On the Calibration of Likelihood Ratios

Daniel Ramos
ATVS – Biometric Recognition Group
Universidad Autonoma de Madrid
daniel.ramos@uam.es
http://atvs.ii.uam.es

WIC-BBfor2 Midwinter Meeting

Outline

- Likelihood Ratio (LR) Framework in Forensic Science
- Assessing LR Performance
- Calibration of LR values
- Some Case Studies
- Challenges and Conclusions

Likelihood Ratio Framework in Forensic Sciences

Likelihood Ratios (LR) in Forensic Science

- Given two materials to compare
  - Evidence (E)
  - *E.g.*, biological samples in crime scene and from a suspect, speech from wire-tapping and from a suspect…
- Relevant hypotheses (at source level)
  - Hypothesis θ_p: materials come from the same source
  - Hypothesis θ_d: materials come from different sources
- Other information in the case (I)

$$\frac{P(\theta_p|E,I)}{P(\theta_d|E,I)} = \underbrace{\boxed{\frac{P(E|\theta_p,I)}{P(E|\theta_d,I)}}}_{LR} \frac{P(\theta_p|I)}{P(\theta_d|I)}$$

Likelihood Ratios in Forensic Science

Forensic Laboratory

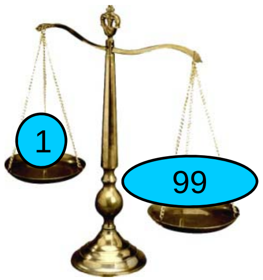
Forensic scientist

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

LR= 1000

Province of the fact finder

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



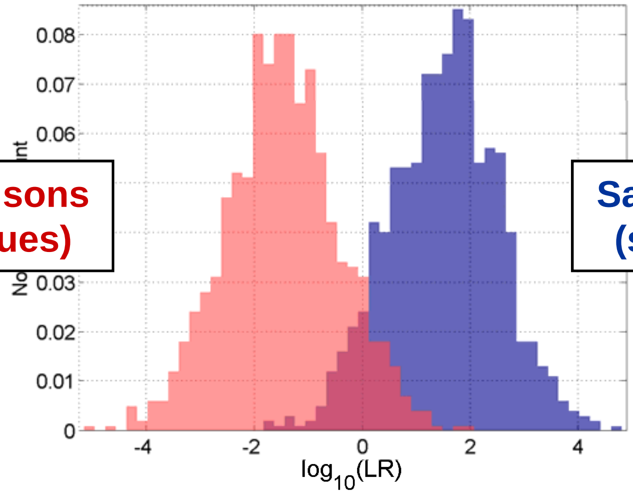
Inference

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

Assessing LR Performance

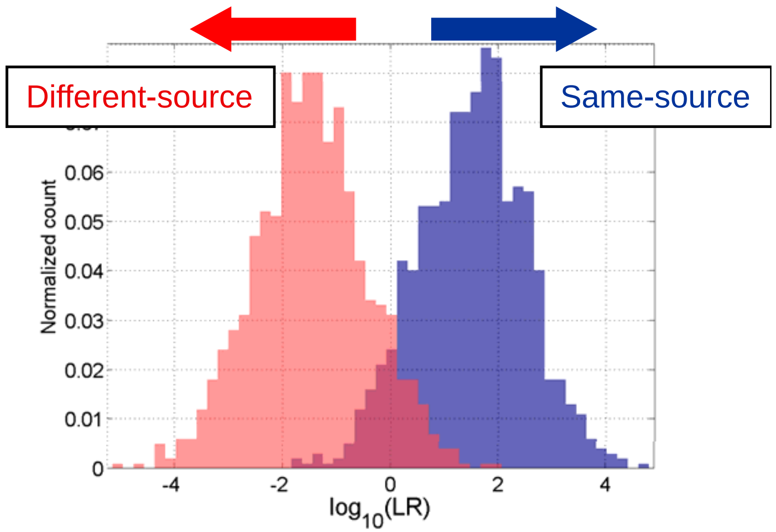
Empirical Assessment of Performance

- Experimental test
  - Database of data with known sources
    - *E.g.*, speech database with known identities of speakers
  - Generate same-source comparisons (θ_p is known to be true)
    - LR values should be higher than 1
  - Generate different-source comparisons (θ_d is known to be true)
    - LR values should be lower than 1



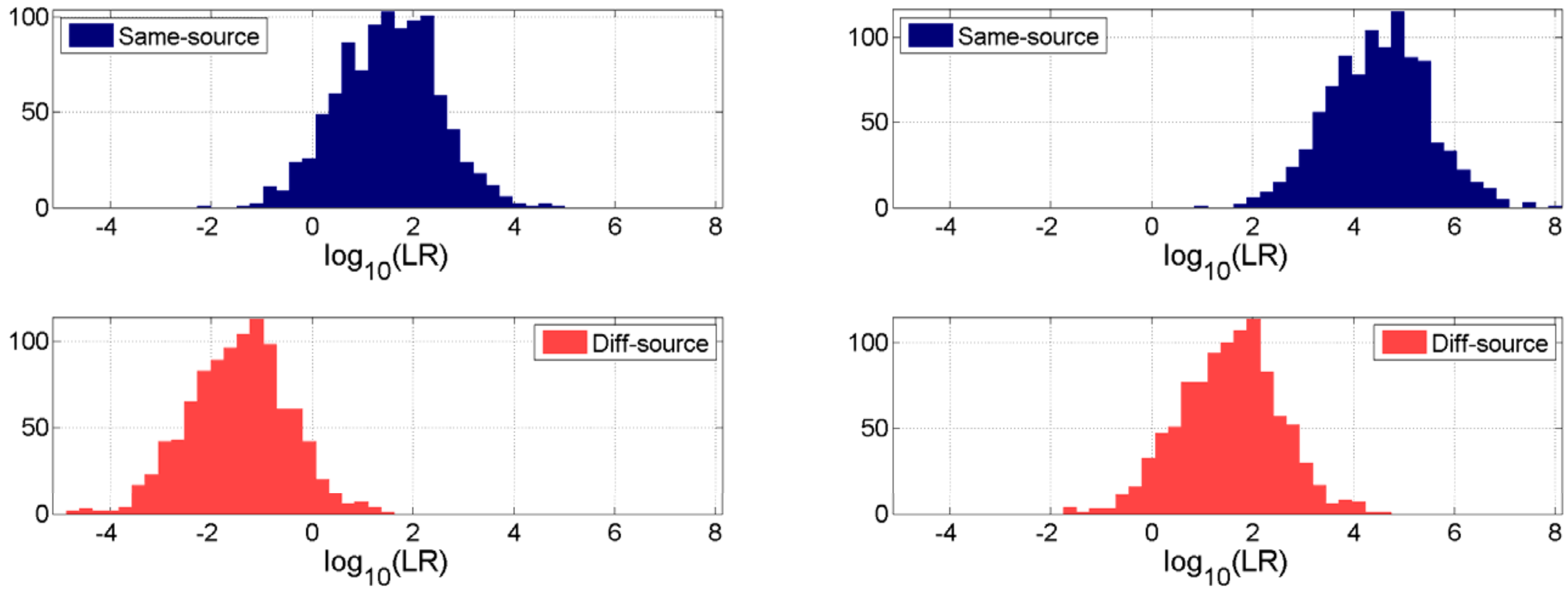
Discriminating Power of the Evidence

- Discriminating power (or simply discrimination) of the evidence is related to the separation (overlapping) among
  - LR values for which θ_p is true
    - Samples come from the same source
  - LR values for which θ_d is true
    - Samples come from different sources
- Good discriminating power means
  - Higher LR values for same-source comparisons
  - Lower LR values for different-source comparisons
- Measured by *e.g.* ROC and DET plots.



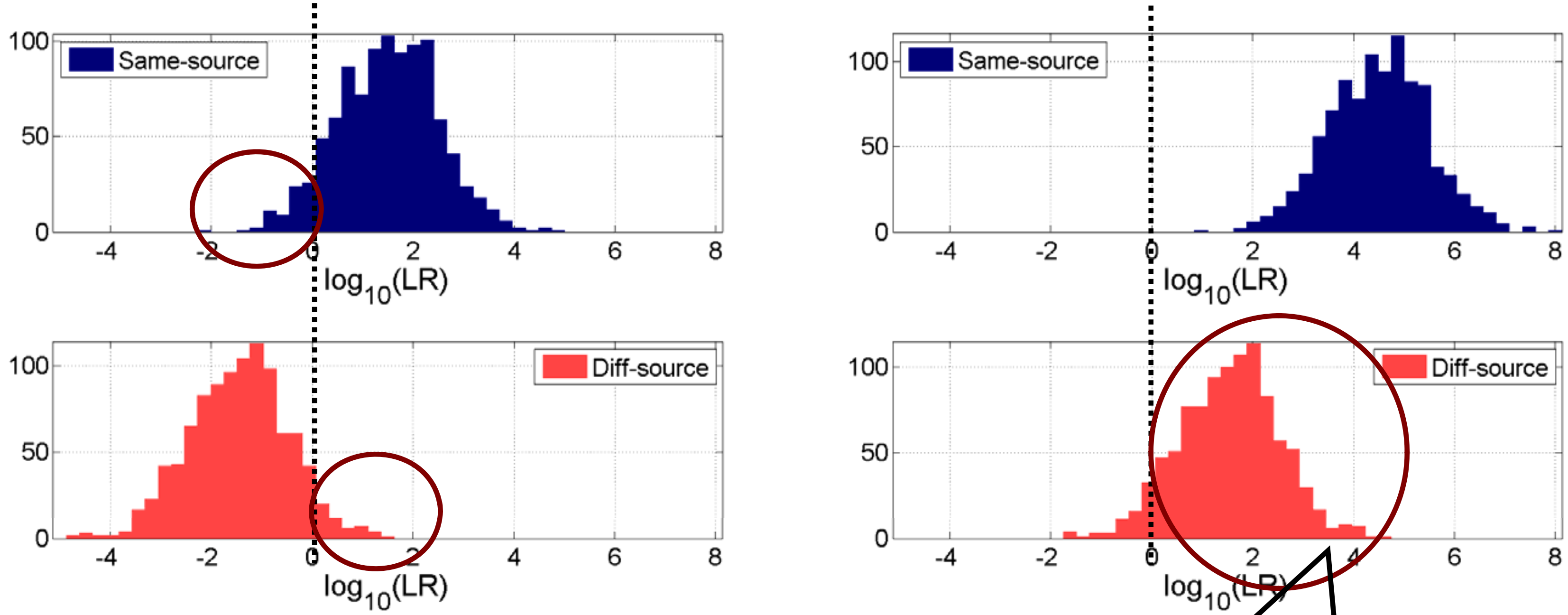
Discrimination is not enough for LR

- Example: two LR sets with the same discrimination
  - Second set is first set +3



Discrimination is not enough for LR

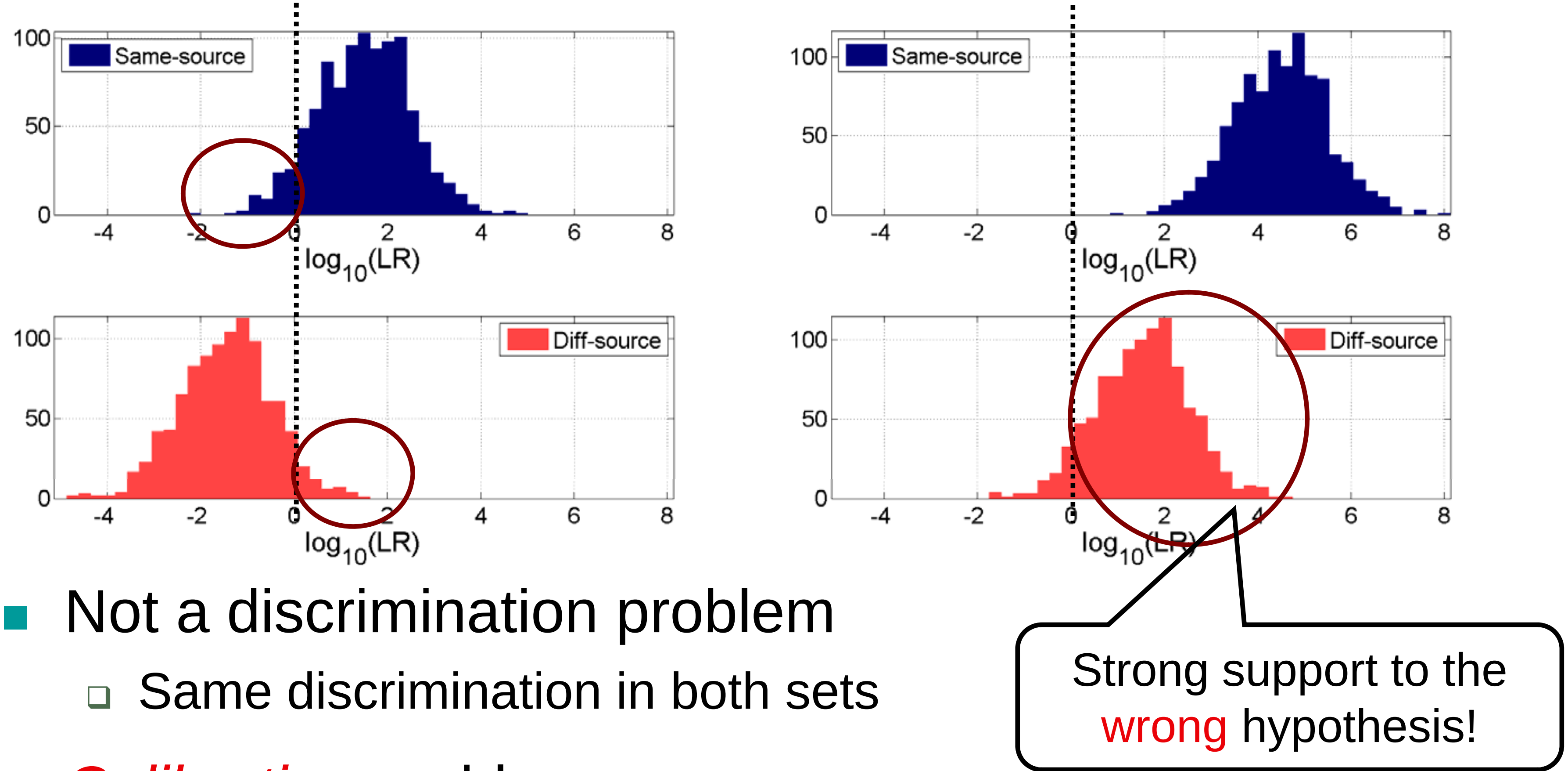
- Example: two LR sets with the same discrimination
  - Second set is first set +3



Strong support to the wrong hypothesis!

Discrimination is not enough for LR

- Example: two LR sets with the same discrimination
  - Second set is first set +3



- Not a discrimination problem
  - Same discrimination in both sets
- *Calibration* problem

Performance of Posterior Probabilities

- Performance of a probabilistic opinion (*forecast*)
  - Classically measured by Strictly Proper Scoring Rules (SPSR)
    - [deGroot82, Dawid07,Gneiting07]
- A SPSR rule assigns a penalty to a probabilistic opinion
  - Depending on which hypothesis is actually true
- In LR-based forensic evidence evaluation, the *forecast* is expressed by the posterior probabilities

$$P\left(\theta_p \middle| E\right)$$

I out from notation (simplicity)

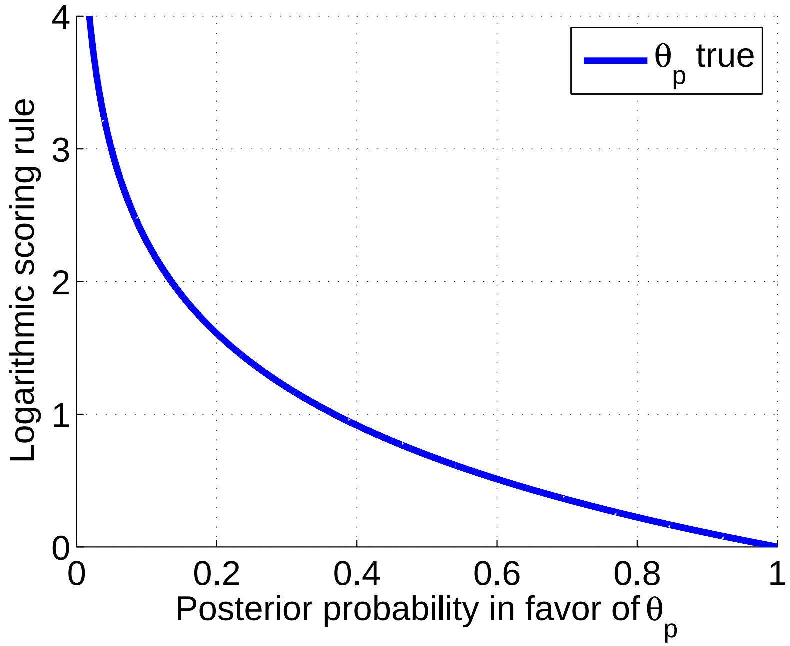
- Prior probabilities, province of the fact finder, are still needed…
  - We will address this issue later

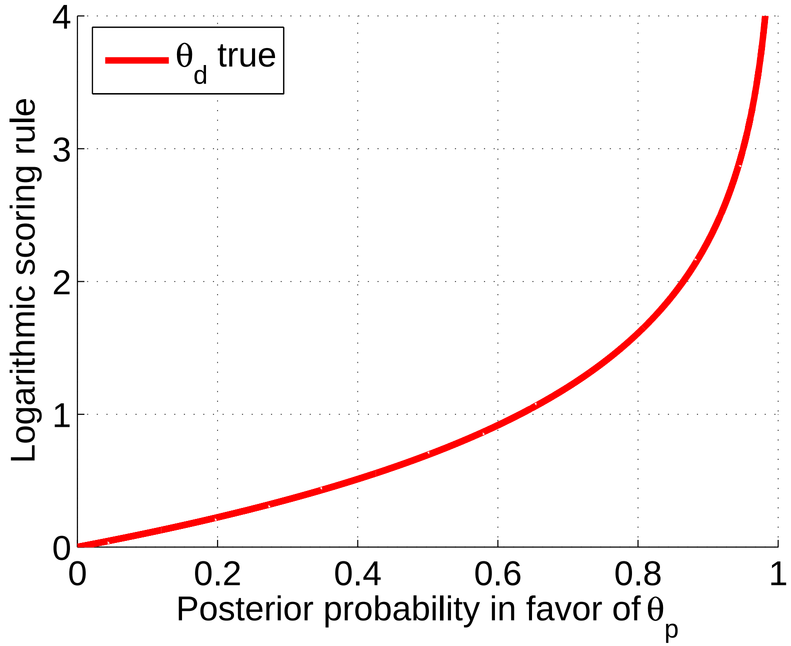
Example: Logarithmic SPSR

- Assigns:

$$-\log_2 P\left(\theta_p \middle| E\right) \qquad \theta_p \text{ is true}$$

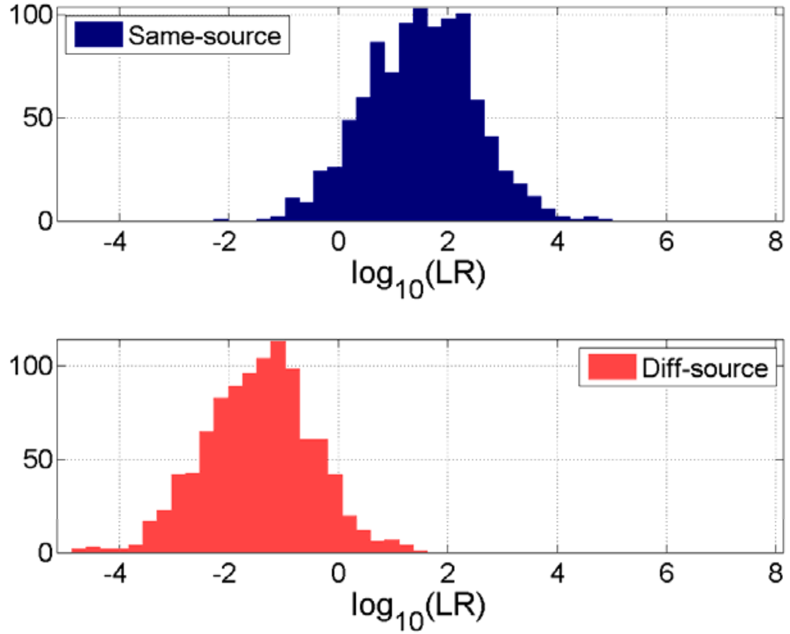
$$-\log_2 P\left(\theta_d \middle| E\right) \qquad \theta_d \text{ is true}$$





Likelihood Ratios (LR) in Forensic Science

- Performance of a set of posterior probabilities (forecasts)
  - Average of a SPSR over different comparisons [deGroot82, Dawid07,Gneiting07]



$$LS = -\frac{1}{N_{ss}} \sum_{i \in \text{same-source}} \log_2 P\left(\theta_p \middle| E_i\right) - \frac{1}{N_{ds}} \sum_{j \in \text{diff-source}} \log_2 P\left(\theta_d \middle| E_j\right)$$

Empirical Cross-Entropy (*ECE*)

- Prior-weighted average of the logarithmic SPSR

$$LS = -\frac{1}{N_{ss}} \sum_{i \in \text{same-source}} \log_2 P(\theta_p | E_i) - \frac{1}{N_{ds}} \sum_{j \in \text{diff-source}} \log_2 P(\theta_d | E_j)$$

$$ECE = -\frac{P(\theta_p)}{N_{ss}} \sum_{i \in \text{same-source}} \log_2 P(\theta_p | E_i) - \frac{P(\theta_d)}{N_{ds}} \sum_{j \in \text{diff-source}} \log_2 P(\theta_d | E_j)$$

- Information-theoretical interpretation [Ramos07]
  - "Average information needed to obtain certainty"
    - Higher ECE means more information needed to know which hypothesis is actually true
    - Using the LR values computed by the forensic scientist

Calibration of LR Values

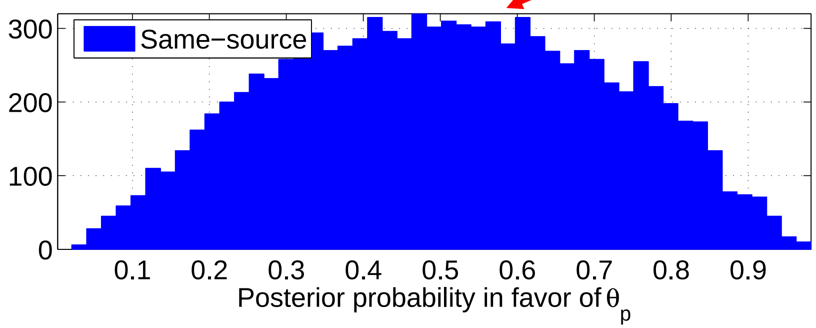
Calibration

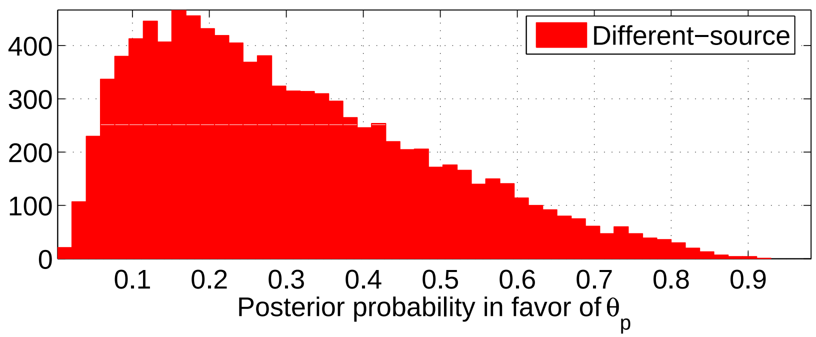
- Given a set of posterior probabilities about hypothesis θ_p, **calibration** means
  - Posterior probabilities of θ_p approximate actual proportions of occurrence of θ_p
- Calibrated probabilities have been dubbed ***reliable*** [deGroot82]
- Calibration improves performance of forecasts
  - Because the average of any SPSR is decomposed [deGroot82]
  - A **refinement loss** component
    - Measure of **discrimination** [Brummer06]
  - A **calibration loss** component

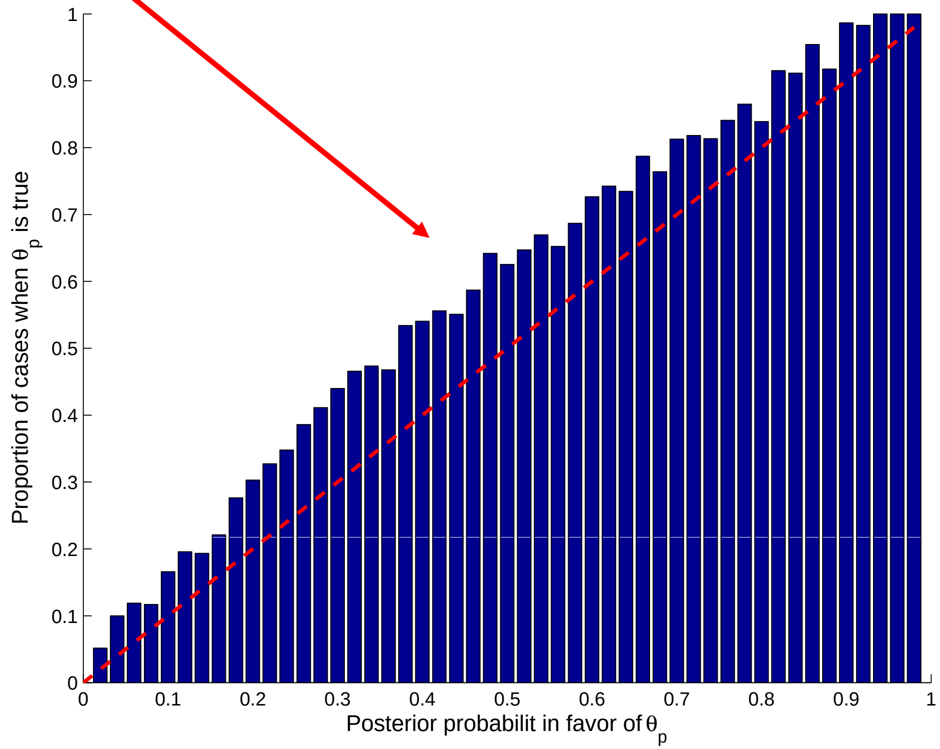
Calibration

- Example: experimental set of posterior probabilities
  - LR values computed by a forensic scientist
  - Fact finder assigns $P(\theta_p)=0.5$

Uncalibrated



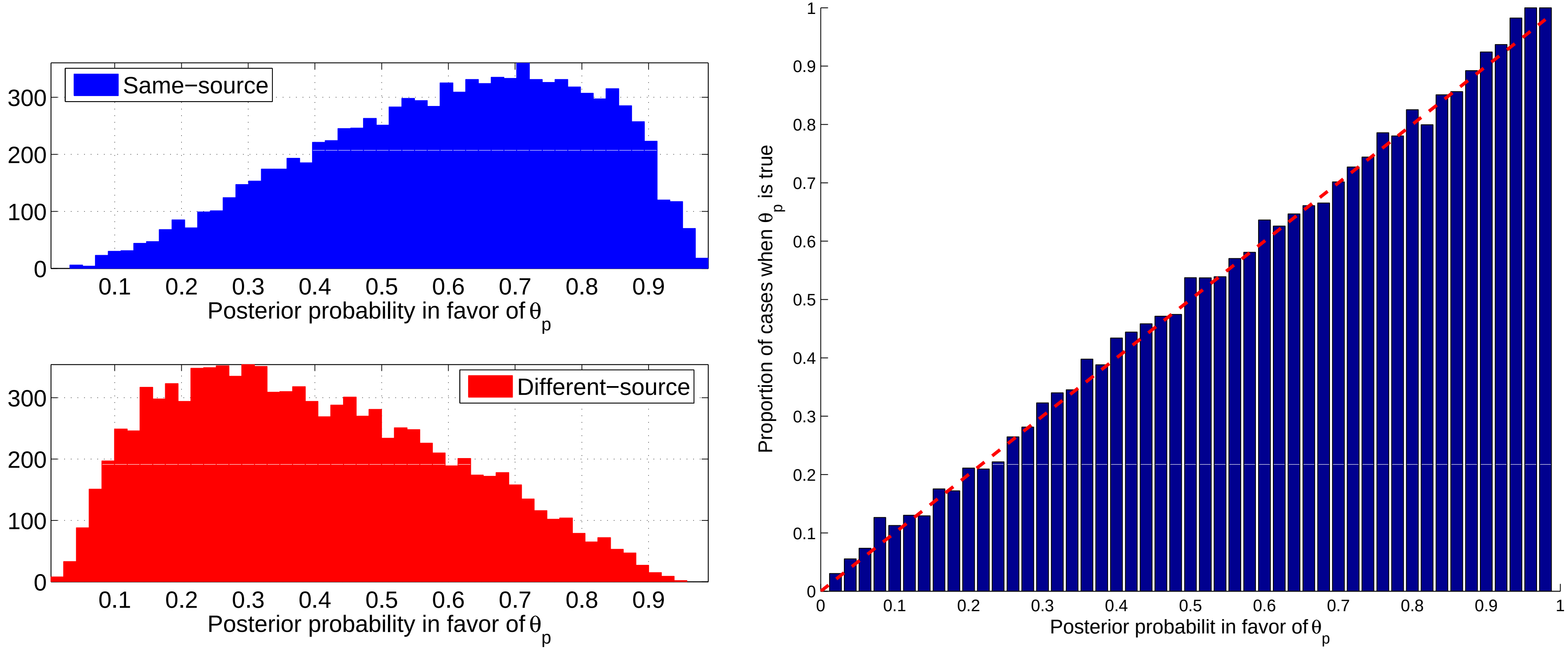




Calibration

- Example: other set of likelihood ratios presenting the same discrimination as before
  - Rest of the conditions unchanged

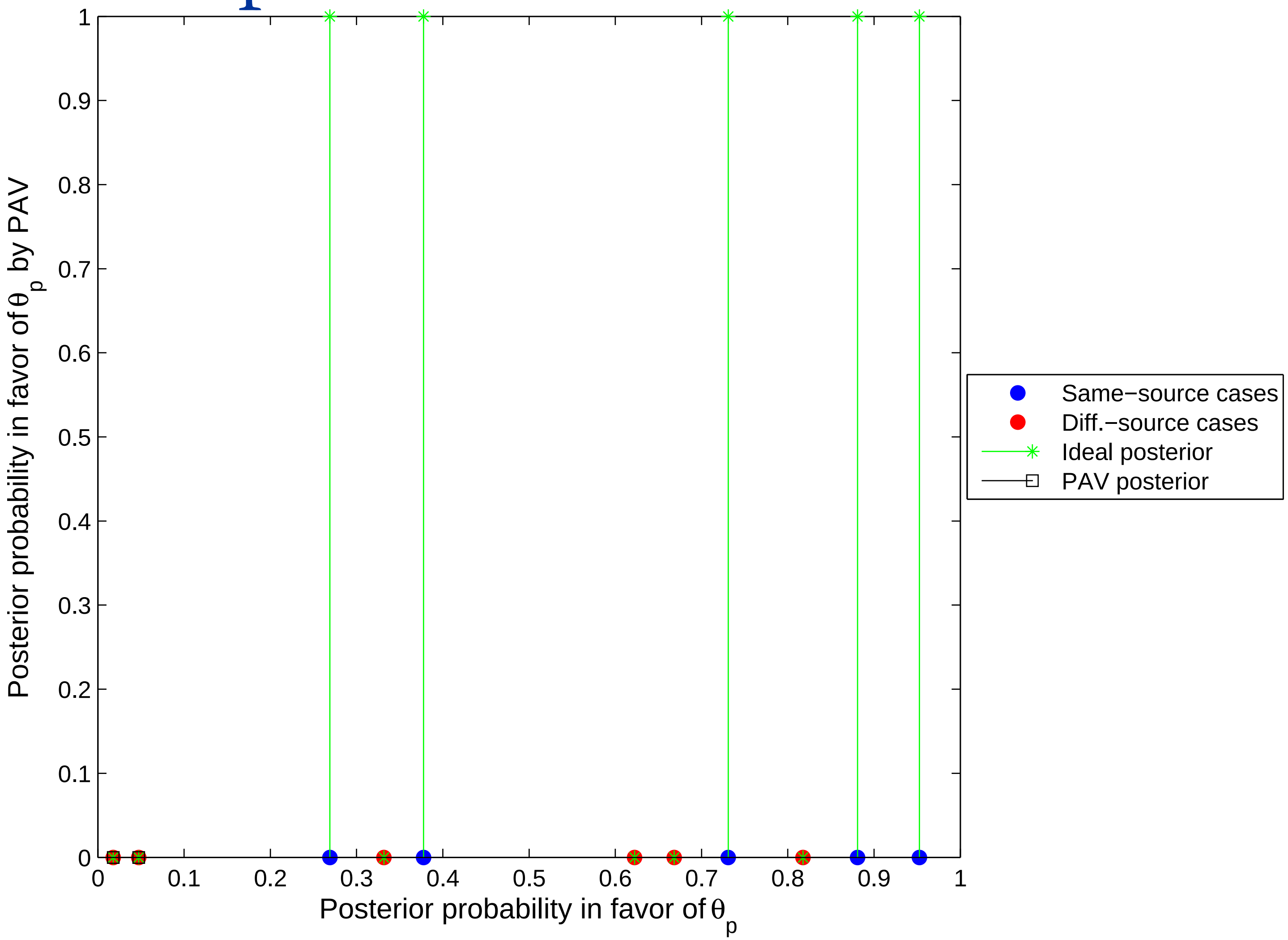
Calibrated



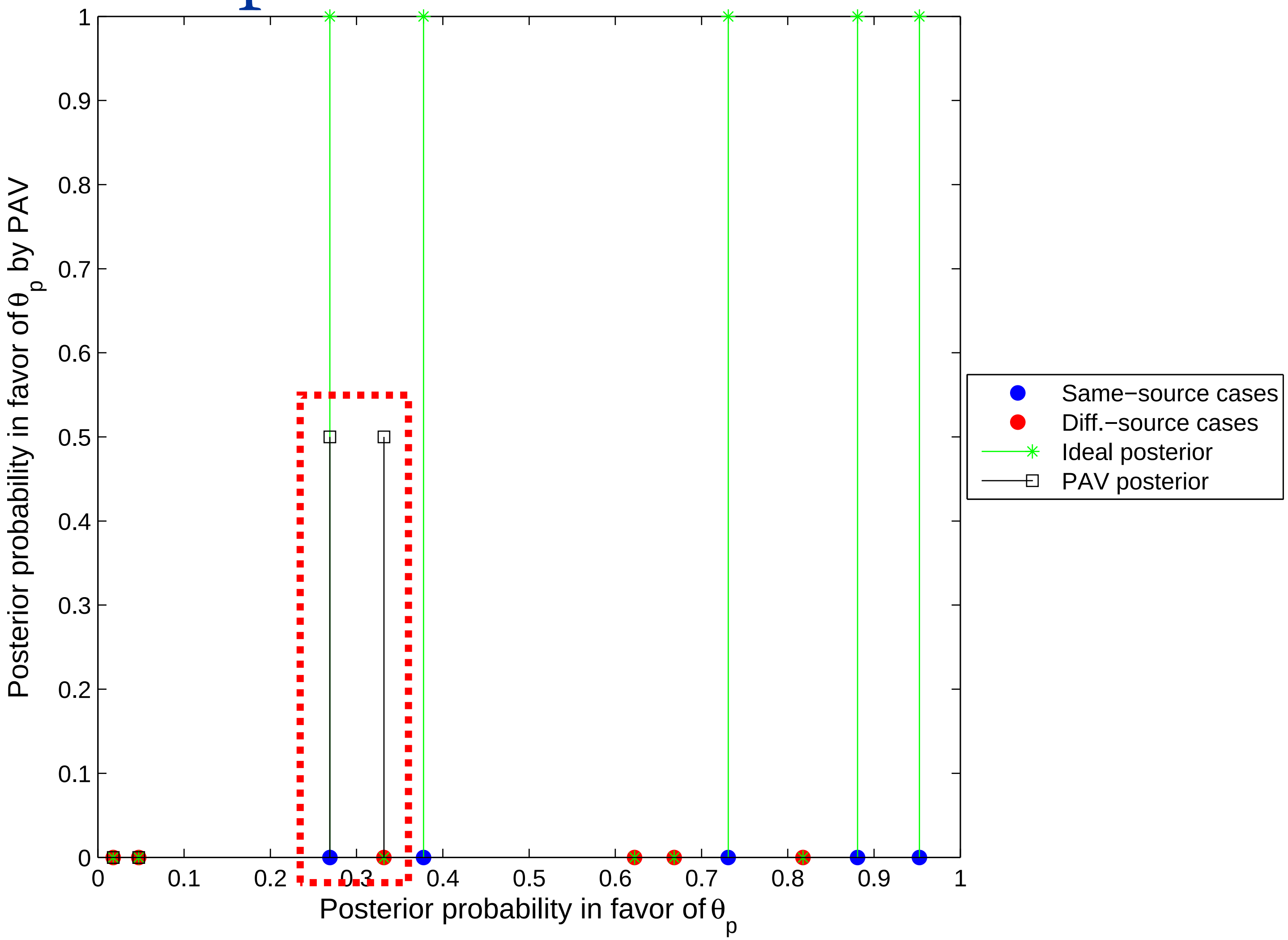
Obtaining calibrated probabilities

- Computing proportions of cases implies binning posterior probabilities
  - How many bins? What bin size?
- A solution: Pool Adjacent Violators Algorithm (PAV) [Brummer06,vanLeeuwen07]
  - Computation of proportions over the experimental set of probabilities (where true answers are known)
  - Monotonically rising (isotonic regression)
    - Preserves discrimination
    - Only calibration is improved

PAV: example

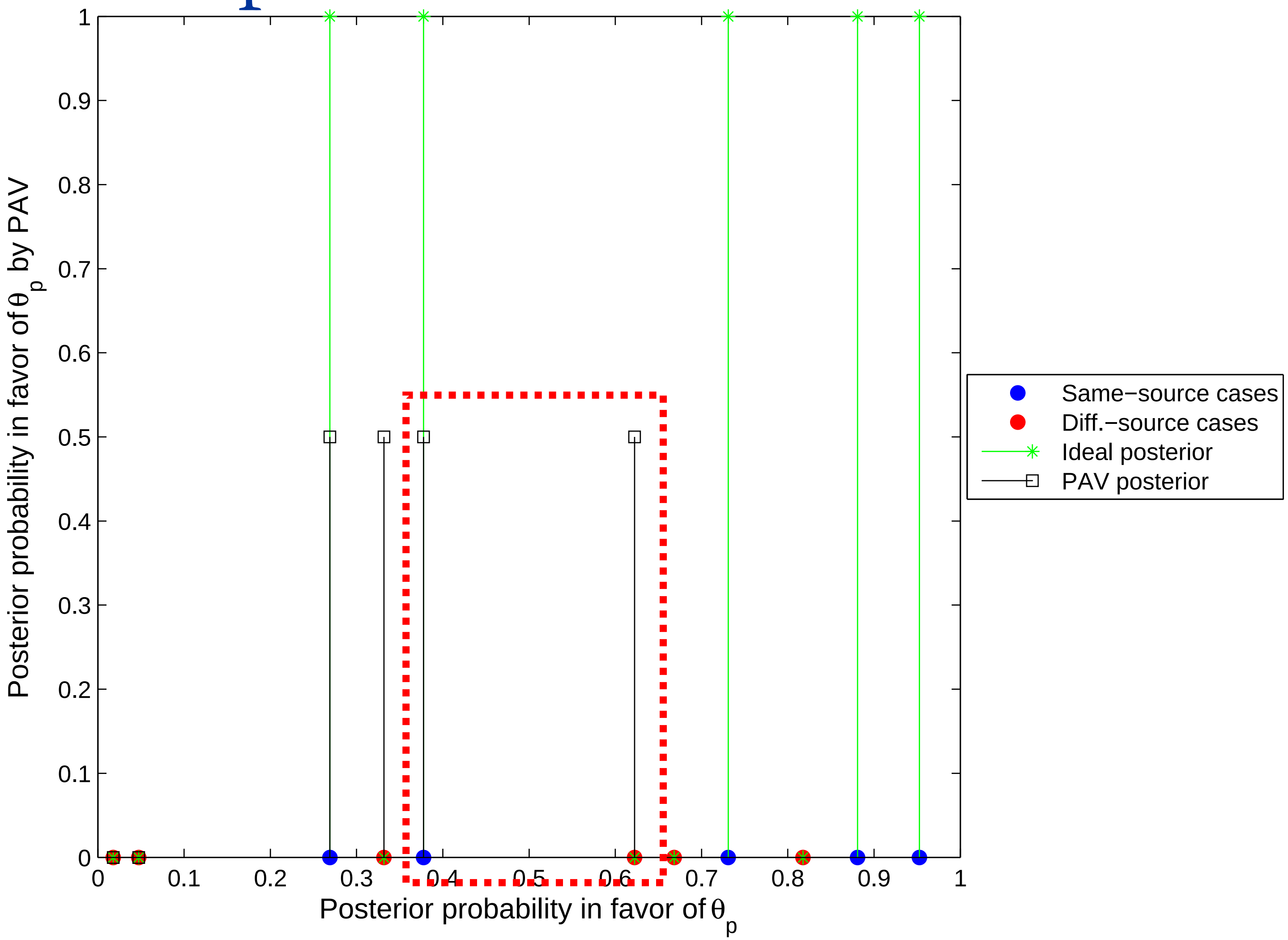

Posterior probability in favor of θp by PAV
Posterior probability in favor of θp
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Same−source cases
Diff.−source cases
Ideal posterior
PAV posterior

PAV: example



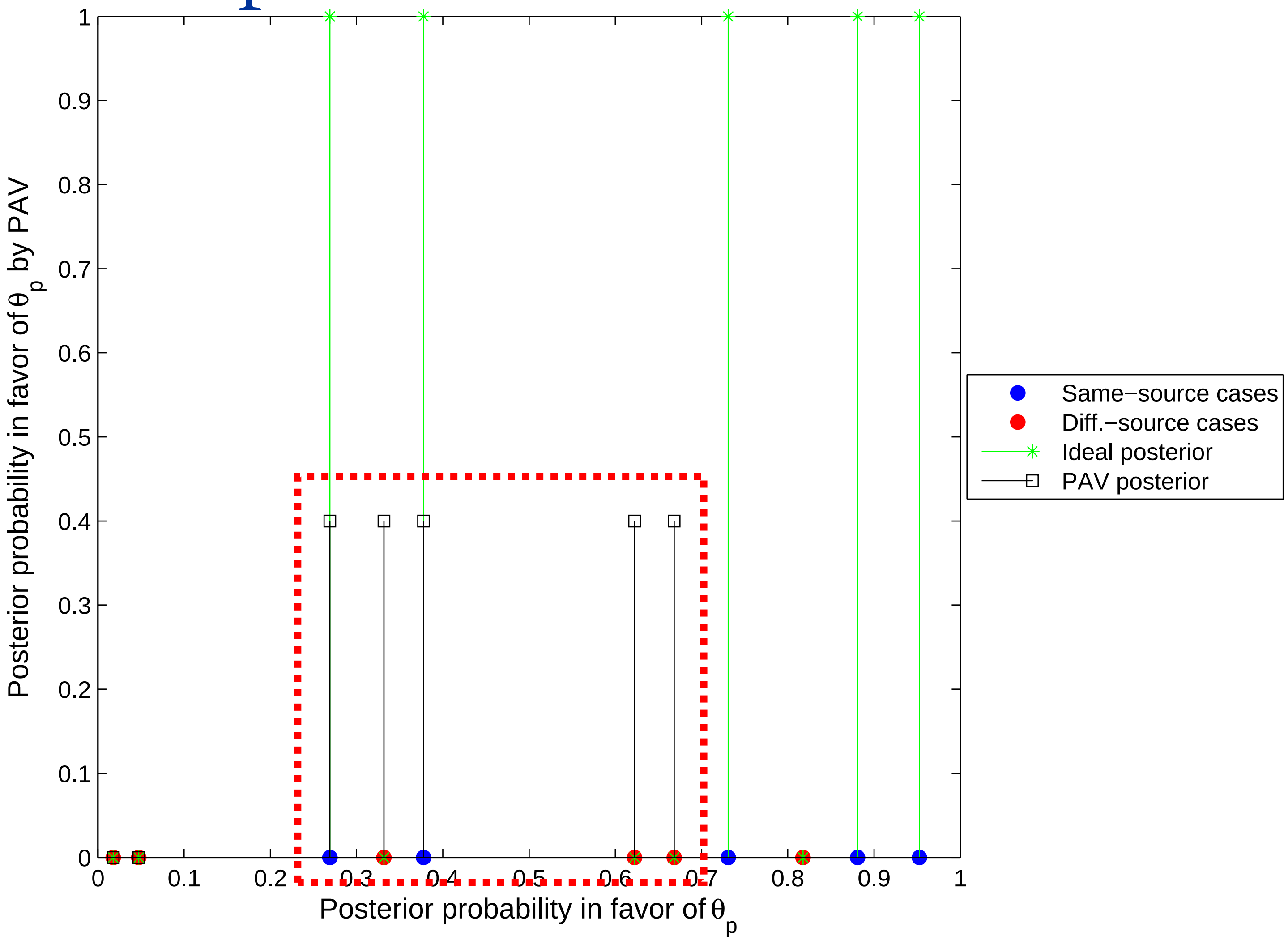
Decreasing "violators": pool them together and average output probabilities

PAV: example



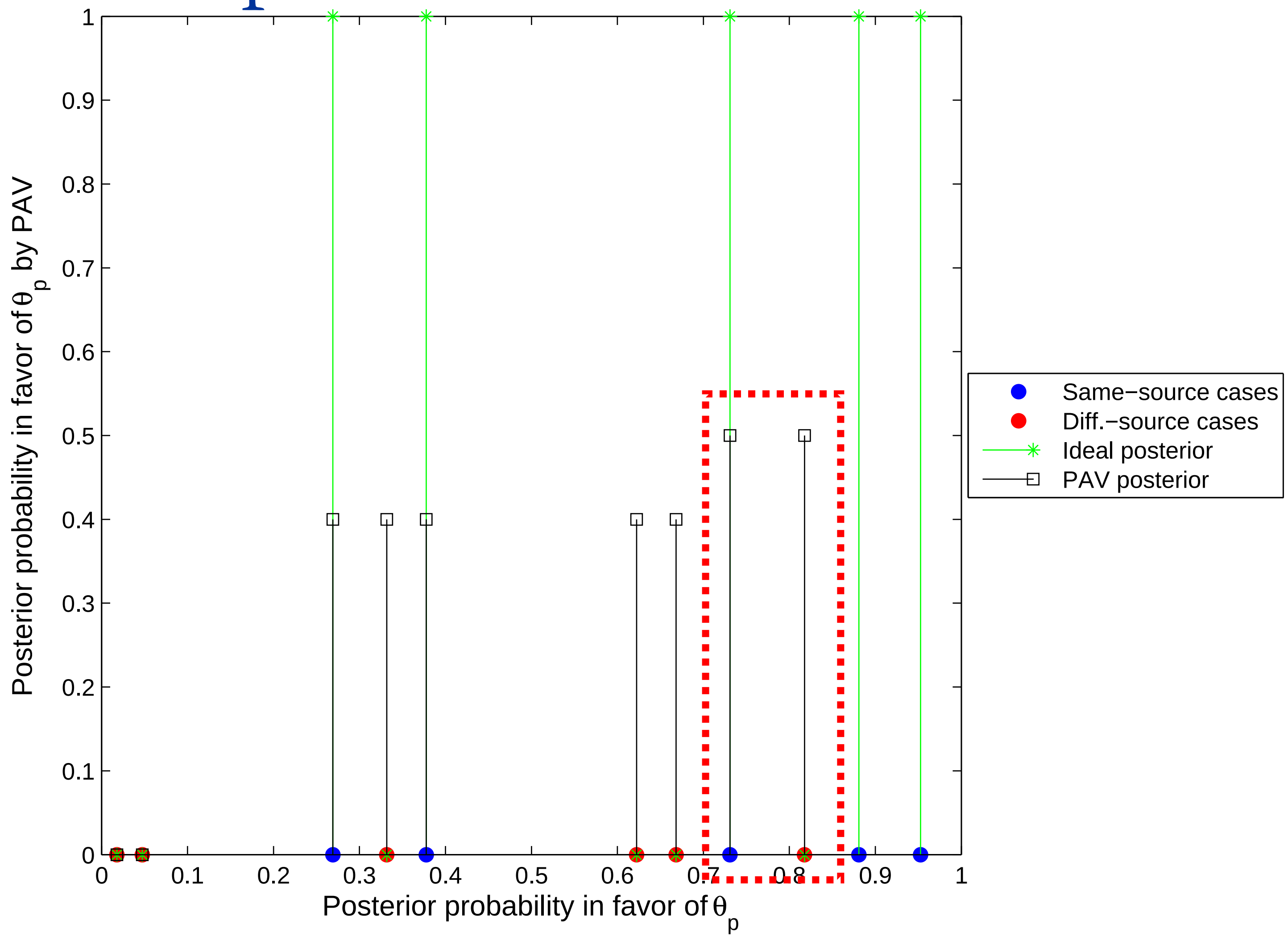
Decreasing "violators": pool them together and average output probabilities

PAV: example



Decreasing "violators": pool them together and average output probabilities

PAV: example

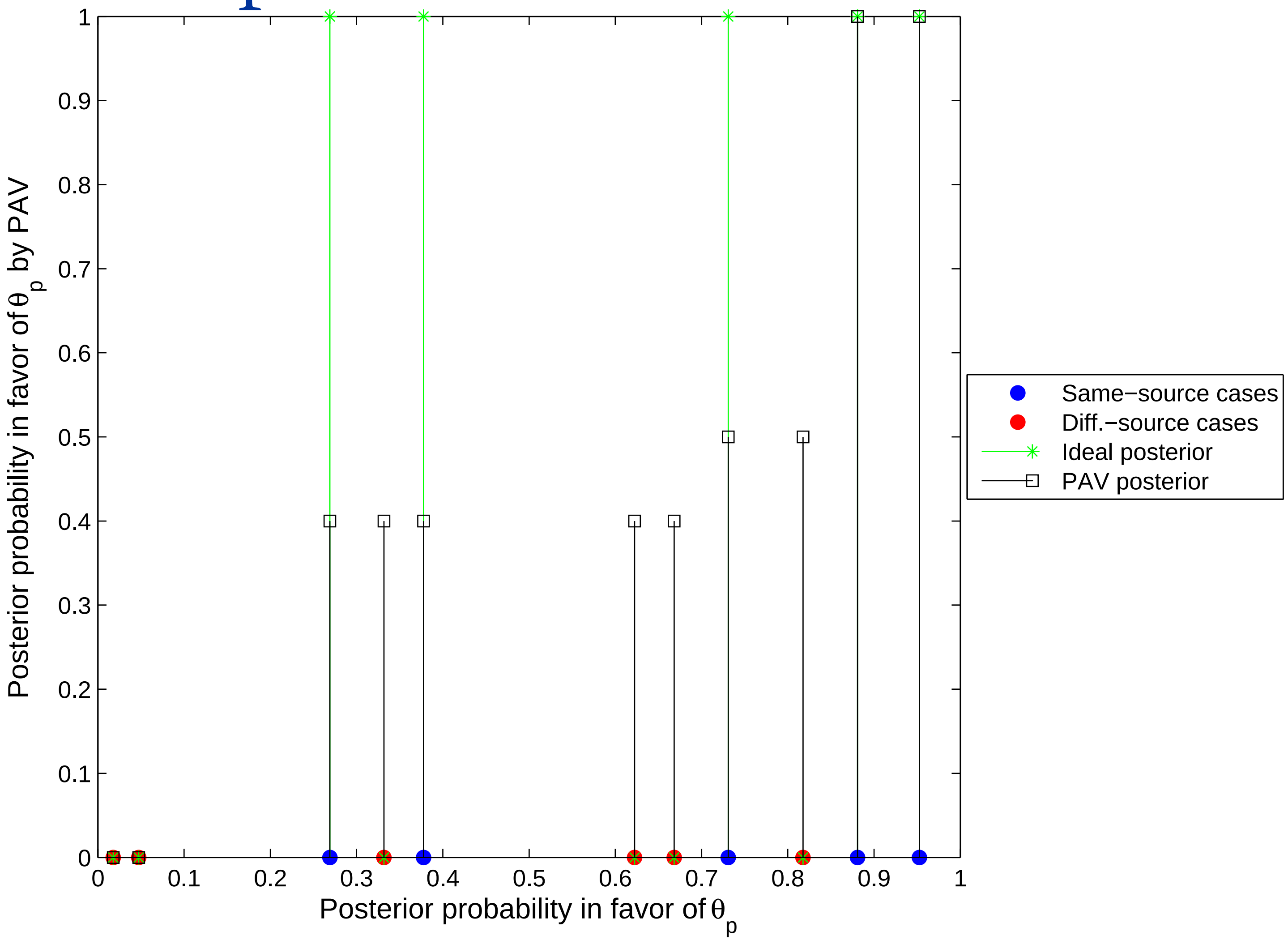


Decreasing "violators": pool them together and average output probabilities

ATVS

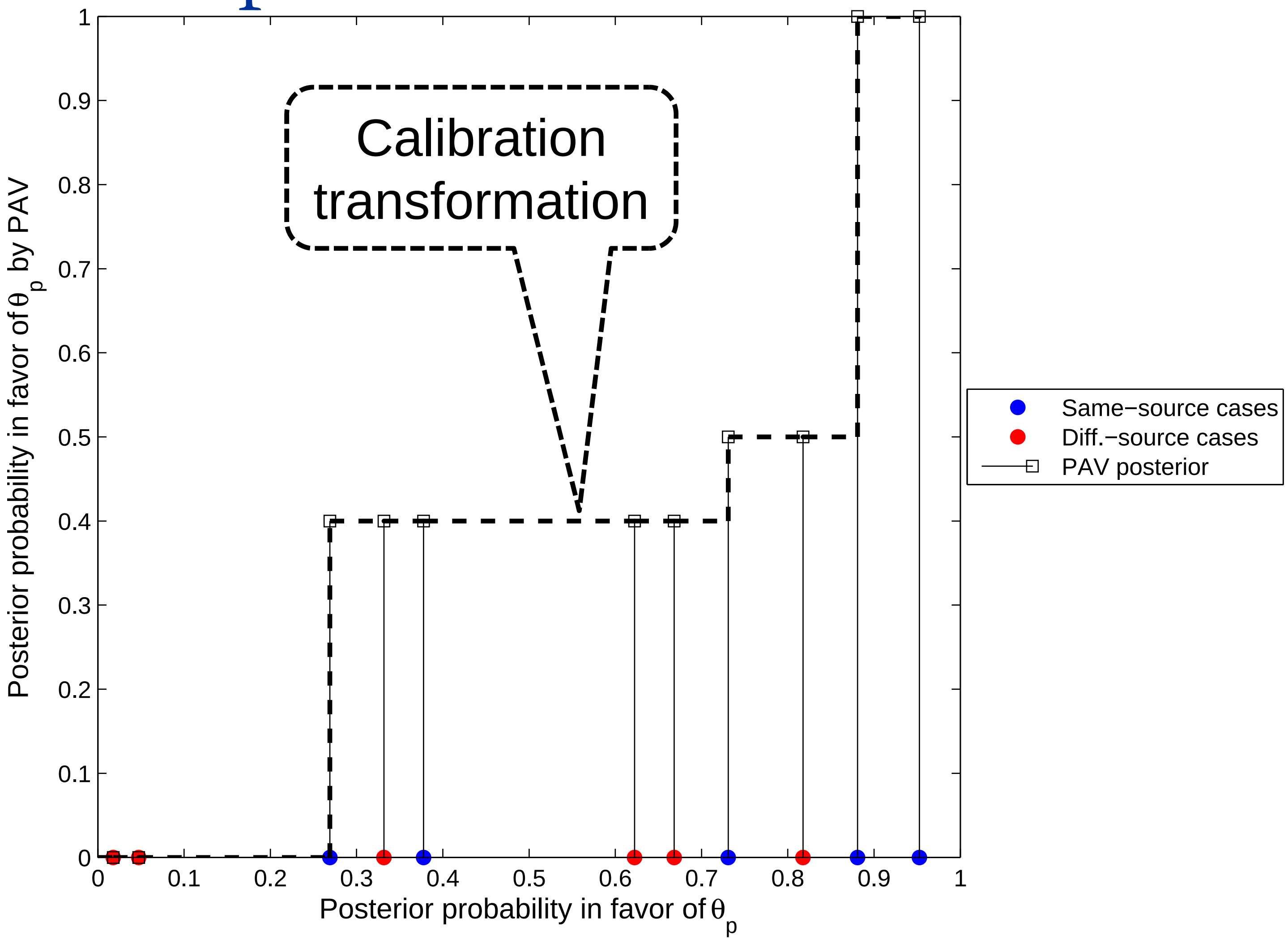
UAM UNIVERSIDAD AUTONOMA DE MADRID

PAV: example



1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Posterior probability in favor of θp by PAV
Posterior probability in favor of θp
Same−source cases
Diff.−source cases
Ideal posterior
PAV posterior

PAV: example


Calibration transformation
Posterior probability in favor of θp by PAV
Posterior probability in favor of θp
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Same−source cases
Diff.−source cases
PAV posterior

Calibration and *ECE*

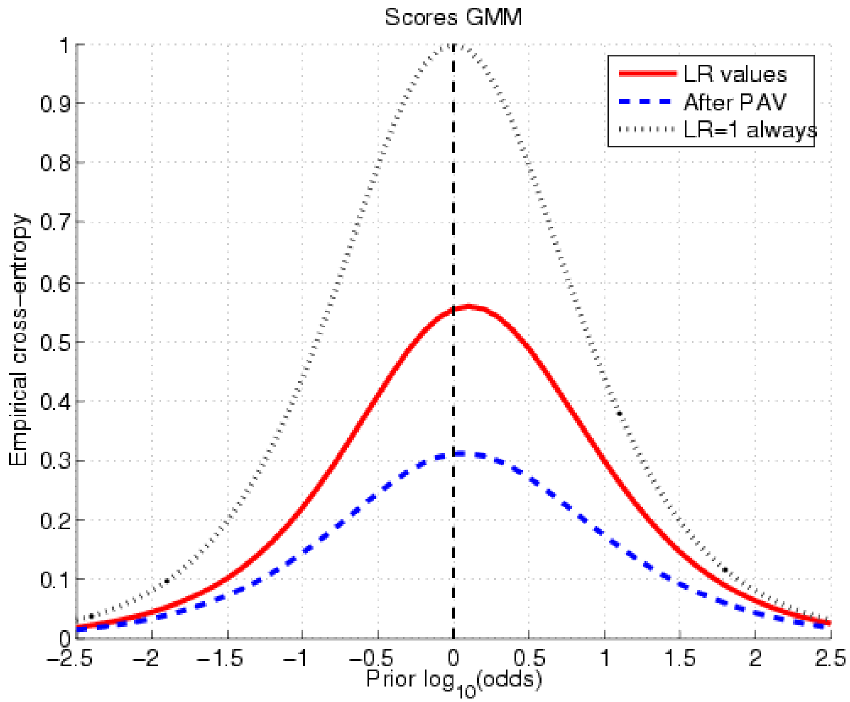
- Improving calibration improves (reduces) *ECE*
  - Because *ECE* decomposes into discrimination + calibration

$$ECE = -\frac{P(\theta_p)}{N_{ss}} \sum_{i \in \text{same-source}} \log_2 P(\theta_p | e_i) - \frac{P(\theta_d)}{N_{ds}} \sum_{j \in \text{diff-source}} \log_2 P(\theta_d | e_j)$$

- However, *ECE* still needs the prior probability…
  - The forensic scientist cannot compute its value in general
- Solution: the *ECE* plot
  - Computing *ECE* for a wide range of priors

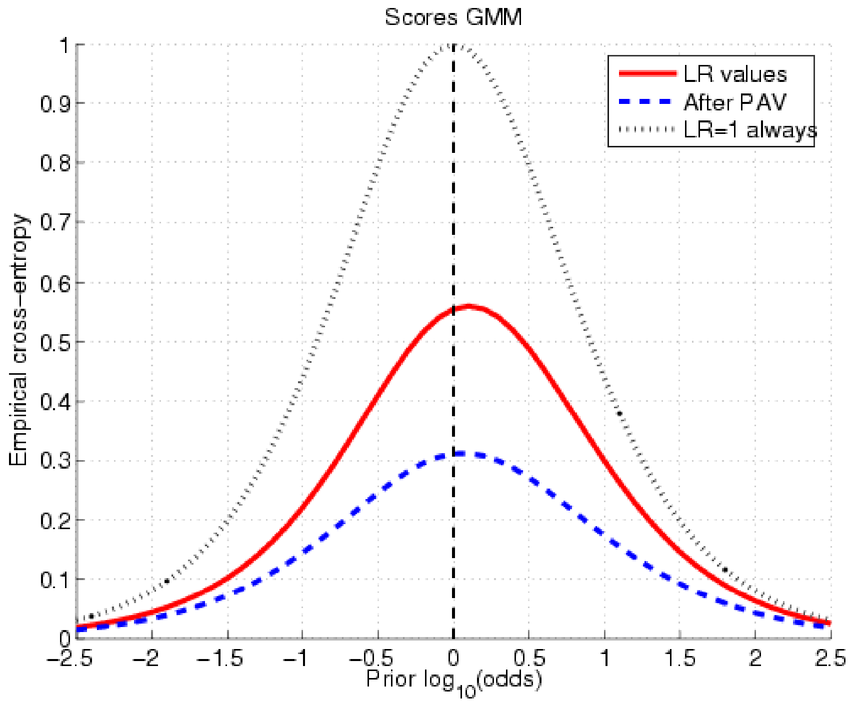
ECE plots: LR performance

- *ECE* of 3 LR sets are represented
  - LR values actually obtained (solid)



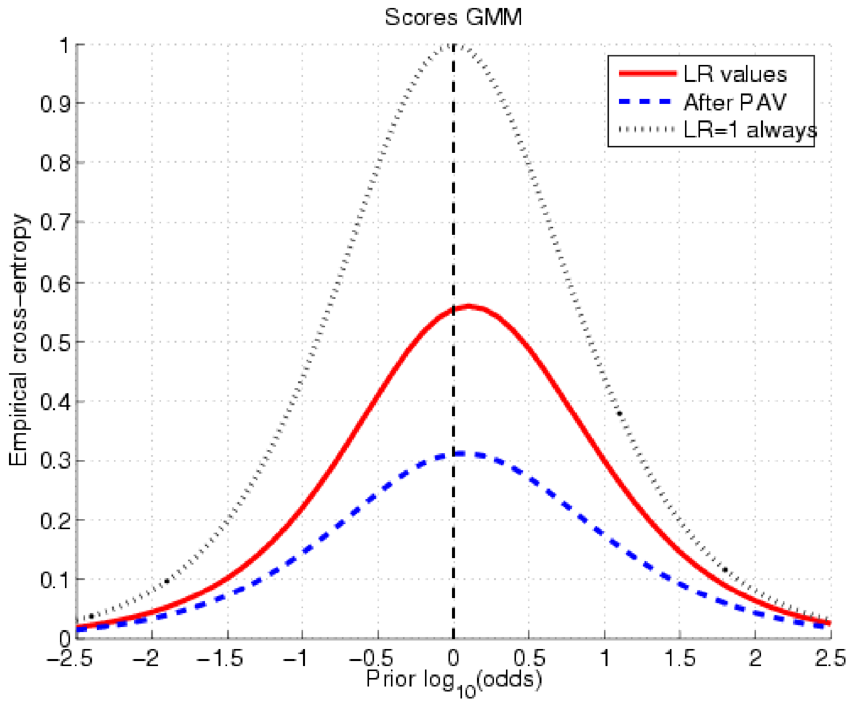
ECE plots: LR performance

- *ECE* of 3 LR sets are represented
  - LR values actually obtained (solid)
  - Always LR=1 (dotted)



ECE plots: LR performance

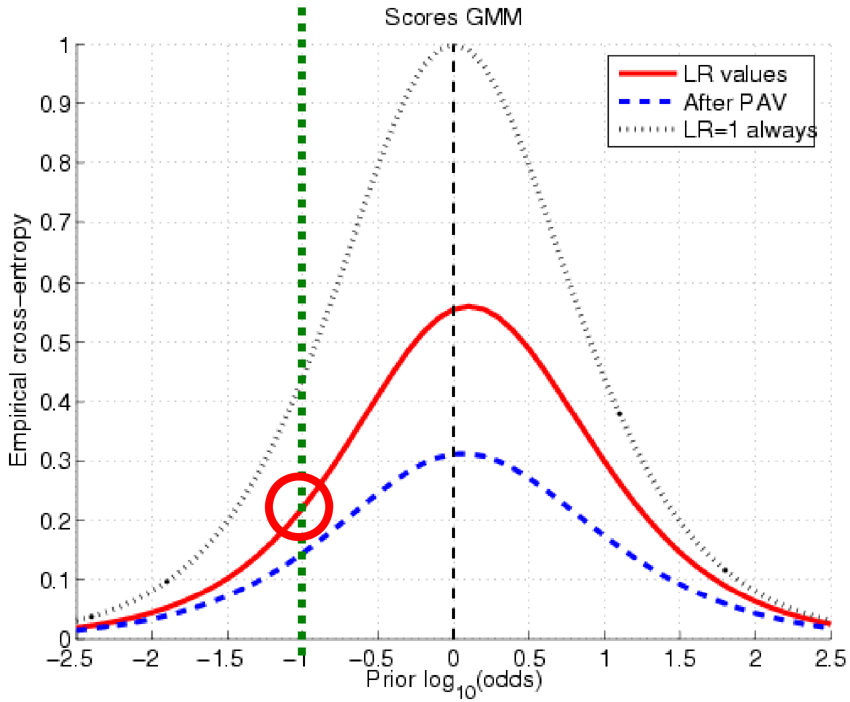
- *ECE* of 3 LR sets are represented
  - LR values actually obtained (solid)
  - Always LR=1 (dotted)
  - Calibrated LR values (dashed)
    - LR values after PAV
    - True answers are needed



ECE plots: LR performance

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

- *ECE* of 3 LR sets are represented
  - LR values actually obtained (solid)
  - Always LR=1 (dotted)
  - Calibrated LR values (dashed)
    - LR values after PAV
    - True answers are needed

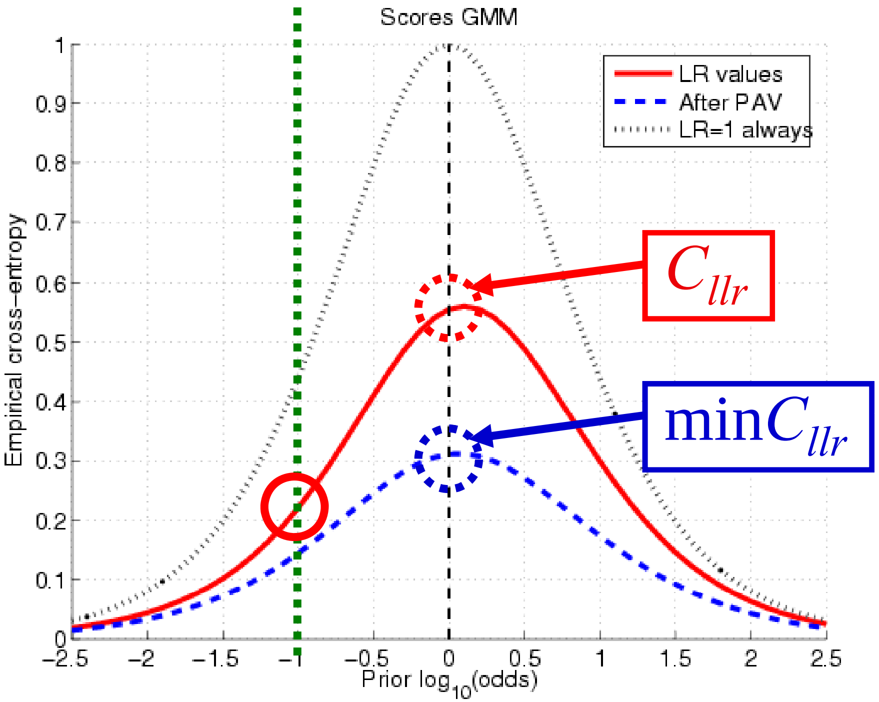


- Separation of roles
  - Forensic scientist: *ECE* computation for a wide range of priors
    - Because the scientist cannot set the prior…
  - Fact finder: prior establishment and measurement of *ECE*

ECE plots: LR performance

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

- *ECE* of 3 LR sets are represented
  - LR values actually obtained (solid)
  - Always LR=1 (dotted)
  - Calibrated LR values (dashed)
    - LR values after PAV
    - True answers are needed
  - C_{llr}: ECE at prior 0.5 [Brummer06]
    - $\min C_{llr}$: after PAV



- Separation of roles
  - Forensic scientist: *ECE* computation for a wide range of priors
    - Because the scientist cannot set the prior…
  - Fact finder: prior establishment and measurement of *ECE*

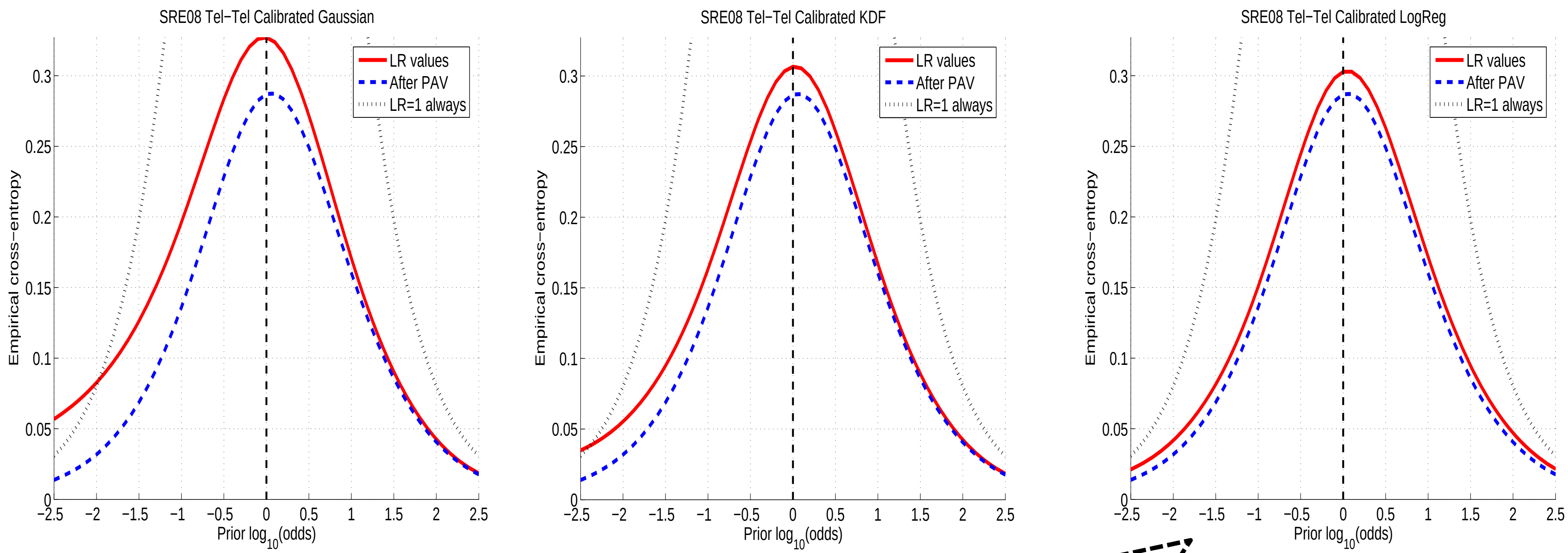
Case studies

Forensic Automatic Speaker Recognition

- Database and protocol: NIST Speaker Recognition Evaluation (SRE) 2008
  - Telephone-only subset
- Comparison of different LR computation methods [Ramos07,Gonzalez07]
  - Gaussian modelling
  - Kernel density functions (KDF)
  - Logistic regression

NIST SRE 2008, telephone-only data

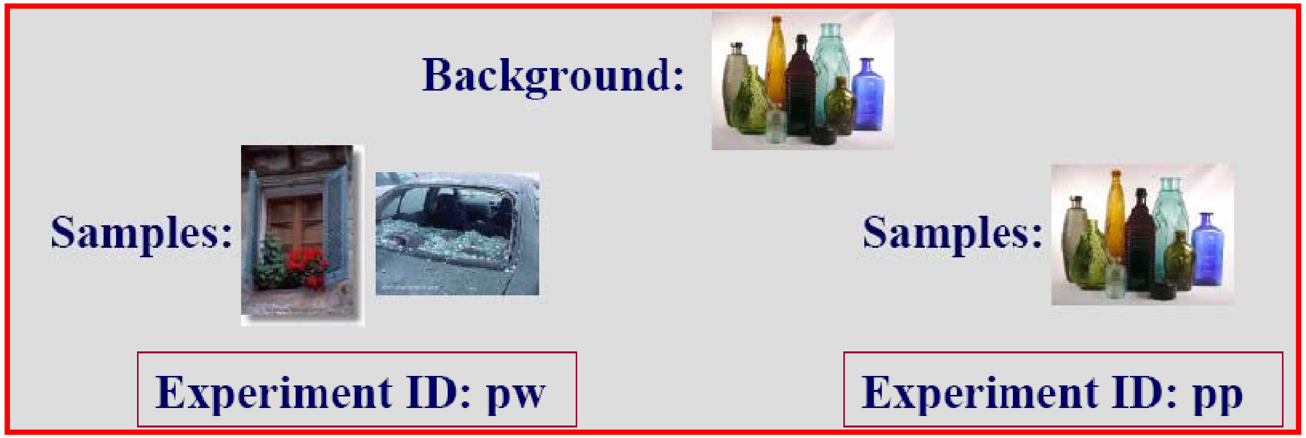
Gaussian | KDF | Logistic regression



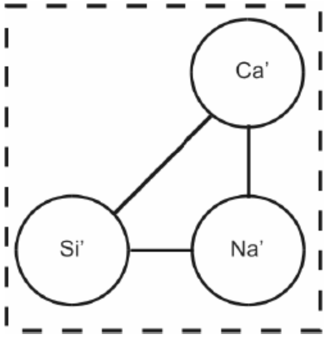
Logistic regression better performance (ECE) and better calibration (ECE – ECE after PAV)

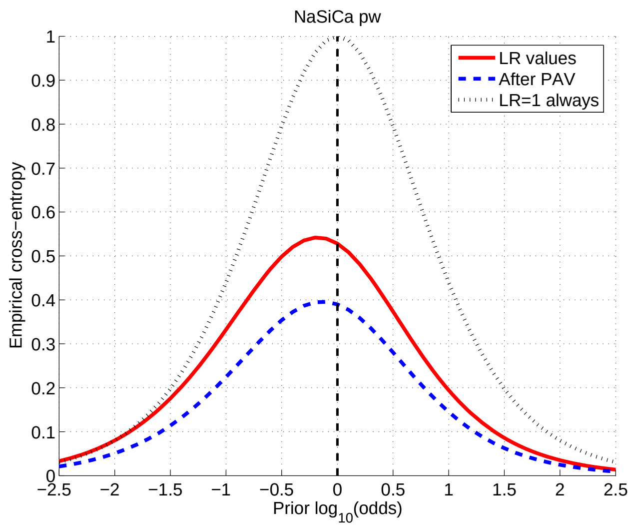
Forensic glass analysis

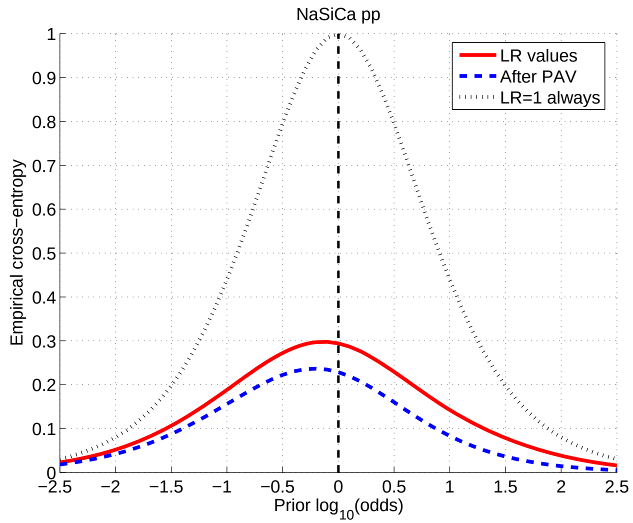
- Database collected by the Institute of Forensic Research (Krakow, Poland)
  - 7 variables (Log of Na, Si, Ca, Al, K, Fe and Mg normalized to O)
- Performance degradation due to population selection
  - [Zadora10]

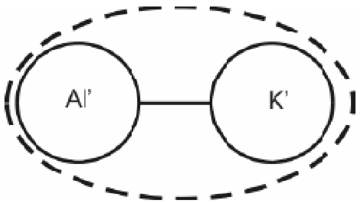


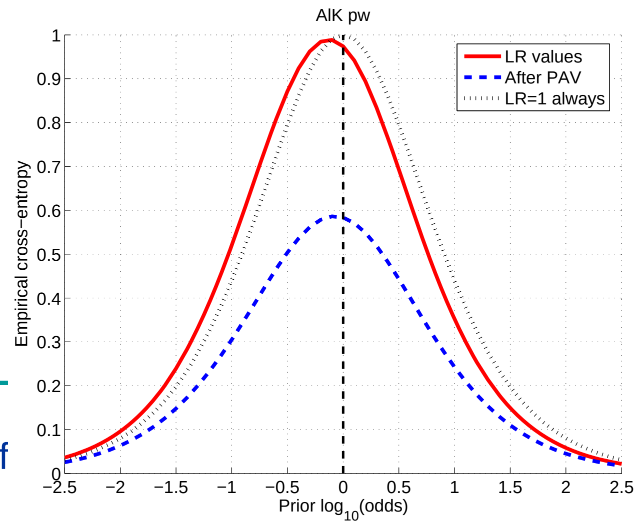
Mismatching background degrades performance

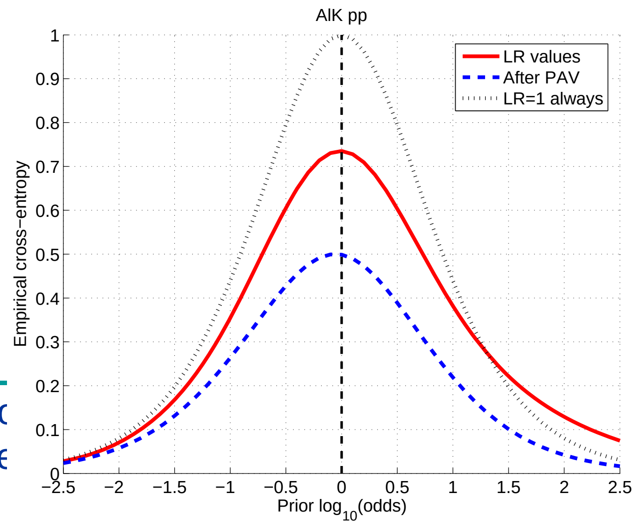












Conclusions

Conclusions

- Importance of Calibration
  - Improves performance of LR values
  - “Reliable” probabilistic interpretation of the LR [deGroot82]
- Measuring calibration: Empirical Cross-Entropy / C_{llr}
  - Information-theoretical interpretation
- ECE / C_{llr} can be applied to any LR-based forensic discipline
- Some challenges still remain…
  - Highly discriminating techniques such as DNA analysis
    - Empirical approach may not be robust or feasible
  - Behavior at extreme values of the prior odds
    - Small-sized experimental sets of LR values may not be robust

Software for Calibration and Assessment

- FoCal toolkit (Niko Brümmer)
  - Tools for assessment
    - C_{llr}
    - Other useful representations such as APE plots [Brummer06]
  - Tools for calibration
  - http://sites.google.com/site/nikobrummer/focal
- Software for drawing ECE plots (Daniel Ramos)
  - http://arantxa.ii.uam.es/~dramos/software.html

References

- [deGroot82] M. H. deGroot and S. E. Fienberg, 1982. "The comparison and evaluation of forecasters." The Statistician, vol. 32, pp. 12–22.
- [Dawid07] A. P. Dawid, 2007. "The geometry of proper scoring rules." Annals of the Institute of Statistical Mathematics 59, 77–93.
- [Gneiting07] T. Gneiting and A. E. Raftery, 2007. "Strictly proper scoring rules, prediction, and estimation." Journal of the American Statistical Association 102, 359-378.
- [Ramos07] D. Ramos, 2007. "Forensic evidence evaluation using automatic speaker recognition systems". Ph.D. Thesis. Universidad Autonoma de Madrid (available at http://atvs.ii.uam.es).
- [Gonzalez07] 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 Trans on Audio, Speech and Language Processing*, Vol. 15, n. 7, pp. 2104-2115.
- [Brummer06] N. Brümmer and J. du Preez, 2006. "Application independent evaluation of speaker detection." Computer Speech and Language, vol. 20, no. 2-3, pp. 230–275.
- [vanLeeuwen07] D. van Leeuwen and N. Brümmer, 2007. "An Introduction to Application-Independent Evaluation of Speaker Recognition Systems." Speaker Classification by Roland Müller (Ed). Lecture Notes in Artificial Intelligence 4343, Springer, Heidelberg.
- [Zadora10] G. Zadora and D. Ramos, "Evaluation of glass samples for forensic purposes - An application of likelihood ratios and an information-theoretical approach.", *Chemometrics and Intelligent Laboratory Systems*, vol. 102, n. 2, pp. 63-83, 2010.

On the Calibration of Likelihood Ratios

Daniel Ramos
ATVS – Biometric Recognition Group
Universidad Autonoma de Madrid
daniel.ramos@uam.es
http://atvs.ii.uam.es

WIC-BBfor2 Midwinter Meeting