Daniel Ramos,[1] Ph.D.; Joaquin Gonzalez-Rodriguez,[1] Ph.D.; Grzegorz Zadora,[2] Ph.D.; and Colin Aitken,[3] Ph.D.

Information-theoretical assessment of the performance of likelihood ratio computation methods[*]

[1] ATVS – Biometric Recognition Group and Forensic Science and Security Institute (ICFS), Universidad Autonoma de Madrid (Spain).

[2] Institute of Forensic Research, Westerplatte 9, 31-033 Krakow, Poland.

[3] School of Mathematics and Joseph Bell Centre for Forensic Statistics and Legal Reasoning. University of Edinburgh, King's Buildings, Edinburgh, EH9 3JZ, UK.

[*] This work has been supported by the Spanish Ministry of Science and Innovation under project TEC2009-14719-C02-01, and co-funded by the Universidad Autonoma de Madrid and the Comunidad Autonoma de Madrid under project CCG10-UAM/TIC-5792.

ABSTRACT

Performance of likelihood ratio (LR) methods for evidence evaluation has been represented in the past using *e.g.* Tippett plots. We propose Empirical Cross-Entropy (ECE) Plots as a metric of accuracy based on the statistical theory of proper scoring rules, interpretable as *information* given by the evidence according to Information Theory, which quantify calibration of LR values. We present results with a case example using a glass database from real casework, comparing performance both with Tippett and ECE plots. We conclude that ECE plots allow clearer comparisons of LR methods than previous metrics, allowing a theoretical criteria to determine if a given method should be used for evidence evaluation or not, which is an improvement over Tippett plots. A set of recommendations for the use of the proposed methodology by practitioners is also given.

Keywords: forensic science; evidence evaluation; likelihood ratio; empirical cross entropy; performance assessment; glass evidence; Information Theory.

Statistical procedures for the evaluation of evidence are at the core of modern forensic science. Scientific methodologies based on databases and statistical analyses are becoming increasingly popular. This is in agreement with the idea that, first, changes in the law and, second, the evidence of errors in disciplines assumed as error-free are motivating a fundamental shift in all procedures followed in forensic science. These ideas are being more accepted by the scientific community of forensic experts, practitioners and statisticians (1)(2)(3). Among the legal factors which have inspired more movements in the field are the American Daubert rules (4), which include the need for common procedures, scientific methods and a clear assessment of results (specifying potential performance in operational conditions) for any piece of evidence to be admitted by the U. S. Supreme Court. This is moving many disciplines to a critical change in procedures from reporting conclusions based on non-repeatable and subjective arguments mainly based on *the experience of the expert* to a probabilistic assessment of the value of the evidence based on representative databases and statistical analysis.

In this changing paradigm, statistics plays a fundamental role for the establishment of scientific procedures. In the past, forensic statistics were focused on assumptions of uniqueness, distance measures and hypothesis testing (5). However, Bayesian methods for evidence evaluation are increasingly popular, since their first statement as a result of the *Dreyfus case* (6). Because of pioneering works such as (7), likelihood ratios LR began to be used for the evaluation of forensic evidence. Thanks to this so-called *LR methodology*, the forensic expert is able to measure statistically the *value* of the evidence. The LR expresses the degree of *support* of the evidence to the relevant propositions present in a given case (8).

Although statistical methods in general and the LR methodology in particular have been fundamental for the establishment of a rigorous framework for forensic evidence evaluation, they do not guarantee the *quality* of the methods in use. There are many problems which can lead to comparisons yielding LR values which provide support, sometimes strong support, for the wrong proposition. Examples of these problems are the variability of the evidential material, which can lead to erroneous models developed from a population which does not represent the control or recovered samples; or the sparsity of the data, which can lead to erroneous models if their robustness to data sparsity is poor. This would lead to evidence evaluation methods which are misleading to the court, in the sense that LR values will tend to support the wrong proposition in a case. Although this is a situation to be avoided, the extent to which this may happen should be assessed and documented by the forensic scientist, in agreement with the spirit of the Daubert rules in the USA concerning the scientific assessment of the performance of the methods; and also with proposals being made in other countries (2)(3)(4)(9).

In this article we propose a framework for the scientific assessment of the performance of forensic evidence evaluation methods which express the value of the evidence in the form of a LR. This framework is based on Information Theory (10), a field which allows an intuitive interpretation of the results of the scientific assessment, which is a highly desirable characteristic when results have to be reported to a court of law. The proposed framework, based on a performance metric called *Empirical Cross-Entropy* (ECE), can be used with any LR-based evidence evaluation method at any level of the propositions stated in the case (source, activity or offence) (11). The main contribution of this work is the full description of the assessment framework in a forensic context, and its generalisation to any forensic discipline.

A convenient graphical representation of the performance is also proposed, namely ECE plots. Although ECE plots have been introduced in (12) in the context of forensic speaker recognition, this article presents novel contributions with respect to such work, with a significant extension at the methodological, application and experimental level. First, we generalise the use of ECE to any forensic field, not only forensic automatic speaker recognition. This represents a fundamental advance in the applicability of the proposed framework, since each different forensic field presents particular types of data and models. In particular, automatic speaker recognition yields continuous, univariate data. On the other hand, glass analysis, as presented here, generates multivariate data. Secondly, this article contributes a set of recommendations for practitioners, which simulate typical scenarios for the use of the proposed assessment tools in daily forensic casework. To this end, the authors have developed publicly available free software implementing ECE plots as proposed in this article (available at http://arantxa.ii.uam.es/~dramos/software.html). The use of this methodology is exemplified by a case example using glass profiles obtained from real forensic cases, where several evidence evaluation methods are compared prior to their application to the case, in order to illustrate the recommendations given.

A remark is in order here. Although we give recommendations for the use of ECE in court, we realize that meaningful interpretation of its value in a legal process seems currently unrealistic until a deeper understanding of the Bayesian approach has been established across all the actors participating in a forensic case. Nevertheless, as it has been stated, ECE presents many other advantages as a performance metric that justify its use for validation purposes in forensic laboratories, even though its use in court reporting currently does not seem to be feasible.

This article is organised as follows. First, the LR methodology for evidence evaluation is described. Secondly, the problem of the scientific assessment of LR-based evidence evaluation methods is introduced, with the presentation of several performance metrics, and the illustration of their main properties and drawbacks. Thirdly, the proposed information-theoretical assessment framework is described, the ECE metric is introduced, its interpretation is stated and its main properties are highlighted. This part also includes the description of the proposed representation of performance, namely the ECE plot, and the algorithms used for it. Fourthly, the proposed recommendations for forensic practitioners are described. A case example is discussed next. Finally, conclusions are given.

Evaluation of the evidence using likelihood ratios

The LR framework for forensic evidence evaluation is stated as follows. Consider the forensic evidence *E*, which includes a *recovered* sample of unknown origin and a *control* sample whose origin is known. The LR expresses the degree of support of the evidence to any of the relevant propositions in the case relative to one another. An example of a pair of propositions at source level (11) (13) is:

- θ_p (also known as the *prosecution* proposition): glass fragments recovered from a suspect's jacket come from a broken window in a burglary.
- θ_d (also known as the *defence* proposition, or the *alternative* proposition): glass fragments recovered from a suspect's jacket do not come from the broken window in the burglary.

The LR is then computed by the forensic examiner from the comparison of measurements of the *control* and recovered materials, and with measurements from a so-called *population database*, representing a *relevant population* for the case (13). For instance, for the pair of propositions above, the relevant population may consist of glass fragments from broken windows of the same type as the broken window in the burglary of the case, which will constitute a so-called *population database*.

In a forensic case, the unobserved variable of interest is the true proposition, $\theta=\{\theta_p,\theta_d\}$, because the fact finder ultimately wants to know its true value. These possible values θ_p and θ_d of the variable θ are complementary within the relevant population. The decision of the fact finder is based on the probability of a given value of θ, conditioned on all the available information in the case. Bayes' theorem relates probabilities before and after the analysis of the evidence:

$$O\left(\theta_p \middle| E,I\right)=\frac{P\left(E \middle| \theta_p,I\right)}{P\left(E \middle| \theta_d,I\right)}\times O\left(\theta_p \middle| I\right) \tag{1}$$

where P denotes probability and I is the background information available in the case apart from the evidence E. The *posterior odds* are defined as the ratio of complementary posterior probabilities, namely $O\left(\theta_p \middle| E,I\right) = \dfrac{P\left(\theta_p \middle| E,I\right)}{P\left(\theta_d \middle| E,I\right)}$; and the *prior odds*, province of the fact finder, are defined as $O\left(\theta_p \middle| I\right) = \dfrac{P\left(\theta_p \middle| I\right)}{P\left(\theta_d \middle| I\right)}$. The likelihood ratio (LR), province of the forensic examiner (14), is defined as

$$LR = \frac{P\left(E \middle| \theta_p, I\right)}{P\left(E \middle| \theta_d, I\right)} \tag{2}$$

It can be easily seen that:

$$P\left(\theta_p \middle| E,I\right) = \frac{LR \times \dfrac{P\left(\theta_p \middle| I\right)}{P\left(\theta_d \middle| I\right)}}{1 + LR \times \dfrac{P\left(\theta_p \middle| I\right)}{P\left(\theta_d \middle| I\right)}} = \frac{LR \times O\left(\theta_p \middle| I\right)}{1 + LR \times O\left(\theta_p \middle| I\right)}$$

(3)

Empirical assessment of the performance of LR-based evidence evaluation methods

Validation databases

A majority of assessment approaches of the performance of evidence evaluation methods, and in particular the ones reviewed and proposed here, are of an empirical nature. Prior to the use of any evidence evaluation method in casework, the forensic scientist should assess the performance of the evidence evaluation method to be used, according to the requirements of modern forensic science. In order to do so, typically a number of hypothetical cases will be simulated in conditions as alike as possible as for a typical case where the evidence evaluation method will be used. We define a *validation* database as the data needed to simulate those hypothetical cases. We use the term *validation* because the performance to be assessed will typically be considered for the validation of the evidence evaluation method prior to its use in casework. The true origin of the data must be known for the validation database, and therefore, in each hypothetical case is known whether θ_p or θ_d is actually true (ground-truth).

It is important to distinguish between the population database and the validation database. The population database is used in a given case to model the variation of the evidential materials in each case in populations in order to compute the LR, as explained above. On the other hand, the validation database consists of hypothetical control and recovered samples that are used to simulate hypothetical cases in conditions similar to those in which the evidence evaluation method will be used. In fact, for each hypothetical case simulated using the validation database, for which a LR value is computed, a different population database may be used. However, these population databases used in the hypothetical cases will typically mimic those used in cases where the evidence evaluation will be used.

Once a validation database has been selected, a number of LR values may be generated for all the defined hypothetical cases. Denote this set of LR values generated from the hypothetical cases as a *validation set of LR values*. Those LR values will then be *representative* of the cases in which that method of evidence evaluation will be used, in the sense that the conditions of the evidence and the population in those cases and in the hypothetical cases were comparable.

Depending on which of the values of the unknown proposition variable θ is true in each hypothetical case, the corresponding comparisons in each hypothetical case will be respectively referred to as *true-*θ_p and *true-*θ_d comparisons. Similarly, the validation LR resulting from the comparisons will be respectively referred to as true-θ_p and true-θ_d LR values. There will be N_p true-θ_p LR values and N_d true-θ_d LR values in the validation set of LR values.

It is important to highlight the fact that the validation of a given evidence evaluation method will be typically carried out before that method is used in a given case. Once the validation of the method has been performed, then the method can be used in subsequent cases in order to present results to a court, and the performance measured can be also reported for the sake of transparency. In relation to this, a validation method should consider the conditions of the materials to compare in the evidence evaluation process, of the population to be used, etc. For instance, if a validation process has been conducted for an evidence evaluation method, and a database of chemical profiles of glass from building windows has been used, it may not be an appropriate method for evidence in the form of chemical profiles of glass from containers. If so, another validation process should be considered before the use of the method in a case involving the latter. These considerations are of critical importance in the context of the validation of evidence evaluation methods, but they are outwith the scope of this article.

Once a set of true-θ_p and true-θ_d LR values have been generated with the validation database, a measure of performance is computed from those values, this measure will be representative of the performance of the evidence evaluation method in casework. Some of the most common ones are reviewed below.

False positive and false negative error rates

False positive and false negative error rates are a common measure for evaluating decisions in forensic science according to decision theory (15), see *e.g.* (16). In a LR context, it may be thought that a false negative error occurs when LR<1 for a true-θ_p comparison, and a false positive error occurs when LR>1 for a true-θ_d comparison. Under this approach the decision would be based only on the LR value. However, decisions in favour of θ_p or θ_d, have to be based on the posterior probabilities $P\left(\theta_p \middle| E,I\right)$ and $P\left(\theta_d \middle| E,I\right)$, which represent the probabilistic opinion about the propositions considering *all* the relevant information of the case (15). Thus, decisions can only be made if the prior odds are known, which is not the case of the forensic examiner in general. In consequence, false positive and false negative rates should not be used as a metric for performance for LR in forensic science. Even though these error rates were represented as Receiver Operating Characteristic curves, they are measures of discriminating power, as we will highlight below, and they are incomplete for the LR framework.

Tippett plots

Tippett plots have been classically used for empirical performance assessment. First used by (17) based on the work by (18), for each value of the logarithm of the likelihood ratio $\log_{10}(\mathrm{LR})$, two values are plotted: the proportion of true-θ_p and of true-θ_d LR values, respectively, in the validation set that are greater than a given $\log_{10}(\mathrm{LR})$ value. Tippett plots also show the so-called *rates of misleading evidence*, defined as the proportion of LR values giving support to the wrong propositions (LR>1 when θ_d is true and LR<1 when θ_p is true). Note that these are similar to false positives and false negatives for a decision threshold at $\log_{10}(\mathrm{LR})=0$. An example of Tippett plots is shown in Figure 1, where the rates of misleading evidence for true-θ_p and true-θ_d LR values are highlighted.

FIGURE 1 - Example of Tippett plots showing the proportion of true-θ_p and true-θ_d LR values in the validation set greater than a given value.

Although Tippett plots are useful and show many important performance indicators for a given validation set of LR values, we identify several problems:

- Comparison among methods presenting different Tippett plots is sometimes difficult. An example of this situation may arise when a forensic expert wants to compare two different background populations to compute a LR value for the comparison of two given fragments of glass (19). In the example in Figure 2, method M_1 presents a lower rate of misleading evidence than M_2 when θ_d is true, but a higher rate when θ_p is true. Moreover, the magnitudes of such rates of misleading evidence are not equal when θ_p is true and when θ_d is true. Under these circumstances, it is difficult for the scientist to decide which method is preferred for a given forensic case.

FIGURE 2: Example of Tippett plots for two different methods of evidence evaluation.

- The impact of misleading evidence is not explicitly measured. It is known that LR values much smaller than 1 when θ_p is true or much bigger than 1 when θ_d is true are undesirable, since they represent *strong* misleading evidence (20). Such LR values should have a higher negative impact than values of LR near LR=1. However, in Tippett plots this impact is not numerically measured.
- It is difficult to visualise the discriminating power, defined as the measurement of the degree of *separation* among true-θ_p and true-θ_d LR values. Discriminating power has been measured in the past by Receiver Operating Characteristic (ROC) or Detection Error Tradeoff (DET) curves (21), and its importance is fundamental for an assessment of the performance of evidence evaluation methods (22), (23). It is not easy to comparatively determine the discriminating power of a given technique

from a Tippett plot. An example of this is given in Figure 3, where two methods M_3 and M_4 with the same discriminating power present very different Tippett plots. The discriminating power of both methods is exactly the same, because the LR values computed with M_4 are just a linearly scaled version of those computed with M_3, a transformation which does not change the discriminating power of a set of LR values (22)(23).

FIGURE 3 - Example of Tippett plots for two different evidence evaluation methods having the same discriminating power.

Assessment of posterior probabilities: calibration and refinement

The concept of performance assessment of posterior probabilities is not new in the statistics literature (24)(25). In (25) it was introduced in order to evaluate and compare posterior probabilities in the context of weather forecasting. There, posterior probabilities were used as degrees of belief about a given proposition (for instance θ_r: *tomorrow it will rain*) against its opposite (for instance θ_{nr}: *tomorrow it will not rain*). This problem can be viewed as equivalent to a forensic case as proposed here, considering posterior probabilities $P\left(\theta_p \middle| E,I\right)$ and $P\left(\theta_d \middle| E,I\right)$ from Equation 1.

The quality of such a forecaster can be assessed by means of *strictly proper scoring rules*. An example of a strictly proper scoring rule is the logarithmic scoring rule. See

(24) for more examples of strictly proper scoring rules. Given the evidence variable E in a forensic case, the logarithmic scoring rule takes the following values:

$$\begin{array}{ll} \theta_p \text{ true:} & -\log_2 P\left(\theta_p \middle| E, I\right) \\ \theta_d \text{ true:} & -\log_2 P\left(\theta_d \middle| E, I\right) \end{array} \tag{4}$$

where θ_p and θ_d are defined as in Section 2. The base of the logarithms is just a scaling factor, and will not have influence in the information-theoretical framework derived below. Here we take base-2 logarithms for convenience and coherence with respect to information-theoretical literature (10). As expressed here, strictly proper scoring rules may be viewed as loss functions which assign a penalty to a given value of the posterior probability. In this context, the penalty is the value of the rule in Equation 3 and depends on: *i)* the value of the posterior probabilities, and *ii)* the true value of the proposition variable θ. For example, if a probabilistic forecast, expressed as a posterior probability, of raining tomorrow is high (value of the forecast) and tomorrow it does not rain (true value of the proposition variable), a strictly proper scoring rule will assign a high penalty to the forecast, and vice-versa.

The overall measure of performance of a forecaster is defined as the average value of a strictly proper scoring rule over many different forecasts, for which the actual value of the proposition variable is known (24)(25). This is equivalent to a validation set of true-θ_p and true-θ_d posterior probabilities. For instance, for the logarithmic scoring rule, this average would be the so-called *logarithmic loss* (L):

$$L = -\frac{1}{N_p}\sum_{i\in\text{true}-\theta_p}\log_2 P\left(\theta_p \middle| E_i, I\right) - \frac{1}{N_d}\sum_{j\in\text{true}-\theta_d}\log_2 P\left(\theta_d \middle| E_j, I\right) \qquad (5)$$

where N_p and N_d is the number of true-θ_p and true-θ_d forecasts in the validation set, and E_i and E_j are particular evidences for each of the forecasts in the validation set. Moreover, it is also demonstrated in (25) that such a measure of performance can be divided into two components:

i. A *calibration loss* component, which measures how probabilistically interpretable are the values of the forecasts obtained (25). Low calibration loss means that for a given range of values of the forecast closely around $P\left(\theta_p \middle| E, I\right) = \rho$, then the proportion of cases where $\theta = \theta_p$ in the validation set tends to be ρ, and similarly where $\theta = \theta_p$.

ii. A *refinement loss* component, which measures how discriminating the forecasts are. According to (25), and roughly speaking, low refinement loss means that if the calibration loss of the forecaster is low, for a given value of the forecast $P\left(\theta_p \middle| E, I\right)$ the relative frequency of trials where $\theta = \theta_p$ in the validation set is either near 0 or near 1, and similarly where $\theta = \theta_p$. Refinement loss is related to the loss of performance due to a non-perfect discriminating power of the LR values in the validation set (23).

Thus, the aim of a good evidence evaluation method will be to reduce the overall loss which consists of calibration loss plus refinement loss. It is important to highlight that calibration and refinement loss are not explicitly separated in Equation 5. In (25), a decomposition of Equation 5 into refinement and calibration loss is given for the case where the values of the forecasts in the validation set are discrete (or discretized). However, this is not general. In this work, and following (23), we propose the use of an algorithmic solution, as explained below.

Information-theoretical assessment of the performance of evidence evaluation methods

An assessment framework is proposed which clearly shows the overall performance of the LR-based evidence evaluation methods under analysis in terms of their calibration and refinement components. Moreover, the proposed methodology has an attractive information-theoretical interpretation. Information theory is a wide area of knowledge which was proposed in the middle of the twentieth century as a framework for measuring and presenting information (26). After more than 50 years, the applications of information theory have been remarkable in many fields like physics, probability theory and economics (10). Under this framework, the uncertainty about an unknown variable is quantified by a magnitude called *entropy*, which is a function of a given probability distribution. It is known that probability is the best way of stating uncertainty about a given value of a variable, and entropy is a function of the probability distribution: it measures with a single number the amount of uncertainty in a probability distribution (10)(26). Additional knowledge about the variables under study will give additional *information* about the unknown variable, and thus it will contribute to the reduction of the entropy. As a simple example, imagine that a fair coin is tossed, with probabilities of obtaining heads equal to the probability of obtaining tails, and equal to 0.5. Under these circumstances, there will be maximum uncertainty about the outcome of the coin toss, and the entropy will be maximal. If we know that the coin is biased, the entropy of the probabilities of head or tails will reduce. In the limit, if we know that the coin has two heads, it is certain that the outcome of the coin toss will be heads, and the entropy will be zero, because the uncertainty is null.

The proposed information-theoretical measure of performance of a validation set of LR values, namely *Empirical Cross-Entropy* (ECE) is stated as follows:

$$ECE=-\frac{P\left(\theta_p\right)}{N_p}\sum_{i\in\mathrm{true}-\theta_p}\log_2 P\left(\theta_p\middle|E_i\right)-\frac{P\left(\theta_d\right)}{N_d}\sum_{j\in\mathrm{true}-\theta_d}\log_2 P\left(\theta_d\middle|E_j\right) \qquad (6)$$

where E_i and E_j denote the evidence in each of the comparisons in the validation set where θ_p or θ_d are respectively true. For simplicity, hereafter I will be eliminated from the notation, but it has to be remembered that it conditions every probability value. From Equation 3, it is straightforward that Equation 6 is equivalent to the following expression:

$$ECE=\frac{P\left(\theta_p\right)}{N_p}\sum_{i\in\mathrm{true}-\theta_p}\log_2\left(1+\frac{1}{LR_i\times O\left(\theta_p\right)}\right)+\frac{P\left(\theta_d\right)}{N_d}\sum_{i\in\mathrm{true}-\theta_d}\log_2\left(1+LR_j\times O\left(\theta_p\right)\right)(7)$$

where LR_i and LR_j in Equation 7 denote a single LR value in the validation set for which θ_p or θ_d are respectively true. By comparison of Equations 5 and 6, it is easily seen that ECE is the average of the logarithmic scoring rule over all the posterior probabilities that would be obtained from the comparisons in the validation set, with an additional weighting term given by the prior probabilities. Therefore, it is a proper measure of performance of posterior probabilities, according to (25).

However, as it is explicitly stated in Equation 7, ECE depends on the prior odds, and it is not possible in general for the forensic scientist to compute its value for a given particular case, because the prior probabilities in such a case are the province of the fact finder. However, it is possible for the forensic scientist to compute and represent ECE for a range of values of the prior probability. Thus, the exact value of the prior may be taken as an unknown parameter by the forensic scientist, and ECE can be represented in a prior-dependent way. An example of such a representation can be seen in Figure 4. We use base-10 logarithms for the prior odds because they are typically used for evidence evaluation. However, base-2 logarithms will be used for information-theoretical values, because they are commonly used in this field. Moreover, the N_p and N_d values used in the computation of ECE should be indicated, in order to give an idea of the balance of the comparisons in the validation LR set, especially if there is a large difference in them.

FIGURE 4 - ECE with respect to the base-10 logarithm of the prior odds. The lower its value, the better the performance of the evidence evaluation method in the given validation set of LR values. This performance will consist of discriminating power plus calibration.

ECE has the following interpretation: it represents the mean additional information, after consideration of the evidence, that the fact finder still needs in order to know the true value of the proposition variable θ. This mean value is computed as an average over the validation set of LR values. If the LR values given by the evidence evaluation method are misleading to the fact finder, then the ECE will increase, and more *information* on average will be needed to know the true value of the proposition variable θ. On the other hand, if the LR values given by the evidence evaluation methods tend to support the correct hypothesis, then ECE will decrease, representing the fact that the amount of *information* about the true value of the proposition θ in the given case has been improved. The term *information* here has a meaning of reduction of uncertainty, in accordance to Information Theory. Thus small values of ECE are good in the sense that less additional information is needed in order to determine the true value of θ.

A detailed formal derivation of ECE, with a justification of its interpretation, can be found in (22) (Chapter 6, Section 6.4).

Optimising ECE of a validation set of LR values: the PAV algorithm

It is important to know the decomposition of ECE into refinement loss and calibration loss. Among other reasons, the discriminating power (related to the refinement) of an empirical set of opinions expressed as posterior probabilities is a desirable characteristic by itself, since it represents the *usefulness* of such opinions as highlighted by (25). Therefore, it is important to have a measure of the discrimination component of ECE to identify whether problems in the methods are because of a calibration problem or a lack of discriminating power. In the former case, the evidence evaluation models can be re-defined in order to obtain better calibration. In the latter case, a better solution may be to explore other ways to extract useful features from the available data in the case, because good calibration will not be of help if good discriminating power is not obtained from the evidence. Moreover, knowing the refinement of a method allows the determination of the calibration of such a method (because both magnitudes are complementary for a given value of ECE), and therefore the calibration of the methods can be explicitly measured.

There is a strategy for approximating the discriminating power of a set of posterior probabilities, proposed in (23), and which is achieved by the use of the so-called Pool Adjacent Violators (PAV) algorithm. The procedure essentially transforms a set of posterior probabilities into a more calibrated set of posterior probabilities, according to the definition of calibration given in (25). The transformation by PAV preserves the discriminating power of the set of LR values, and therefore, after its application, the value of ECE represents the loss approximately due to a non-perfect discriminating power, because the calibration component of ECE has been reduced to its minimum.

The PAV algorithm is described in depth in (27)(28).

Representing assessment results: the ECE Plot

In this work we propose a representation of ECE as a function of $P\left(\theta_p \middle| E\right)$ in an *ECE plot*. Posterior probabilities are computed using the validation set of LR values for each prior probability in a fine set of values of the [0,1] range. An ECE plot shows three comparative performance curves together (Figure 5):

i. The solid curve is the ECE (average information loss) of the LR values in the validation set. The higher this ECE curve, the more *information* is needed in order to know the true propositions on average for the LR values in the validation set, and therefore the worse the method. This is the same representation as shown in Figure 4.

ii. The dashed curve represents the comparative performance of the *calibrated* method. This curve is the ECE of the validation set of LR values after being transformed using the PAV algorithm. Therefore, this shows the performance of a method that has the same discriminating power as the original one, but optimises its calibration. This dashed curve can only be obtained if the correct values of θ are known, and therefore it represents a ceiling of performance rarely achievable in practice.

iii. The dotted curve represents the comparative performance of a so-called *neutral* evidence evaluation method, defined as the one which always delivers LR=1 for each case. For this *neutral* method, $P\left(\theta_p \middle| E\right)$ is always equal to $P\left(\theta_p\right)$, and the evidence has no value.

FIGURE 5 - Example of ECE plot.

In an ECE plot, the two comparative performance curves play an important role.

- The performance of the calibrated method represents the component of the ECE arising from the non-perfect discriminating power of the validation set of LR values under analysis, because the component of ECE due to calibration has been minimised.
- Neutral method. If the ECE of the validation set of LR values under analysis is greater than the performance of the neutral method, then it will perform even worse than not using the evidence at all.

Using ECE plots: recommendations for forensic scientists

Several applications of ECE plots are reported in the form of recommendations for forensic scientists for different scenarios in casework.

The proposed techniques are available for use by forensic scientist with a toolkit for drawing ECE plots in Matlab^{TM} that has been developed by the first author of this work. The software is freely available, and can be downloaded from http://arantxa.ii.uam.es/~dramos/software.html.

Deciding which evidence evaluation method to choose

In forensic practice, and prior to casework, it may be the case to have different methods for evaluating the same type of evidence. As examples two different software packages may be available for Automatic Fingerprint Identification; or two different selection strategies for populations may be available for glass analysis (19).

Assume that the forensic examiner has available for use two different methods for evidence evaluation, M_1 and M_2, say. A validation database is then set-up in order to assess the accuracy of their methods prior to their use in casework. The scientist then computes the ECE for both methods using the common validation database. The ECE values are denoted ECE_{M1} and ECE_{M2} respectively. The question to answer is then: *Which method should be used in subsequent casework?* We identify the following scenarios in this context:

- Assume that $ECE_{M1} < ECE_{M2}$ for a region R_1 of possible values of the prior odds, and $ECE_{M1} > ECE_{M2}$ elsewhere. In this situation, the information in the case that is not related to the evidence and is summarised in the prior odds determines which method should be used. If the prior odds fall in R_1 then M_1 should be used, otherwise M_2 should be used. If R_1 is all possible values of the prior odds, then M_1 will be preferred to M_2. If the prior odds are not known, as it is usually the case, an option would be to evaluate the evidence using both methods, and to clearly inform the court about this.
- $ECE_{M1} = ECE_{M2}$ for all values of the prior odds. In this case, the value of ECE is the same for both methods for every value of

the prior probability. Either method can be used for subsequent casework, since their performance is the same.

Deciding whether a given method is adequate to evaluate the evidence

A typical scenario in forensic casework is the need to evaluate the evidence when the available materials are of poor quality (*e.g.*, low-copy DNA profiles, extremely degraded fingermarks, too small an amount of glass from the suspect clothes or from the evidential clothing, etc.). In these cases, the practitioner may wonder whether there is any value at all in the evidence.

ECE plots help answer this question, because they establish a theoretical limit for what is understood as a *minimum* performance. After selecting a proper validation database in conditions comparable to the materials in the cases where some evidence evaluation method is to be used, comparisons are generated using the database. The corresponding validation set of LR values is then computed, and the corresponding ECE plot can be also be computed. There may be a region of prior probabilities where the value of ECE is greater than the performance of the neutral method (the dotted curve in the ECE plot). If this happens, the evidence evaluation method at hand performs even worse in terms of *information* (in an Information-Theoretical sense) than not evaluating the evidence at all (neutral method). In the regions of prior probabilities of an ECE plot where that happens, the performance of the evidence evaluation method is deemed inadequate, because using the method is worse than not evaluating the evidence (neutral method).

This gives practitioners advice as to whether or not it is worthwhile to use the given evidence evaluation method in some casework scenario.

In the case where the ECE is lower than the neutral method for all the values of the prior probabilities, then the performance of the evidence evaluation method being evaluated is better than the neutral method, independently of the value of the prior probabilities. This is the aim of all evidence evaluation methods, since the prior probabilities are usually unknown to the forensic scientist, and therefore methods should perform in this manner for all possible prior probabilities. Thus, this situation should be encouraged in forensic practice as a condition for the validation of the methods in use. This is an advantage of the use of the proposed evidence evaluation framework over other approaches such as tippet plots, because it allows the practitioner to assess whether this condition holds.

Detecting problems in evidence evaluation methods

ECE can be used effectively to detect certain LR values that degrade the overall performance of a given validation set. This is especially useful to identify outliers in the set and to investigate possible causes in order to seek possible problems in the evidence evaluation methods used in casework.

Because of the averaging process in ECE (see Equation 7), each LR value obtained from a single comparison performed using a validation database contributes to the overall value of ECE. This allows the impact on the ECE of each of the LR values to be identified, and to detect certain LR values which represent a higher degradation of the value of the ECE. Figure 6 shows an example with artificial data, where the analysis of the individual contributions to ECE enables detection of the LR values which highly degrade performance. The dataset used to generate Figure 6 is artificial for the purpose of illustration. A single true-θ_d comparison was generated with a very high LR value to simulate an outlier. In Figure 6(a) and 6(b) the Tippett plots and ECE plots of this validation set of LR values are shown. From the Tippett plots, the impact of a non-negligible proportion of true-θ_d LR values of around LR=1000 ($\log_{10}$(LR)=3) is not easily shown. In Figures 6(c) and 6(d) the individual contribution of each LR value to ECE is shown respectively for true-θ_p and true-θ_d comparisons. Figure 6(d) clearly shows that there is a single comparison contributing to ECE much more strongly than the rest. Thus, a single LR value which degrades the performance has been detected. The forensic scientist can then investigate that result in more detail in search of the causes of problems.

FIGURE 6 - Detection of single LR values in the validation set which seriously affect performance. For all cases, N_p=N_d=1000. In Tippet plots (a) the impact of some high LR values when θ_d is true is not highlighted. However, ECE plots (b) show a bad behaviour in the area of lower absolute values of the prior odds (ECE is over the neutral LR set in that area). The analysis of the individual contribution of LR values for true-θ_p (c) and *true-θ_d* (d) cases, shows that the bad behavior is caused by a single true-θ_d LR. In vertical axes, the ECE of each value is multiplied by the number of LR values respectively for true-θ_p (c) and *true-θ_d* (d) cases in order to make the representation of the individual contributions insensitive to the sample size.

Reporting performance of evidence evaluation methods to court

Once it is decided that it is worth using a given evidence evaluation method in casework, it may be necessary to report the validation results in court. In this situation, the information-theoretical interpretation of ECE could be used. Imagine a case in court where control and recovered materials are presented as evidence. The fact finder asks for the forensic evaluation of such evidence. The forensic scientist compares the control and recovered materials with respect to a relevant population and computes a LR value using a given method.

Now assume that the fact finder asks for the performance of the evidence evaluation method used in the case. Suppose that the fact finder has a prior probability $P(\theta_p)$ for the prosecutor proposition θ_p before the analysis of the evidence. Thus, the ECE value in the plot at the given value of $P(\theta_p)$ should be stated as *the average information, once the evidence under consideration has been analysed, that the fact finder still needs in order to know which proposition is actually true in the case*, for the given value of the prior probabilities. This means that the smaller the value of the ECE, the better the method, because the fact finder needs less information after the evidence evaluation. Moreover, the ECE should be as lower as possible than the neutral method.

For instance, imagine that in a given case at trial the ECE of the validation set of LR values is 0.6 in the ECE plot, and the ECE of the neutral method is 1 (its maximum value). That means that the amount of information about the case once the evidence is known has increased by 40% compared with the situation before the evidence was presented.

In the previous example, the exact expression of ECE with values equal to 0.6 or 1 is only possible if the prior odds are known. As a report to the court is usually explained by the practitioner at trial, information about the prior probabilities can be given there to the forensic practitioner. However, in many cases such prior probabilities are not even stated by the fact finder. In this case, we remark that the information-theoretical interpretation of ECE can be expressed for any possible value of the prior. But the use of evidence evaluation methods that perform better than the neutral method for all possible values of the prior probabilities is highly recommended, in order to be able to report informative results in court regardless of whether the prior odds are stated or not.

Again, we highlight that the current methodology will be only possible to use in court when judges and advocates will be ready to understand the meaning of the likelihood ratio paradigm with respect to all the rest of elements in a decision framework. Despite the efforts to that respect, we do not consider this possibility in the short-term.

Case example: evaluation of glass evidence

Here we present a case example with forensic glass analysis, which illustrates the proposed methodology and the recommendations given to practitioners with the methods and databases used in real forensic practice. The objective of this section is to show the use of the proposed assessment methodology in comparison to Tippett plots. To this end, we use several evidence evaluation methods previously proposed in the literature.

The importance of glass as evidence was recognised many years ago as very small glass fragments (of linear dimension 0.1 - 0.5 mm) that arise during car accidents, burglaries, fights, *etc*., could be carried on the clothes, shoes and hair of participants (29). Because of their very small size they are analysed by the application of various analytical methods which give reliable data for small objects and yield various kinds of physicochemical information. Scanning Electron Microscopy coupled with an Energy Dispersive X-ray spectrometer (SEM-EDX) is one of these methods and this is routinely used in many forensic institutes for solving various forensic problems. Results of glass analyses by the application of this method were used here.

Description and context of the case

A burglary has occurred. A window has been broken to get into a house. There are no eyewitnesses. The police has been advised by an anonymous informant, and shortly after the burglary a suspect has been arrested close to the scene of the crime. Some fragments of glass are recovered from the jacket of the suspect, in an appropriate amount and size elemental composition may be successfully measured using the SEM-EDX method. The suspect gives an interview without comments; under such circumstances the scientist considers that he is unable to address activity- or offence-level propositions. He therefore concentrates on establishing source-level propositions relating the fragments recovered on the suspect and a control glass from the broken window at the scene of the crime. The following propositions are stated:

- θ_p: the glass recovered from the suspect comes from the broken window in the scene of the crime.
- θ_d: the glass recovered from the suspect comes from another window with similar physicochemical characteristics as the broken window at the scene of the crime.

The forensic scientist will compute a LR to express the value of the evidence in the case at hand. In order to compute the LR, there are three different models that may be used, these are identified as MVLR-Full, MVLR-NaSiCa and GMF and are described below. The forensic scientist wants to select the best model for the case at hand.

Validation of evidence evaluation methods in the context of the case

We assume that the forensic laboratory in charge of evidence evaluation of the case at hand has never conducted a validation experiment in order to determine the performance of the evidence evaluation methods to be used in a case such as this. Therefore, the practitioner needs to conduct such a validation experiment if they want to know which is the best method to be used in the case. It is important to highlight that this will not be the typical case in forensic practice, where the validation procedure may have been conducted before the methods are even considered to be used in casework.

However, for the sake of illustration here, knowing the case in which the evidence evaluation methods are to be used will be useful in order to describe the selection of a proper validation database.

Prior to their use in casework the forensic practitioner decides to use Empirical Cross-Entropy as the criterion to establish the best evidence evaluation method. He then needs to select a proper validation database. The steps needed to do this are described in the following sections.

Available data

For the given case, the data available to the forensic scientist are as follows (the data used in this paper have been collected at the Institute of Forensic Research in Krakow, Poland): they consist of 165 glass-objects, namely 87 car windows and 78 building windows. Four glass fragments from each glass object were analysed by the SEM-EDX method. Each of the four glass fragments selected for analysis was measured three times and the mean of the three measurements was taken for each fragment. Therefore, each glass object was described by four vectors (one for each fragment) of elemental composition of oxygen (*O*), sodium (*Na*), magnesium (*Mg*), aluminum (*Al*), silicon (*Si*), potassium (*K*), calcium (*Ca*) and iron (*Fe*). For each measurement, the logarithm to the base 10 of various elemental concentrations divided by the concentration of oxygen were analysed, leading to seven variables: *Na'*, *Mg'*, *Al'*, *Si'*, *K'*, *Ca'* and *Fe'* (*e.g.*,

$Na'=\log_{10}(Na/O)$). More information about the SEM-EDX procedure can be found in (30).

Selection of a population database

As explained above, a population database will be used for the case under investigation. The database will be used to help determination of the parameters of the evidence evaluation models for the LR computation using the given control and recovered data. The model parameters will be determined from this population database; after that the statistics from the control and recovered data will be used for the computation of the LR.

As the alternative proposition θ_d states that the potential sources of the recovered glass are windows of similar physicochemical characteristics as the broken window in the scene of the crime, the forensic scientist selects a database of glass fragments from car and building windows, which are known to present similar physicochemical properties as the broken window at the scene of the crime, and which are known to behave similarly when analyzed with SEM-EDX methods (31). Thus, the forensic scientist decides that the population database will consist of all the available data described above.

Selection of a proper validation database

The validation database will be selected prior to the LR computation in the case at hand. As previously described, the aim of the validation database is to generate hypothetical cases by the use of hypothetical control and recovered data, for which θ_p and θ_d, respectively, will be true. Then, true-θ_p and true-θ_d LR values are computed from those hypothetical cases to generate a validation set.

As a consequence, the forensic scientist should establish the conditions of the control and recovered glass, in which will then be used to generate appropriate data for the validation database. In this case example, the forensic scientist knows the physicochemical characteristics of the control materials, because their origin is known. For the recovered materials, the forensic scientist may consider the circumstances of the case and any statements of the suspect (defence proposition) to determine the conditions of the recovered glass fragments. In a case like this one, for which the type of the recovered materials is not clear, the forensic scientist may establish the type of the recovered glass by the use of classification methods, which have been demonstrated to have high accuracy for SEM-EDX analysis in order to distinguish car and building windows from other glass types (31)(32). The forensic scientist uses glass classification techniques to determine that the conditions of the recovered glass fragments are clearly most likely to be those of car and building windows. Then, they assume these conditions in the recovered glass in order to select a proper validation database. In some other forensic disciplines, the procedure to establish the type of the recovered materials may be different, but the aim is similar. For instance, for speech evidence the transmission channel of the recovered materials can be obtained from police information, and the noise conditions can be determined from the analysis of the speech itself.

Therefore, each comparison performed with the hypothetical cases generated with the validation database will have to be done with hypothetical control and recovered materials coming from car or building windows. The forensic scientist considers that, using all the available data described above, the hypothetical cases necessary to compute the ECE plots can be properly simulated, according to the comparison protocol described below. Therefore, the validation database will be the whole dataset described above.

Notice that the population and the validation database in this example are the same dataset. In general, this will not be the situation: the validation database will be used to measure performance before casework, and then a different population database will be used for each particular case. We clarify the reasons for this below in next section.

Defining the comparison protocol to generate the validation LR set

As the available data in the validation database for this example are sparse, a so-called cross-validation procedure is used to perform the necessary comparisons. The procedure is described as follows, depending on whether true-θ_p or true-θ_d LR values are generated:

- For each true-θ_p hypothetical case, a single glass object is taken from the database, and the hypothetical control and recovered samples are selected from the SEM-EDX profiles for this object. The population database used with such a hypothetical case will consist of the rest of the glass objects in the database. The process is repeated for all the objects in the validation database, *i.e.*, a hypothetical case is simulated for each object in the validation database, with the appropriate population database on each occasion begin the remainder of the objects.
- For each true-θ_d hypothetical case, two different glass objects are selected from the database. The hypothetical control sample is taken from one of the objects, the hypothetical recovered sample is selected from the other. The population database used with such a hypothetical case will consist of the rest of the glass objects in the database. The process is repeated for all possible combinations of two glass objects in the validation database, *i.e.*, a hypothetical case is simulated for each possible combination of two different glass objects in the validation database, with the appropriate population database on each occasion being the remainder of the objects after removal of the two objects to act as control and recovered objects.

In particular, for true-θ_p comparisons, LR values in each hypothetical case were calculated using two of the four fragments of the elemental composition from each object as control data, and the other two fragments of the elemental composition were used as recovered data. Therefore, the number of hypothetical cases simulated equals the number of objects M=N_p=165 present in the database. For true-θ_d comparisons the first two fragments of the elemental composition of each object were used as control and recovered data, respectively. The number of hypothetical cases simulated is therefore the combination without replacement of the total number of pairs of different objects *M*, *i.e.*, $N_d = M \times \frac{M-1}{2} = 13530$, where $M = 165$.

With this cross-validation procedure, for each hypothetical case involving two glass objects, the rest of the objects in the validation database were used as the population database. In this way, the available database is used efficiently to simulate the comparison conditions in the actual case. Moreover, the control and the recovered materials in each hypothetical case are of similar physicochemical conditions as the materials in the actual case, and the population database used in each hypothetical case is similar to the one used in the actual case (it will differ on only one or two glass objects). Therefore, this procedure for generating a validation set of LR values complies with the requirements of a proper empirical assessment of performance, as described earlier.

The reason for using a cross-validation procedure is because the size of the available dataset in this glass example is small. For cases or disciplines where a much bigger dataset will be available, the use of a cross-validation procedure may not be necessary, and both the population and the validation database may be taken as different subsets of the whole database. The cross-validation procedure enables the same database to be used for the two purposes of a validation database and a population database.

Evidence evaluation methods for comparison

The forensic scientist wants to compare the performance of three evidence evaluation methods, MVLR-Full, MVLR-NaSiCa and GMF. These evidence evaluation methods will each be used on the whole validation set to obtain three sets of LR values.

For simplicity, the methods are not described in detail; the interested reader can study the references given for further detail. The three methods for comparison are as follows.

- The first two methods are different versions of a Multivariate LR (MVLR) method, proposed by (33). The method assumes a multivariate model for all the SEM-EDX variables. Further details can be found in (33). In this work we compare the use of this approach for two methods using a different number of variables (dimensions):
    - The whole set of seven variables in the database are modelled. This method will be referred to as MVLR-Full.
    - A reduced set of 3 variables, namely *Na'*, *Si'* and *Ca'* are considered. This method will be referred to as MVLR-NaSiCa.
- The third method will be referred to as Graphical Model Factorisation (GMF), and has been proposed by (30). The aim of the approach is a reduction in the dimensionality of the model while still using all the available variables. This may be done with a graphical model. Further details can be found in (30).

LR values close to zero affect the factorisation of the models. In order to avoid any associated problems, the minimum value given to all LR values will be limited to 10^{-12}.

Comparative analysis of the performance of evidence evaluation methods: deciding which evidence evaluation method to choose

As a preliminary analysis, the forensic scientist draws Tippett plots with the validation set of LR values, in order to compare the proposed methods, namely MVLR-NaSiCa, MVLR-Full and GMF; see Figure 7. It can be seen that the MVLR-NaSiCa method (Figure 7(a)) presents limited rates of misleading evidence with a moderate value of strong misleading evidence. Rates of misleading evidence have been defined before in the introduction of Tippett plots as the proportion of LR values supporting the wrong proposition. Strong misleading evidence is defined as LR values strongly supporting the wrong proposition. On the other hand, MVLR-Full and GMF methods (Figures 7(b) and (c) respectively) present lower rates of misleading evidence in general, but they also present a much higher proportion of very strong misleading evidence. In this scenario, it is not clear which of the proposed methods is better. Moreover, the discriminating power of the methods cannot be easily compared.

The forensic scientist then compares the three methods with ECE plots, as shown in Figure 8, where the following conclusions can be drawn:

- The method with the lowest value of ECE (solid curve) for the entire range of prior probabilities is the MVLR-NaSiCa method, followed by the GMF and finally the MVLR-Full method. This means that the method that performs best is MVLR-NaSiCa. This can be justified because the data used for evidence evaluation, namely the control and recovered samples and the population database, are too sparse for a high-dimensional multivariate problem. Data sparsity is known to affect the reliability of the performance seriously when the dimensionality of the problem increases (the so-called curse of dimensionality). The MVLR-NaSiCa model uses only 3 variables, compared to the dimensionality of 7 of the MVLR-Full and the product of several two- and lower-dimensional distributions in the GMF model. Therefore MVLR-NaSiCa is more robust for data sparsity conditions. As the validation database has been designed to mimic the conditions of the LR computation process in the case, the forensic scientist is justified in concluding that the MVLR-NaSiCa method should be chosen for this case.

- The GMF method is much better than the MVLR-Full method, because its value of ECE is much lower. This is due to the strong requirements of data of the MVLR-Full method in high dimensionality. On the other hand, GMF reduces the dimension of the model. Thus, when the dimensionality increases, reliability in the performance of the MVLR-Full method can decrease. In contrast, the GMF performance is more reliable, using distributions with only one or two variables and thus being less susceptible to the curse of dimensionality.

- The discriminating powers (dashed curve obtained using the PAV algorithm) of the GMF and MVLR-Full methods are similar, outperforming that of the MVLR-NaSiCa method. This means that the GMF and MVLR-Full methods are better at extracting discriminating information from the evidence. This is a reasonable conclusion, since they use seven variables (sources of information) rather than the three variables used by MVLR-NaSiCa. This justifies the use of as much information (variables) as possible for evidence evaluation and also supports the use of dimensionality reduction techniques such as the GMF model to handle calibration problems.

- Although there is good discriminating power (dashed curve), the GMF and MVLR-Full methods present a high value of ECE curve (solid curve) which means a bad overall performance. This indicates a calibration problem. For the MVLR-Full model, there is a clear cause of such a problem, since the model is highly sensitive to data sparsity. However, in the case of GMF such results indicate the need for more research in order to adapt the model to situations with sparse data, situations which are beyond the scope of this paper.

FIGURE 7 - Tippett plots for the proposed methods MVLR-NaSiCa (a), MVLR-Full (b) and GMF (c). For all plots, N_p=165, N_d=13530.

FIGURE 8 - ECE plots for the proposed methods MVLR-NaSiCa (a), MVLR-Full (b) and GMF (c). For all plots, N_p=165, N_d=13530.

The ECE plots in Figure 8 illustrate a maximum value of ECE=1. This is because the maximum value of the ECE of the neutral method is 1, and therefore any value for the ECE of any validation set of LR values above 1 for any range of the prior probabilities means that the evidence evaluation method is deemed inadequate in such a range as it is worse than the neutral method. Thus, it will be less interesting in general to analyze the ECE performance above the value of 1. Moreover, this allows a clearer representation of the performance when the value of ECE is between 0 and 1, which is the ECE range where the evidence can yield meaningful information.

It can be observed in Figure 7, Tippett plots present a stair-wise shape for the true-θ_d curve, which is due to the small size of the database: as N_p is only 165, the cumulative distributions show that effect, because variation in the x-axis make cumulative distributions vary only in occurrences of a LR in the validation set. It is not the case of ECE plots in Figure 8, because the variation of ECE in the x-axis is in the prior-log-odds range, and therefore it is not affected by the small sample size (see Equation 7).

As it can be seen, ECE plots allow forensic practitioners to extract much more useful conclusions than Tippett plots about the comparative performance of evidence evaluation methods in a case. Moreover, they provide a more thorough analysis of the strengths and the weaknesses of the methods under analysis, which is of great value in forensic practice.

Comparative analysis of the performance of evidence evaluation methods: checking the adequacy of evidence evaluation methods

Tippett plots, as in Figure 7, do not provide a clear measure on whether a particular evidence evaluation is adequate or not to be used in casework. However, from ECE plots in Figure 8 the forensic scientist can reason as follows:

- MVLR-NaSiCa: for this method (Figure 8(a)) the value of ECE (solid curve) is higher than the value of the neutral method (dotted curve) for all values of the prior-log-odds axis above 0.6. This evidence evaluation method will not be adequate if the base-10 logarithm of the prior odds is higher than 0.6.
- MVLR-Full: this method (Figure 8(b)) yields a poor performance, with a value of ECE much higher than the neutral method for all prior odds greater than 1 (log prior odds greater than 0), and with a similar performance to the neutral method for prior odds less than 1. This means that, in the best of cases, this method will be almost the same as not evaluating the evidence at all, and it can be even worse if the logarithm of the prior odds is higher than 0. Therefore, it seems clear that the method is not suitable for evidence evaluation in these conditions.

- GMF: for the GMF method (Figure 8(c)), the value of ECE is higher than the value of the neutral method (dotted curve) for all values of the prior-log-odds above 0.3. Therefore, if the base-10 logarithm of the prior odds is higher than 0.3 the method provides no useful information for evidence evaluation.

It is worth noting that no method from the three analysed has achieved a better performance than the neutral method over the whole range of prior probabilities. This means that, if the prior odds are not known by the forensic scientist, as it happens in many cases, there will be no way to determine if a method is appropriate for evidence evaluation in this case. We highlight two facts in relation to this. First, this circumstance should be stated in court if the technique is going to be used in the case, because the fact finder must know that the technique may be misleading in some situations (some values of the prior odds). Second, and more important, it is recommended that the forensic practitioner only consider the use of evidence evaluation methods that provide useful information for a wide range of values of the prior odds. Thus, in the given case example, further work is required to improve the calibration of the methods used so that the value of ECE may be reduced to be lower than the neutral method for a wider range of prior probabilities. Without the aid of the ECE methodology proposed, this important weakness of the methods in use would have been much more difficult to detect and characterize.

Comparative analysis of the performance of evidence evaluation methods: detecting problems in evidence evaluation methods

Figure 9 shows the contribution of each LR value in the validation set to the total value of ECE, for the three methods analysed. It can be seen that for all methods the contribution of true-θ_d LR values follows a comparable trend, with no one being strongly misleading. However, for true-θ_p comparisons all the methods present some LR values which strongly contribute to the final value of ECE compared with the majority of comparisons, especially for the MVLR-Full and GMF methods. This suggests that the models working at a higher dimension (7 rather than 3) tend to produce more strongly misleading true-θ_p LR values. Furthermore, a deeper analysis of such comparisons (which is outwith the scope of this work) may lead the forensic scientist to identify outliers or problems with the models in use, as a first step to improve their methodologies. The use of the proposed ECE methodology has eased the detection of these sources of problems and the measurement of their impact in the performance.

FIGURE 9 - ECE plots with the individual contribution of each comparison in the validation set of LR values, for the proposed methods MVLR-NaSiCa (a), MVLR-Full (b) and GMF (c). For all plots, N_p=165, N_d=13530. Comparisons of glass objects yielding the highest contributions to ECE are labelled in the figure.

Discussion

Although the use of a wealth of validation databases to determine the performance of methods is common practice in forensic fields such as forensic speaker recognition (34) or forensic biometric systems (35), the impulse of the requirements of the so-called coming paradigm shift is motivating their use in other disciplines where classically performance assessment of comparison methods was not so popular or typical. In the United States in particular, the need of the measurement of the performance in realistic forensic conditions motivated by the Daubert rules is fostering the development of the construction of these validation databases and protocols. It is the case, for example, of the field of fingerprint identification, where, for instance, the FBI has recently conducted a massive campaign for measuring the performance of an important number of fingerprint experts across the United States (16). As a part of this initiative, a validation database consisting of fingermarks and prints in forensic realistic conditions has been built, with a detailed protocol that tries to mimic simulated forensic cases in realistic conditions. Although in such a work the standard procedures for fingerprint identification do not generally consider the use of a population database for any kind of statistical procedure, the philosophy in the construction of a validation database follows the same ideas as described in this work. The development of such validation corpora will play a critical role in the establishment of scientific procedures for the analysis of evidence evaluation methods in the future.

The construction of validation databases is also affected by the amount of data in the forensic laboratory. Evidence evaluation using a likelihood ratio approach needs a population database in order to model the alternative proposition in a given case. It is hoped that the chosen database is a sufficiently large size that reliable estimates of parameters for the models may be made. If a validation database is also to be used to measure performance of the methods at hand, then the forensic scientist has to choose how to use the available data in the laboratory to build validation databases that are representative of the cases in which the evidence evaluation methods will be used. However, if these data are sparse, then the data division of the validation data to simulate populations in hypothetical cases might lead to database sizes that are not large enough to give significant measures of performance. This is of course solved by the provision of more data in the laboratory, an obvious desire for any scientific procedure. However, in cases where data collection is time-consuming or expensive, the use of cross-validation methods, as in the example above, may be of help until more data are available.

Conclusions

The need for the assessment of the performance of evidence evaluation methods is increasing, as exemplified by the requirements of what has been dubbed *the coming paradigm shift in forensic science* (1). In agreement with such ideas, this work has presented a methodology for the assessment of the performance of forensic evidence evaluation methods which express the value of the evidence in the form of likelihood ratios (LR). The proposed framework constitutes a step forward with respect to other popular assessment techniques such as Tippett plots, giving much more useful information about the quality of the LR values and their impact on the performance of the methods. This provides forensic scientists with a more useful tool to assess the performance of their methods, and also to identify problems in statistical models for evidence evaluation. The proposed methodology is based on information theory, and allows the interpretation of performance in terms of *information*. The main contribution of the work is the proposed performance metric, namely Empirical Cross-Entropy (ECE), its use in a LR-based evidence evaluation context, and also several useful tools derived from it, such as ECE plots. Recommendations are also provided for forensic scientists in order to use the proposed performance assessment methodologies in a variety of scenarios. The method is illustrated with a case example, where a database of glass chemical profiles collected from real cases is used. The example illustrates how the best method among several options may be selected for evidence evaluation by the forensic scientist for the case at hand. The usefulness of the recommendations proposed in this work is also illustrated. Future work includes the use of the proposed assessment methodology in other forensic disciplines, and considering propositions at different levels (activity or offence), in order to show its adequacy to different casework scenarios. Moreover, this methodology assumes the definition of two mutually exclusive propositions, which may not be the general case. The exploration of more general frameworks is considered as a future line of research.

Acknowledgements

The authors thank Niko Brümmer for his support and for many fruitful and inspiring discussions about calibration. The authors also thank the reviewers of this article, whose comments have greatly improved the quality of the final work. This work has been partially funded by Universidad Autonoma de Madrid and DGUI of the Comunidad Autonoma de Madrid, under grant CCG10-UAM/TIC-5792.

(1) Saks MJ, Koehler JJ. The coming paradigm shift in forensic identification. Science 2005;309(5736):892-895.

(2) National Research Council, The National Academies. Strengthening Forensic Science in the United States: a path forward. Washington, DC: National Academies Press (US), 2009.

(3) Government of the United Kingdom. The Law Commission Consultation Paper No 190. The admissibility of expert evidence in criminal proceedings in England and Wales - A new approach to the determination of evidentiary reliability. Available at: http://www.lawcom.gov.uk/docs/cp190.pdf (accessed January 17, 2011)

(4) U.S. Supreme Court. Daubert v. Merrel Dow Pharmaceuticals. 509 U.S. 579; 1993.

(5) Parker JB, Holford A. Optimum test statistics with particular reference to a forensic science problem. J R Stat Soc Ser C Appl Stat; 1968;17;237-251.

(6) Darboux JG, Appell PE, Poincaré JH. Examen critique des diverses systémes ou études graphologiques auxquels a donné lieu le bordereau. Enquête de la chambre criminelle de la Cour De Cassation, 5 mars-19 novembre 1904, Tome 3. Ligue française de droits de l'homme et du citoyen, Paris 1908: 499-600.

(7) Lindley DV. A problem in forensic science. Biometrika 1977;64(2):207-213.

(8) Evett IW. Towards a uniform framework for reporting opinions in forensic science casework. Sci Justice 1998; 38(3):198-202.

(9) Champod C, Vuille J. Scientific evidence in Europe - Admissibility, appraisal and equality of arms; Comparative study on scientific evidence drawn up for the Bureau of the Council of Europe's European Committee on Crime Problems. Tech. Rep., Council of Europe - European Committee on Crime Problems, Strasbourg, 2010.

(10) Cover TM, Thomas JA. Elements of Information Theory, 2nd ed. Wiley Interscience, 2006.

(11) Cook R, Evett IW, Jackson G, Jones PJ, Lambert JA. A hierarchy of propositions: deciding which level to address in casework. Sci Justice 1998; 38(4): 231-239.

(12) Ramos D, Gonzalez-Rodriguez J. Cross-entropy Analysis of the Information in Forensic Speaker Recognition. Proc. of Odyssey, Stellenbosch, South Africa, 2008.

(13) Cook R, Evett IW, Jackson G, Jones PJ, Lambert JA. A model for case assessment and interpretation. Sci Justice 1998;38: 151-156.

(14) Aitken C, Taroni F. Statistics and the evaluation of evidence for forensic scientists. 2nd ed. Chichester: J. Wiley & Sons, 2004.

(15) Taroni F, Bozza F, Biedermann A, Garbolino P, Aitken C. Data Analysis in Forensic Science: A Bayesian Decision Perspective. Wiley, 2010.

(16) Ulery BT, Hicklin RA, Buscaglia J, Roberts MA. Accuracy and Reliability of Forensic Latent Fingerprint Decisions. Proc Natl Acad Sci U S A 2011. Doi: doi/10.1073/pnas.1018707108.

(17) Evett IW, Buckleton J. Statistical Analysis of STR Data. Advances in Forensic Haemogenetics, Springer-Verlag, Heilderberg 1996; 6: 79-86.

(18) Tippett CF, Emerson VJ, Fereday MJ, Lawton F, Richardson A, Jones LT et al. The evidential value of the comparison of paint flakes from sources other than vehicles. J Forensic Sci Soc 1968;2: 61-65.

(19) Zadora G, Ramos D. Evaluation of glass samples for forensic purposes - an application of likelihood ratio model and information-theoretical approach. Chemometr Intell Lab 2010;102:62-73.

(20) Royall R. On the probability of observing statistical misleading evidence. J Am Stat Assoc 2000;95(451):760-768.

(21) Martin A, Doddington G, Kamm T, Ordowski M, Przybocki M. The DET curve in assessment of decision task performance. Proc. of Eurospeech 1997:1895-1898.

(22) Ramos D. Forensic evaluation of the evidence using automatic speaker recognition systems. Ph.D. Thesis, Depto. de Ingenieria Informatica, Escuela Politecnica Superior, Universidad Autonoma de Madrid, Madrid, Spain, 2007. Available at http://atvs.ii.uam.es.

(23) Brümmer N, duPreez J. Application Independent Evaluation of Speaker Detection. Comput Speech Lang 2006;20(2-3): 230-275.

(24) Gneiting T, Raftery A. Strictly Proper Scoring Rules, Prediction and Estimation. J Am Stat Assoc 2007; 102: 359-378.

(25) deGroot MH, Fienberg SE. The Comparison and Evaluation of Forecasters. Statistician 1983;32:12-22.

(26) Shannon CE. A Mathematical Theory of Communication. AT&T Tech J 1948;27:379-423,623-656.

(27) Fawcett T, Niculescu-Mizil A. PAV and the ROC convex hull. Mach Learn 2007;68(1): 97-106.

(28) Brummer N. Measuring, refining and calibrating speaker and language information extracted from speech. Ph.D. Thesis, School of Electrical Engineering, University of Stellenbosch, Stellenbosch, South Africa, 2010. Available at http://sites.google.com/site/nikobrummer/.

(29) Curran JM, Hicks TN, Buckleton JS. Forensic Interpretation of Glass Evidence. CRC Press, 2000.

(30) Aitken CGG, Zadora G, Lucy D. A two-level model for evidence evaluation. J Forensic Sci 2007;52(2): 412-419.

(31) Zadora G, Neocleous T. Likelihood ratio model for classification of forensic evidences. Anal Chim Acta 2009; 642(1-2):266-278.

(32) Zadora G. Classification of Glass Fragments Based on Elemental Composition and Refractive Index. J Forensic Sci 2009;54:49-59.

(33) Aitken CGG, Lucy D. Evaluation of trace evidence in the form of multivariate data. J R Stat Soc Ser C Appl Stat 2004;53:109-122 with corrigendum 665-666.

(34) Gonzalez-Rodriguez J, Rose P, Ramos D, Toledano DT, Ortega-Garcia J. Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition. IEEE Trans Audio Speech Lang Processing 2007;15(7):2104-2115.

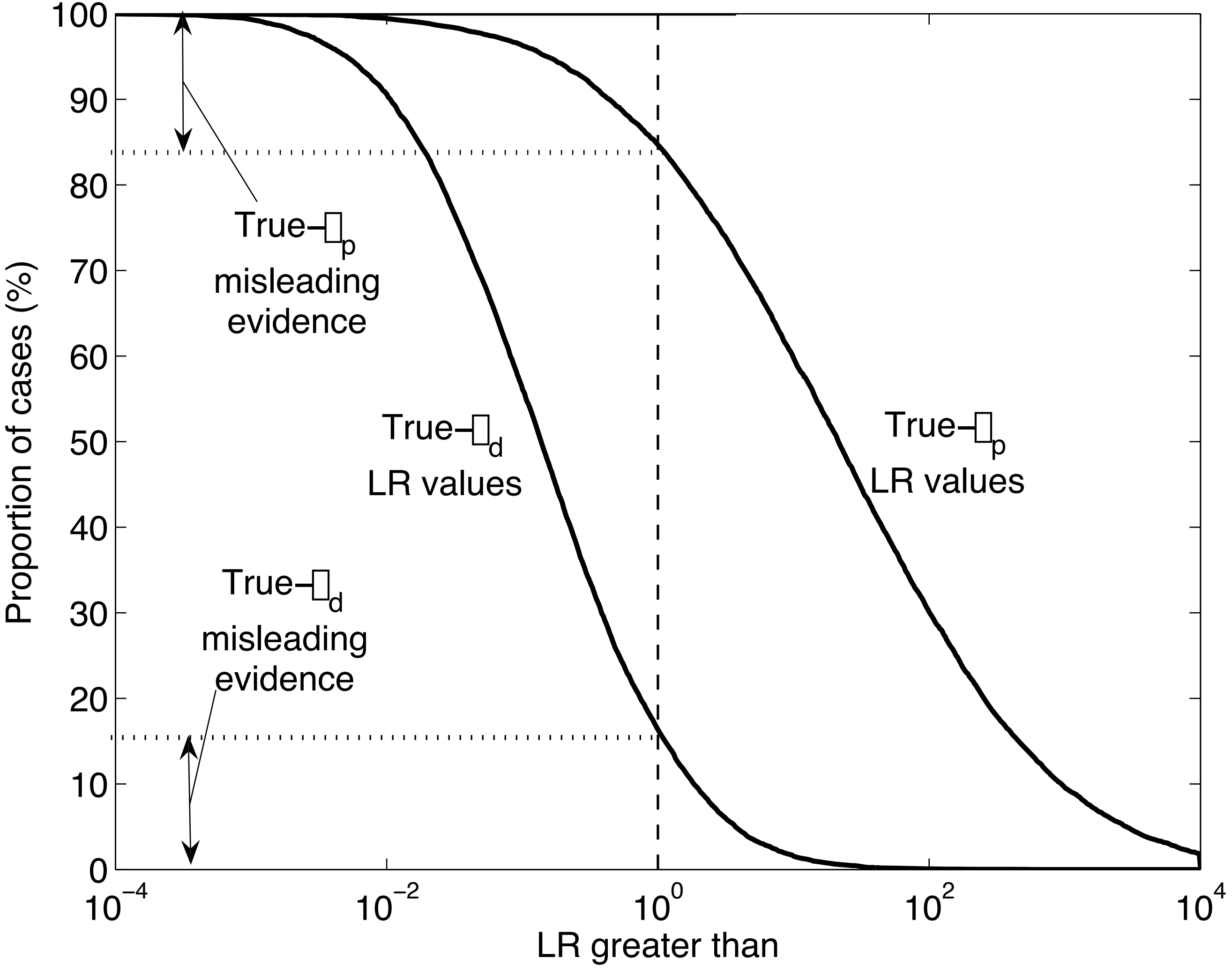
(35) National Institute of Standards and Technologies. Evaluation of Latent Fingerprint Technologies. Tech. rep., http://fingerprint.nist.gov/latent/elft07.

Additional information and reprint requests:

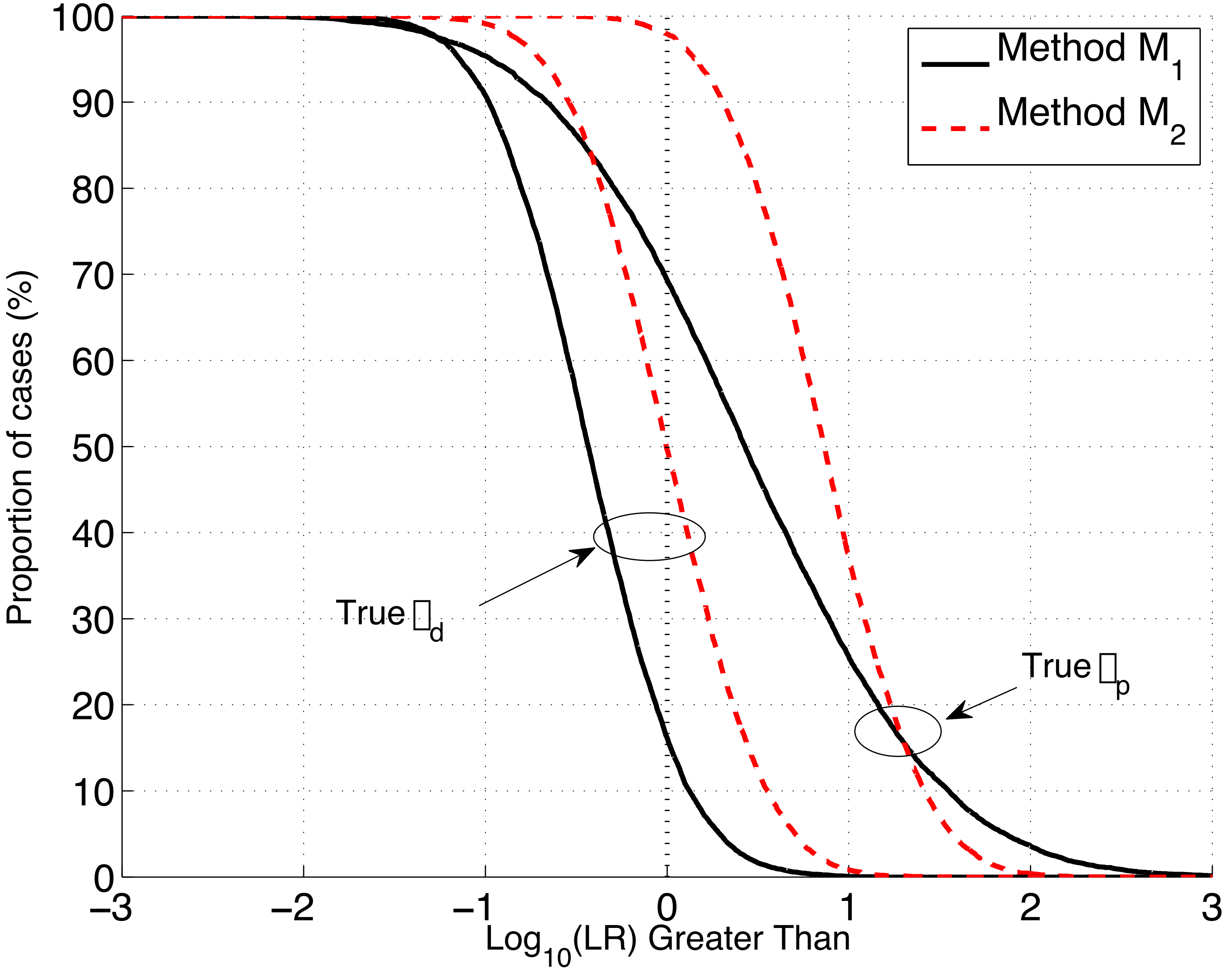
Daniel Ramos, Ph.D.

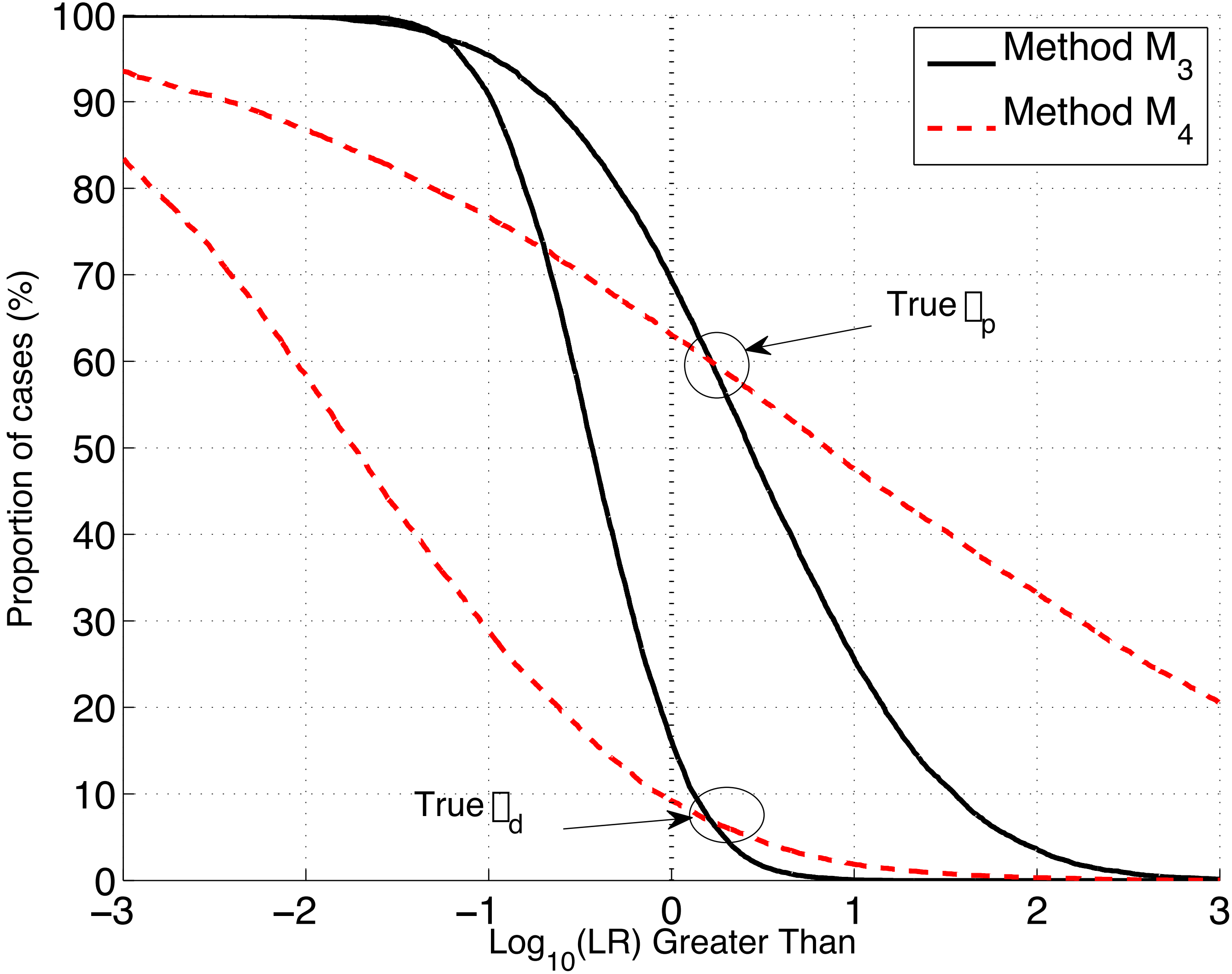
ATVS – Biometric Recognition Group and Forensic Science and Security Institute (ICFS). Escuela Politecnica Superior. Universidad Autonoma de Madrid. C/ Francisco Tomás y Valiente 11, 28049 Madrid, Spain.

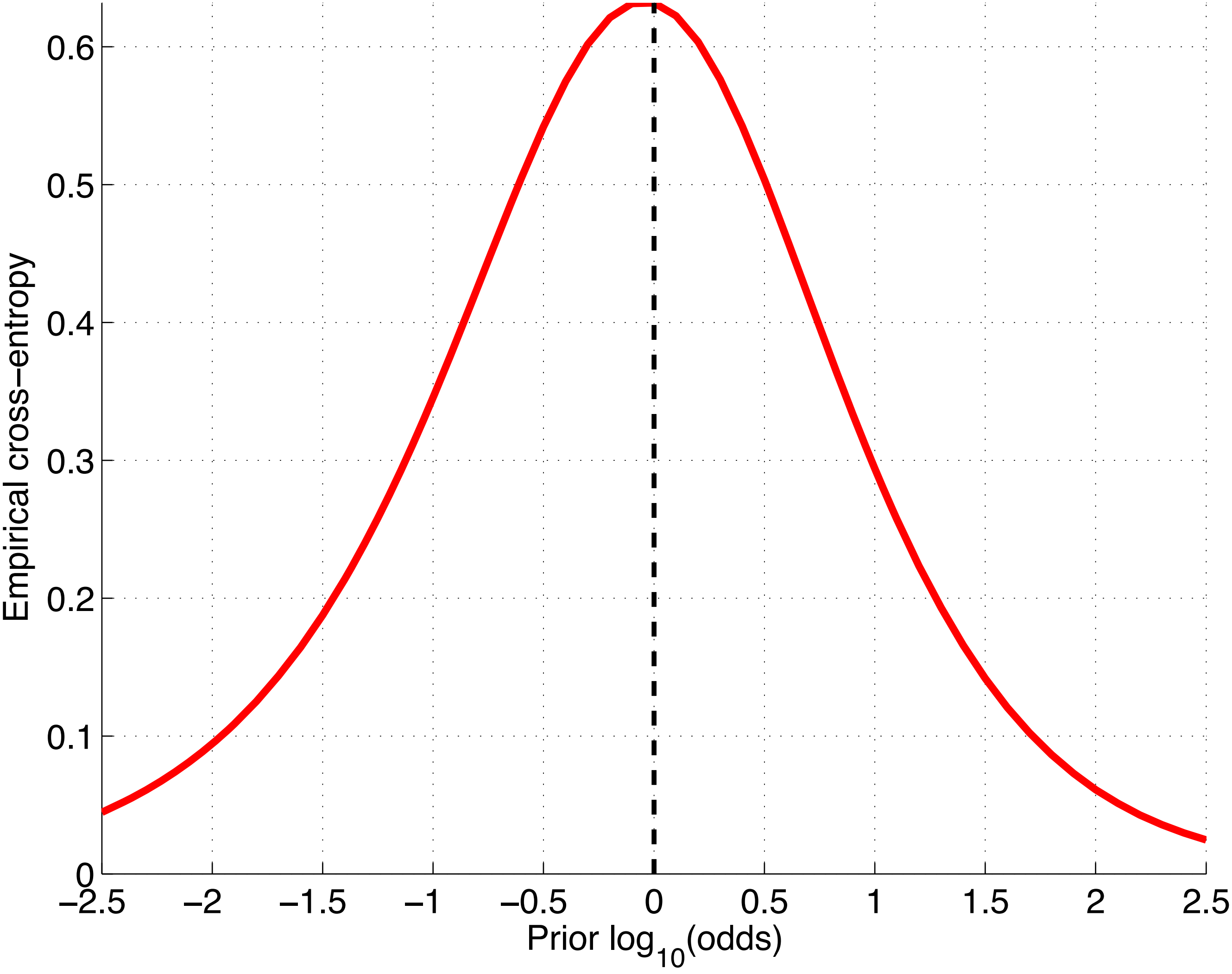
e-mail: daniel.ramos@uam.es

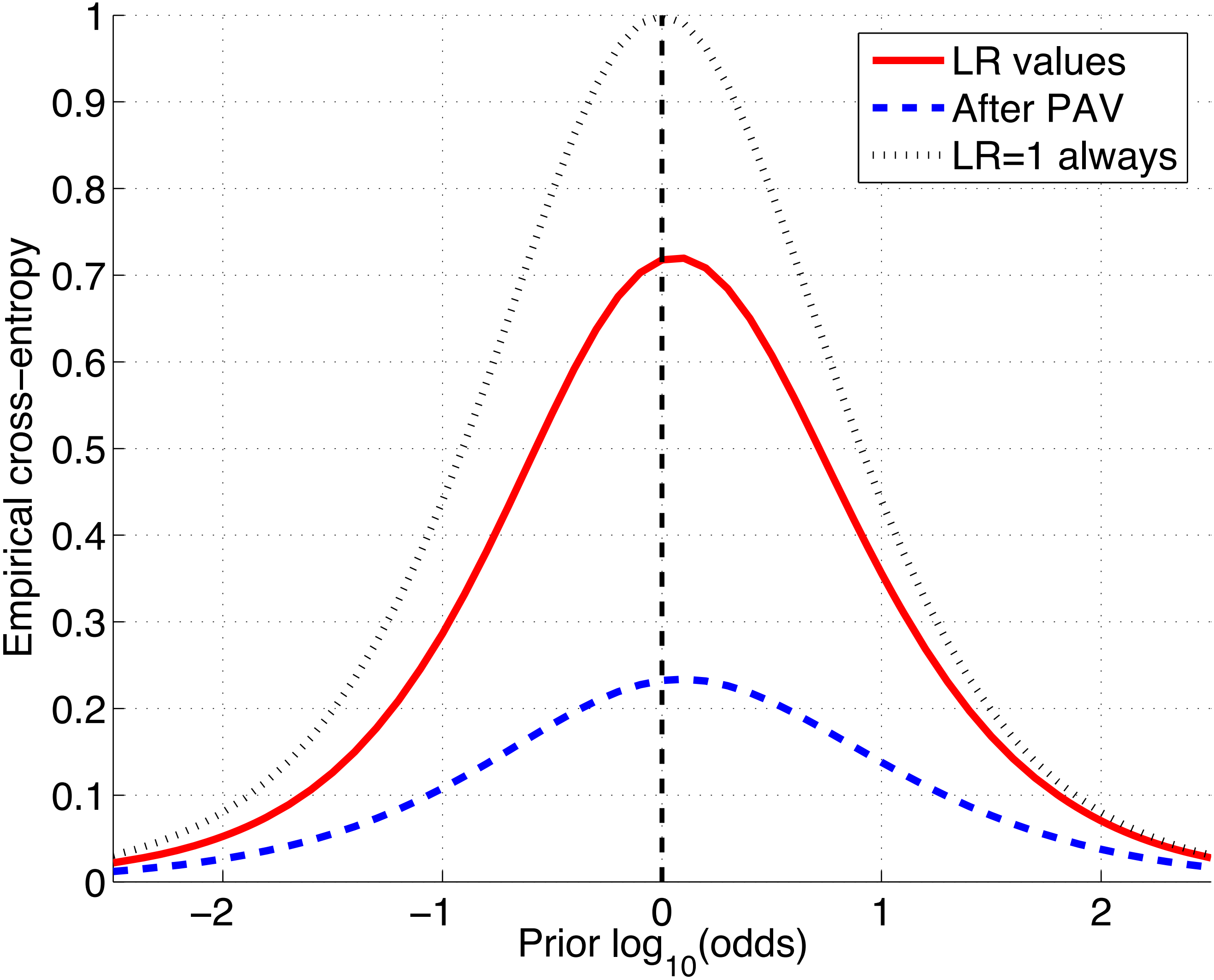


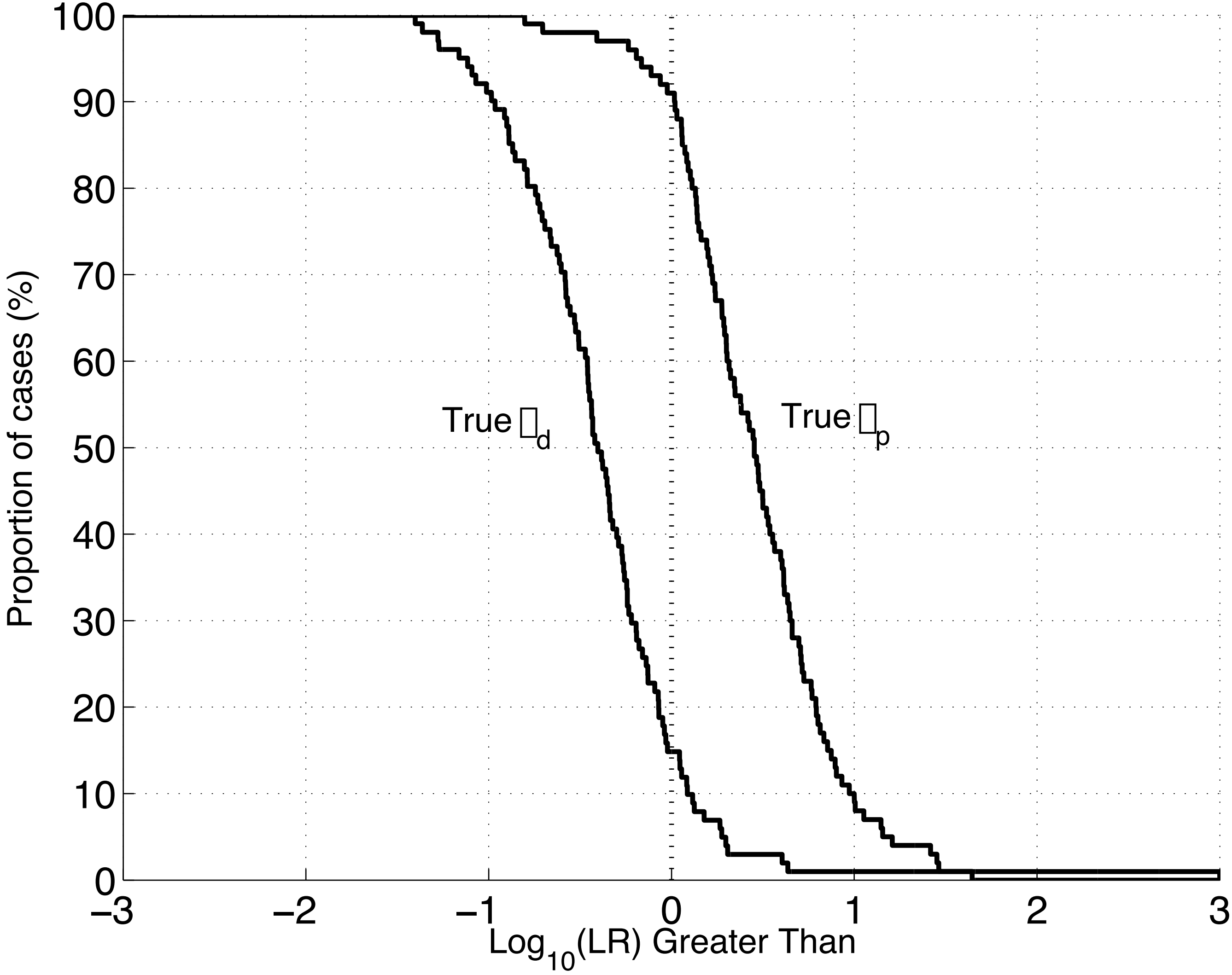
100
90
80
70
60
50
40
30
20
10
0
Proportion of cases (%)
True−□p
misleading
evidence
True−□d
LR values
True−□p
LR values
True−□d
misleading
evidence
10^−4
10^−2
10^0
10^2
10^4
LR greater than

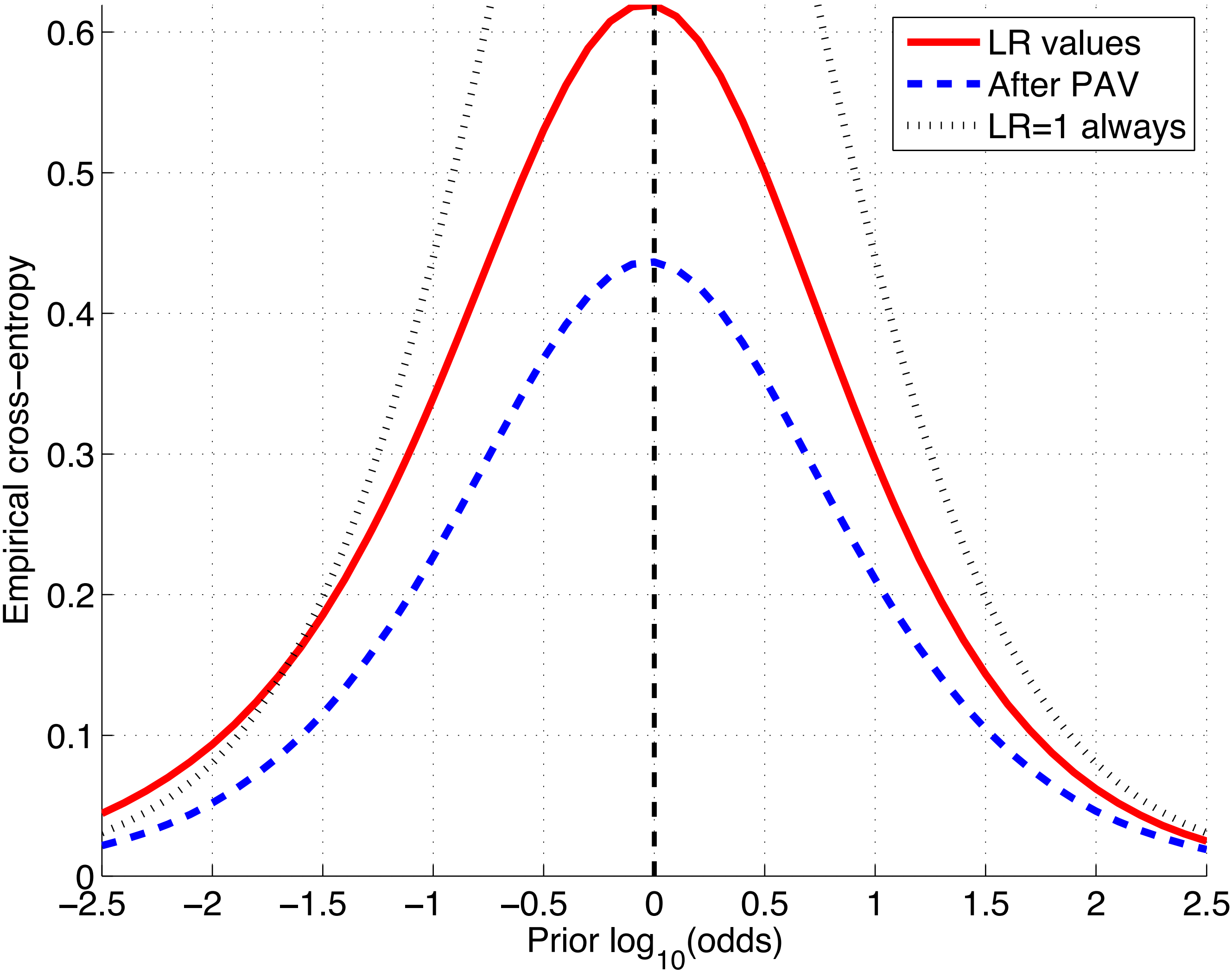

Method M_1
Method M_2
True □$_d$
True □$_p$
Proportion of cases (%)
100
90
80
70
60
50
40
30
20
10
0
−3
−2
−1
0
1
2
3
Log_{10}(LR) Greater Than

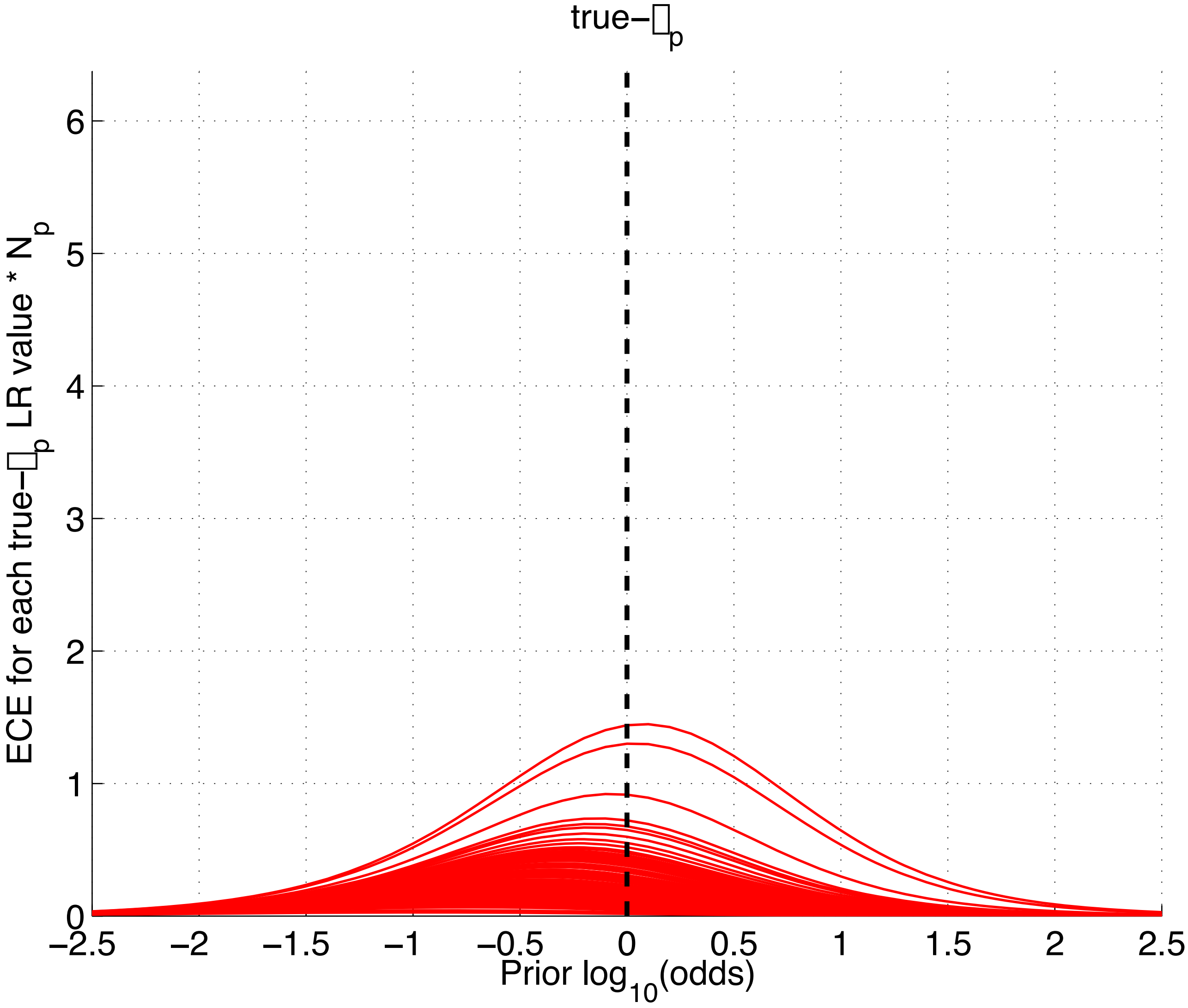

Method M_3
Method M_4
True □$_p$
True □$_d$
100
90
80
70
60
50
40
30
20
10
0
−3
−2
−1
0
1
2
3
Proportion of cases (%)
Log_{10}(LR) Greater Than

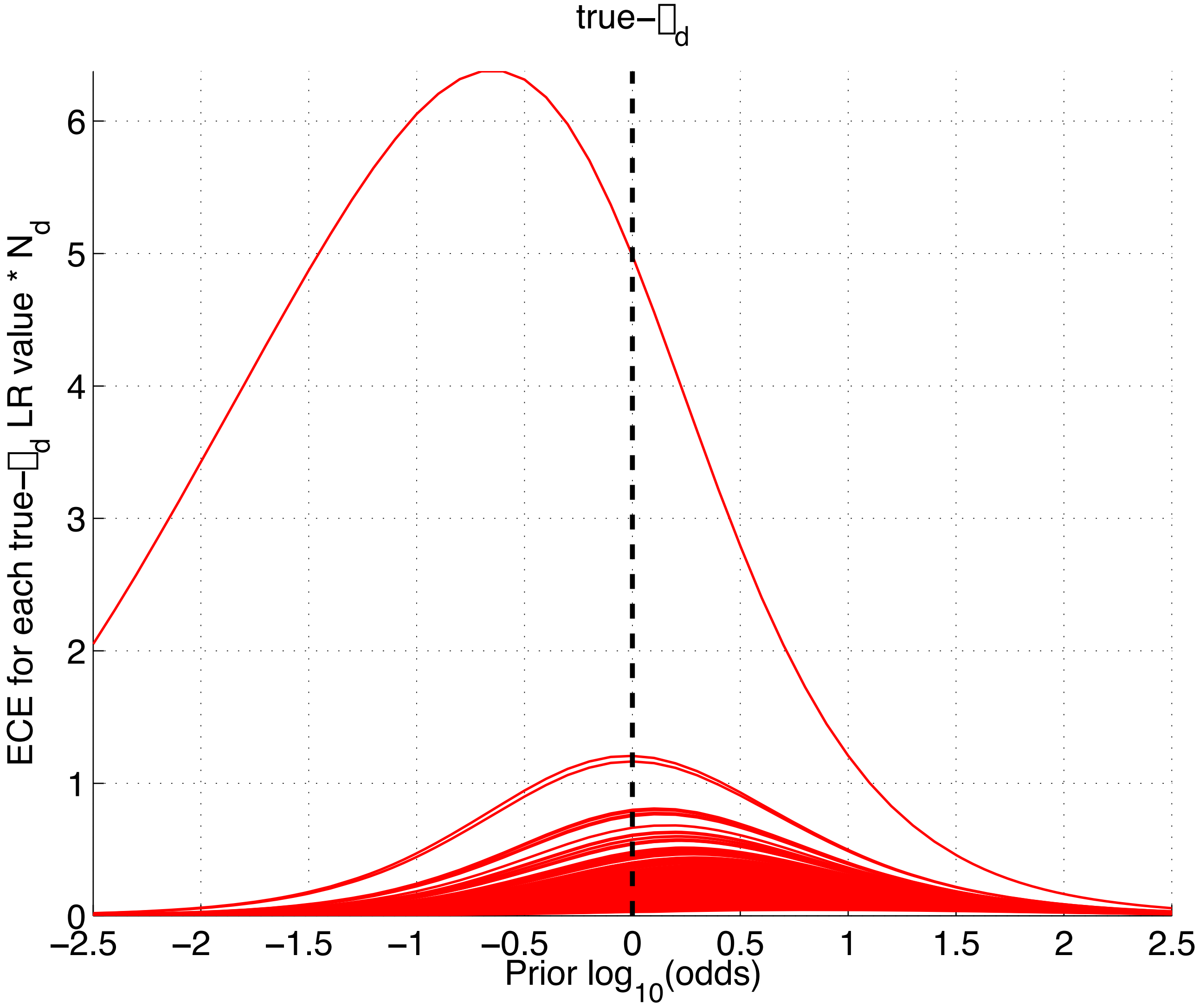

0.6
0.5
0.4
0.3
0.2
0.1
0
Empirical cross−entropy
−2.5
−2
−1.5
−1
−0.5
0
0.5
1
1.5
2
2.5
Prior $\log_{10}$(odds)

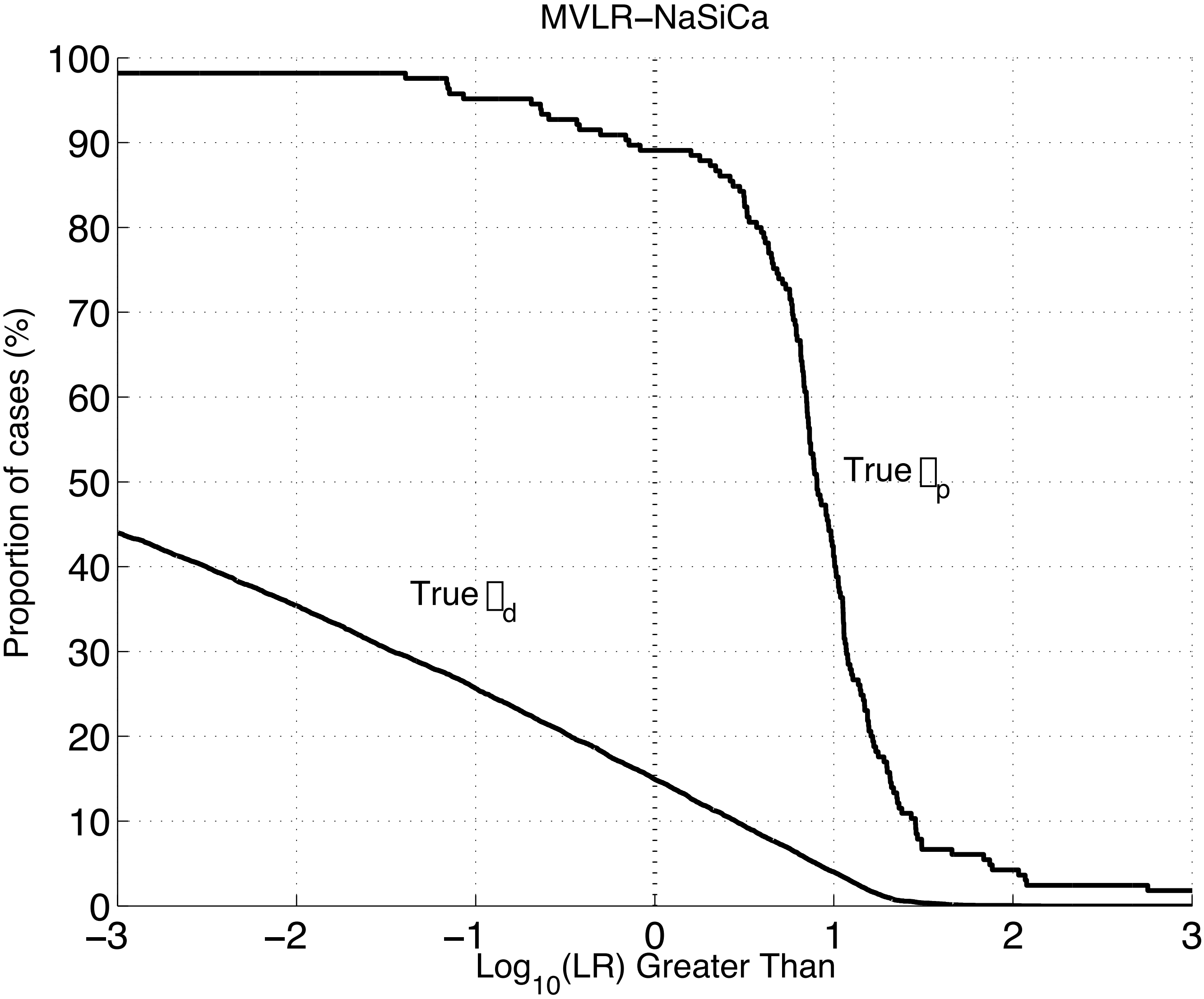

1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Empirical cross−entropy
−2
−1
0
1
2
Prior $\log_{10}$(odds)
LR values
After PAV
LR=1 always

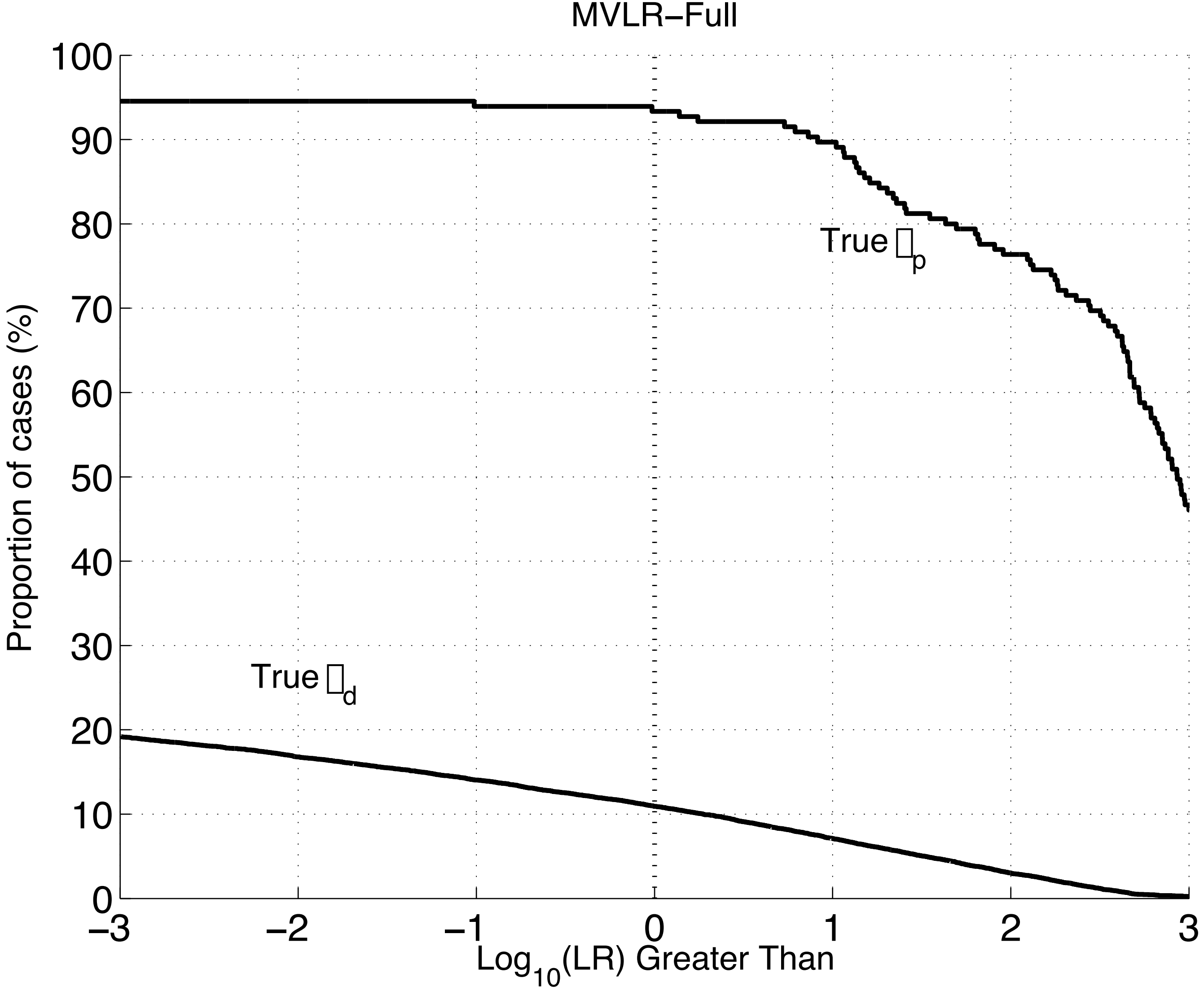

100
90
80
70
60
50
40
30
20
10
0
Proportion of cases (%)
−3
−2
−1
0
1
2
3
Log$_{10}$(LR) Greater Than
True H_d
True H_p

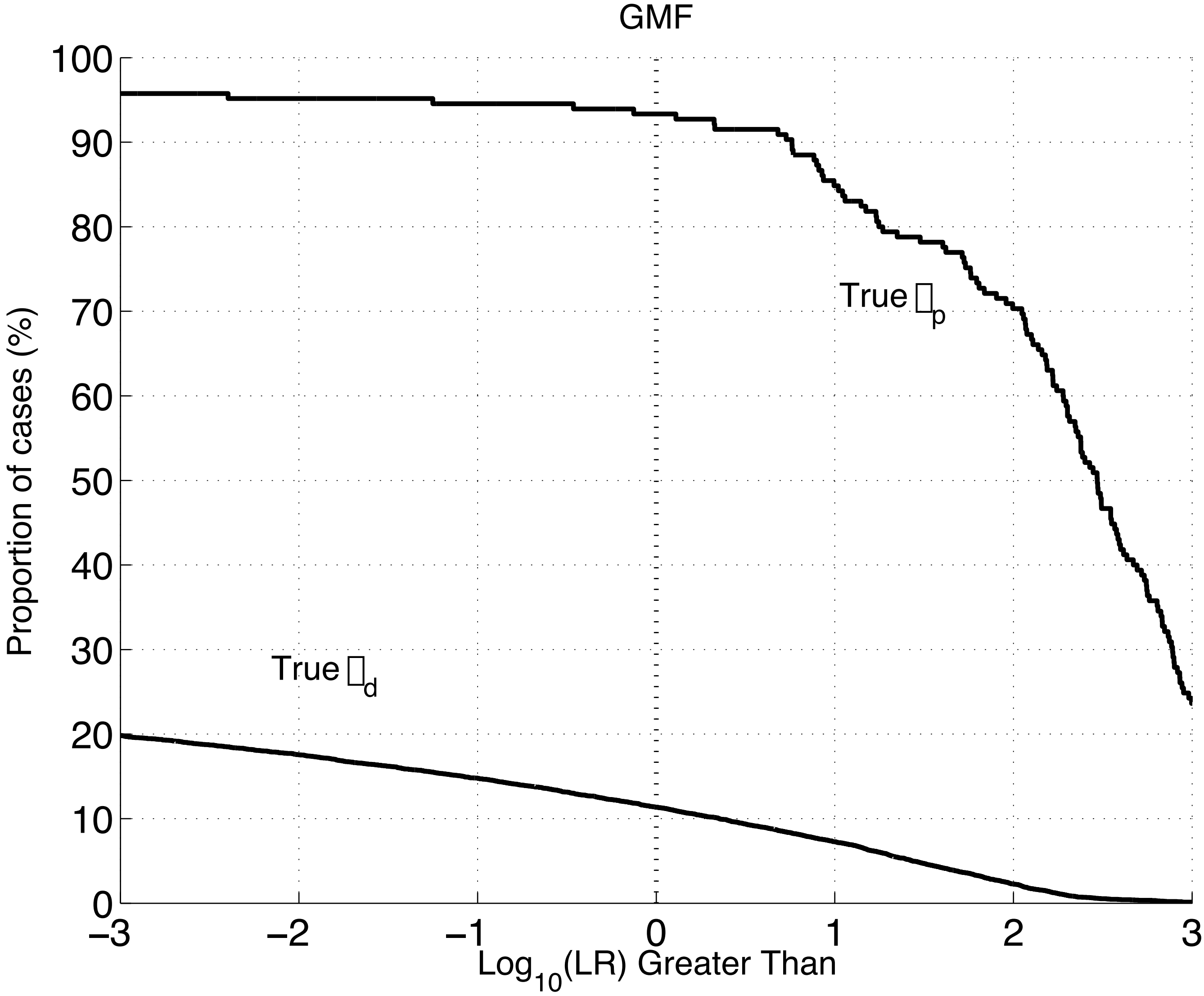

LR values
After PAV
LR=1 always
Empirical cross−entropy
Prior log$_{10}$(odds)
0
0.1
0.2
0.3
0.4
0.5
0.6
−2.5
−2
−1.5
−1
−0.5
0
0.5
1
1.5
2
2.5

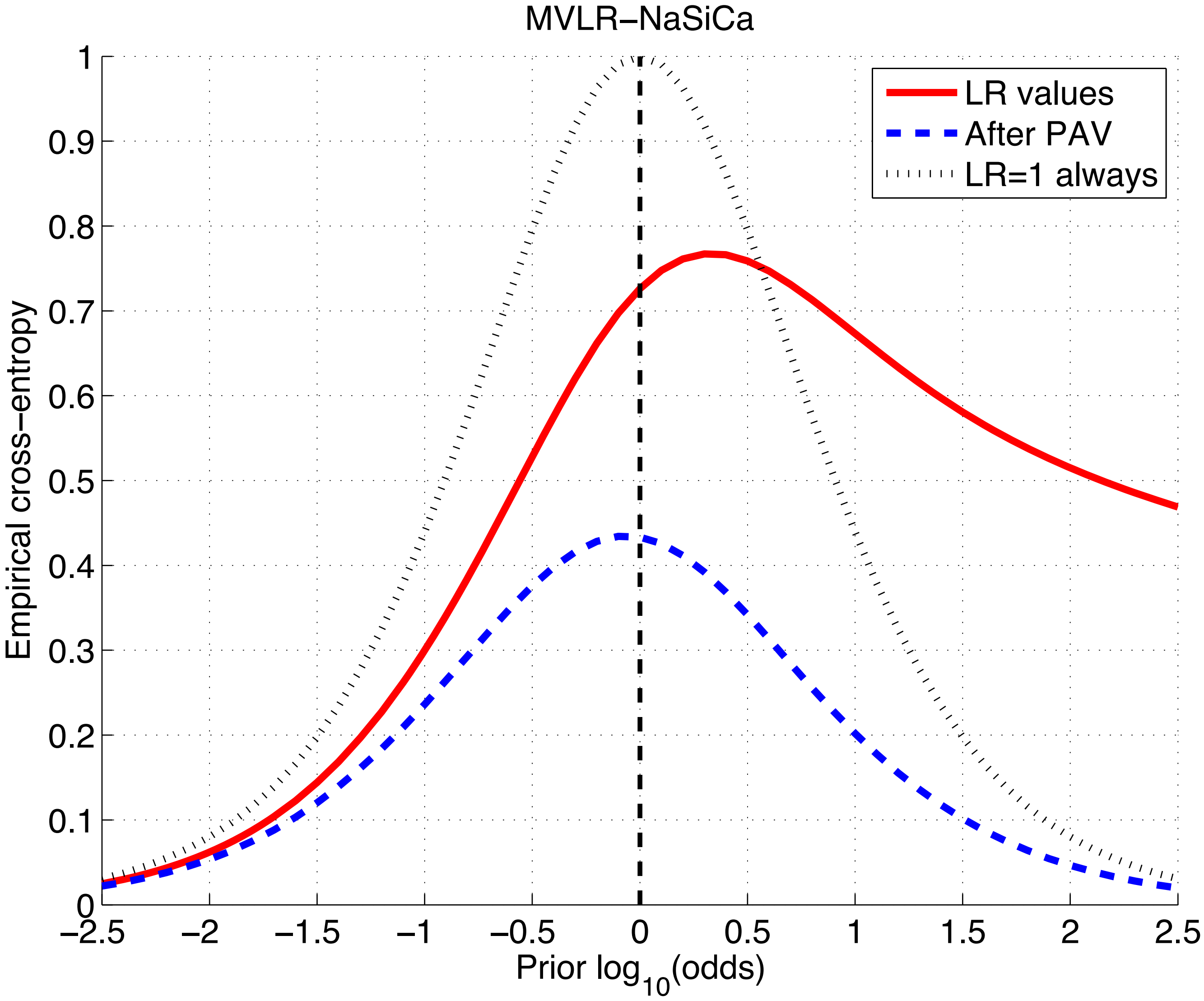

true−□p
ECE for each true−□p LR value * Np
0
1
2
3
4
5
6
−2.5
−2
−1.5
−1
−0.5
0
0.5
1
1.5
2
2.5
Prior log10(odds)

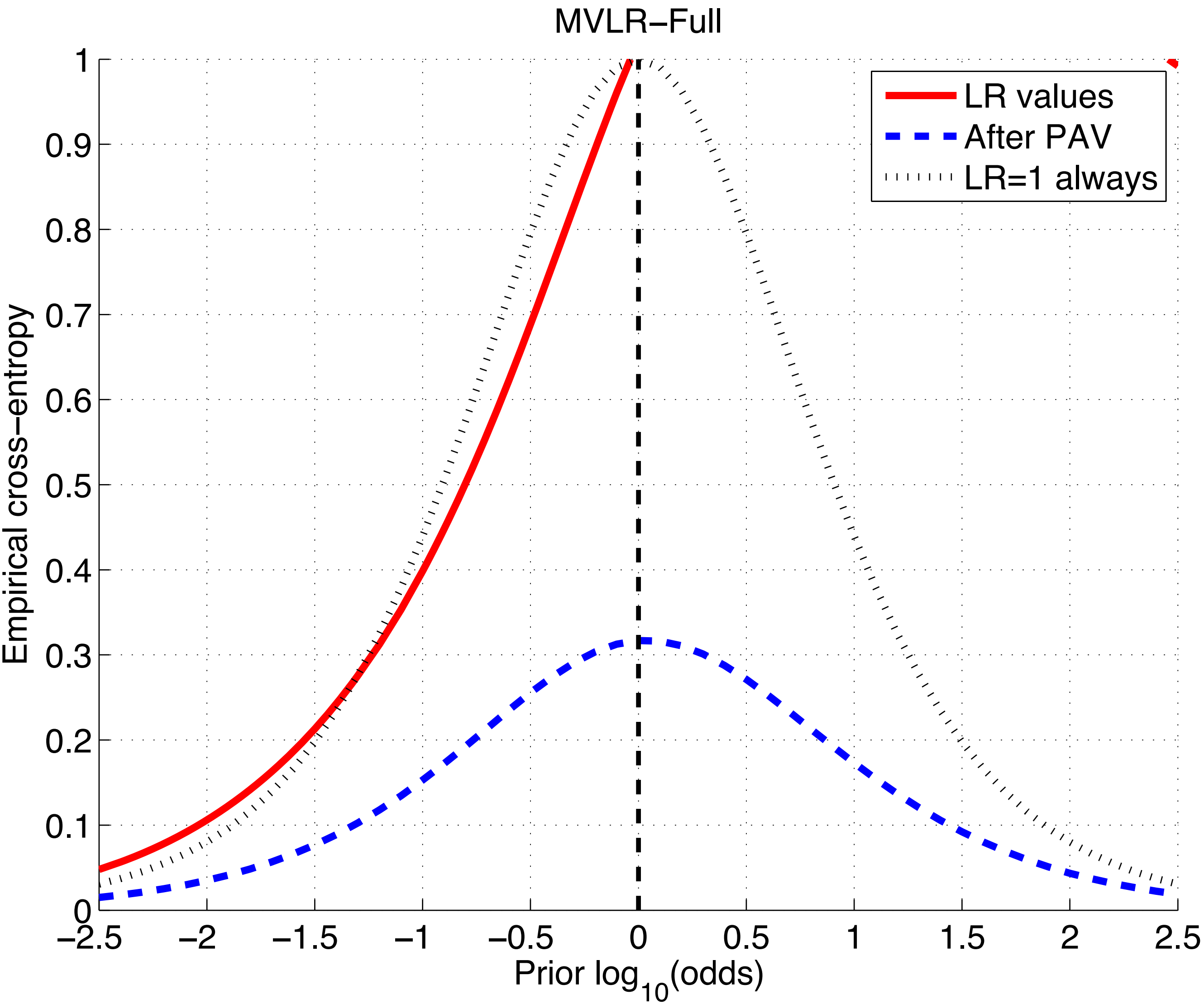

true−□$_d$
ECE for each true−□$_d$ LR value * N$_d$
Prior log$_{10}$(odds)
0
1
2
3
4
5
6
−2.5
−2
−1.5
−1
−0.5
0
0.5
1
1.5
2
2.5

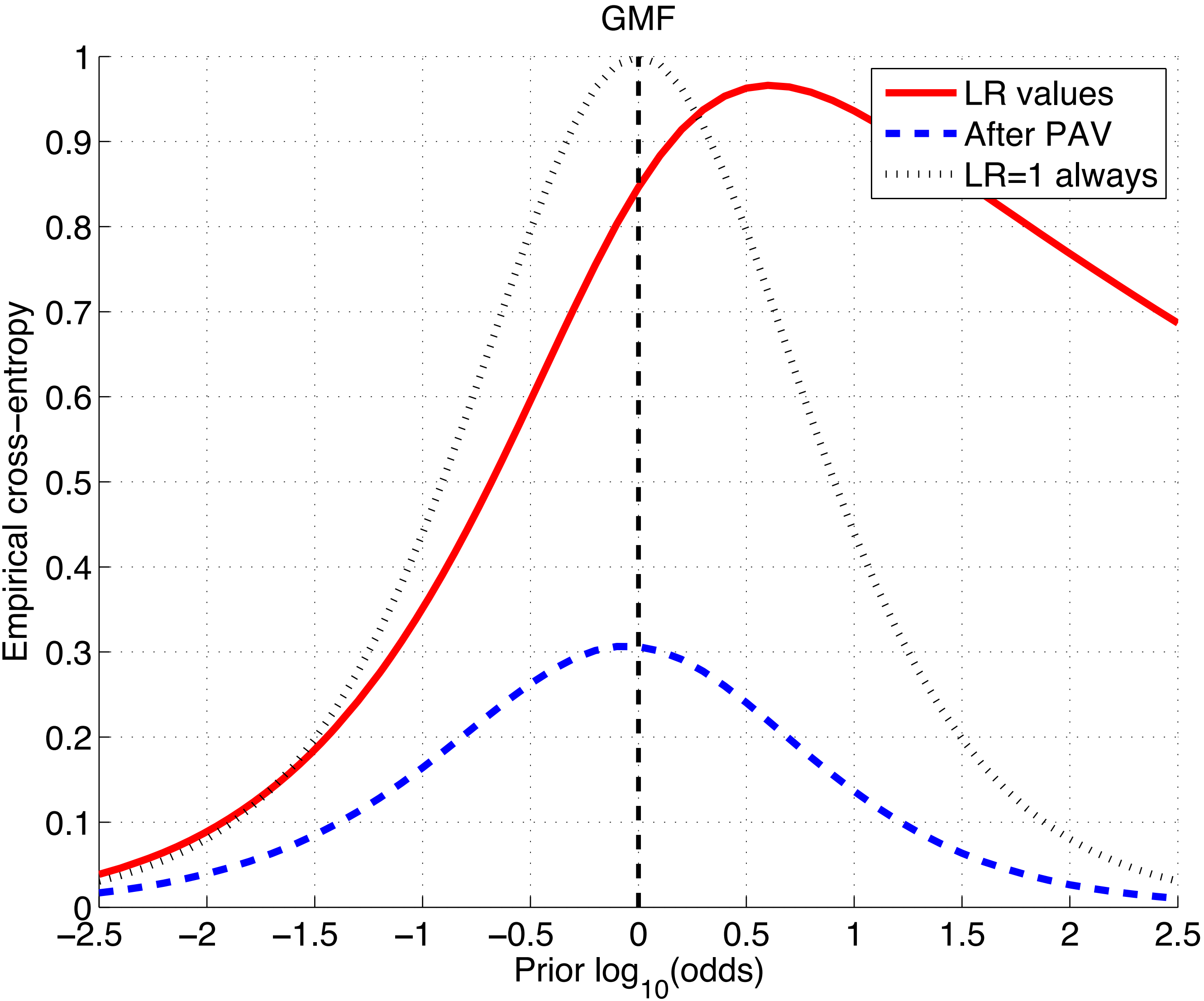

MVLR−NaSiCa
Proportion of cases (%)
100
90
80
70
60
50
40
30
20
10
0
−3
−2
−1
0
1
2
3
Log$_{10}$(LR) Greater Than
True □$_p$
True □$_d$

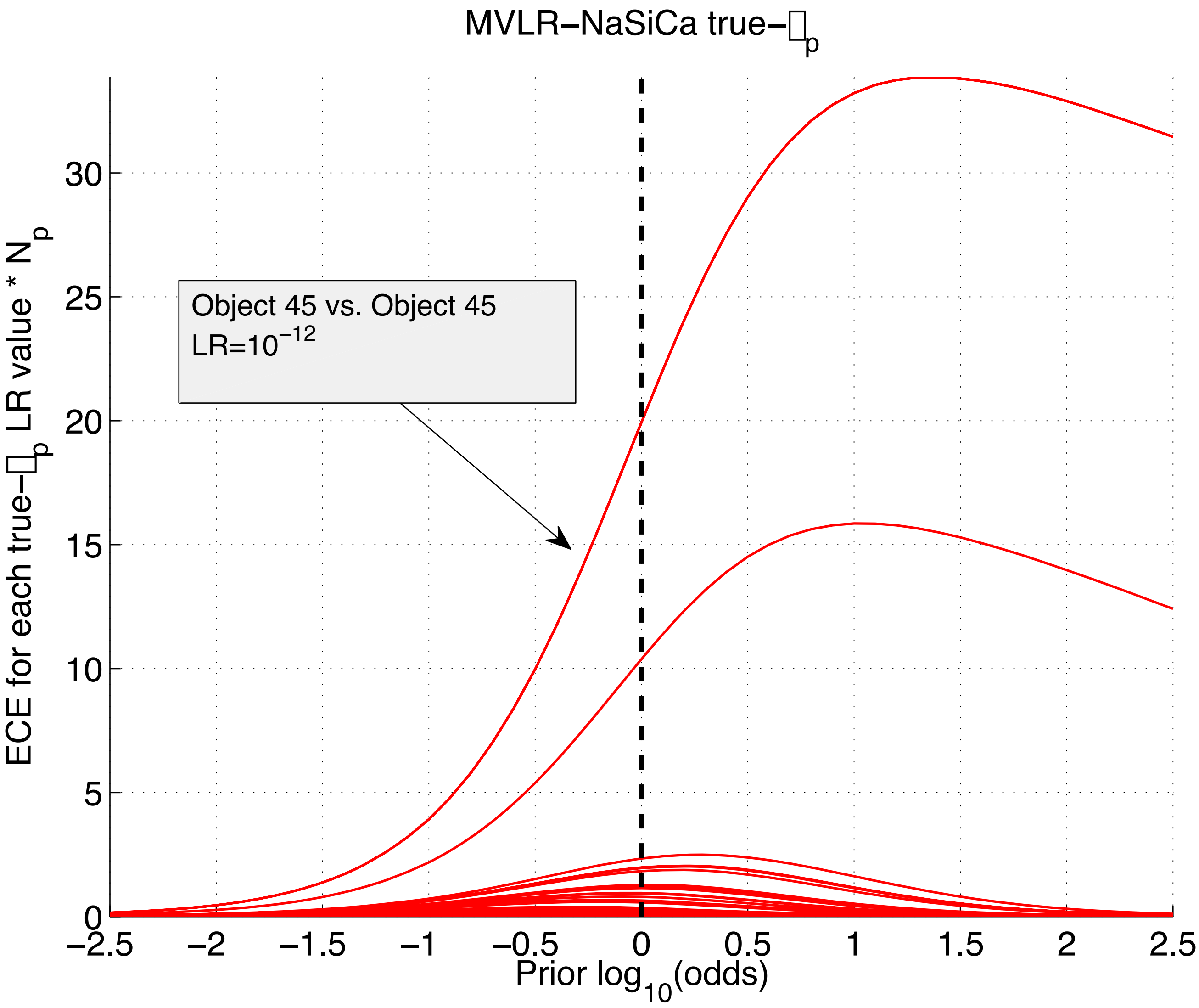

MVLR−Full
Proportion of cases (%)
100
90
80
70
60
50
40
30
20
10
0
−3
−2
−1
0
1
2
3
Log_{10}(LR) Greater Than
True □$_p$
True □$_d$

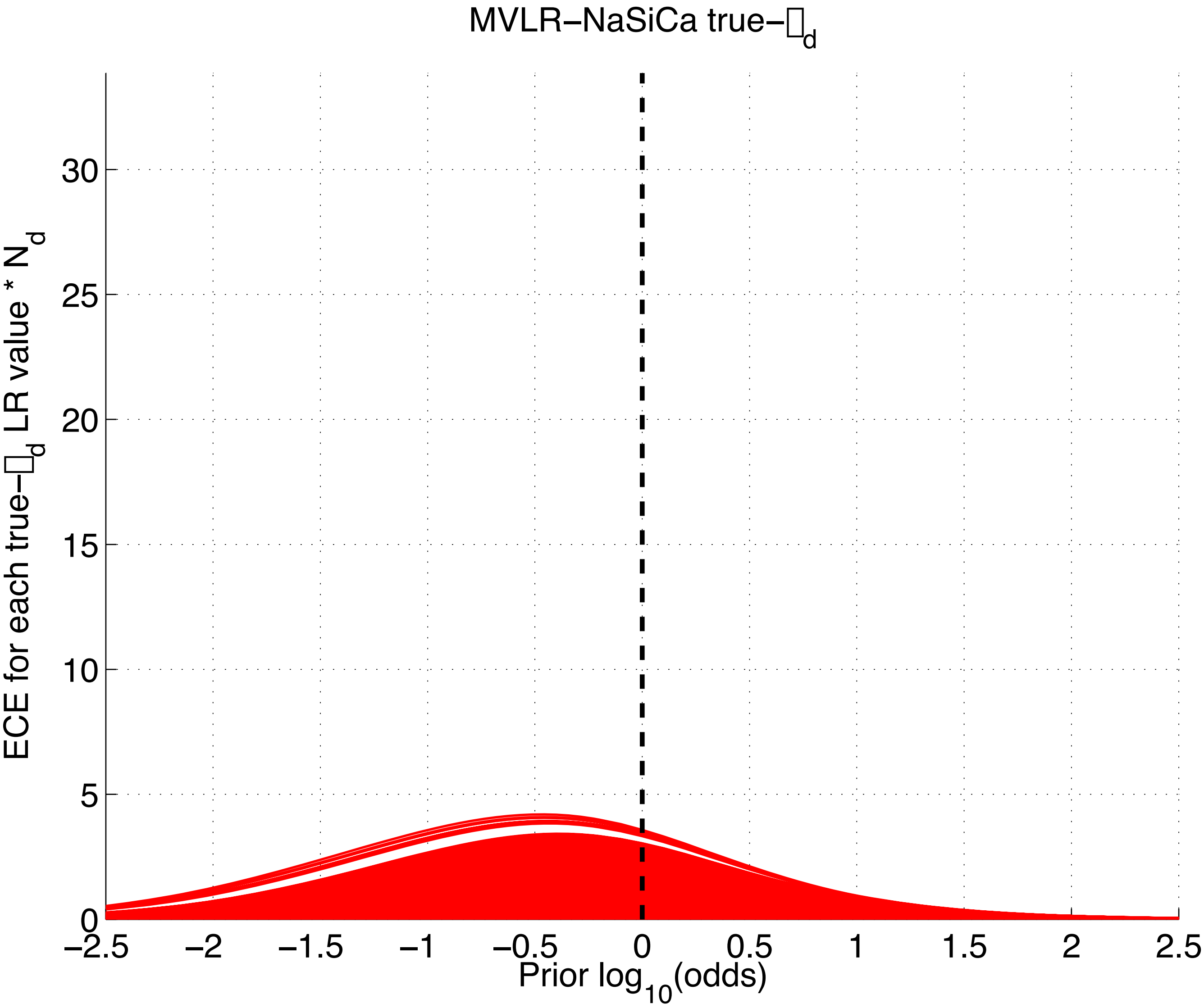

GMF
Proportion of cases (%)
100
90
80
70
60
50
40
30
20
10
0
True $\square_p$
True $\square_d$
−3
−2
−1
0
1
2
3
Log$_{10}$(LR) Greater Than

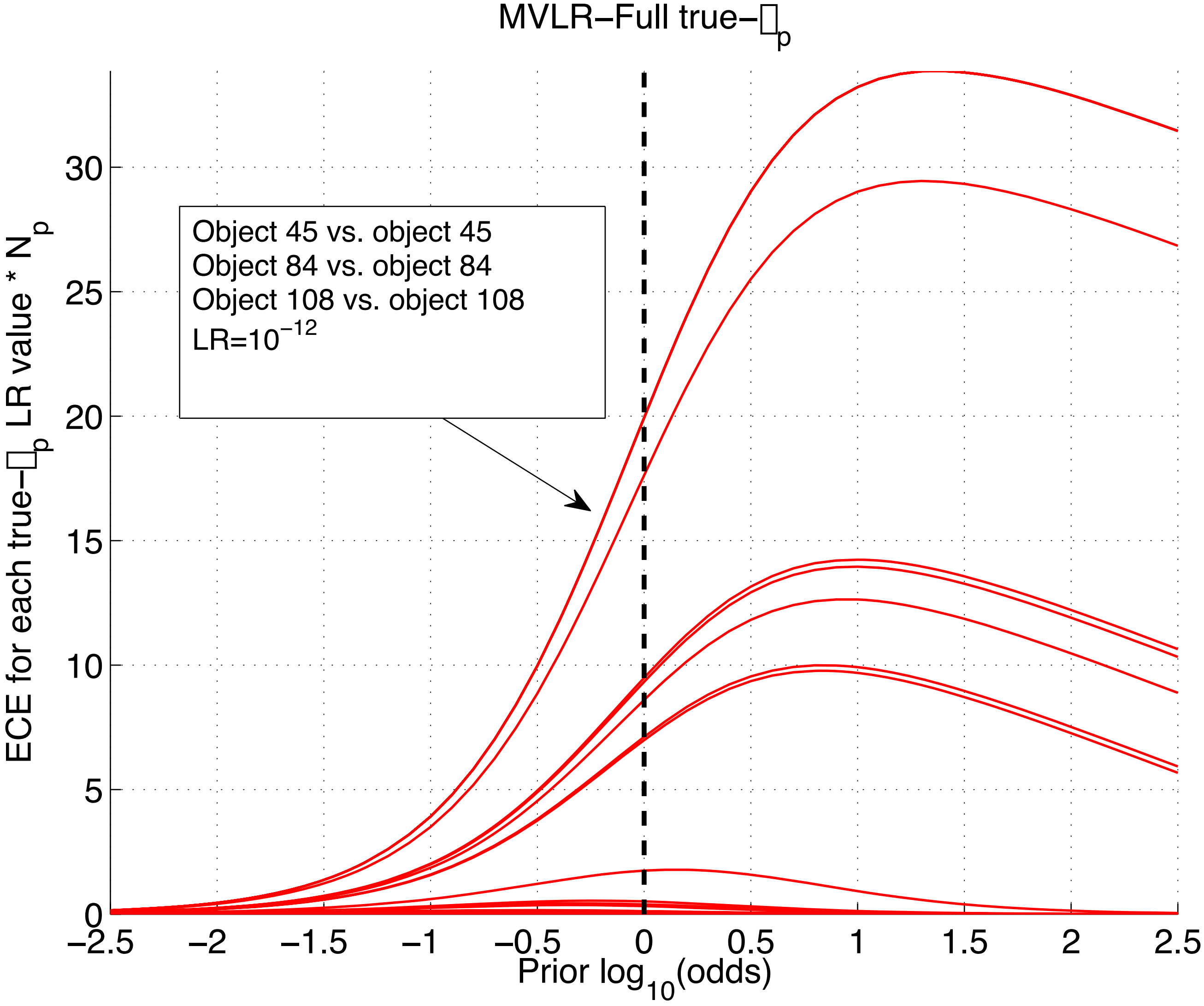

MVLR−NaSiCa
Empirical cross−entropy
Prior $\log_{10}$(odds)
LR values
After PAV
LR=1 always
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
−2.5
−2
−1.5
−1
−0.5
0
0.5
1
1.5
2
2.5

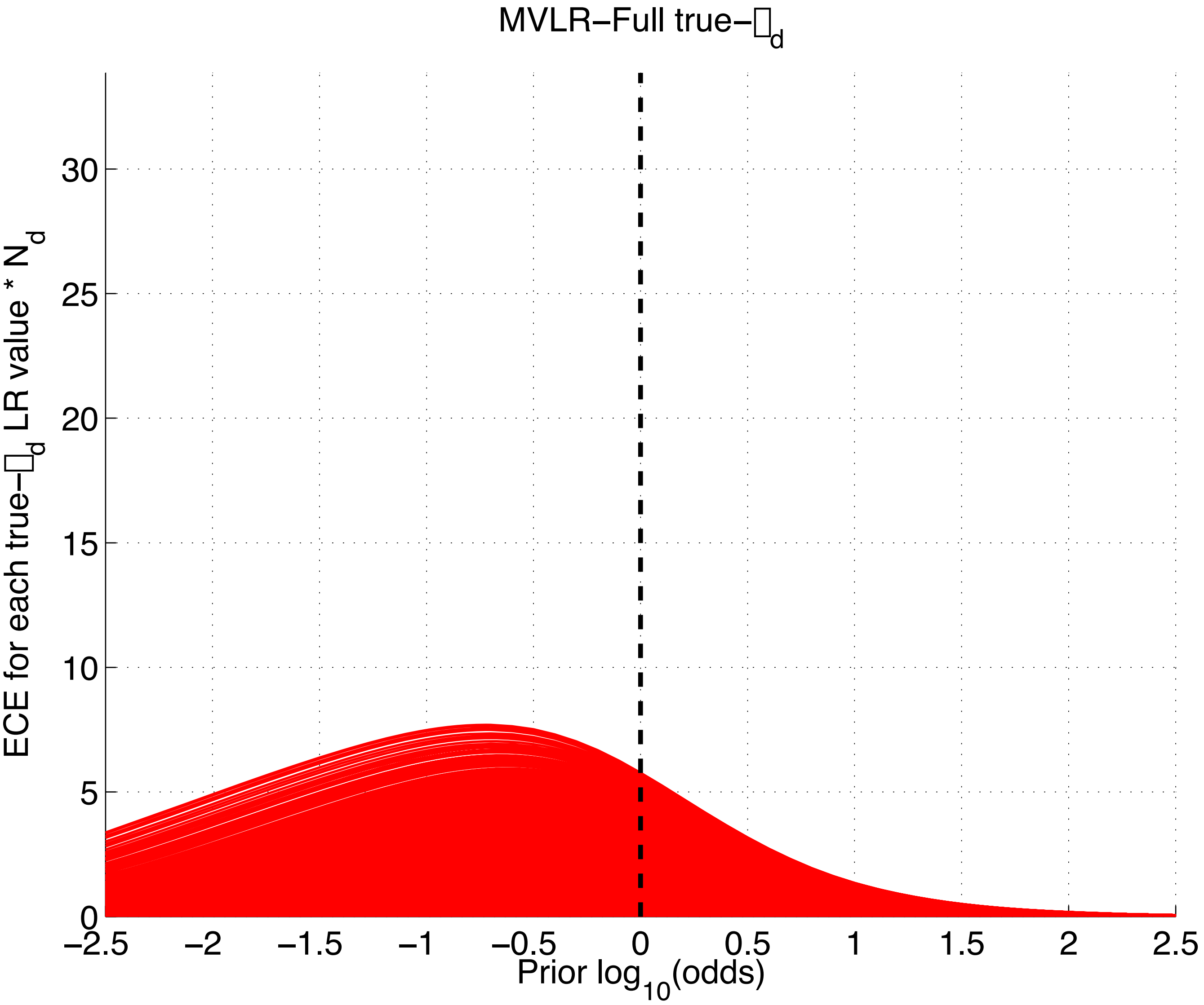

MVLR−Full
Empirical cross−entropy
Prior log10(odds)
LR values
After PAV
LR=1 always
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
−2.5
−2
−1.5
−1
−0.5
0
0.5
1
1.5
2
2.5

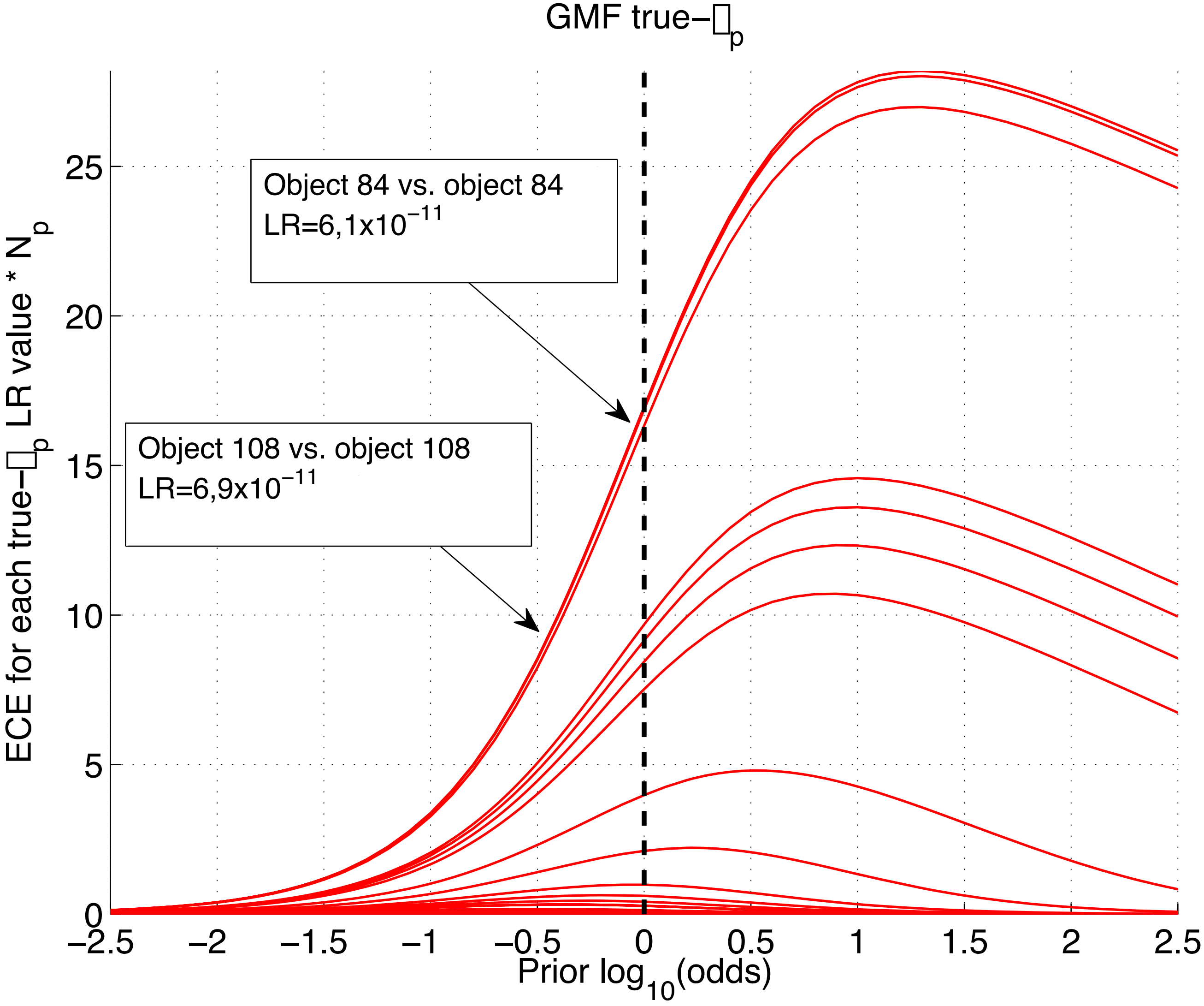

GMF
Empirical cross–entropy
Prior log$_{10}$(odds)
LR values
After PAV
LR=1 always
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
−2.5
−2
−1.5
−1
−0.5
0
0.5
1
1.5
2
2.5

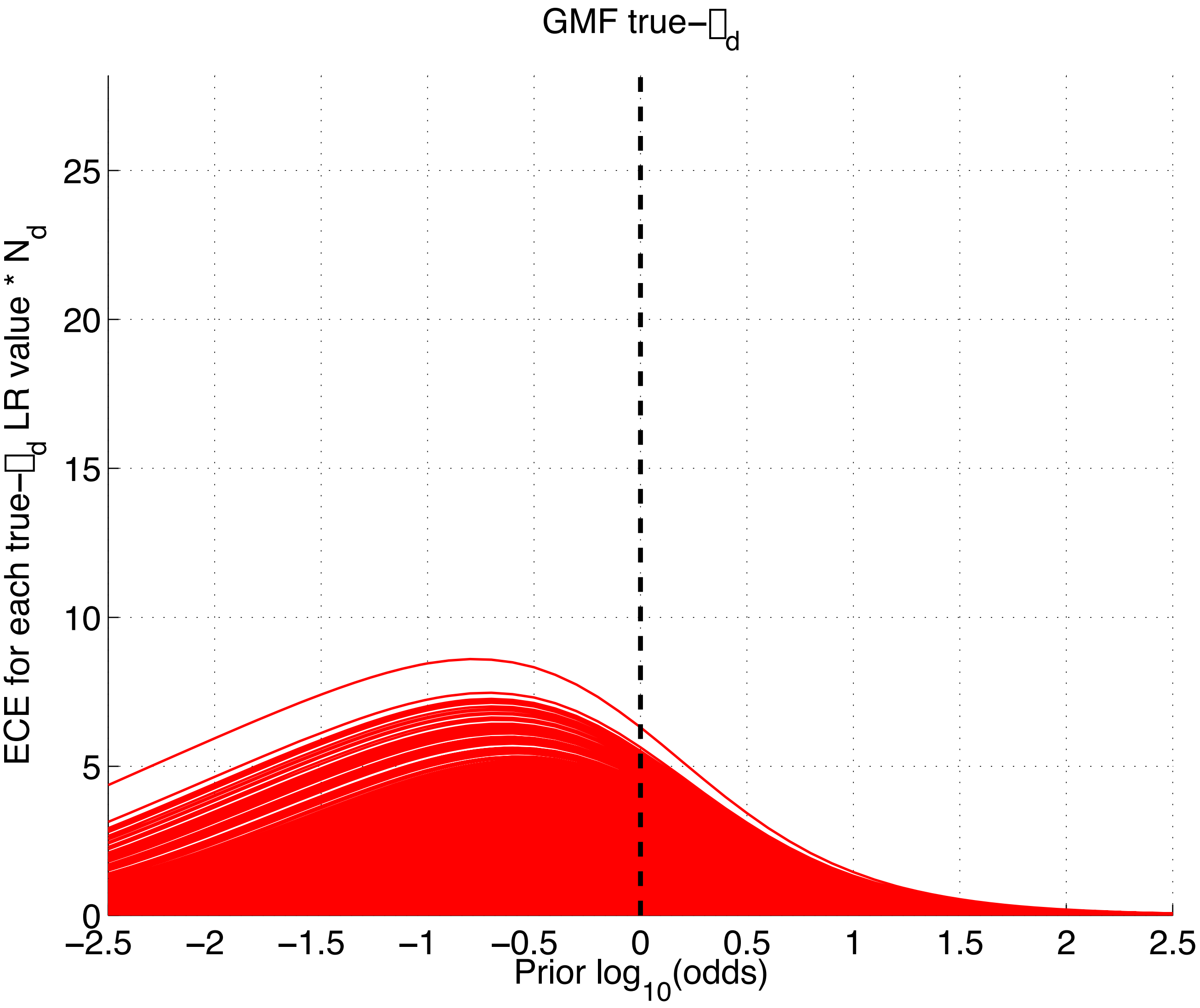

MVLR−NaSiCa true−□p
ECE for each true−□p LR value * Np
Prior log10(odds)
Object 45 vs. Object 45
LR=10−12
0
5
10
15
20
25
30
−2.5
−2
−1.5
−1
−0.5
0
0.5
1
1.5
2
2.5


MVLR−NaSiCa true−□$_d$
ECE for each true−□$_d$ LR value * N$_d$
30
25
20
15
10
5
0
−2.5
−2
−1.5
−1
−0.5
0
0.5
1
1.5
2
2.5
Prior log$_{10}$(odds)


MVLR−Full true−□$_p$
ECE for each true−□$_p$ LR value * N$_p$
Prior log$_{10}$(odds)
Object 45 vs. object 45
Object 84 vs. object 84
Object 108 vs. object 108
LR=10^{-12}
0
5
10
15
20
25
30
−2.5
−2
−1.5
−1
−0.5
0
0.5
1
1.5
2
2.5


MVLR–Full true–□$_d$
ECE for each true–□$_d$ LR value * N$_d$
Prior log$_{10}$(odds)
0
5
10
15
20
25
30
–2.5
–2
–1.5
–1
–0.5
0
0.5
1
1.5
2
2.5


GMF true−□p
ECE for each true−□p LR value * Np
Prior log10(odds)
Object 84 vs. object 84
LR=6,1x10−11
Object 108 vs. object 108
LR=6,9x10−11
0
5
10
15
20
25
−2.5
−2
−1.5
−1
−0.5
0
0.5
1
1.5
2
2.5


GMF true−□d
ECE for each true−□d LR value * Nd
Prior log10(odds)
0
5
10
15
20
25
−2.5
−2
−1.5
−1
−0.5
0
0.5
1
1.5
2
2.5