Calibration of Likelihood-Ratios in Automatic Speaker Recognition

Applicability to other Forensic Technologies

Niko Brümmer, AGNITIO Research, South Africa

IDIAP, Martigny

BBfor2 Workshop, 14 December 2011

Agenda

The purpose of this presentation is to:

- Explain the role of calibration of likelihood-ratios in the Bayesian paradigm for presenting forensic evidence in court.
- Discuss solutions for both measurement and optimization of the calibration of automatic speaker recognizers.
- Propose that these solutions are applicable also to other forensic disciplines.

Outline

Outline

LR and the Bayesian paradigm

We review basics of the Bayesian paradigm (hopefully old news), in order to:

- Establish a common terminology, for explanation of new material.
- Emphasize key concepts.

LR and the Bayesian paradigm

In his book, WEIGHT-OF-EVIDENCE FOR FORENSIC DNA PROFILES, Balding states:

> *"Evidence is of value inasmuch as it alters the probability that the defendant is guilty."*

We will be interested in Bayesian calculations where the evidence contributes to the probability for guilt via the vehicle of the likelihood-ratio (LR).

The canonical forensic problem

We are given evidence:

$$e = (\text{trace at crime scene}, \text{sample from suspect}).$$

which could (in principle) be used to compute a posterior of the form:

$$P(H|e, \mathcal{C}),$$

where:

- H denotes the prosecution hypothesis that the trace originated from the suspect and
- $\mathcal{C}$ denotes a collection of other information and assumptions.

Bayes decisions

- The purpose of the forensic analysis of the evidence is to contribute (via the LR) to computing the posterior $P(H|e,\mathcal{C})$.
- But the end goal of the court case is not this posterior.

The court has to make a decision!

Bayes decision theory turns posteriors into decisions.

Bayes decisions
Odds notation

Let $h \in \{H, \neg H\}$, where:

$$H \equiv \text{prosecution hypothesis}, \qquad \neg H \equiv \text{defence hypothesis}$$

Any binary distribution, $P(h|\cdot)$, can be represented as odds:

$$\text{odds} = \frac{P(H|\cdot)}{P(\neg H|\cdot)} = \frac{P(H|\cdot)}{1 - P(H|\cdot)}$$

The odds represents the whole distribution, because:

$$P(H|\cdot) = \frac{\text{odds}}{\text{odds} + 1}, \qquad P(\neg H|\cdot) = \frac{1}{\text{odds} + 1}$$

Bayes decisions

Beyond a reasonable doubt

The court should decide in favour of the prosecution, if:

$$e, \mathcal{C} \text{ imply: } H \text{ is true beyond a reasonable doubt.}$$

For argument's sake, let reasonable doubt $\equiv$ 10 000 to 1. The decision is H if:

$$\text{posterior odds} = \frac{P(H|e,\mathcal{C})}{P(\neg H|e,\mathcal{C})} > \frac{10\,000}{1}$$

or equivalently:

$$P(\neg H|e,\mathcal{C}) < \frac{1}{10\,001}$$

Bayes decisions
Minimum expected cost

Or, the court could decide in favour of the prosecution, if:

expected cost of false acquit $>$ expected cost of false convict

For argument's sake, let:

$$\frac{\text{cost of false convict}}{\text{cost of false acquit}} = \frac{10\,000}{1}$$

Then, the minimum-expected-cost Bayes decision is H, if

$$P(H|e,\mathcal{C}) \times 1 > P(\neg H|e,\mathcal{C}) \times 10\,000$$

or equivalently,

$$\text{posterior odds} = \frac{P(H|e,\mathcal{C})}{P(\neg H|e,\mathcal{C})} > \frac{10\,000}{1}$$

Bayes decisions
Equivalence

Notice: The reasonable doubt criterion and the minimum-expected-cost criterion are equivalent in the sense that they lead to identical formulas and identical decisions.

We shall base the following analysis on minimum-expected-cost decisions, but because of this equivalence, the results are equally valid for the reasonable doubt criterion.

Outline

The likelihood-ratio

Now that we can use the posterior to make Bayes decisions, let us:

1. Compute the posterior from the LR.
2. Show how this is equivalent to making decisions directly with the LR.

The likelihood-ratio

Disassembling the posterior

If the prior information and the probability model for the evidence are independent, we can factor the posterior as:

$$\text{posterior odds} = \text{likelihood ratio} \times \text{prior odds}$$

We confine attention to this case and analyse it in more detail.

The likelihood-ratio
Disassembling the posterior

We confine attention to the special case where the posterior conditioning, $\mathcal{C}$, can be disassembled as:

$$P(H|e,\mathcal{C}) = P(H|e,\mathcal{M},\pi)$$

subject to the conditional independence assumptions as represented in directed graphical model notation:

$$\mathcal{M} \rightarrow e \leftarrow h \leftarrow \pi$$

$\mathcal{M}$: is a probabilistic model, defining the likelihood: $P(e|h,\mathcal{M})$.

π: is other information, defining the prior:
$P(H|\pi) = 1 - P(\neg H|\pi)$.

The likelihood-ratio

Disassembling the posterior

Under the assumption, $\mathcal{M} \rightarrow e \leftarrow h \leftarrow \pi$, we can factor the posterior odds as a product of LR and prior odds:

$$\frac{P(H|e,\pi,\mathcal{M})}{P(\neg H|e,\pi,\mathcal{M})} = \frac{P(e|\mathcal{M},H)}{P(e|\mathcal{M},\neg H)} \times \frac{P(H|\pi)}{P(\neg H|\pi)}$$

The likelihood-ratio

Bayes decision

We have:

$$\text{posterior odds} = \text{LR} \times \text{prior odds} \text{ and}$$

$$\text{posterior odds} > 10\,000$$

This gives:

$$\text{LR}(e) = \frac{P(e|H,\mathcal{M})}{P(e|\neg H,\mathcal{M})} > \text{threshold} = \frac{10\,000}{\text{prior odds}}$$

For any cost ratio (or reasonable doubt factor) and for any prior odds, the decision can be made by comparing the LR against some threshold.

Outline

LR and the Bayesian paradigm

Summary

- In general, the Bayesian paradigm allows minimum-expected-cost / beyond-a-reasonable-doubt decisions to be made via the posterior $P(H|e,\mathcal{C})$.
- We restrict attention to $\mathcal{C}$, such that:

$$\text{posterior odds} = \text{LR} \times \text{prior odds}$$

- Decide in favour of the prosecution, if:

$$\text{LR} > \text{threshold} = \frac{\text{cost ratio}}{\text{prior odds}}$$

LR and the Bayesian paradigm

Let's be realistic

In a real court case, it is unlikely that

- numerical values will be assigned to prior odds and cost ratio,
- the posterior will be explicitly calculated.

However, if we want to analyse the function of the LR, we need a mathematical model of how it could ideally be used. For this purpose, we adopt the Bayes decision paradigm.

Outline

Ideal LR calculation

We have motivated why we need to compute the LR. So let's compute one!

Ideal LR calculation

The probabilistic model

Let $\mathcal{M}$ be a probabilistic model relating vocal tracts, $\mathbf{v}$, and speech samples, $\mathbf{s}$.

We assume conditional independence of the form:

- For different speakers:

$$P(\mathbf{v}_1, \mathbf{v}_2|\mathcal{M}) = P(\mathbf{v}_1|\mathcal{M})P(\mathbf{v}_2|\mathcal{M})$$

- For different speech samples of the same speaker:

$$P(\mathbf{s}_1, \mathbf{s}_2|\mathbf{v}, \mathcal{M}) = P(\mathbf{s}_1|\mathbf{v}, \mathcal{M})P(\mathbf{s}_2|\mathbf{v}, \mathcal{M})$$

Ideal LR calculation

Vocal tract measurement

We cannot directly observe **v**. If asked:

> *What are the vocal tract parameters,* **v***, of the speaker who spoke* **s***?*

The answer is the posterior: $P(\mathbf{v}|\mathbf{s},\mathcal{M})$. A simple example is:

$$P(\mathbf{v}|\mathbf{s},\mathcal{M}) = \mathcal{N}(\mathbf{v}|\boldsymbol{\mu},\mathbf{C}), \qquad (\boldsymbol{\mu},\mathbf{C}) = f_{\mathcal{M}}(\mathbf{s})$$

where the function $f_{\mathcal{M}}$ is defined by $\mathcal{M}$. We can interpret:

- $\boldsymbol{\mu}$ as a point estimate for the value of **v**
- **C** as the uncertainty around this estimate.

Ideal LR calculation

Which hypothesis?

If given two speech samples as evidence, $e = (\mathbf{s}_1, \mathbf{s}_2)$, and we are asked:

> *Do $\mathbf{s}_1, \mathbf{s}_2$ come from the same speaker (H), or from two different speakers ($\neg H$)?*

The answer is the LR:

$$\frac{P(e|H,\mathcal{M})}{P(e|\neg H,\mathcal{M})} = \int_{\mathcal{V}} \frac{P(\mathbf{v}|\mathbf{s}_1,\mathcal{M})P(\mathbf{v}|\mathbf{s}_2,\mathcal{M})}{P(\mathbf{v}|\mathcal{M})}\, d\mathbf{v}$$

Ideal LR calculation

Comment

We have:

$$\mathrm{LR}_{\mathcal{M}} = \int_{\mathcal{V}} \frac{P(\mathbf{v}|\mathbf{s}_1, \mathcal{M}) P(\mathbf{v}|\mathbf{s}_2, \mathcal{M})}{P(\mathbf{v}|\mathcal{M})} \, d\mathbf{v}$$

Notice: $P(\mathbf{v}|\mathbf{s}_i, \mathcal{M})$ is just the answer to the vocal tract measurement problem. The LR automatically takes account of the uncertainty in the vocal tract measurement.

Compare this to DNA, where the measurement uncertainty is often assumed to be negligible. Then the LR calculation is based purely on the analogue of $P(\mathbf{v}|\mathcal{M})$, the rarity of the DNA profile in the population.

The problem

Ideal LR makes bad decisions

In some of the latest speaker recognition systems, we do use calculations of the form $\mathrm{LR}_{\mathcal{M}} = \int_{\mathcal{V}} \frac{P(\mathbf{v}|\mathbf{s}_1,\mathcal{M})P(\mathbf{v}|\mathbf{s}_2,\mathcal{M})}{P(\mathbf{v}|\mathcal{M})}\, d\mathbf{v}$.

The problem is, it doesn't work as intended! If we make Bayes decisions with the rule $\mathrm{LR}_{\mathcal{M}} > \frac{\text{cost ratio}}{\text{prior odds}}$, we find that on average the decisions are very bad.

The problem
Analysis

Why does $\text{LR}_{\mathcal{M}}$ make bad decisions? Because $\mathcal{M}$ does not model the data accurately enough.

Do we throw away $\mathcal{M}$ and start again? No:

- Sufficiently accurate modelling is too complex. If we try a different modelling approach, we are virtually guaranteed to have the same problem.
- $\text{LR}_{\mathcal{M}}$ is intrinsically very accurate—you can see this using ROC/DET-curves.
- $\text{LR}_{\mathcal{M}}$ is just badly calibrated.
- We can fix $\text{LR}_{\mathcal{M}}$ by re-calibration. (See last section).

The problem
Other technologies

With the (possible?) exception of DNA, other forensic technologies have the same problem.

If this problem has not been observed, it is because the goodness of Bayes decisions has not been empirically tested.

Outline

Defining calibration

"Calibration" can have two meanings:

1. A measure of goodness of an (automatic) system that provides outputs in LR form.
2. A procedure (i.e. re-calibration) for improving the above quality.

In this section, were are concerned with the measure of goodness.

Defining calibration
Classical definition

DeGroot & Fienberg, 1983: A weather forecaster is well calibrated, if:

- out of all the times he predicts rain with probability p%,
- rain does materialize p% of the time.

Defining calibration

More generally: A probabilistic model, $\mathcal{M}$, is well calibrated from the point of view of an evaluator, $\mathcal{E}$, if for every e:

$$P\Big(H\Big|\mathcal{E}, P(H|e,\mathcal{M}) = p\Big) = p$$

or more realistically:

$$P\Big(H\Big|\mathcal{E}, P(H|e,\mathcal{M}) = p\Big) \approx p$$

Defining calibration

More practically: $\mathcal{E}$ sees $\mathcal{M}$ as well calibrated if the expected KL-divergence between their posteriors is small:

$$\Big\langle \mathrm{KL}\big([q, 1-q]; [p, 1-p]\big) \Big\rangle_{P(e|\mathcal{E})}$$

where

$$q = P(H|\mathcal{E}, p), \qquad p = P(H|\mathcal{M}, e)$$

Defining calibration

Why KL-divergence?

- KL-divergence compares probability distributions.
- However (Dawid 1998, Brümmer 2010),

$$\mathrm{KL}(p;q) \in \text{a family of generalized divergences},$$

  which are formed by comparing the goodness of Bayes decisions made by either p or q.
- This is just what we need! Our application is to make Bayes decisions. So, let's use the whole family.

Defining calibration
In terms of decisions

We now define calibration in terms of Bayes decisions:

A well-calibrated probability model gives LRs and posteriors that make good Bayes decisions.

- Since LR-based Bayes decisions depend on priors and costs,
- a well-calibrated model should make good decisions for all priors and costs.

Outline

Calibration
Summary

In summary of this section:

- Good calibration $\equiv$ the ability to make good Bayes decisions.
- Open-loop probabilistic modelling tends to give badly calibrated LRs, that make bad Bayes decisions.
- This can be fixed with re-calibration.

To detect and fix calibration problems, we need to be able to measure goodness of calibration.

Outline

Measuring calibration
Background

Measuring LR calibration is easy—just use the LRs to make Bayes decisions. Why does this methodology not enjoy more widespread use?

- Perhaps calibration has been widely ignored on purpose?
- The Bayes decisions involve costs and prior. How can we test all combinations of costs and priors?

We elaborate on these points in the next two subsections.

Measuring calibration
Background

Most probabilistic recognizers are not naturally well calibrated; calibration is notoriously sensitive to dataset shift; some applications don't even require calibration; etc.

- This has motivated the development of an array of measurement methods that ignore calibration!
- Examples: ROC/DET curves; equal-error-rate (EER); CMC curves; precision-recall-break-even-point (PRBEP); etc.
- These are very valuable tools for basic algorithm development. But low EER does not guarantee good decisions.

Measuring calibration
Background

In the regular NIST Speaker Recognition Evaluations, there are two complementary evaluation criteria:

DET-curve: does not measure calibration; spans operating points by sweeping decision threshold.

DCF: weighted combination of false-acquit and false-convict error-rates; effectively measures calibration at a single operating point (weights are fixed); does not span operating points.

(In other biometric fields, HTER is similar to DCF, but with equal weights.)

Measuring calibration
Background

NIST's DET + DCF evaluation methodology is powerful and has driven speaker recognition research for more than a decade. But:

- It is not quite adequate for evaluation of LR calibration.
- We need a single criterion, which combines the benefits of DET and DCF into one—we need a tool that is calibration sensitive over a wide range of operating points.

In what follows, we construct such a criterion.

Empirical Bayes risk

Definition: Empirical Bayes risk evaluates the goodness of a system $\mathcal{S}$, which outputs $\text{LR}_{\mathcal{S}}(e)$ for every e in a supervised evaluation database, as:

$$\mathcal{R}(\mathcal{S}|\pi, C_{fa}, C_{fc}) = \pi C_{fa} P_{fa}(\theta) + (1-\pi) C_{fc} P_{fc}(\theta)$$

prior: $0 \leq \pi \leq 1$

costs of false-acquit and false-convict: $C_{fa}, C_{fc} > 0$

Bayes decision threshold: $\theta = \frac{C_{fc}}{C_{fa}} \times \frac{1-\pi}{\pi}$

error-rates: $P_{fa}(\theta)$ and $P_{fc}(\theta)$, when thresholding at $LR_S(e) > \theta$ for all e in the supervised evaluation database.

Empirical Bayes risk

The default system

Definition: The default system, $\mathcal{S}_{\text{def}}$, is the one that outputs $\text{LR}(e) = 1$, for any e.

The default system is equivalent to disregarding the evidence and making decisions with the prior, π, only. It has empirical Bayes risk:

$$\mathcal{R}(\mathcal{S}_{\text{def}}|\pi, C_{fa}, C_{fc}) = \min\Big(\pi C_{fa}, (1-\pi)C_{fc}\Big)$$

.

Empirical Bayes risk

The default system

The default system serves as a reference value for goodness of calibration:

- We consider calibration to have failed for a system $\mathcal{S}$, at operating point π, C_{fa}, C_{fc}, if:

$$\mathcal{R}(\mathcal{S}|\pi, C_{fa}, C_{fc}) > \mathcal{R}(\mathcal{S}_{\text{def}}|\pi, C_{fa}, C_{fc})$$

Empirical Bayes risk

The problem

The problem is:

- Bayes decisions are functions of costs and priors.
- $\text{LR}_{\mathcal{S}}$ should be able to make good decisions for all costs and priors, because these parameters vary from case to case.
- Bayes risk, $\mathcal{R}(\mathcal{S}|\pi, C_{fa}, C_{fc})$, measures at a fixed value of these parameters.

Can we instead test decisions made under all combinations of costs and priors? Yes. The solution is to simplify the evaluation criterion via equivalence of evaluation criteria.

Equivalence of evaluation criteria
Definition

Definition. Equivalence of evaluation criteria:

- Let $\mathcal{S}$ and $\mathcal{S}'$ be any systems under evaluation.
- Let $\mathcal{E}(\mathcal{S})$ and $\mathcal{E}'(\mathcal{S})$ be two different evaluation criteria.

$\mathcal{E}$ and $\mathcal{E}'$, are equivalent, if:

$$\mathcal{E}(\mathcal{S}) \leq \mathcal{E}(\mathcal{S}') \qquad \text{if and only if} \qquad \mathcal{E}'(\mathcal{S}) \leq \mathcal{E}'(\mathcal{S}')$$

Equivalence of evaluation criteria

This means:

Criteria $\mathcal{E}$ and $\mathcal{E}'$ are equivalent if they always agree when comparing the relative goodness of two systems.

- Relative goodness between systems under evaluation is sufficient for our purposes.
- Specifically, to judge calibration, we want to compare any system $\mathcal{S}$ to the default system, $\mathcal{S}_{\text{def}}$.

Error-rate and risk are equivalent

Theorem

Definition: The effective prior is:

$$\tilde{\pi} = \frac{\pi C_{fa}}{\pi C_{fa} + (1-\pi) C_{fc}}$$

Theorem: Criteria, $\mathcal{R}(\mathcal{S}|\tilde{\pi}, 1, 1)$ and $\mathcal{R}(\mathcal{S}|\pi, C_{fa}, C_{fr})$ are equivalent.

Proof: Their ratio, $\pi C_{fa} + (1-\pi) C_{fc} > 0$, is independent of $\mathcal{S}$. ■

Corollary: If $\mathcal{R}(\mathcal{S}|\tilde{\pi}, 1, 1) \leq \mathcal{R}(\mathcal{S}_{def}|\tilde{\pi}, 1, 1)$ for every $0 \leq \tilde{\pi} \leq 1$, then also $\mathcal{R}(\mathcal{S}|\pi, C_{fa}, C_{fr}) \leq \mathcal{R}(\mathcal{S}_{def}|\pi, C_{fa}, C_{fr})$ for all costs and priors.

Empirical Bayes error-rate

Definition

Definition. Empirical Bayes error-rate is $\mathcal{E}(\mathcal{S}|\tilde{\pi}) = \mathcal{R}(\mathcal{S}|\tilde{\pi}, 1, 1)$, so that:

$$\mathcal{E}(\mathcal{S}|\tilde{\pi}) = \tilde{\pi} P_{fa}\left(\frac{1-\tilde{\pi}}{\tilde{\pi}}\right) + (1-\tilde{\pi}) P_{fc}\left(\frac{1-\tilde{\pi}}{\tilde{\pi}}\right)$$

where P_{fa}, P_{fc} are the empirical error-rates obtained by the decision rule $\text{LR}_{\mathcal{S}}(e) > \frac{1-\tilde{\pi}}{\tilde{\pi}}$.

Notice $\tilde{\pi}$ plays 2 roles:

- It weights the error-rates.
- It forms the decision threshold.

Error-rate and risk are equivalent
Summary

In summary, to judge calibration:

- We can set costs to unity, without loss of generality, so that
- Bayes risk simplifies to Bayes error-rate.
- We can now perform a full evaluation by sweeping over the scalar range of operating points: $0 \leq \tilde{\pi} \leq 1$.

Below, we develop a graphical tool to do this.

Outline

Normalized Bayes error-rate plot

The default system

As noted, the default system serves as a reference value for goodness of calibration.

- The default error-rate is: $\mathcal{E}(\mathcal{S}_{\text{def}}|\tilde{\pi}) = \min(\tilde{\pi}, 1 - \tilde{\pi})$.
- We consider calibration to have failed for a system $\mathcal{S}$, at operating point $\tilde{\pi}$, if:

$$\mathcal{E}(\mathcal{S}|\tilde{\pi}) > \mathcal{E}(\mathcal{S}_{\text{def}}|\tilde{\pi})$$

Normalized Bayes error-rate plot

Let us now turn this criterion into a graphical analysis tool. The criterion is:

$$\tilde{\pi} P_{fa}(\tilde{\pi}) + (1 - \tilde{\pi}) P_{fc}(\tilde{\pi}) \leq \min(\tilde{\pi}, 1 - \tilde{\pi})$$

or, equivalently:

$$y = \frac{\tilde{\pi} P_{fa}(\tilde{\pi}) + (1 - \tilde{\pi}) P_{fc}(\tilde{\pi})}{\min(\tilde{\pi}, 1 - \tilde{\pi})} \leq 1$$

The LHS is the normalized Bayes error-rate. This will be the y-axis of the plot.

Normalized Bayes Error-Rate Curve

Y-axis amplification

The normalization, $\min(\tilde{\pi}, 1 - \tilde{\pi})$, amplifies the y-axis of the plot near $\tilde{\pi} = 0$ and $\tilde{\pi} = 1$, where the un-normalized Bayes error-rate becomes small to the point of invisibility.

We also need to amplify the x-axis.

Normalized Bayes Error-Rate Curve

X-axis amplification

With the probability scale, $0 \leq \tilde{\pi} \leq 1$, important parts of the plot would be compressed against 0 and 1 on either side.

- We expand the x-axis to the whole real line, by letting:

$$x = \text{logit}\,\tilde{\pi} = \log \frac{\tilde{\pi}}{1-\tilde{\pi}} \qquad \Leftrightarrow \qquad \tilde{\pi} = \frac{1}{1+\exp(-x)}$$

- The Bayes decision threshold is just:

$$\log \text{LR} > -x$$

The x-axis represents both effective prior-log-odds and log-likelihood-ratio threshold.

Normalized Bayes Error-Rate Curve

Examples

Below we show some examples of normalized Bayes error-rate plots, for:

- Synthetic data, to show the effects of deliberate miscalibration.
- LRs from three real automatic speaker recognition systems, from NIST SRE 2010.

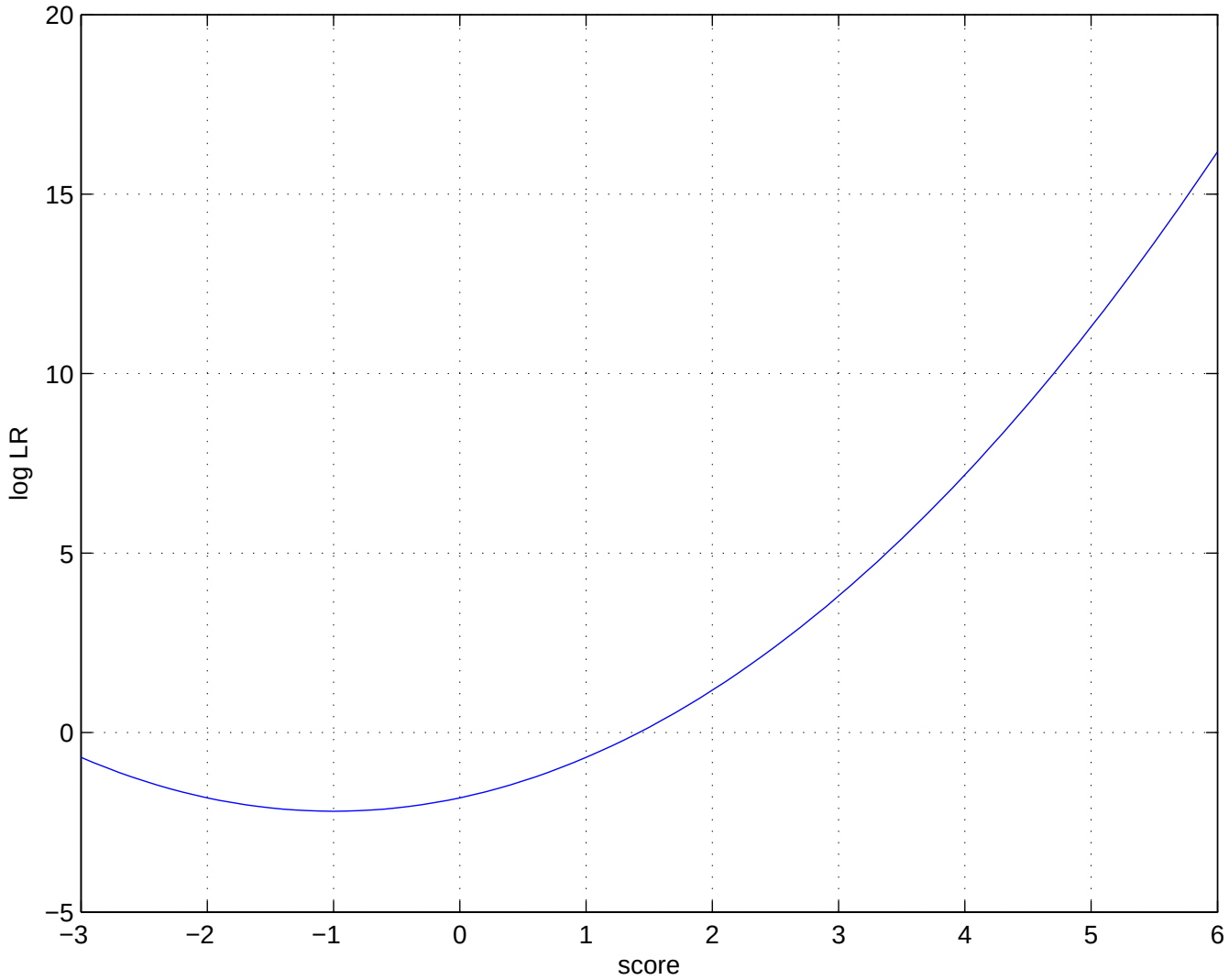
Synthetic example

- The evidence is a 'score', s, with likelihoods:

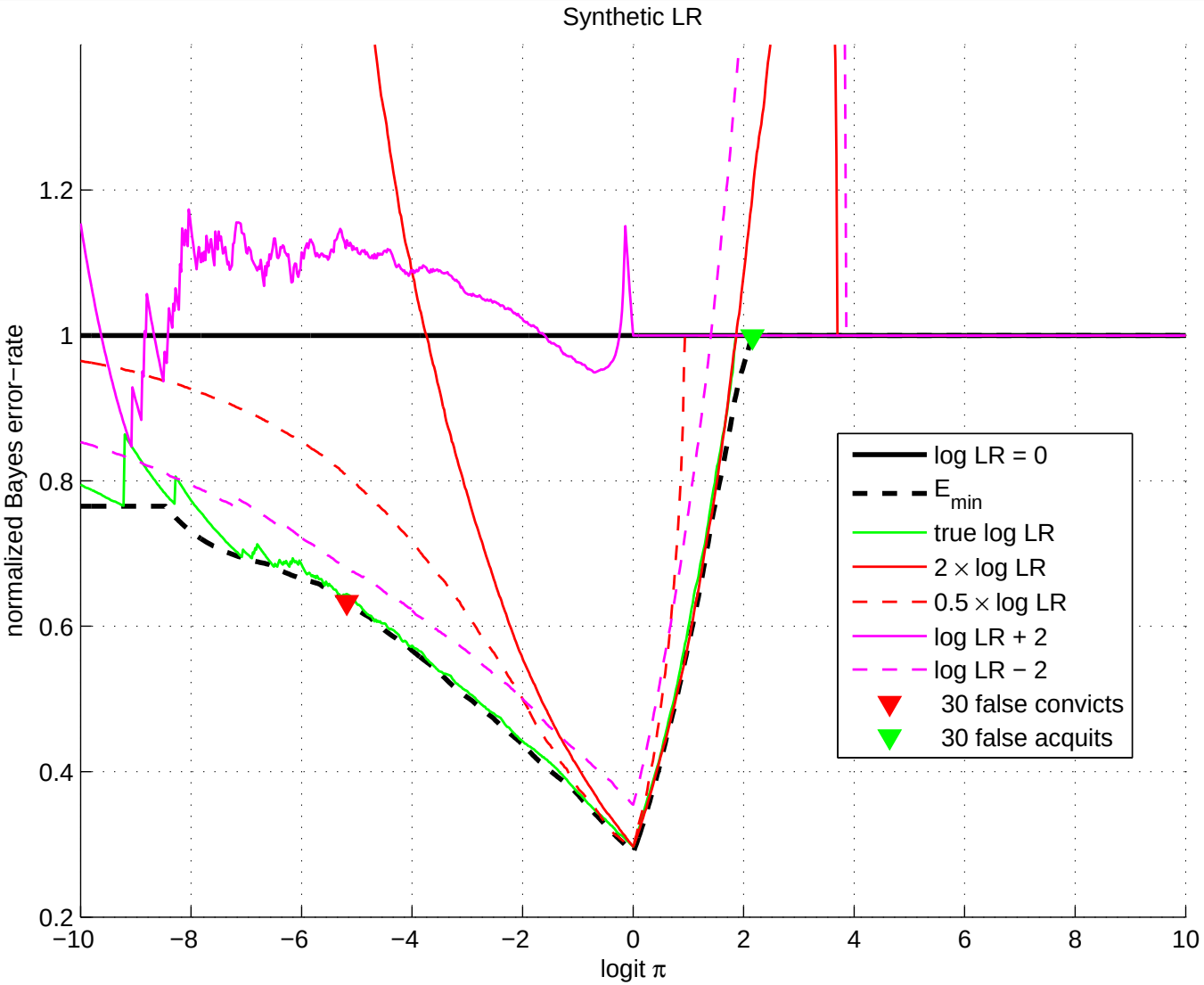
$$P(s|H) = \mathcal{N}(s|3,2), \qquad P(s|\neg H) = \mathcal{N}(s|0,1)$$

- $\log \mathrm{LR}(s)$ is a parabola, with turning point at $\log \mathrm{LR} \approx -2$. See plot.

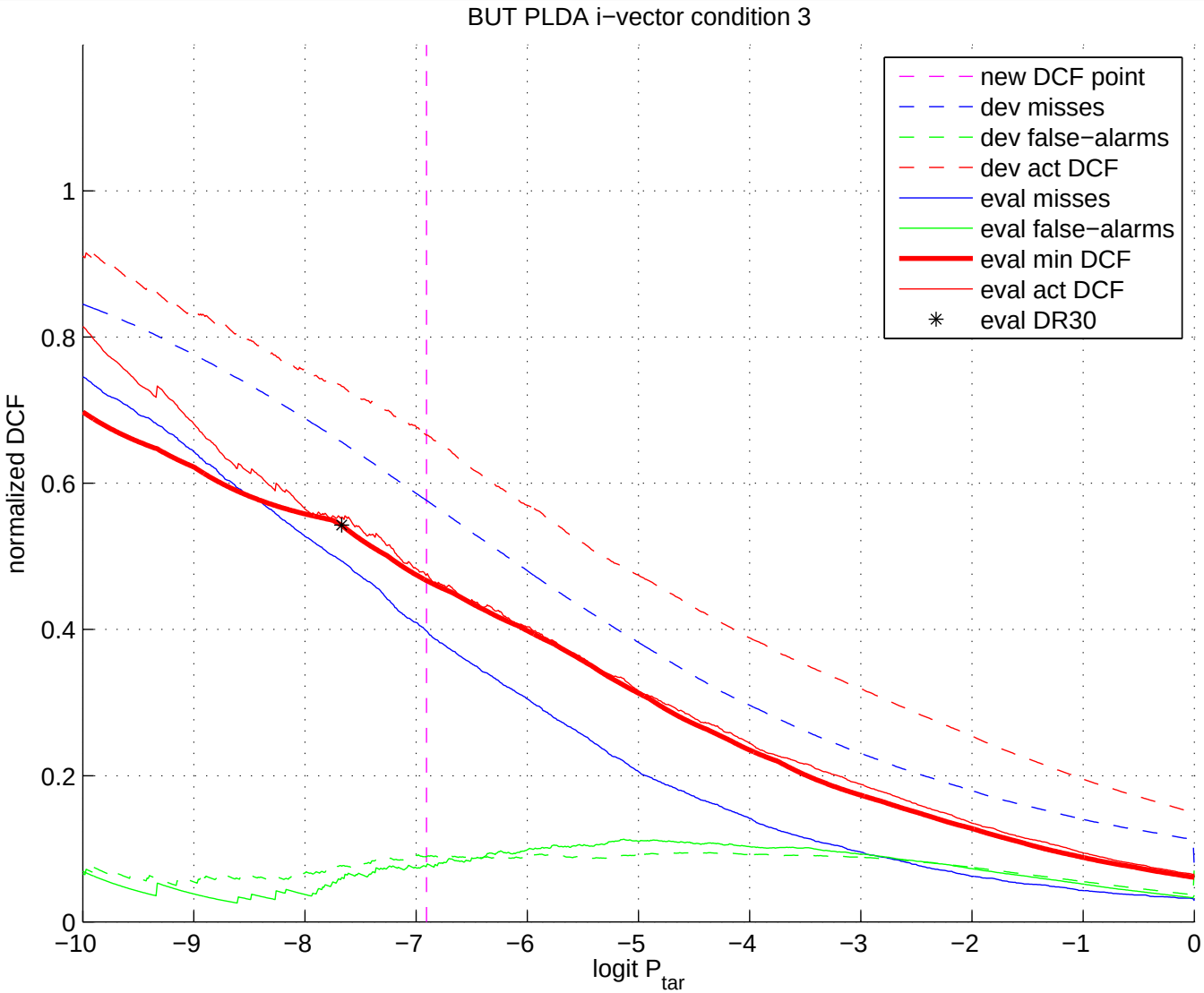
Log LR for synthetic Gaussian scores



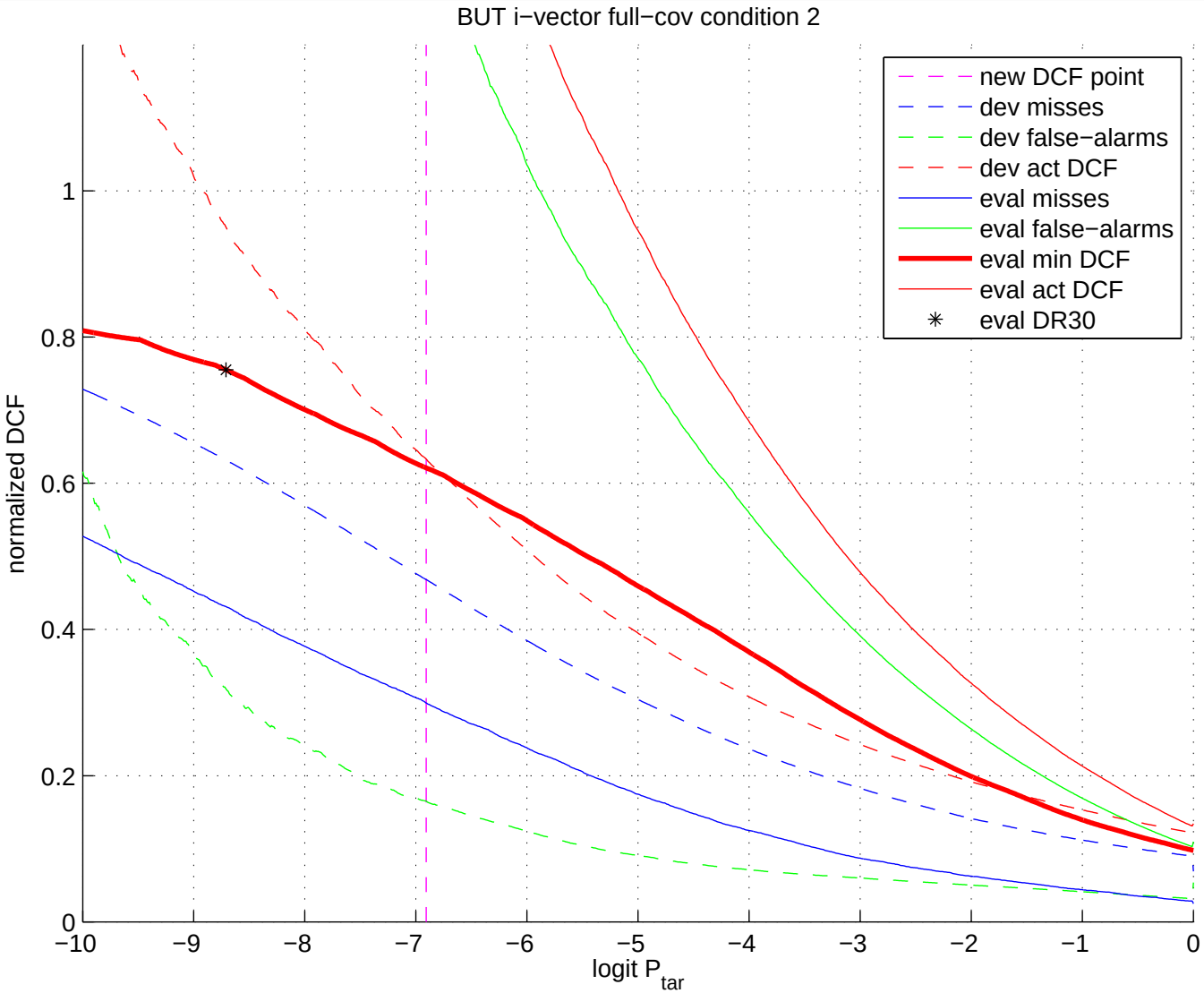
Normalized Bayes Error-Rate Plot



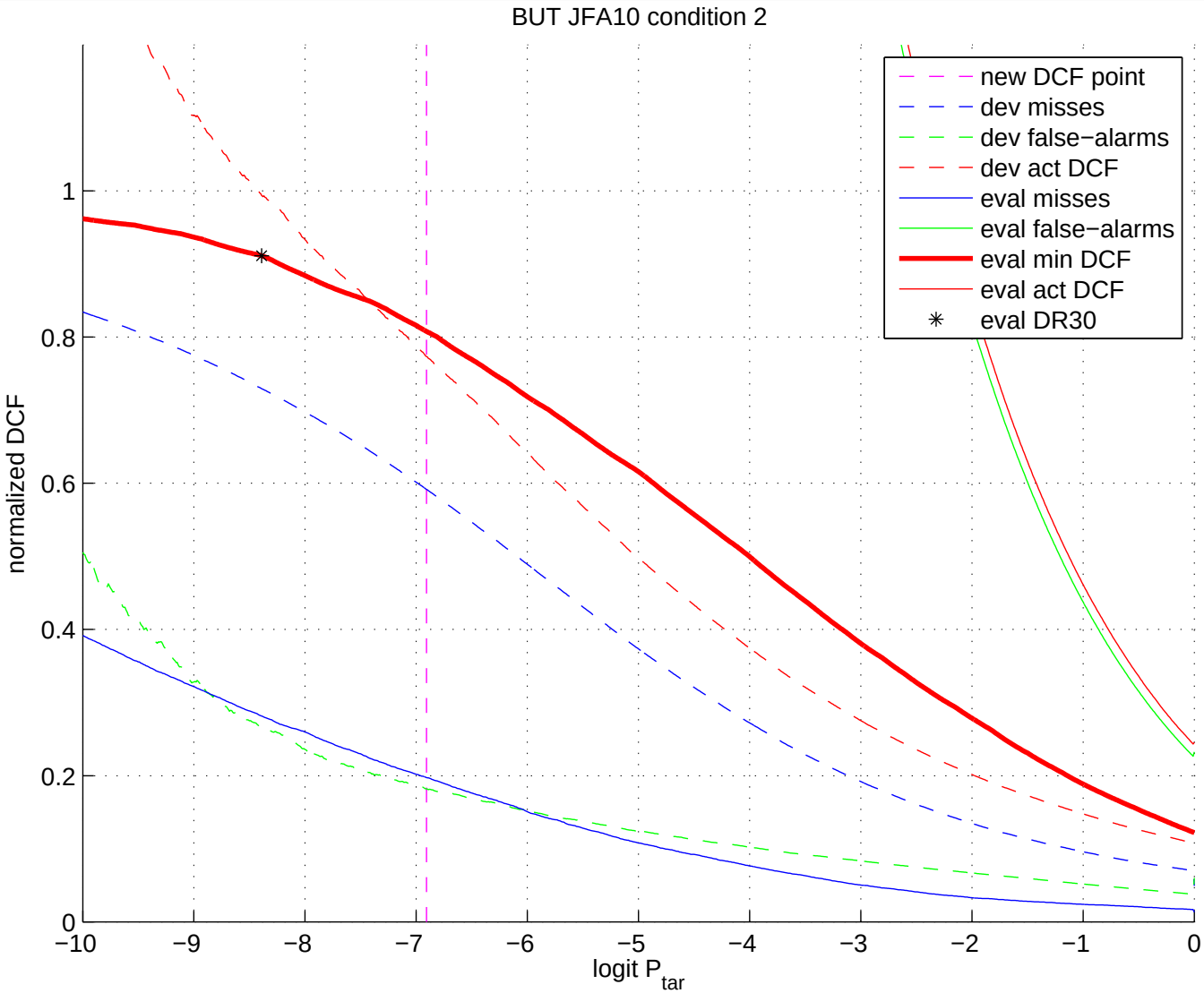
NIST SRE10: Example of Excellent Calibration



NIST SRE10: Example of Bad Calibration



NIST SRE10: Example of Worse Calibration



Measuring calibration
Summary

Normalized Bayes error-rate plots can do calibration analysis over all effective operating points of interest.

There is a freely available MATLAB toolkit to make these plots (see conclusion).

Outline

Re-calibration

Once you know how to measure calibration, you will most probably find the calibration of your system (or method) is bad at some operating points.

How can we make this better?

Re-calibration via probabilistic modelling

Bad calibration of LRs is due to inaccurate probabilistic modelling of the original evidence.

Calibration can be improved by an extra, simpler modelling step.

Re-calibration via probabilistic modelling

Score modelling

'Demote' the original badly calibrated LRs and just call them scores. Re-calibration is just probabilistic modelling of the scores.

- One expects larger scores to favour H and smaller scores to favour $\neg H$, but
- one does not expect scores to be able to make good decisions at the Bayes decision threshold.
- The score is a statistic, or a feature that we extract from the original evidence,
- which is easier to model than the original evidence.

Re-calibration
Score modelling

Uncalibrated score, from complex, difficult to calibrate, evidence model, $\mathcal{M}$:

$$s = \frac{P(e|H, \mathcal{M})}{P(e|\neg H, \mathcal{M})}$$

Re-calibrated LR, from simpler, easier to calibrate, score model $\mathcal{M}'$:

$$\text{LR} = \frac{P(s|H, \mathcal{M}')}{P(s|\neg H, \mathcal{M}')}$$

Probabilistic score modelling
Flavours

Various flavours of probabilistic score modelling are possible:

- parametric vs non-parametric
- generative vs discriminative
- plug-in vs fully Bayesian

I'm currently researching fully Bayesian methods, but my favourite workhorse is still logistic regression, which

- is a parametric, discriminatively trained, plug-in model.
- It is theoretically crude, but works very well in practice.

Probabilistic score modelling

Parameter learning, data requirements

Whatever the modelling paradigm, the model has unknown parameters, which have to be learnt from a database, which should be:

- large enough,
- (at least partially) supervised,
- (ideally) independent of data that was used to learn scoring model parameters
- independent of test data that will be used to judge the final calibration

Re-calibration
Summary

In summary:

- Demote the result of the more complex evidence model, from LR to uncalibrated score.
- The score may be uncalibrated, but nevertheless contains good discriminative information,
- which can be extracted by a subsequent modelling stage,
- which can be much simpler and therefore accurate enough to be well-calibrated.

Conclusion

Calibration is the ability to make good Bayes decisions.

- State-of-the-art evidence models (even when they give good discrimination) are usually not well-calibrated.
- It is important to test (i.e. measure) calibration, e.g. with Bayes error-rate on a supervised evaluation database.
- Bad calibration can be fixed by re-calibration, i.e. data-driven, probabilistic score modelling.

Please consider applying these ideas also to other forensic/biometric technologies. See below for MATLAB toolkit.

Conclusion
BOSARIS Toolkit

A MATLAB implementation of most of the techniques described in this talk are available in the freely available BOSARIS Toolkit.

- Calibration analysis: Normalized Bayes error-rate plots
- Discrimination analysis: ROC/ROCCH/DET curves.
- Re-calibration: non-parametric (PAV), or parametric (logistic regression).
- User guide, with detailed explanation of normalized Bayes error-rate plots.

`http://sites.google.com/site/bosaristoolkit`

References

- Niko Brümmer, *Measuring, refining and calibrating speaker and language information extracted from speech*, Ph.D. dissertation, University of Stellenbosch, October 2010.
- A. Philip Dawid, "Coherent measures of discrepancy, uncertainty and dependence, with applications to Bayesian predictive experimental design," Technical Report 139, Department of Statistical Science, University College London, Aug. 1998, Online: www.ucl.ac.uk/Stats.
- Morris H. DeGroot and Stephen E. Fienberg, "The comparison and evaluation of forecasters," *The Statistician*, vol. 32, pp. 14-22, 1983.