# Assessing and comparing evidence evaluation methods using information theory



**Daniel Ramos**, Joaquin Gonzalez-Rodriguez

ATVS – Biometric Recognition Group

Universidad Autonoma de Madrid

http://atvs.ii.uam.es


Grzegorz Zadora

Institute of Forensic Research, Krakow, Poland


Colin G. G. Aitken

School of Mathematics, University of Edinburgh

# Outline

- LR framework for evidence evaluation
- Assessing LR accuracy
  - Strictly proper scoring rules
  - Information-theoretical framework
    - Empirical Cross-Entropy (ECE)
- Calibration and refinement
  - The PAV algorithm
- Accuracy representation: ECE plots
- Case studies
  - Glass analysis
  - Automatic speaker recognition
- Conclusions

# Evidence evaluation: Likelihood Ratio (LR) framework

# LR: support to a previous opinion

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

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


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

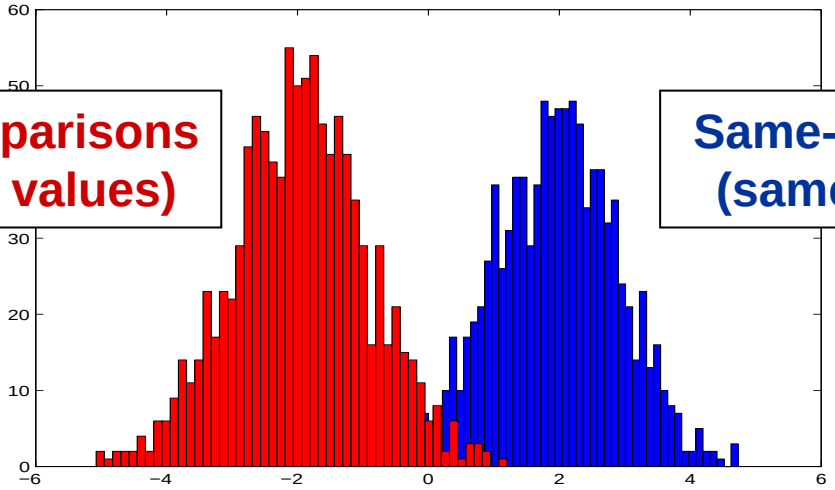

ATVS

UAM UNIVERSIDAD AUTONOMA DE MADRID

# Assessing the accuracy of LR values: Information-theoretical framework

# Empirically measuring accuracy

- Experimental test
  - Database of data with known sources
    - E.g., fingerprint database with known identities
  - Generate same-source comparisons ($\theta_p$ is true)
  - Generate different-source comparisons ($\theta_d$ is true)

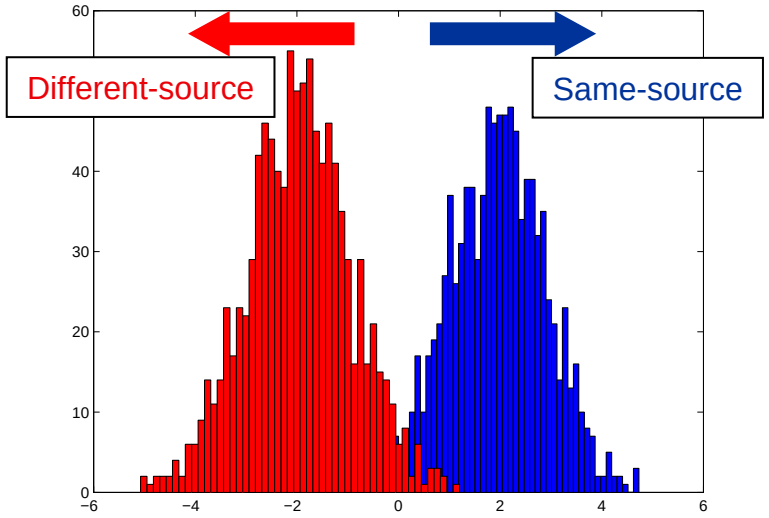



**Different-source comparisons (different-source LR values)**

**Same-source comparisons (same-source LR values)**

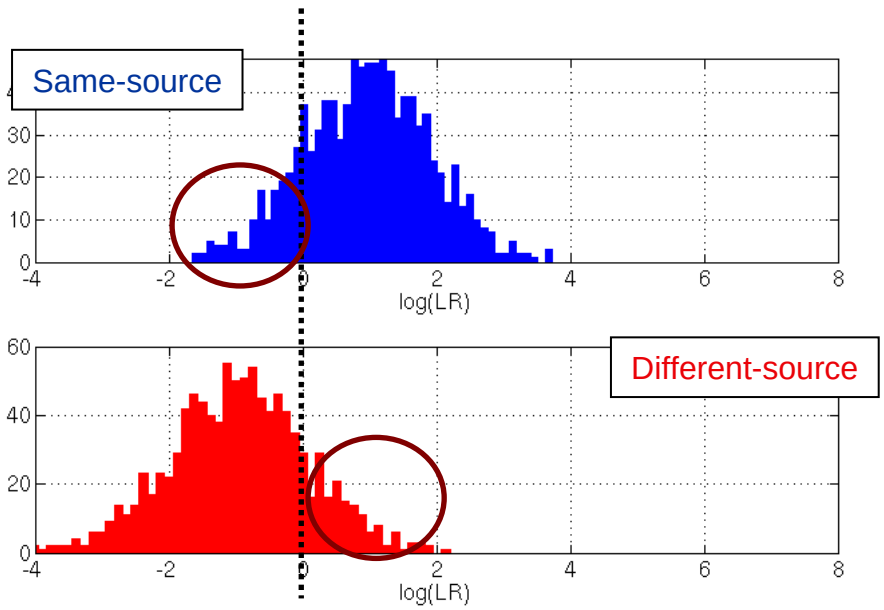
# Discriminating power

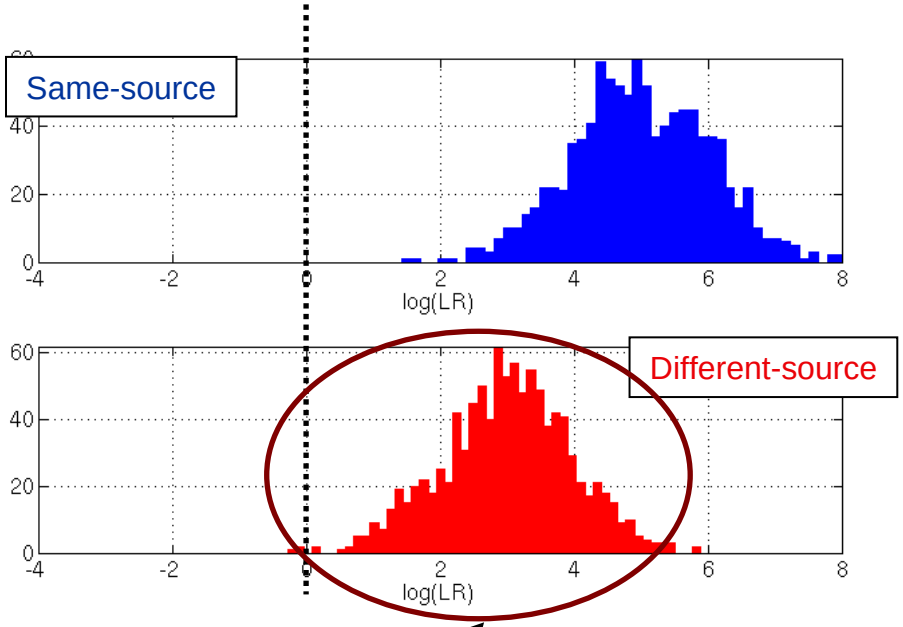
- *Discriminating* objects in the light of the evidence
  - Discriminating power (or simply discrimination) can be defined as the separation between
    - LR values for which $\theta_p$ is true
      - Samples come from the same source
    - LR values for which $\theta_d$ is true
      - Samples come from different sources
  - Good discriminating power means:
    - Higher log-LR values for same-source comparisons
    - Lower log-LR values for different-source comparisons
  - Measured by ROC and DET plots, etc.



# Discrimination is not enough for LR

- Example: two techniques with the same discrimination




- Not a discrimination problem
  - The same in both of them
- *Calibration* problem [deGroot82]

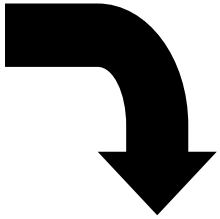
Strong support to the wrong hypothesis! Will lead to errors

# Defining LR accuracy: roadmap

Classical assessment of forecasts (posterior): Strictly Proper Scoring Rules (SPSR) [deGroot82,Dawid07]

# Defining LR accuracy: roadmap

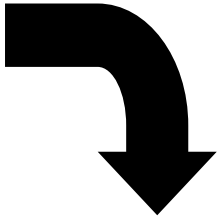
Classical assessment of forecasts (posterior): Strictly Proper Scoring Rules (SPSR) [deGroot82,Dawid07]



Accuracy: average of a SPSR over forecasts [deGroot82]

# Defining LR accuracy: roadmap

Classical assessment of forecasts (posterior): Strictly Proper Scoring Rules (SPSR) [deGroot82,Dawid07]

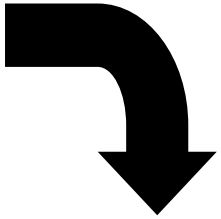


Accuracy: average of a SPSR over forecasts [deGroot82]

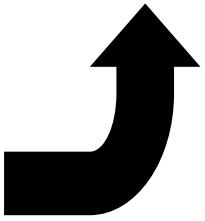
Proposed information-theoretical Framework for assessing LR accuracy [Ramos07]

# Defining LR accuracy: roadmap

Classical assessment of forecasts (posterior): Strictly Proper Scoring Rules (SPSR) [deGroot82,Dawid07]

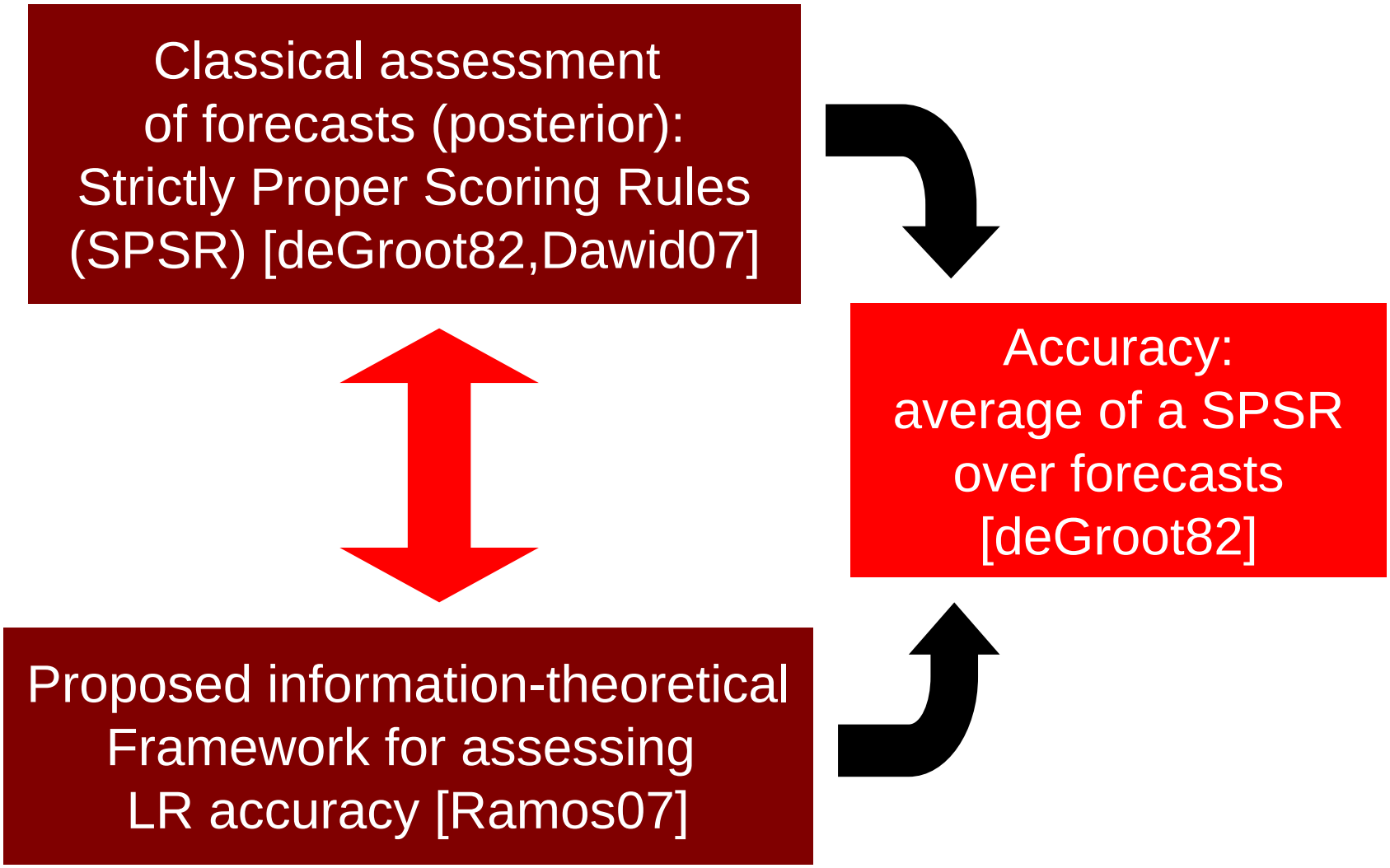


Accuracy: average of a SPSR over forecasts [deGroot82]



Proposed information-theoretical Framework for assessing LR accuracy [Ramos07]

# Defining LR accuracy: roadmap



# Accuracy of the LR

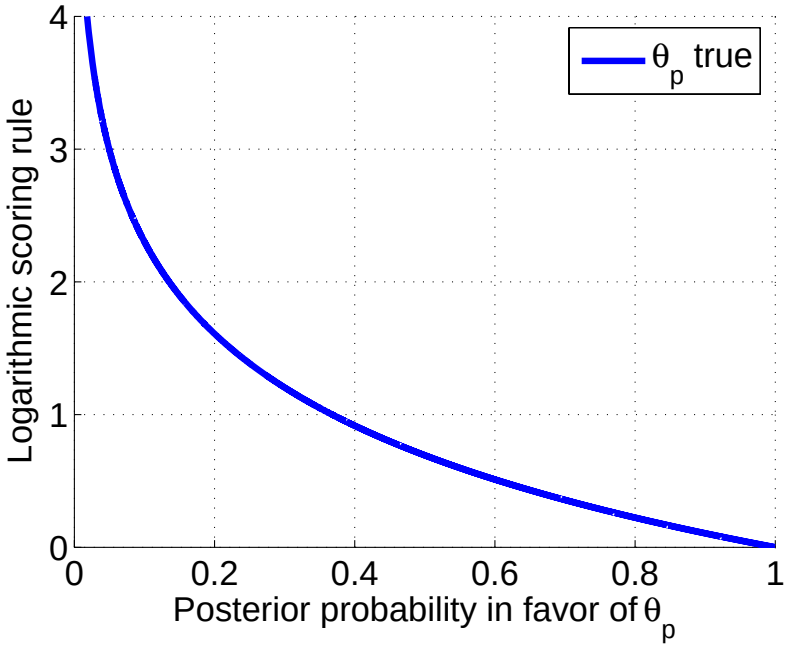
- Accuracy of a probabilistic opinion (*forecast*)
  - Classically measured by Strictly Proper Scoring Rules (SPSR) [deGroot82, Dawid07]
- A scoring rule assigns a penalty to each probabilistic opinion
  - Depending on whether each hypothesis is true or not
  - It is needed to know the true answer to evaluate
- Forecasts can be expressed by posterior probabilities

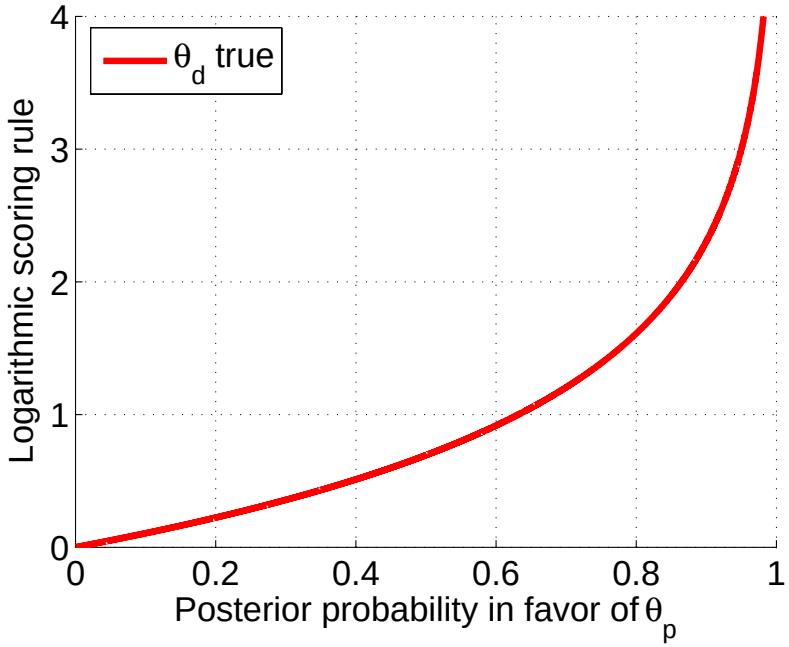
$$P\left(\theta_p \middle| E, I\right)$$

# Logarithmic scoring rule

- Assigns: $-\log_2 P(\theta_p | e_j)$ $\theta_p$ is true

  $-\log_2 P(\theta_d | e_j)$ $\theta_d$ is true

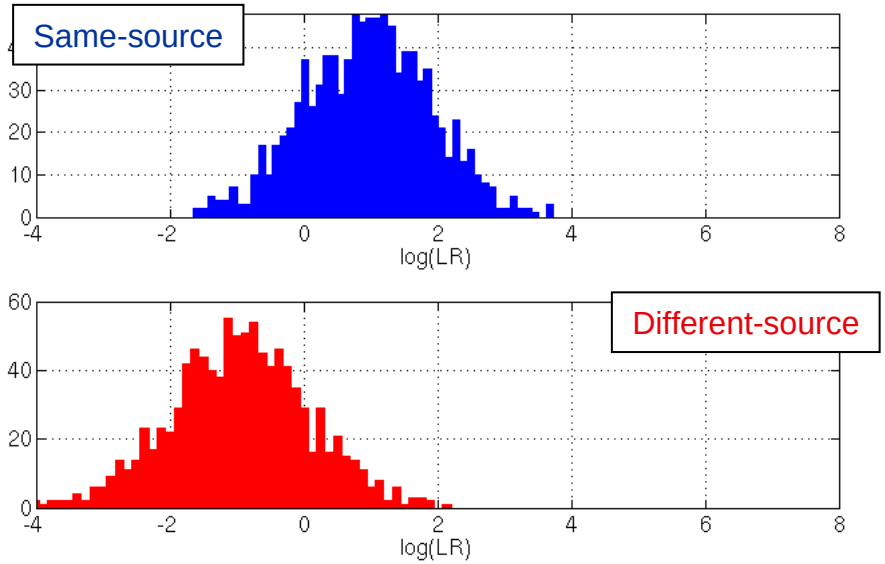



  - Maximum penalty: infinite for categorical and erroneous opinions
    - Such as erroneous identifications/exclusions

# Accuracy: empirical computation

- Accuracy of a set of opinions from comparisons
  - Average of a strictly proper scoring rule over comparisons




$$LS = -\frac{1}{N_p} \sum_{j \in \text{same-source}} \log_2 P\left(\theta_p \middle| e_j\right)$$

$$-\frac{1}{N_d} \sum_{j \in \text{diff-source}} \log_2 P\left(\theta_d \middle| e_j\right)$$

# Empirical Cross-Entropy (*ECE*)

- Recently proposed in [Ramos07]
- Prior-weighted average of the logarithmic SPSR

$$LS = -\frac{1}{N_p} \sum_{j \in \text{same-source}} \log_2 P\left(\theta_p \middle| e_j\right) \qquad ECE = -\frac{P\left(\theta_p\right)}{N_p} \sum_{j \in \text{same-source}} \log_2 P\left(\theta_p \middle| e_j\right)$$

$$-\frac{1}{N_d} \sum_{j \in \text{diff-source}} \log_2 P\left(\theta_d \middle| e_j\right) \qquad -\frac{P\left(\theta_d\right)}{N_d} \sum_{j \in \text{diff-source}} \log_2 P\left(\theta_d \middle| e_j\right)$$

  - "Average information needed to obtain certainty"
    - The higher the cross-entropy, more information needed to be certain
    - Using the LR values computed by the scientist

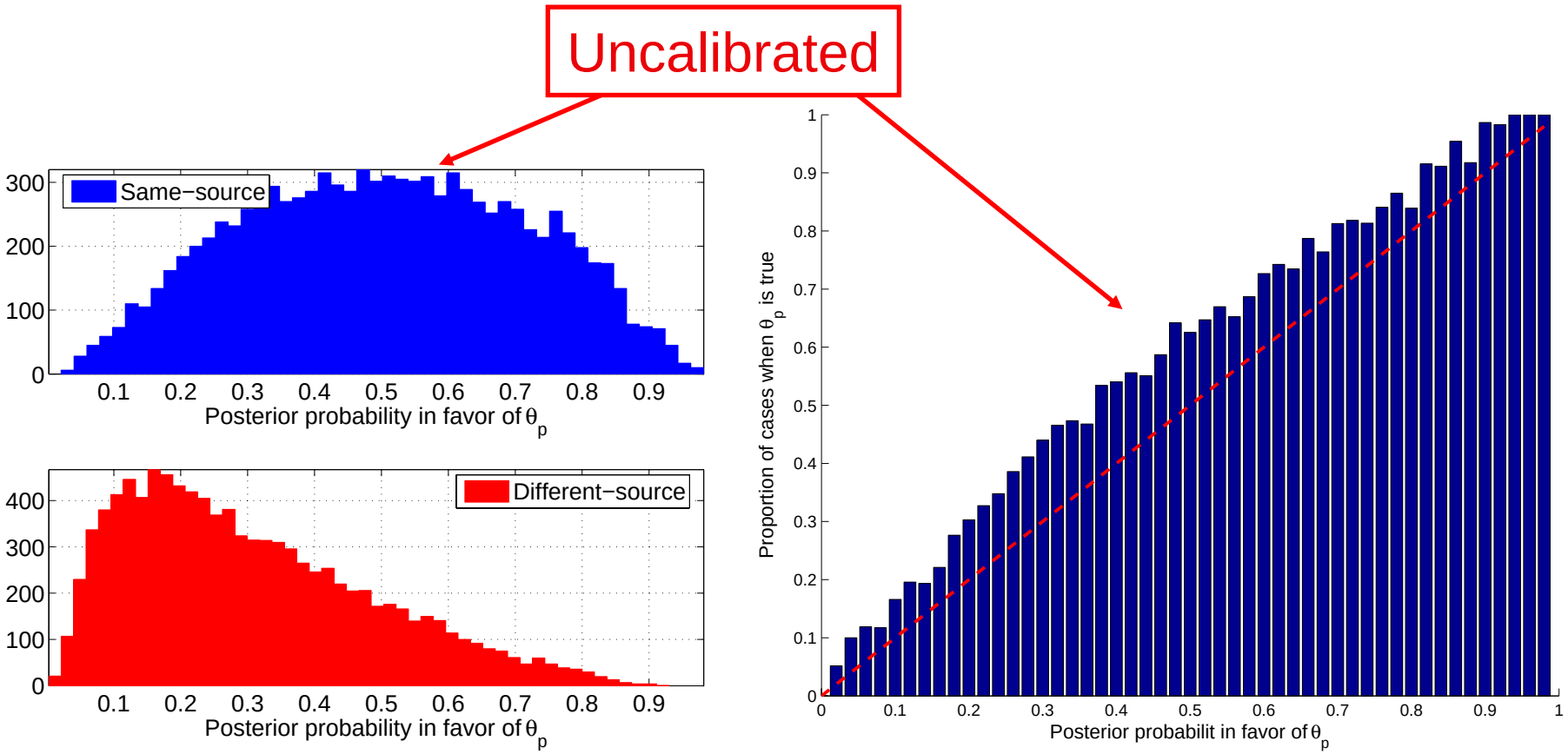
# Calibration of LR values

# Calibration

- Given a set of posterior probabilities about hypothesis $\theta_p$, calibration means
  - Posterior probabilities of $\theta_p$ approximate actual proportions of occurrence of $\theta_p$
- Calibrated probabilities have been dubbed *reliable* [deGroot82]
  - Probabilities converge to the actual proportions of occurrence
- Calibration improves accuracy
  - Because average of SPSR are decomposed [deGroot82]
  - A refinement component
    - Measure of discrimination [Brummer06]
  - A calibration component

# Calibration

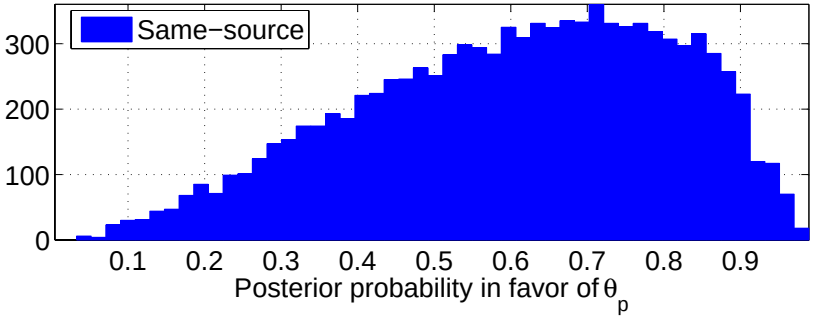
- Example: experimental set of posterior probabilities
  - Obtained from the LR values computed by the forensic scientist
  - Choosing a value for the prior probability ($P(\theta_p)=0.5$)

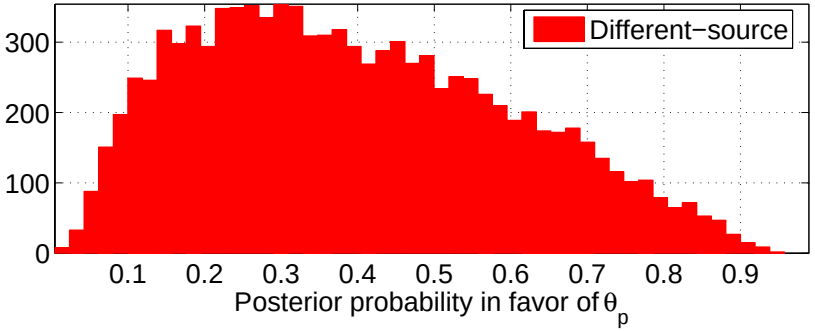


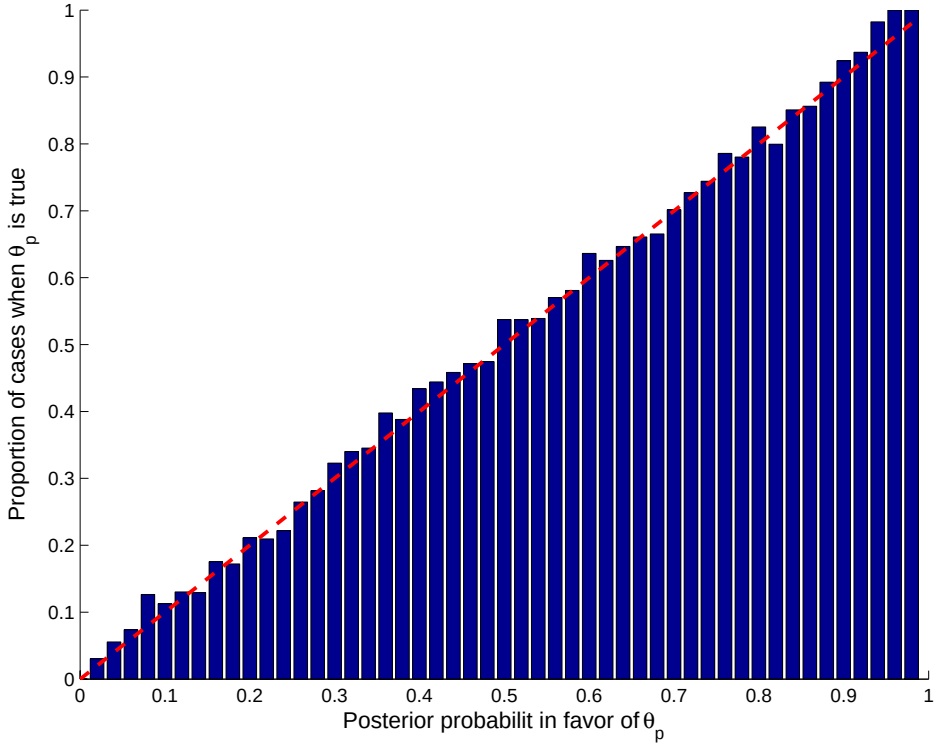
# Calibration

- Example: other set of posterior probabilities presenting the same discrimination as before
  - $P(\theta_p)$=0.5

Calibrated



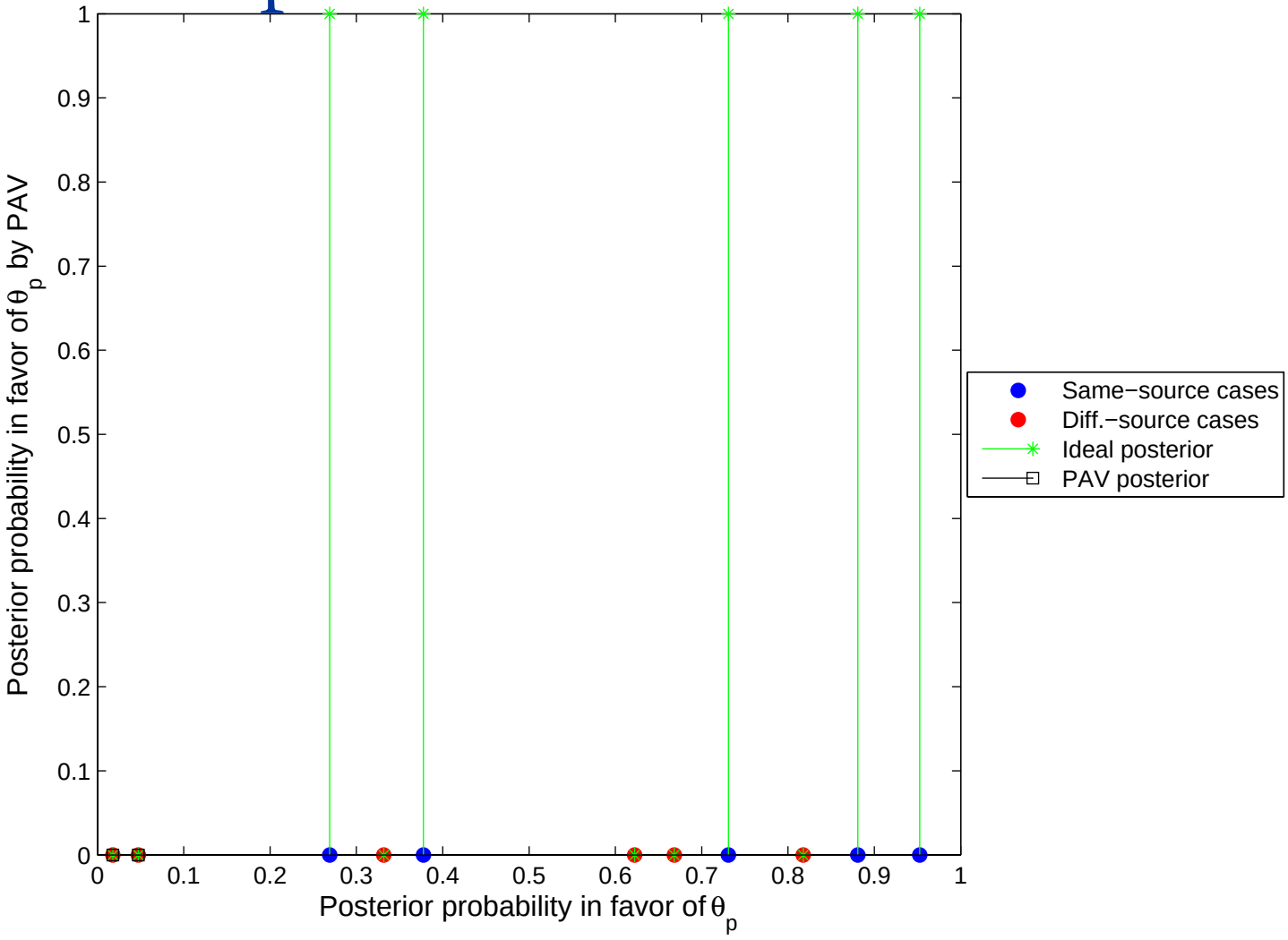



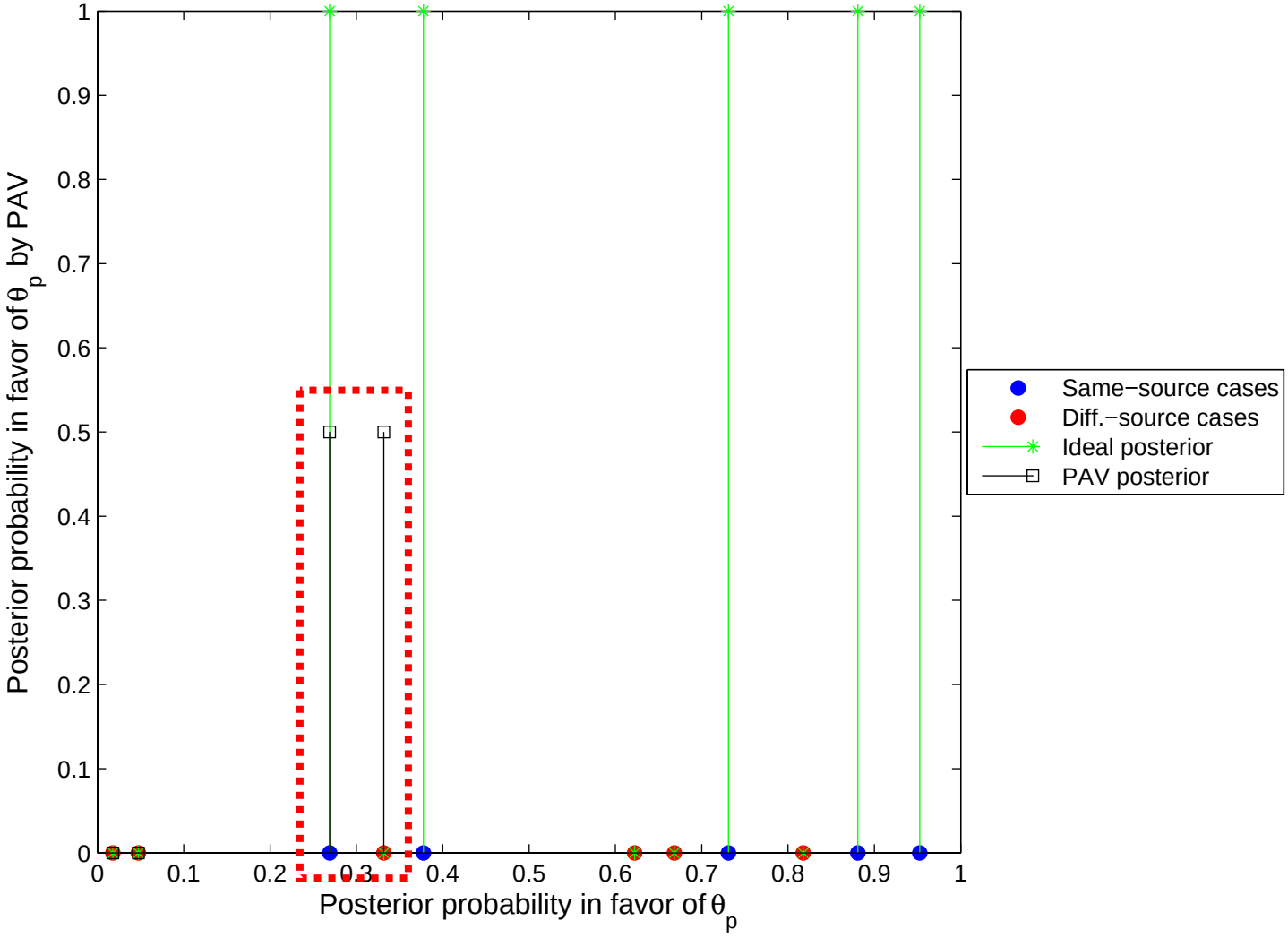



# Obtaining calibrated probabilities

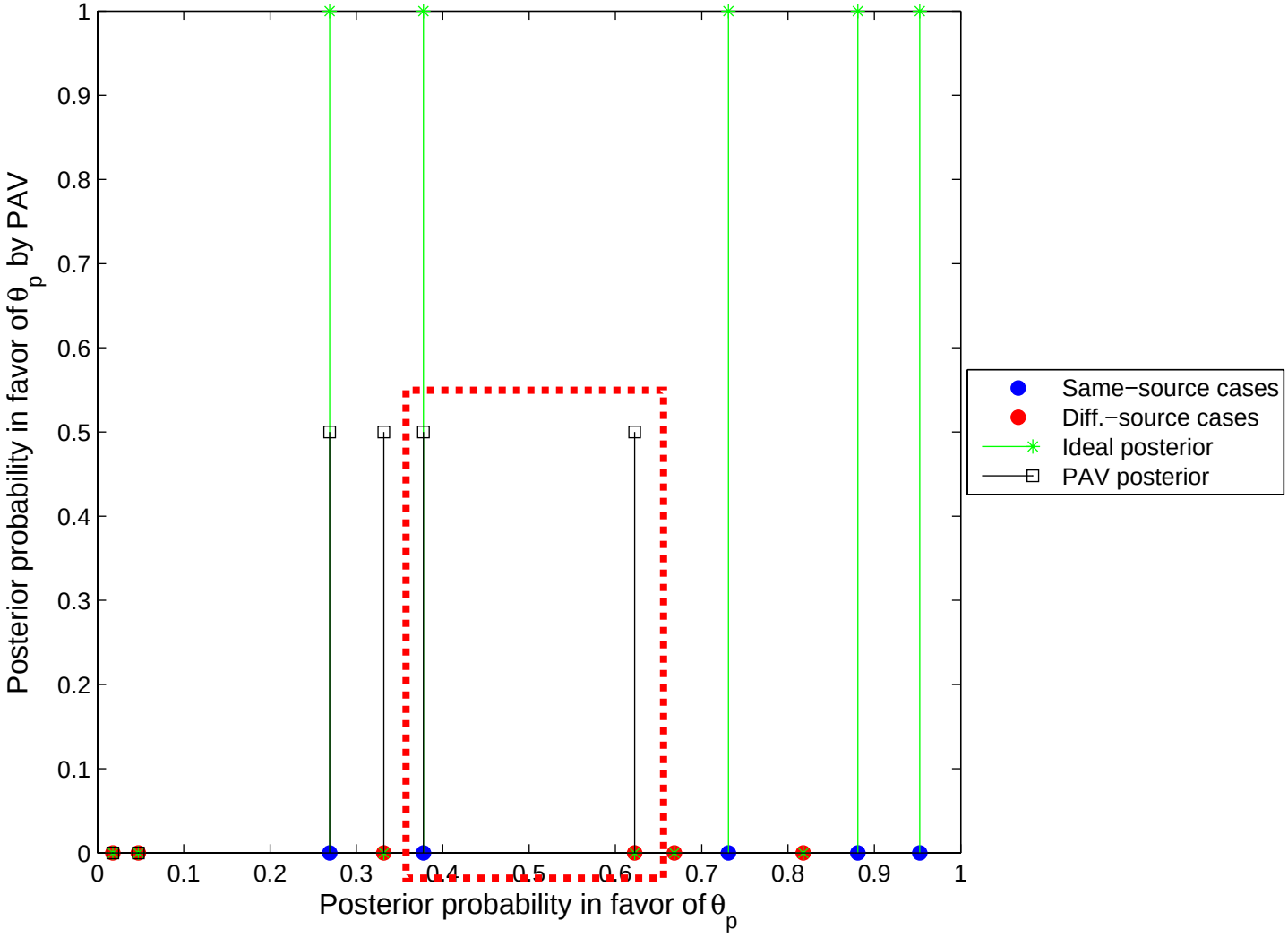
- Computing proportions of cases imply binning cases
  - How many bins? What bin size?
- Solution: Pool Adjacent Violators Algorithm (PAV) [Brummer06]
  - Computation of proportions over the given set of probabilities
  - Monotonically rising (isotonic regression)
    - Preserves discrimination
    - Only calibration is improved
- Yields a calibrating transformation
  - Input: probabilities from the forensic scientist
  - Output: calibrated probabilities preserving the discrimination
- True answers to the hypotheses are needed!

# PAV: example

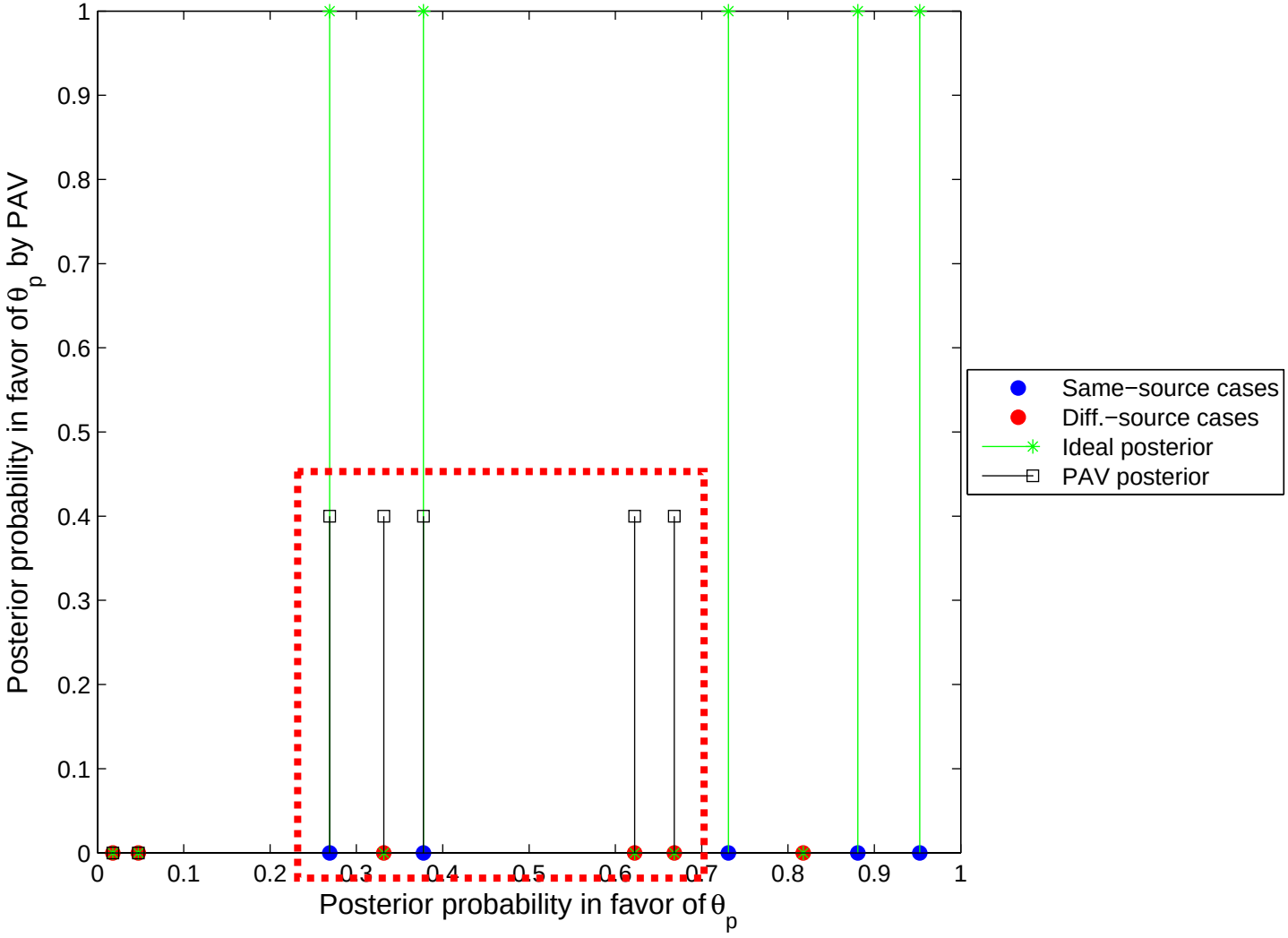

Posterior probability in favor of $\theta_p$ by PAV
Posterior probability in favor of $\theta_p$
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Same−source cases
Diff.−source cases
Ideal posterior
PAV posterior



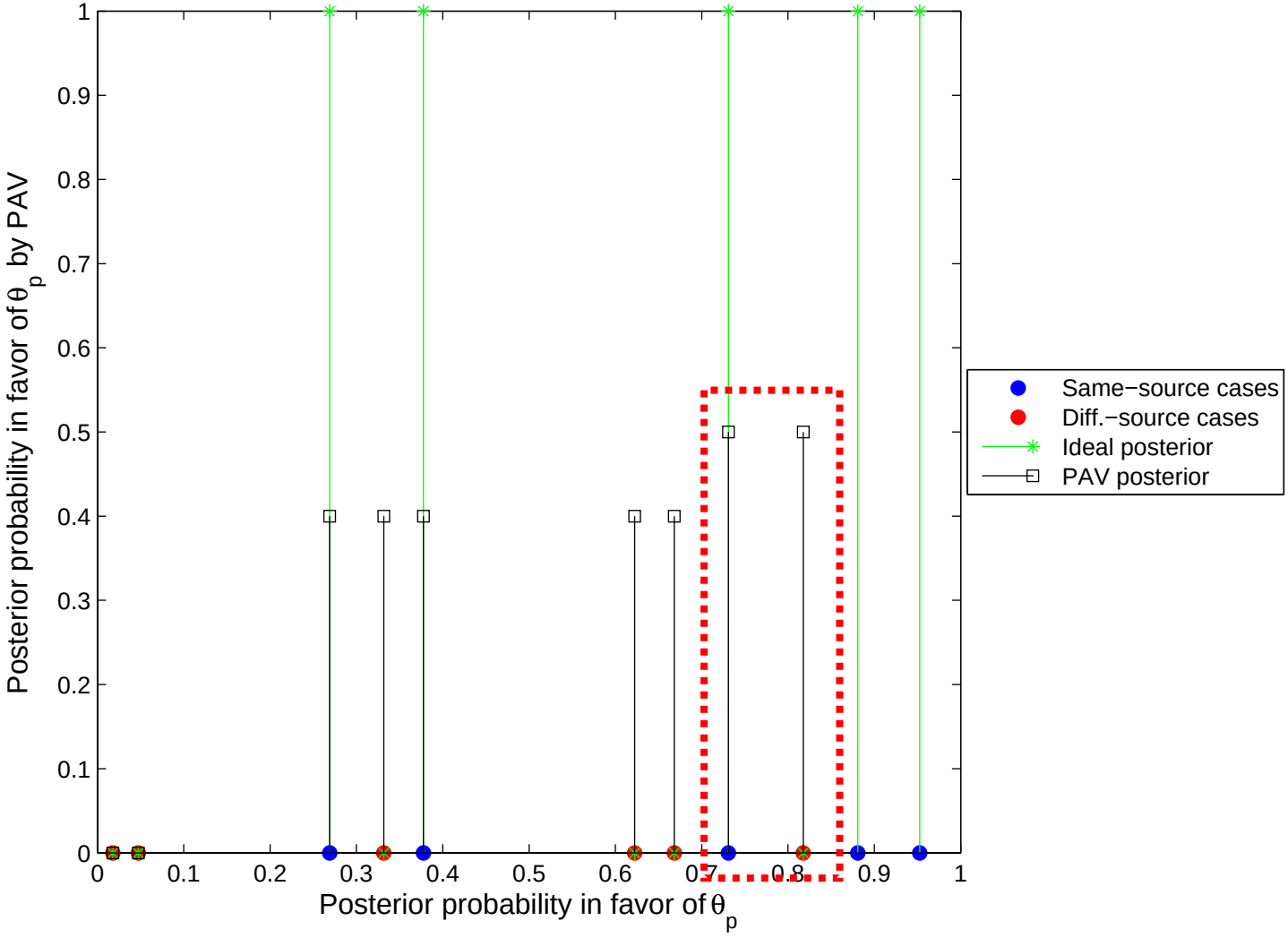

Decreasing (violators): pool them together and average output probabilities



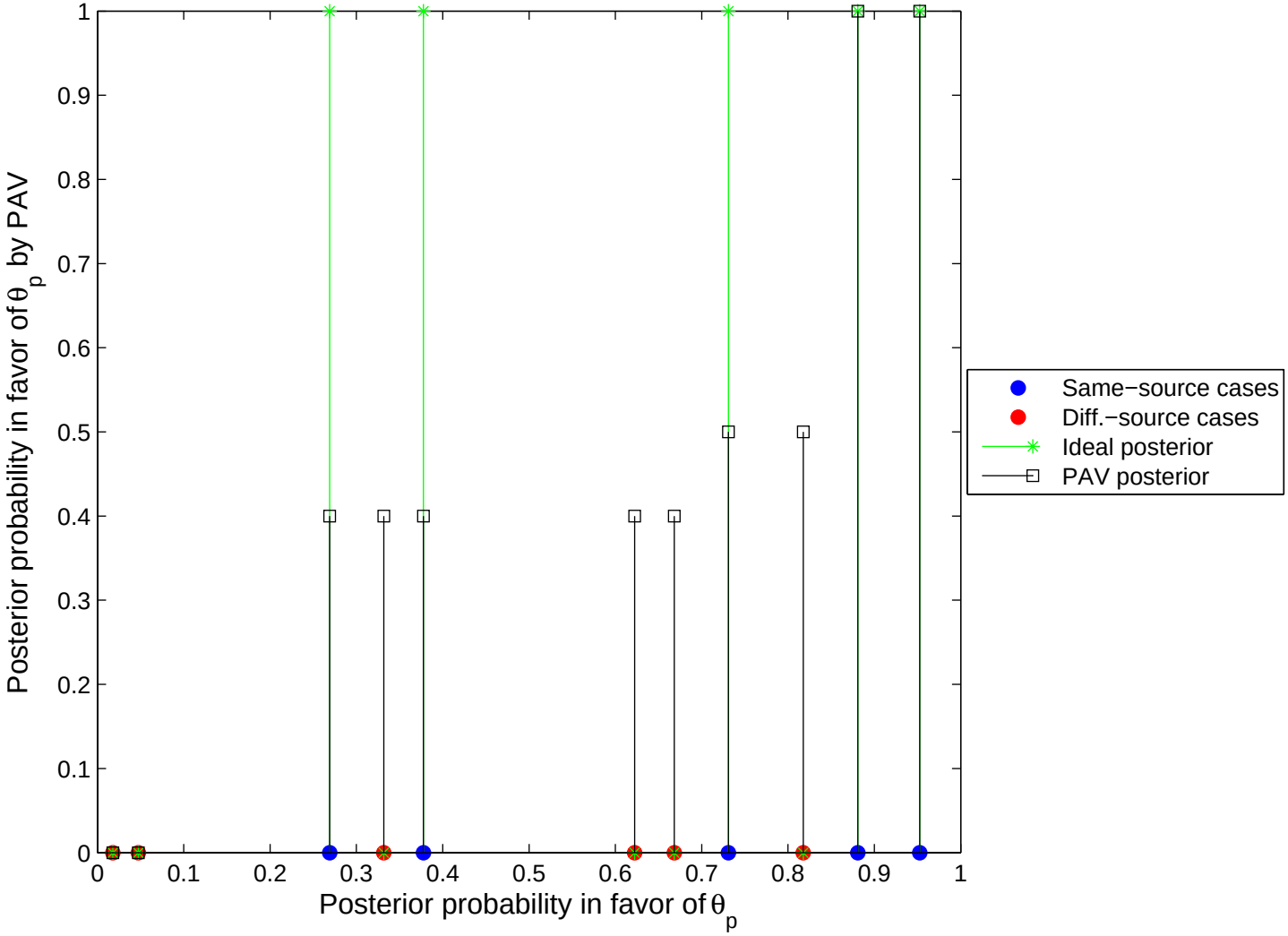

Decreasing (violators): pool them together and average output probabilities

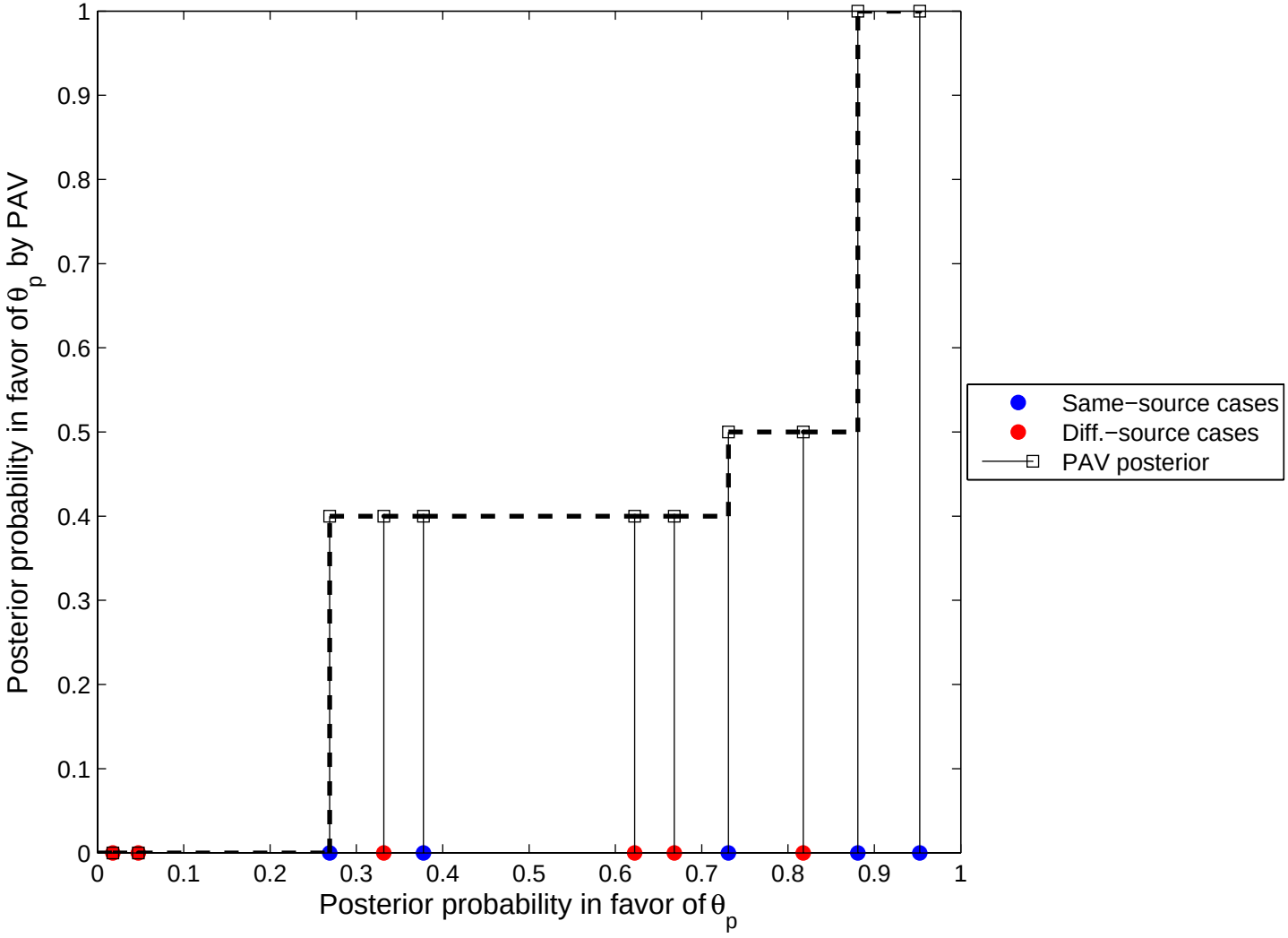



Decreasing (violators): pool them together and average output probabilities




Decreasing (violators): pool them together and average output probabilities


Posterior probability in favor of $\theta_p$ by PAV
Posterior probability in favor of $\theta_p$
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


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 *ECE*
  - (Prior-weighted) average value of SPSR
  - Calibrated posteriors need less information to obtain certainty
    - Accuracy is improved

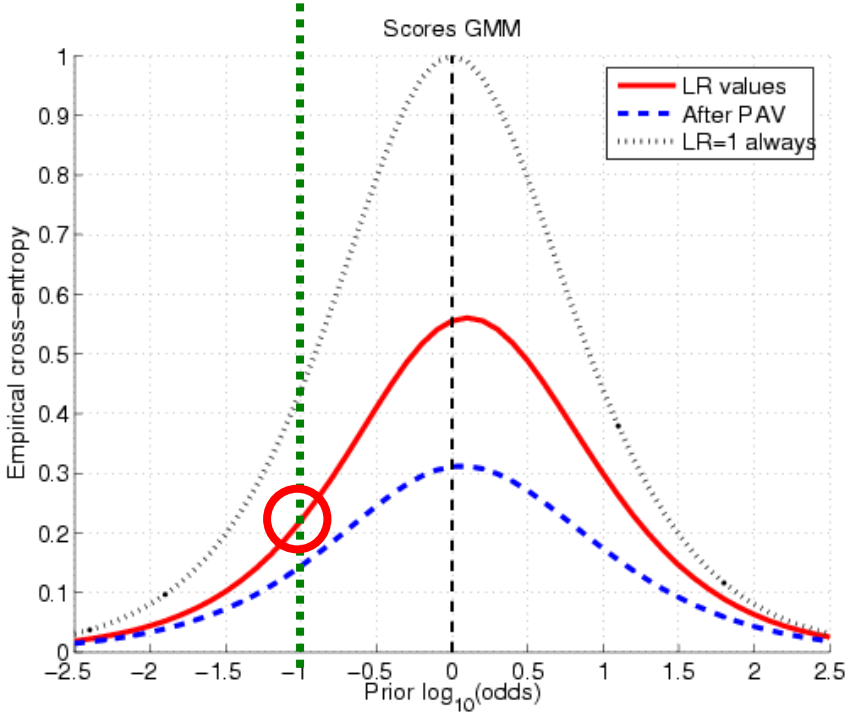
$$ECE = -\frac{P(\theta_p)}{N_p} \sum_{j \in \text{same-source}} \log_2 P(\theta_p | e_j) - \frac{P(\theta_d)}{N_d} \sum_{j \in \text{diff-source}} \log_2 P(\theta_d | e_j)$$

- However, PAV and *ECE* both need the prior probability…
  - The forensic scientist cannot compute its value
- Solution: the *ECE* plot
  - Computing *ECE* and PAV transformation for a wide range of priors

# *ECE* plots: LR accuracy

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

- 3 sets of LR values are represented
  - LR values from the forensic scientist (solid)
  - Always LR=1 (dotted)
  - Calibrated LR values (dashed)
    - After applying PAV transformation
    - Best accuracy for the given discrimination
    - True answers are needed




- Separation of roles
  - Forensic scientist: *ECE* computation for a wide range of priors
  - Fact finder: prior establishment and measure of *ECE* in the plot

ATVS

UAM UNIVERSIDAD AUTONOMA DE MADRID

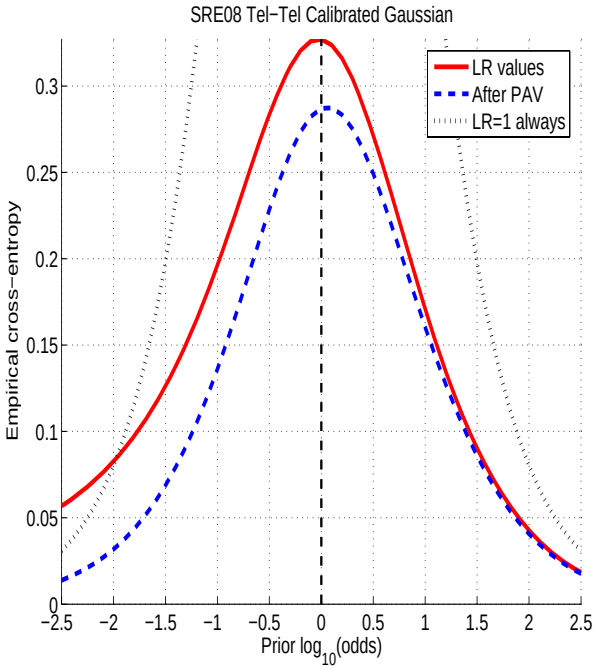
# Case studies

# Forensic Automatic Speaker Recognition

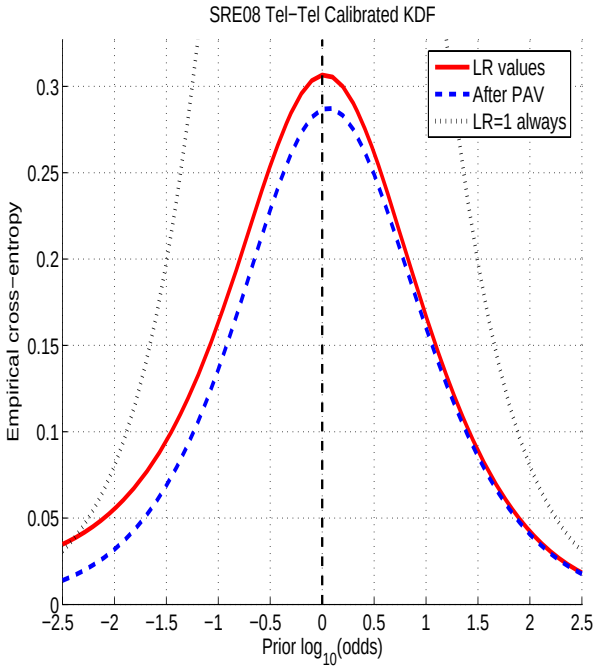
- Database and protocol: NIST Speaker Recognition Evaluation (SRE) 2008
  - Telephone-only subset
  - Speech from different channels, languages, environmental conditions
  - Strongly variable and mismatching conditions
    - Among recovered and control speech
    - Among background and testing databases
- Comparison of different LR computation methods [Ramos07]
  - Gaussian modelling
  - Kernel density functions (KDF)
  - Logistic regression

# NIST SRE 2008, telephone-only data

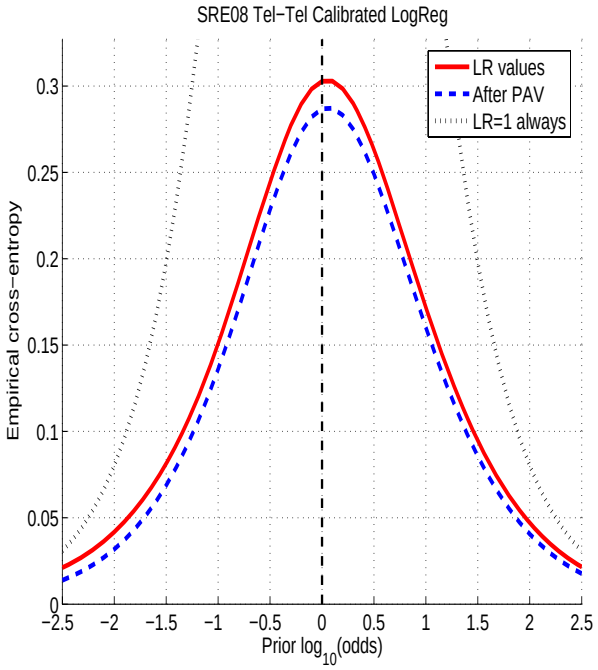
Gaussian




KDF




Logistic regression




- Logistic regression slightly better accuracy
  - Needs less information for obtaining certainty
- Discriminating power (blue curve) is almost unchanged

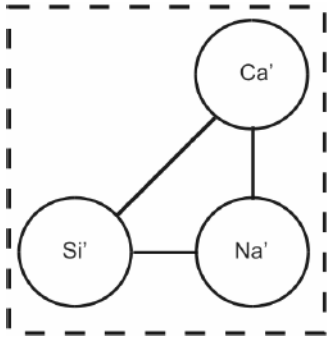
ATVS

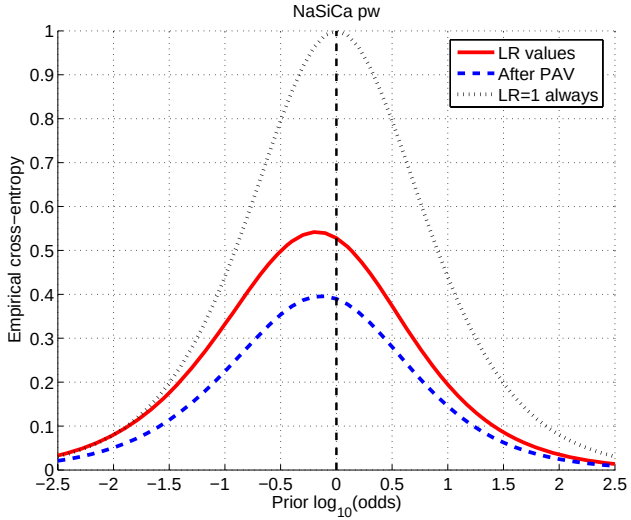
UAM UNIVERSIDAD AUTONOMA DE MADRID

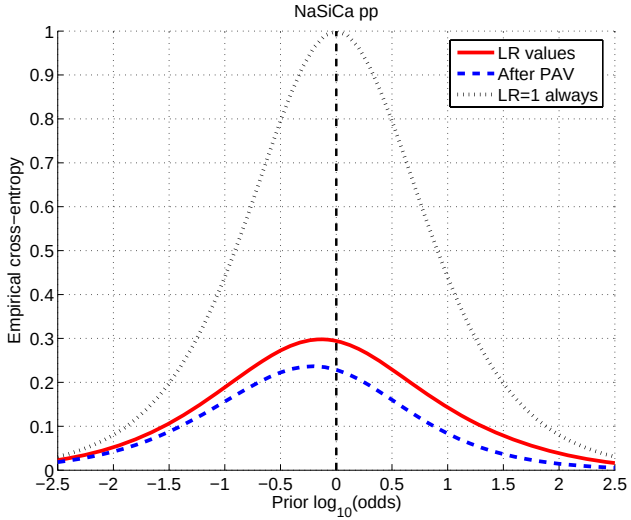
# Forensic glass analysis

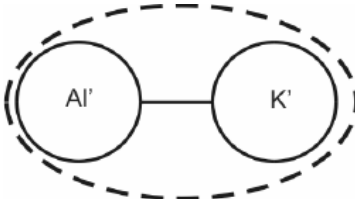
- Database collected by the Institute of Forensic Research (Krakow, Poland)
  - 7 variables (Log of Na, Si, Ca, Al, K, Fe and Mg normalized to O)
  - 175 objects from car and building windows (w data)
  - 57 objects from containers (p data)
- Exploring variability due to population selection
  - [Zadora08]

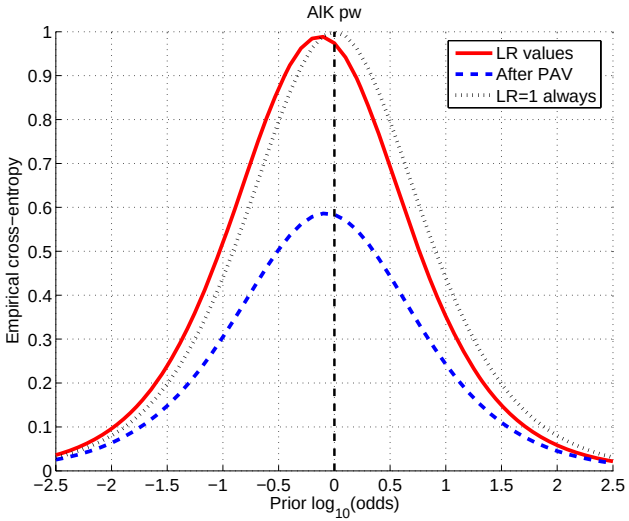
# Mismatching background degrades accuracy

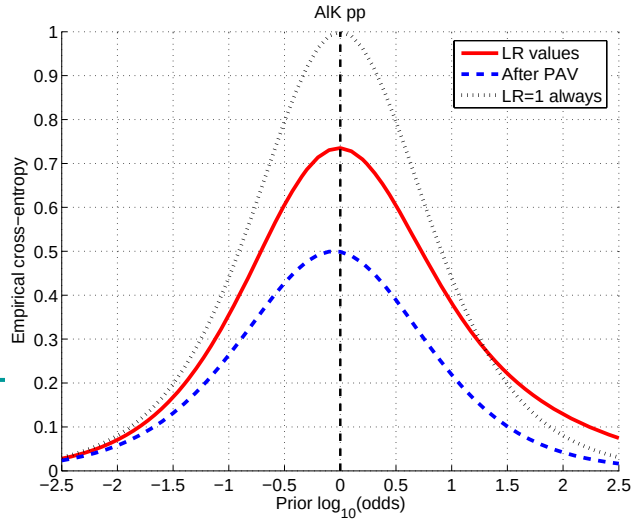



# Conclusions

# Conclusions

- Proposed framework for assessing LR accuracy
  - Empirical approach
    - Experimental database and comparison protocol
  - Information-theoretical interpretation
    - Average information for obtaining certainty using a set of LR values
    - Understandable and intuitive
- Importance of calibration
  - Calibration improves accuracy
    - Calibrated LR values need less information for obtaining certainty
- Methodology can be applied to any LR-based forensic discipline
    - Forensic automatic speaker recognition
    - Glass analysis

# References


- [deGroot82] M. H. deGroot and S. E. Fienberg, 1982. “The comparison and evaluation of forecasters.” The Statistician, vol. 32, pp. 12–22.
- [Dawid07] Dawid, A. P., 2007. “The geometry of proper scoring rules.” Annals of the Institute of Statistical Mathematics 59, 77–93.
- [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).
- [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.
- [Zadora08] G. Zadora et al., 2008. “The effects of variability of elemental composition of glass in the accuracy of the evidence evaluation process.” International Conference of Forensic Inference and Statistics (ICFIS 2008). Lausanne, Switzerland.