Accuracy degradation of LR-based evidence evaluation: an experimental study with glass evidence

Daniel Ramos
daniel.ramos@uam.es

ATVS – Biometric Recognition Group
http://atvs.ii.uam.es
Escuela Politecnica Superior
Universidad Autonoma de Madrid

Outline

- Accuracy of likelihood-ratio (LR) based evidence evaluation
  - Empirical cross-entropy (ECE)
- Detecting LR values which degrade accuracy
  - Hypothesis-dependent histograms
  - Contribution of “bad” LR values to ECE
- Examples with glass evidence
  - Finding the problems
  - Possible solutions
- Conclusions

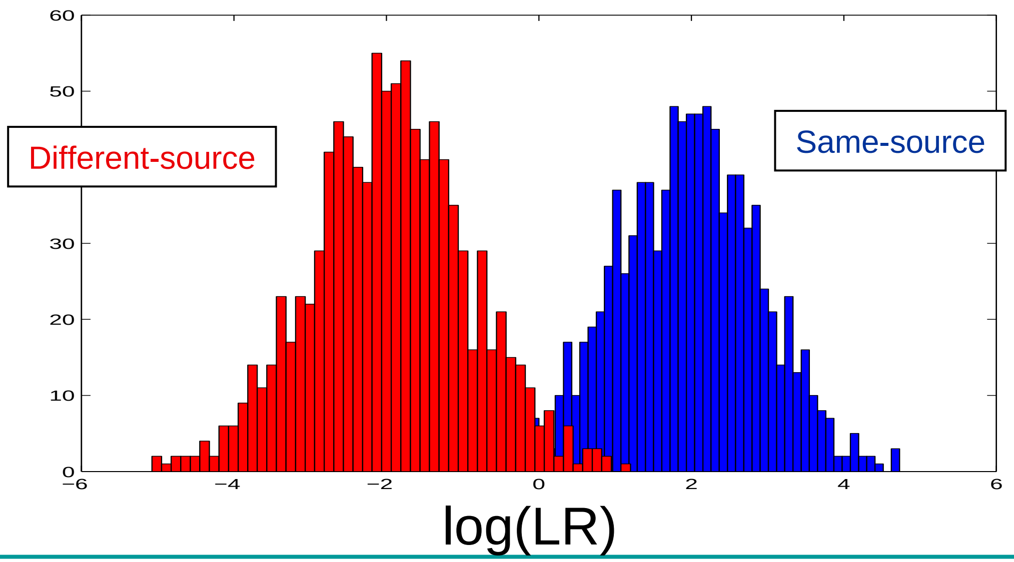
Accuracy of likelihood-ratio-based evidence evaluation

Accuracy of the LR

- The LR has a *meaning* by itself
  - *Degree of support* to the previous opinion
  - LR is the weight of the evidence E
- Inferred posterior probabilities must be accurate
- But what's accuracy?

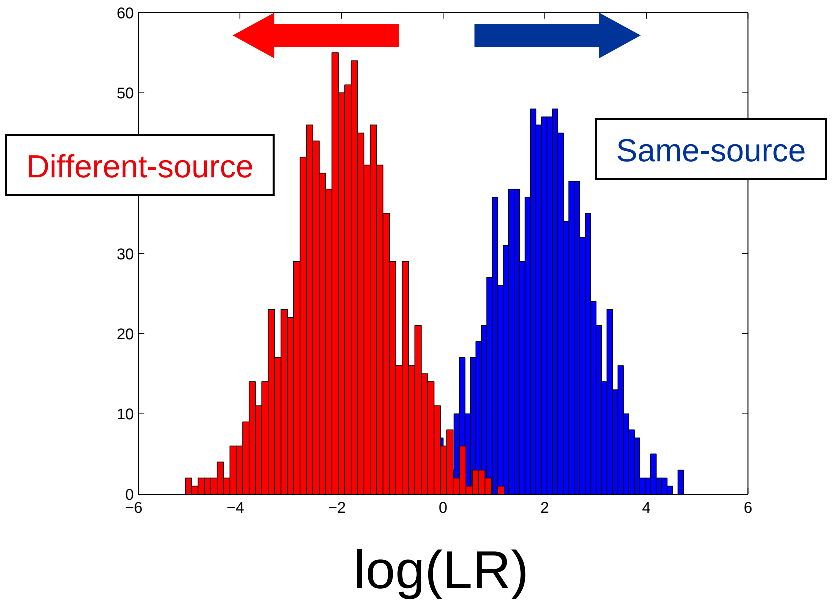
Empirically measuring accuracy

- Experimental test
  - Database of data with known sources
    - E.g., glass chemical profiles
    - The object where each profile has been measure is known
  - Generate same-source comparisons (θ_p is true)
  - Generate different-source comparisons (θ_d is true)



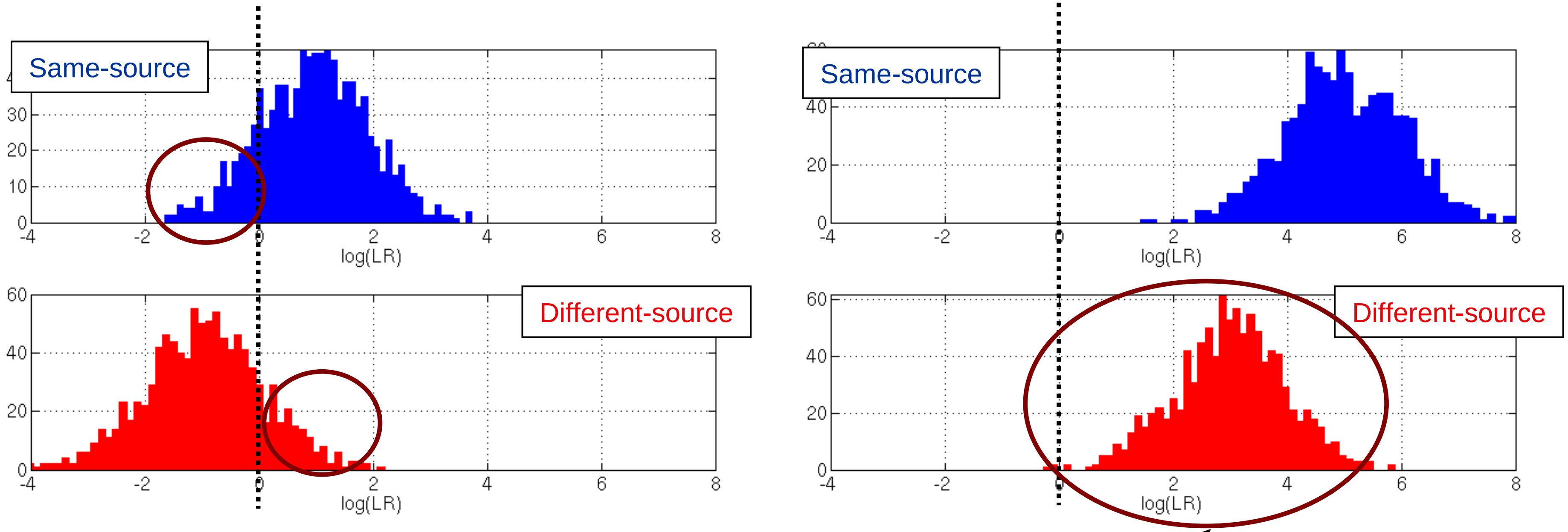
Discriminating power

- *Discriminating* objects in the light of the evidence
  - Discriminating power (or simply discrimination) can be difined as the separation between
    - LR values for which θ_p is true
      - Control and recovered samples come from the same source
    - LR values for which θ_d is true
      - Control and recovered 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



Discrimination is not enough

- Example: two techniques with the same discrimination

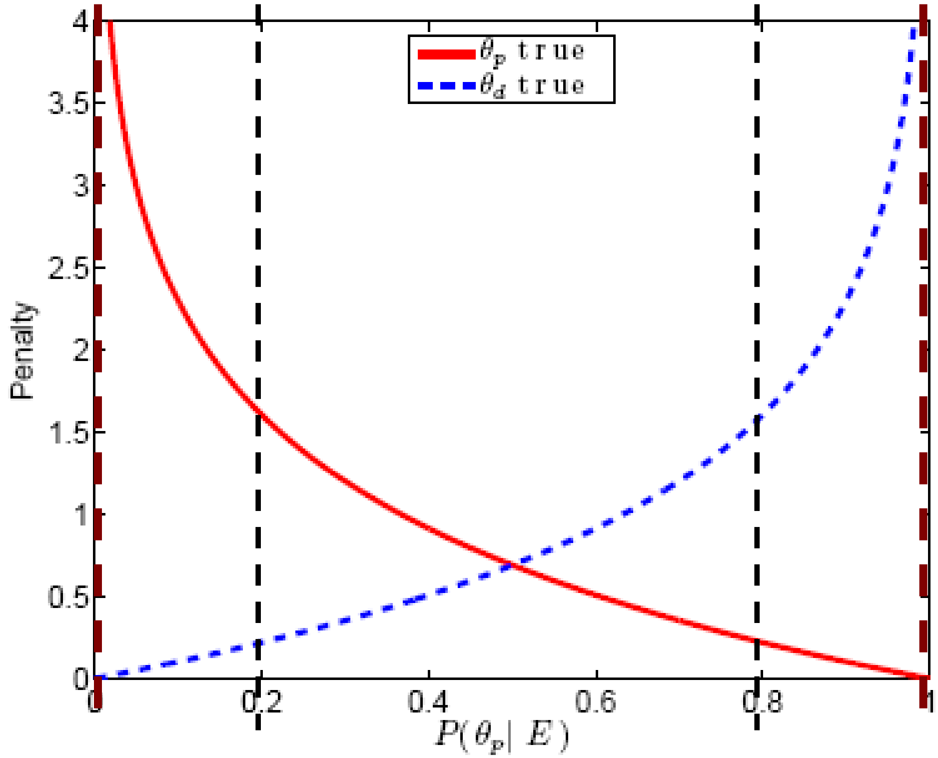


- Not a disrcrimination problem
  - The same in both of them
- *Calibration* problem [deGroot 1982]

Strong support to the wrong hypothesis! Will lead to errors

Accuracy of the LR

- Accuracy of a probabilistic opinion (*forecast*)
  - Classically measured by Strictly Proper Scoring Rules (SPSR) [deGroot 1982]



Accuracy: average value of the SPSR 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)

- ECE is the prior-weighted average value of a SPSR
- Empirical approach: experimental test
    - Generate same-source comparisons (θ_p is true)
    - Generate different-source comparisons (θ_d is true)

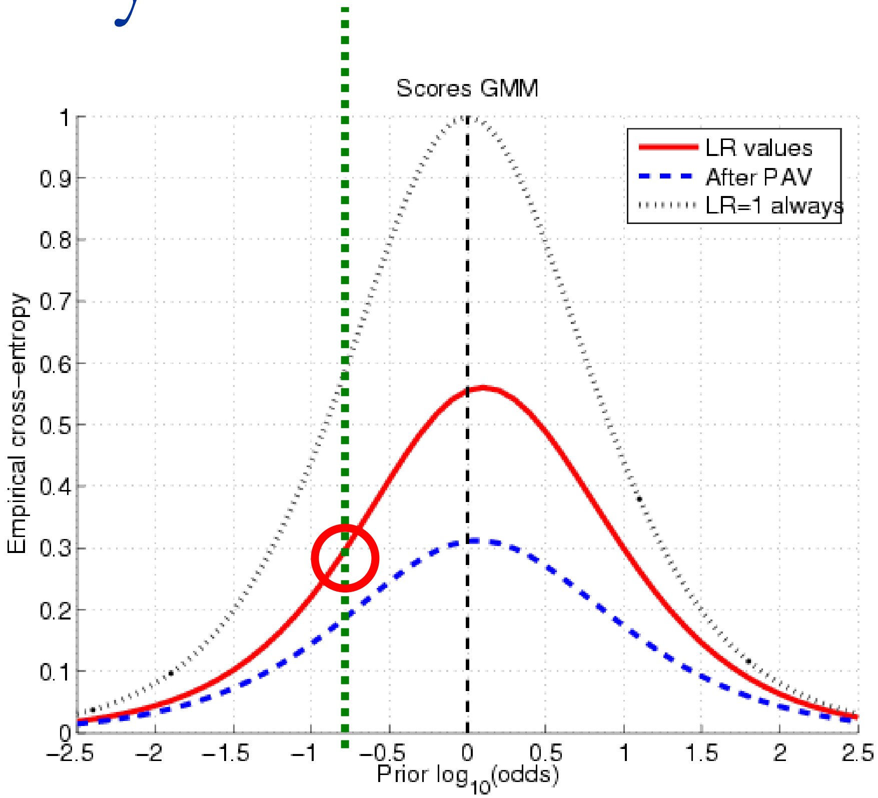
$$ECE = -P(\theta_p)\frac{1}{N_p}\sum_{j\in\text{same-source}}\log_2 P(\theta_p|e_j)$$

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

- However, it depends on the prior
    - The forensic scientist cannot compute its value
- Solution: the ECE plot
    - Prior-dependent representation

ECE plots: LR accuracy

- ECE depends on the prior
  - Compute it for every prior
- ECE is the red curve
- The higher the ECE:
  - Infromation loss!
  - Blue curve: best calibrated
  - Dotted curve: neutral (LR=1)



- 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 measure of *ECE* in the plot

More on ECE and LR accuracy

Information-theoretical comparison of likelihood ratio methods of forensic evidence evaluation

Daniel Ramos and Joaquin Gonzalez-Rodriguez,
ATVS - Biometric Recognition Group, Universidad Autonoma de Madrid, Spain; daniel.ramos@uam.es

Grzegorz Zadora and Janina Zieba-Palus,*
Institute of Forensic Research, Westerplatte 9, 31-033 Krakow, Poland; gzadora@ies.krakow.pl

Colin Aitken,
School of Mathematics and Joseph Bell Centre for Forensic Statistics and Legal Reasoning
University of Edinburgh, King's Buildings, Edinburgh, EH9 3JZ, UK; c.g.g.aitken@ed.ac.uk

Abstract

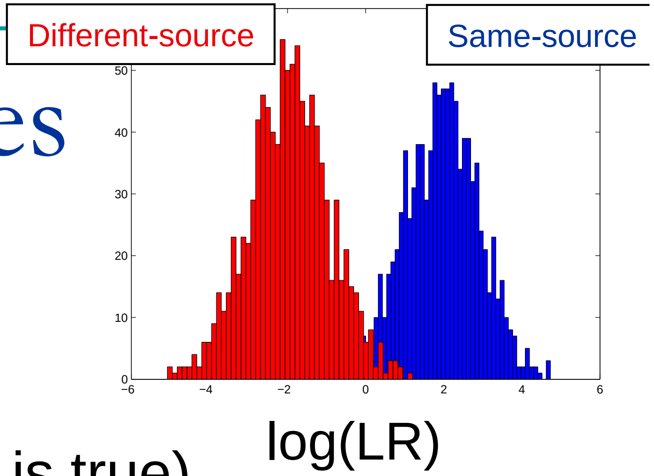
Forensic evidence in the form of two-level hierarchical

uated by consideration of their physico-chemical features (*e.g.*, chemical composition).

Evidence evaluation requires consideration of the vari-

Detecting LR values which degrade accuracy

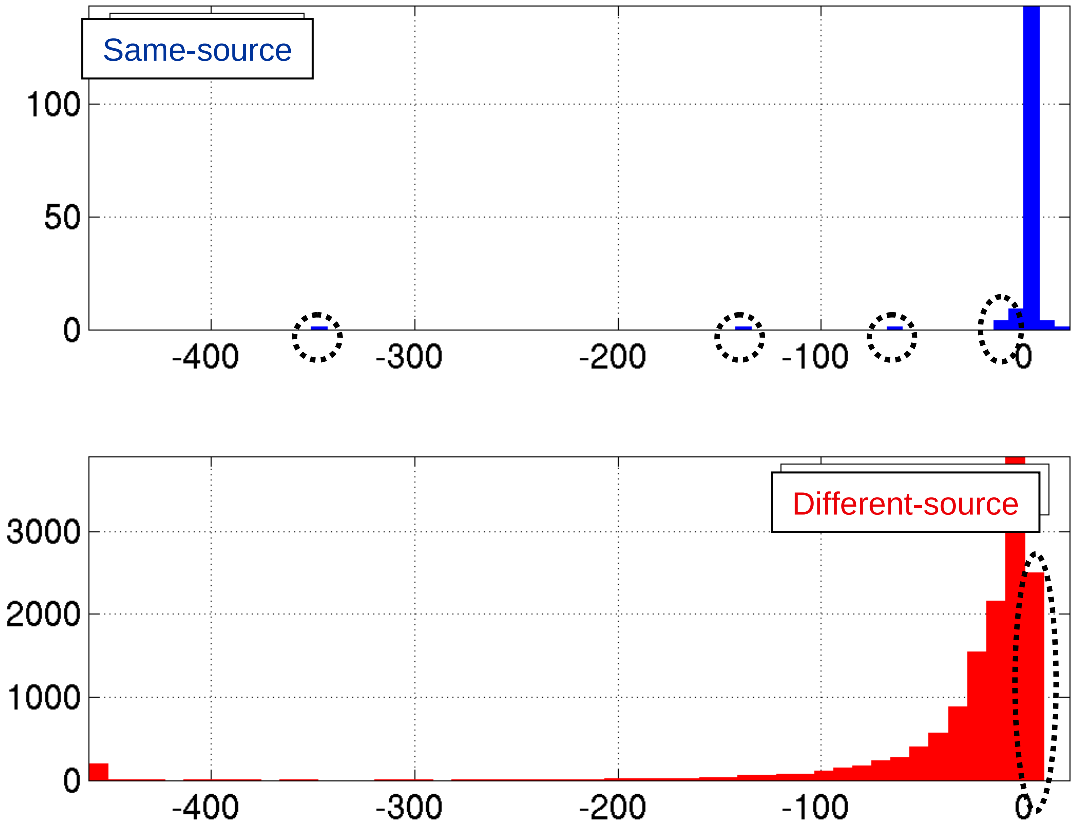
Detecting "bad" LR values



- Experimental test
  - Database of data with known sources
  - Generate same-source comparisons (θ_p is true)
  - Generate different-source comparisons (θ_d is true)
- For each of ones:
  - "Worst" LR values will be the most misleading
    - Lower values when θ_p is true (same-source)
    - Higher values when θ_d is true (different-source)
- "Worst" LR values will increase ECE the most
  - Accuracy degradation

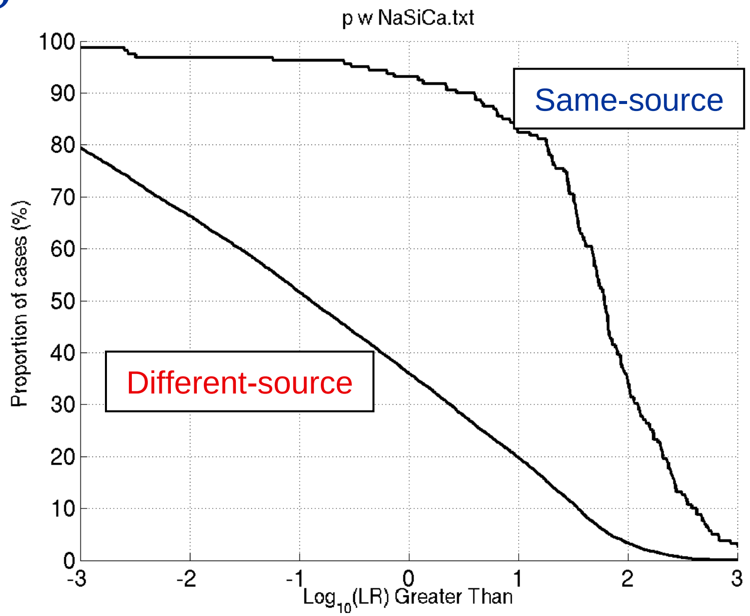
Hypothesis-dependent Histograms

- They help on detecting the worst LR values

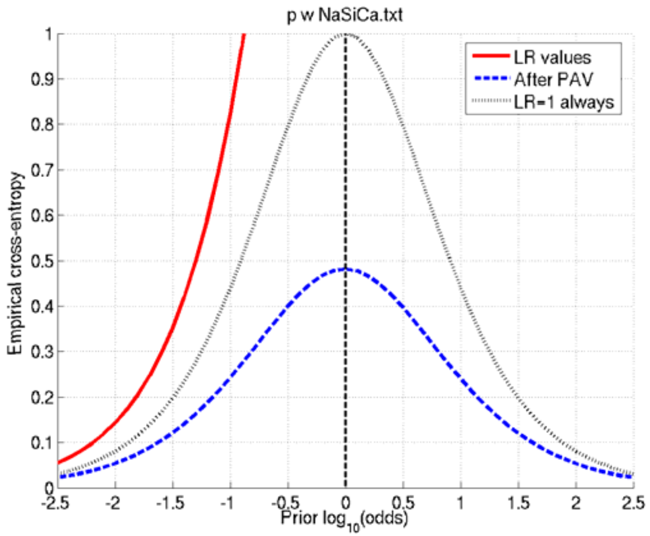


ECE and Histograms

- Histograms:
  - Simple
  - Easy to understand
  - Tippett plots are in fact cumulative histograms...



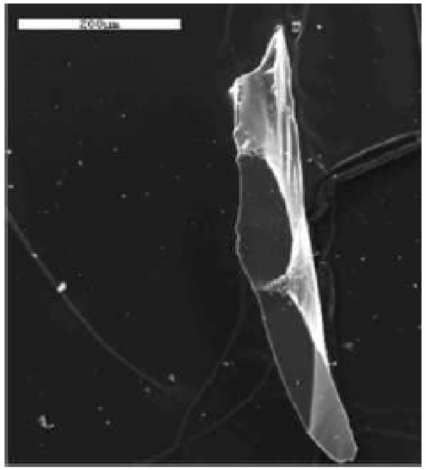
- ECE
  - It tells to what extent a bad LR value affects accuracy
    - It helps detecting extremely pathological cases
    - In terms of information loss



Examples with glass evidence

Database (Institute of Forensic Research, Krakow, Poland)

- 164 glass items coming from windows (w)
- 56 glass items comings from containers (p)
- 4 measurements of elemental composition per object
- 3 selected variables (7 variables in the database):
  - log(Na/O), or Na’
  - log(Si/O), or Si’
  - log(Ca/O), or Ca’



Experimental protocol

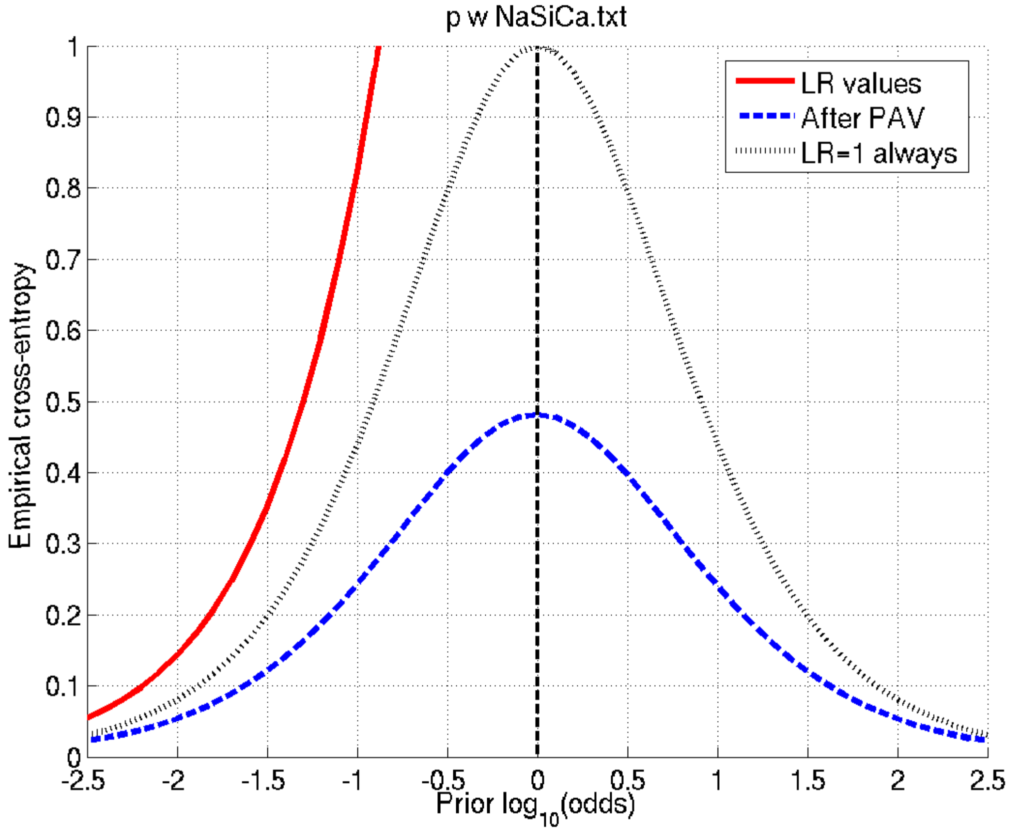
- Control and recovered data
- For same-source trials
  - 2 measurements per w object as recovered data
  - 2 measurements of the same w object as control data
- For different-source trials
  - 4 measurements per w object as recovered data
  - 4 measurements of a different w object as control data

Experimental protocol

- Sample experiment (mismatch)
  - p items used as background modelling
    - Used to compute within- and between-source variation
  - w items used as control-recovered data
- LR values computed using multivariate model as in [Aitken and Lucy 2004]
  - Normal density for within-source
  - Kernel density for between-source

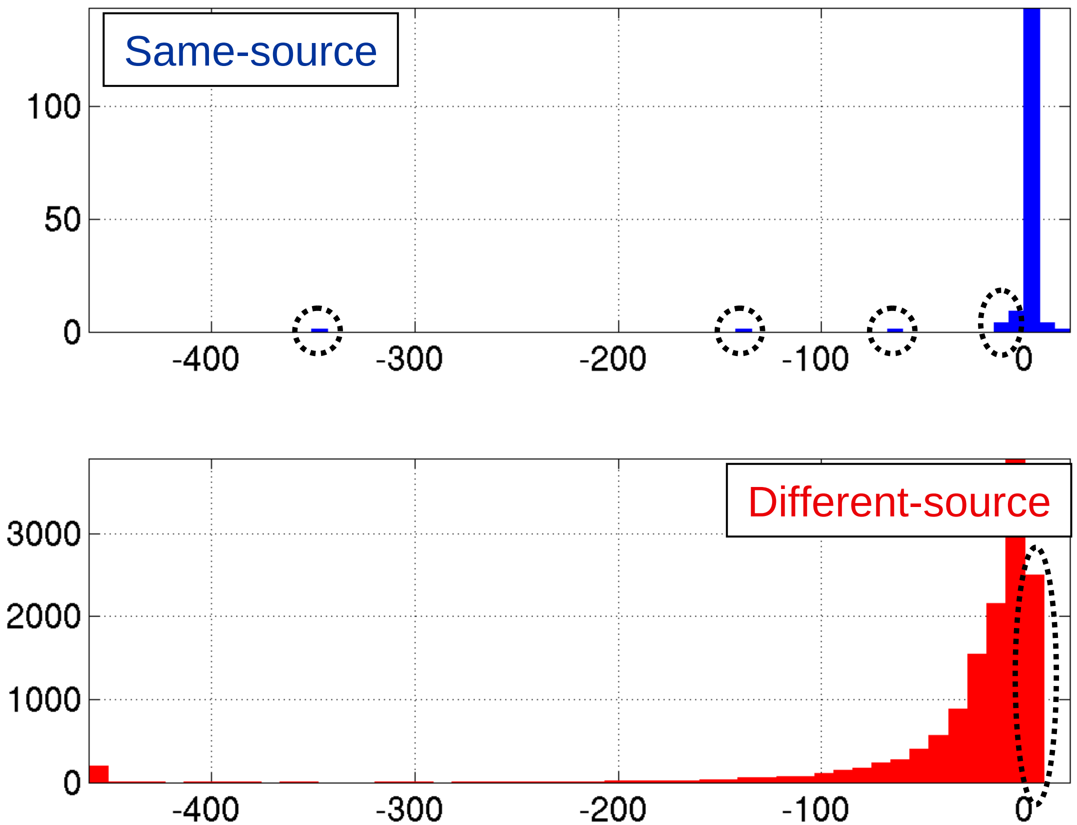
Accuracy

- ECE plots denote something bad is happening



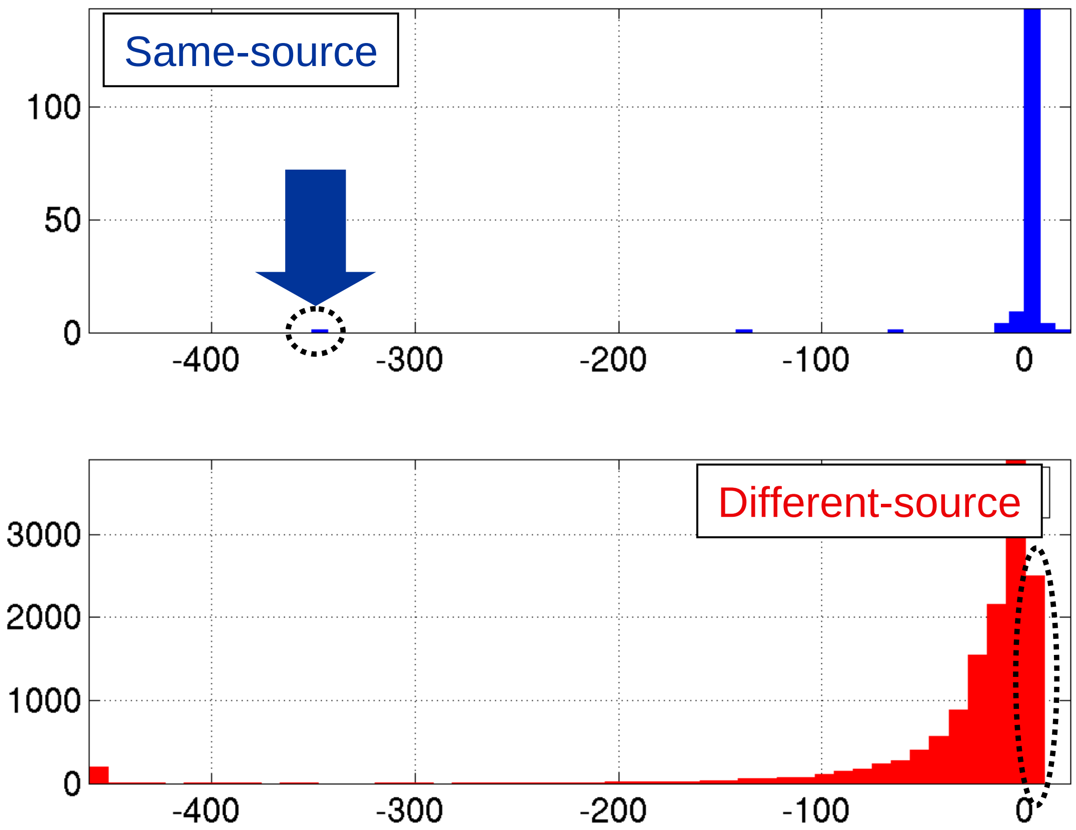
Hypothesis-dependent histograms

- We identify "bad" LR values
  - Lower same-source LR values
  - Higher different-source LR values

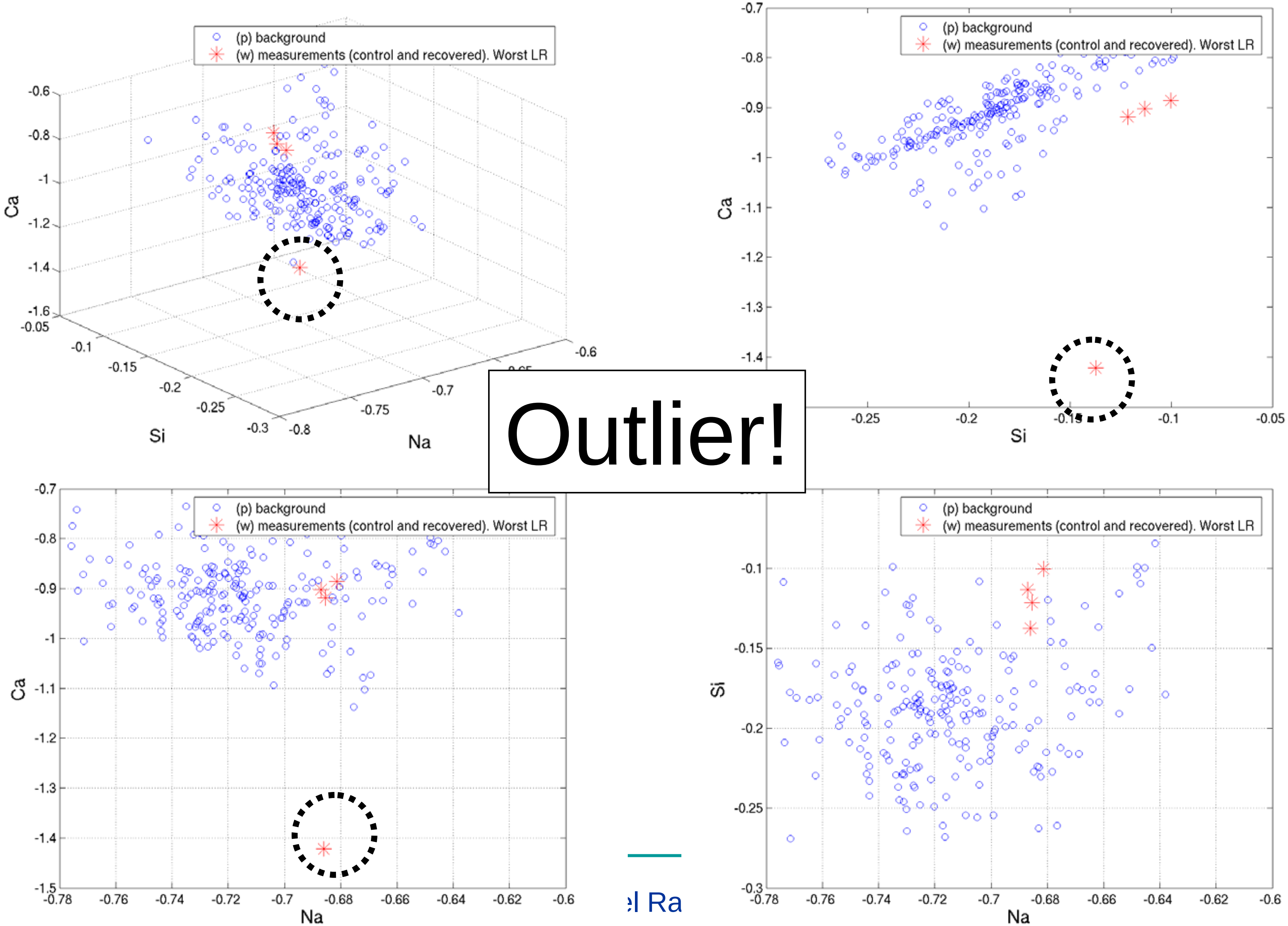


Same-source experiments

- What's happening with the worst same-source LR value?

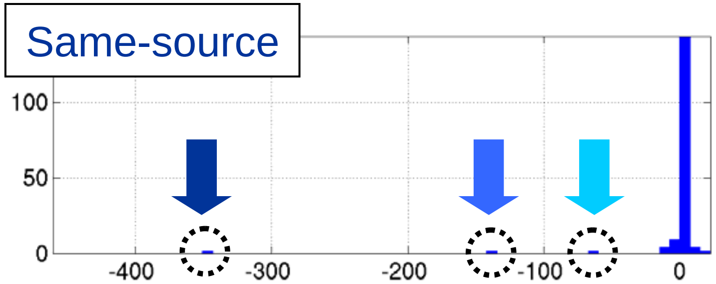


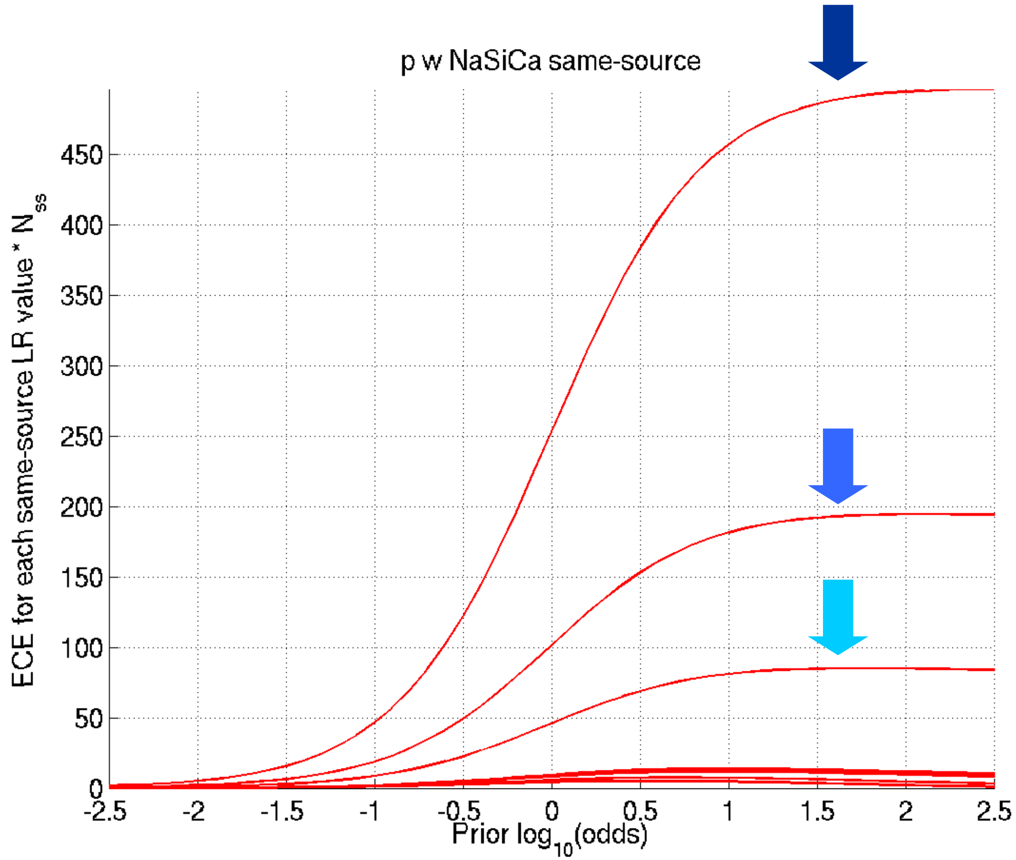
Same-source experiments



Contribution to cross-entropy

- The "bad" same-source LR values enormously degrade empirical cross-entropy...



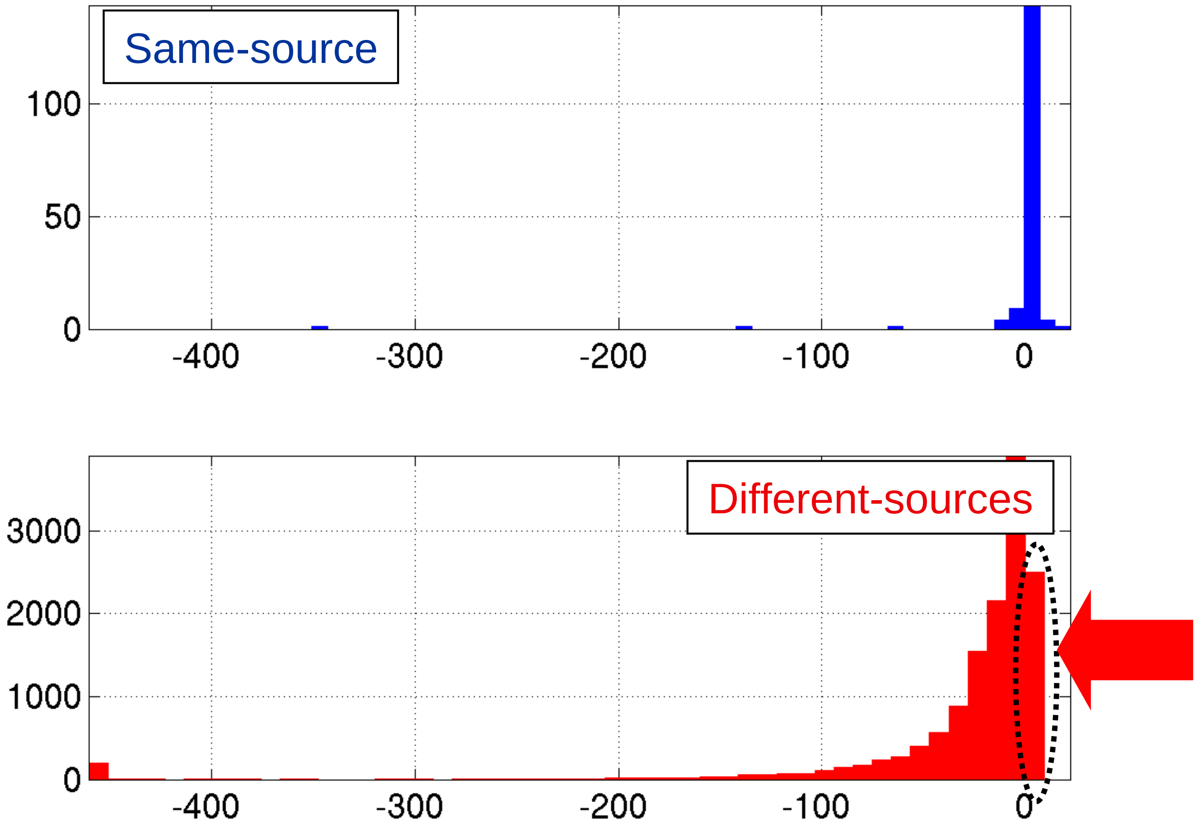


Problems and solutions: same source

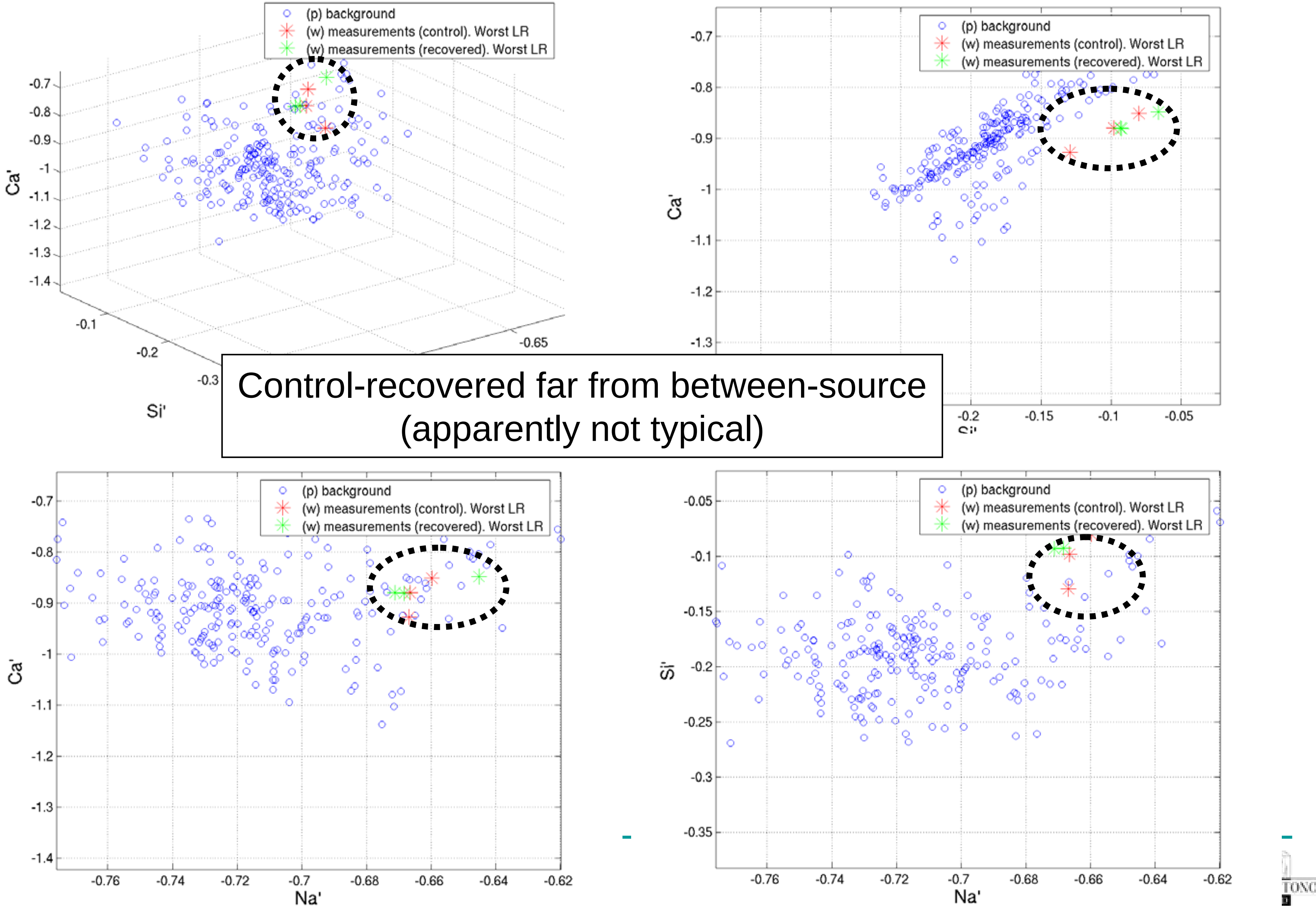
- Model is sensitive to outliers in control-recovered data
  - More control-recovered data should be collected
  - An outlier detection / compensation strategy should be used
  - Ca’ variable should be avoided
  - ...

Different-source experiments

- What is happening with the worst different-source LR value?

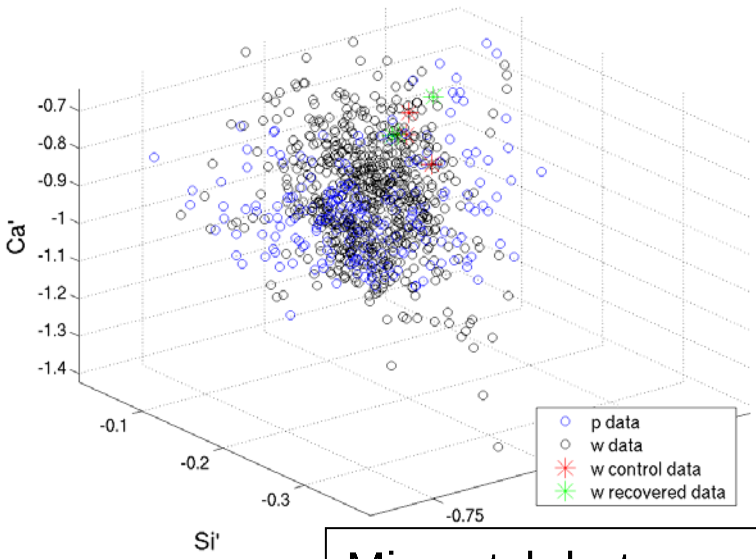


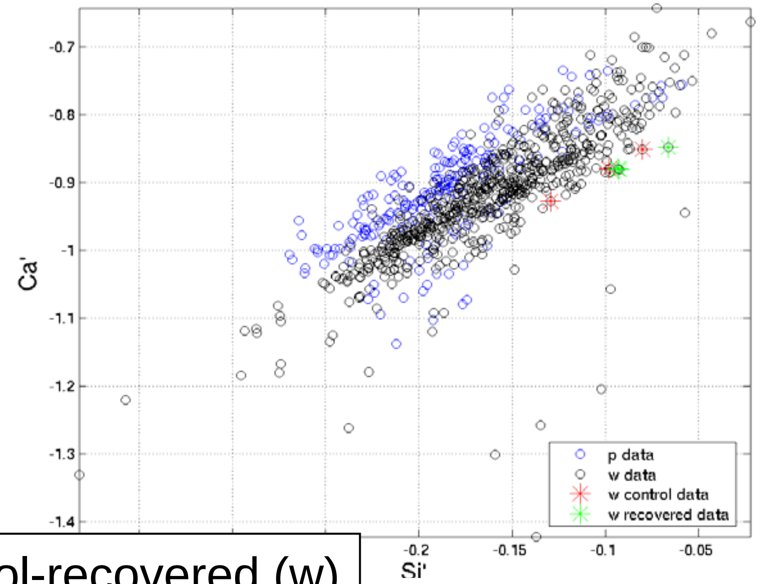
Different-source experiments



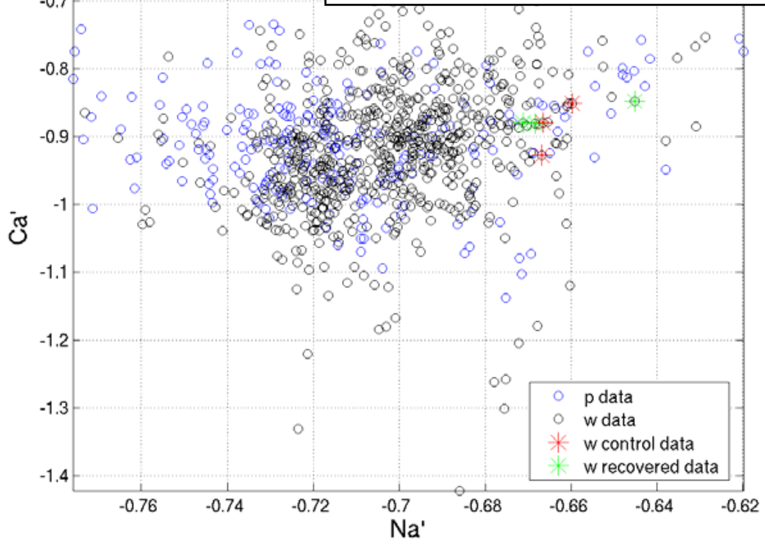
Control-recovered far from between-source
(apparently not typical)

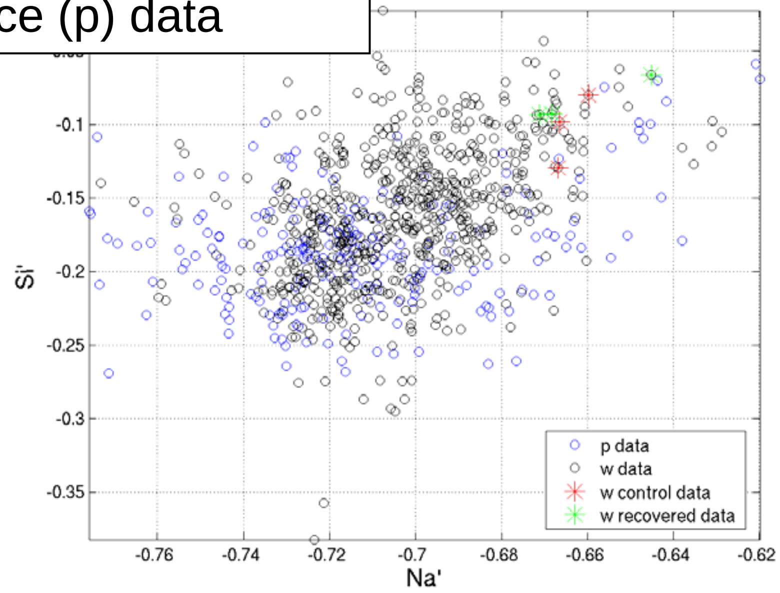
Different-source experiments





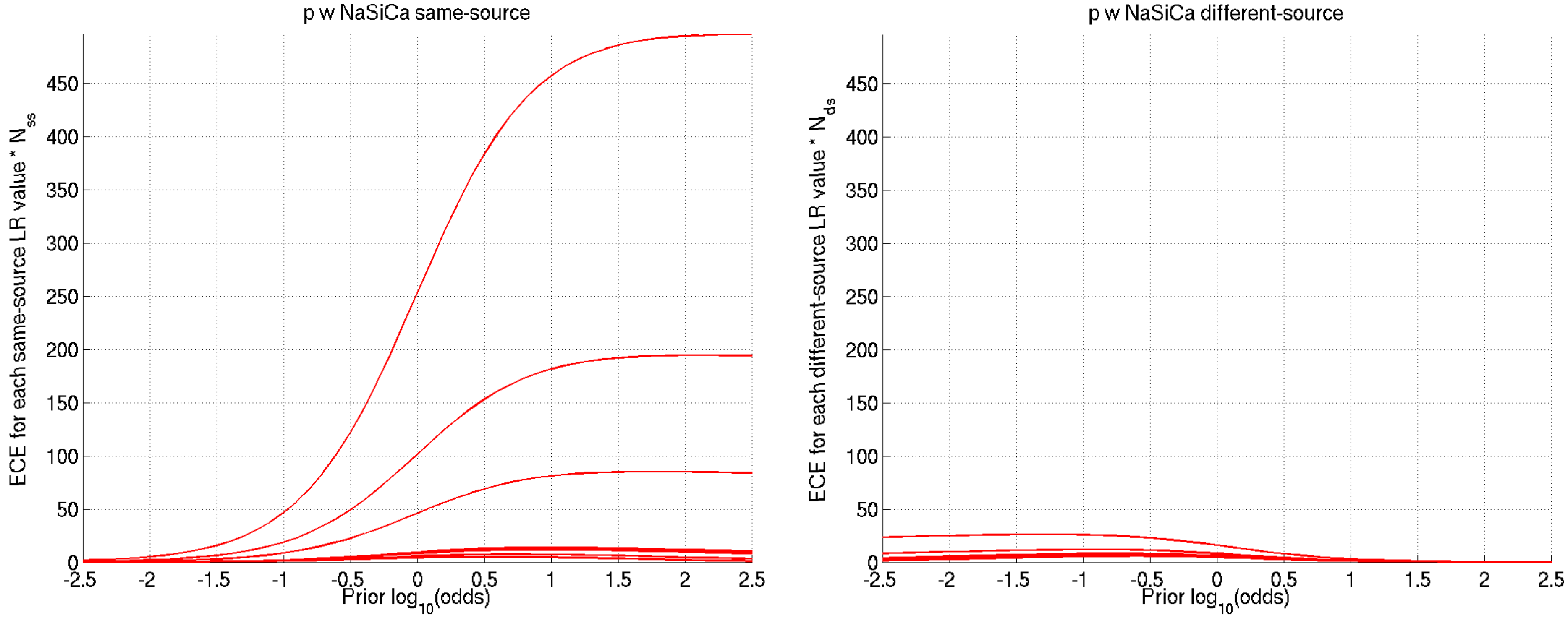
Mismatch between control-recovered (w) and between-source (p) data





Contribution to cross-entropy

- Degradation of accuracy is smaller for different-source LR values than for same-source LR values
  - LR values are not so “bad” for different-source comparisons



Problems and solutions: different source

- Between-source modelling with a non-proper population
  - Different-source comparisons unrealistically assumed as non-typical
  - LR values get high
- A proper population is important in glass analysis
  - Representative of control-recovered type of data
  - Study in Zadora et al. 2008 (ICFIS Lausanne, to appear)

Conclusions

Conclusions

- Problems in evidence evaluation can be detected
  - “Bad” LR values are easily seen in
    - Hypothesis-dependent histograms
    - Contribution to ECE
- Deeper analysis starting from “bad” LR values show the causes of the problem
  - Glass experiments show typical problems
    - Outliers
    - Mismatch between population and control-recovered data
- ECE measures the impact of such problems in the accuracy of the evidence evaluation methods

Thanks!

Acknowledgements:
Grzegorz Zadora *for providing glass data, and for review and comments.*
Joaquin Gonzalez-Rodriguez *and* ***Colin Aitken*** *for review and comments.*

Daniel Ramos
daniel.ramos@uam.es