EVALUACIÓN DE SISTEMAS DE RECONOCIMIENTO FORENSES

JOSE JUAN LUCENA MOLINA

CORONEL (R) DE LA GUARDIA CIVIL

Fecha de recepción: 26/01/2023. Fecha de aceptación: 21/03/2023

RESUMEN

El eco del informe PCAST 2016 entre científicos forenses principalmente europeos con gran influencia en ENFSI (*European Network of Forensic Science Institutes*) se puede encontrar en un artículo publicado en la revista *Forensic Science International* en 2017. Los autores señalan dos términos clave que eligen como los que describen la mejora científica que la ciencia forense necesita experimentar hacia el futuro: inferencia lógica y calibración. En este artículo se ofrece un itinerario pedagógico tanto para la comprensión de aspectos clave de la teoría bayesiana de la decisión como para la comprensión de la noción de calibración y, en consecuencia, de la evaluación de los sistemas de reconocimiento forenses. El grupo de investigación AUDIAS, de la Universidad Autónoma de Madrid, ha contribuido científicamente al aprecio de esa noción en ENFSI y en el desarrollo de la Guía de ENFSI sobre conclusiones evaluativas.

Palabras clave: inferencia lógica; teoría bayesiana de la decisión; calibración.

ABSTRACT

The echo of the PCAST 2016 report among mainly European forensic scientists with great influence in ENFSI (*European Network of Forensic Science Institutes*) can be found in an article published in the journal *Forensic Science International* in 2017. The authors point to two key terms as descriptors of the scientific improvement that forensic science needs to experience into the future: logical inference and calibration. This article offers a pedagogical itinerary both for the understanding of key aspects of Bayesian decision theory and for the understanding of the notion of calibration and, consequently, of the evaluation of forensic recognition systems. The AUDIAS research group, from the Autonomous University of Madrid, has contributed scientifically to the appreciation of this notion in ENFSI and in the development of the ENFSI Guide on Evaluative Reporting.

Keywords: logical inference; Bayesian decision theory; calibration.

1. HACIA LA COMPRENSIÓN DE LA NOCIÓN DE CALIBRACIÓN

La insistencia del artículo de Evett et al. (2017) en la relevancia tanto en la inferencia lógica como de la calibración, a la hora de expresar la fuerza de la evidencia científica mediante una relación de verosimilitudes, está relacionada con la medición del rendimiento de los sistemas de reconocimiento de características forenses.

Para adentrarnos en la noción de calibración, seguiremos las explicaciones de uno de sus principales conocedores y difusores en el ámbito académico y profesional forense europeo, el doctor Daniel Ramos Castro, investigador del grupo AUDIAS, quien presentó su tesis doctoral en reconocimiento de voz forense bajo la cobertura del Convenio de Colaboración con la Guardia Civil con la Universidad Autónoma de Madrid en el año 2007. En la bibliografía se incluyen artículos de los ingenieros de telecomunicaciones Joaquín González Rodríguez y Daniel Ramos Castro, entre otros coautores, que contienen explicaciones más detalladas que el resumen pedagógico que se ofrece en este artículo.

El esquema que vamos a seguir es el siguiente:

- Algunas cuestiones clave de la teoría bayesiana de la decisión.
- Las denominadas reglas de puntuación (*scoring rules*).
- Distinción entre discriminación y calibración.
- La evaluación de sistemas de reconocimiento forenses, distinguiendo métodos que no evalúan la calibración (*DET*, Tippett), de los que sí la evalúan (C_{llr}, *APE*).

Antes de adentrarnos en el tema, es importante resaltar que este camino que queremos recorrer se necesita para superar las nociones de falsos positivos y negativos como caracterizadoras del funcionamiento de un sistema clasificador, nociones en las que se fundamenta el informe PCAST pero que se alejan del estado del arte en inferencia lógica, tal y como Evett y los demás coautores de su artículo resaltan.

2. ALGUNAS CUESTIONES CLAVE DE LA TEORÍA BAYESIANA DE LA DECISIÓN

Decidir es formar un juicio definitivo sobre algo dudoso y contestable. Para decidir es necesario evaluar el grado de duda o incertidumbre sobre lo que se decide. Centrándonos en un cotejo forense cualquiera y dentro del marco lógico de referencia al que alude el artículo de Evett, hay dos proposiciones alternativas entre las que hay que tomar una decisión: la que afirma que la coincidencia observada entre las características analizadas es consecuencia de que las muestras comparadas tienen un mismo origen (que denotamos por H_p) y la que asevera que es consecuencia del azar (que denotamos por H_d).

Al decidir, evaluamos el grado de incertidumbre sobre lo que se decide (es decir, en qué medida son ciertas cada una de las proposiciones mencionadas) de esta forma:

$$\frac{P(H_p \mid E, I)}{P(H_d \mid E, I)}$$

La letra *P* significa probabilidad, H_p y H_d ya las conocemos, *E* significa evidencia y, en nuestro caso, sería el resultado de la coincidencia entre las características comparadas (un 'match'), y la *I* es la información de contexto. La coma entre *E* e *I* significa

que ambos sucesos están interseccionados, es decir, ocurren a la vez. Por tanto, esa expresión puede enunciarse diciendo que es una relación entre dos probabilidades condicionales compuestas: (1) la probabilidad de que la hipótesis de la acusación sea cierta —que las características observadas y coincidentes en las muestras comparadas procedan del sospechoso— y (2) la probabilidad de que la hipótesis de la defensa sea cierta —que las características observadas y coincidentes en las muestras comparadas no procedan del sospechoso—, dada la información de contexto del caso.

Por tanto, la información de contexto del caso, al igual que la coincidencia observada, constituyen los sucesos que consideramos ciertos, conocidos. Y lo que queremos averiguar son las probabilidades de que cada una de las hipótesis sea cierta, relacionándolas entre sí por medio de una división. Si las probabilidades condicionales que nos interesan podemos expresarlas mediante un valor numérico entre *0* y *1*, al ser los sucesos de los que estamos interesados en conocer las probabilidades (las hipótesis de la acusación y de la defensa) complementarios —es decir, si la probabilidad del primero vale *x*, la del segundo, será *(1 - x)* — el rango posible de esa división es de *0* a infinito.

Ese rango es el propio de las apuestas. Además de las probabilidades de un suceso que cuantificamos entre *0* y *1*, podemos cuantificar las divisiones entre probabilidades que así, podrán valer entre *0* e infinito.

Ya hemos visto que esa incertidumbre se expresa así:

$$\frac{P(H_p \mid E, I)}{P(H_d \mid E, I)}$$

pero la decisión sobre las hipótesis de la acusación y de la defensa es responsabilidad del juez. Solo él puede obtener la probabilidad *a posteriori* porque solo él puede determinar la probabilidad *a priori*. Esto es el teorema de Bayes en forma de apuestas. Por tanto, conocer ese teorema y enunciarlo en el contexto de una inferencia forense, es decir, (1) con hipótesis sustentadas por cada una de las partes sobre el origen de las muestras; (2) con un resultado fruto de la aplicación de técnicas de análisis a dos muestras (dubitada e indubitada) tras compararlas entre sí (por ejemplo, un 'match', o sea, una coincidencia entre características); y (3) con una información de contexto determinada, conocida, que afecta a ambas probabilidades, nos lleva a la siguiente expresión:

$$\frac{P(H_p \mid E, I)}{P(H_d \mid E, I)} = \frac{P(E \mid H_p, I)}{P(E \mid H_d, I)} \times \frac{P(H_p \mid I)}{P(H_d \mid I)}$$

Esta expresión, que se deduce de la que está a la izquierda y de la que hemos hablado hasta ahora, es el desarrollo de las probabilidades condicionales aplicando la tercera ley de la probabilidad o ley del producto. Por tanto, es una expresión cuya relación entre sus términos es lógicamente deductiva.

$$\frac{P(H_p \mid E,I)}{P(H_d \mid E,I)}[1] = \frac{P(E \mid H_p,I)}{P(E \mid H_d,I)}[2] \times \frac{P(H_p \mid I)}{P(H_d \mid I)}[3]$$

Entre corchetes y con un número queremos ayudar a distinguir estos términos de la igualdad. No forman parte de la fórmula, sino que solo son indicadores para ayudar a entender la explicación. El [1] ya lo conocemos y lo llamamos "apuesta *a posteriori*" (apuesta porque es una división entre probabilidades sobre hipótesis y *"a posteriori*" porque para su cálculo de requiere, forzosamente, que resolvamos la parte de la derecha de la igualdad en la ecuación). La parte derecha de la igualdad en la ecuación está dividida en dos factores que hemos etiquetado con el [2] y el [3], respectivamente. El primer factor, o sea [2], se llama relación de verosimilitudes (conocido como LR por sus iniciales en inglés: *likelihood ratio*) y el segundo, o sea, [3], se llama "apuesta *a priori*".

Ahora ya estamos en condiciones de poder entender ulteriores explicaciones que aclaran importantes aspectos de la teoría de la decisión bayesiana. Nos valemos de un ejemplo expuesto por el profesor Daniel Ramos en una explicación pedagógica a sus alumnos de la Universidad sobre este tema: lo denomina el ejemplo del ladrón lapón rural.

Imaginemos que el juez ha decidido sobre las hipótesis mencionadas y obtiene este valor numérico:

$$\frac{P(H_p \mid E,I)}{P(H_d \mid E,I)} = \frac{20}{99}$$

Obviamos cómo se ha llegado a esa cifra para centrar el discurso en lo que queremos explicar. Quienes quieran saber cómo se ha alcanzado ese valor pueden consultar el Apéndice 1.

Habiendo logrado cuantificar esa división entre probabilidades condicionales compuestas, estamos en condiciones de tomar una decisión —eso es lo que haría el juez, que es a quién corresponde—.

El valor numérico indica que es, aproximadamente, cinco veces más probable que la proposición de la defensa sea cierta que lo sea la proposición de la acusación. Así, cualquier juez dictaminaría una sentencia o resolución favorable a la defensa.

Imaginemos que diera igual equivocarse decidiendo a favor de la acusación que a favor de la defensa —algo que, claramente, contradice el principio *in dubio pro reo*—. En ese caso, el umbral de decisión sería el valor igual a *1*. Cualquier cifra que superase ese valor, a favor de la acusación o de la defensa, inclinaría al juez a elegir la opción con mayor valor numérico sin más complicaciones.

Este ejemplo sencillo centra la decisión una vez que se determina el valor de la apuesta *a posteriori*, que algunos llaman, también, pronóstico *a posteriori*.

Sin embargo, hagamos una descomposición del pronóstico *a posteriori* de la siguiente manera:

$$\frac{P(H_p \mid E,I)}{P(H_d \mid E,I)}[1] = LR[2] \times \frac{P(H_p \mid I)}{P(H_d \mid I)}[3]$$

Si el valor numérico de [2], o sea el LR, supera al valor inverso de [3], el inverso de la apuesta *a priori*, resulta que la apuesta *a posteriori* —el resultado de la multiplicación— será mayor que *1*, y así: $P(H_p \mid E,I) > P(H_d \mid E,I)$. Vemos, por tanto, que la decisión se apoya en el valor del LR, pero el umbral de decisión depende de la probabilidad *a priori*. Por tanto, de acuerdo con nuestro marco lógico de referencia bayesiano, es clave ser conscientes de que solo el juez puede establecer el umbral del LR. En definitiva:

1. si el valor del $LR > P(H_d \mid I)/P(H_p \mid I)$, entonces $P(H_p \mid E,I) > P(H_d \mid E,I)$ —la tesis de la acusación tiene más probabilidades de ser cierta que la de la defensa—;
2. si el valor del $LR < P(H_d \mid I)/P(H_p \mid I)$, entonces $P(H_d \mid E,I) > P(H_p \mid E,I)$ —la tesis de la defensa tiene más probabilidades de ser cierta que la de la acusación—;
3. y si el *LR* es igual a $P(H_d \mid I)/P(H_p \mid I)$, entonces $P(H_p \mid E,I) = P(H_d \mid E,I)$. Ninguna hipótesis prevalece sobre la otra.

Siempre que se toma una decisión se pueden producir errores. En este caso, en que la decisión es binaria (entre dos hipótesis alternativas —si una es cierta, la otra es falsa— y exhaustivas —no hay más posibilidades hipotéticas que las dos contempladas—), si el juez decide a favor de H_p y la hipótesis verdadera es H_d, comete un falso positivo; y si decide a favor de H_d siendo la hipótesis verdadera H_d, comete un falso negativo.

Sabemos que, en el ámbito judicial, los errores posibles no son equivalentes. Un falso positivo es mucho peor que un falso negativo.

La teoría de la decisión bayesiana tiene en cuenta este aspecto y asigna un coste a cada tipo de error. La notación que emplearemos para describir los costes es la siguiente: llamamos *C* a la función matemática que representa un coste cualquiera y le asignamos dos subíndices separados por una coma —el primero representa la decisión y el segundo la realidad—. Así, $C_{d,p}$ significa el coste de decidir a favor de la hipótesis H_d, cuando la hipótesis verdadera es H_p (sería el coste de un falso negativo). Y $C_{p,d}$ significa el coste de decidir a favor de la hipótesis H_p, cuando la hipótesis verdadera es H_d (sería el coste de un falso positivo).

Por tanto, ahora contemplamos errores y costes por cometer errores y, en ambos casos, tenemos en cuenta las decisiones que sean falsos positivos y los falsos negativos.

La teoría bayesiana establece que si:

$$\frac{P(H_p \mid E,I)}{P(H_d \mid E,I)} > \frac{C_{p,d}}{C_{d,p}}$$

es decir, si la apuesta *a posteriori* es mayor, según el juez, que la relación de costes entre el falso positivo y el negativo, también fijados por el juez, entonces la decisión final será a favor de la hipótesis de la acusación; en otro caso será a favor de la hipótesis de la defensa.

Por ejemplo, si consideramos 10 veces más grave encarcelar a un inocente que soltar a un culpable:

$$\frac{C_{p,d}}{C_{d,p}} = \frac{10}{1}$$

entonces, en nuestro ejemplo:

$$\frac{P(H_p \mid E,I)}{P(H_d \mid E,I)} = \frac{20}{99} \not> \frac{10}{1}$$

(el signo $\not>$ quiere decir “no mayor que”)

y, por tanto, la decisión se toma a favor de H_d.

En un supuesto hipotético no respetuoso con el principio de presunción de inocencia en el que considerásemos 10 veces más grave soltar a un culpable que encarcelar a un inocente:

$$\frac{P(H_p \mid E,I)}{P(H_d \mid E,I)} = \frac{20}{99} > \frac{1}{10}$$

y, por consiguiente, la decisión se toma a favor de H_p.

Una decisión mal tomada genera un coste bayesiano o riesgo medio, el cual se obtiene empleando la regla de Bayes: este coste es una especie de residuo que depende de la bondad del clasificador.

$$\Re = P(H_p)C_{d,p}P(error \mid H_p) + P(H_d)C_{p,d}P(error \mid H_d)$$

Se puede demostrar matemáticamente que ese coste es mínimo, es decir, no se pueden tomar decisiones mejores que las de Bayes una vez fijada la apuesta *a posteriori*.

En resumen:

- El juez es el que siempre toma las decisiones.

- El coste mínimo se obtiene para probabilidades *a posteriori* fijas (o apuesta fija *a posteriori*). El modo en que funciona un sistema clasificador se determina fijando los valores de las probabilidades *a posteriori*.
- Si las probabilidades *a posteriori* cambian, el coste cambiará. Es decir, si cambiamos el modo de funcionamiento de un sistema el coste cambiará.
- Si distintos sistemas funcionan con las mismas probabilidades *a posteriori*, no podremos tomar mejores decisiones que las bayesianas.
- Todo esto se cumple si las probabilidades se corresponden con lo que sabemos del problema. Con otras probabilidades el coste será mayor. Si ignoramos lo que sabemos del problema, también.

3. LAS DENOMINADAS REGLAS DE PUNTUACIÓN (*SCORING RULES*)

Los sistemas biométricos comparadores, como, por ejemplo, el sistema automático de identificación dactilar, calculan una puntuación de similitud entre las características que se comparan en una búsqueda que se denomina, en inglés, *score* (puntuación).

El *score* será numéricamente mayor cuanto más apoye el sistema de clasificación la hipótesis de la acusación y viceversa, pero desde un *score* no se pueden calcular probabilidades. Sin embargo, a partir del *score*, se pueden calcular relaciones de verosimilitudes y, desde esos valores sí se pueden calcular probabilidades. Las relaciones de verosimilitudes se pueden usar también como un *score* porque cuanto mayor sea su valor, más similares son las muestras comparadas (dubitada e indubitada).

Los costes se asocian a las decisiones: castigan las decisiones incorrectas (a favor de H_p: falsos positivos, que los designamos mediante $C_{p,d}$; y a favor de H_d: falsos negativos, que los designamos mediante $C_{d,p}$), y pudieran hacerlo con las decisiones correctas, pero no tendría sentido, por lo que les asignamos el valor de coste igual a cero.

Imaginemos un caso en el que sea *10* veces más costoso encarcelar a un inocente que soltar a un culpable (respetuoso con la presunción de inocencia). Por tanto, $C_{p,d}=10$ y $C_{d,p}=1$. Obsérvese que ahora penalizamos cualquier error. El contexto imaginario en este ejemplo es un cotejo de voces y estamos realizando una toma de voz indubitada.

El coste elegido determina nuestro umbral de decisión:

$$Si\ \frac{P(H_p \mid E,I)}{P(H_d \mid E,I)}\ ha\ de\ ser\ > 10$$

$$P(H_p \mid E,I) > 0.91$$

El umbral exigido por el coste conlleva que, para que decidamos a favor de la hipótesis de la acusación, su probabilidad ha de ser superior al 91%.

En las gráficas siguientes de función de coste el coste igual a 0 indica que la decisión (H_p es cierta) es correcta, porque se supera un umbral de decisión relacionado con la probabilidad de que H_p sea cierta, dado que E e I lo sean. El umbral es 0.91 y si probabilidad de que H_p sea cierta lo supera, ya hemos señalado que no hay coste; en otro caso, el coste es igual a 1, el mínimo, porque la decisión ha sido correcta pero no respeta el criterio entre costes establecido (10 a 1) del modo más leve.

COSTES DE DECISIONES (figura 1)

Si se cumple H_p (el sospechoso es el autor de la toma dubitada) y, además:

- Si $P(H_p / E,I) > 0.91$, coste 0;
- Si $P(H_p / E,I) < 0.91$, coste 1.

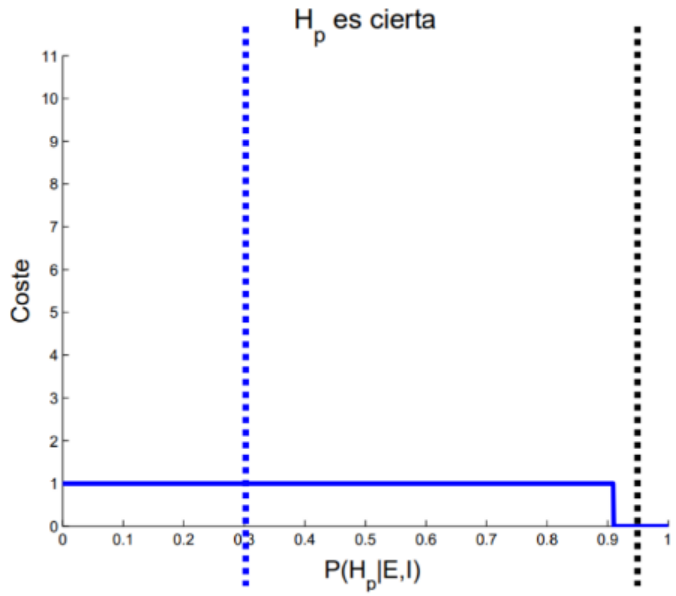


Figura 1 - Coste de decisiones.

COSTES DE DECISIONES (figura 2)

Si se cumple H_d (otra persona es el autor de la toma dubitada) y, además:

- Si $P(H_p / E,I) < 0.91$, coste 0;
- Si $P(H_p / E,I) > 0.91$, coste 10.

En la figura 2, el coste igual a 10 indica que la decisión es errónea (ahora H_d es cierta) porque se supera un umbral de decisión relacionado con la probabilidad de que H_p sea cierta, dado que E e I lo sean. El umbral es 0.91 y si probabilidad de que H_p sea cierta es menor, no hay coste; en otro caso, el coste es igual a 10, el máximo, porque la decisión ha sido errónea e infringe el criterio entre costes establecido (10 a 1) del modo más grave.

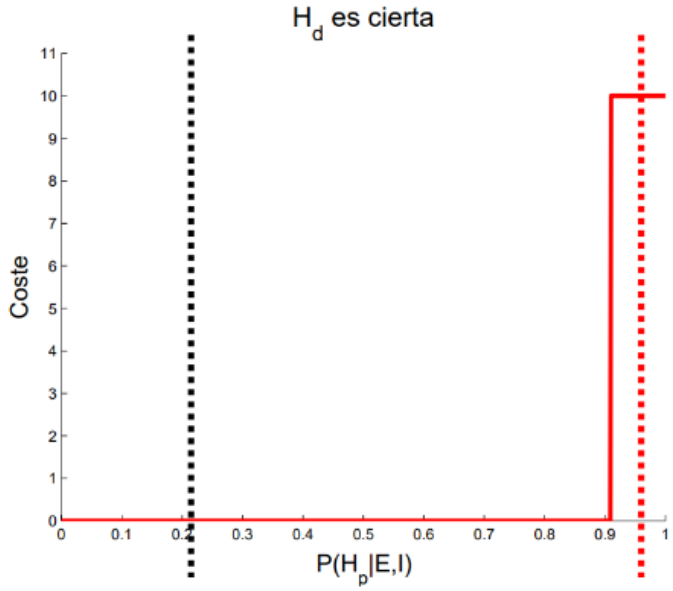


Figura 2 – Coste de decisiones.

Conviene advertir que el valor de las probabilidades a posteriori es indiferente en la región de decisión en ambas gráficas (figuras 1 y 2). Se penalizan las decisiones erróneas porque el coste se asocia a la decisión.

COSTES DE DECISIONES (figura 3)

- Si el coste se da a la decisión, se penalizan decisiones erróneas.
- El valor de la probabilidad a posteriori es indiferente en las regiones de decisión.
- Ejemplo: si H_d es cierta
  1. $P(H_p / E,I) = 0.92$, coste 10;
  2. $P(H_p / E,I) = 1$, coste 10.

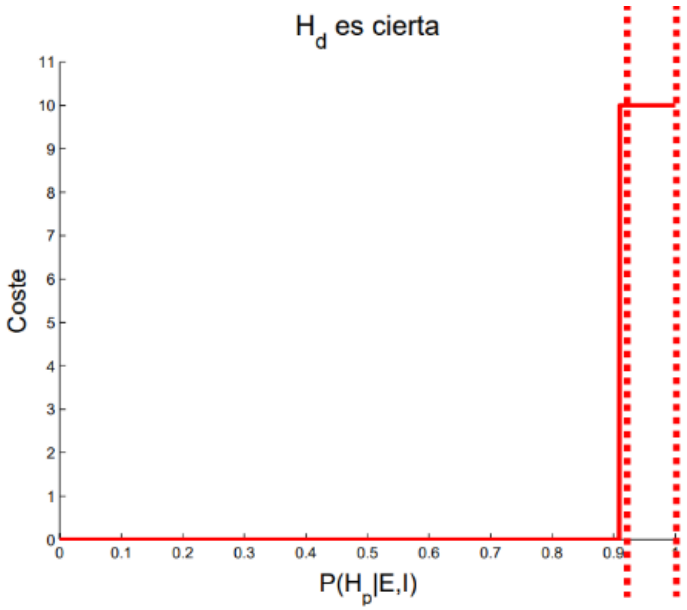


Figura 3 – Coste de decisiones.

Llegados a este punto, surge una cuestión interesante relacionada con la noción de identificación y la interpretación de la noción de coste si se comete un error al decidir, y esta vez con la noción de identificación propiamente interpretada porque se adjudica a una decisión judicial.

Si la probabilidad a posteriori de la hipótesis de la acusación, dada la evidencia y la información de contexto, es igual a 1 estamos diciendo que tenemos certeza de la veracidad de esa hipótesis. Y, si la probabilidad a priori no es ni uno ni cero, el LR tiene que valer forzosamente infinito al serlo la apuesta a posteriori.

Si la decisión a favor de la hipótesis de la acusación igual a 1 es errónea, nos preguntamos si no debería penalizarse esa decisión de una forma mucho más contundente que un coste de 10.

Y es aquí donde entran en juego las llamadas reglas de puntuación (scoring rules), porque una regla de puntuación asigna un coste a cada valor de probabilidad a posteriori. Es, por tanto, una función de la probabilidad a posteriori. La regla nos será útil si asigna costes altos a valores de probabilidad a posteriori muy erróneos, que es lo que queríamos conseguir.

Hay una regla de puntuación muy empleada en meteorología y para otras aplicaciones que se llama regla de Brier o del error cuadrático. La regla asigna un coste cuadrático que crece a medida que la probabilidad es más errónea y cuyo valor máximo es igual a 1.

REGLA DE BRIER / COSTE MÁXIMO = 1

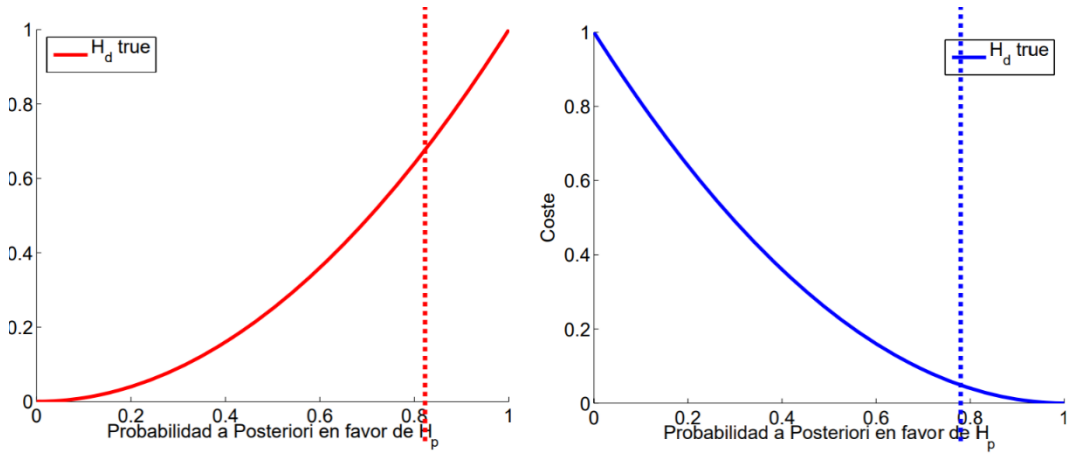


Figura 4 - Regla de Brier.

El coste de Brier se expresa matemáticamente de la siguiente forma:

$$C_{Brier} = \begin{Bmatrix} \left(1 - P\left(H_p \mid E, I\right)\right)^2, H_p\ cierta \\ \left(P\left(H_p \mid E, I\right)\right)^2, H_d\ cierta \end{Bmatrix}$$

Si nos fijamos en las gráficas de la regla de Brier, observamos que la probabilidad a posteriori a favor de Hp es el eje de abcisas y el coste de Brier el de ordenadas.

La fórmula cuadrática $(1 – P(H_p / E,I))^2$ se corresponde con la curva azul. El coste es 0 cuando $(1 – P(H_p / E,I))^2=0$, es decir, cuando la probabilidad a posteriori de la hipótesis de la acusación es igual a 1 y H_p es cierta. El coste es igual a 1 cuando la probabilidad a posteriori de la hipótesis de la defensa es igual a 1 y H_d es cierta.

La regla asigna un coste creciente, en forma cuadrática, a medida que el error en la decisión en función de la probabilidad a posteriori calculada en favor de H_p se aleje de la realidad. Tiene mucho sentido que si la probabilidad a posteriori que calculemos es más baja, el error que pudiéramos cometer tenga menos importancia, y al revés.

Hay otra regla muy utilizada, la logarítmica, que tiene la propiedad de que asigna un coste logarítmico. En lugar de que el coste sea igual a 1 en caso de error máximo, ahora se le asigna un valor infinito.

REGLA LOGARÍTMICA / COSTE MÁXIMO INFINITO / *A POSTERIORI* = 1 o erróneo, CASTIGO INFINITO

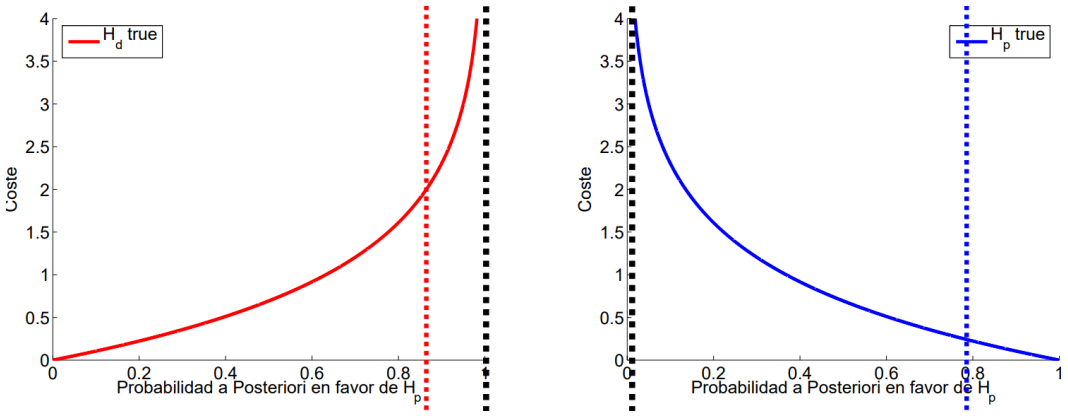


Figura 5 – Regla logarítmica.

Matemáticamente se expresa así:

$$C_{log} = \begin{Bmatrix} -\log\left(P\left(H_p \mid E,I\right)\right), H_p\ cierta \\ -\log\left(1 - P\left(H_p \mid E,I\right)\right), H_d\ cierta \end{Bmatrix}$$

Dando un paso más para acercarnos a las reglas de coste que se emplean en los sistemas automáticos de reconocimiento biométrico, necesitamos entender qué son las reglas estrictamente propias.

Apoyándonos en las nociones de variable aleatoria y esperanza matemática, imaginemos que necesitamos una regla de puntuación con respecto a la capacidad de predecir correctamente que pueda llover al día siguiente en nuestra ciudad. Podemos calcular la esperanza de la regla de puntuación con respecto a una probabilidad determinada *Q*. Llamamos *P* a la probabilidad de que llueva mañana. *Q* sería la probabilidad de que llueva mañana más fiable, porque se apoya en datos estadísticos de lluvia ese día en la ciudad en años anteriores. Lo que queremos conocer es la media de la regla de puntuación si asumimos que la proporción de días que llueve en nuestra ciudad viene dada por *Q*.

Esa media o esperanza matemática de la variable aleatoria consistente en la regla de puntuación R(P) se expresa así:

$$Q \cdot \Re(P) + (1 - Q) \cdot \Re(1 - P)$$

Una regla propia se minimiza si *P* y *Q* son iguales. Es decir, si nuestras probabilidades *a posteriori* se acercan a las probabilidades *a posteriori* de referencia, vamos bien y la regla de puntuación nos penaliza menos en media.

Una regla estrictamente propia solo se minimiza si *P* y *Q* son iguales. Es decir, si nuestras probabilidades *a posteriori* se acercan a las probabilidades *a posteriori* de referencia y, solo entonces, vamos bien y la regla de puntuación nos penaliza menos en media.

La regla logarítmica es estrictamente propia. La regla de Brier también lo es, pero no penaliza lo suficiente.

Esta es la expresión de la regla estrictamente propia logarítmica:

$$\underset{P}{\arg\min}[-Q \cdot \log(P) - (1 - Q) \cdot \log(1 - P)] = Q$$

La regla logarítmica esperada se puede promediar sobre un conjunto grande de *scores*. Se promedia ponderando sobre la probabilidad *a priori*. Se asumen probabilidades *a priori* de 0,5 y el número total resultante del cálculo se llama coste C_{llr}.

$$C_{llr} = \frac{1}{2 \cdot N_p} \sum_{i\ de\ H_p} log_2\left(1 + \frac{1}{LR_i}\right) + \frac{1}{2 \cdot N_d} \sum_{j\ de\ H_d} log_2(1 + LR_j)$$

Nos conviene elegir esta regla de coste porque (1) está basada en la regla logarítmica; (2) es estrictamente propia; (3) penaliza infinitamente las identificaciones erróneas; (4) se puede demostrar que C_{llr} es el coste medio de las decisiones tomadas por cualquier *a priori* y con cualquier coste.

Si minimizamos C_{llr}, reducimos el coste de las decisiones que tomamos. Sabemos, por la teoría de la decisión bayesiana, que el coste mínimo es el de Bayes. Por tanto, minimizar C_{llr} implica acercarnos a las decisiones de Bayes, es decir, las decisiones óptimas.

4. DISTINCIÓN ENTRE DISCRIMINACIÓN Y CALIBRACIÓN

Un sistema biométrico automático que calcule *scores* tiene la capacidad de discriminar a los usuarios que lo utilicen para, por ejemplo, permitir su entrada en un local. Tiene, igualmente, la capacidad de buscar en una base de datos a un individuo a partir de su información biométrica indubitada. Esa capacidad discriminativa se fundamenta en el estudio estadístico de los *scores*.

Dos conjuntos de scores *A* y *B* se dice que tienen la misma capacidad de discriminación si, para todo umbral en el conjunto *A*, podemos encontrar un umbral en el conjunto *B* con el mismo porcentaje de falsos positivos y falsos negativos.

Veamos dos ejemplos de conjuntos de *scores* en los que se conserva la capacidad de discriminación, aunque varíen las puntuaciones.

DISCRIMINACIÓN (caso 1)

Ejemplo: una transformación consistente en sumar un número a todos los *scores* (desplazamiento), no altera la discriminación.

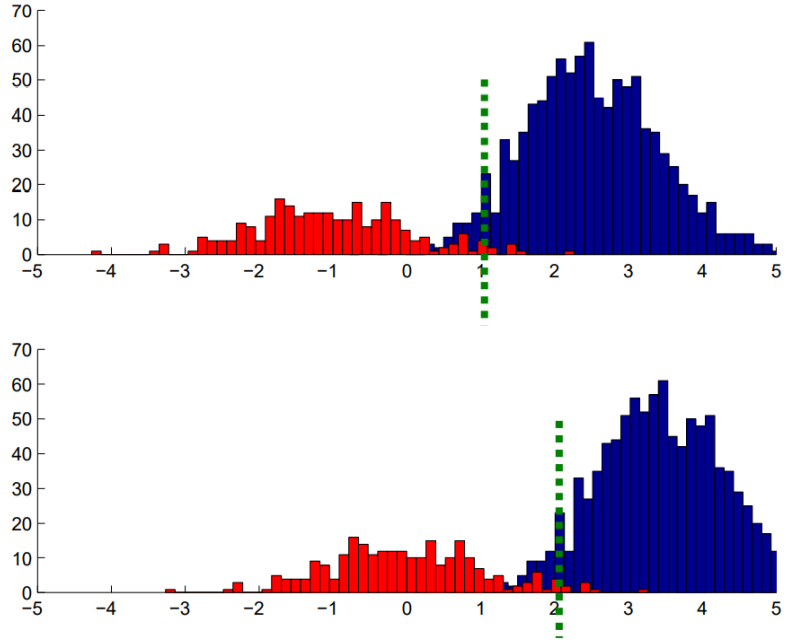


Figura 6 - Desplazamiento.

DISCRIMINACIÓN (caso 2)

Ejemplo: una transformación consistente en multiplicar todos los *scores* (escalado), no altera la discriminación.

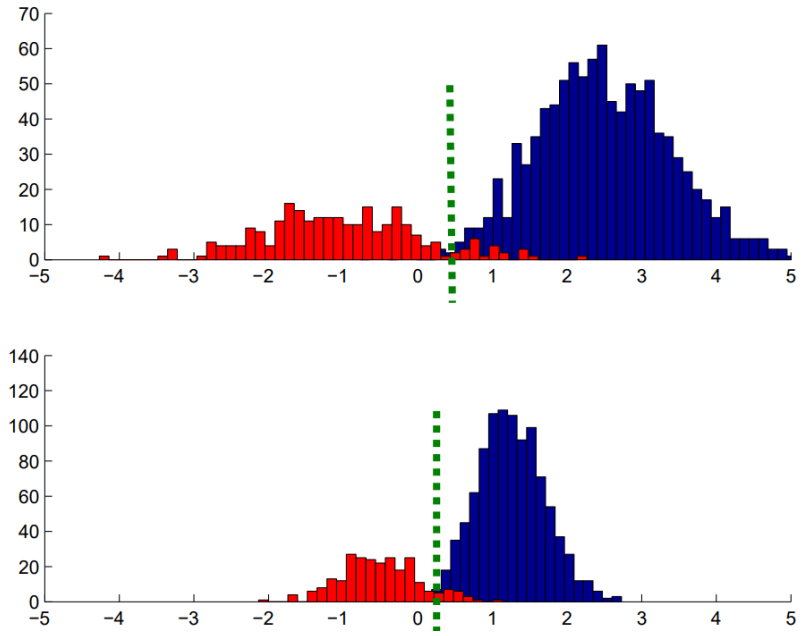


Figura 7 - Escalado.

Por tanto, cualquier alteración monótona conserva la capacidad de discriminación. Lo que altera esa capacidad es cualquier transformación que cambie el orden de los *scores* entre sí.

Por tanto, la calibración de los scores repercute en las prestaciones del sistema. Las reglas de puntuación estrictamente propias tienen en cuenta esto: cuanto más se alejen las probabilidades a posteriori de los resultados ciertos, mayor penalización infligirá la regla de decisión. Eso es justo lo que hace el coste C_{llr}. Podemos entender la noción de calibración como el castigo o penalización que se nos impone por lo que no es estrictamente efecto de la capacidad de discriminación propia del sistema clasificatorio, que será siempre limitada.

Imaginemos el siguiente ejemplo en el disponemos de dos casos diferentes: el caso *A* en el que la hipótesis H_p es cierta; y el caso *B* en el que la hipótesis H_d es cierta. Disponemos, además, de dos conjuntos de valores de *LR*: en el conjunto 1 obtenemos un valor de 0.1 para el caso *A* y 10 para el caso *B*; en el conjunto 2 obtenemos un valor de 10 para el caso *A* y 1000 para el caso *B*. Ambos conjuntos separan de la misma forma. Podemos trazar una frontera entre los valores de *LR* de ambos conjuntos y comprobar que separan de la misma forma, por tanto, permiten una plena discriminación, es decir, el mismo poder discriminante.

Sin embargo, nos conducen a probabilidades a posteriori muy diferentes. Por ejemplo, si la probabilidad a priori es 0.5, las probabilidades a posteriori serán las siguientes: para el conjunto 1: 9.1% para el caso *A* y 91,1% para el caso *B*; y para el conjunto 2: 91,1% para el caso *A*, 99,9% para el caso *B*.

Observamos que estamos dando un 91,1% de probabilidad *a posteriori* de que el sospe-choso sea la fuente de la muestra dubitada cuando el sospechoso es inocente. Lógi-camente, se nos va a penalizar. ¡Y todo eso con la misma discriminación! Este es el grave problema de la falta de calibración de los *scores*.

¿Podemos encontrar el mejor sistema clasificatorio que conserve la discriminación? Ese sistema diríamos que está bien calibrado. No se trata de que no falle al clasificar —porque todo sistema real clasifica limitadamente— sino que no lo haga por descalibración de los *scores*, es decir, solo por su inherente limitación clasificatoria. La transformación monótona que minimice una regla de puntuación estrictamente propia será la que nos aporte el mejor sistema que conserve la discriminación. Existe un algoritmo para conseguir esto que se llama por sus siglas en inglés: *PAV*. De este modo, obtenemos *scores* perfectamente calibrados que conservan la capacidad de discriminación de los *scores* originales.

Definimos una descomposición $C_{llr} = minC_{llr} + calC_{llr}$, donde $minC_{llr}$ es el valor de C_{llr} obtenido para el sistema óptimo que conserva la capacidad de discriminación. Lo obtenemos transformando los scores bajo análisis mediante el algoritmo *PAV* y calculando el C_{llr} de los scores transformados. El coste $calC_{llr}$ se obtiene así: $calC_{llr} = C_{llr} - minC_{llr}$.

El coste C_{llr} sabemos que una penalización por probabilidades a posteriori erróneas. Por tanto, nos informa sobre la bondad del sistema. Un C_{llr} alto es un sistema malo y viceversa. El coste $minC_{llr}$ es la penalización óptima que conserva el nivel de discriminación. Eso quiere decir que, si hubiésemos calibrado los scores, hubiésemos obtenido un $C_{llr} = minC_{llr}$.

El coste $calC_{llr}$ es la diferencia con el nivel óptimo de discriminación alcanzable debido a la falta de calibración.

Como nos alejamos de las probabilidades a posteriori de referencia -las de mínima discriminación alcanzable-, hay una penalización debido a la falta de calibración.

Ya tenemos todos los costes necesarios que nos permiten abordar la siguiente etapa, que es la que realmente buscábamos: cómo evaluar un sistema de reconocimiento forense que calcula relaciones de verosimilitudes en cuanto a su rendimiento.

5. LA EVALUACIÓN DE SISTEMAS FORENSES

Los sistemas forenses que evalúan la evidencia (por ejemplo, un “match”) mediante relaciones de verosimilitudes (LR) pueden, a su vez, ser evaluados en cuanto a su rendimiento.

Sabemos que la discriminación es importante, es decir, que el sistema pueda distinguir, eficazmente, parejas de muestras dubitadas e indubitadas comparadas entre sí cuando la hipótesis de la acusación H_p sea cierta de parejas de muestras análogas comparadas entre sí cuando la hipótesis de la defensa H_d sea cierta.

No obstante, sabemos que eso no es suficiente para garantizar el buen rendimiento del sistema que intentamos evaluar. Las probabilidades *a posteriori* que el juez obtenga pueden verse muy sesgadas por el efecto de la descalibración de las relaciones de verosimilitudes.

Para reglas estrictamente propias (como C_{llr}), calibrar significa acercarse a las probabilidades a posteriori de referencia. Si esas probabilidades a posteriori son de coste mínimo, estaremos acercándonos a las decisiones bayesianas.

Se puede demostrar que C_{llr} es el coste medio de las decisiones tomadas, por tanto calibrar significa acercarse a las decisiones de coste mínimo, es decir, minimizar $calC_{llr}$ convierte a los scores en algo más parecido a un *LR*.

La consecuencia de todo esto es que medir la calibración de los scores es también importante, no solo su discriminación.

Los sistemas de reconocimiento de patrones automáticos se han evaluado mediante unas curvas que muestran su rendimiento en discriminación que se llaman *DET* (*Detection Error Tradeoff*).

Se trata de un gráfico de tasas de error para sistemas de clasificación binarios. Sus ejes representan las probabilidades de falsos positivos (eje de abcisas) y falsos negativos (eje de ordenadas).

Los valores que configuran la curva están expresados log arítmicamente, convirtiendo la curva *ROC* (*Receiving Operating Characteristic*) con valores lineales en una recta, facilitando así la visualización de la separación entre las curvas de distintos sistemas.

La curva *DET* representa la relación entre las probabilidades de falsos positivos y falsos negativos para cualquier umbral de discriminación.

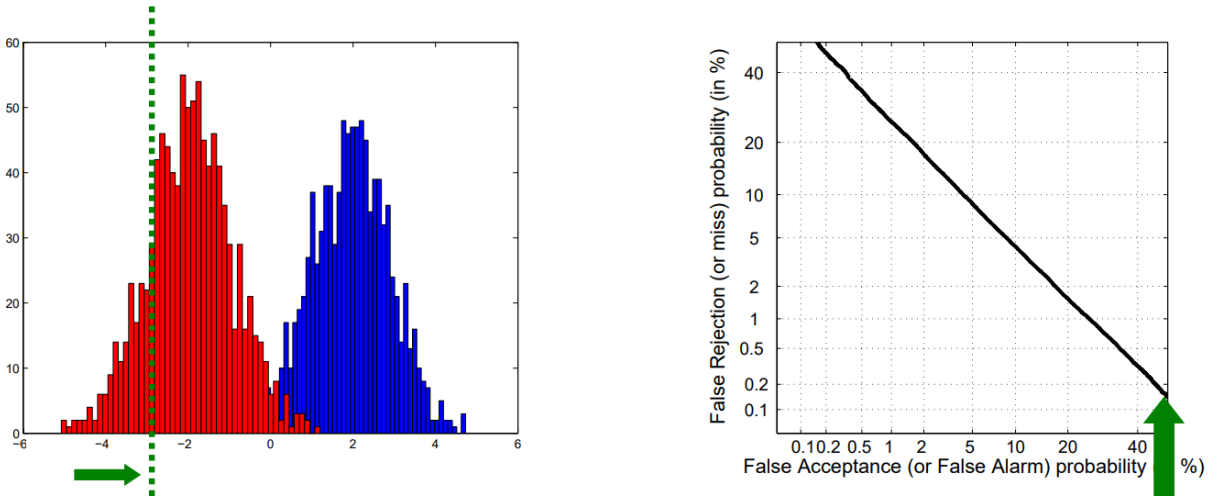


Figura 8 – Curva DET.

Esta curva solo nos permite evaluar la discriminación de los sistemas de reconocimiento que proporcionen *scores*. Por tanto, no permite que evaluemos la calibración de los *scores*.

Las curvas Tippet representan las distribuciones de los valores de *LR* obtenidos en una prueba de clasificación que intenta medir el rendimiento de un sistema. Nos informan sobre la discriminación, pero también nos permiten visualizar la falta de calibración en los valores de *LR* que el sistema calcula. Descalibración que procede de la descalibración, a su vez, de los scores.

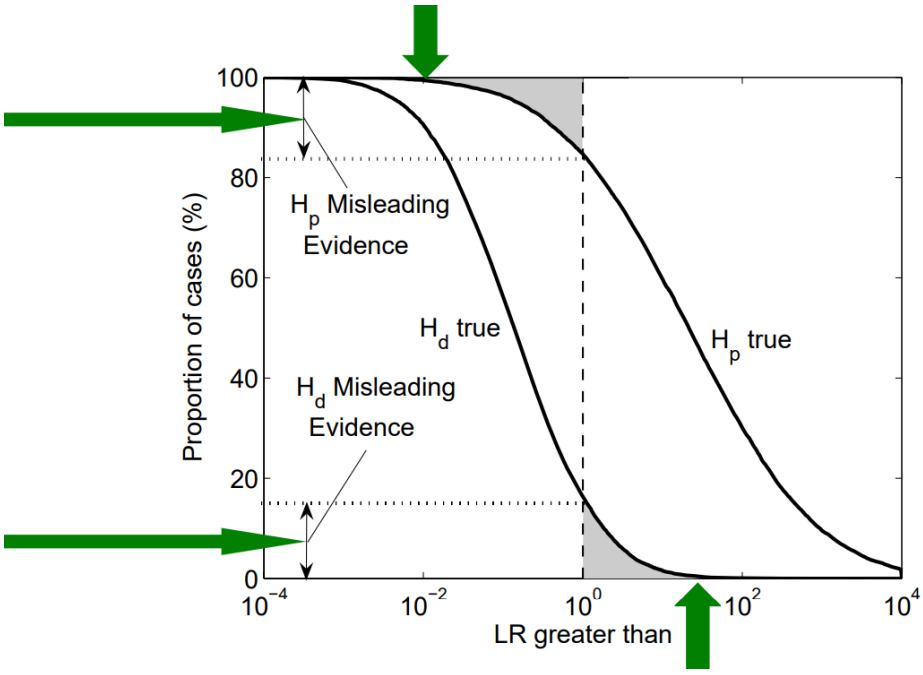


Figura 9 – Curva Tippet.

Si el LR mide evidencia, es decir, mide de qué modo los datos observados (por ejemplo, un “match”) apoyan la hipótesis de la común procedencia de las muestras comparadas frente a la hipótesis alternativa o, lo que es lo mismo, mide cómo de probable es la observación de un “match” si la hipótesis de la común procedencia de las muestras comparadas es verdadera frente a que lo sea la hipótesis alternativa, un LR

que apoya más la hipótesis no verdadera que la verdadera es el que recibe el calificativo de erróneo o engañoso.

Las curvas Tippet permiten observar la discriminación del sistema de reconocimiento en la separación entre las curvas de LR calculados conociendo que las muestras comparadas proceden del mismo origen (curva H_p cierta) y conociendo que las muestras comparadas proceden de orígenes distintos (curva H_d cierta).

En las zonas grises de la gráfica de la Tippet se aprecian los valores de LR erróneos o engañosos. Se calculan a partir de experimentos controlados, en los que *a priori* es necesario conocer los resultados correctos de cada comparación en lo referente a la discriminación.

La apreciación de valores exagerados de LR erróneo o engañoso se produce en ocasiones a simple vista, por lo que estas curvas, en principio, permiten visualizar que los LR que calcula un sistema están descalibrados. Cuanta mayor sea la descalibración, lógicamente mejor se apreciará en la curva esta deficiencia. Pero eso muchas veces no es tan fácil de apreciar. Por tanto, estas curvas son útiles y una ayuda, pero es mucho más seguro y preciso calcular los costes de los que hemos hablado (C_{llr}, $minC_{llr}$, *y* $calC_{llr}$).

Hemos visto que el coste C_{llr} mide discriminación y calibración. Es, sin duda, la medida óptima para evaluar el rendimiento de un sistema de reconocimiento. Es un valor escalar, por lo que permite ordenar los sistemas de acuerdo con su valor. Ya hemos dicho que podemos distinguir entre C_{llr} total (coste por discriminación + coste por calibración); $minC_{llr}$ (coste por discriminación); y $calC_{llr}$ (coste por calibración).

Además, existe una interpretación muy interesante de los costes de acuerdo con la teoría de la información que veremos más adelante.

Pero, antes de llegar a ella, conviene explicar las ventajas de representar el rendimiento de un sistema mediante las denominadas curvas *APE* (*Applied Probability of Error*).

Estas curvas se aplican a los valores de LR que nos entrega un sistema cuando lo evaluamos.

Podemos representar la probabilidad de error media, que se representa matemáticamente así:

$$E\{P_e\} = P(H_p) \cdot P(error \mid H_p) + P(H_d) \cdot P(error \mid H_d)$$

Se comete un error cuando las decisiones que se toman con el LR de acuerdo con el umbral de decisión —fijado por las probabilidades *a priori* y los costes— son erróneas. Se trata de una medida intuitiva. Si nos equivocamos más veces en media, el sistema de reconocimiento funciona peor.

Esa medida depende del umbral de decisión del LR, de las probabilidades *a priori* y de los costes.

A continuación, mostramos una curva *APE*:

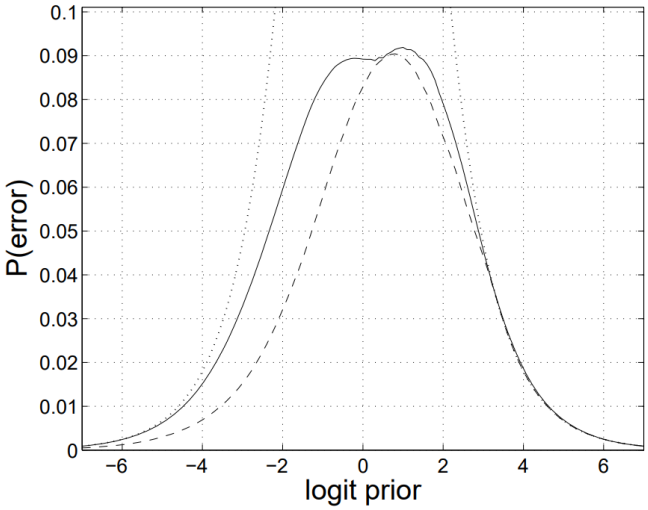


Figura 10 – Curva APE.

En esta curva representamos la probabilidad de error de los LR (también podemos llamarlos LR engañosos, como sabemos), en el eje de ordenadas, y las probabilidades *a priori* en el eje de abcisas, aunque representadas logarítmicamente.

Podemos apreciar tres curvas según el trazo: la curva de trazo continuo es la probabilidad de error de los LR tal y como los ha calculado el sistema de reconocimiento; la curva de trazo discontinuo bien visible es la probabilidad de error de los LR calculados por el sistema óptimamente calibrados con el algoritmo *PAV*; y la curva de trazo discontinuo apenas perceptible es la probabilidad de error de un sistema con valores de *LR=1*, es decir, con una valoración de la evidencia neutral respecto al apoyo a cualquiera de las hipótesis que se barajen en el caso. Como puede verse, la curva no fija los valores *a priori*, por lo que los tiene a todos en cuenta.

En la gráfica se explica un caso práctico en que se aplica una curva *APE* representativa del funcionamiento de un sistema de reconocimiento.

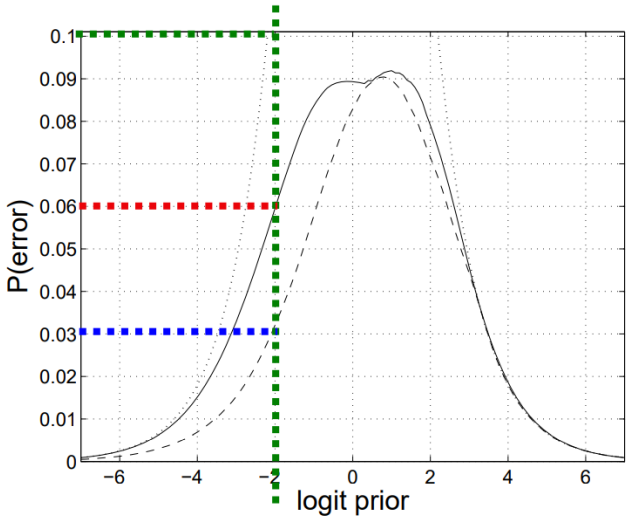


Figura 11 – Ejemplo de curva APE.

Si el perito utiliza esa curva para interpretar los resultados de LR de la prueba, ha de pedir al juez que fije una probabilidad *a posteriori* de error media máxima que desee aceptar en los LR si los usa. En el ejemplo se ha fijado el 6% de probabilidad *a posteriori* de error sobre las hipótesis que se barajan en el caso.

En el supuesto de que el sistema estuviera bien calibrado, la probabilidad a posteriori de error se reduciría al 3%. Y si el sistema que calcula los LR no aportara información (*LR = 1*), la probabilidad a posteriori de error sería del 10%.

Se puede demostrar: (1) que el área bajo la curva sólida de la *APE* es igual al valor del coste C_{llr}; (2) que el área sobre la curva rayada de la *APE* es $minC_{llr}$; y que la diferencia entre ambas áreas $calC_{llr}$. Esta demostración confirma que minimizar C_{llr} nos lleva a tomar decisiones mejores. La curva *APE* y el valor de C_{llr} suelen presentarse juntos. El coste C_{llr} es el valor escalar que resume la curva *APE*.

Habíamos comentado que hay otra curva de gran interés para evaluar el rendimiento de un sistema que se fundamenta en la teoría de la información.

Esa curva se denomina en inglés *Empirical Cross Entropy (ECE)*. Basada en la regla de puntuación logarítmica, evalúa el rendimiento de un sistema mediante un valor escalar. Cuanto mayor es el valor, mejor es el rendimiento del sistema.

Permite una fácil comparación de los métodos, tiene en cuenta la evidencia errónea, el poder de discriminación de los sistemas queda claramente establecido y permite realizar una interpretación del rendimiento de los sistemas de acuerdo a la teoría de la información de Shannon (Ramos et al., 2013b).

La curva *ECE* se diferencia esencialmente de la *APE* en los valores representados en el eje de ordenadas. En abcisas observamos las probabilidades *a priori*, y tenemos las tres curvas relacionadas con los valores de LR entregados por el sistema (línea continua en rojo), los valores optimizados con el algoritmo *PAV* (línea discontinua en azul) y valores de *LR = 1* que representa el caso en el que el sistema es neutral, es decir, no aporta nada al rendimiento en el reconocimiento. La descalibración producida en los LR es claramente observable y medible.

En las siguientes figuras puede verse cómo puede representarse el rendimiento de un sistema de reconocimiento de acuerdo con los métodos de evaluación propuestos.

Las figuras proceden de la tesis doctoral de D. Ramos (2007) y contemplan todos los tipos de gráficas vistos hasta ahora (curvas *DET*, Tippet, *APE* y *ECE*). El sistema que se evalúa es un sistema de reconocimiento automático de cotejo de voces.

Puede verse cómo mejora el rendimiento del sistema de reconocimiento comparando las gráficas relacionadas con dos procedimientos de medición de la similitud entre las voces comparadas (a partir de los *scores* de un sistema de reconocimiento basado en modelos de mezclas de gaussianas que se denomina *'Scores GMM*' y a partir de los valores de LR tras un proceso de adaptación sobre el procedimiento inicial que se denomina *'LR* adapted*'* que significa 'cálculo de LR adaptado al sospechoso'). Nos importa fijarnos en las gráficas y podemos decir, resumidamente, lo siguiente:

(1) la curva *DET* ligada al procedimiento '*LR* adapted' baja su tasa de igual error (*Equal Error Rate*) -observable tenuemente en la línea roja- (figura 12 (a));

(2) la curva Tippet azul a trazos tiene mejor discriminación (mayor separación entre las curvas H_p y H_d) y menor tasa de evidencia errónea (figura 12 (b));

(3) la curva *APE* y el gráfico de caja del coste C_{llr} se aprecia cómo mejora la calibración de los LR adaptados al sospechoso con respecto a los procedentes de los scores (figura 13 (c)); y

(4) en la curva *ECE* se aprecia, igualmente, la mejora en calibración de los LRs acercándose los resultados a la curva azul tras la aplicación del algoritmo *PAV* (figura 14 (d)).

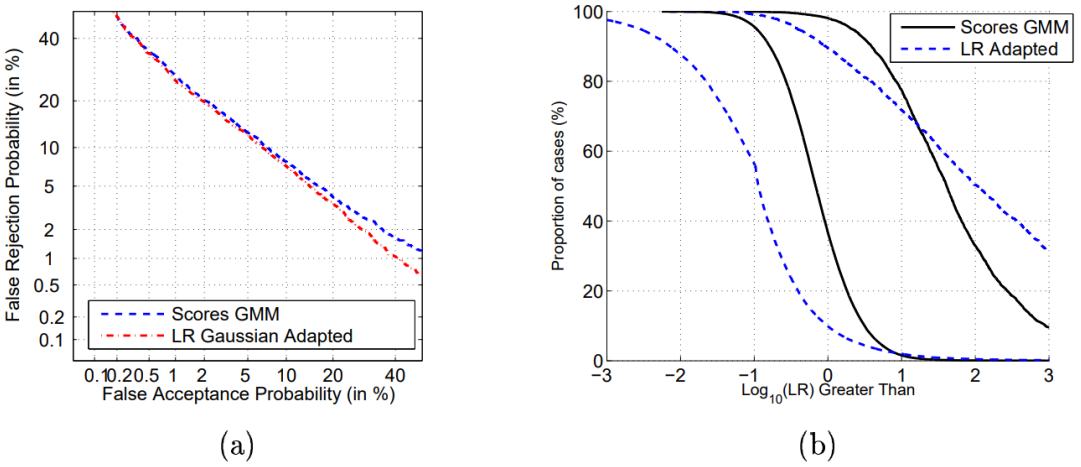


Figura 12 – EER, curva DET y curva Tippet.

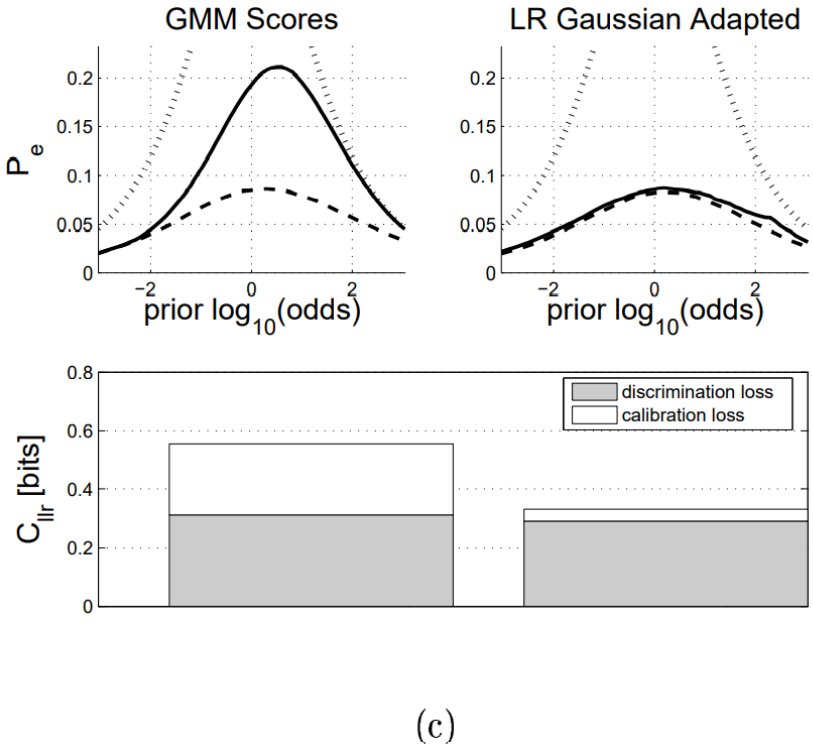
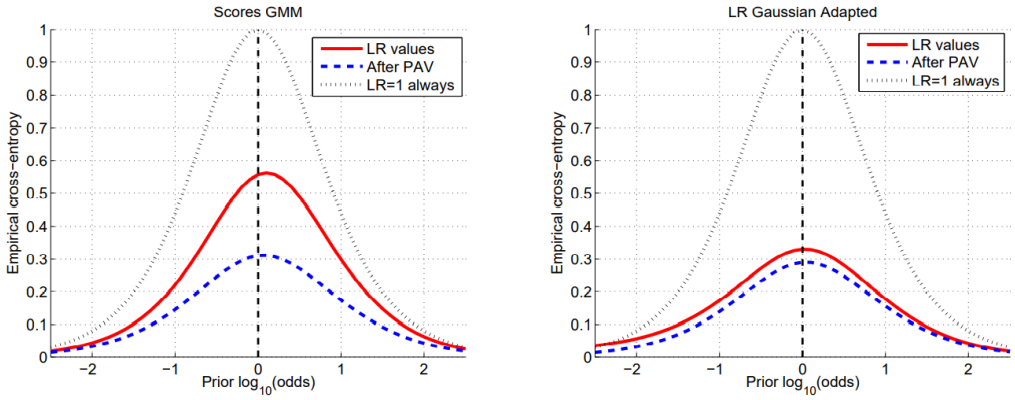


Figura 13 - Cllr.



(d)

Figura 14 – Curva ECE.

Para ilustrar cómo hacer uso de una curva *ECE*, imaginemos que la probabilidad a priori se fija por el juez en el valor 0.1 (coincidente con el valor -1 en el eje de abcisas). La curva gris punteada en el valor -1 en el eje de abcisas tiene el valor 0.45 en el eje de ordenadas. Los valores de la entropía cruzada empírica se expresan en bits. Su rango de valores es de 0 a 1 bit. El valor 0.45 bits significa que es el valor medio de la información que se necesita para conocer las hipótesis verdaderas promediando entre casos similares. Cuanto más alta sea la curva *ECE* más información nos hará falta.

Una vez analizado el peso de la evidencia, se obtiene más información, y, consecuentemente, se necesita un valor medio de información más bajo para conocer las hipótesis verdaderas promediando entre casos similares. En concreto, se necesitan 0.18 bits (cruce de una línea vertical imaginaria en el valor -1 en el eje de abcisas con la curva roja).

Y si se hubiera utilizado el sistema calibrado, se necesitaría un valor medio de información de 0.14 bits para conocer las hipótesis verdaderas promediando entre casos similares. Esta valoración no puede utilizarse en la práctica porque se necesitaría conocer qué hipótesis son las verdaderas en cada caso concreto. Cada caso real es incierto y solo es posible valorar su ocurrencia probabilísticamente.

6. CONCLUSIONES

La evaluación de sistemas de reconocimiento forenses está regida por la inferencia lógica y la calibración. No basta conocer la capacidad de discriminación que tenga un sistema de reconocimiento, sino que es imprescindible conocer si sus *scores* están bien calibrados.

Una regla de puntuación *(scoring rule*) asigna un coste a cada valor de probabilidad *a posteriori* que obtengamos en un proceso de evaluación del rendimiento de un sistema. Nos será útil si asigna costes altos a valores de probabilidad *a posteriori* muy erróneos.

Los sistemas de reconocimiento forenses se evalúan en su capacidad de discriminación generando clásicamente curvas ROC, DET o Tippet. Para medir la calibración de los scores necesitamos nuevas curvas como las APE o ECE, aunque la

curvas Tippet también ayudan a descubrir descalibración. El coste C_{llr} permite clasificar sistemas de reconocimiento teniendo en cuenta su calibración y discriminación.

Los avances logrados por la teoría de la decisión bayesiana aplicados a la evaluación de sistemas de reconocimiento forenses permiten que las clásicas técnicas utilizadas en los laboratorios para medir el rendimiento de los sistemas clasificatorios, fundamentadas en mediciones de tasas de error en el mejor de los casos —con frecuencia ni siquiera se calculaban experimentalmente—, sean sustituidas por modernas técnicas de inferencia estadística y de calibración.

Por tanto, los laboratorios forenses debieran priorizar sus esfuerzos para conseguir mejorar científicamente la forma en que evalúan sus sistemas de reconocimiento.

En este sentido, la Guía de ENFSI de informes evaluativos emitida en el año 2015 (https://enfsi.eu/documents/forensic-guidelines) ha sido una iniciativa pionera entre las redes de laboratorios forenses de todo el mundo, y en la que la Guardia Civil institucionalmente participó.

El mencionado documento proporciona una guía práctica para implementar en los laboratorios los informes evaluativos.

Apéndice 1

Zaragoza en 2006 tiene 500.000 habitantes. Uno de sus días, un varón sin identificar roba un banco. Lo único que se sabe de él es que es moreno y que no es foráneo. La proporción de morenos en Zaragoza es del 50%.

La policía arresta a un sospechoso al azar entre la población y resulta que es moreno. ¿Cuál es la probabilidad de que el sospechoso sea el autor del robo?

Hipótesis que se manejan:

- por parte de la acusación (H_p): “El sospechoso es el autor del robo;
- por parte de la defensa (H_d): “Cualquier otro zaragozano es el autor del robo”.

Evidencia (E): “El sospechoso es moreno”.

Pregunta a resolver por el tribunal: ¿cuál es la probabilidad de que, a la luz de la evidencia y del resto de información acerca del caso, el sospechoso sea el autor del robo?

En términos matemáticos: $P(H_p / E,I)$.

Disponemos de dos tipos de información acerca del caso:

- información sobre la evidencia (E);
- información de contexto (I) o relevante en el caso pero que no tiene que ver con la evidencia (testimonios de testigos, número de potenciales causantes del robo, coche que empleó en la huida, etc.)

Del perito se espera que valore el peso de la evidencia (E) y del juez que valore el peso de la evidencia teniendo en cuenta las tesis de las partes en el proceso y la información disponible en el caso.

La respuesta que el juez necesita la encuentra aplicando al caso el teorema de Bayes en forma de apuestas:

$$[apuesta\ a\ posteriori] = [LR] \times [apuesta\ a\ priori]$$

$$\frac{P(H_p \mid E, I)}{P(H_d \mid E, I)} = \frac{P(E \mid H_p, I)}{P(E \mid H_d, I)} \times \frac{P(H_p \mid I)}{P(H_d \mid I)}$$

El juez puede calcular la apuesta *a priori* apoyándose en datos o juicios subjetivos sobre las hipótesis.

La probabilidad se interpreta, en cualquiera de esos dos casos, como una opinión, cuya veracidad se gradúa. Se trata de una noción de probabilidad subjetiva.

La noción subjetiva de probabilidad permite asignar probabilidades a sucesos únicos, es decir, no repetibles, de los cuales podemos tener una cierta convicción sobre su ocurrencia o veracidad por el conocimiento, información y experiencia previas de que dispongamos al respecto.

Una vez que hemos presentado la herramienta lógico-matemática para abordar el problema, lo resolvemos así:

Del enunciado del problema, sabemos esto:

- la probabilidad de que el sospechoso sea el autor del robo, si solo sabemos que es zaragozano, es la que le corresponde a él entre todos los zaragozanos, es decir: $P(H_p / I)$ = 1/500.000;
- la probabilidad de que el sospechoso no sea el autor del robo, si solo sabemos que es zaragozano, es la complementaria de la anterior: : $P(H_d / I)$ = 499.999/500.000;
- si el sospechoso es el autor del crimen, entonces es moreno, es decir: $P(E / H_p, I)$= 1;
- si el autor del crimen es otro individuo, la probabilidad de que sea moreno es la proporción de morenos en la población: : $P(E / H_d, I)$= 0,5.

Y entonces, podemos calcular el valor del LR (peso de la evidencia) así:

$$LR = \frac{P(E \mid H_p, I)}{P(E \mid H_d, I)} = \frac{1}{0.5} = 2.$$

La evidencia apoya a la hipótesis de la acusación dos veces más que a la hipótesis de la defensa.

Sin que conociéramos la evidencia (que el autor del crimen era moreno), la apuesta *a priori* era la siguiente:

$$\frac{P(H_p \mid I)}{P(H_d \mid I)} = \frac{1}{500.000} \Big/ \frac{499.999}{500.000} = 1/499.999.$$

Sin conocer la evidencia, apostamos 499.999 a 1 que el sospechoso no es el autor del robo.

El hecho de que conozcamos que el autor del crimen sea moreno, supone apoyar 2 a 1 que el sospechoso sea el autor del robo.

Y el juez, teniendo en cuenta la apuesta *a priori* y la relación de verosimilitudes (el LR), aplicando el teorema de Bayes en forma de apuestas, llega a esta apuesta *a posteriori*:

$$\frac{P(H_p \mid E, I)}{P(H_d \mid E, I)} = 2 \times \left(\frac{1}{499.999}\right) = 2/499.999.$$

Es decir, habiendo observado la evidencia, apostamos 499.999 a 2 que el sospechoso no es el autor del robo.

Y sobre esa apuesta *a posteriori*, el juez decide.

Pues bien, ahora damos un paso más y nos encontramos ante un caso de un ladrón en un pueblo de Zaragoza junto al Moncayo, de 100 habitantes.

Sin conocer la evidencia, la apuesta *a priori* era la siguiente:

$$\frac{P(H_p \mid I)}{P(H_d \mid I)} = \frac{1}{100} \Big/ \frac{99}{100} = 1/99.$$

El hecho de conocer que el autor del crimen era moreno, supone una relación de verosimilitudes igual a 2.

Y el juez, teniendo en cuenta la apuesta *a priori* y la relación de verosimilitudes (el LR), aplicando el teorema de Bayes en forma de apuestas, llega a esta apuesta *a posteriori*:

$$\frac{P(H_p \mid E, I)}{P(H_d \mid E, I)} = 2 \times \left(\frac{1}{99}\right) = 2/99.$$

Es decir, habiendo observado la evidencia, apostamos 99 a 2 que el sospechoso no es el autor del robo.

Y sobre esa apuesta *a posteriori*, el juez decide.

Y, por fin, llegamos a la situación de saber que el autor del crimen era un lapón rural. Siendo de Laponia no es esperable que haya muchos morenos y, tras consultar estadísticas oficiales, el porcentaje de morenos en la población lapona rural es del orden del 5%.

Calculamos el peso de la evidencia en ese contexto:

$$LR = \frac{P(E \mid H_p, I)}{P(E \mid H_d, I)} = \frac{1}{0.05} = 20.$$

El hecho de conocer que el autor del crimen era un varón moreno de Laponia, supone una relación de verosimilitudes igual a 20.

Y el juez, teniendo en cuenta la apuesta *a priori* y la relación de verosimilitudes (el LR), aplicando el teorema de Bayes en forma de apuestas, llega a esta apuesta *a posteriori*:

$$\frac{P\left(H_p \mid E, I\right)}{P(H_d \mid E, I)} = 20 \times \left(\frac{1}{99}\right) = 20/99.$$

Es decir, habiendo observado la evidencia, apostamos 99 a 20 que el sospechoso no es el autor del robo.

Y sobre esa apuesta *a posteriori*, el juez decide.

Si en lugar de fijarnos en las apuestas *a posteriori*, con las que no estamos muy familiarizados, y hallamos las probabilidades *a posteriori* de esos tres casos a favor de que el sospechoso sea el autor del robo, obtenemos estos datos:

En el caso del ladrón zaragozano:

$$\frac{P\left(H_p \mid E, I\right)}{P(H_d \mid E, I)} = \frac{2}{499.999} \Rightarrow P\left(H_p \mid E, I\right) = \frac{\frac{2}{499.999}}{1 + \frac{2}{499.999}} = \frac{2}{500.001}$$

$$= 0.000004\%$$

Y en el del ladrón zaragozano rural:

$$\frac{P\left(H_p \mid E, I\right)}{P(H_d \mid E, I)} = \frac{2}{99} \Rightarrow P\left(H_p \mid E, I\right) = \frac{\frac{2}{99}}{1 + \frac{2}{99}} = \frac{2}{101} = 1{,}98\%$$

Y en el del lapón rural:

$$\frac{P(H_p \mid E,I)}{P(H_d \mid E,I)} = \frac{20}{99} \Rightarrow P(H_p \mid E,I) = \frac{\frac{20}{99}}{1+\frac{20}{99}} = \frac{2}{119} = 16{,}81\%$$

BIBLIOGRAFÍA

Aitken C.G.G., Taroni F., & Bozza S. (2021). Statistics and the Evaluation of Evidence for Forensic Scientists. 3rd ed., Wiley, Chichester (UK).

Brümmer N., du Preez J. (2006). Application-independent evaluation of speaker detection. *Computer Speech and Language*, *20*, 230-275.

Cook R., Evett I.W., Jackson G., Jones P.J., & Lambert J.A. (1998). A hierarchy of propositions: deciding which level to address in casework. *Science and Justice 38*(4), 231-240.

ENFSI GUIDELINE FOR EVALUATIVE REPORTING IN FORENSIC SCIENCE (2015). https://enfsi.eu/documents/forensic-guidelines

Evett I.W. (2009). Evaluation and professionalism. *Science & Justice*, *49*(3), 159-160.

Evett, I.W. et al. (2017). Finding the way forward for forensic science in the US—A comentary on the PCAST report. *Forensic Science International*, *278*, 16-23. http://dx.doi.org/10.1016/j.forsciint.2017.06.018.

González Rodríguez J., Drygajlo A., Ramos Castro D., García Gomar M., & Ortega García J. (2006). Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. *Computer Speech and Language*, *20*(2-3), 331-355.

González Rodríguez J., Rose P., Ramos Castro D., Toledano D.T., & Ortega García J. (2007). Emulating DNA: rigorous quantification of evidential weight in transparent and testable forensic speaker recognition. *IEEE Transactions on Audio, Speech and Language Processing*, *15*(7), 2104-2115.

González Rodríguez J. (2014). Evaluating Speaker Recognition systems: An overview of the NIST Speaker Recognition Evaluations (1996-2014). *Loquens*, *1*(1), e007. http://loquens.revistas.csic.es/index.php/loquens/article/download/9/21

Morrison G.S. (2012). Response to DR AS 5388.3 Forensic Analysis – Part 3 – Interpretation. https://geoff-morrison.net/documents/Morrison,%20et%20al%20(2012)%20Response%20to%20Australian%20Draft%20Standards%20DR%20AS%205388.3%20Forensic%20analysis%20-%20Part%203-%20Interpretation.pdf

National Research Council (2009). Strengthening Forensic Science in the United States: A Path Forward. The National Academies Press, Washington DC. https://www.ncjrs.gov/pdffiles1/nij/grants/228091.pdf

President's Council of Advisors on Science and Technology (2016). Report to the president Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods, Washington DC. https://obamawhitehouse.archives.gov/administration/eop/ostp/pcast/docsreports

Ramos Castro D. (2007). Forensic evaluation of the evidence using automatic speaker recognition systems. Tesis doctoral. Departamento de Ingeniería Informática, Escuela Politécnica Superior, Universidad Autónoma de Madrid.

Ramos Castro D., González Rodríguez J. (2013a). Reliable Support: Measuring Calibration of Likelihood Ratios. *Forensic Science International*, *230*, 156-159.

Ramos Castro D., González Rodríguez J., Zadora G., & Aitken C. (2013b). Information-theoretical assessment of the performance of likelihood ratio computation methods. *Forensic Science International*, *58*(6), 1503-1518.

Robertson B., Vignaux G.A. (2016). Interpreting Evidence – Evaluating Forensic Science in the Courtroom. 2nd edition, John Wiley & Sons, Chichester (UK).

Royall R. (1997). Statistical Evidence: A Likelihood Paradigm. Chapman&Hall, London (UK).

Taroni F., Aitken C.G.G., & Garbolino P. (2001). De Finetti's subjectivism, the assessment of probabilities and the evaluation of evidence: a commentary for forensic scientists. *Science and Justice*, *41*(3), 145-150.

Taroni F., Bozza S., Biedermann A., Garbolino P., & Aitken C. (2010). Data Analysis in Forensic Science. J. Wiley & Sons, Statistics in Practice, Chichester (UK).

Taroni F., Biedermann A., Bozza S., Garbolino P., & Aitken C. (2014). Bayesian Network for Probabilistic Inference and Decision Analysis in Forensic Science. 2nd edition, John Wiley & Sons, Statistics in Practice, Chichester (UK).

Taroni F., Garbolino P., Biedermann A., Aitken C., & Bozza S. (2018). Reconciliation of subjective probabilities and frequencies in forensic science. *Law, Probability and Risk*, *17*(3), 243-262.