la Guardia Civil
Ministerio del Interior, Spain

Outline

- Assessment of evidence evaluation methods
  - Motivation
  - Likelihood ratios for evidence evaluation and interpretation
- Assessment of likelihood-ratio-based evidence evaluation
  - Empirical approach
  - Assessment methods
    - Rates of misleading evidence
    - Tippett plots
    - Empirical Cross-Entropy
    - Limit Tippett plots (novel assessment methodology)
- Experiment with speaker recognition systems
- Conclusions

Assessment of Evidence Evaluation Methods

Assessment of Performance: Motivation

- Increasing interest for the scientific assessment of any processes involved in forensic science
  - The effect of Daubert rules
  - Two recent and important references found in the literature

Committee on Identifying the Needs of the Forensic Sciences Community, "Strengthening Forensic Science in the United States: A Path Forward, National Research Council, National Academy of Sciences, 2009.

The Law Comission, The admissibility of Expert Evidence in Criminal Proceedings in England and Wales. A New Approach to the Determination of Evidentiary Reliability. Consultation paper no. 190, 2009.

- Assessment of evidence evaluation methods is a key point towards this aim

Evidence Evaluation with Likelihood Ratios

Forensic Laboratory

Forensic scientist

$$LR = \frac{p\left(E \middle| \theta_p, I\right)}{p\left(E \middle| \theta_d, I\right)}$$

LR=1000

Province of the court

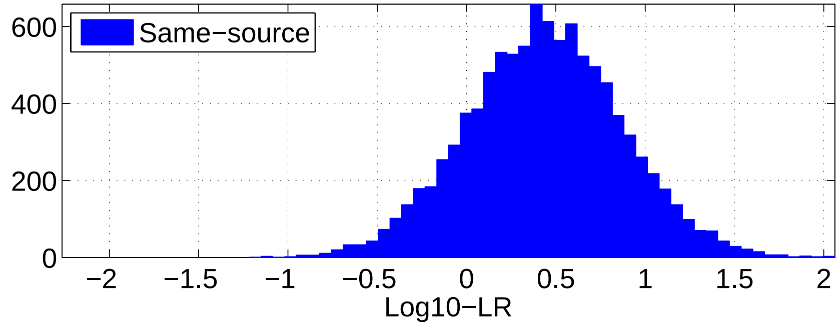
$$\frac{P\left(\theta_p \middle| I\right)}{P\left(\theta_d \middle| I\right)} = \frac{1}{99}$$

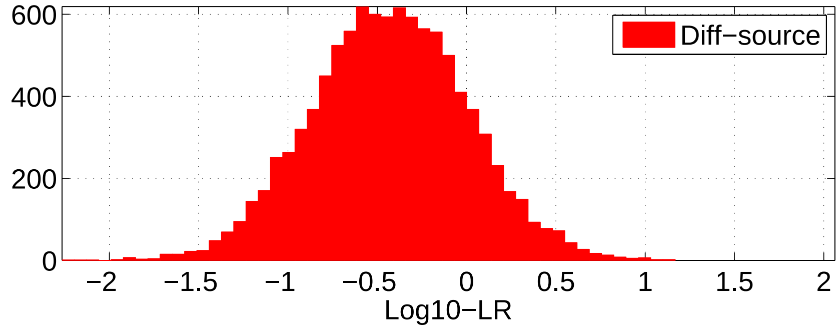
$$\frac{P\left(\theta_p \middle| E, I\right)}{P\left(\theta_d \middle| E, I\right)} = \frac{1000}{99}$$

Assessing Performance of Likelihood-Ratio-Based Evidence Evaluation Methods

Empirically Measuring Performance

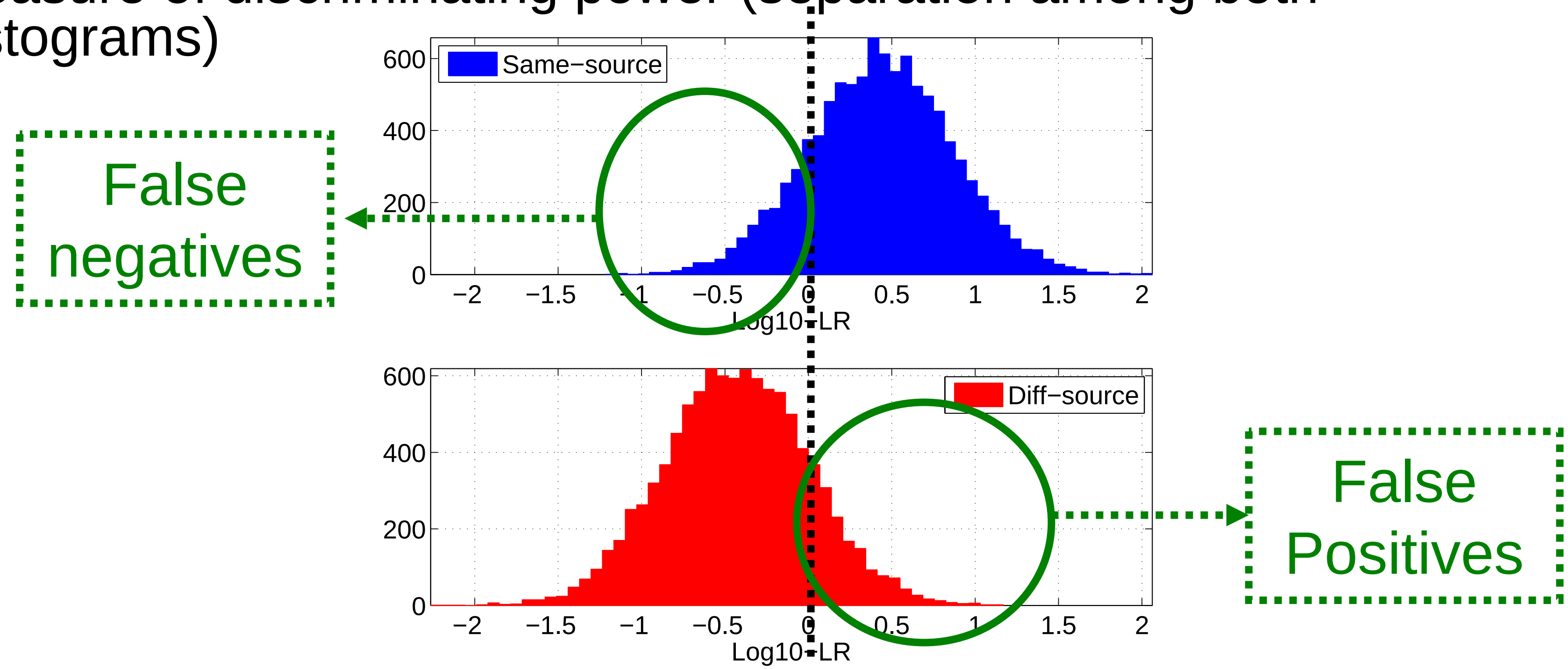
- Experimental test
  - Database of materials with known sources
  - *E.g.*, speech utterances of known origin
- Generate comparisons (LR values) where:
  - Both materials to compare come from the same source
    - θ_p is true
  - Materials to compare come from different sources
    - θ_d is true





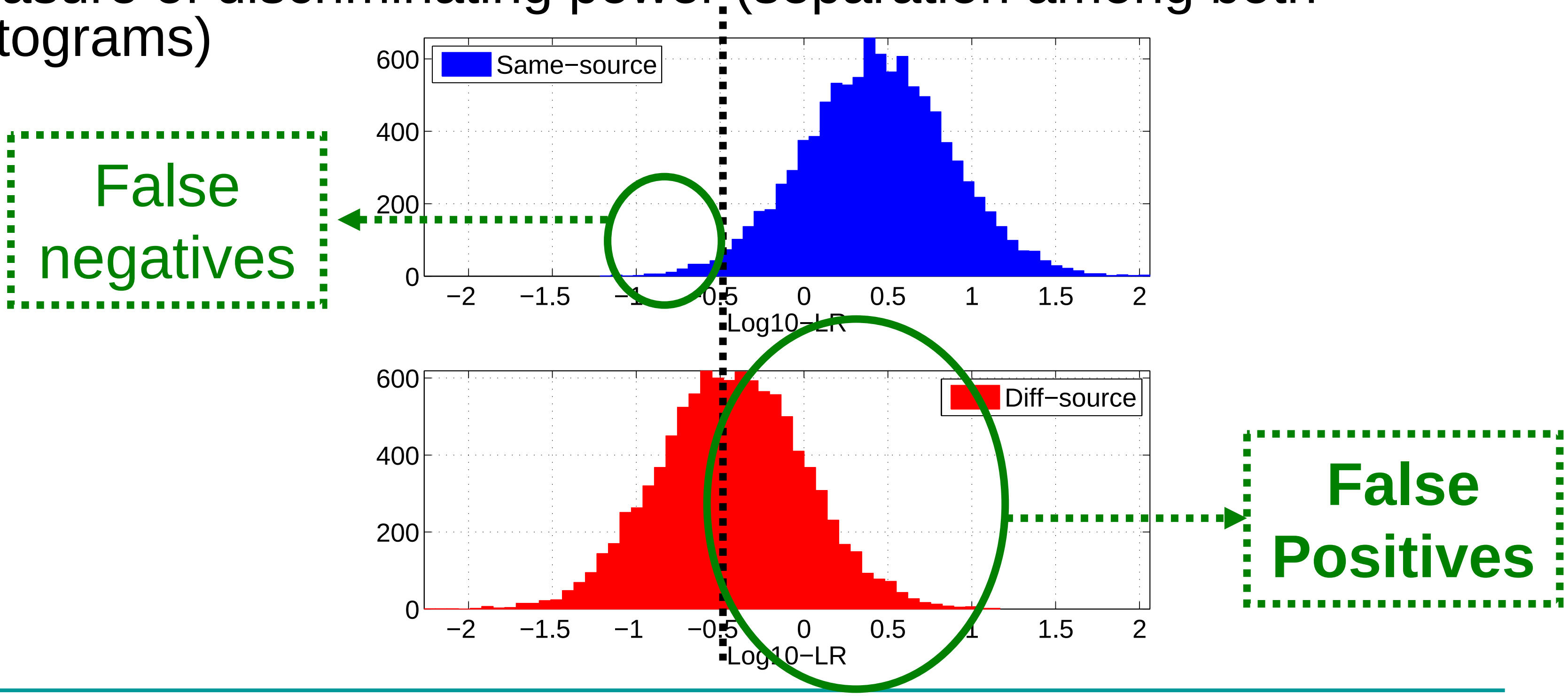
False Positive and False Negative Rates

- Classical measure of performance
- For a given decision threshold, percentage of false positive and false negative cases
  - They depend on the decision threshold (typically LR=1)
- Measure of discriminating power (separation among both histograms)



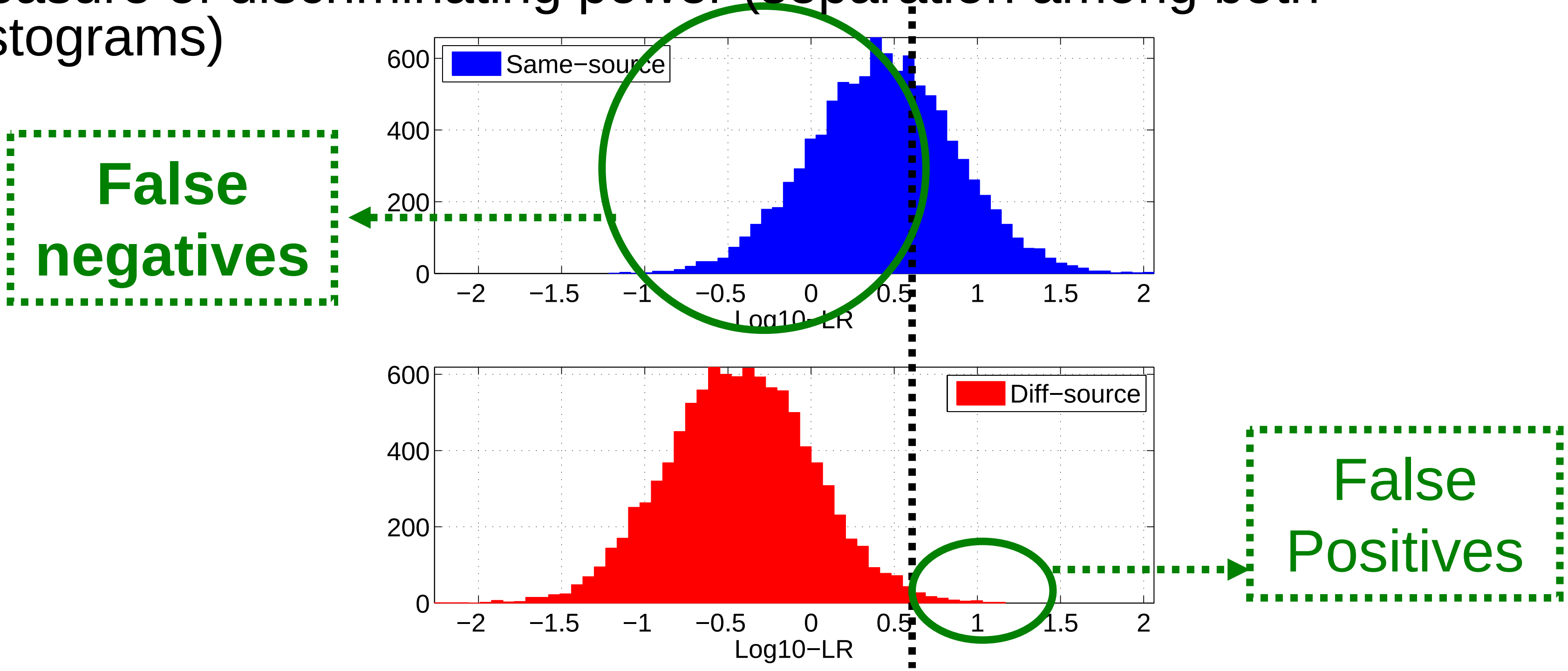
False Positive and False Negative Rates

- Classical measure of performance
- For a given decision threshold, percentage of false positive and false negative cases
  - They depend on the decision threshold (typically LR=1)
- Measure of discriminating power (separation among both histograms)



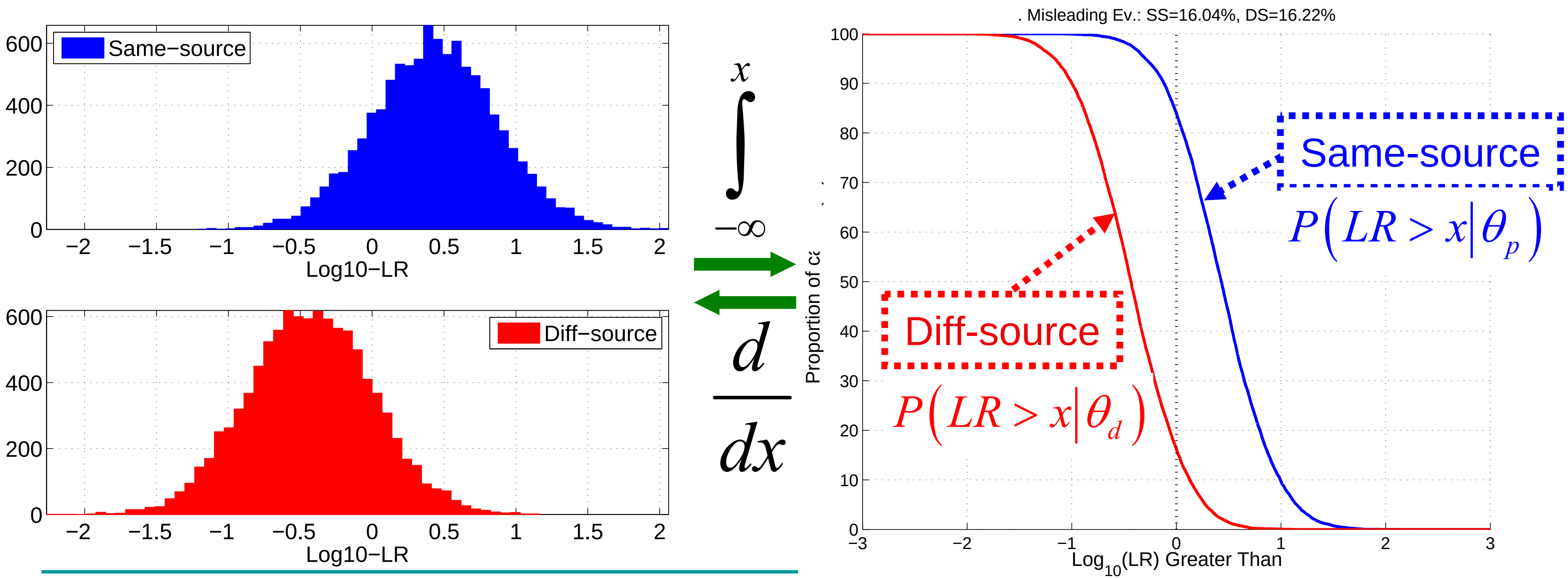
False Positive and False Negative Rates

- Classical measure of performance
- For a given decision threshold, percentage of false positive and false negative cases
  - They depend on the decision threshold (typically LR=1)
- Measure of discriminating power (separation among both histograms)



Tippett Plots

- Cumulative histograms of same-source and different-source LR values
  - Interpretable as probabilities
    - Probability of finding LR values greater than… (value in the x-axis)
- Equivalent to plot false positive and (the complementary of) false negative rates for any threshold in the x-axis



Performance in Tippett Plots: ROME

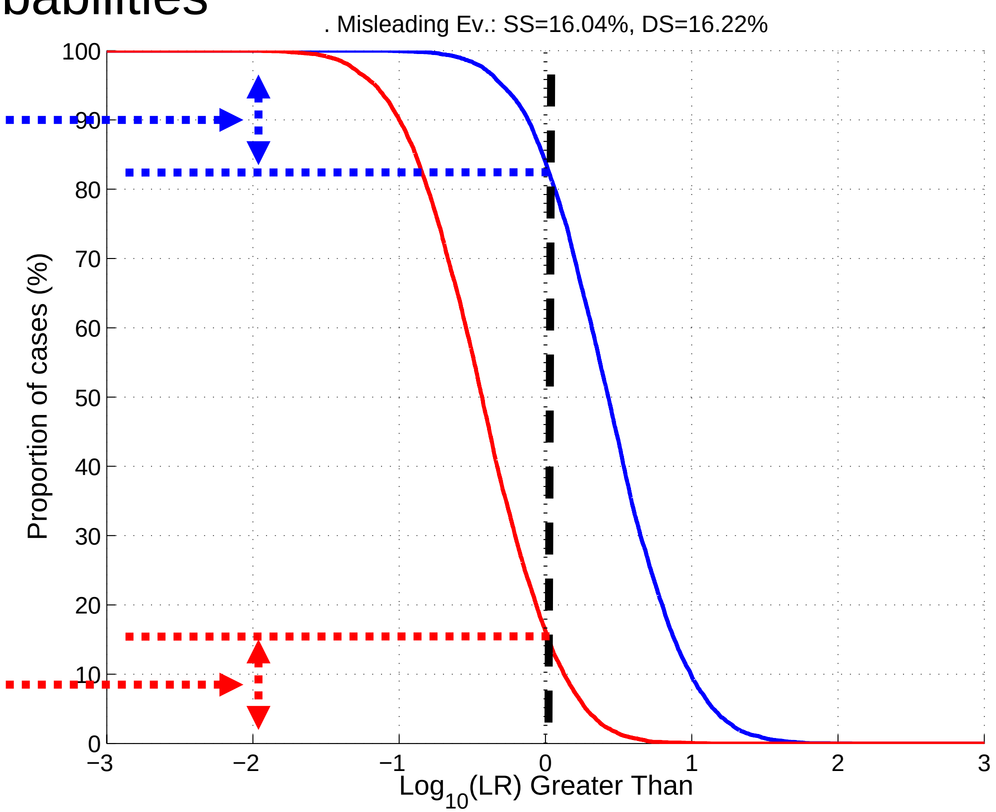
- Rates of Misleading Evidence (ROME)
  - "Proportion of LR values that will give support to the wrong hypothesis"
    - Can be interpreted as probabilities

Same-source ROME

$$P\left(LR<1\middle|\theta_p\right)=15\%$$

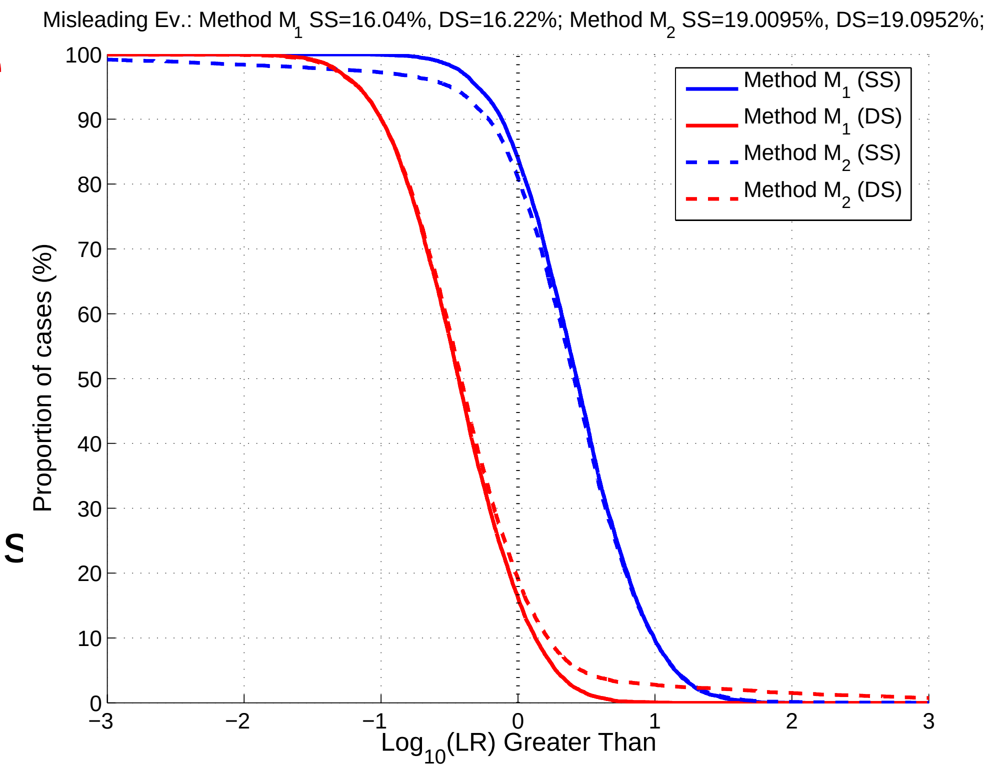
$$P\left(LR>1\middle|\theta_d\right)=14\%$$

Diff-source ROME



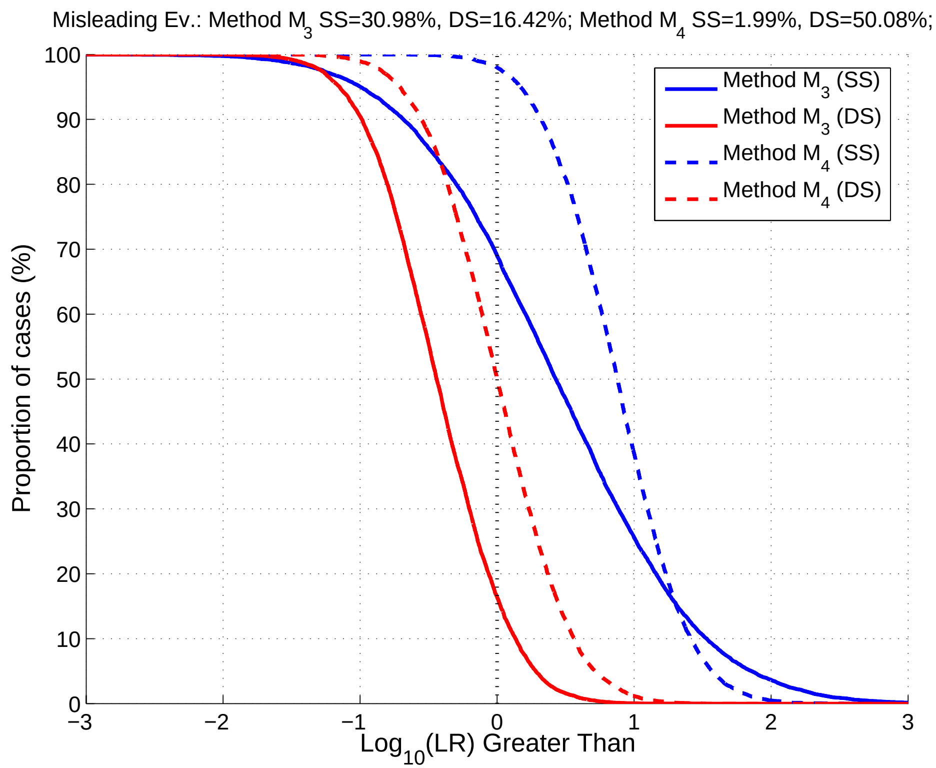
Problems with Tippett Plots

- ROME is far from enough to determine performance by itself
  - Strong misleading evidence is also very important
- Methods M_1 and M_2 have very similar ROME
  - But M_2 presents much higher strong misleading evidence
  - M_1 should be much better
    - ROME do not highlight this
    - And Tippett plots do not numerically measure that



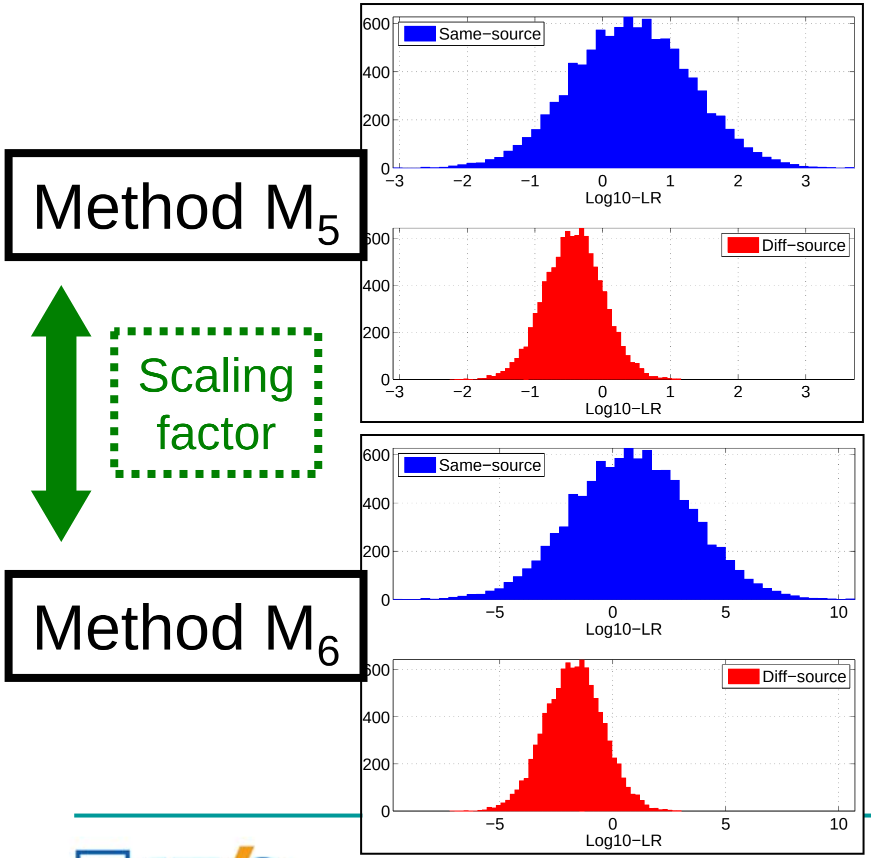
Problems with Tippett Plots

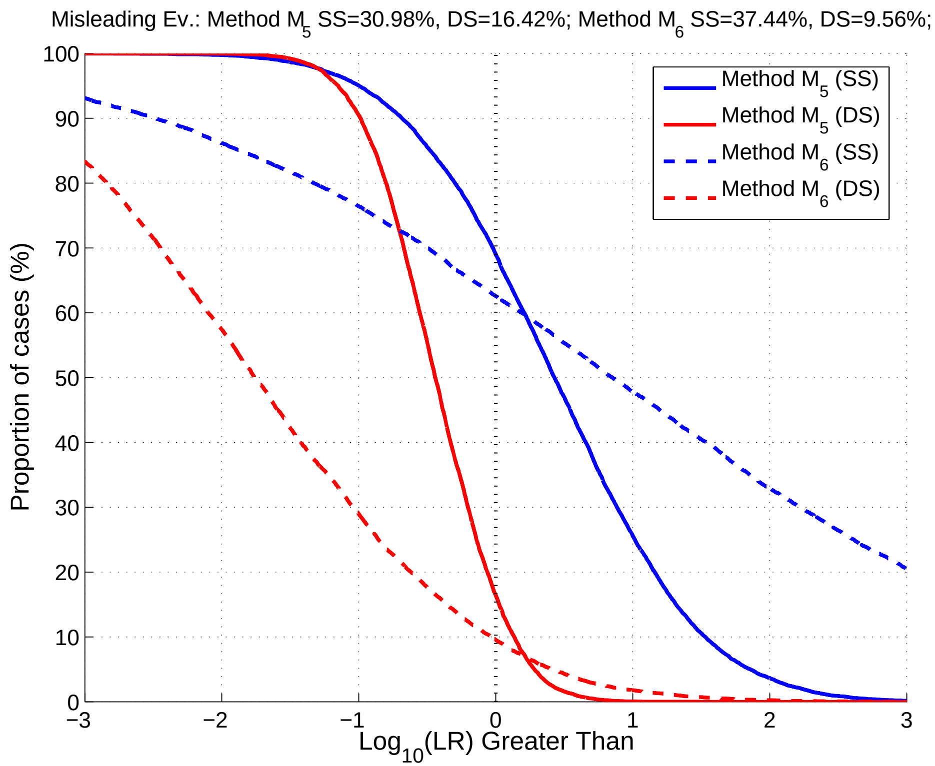
- Performance with Tippett plots is not numerically measured
  - Sometimes their interpretation is subjective
  - And sometimes it is difficult to identify the best method
- Which method is better among M_3 and M_4?
  - ROME is not conclusive
    - M_3: same-source is worse
    - M_4: different-source is worse
  - Strong misleading evidence is not conclusive
    - M_3: same-source is stronger
    - M_4: different-source is stronger



Problems with Tippett Plots

- Discriminating power is not easily comparable
  - Methods M_5 and M_6 have the same discriminating power
    - They have very different Tippett plots





Empirical Cross-Entropy (ECE)

- Objective measure of performance: numerical value
  - The higher its value, the worse the evidence evaluation method
  - Allows easy comparison of methods
- Discriminating power is clearly stated
- Takes into account strong misleading evidence
- Based on the logarithmic scoring rule
- Information-theoretical interpretation
  - Intuitive and understandable

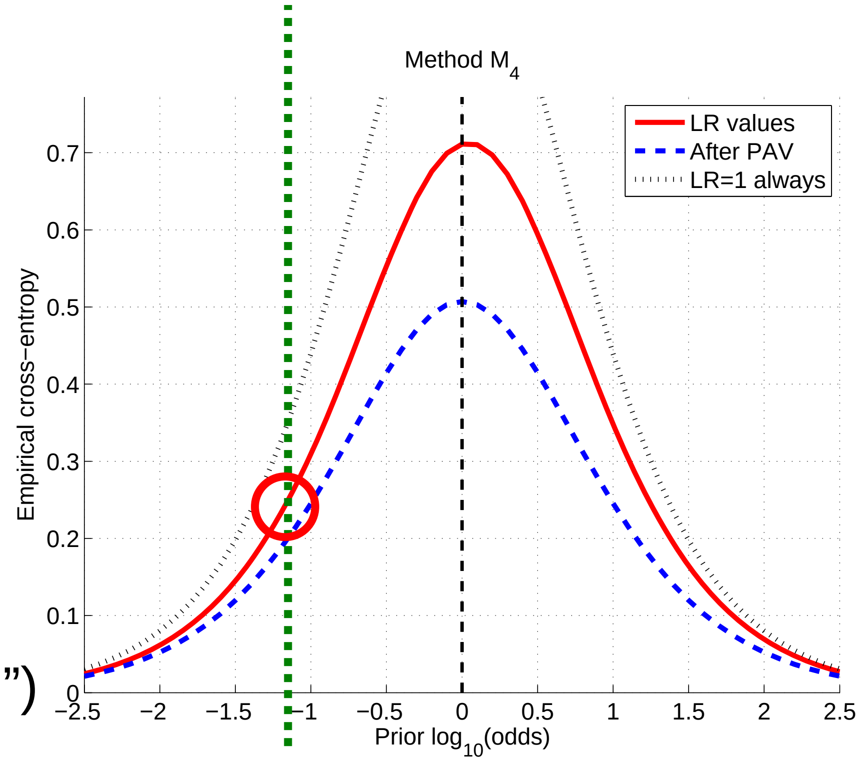
D. Ramos, J. Gonzalez-Rodriguez, G. Zadora, J. Zieba-Palus and C. G. G. Aitken (2007). "Information-theoretical comparison of likelihood ratio methods of forensicevidence evaluation". Proceedings of International Workshop on Computational Forensics (in IAS 2007), pp. 411-416.

D. Ramos (2007). "Forensic Evidence Evaluation Using Automatic Speaker Recognition Systems". Ph.D. Thesis, Dept. of Computer Science, Univ. Autonoma de Madrid.

ECE Plots: LR Performance

$$\frac{P(\theta_p|I)}{P(\theta_d|I)} = \frac{1}{10}$$

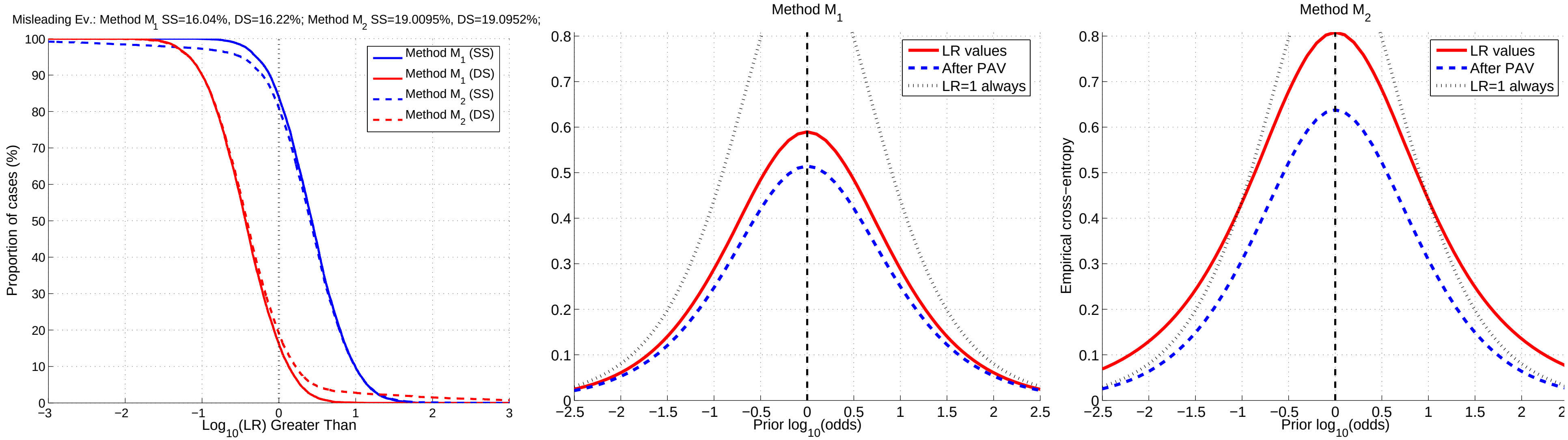
- 3 curves are represented
  - ECE (solid): overall performance
    - The higher its value, the worse the method
  - Calibrated (dashed): discriminating power
    - Difference among ECE & Calibrated is the calibration performance
  - Always LR=1 (dotted)
    - A method that does not take into account the evidence (“does nothing”)



- Separation of roles
  - Forensic scientist: ECE computation for a wide range of priors
  - Fact finder: prior establishment (allows measuring ECE)

ECE Plots vs. Tippett Plots

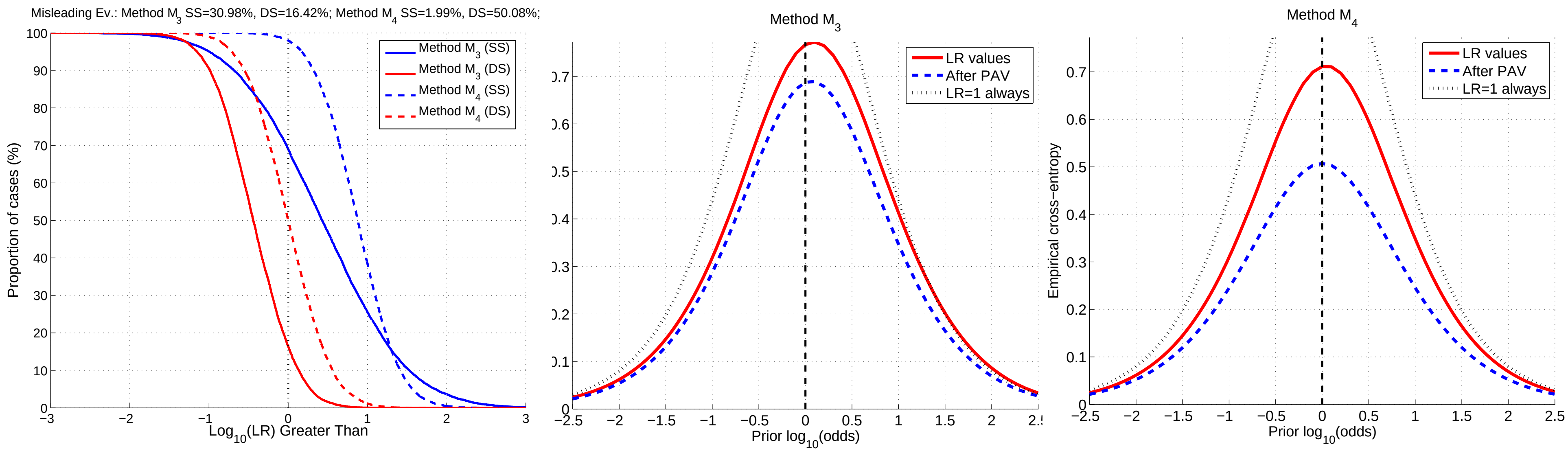
- ECE plots solve many problems of Tippett plots
  - Takes into account <u>strong misleading evidence</u>
    - Strong misleading evidence in M_2 makes ECE (solid curve) grow
      - In fact, using M_2 is even worse than not evaluating the evidence (dotted curve) at extreme prior probabilities
    - It also degrades calibration performance for M_2
      - Difference among solid and dashed curves increases



UNIVERSIDAD AUTONOMA DE MADRID

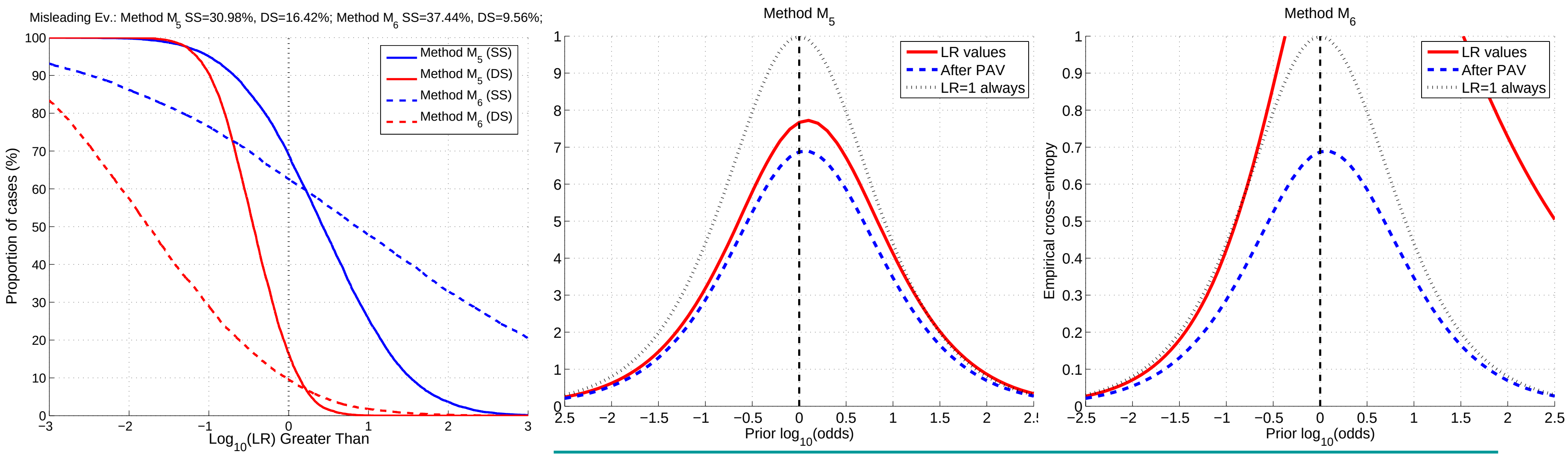
ECE Plots vs. Tippett Plots

- Which method is better, M_3 or M_4?
  - M_3 is slightly worse (slightly higher ECE, solid curve)
  - However, ECE (solid curve) similar in M_3 and M_4
    - Both methods perform similarly
  - Overall performance (ECE, solid curve) is not outstanding
    - Solid curve near dotted curve (not evaluating evidence) in both M_3 and M_4
  - Calibration (difference among solid and dashed curves) is bad in M_4



ECE Plots vs. Tippett Plots

- Discriminating power (dashed curve) is easily seen and compared
  - M_5 and M_6 have the same discriminating power (dashed curve)
  - M_6 has a big calibration problem (big difference among solid and dashed curves)
    - That makes M_6 to be even worse than not evaluating the evidence (solid curve higher than dotted curve)
      - Conclusion: do not use M_6 for evidence evaluation



Limit Tippett Plots: Novel Assessment Tool

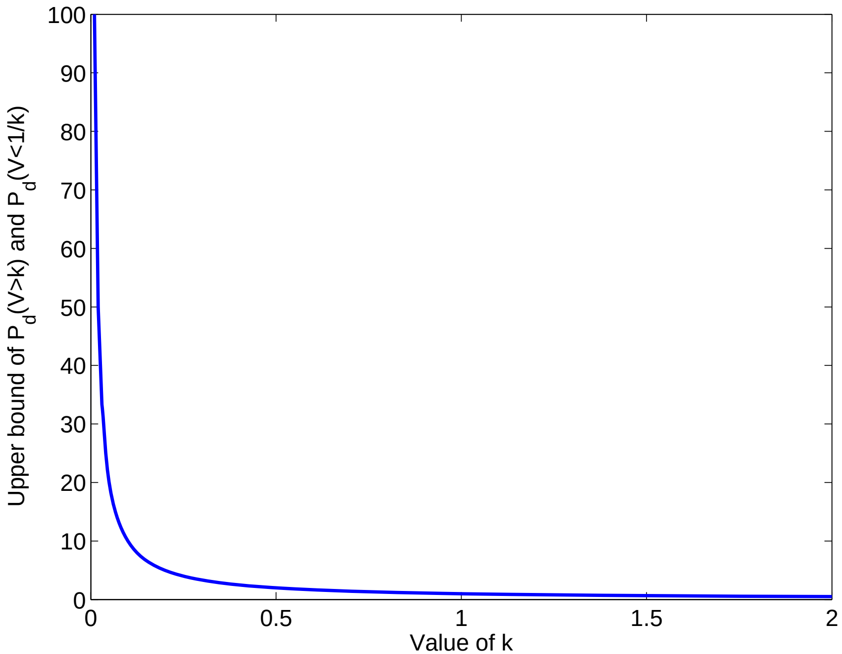
- Let assume that LR values are computed properly
- Then, there is a universal bound for the probability of strong misleading evidence

$$LR = \frac{P(E|\theta_p)}{P(E|\theta_d)}$$

Universal upper bound

$$P(LR > k|\theta_d) < 1/k$$

$$P(LR < k|\theta_p) < 1/k$$



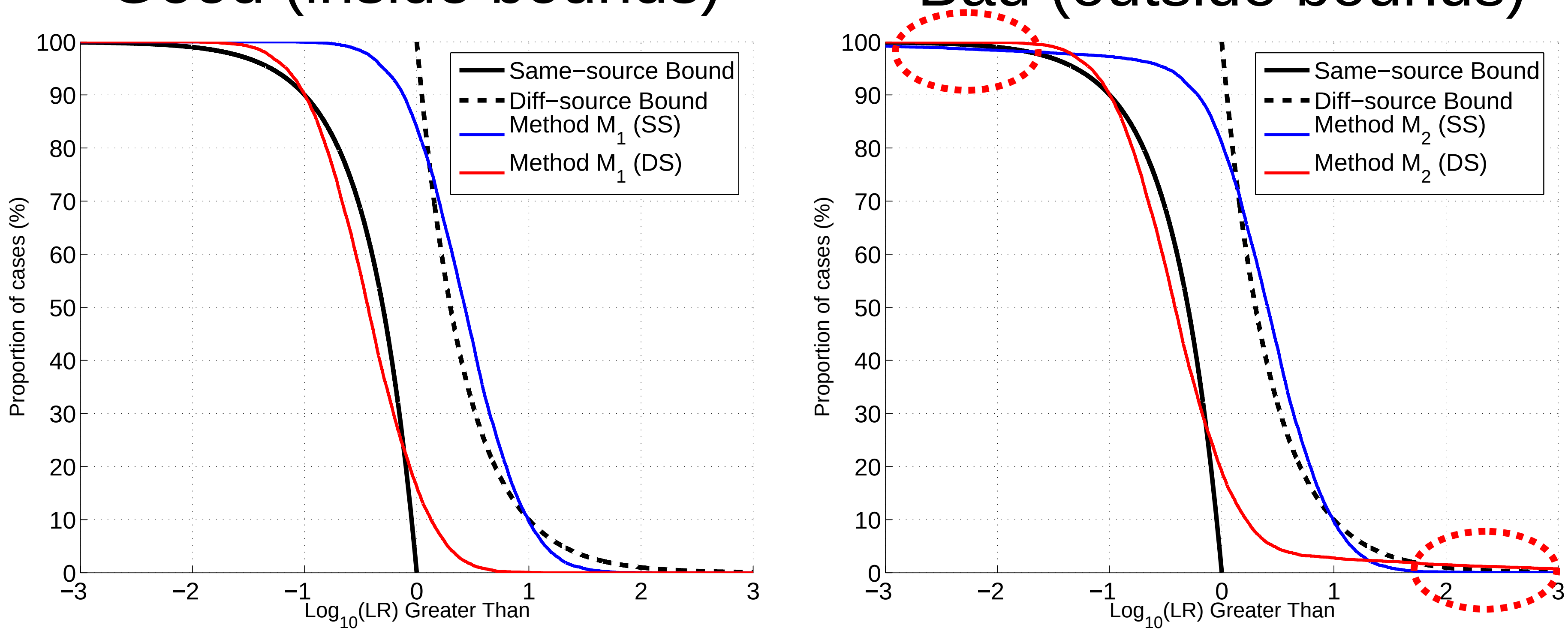
R. Royall, 2000. "On the probability of observing misleading evidence." Journal of the American Statistical Association, v. 95(451), pp. 760-768.

UNIVERSIDAD AUTONOMA DE MADRID

Limit Tippett Plots: Novel Assessment Tool

- Such limits can be drawn in Tippett plots
  - Way of detecting if LR values are correctly obtained

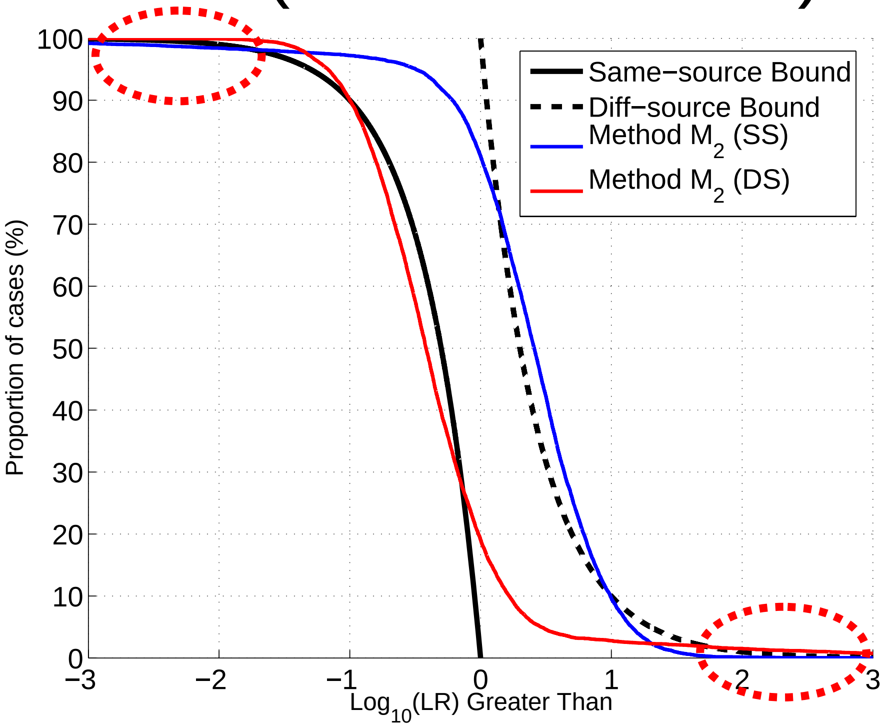
Good (inside bounds) Bad (outside bounds)



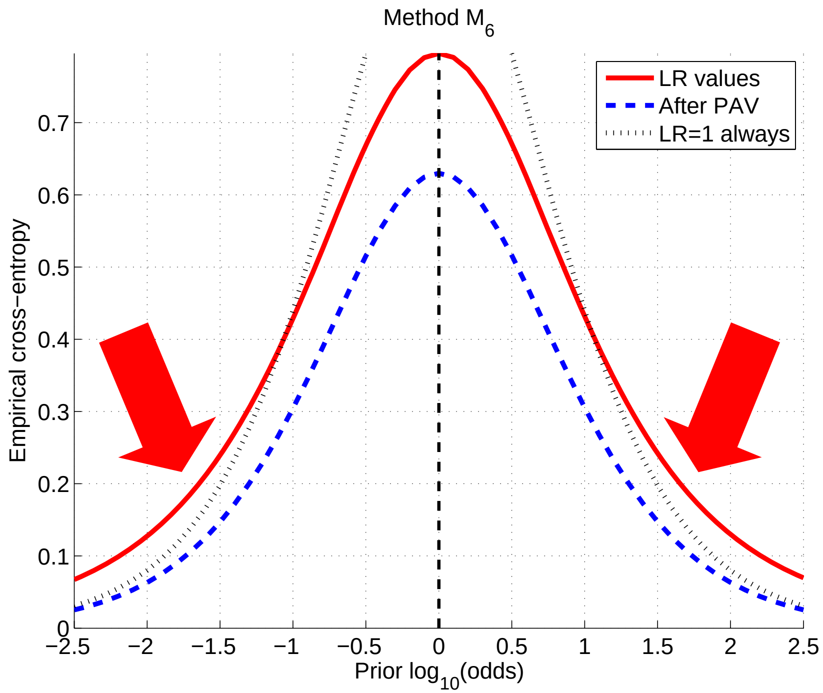
Limit Tippett Plots: Novel Assessment Tool

- Violation of universal bounds related with bad calibration
  - Can be seen in ECE plots
- Limit Tippett plots useful to detect calibration problems

Bad (outside bounds)



Bad calibration



UNIVERSIDAD AUTONOMA DE MADRID

Experimental Example: Forensic Automatic Speaker Recognition

Example with Forensic Speaker Recognition

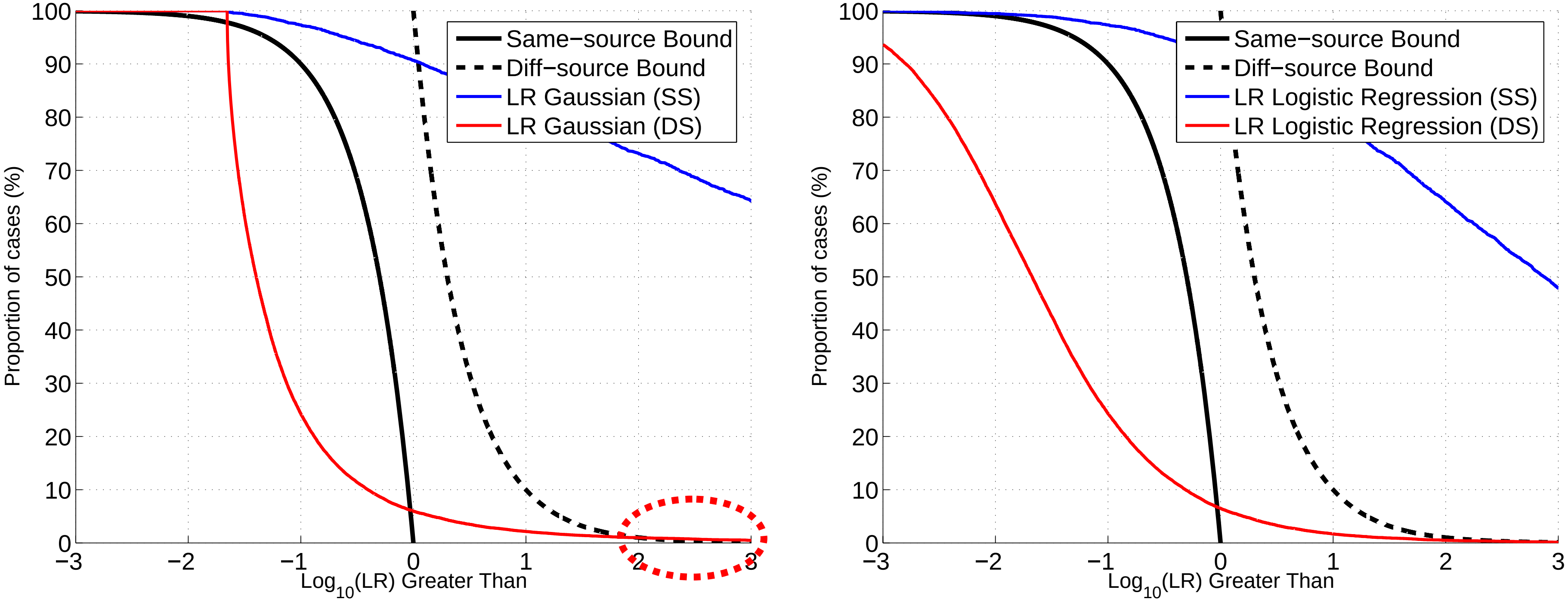
- Common database and protocol for comparisons
  - NIST Speaker Recognition Evaluation (SRE) 2008
  - More than 100,000 comparisons…
- Background data for model tuning
  - Past NIST SRE databases
- Two different evidence evaluation methods for score-based biometric systems
  - Gaussian modelling
  - Logistic Regression

D. Ramos (2007). "Forensic Evidence Evaluation Using Automatic Speaker Recognition Systems". Ph.D. Thesis, Dept. of Computer Science, Univ. Autonoma de Madrid.

J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D. T. Toledano and J. Ortega-Garcia (2007). "Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition." IEEE Transactions on Audio, Speech and Language Processing, 15(7), pp. 2072-2084.

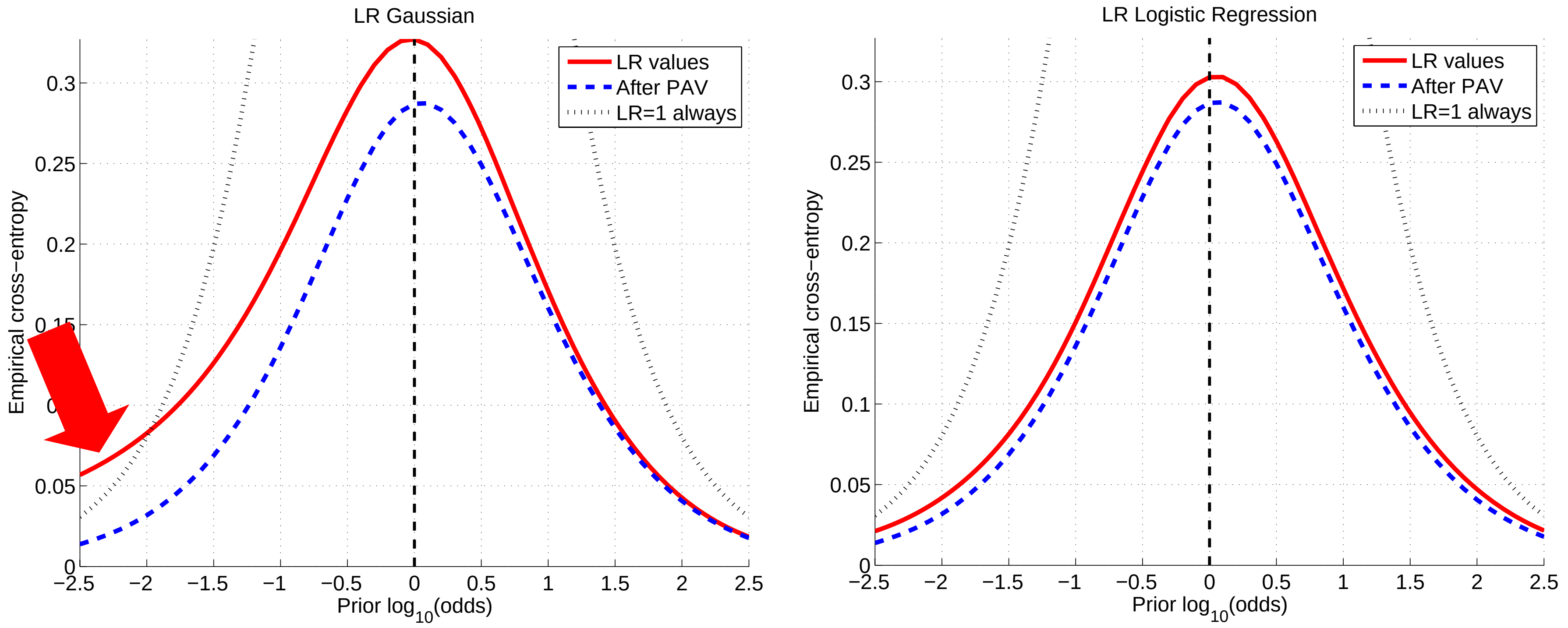
Example with Forensic Speaker Recognition

- Limit Tippett plots
  - Gaussian method slightly out from theoretical bounds
    - Reason: distributions in testing data were not exactly Gaussian



Example with Forensic Speaker Recognition

- ECE plots
  - Calibration is not optimal for Gaussian method
    - Limit Tippett plots detected a calibration problem



Conclusions

- The importance of scientific and objective performance assessment of forensic evidence evaluation methods is recently increasing
  - *How good are we?*
- Likelihood-ratio-based evidence evaluation methods have been assessed in several ways in the literature, e.g.:
  - False positive and false negative rates
  - Tippett plots
- We have reviewed such frameworks, identified their problems and proposed alternatives and improvements
  - ECE plots
  - Limit Tippett plots

Accuracy Assessment Methods for Likelihood-Ratio-based Evidence Evaluation

Daniel Ramos and Joaquin Gonzalez-Rodriguez
ATVS – Biometric Recognition Group
Universidad Autónoma de Madrid
http://atvs.ii.uam.es

Jose-Juan Lucena-Molina
Statistics Department, Criminalistics Service
Dirección General de la Policía y de la Guardia Civil
Ministerio del Interior, Spain