Measuring Performance of Likelihood-Ratio-Based Evidence Evaluation Methods

Daniel Ramos
ATVS – Biometric Recognition Group
daniel.ramos@uam.es
http://atvs.ii.uam.es
Universidad Autonoma de Madrid, Spain
Introduction and Motivation: A Validation Guideline For LR-Based Forensic Evaluation Methods
A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation

Didier Meuwly\textsuperscript{a,b,*}, Daniel Ramos\textsuperscript{c}, Rudolf Haraksim\textsuperscript{d}

\textsuperscript{a,b,*} Netherlands Forensic Institute, Laan van Ypenburg 6, 2497GB The Hague, The Netherlands
\textsuperscript{c} University of Twente, Drienerlolaan 5, 7522NB Enschede, The Netherlands
\textsuperscript{d} ATVS – Biometric Recognition Group, Escuela Politecnica Superior, Universidad Autonoma de Madrid, C/Juan Francisco Tomas y Valiente 11, 28049 Madrid, Spain
\textsuperscript{d} LTISS – Signal Processing Laboratory, École Polytechnique Fédérale de Lausanne, Faculty of Electrical Engineering, Station 11, CH-1015 Lausanne, Switzerland
Guideline: Validation of Forensic LR Methods

- **Objective**
  - Determine if a LR method is valid to be used in casework
Guideline: Validation of Forensic LR Methods

- **Objective**
  - Determine if a LR method is valid to be used in casework
  - All the validation process should be documented for transparency
Guideline: Validation of Forensic LR Methods

Objective

- Determine if a LR method is valid to be used in casework
- All the validation process should be documented for transparency
- Towards standardization of procedures (for biometrics)
Guideline: Validation of Forensic LR Methods

- **Objective**
  - Determine if a LR method is valid to be used in casework
  - All the validation process should be documented for transparency
  - Towards *standardization* of procedures (for biometrics)

- **Validation process**
  - Based on *Empirical Testing*
    - Data: still an issue
Guideline: Validation of Forensic LR Methods

Objective
- Determine if a LR method is valid to be used in casework
- All the validation process should be documented for transparency
- Towards standardization of procedures (for biometrics)

Validation process
- Based on Empirical Testing
  - Data: still an issue

Emphasis on Forensic Data
- Lab (development) performance
- Followed by forensic performance
Guideline: Validation of Forensic LR Methods

- **Objective**
  - Determine if a LR method is valid to be used in casework
  - All the validation process should be documented for transparency
  - Towards *standardization* of procedures (for biometrics)

- **Validation process**
  - Based on *Empirical Testing*
    - Data: still an issue
  - Performance assessment
    - Performance *characteristics*
      - *What aspect of performance should be measured?*
    - Performance *metrics*
      - *How to measure a characteristic?*
    - Performance graphical representations
      - *Ok, show me an illustrating plot*

---

Emphasis on Forensic Data
- Lab (development) performance
  - Followed by forensic performance
Guideline: Validation of Forensic LR Methods

Objective
- Determine if a LR method is valid to be used in casework
- All the validation process should be documented for transparency
- Towards standardization of procedures (for biometrics)

Validation process
- Based on Empirical Testing
  - Data: still an issue
- Performance assessment
  - Performance characteristics
    - What aspect of performance should be measured?
  - Performance metrics
    - How to measure a characteristic?
  - Performance graphical representations
    - Ok, show me an illustrating plot

Emphasis on Forensic Data
- Lab (development) performance
- Followed by forensic performance

Terminology

<table>
<thead>
<tr>
<th>Performance Characteristic</th>
<th>Performance Metric</th>
<th>Graphical Representation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accuracy</strong></td>
<td>Cllr</td>
<td>ECE plot</td>
</tr>
<tr>
<td><strong>Discriminating power</strong></td>
<td>Cllr_{min} \ EER</td>
<td>ECE_{min} plot \ DET plot</td>
</tr>
<tr>
<td><strong>Calibration</strong></td>
<td>Cllr_{cal}</td>
<td>Tippett plot</td>
</tr>
<tr>
<td><strong>Robustness</strong></td>
<td>Cllr \ EER \ LR range</td>
<td>ECE plot \ DET plot \ Tippett plot</td>
</tr>
<tr>
<td><strong>Coherence</strong></td>
<td>Cllr \ EER</td>
<td>ECE plot \ DET plot \ Tippett plot</td>
</tr>
<tr>
<td><strong>Generalization</strong></td>
<td>Cllr \ EER</td>
<td>ECE plot \ DET plot</td>
</tr>
<tr>
<td>Performance Characteristic</td>
<td>Performance Metric</td>
<td>Graphical Representation</td>
</tr>
<tr>
<td>----------------------------</td>
<td>--------------------</td>
<td>--------------------------</td>
</tr>
<tr>
<td><strong>Accuracy</strong></td>
<td>Cllr</td>
<td>ECE plot</td>
</tr>
<tr>
<td><strong>Discriminating power</strong></td>
<td>Cllr\textsubscript{min} EER</td>
<td>ECE\textsubscript{min} plot DET plot</td>
</tr>
<tr>
<td><strong>Calibration</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Robustness</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Coherence</strong></td>
<td>Cllr EER</td>
<td>ECE plot DET plot Tippett plot</td>
</tr>
<tr>
<td><strong>Generalization</strong></td>
<td>Cllr EER</td>
<td>ECE plot DET plot</td>
</tr>
</tbody>
</table>

\textbf{Guideline for Validation: Performance}

\textbf{Guideline is intended to be open in this sense}
Aim of This Talk

- Introducing a general Bayesian framework for performance assessment of LRs
- Describe some of the performance metrics
- Show some examples using real forensic data
- Introducing a Validation Toolbox for forensic researchers
Performance of the LR: A Bayesian Decision Framework
Bayesian Decisions

- Inference
  - Prior probability (odds), before knowing the evidence
  - Posterior probability (odds) after observing the evidence
  - LR is the value of the evidence: from prior to posterior odds
Bayesian Decisions

- Costs
  - Penalty of making a wrong decision towards $H_p (C_{fp})$ or $H_d (C_{fd})$.
  - Can be different
    - Example at the offence level: is it better to condemn an innocent (cost $C_{fp}$) or to release a guilty (cost $C_{fd}$)?
Bayesian Decisions

- Decision
  - Categorical, decide between propositions $H_p$ or $H_d$
  - Based on the posterior odds, and the costs
  - Bayes’ decision rule: $LR \geq \frac{P(H_d)C_{fd}}{P(H_p)C_{fp}}$
    Minimizes the expected cost (risk)
Bayesian Decisions

- Ultimate aim of the Forensic Scientist
  - Improve decisions of the fact finder

Responsibility of the Forensic Scientist

Responsibility of the Fact Finder

Prior → Inference → Posterior → Decision

Decision $H_p$ or $H_d$?

Costs $C_{fp}, C_{fd}$

LR

Statistical Modelling of Forensic Evidence
INI Cambridge, UK, 7-11 November
Better Decisions and The LR
**Performance of Decisions**

- Bayesian decisions aim at minimizing the risk (expected cost)
- Elements of risk
  - Posterior probabilities
  - Decision costs

\[ LR \geq \frac{P(H_d)C_{fd}}{P(H_p)C_{fp}} \]
Performance of Decisions

- Bayesian decisions aim at minimizing the risk (expected cost)
- Elements of risk
  - Posterior probabilities
  - Decision costs

But a forensic LR method does not deal with priors or costs... (thus, the forensic method cannot make decisions)

\[ LR \equiv \frac{P(H_d) C_{fd}}{P(H_p) C_{fp}} \]
Accuracy: A Definition for the LR

- A LR is more accurate when...
  - It supports $H_p$ in a stronger way when $H_p$ is actually true
  - It supports $H_d$ in a stronger way when $H_d$ is actually true
Accuracy: A Definition for the LR

A LR is more accurate when...
- It supports $H_p$ in a stronger way when $H_p$ is actually true
- It supports $H_d$ in a stronger way when $H_d$ is actually true

Increasing accuracy of LRs reduces the expected cost of decisions for fixed priors and costs

Should lead to minimum-cost decisions!
Why Accuracy of LRs?

Accuracy of LRs is proposed as the main performance characteristic

- Because LR accuracy is maximum when the LR is perfect
  - LR = infinity when $H_p$ is true
  - LR = 0 when $H_d$ is true
Why Accuracy of LRs?

Accuracy of LRs is proposed as the main performance characteristic

- Because LR accuracy is maximum when the LR is perfect
  - LR = infinity when $H_p$ is true
  - LR = 0 when $H_d$ is true

- Because LR accuracy focuses on the LR, not on posteriors/decisions
  - Independently of priors and costs
    - Not the responsibility of the LR method
Why Accuracy of LRs?

Accuracy of LRs is proposed as the main performance characteristic

- Because LR accuracy is maximum when the LR is perfect
  - LR = infinity when $H_p$ is true
  - LR = 0 when $H_d$ is true

- Because LR accuracy focuses on the LR, not on posteriors/decisions
  - Independently of priors and costs
    - Not the responsibility of the LR method

- Because LR accuracy addresses the ultimate aim of the LR
  - Aiding the fact finder to make better decisions
A Metric for LR Accuracy Using Strictly Proper Scoring Rules (SPSR)
Strictly Proper Scoring Rules (SPSR)

- Assigns a penalty to a posterior probability
- Average of a SPSR: “goodness” of probabilities
  - More accurate LRs imply lower average values of the SPSR

Example: logarithmic SPSR:

\[
P(H_p|E)
\]

\[
P(H_d|E)
\]
Goodness of Forensic Posteriors

- Posterior probability in forensic evaluation

\[ P(H_p|E,I) \]

“Probability of \( H_p \) given \( E,I \)”

- \( H_p \): prosecutor hypothesis
- \( H_d \): defense hypothesis
- \( E, I \): available knowledge
  (evidence + other information)

- But the forensic examiner must not assess the prior!
  - Therefore, she or he cannot use the posterior!
  - How to measure “goodness” then with a SPSR?
LR Accuracy in Forensic Evaluation

- **Step 1**: set-up a validation experiment
  - Compute LR values
  - Using a validation database
  - This is done *for validation, not for casework*
LR Accuracy in Forensic Evaluation

- **Step 1:** set-up a validation experiment
  - Compute LR values
  - Using a validation database
  - This is done for validation, not for casework

Ground-truth:
- Either \( H_p \) is true
- Or \( H_d \) is true

Validation database (known identities)

- Calculate True-\( H_p \) LRs
- Calculate True-\( H_d \) LRs
LR Accuracy in Forensic Evaluation

- **Step 1:** set-up a validation experiment
  - Compute LR values
  - Using a validation database
  - This is done for validation, not for casework

**Ground-truth:**
- Either $H_p$ is true
- Or $H_d$ is true

**Empirical Validation LR set**
- True-$H_d$ LR set
- True-$H_p$ LR set

**Validation database (known identities):**
- Compute True-$H_p$ LR set
- Compute True-$H_d$ LR set
LR Accuracy in Forensic Evaluation

- **Step 2:** consider the prior/costs as **unknown** parameters
  - Do not assess its value in any case!
  - But vary it over a wide range **within the experiment**
    - In casework, however, you will just compute the LR!

- Compute and represent accuracy (average of SPSR) for all the priors in that range

![Graph showing accuracy as a function of priors/costs](image)

- **Accuracy as a function of the priors/costs**
- **Vary priors and costs**
- **True-$H_d$ LRs**
- **True-$H_p$ LRs**
**LR Accuracy in Forensic Evaluation**

- **Step 2:** consider the prior/costs as *unknown* parameters
  - Do not assess its value in any case!
  - But vary it over a wide range within the experiment
    - In casework, however, you will just compute the LR!

- Compute and represent accuracy (average of SPSR) for all the priors in that range

---

*In this simulated experiment, if we assume priors/costs varying in a range, what would be the accuracy of LRs as a function of those priors/costs?*
LR Accuracy: Empirical Cross-Entropy

- Proposed choice of SPSR: logarithmic SPSR
  - It can be argued that it has nice properties
- Accuracy: Empirical Cross-Entropy

\[
ECE = -P(H_p) \text{Average}[\log_2 P_i(H_p|E)] - P(H_d) \text{Average}[\log_2 P_i(H_d|E)]
\]

Proposed choice of SPSR: logarithmic SPSR
- It can be argued that it has nice properties

Accuracy: Empirical Cross-Entropy
- We only vary prior odds in the validation experiment
- In casework only the LR will be reported (as usual)

\[
ECE = -P(H_p) \text{Average} [\log_2 P_i(H_p|E)] - P(H_d) \text{Average} [\log_2 P_i(H_d|E)]
\]

Calibration
Calibration

- Extremely important property of a set of probabilistic assessments
- Studied in classical Bayesian statistics literature
  - Dubbed as “reliability” of probabilistic assessments

- More recently, calibration of LRs

---

**The Statistician** 32 (1983)

*The Comparison and Evaluation of Forecasters*†

MORRIS H. DeGROOT and STEPHEN E. FIENBERG

**Journal of the American Statistical Association**

September 1982, Volume 77, Number 379

*The Well-Calibrated Bayesian*  
A. P. DAWID*

---

**Niko Brümmer** a,b,*, Johan du Preez b
Application-independent evaluation of speaker detection

**The distribution of calibrated likelihood-ratios in speaker recognition**

David A. van Leeuwen1 and Niko Brümmer2


---

**Forensic Science International**

Reliable support: Measuring calibration of likelihood ratios^2

Daniel Ramos*, Joaquin Gonzalez-Rodriguez

---

Statistical Modelling of Forensic Evidence  
INI Cambridge, UK, 7-11 November
SPSR = Discrimination + Calibration

- A property of the average value of a SPSR
- E.g., ECE can be decomposed into
  - Discriminating power of the LR set
    - Distinguish between true-$H_p$ and true-$H_d$ cases
  - Calibration of the LR set
    - “Reliability” of the LR

\[ \text{ECE of LRs} = \text{Discriminating Power of LRs} + \text{Calibration of LRs} \]

Degree of overlap (roughly speaking)
Some Performance Metrics and Representations (Included in the Guideline)
## Guideline for Validation: Performance

<table>
<thead>
<tr>
<th>Performance Characteristic</th>
<th>Performance Metric</th>
<th>Graphical Representation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accuracy</strong></td>
<td>Cllr</td>
<td>ECE plot</td>
</tr>
<tr>
<td><strong>Discriminating power</strong></td>
<td>Cllr\text{min} EER</td>
<td>ECE\text{min} plot DET plot</td>
</tr>
<tr>
<td><strong>Calibration</strong></td>
<td>Cllr\text{cal}</td>
<td>Tippett plot</td>
</tr>
<tr>
<td><strong>Robustness</strong></td>
<td>Cllr EER LR range</td>
<td>ECE plot DET plot Tippett plot</td>
</tr>
<tr>
<td><strong>Coherence</strong></td>
<td>Cllr EER</td>
<td>ECE plot DET plot Tippett plot</td>
</tr>
<tr>
<td><strong>Generalization</strong></td>
<td>Cllr EER</td>
<td>ECE plot DET plot</td>
</tr>
</tbody>
</table>
## Guideline for Validation: Performance

<table>
<thead>
<tr>
<th>Performance Characteristic</th>
<th>Performance Metric</th>
<th>Graphical Representation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Primary</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>Cllr</td>
<td>ECE plot</td>
</tr>
<tr>
<td>Discriminating power</td>
<td>Cllr&lt;sub&gt;min&lt;/sub&gt;</td>
<td>ECE&lt;sub&gt;min&lt;/sub&gt; plot</td>
</tr>
<tr>
<td></td>
<td>EER</td>
<td>DET plot</td>
</tr>
<tr>
<td>Calibration</td>
<td>Cllr&lt;sub&gt;cal&lt;/sub&gt;</td>
<td>Tippett plot</td>
</tr>
<tr>
<td></td>
<td>Cllr</td>
<td>ECE plot</td>
</tr>
<tr>
<td></td>
<td>EER</td>
<td>DET plot</td>
</tr>
<tr>
<td></td>
<td>LR range</td>
<td>Tippett plot</td>
</tr>
<tr>
<td>Robustness</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coherence</td>
<td>Cllr</td>
<td>ECE plot</td>
</tr>
<tr>
<td></td>
<td>EER</td>
<td>DET plot</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tippett plot</td>
</tr>
<tr>
<td>Generalization</td>
<td>Cllr</td>
<td>ECE plot</td>
</tr>
<tr>
<td></td>
<td>EER</td>
<td>DET plot</td>
</tr>
</tbody>
</table>
ABSTRACT:

The DET Curve in Assessment of Detection Task Performance

G. Doddington, T. Kamm, A. Martin, M. Ordowski, M. Przybocki

Introduction

Many detection tasks can be viewed as involving a tradeoff between two types of error: missed detections and false alarms. In speech processing this is particularly true when the task is to recognize the person who is speaking, or to recognize the language being spoken. A recognition system may fail to detect a target speaker or language known to the system, or it may declare such a detection when the target is not present.

When there is a tradeoff of error types, a single performance number is inadequate to represent the capabilities of a system. Such a system has many operating points, and is best represented by a performance curve. The ROC (Receiver Operating Characteristic) Curve [1] is traditionally used for this purpose. We have found it useful in speech applications to use a variant of this which we call the DET (Detection Error Tradeoff) Curve.

We have focused our evaluations on what we regard as the basic generic detection task. Participants are given a set of known targets (speakers or languages) for which their systems have trained models, and a set of unknown speech segments. For each segment and target the system must determine whether the segment is an instance of the target. The system output is a likelihood that the segment is an instance of the target. The scale of the likelihood is arbitrary but should be consistent across all decisions, with larger values indicating greater likelihood. This likelihood is then used to generate the DET curve.

Examples

Figure 1 shows DET Curves for several systems from the evaluation of a speaker recognition task. Figure 2 contrasts this with traditional ROC Curves for the same data. Some Features of note for the DET Curve, are reviewed below. Further examples of DET Curves may be obtained from NIST via the URL: [ftp://jaguar.ncsl.nist.gov/speaker/mar96/graphs/site_compare/].

Figure 1

0.1
0.2
0.5
1
2
5
10
20
40

Miss probability (in %)

Figure 2

0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9

0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9

False Alarm probability (in %)

Correct Detection (in %)


Summarizing metric: Equal Error Rate (EER)
Tippett Plots

- Cumulative distribution of LR values

Accuracy

Calibration

Discrimination

WARNING: Tippett plots do not measure them explicitly!

- Summarizing metric:
  - Rates of misleading evidence
Empirical Cross-Entropy and $C_{llr}$


Empirical Cross-Entropy and $C_{llr}$

Accuracy

Calibration

Discrimination


- Summarizing metric: $C_{llr}$

Niko Brümmer a,b,* , Johan du Preez b
Application-independent evaluation of speaker detection
Some Performance Metrics and Representations
(Not Included in the Guideline)
Some Performance Metrics and Representations
(Not Included in the Guideline)
(Well... Not yet)
Normalized Bayes Error Plots (NBE)

- Normalized error rates
  - Dependent on Bayes thresholds

Summarizing metric:
- $C_{llr}$ (relates to NBE by integration)

Accuracy

Calibration

Discrimination

https://sites.google.com/site/bosaristoolkit
Expected Value of the LR under $H_p$ or $H_d$

- A property of calibration
- Used as a performance metric

\[
1 = E_{\text{True-}H_d} \left[ LR \right] \approx \frac{1}{N_d} \sum_{i \in \text{True-}H_p} LR_i
\]

\[
1 = E_{\text{True-}H_p} \left[ \frac{1}{LR} \right] \approx \frac{1}{N_p} \sum_{j \in \text{True-}H_p} \frac{1}{LR_j}
\]

- The better-calibrated the LR, the closer the expected values will be to 1
An Example in Forensic Evaluation of Physicochemical Evidence
Evidence Evaluation with Physicochemical Multivariate Data

RESEARCH ARTICLE
Gaussian Mixture Models of Between-Source Variation for Likelihood Ratio Computation from Multivariate Data
Javier Franco-Pedroso*, Daniel Ramos, Joaquin Gonzalez-Rodriguez
ATVS-Biometric Recognition Group, Universidad Autonoma de Madrid, Madrid, Spain
Evidence Evaluation with Physicochemical Multivariate Data

Comparison of two models:

1. Evaluation of trace evidence in the form of multivariate data
   
   C. G. G. Aitken and D. Lucy
   

2. Proposal: replace KDF between-source distribution with a Gaussian Mixture Model (GMM)
Evidence Evaluation with Physico-Chemical Multivariate Data

- ECE plots to compare performance

---

Fig 14. ECE plots for the KDF and GMM approaches on the car-paints dataset when applying the cross-validation protocol. GMM is trained by maximizing Eq 36.
Validation Toolbox
Freely Downloadable in:
https://sites.google.com/site/validationtoolbox/
An Easy GUI to Measure LR Performance
Conclusions
Conclusions

- Recent Guideline for Validation of Forensic LR Methods
  - Empirical testing
  - Performance measurement
Conclusions

- Recent Guideline for Validation of Forensic LR Methods
  - Empirical testing
  - Performance measurement

- Definition of performance of LR values in a forensic context
  - Independent of the priors and costs
  - The LR aids in the decision process
Conclusions

- Recent Guideline for Validation of Forensic LR Methods
  - Empirical testing
  - Performance measurement

- Definition of performance of LR values in a forensic context
  - Independent of the priors and costs
  - The LR aids in the decision process

- Importance of calibration of LRs
  - “Reliability” of decisions made using the LR
Conclusions

- Recent Guideline for Validation of Forensic LR Methods
  - Empirical testing
  - Performance measurement

- Definition of performance of LR values in a forensic context
  - Independent of the priors and costs
  - The LR aids in the decision process

- Importance of calibration of LRs
  - “Reliability” of decisions made using the LR

- A plethora of performance representations and metrics
  - Guideline is open for future proposals
  - Continuous improvement!
Measuring Performance of Likelihood-Ratio-Based Evidence Evaluation Methods

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