UNIVERSIDAD
POLITECNICA
DE VALENCIA

UNIVERSITAT POLITÉCNICA VALÉNCIA

Departamento de comunicaciones

Fusión de datos estadísticamente dependientes en sistemas de detección

TESIS DOCTORAL

Antonio Soriano Tolosa

Director:

Dr. Luis Vergara Domínguez

Valencia, Noviembre 2013

"Basta un poco de espíritu aventurero para estar siempre satisfechos, pues en esta vida, gracias a dios, nada sucede como deseábamos, como suponíamos, ni como teníamos previsto."

Agradecimientos

Mientras el río corra, los montes hagan sombra y en el cielo haya estrellas, debe durar la memoria del beneficio recibido en la mente del hombre agradecido.

– Virgilio –

Antes que nada quiero expresar mi gratitud a Luis Vergara Domínguez por brindarme la oportunidad de realizar mi trabajo de investigación en el seno del grupo de tratamiento de señal (*GTS*). Darle las gracias por su trato hacia mi persona, siempre amable, siempre dispuesto a utilizar su inmenso conocimiento para asesorarme, guiarme y darme consejos durante todo el proceso de investigación que ha culminado en el presente documento.

También me gustaría agradecer a mis compañeros Jorge, Arturo, Gonzalo, Guille, Vicente y Alicia por su buena acogida en el laboratorio, por estar siempre ahí, dispuestos a ayudar en cualquier cosa, por hacer más llevadero el trabajo con historias, bromas, celebraciones y por tantos buenos momentos que me han hecho vivir.

A lo largo de estos años muchos han sido los amigos que me han aportado su cariño, su compañía y su apoyo, tanto en los buenos momentos, como en lo no tan buenos. Gracias por su inestimable ayuda, tanto por permitirme evadirme del trabajo cuando me ha sido necesario, como por proporcionarte las fuerzas y el apoyo necesario cuando me ha tocado centrarme en él. Gracias Rubén, Imelda y Alex, que además de amigos, habéis sido unos excelentes compañeros de piso; gracias Miriam, Sabi y Pedro que, junto a Rubén e Imelda, sabemos que "el cedro mola más"; gracias a "los gitanos gaditanos" Nacho, Jezú y Diego, gracias Laura, Maite, Cristina, Lidi, Vero y Gualda, que aunque de "mordor norte", nos hemos conseguido juntar en tantos buenos momentos. Gracias a todos mis amigos del pueblo, Javi, Fran, Chusin, Mily, Pocholilla, Anita, gracias por ser mis amigos de toda la vida, amigos que, aunque haya pasado mucho tiempo fuera, nunca os habéis olvidado de mí (sobre todo agradecérselo a Anita, la encargada de recordármelo en innumerables ocasiones). De entre todos ellos quiero destacar a mi gran amigo Cuevas, sin duda, una gran persona, sin duda, siempre un gran apoyo.

Por último, y no por ello menos importante, quiero agradecer el inmerso apoyo, la confianza y cariño que siempre he recibido por parte de todos y cada uno de los miembros de mi familia. Gracias a mi hermanita Ana y a mi cuñado Pascual, a mis primos Juanan, Isa, Joseda, Anly y Adrí, que en verdad os tengo casi por hermanos. Gracias a todos mis tíos, aunque tengo que destacar a mi tío "preforito" Juan Antonio, que por ser el único miembro de la familia que es capaz de entender algo del trabajo que realizo ha tenido que aguantar algún que otro "tostón" por mi parte. Ante todo, dar las gracias tanto a mis padres, Antonio y Ana, como a todos mis abuelitos, gracias por vuestro amor, gracias por vuestro sacrificio, gracias por la educación que me habéis dado, gracias por vuestros consejos… en fin, gracias por tantas cosas que habéis aportado a mi vida y que me han hecho llegar a ser la persona que soy hoy en día.

Resumen

La presente tesis se centra en la problemática existente a la hora de implementar un sistema de detección cuando es necesario combinar, integrar o fusionar diversas **fuentes de información dependientes y heterogéneas** entre sí. Las técnicas de fusión de datos tratan de combinar múltiples fuentes de información para alcanzar la exactitud y precisión en la toma de decisiones que no sería posible conseguir con el uso de una sola fuente de información de forma aislada.

En un sistema de detección se pueden diferenciar varios niveles de fusión: en la etapa de pre-detección encontramos los niveles de fusión de sensores y de características, donde se combinan varios flujos de muestras proporcionados por una serie de sensores o diferentes características obtenidas del procesado estos; en la etapa de post-detección, se realiza la combinación de diferentes detectores a través de la fusión de valoraciones continuas o de decisiones individuales aportadas por cada uno de ellos. Según el tipo de datos a combinar diferenciamos dos grupos: **fusión soft**, donde se combinan variables aleatorias (*v.a*) continuas, caracterizadas mediante funciones de densidad de probabilidad (*PDFs*), o **fusión hard**, asociada a la combinación de las decisiones individuales proporcionadas por diversos detectores, donde se combinan datos binarios modelados mediante *v.a* discretas, caracterizadas por funciones de masa de probabilidad (*PMFs*). Se destaca la **fusión de scores** como un caso particular de fusión soft asociada a la fusión de detectores, en donde los datos a combinar presentan buenas propiedades discriminatorias y se encuentran definidos en un mismo rango normalizado $[0,1]$.

En el presente trabajo se ha realizado una completa revisión del estado del arte en cuanto a técnicas de fusión y combinación de datos aplicadas en problemas de detección donde los datos pueden ser heterogéneos y dependientes entre sí. Se realiza una revisión en mayor profundidad de la técnica de estimación de *PDFs* basada en la **teoría de cópulas,** debido tanto a su novedad e incipiente uso en el campo del procesado de señal, como por su adecuación de uso en problemas de detección. Nos permite modelar de forma aislada las funciones marginales de los datos y la estructura de dependencia presente entre ellos, simplificando el problema de modelado de *PDFs* de datos heterogéneos y dependientes.

Se ha propuesto una nueva técnica de fusión soft denominada **integración-α**, basada en una función de media-α, la cual aporta un mayor grado de flexibilidad y de adaptación, siendo capaz de mejorar las prestaciones que se pueden obtener con respecto al resto de técnicas subóptimas utilizadas comúnmente en problemas de fusión de scores. Se ha derivado un novedoso **método de entrenamiento** basado en el **criterio de maximización parcial del área bajo la curva *ROC*.**

Se han utilizado diversas bases de datos públicas para poder testear y comprobar el correcto funcionamiento de las técnicas de fusión propuestas en problemas de **autentificación multibiométrica**. También se han aplicado algunas de las técnicas de fusión en la mejora de un sistema de detección de eventos acústicos. Se ha propuesto un **nuevo tipo de detector** basado en la teoría de cópulas denominado ***COCD*** para lidiar con el problema de la **detección de señal desconocida en presencia de ruido aleatorio** dependiente y no Gaussiano, centrándonos en su utilización para una aplicación de detección de eventos sonoros desconocidos. También se realiza un estudio de **fusión de más de un canal de audio** (utilizando más de un micrófono para captar diferentes señales) como método para incrementar las prestaciones obtenidas.

Resum

La present tesi se centra en la problemàtica existent a l'hora d'implementar un sistema de detecció o classificació binària quan és necessari combinar, integrar o fusionar diverses fonts d'informació que poden ser dependents i heterogènies entre si. Les tècniques de fusió de dades tracten de combinar múltiples fonts d'informació per a aconseguir l'exactitud i precisió en la presa de decisions que no seria possible aconseguir amb l'ús d'una sola font d'informació de forma aïllada. En el present treball s'ha realitzat una completa revisió de l'estat de l'art quant a tècniques de fusió i combinació de dades aplicades en problemes de detecció on les dades poden ser heterogenis i dependents entre si.

Atenent al tipus de dades a combinar trobem dos grups: **fusió soft**, on es combinen dades modelatges per mitjà de variables aleatòries contínues, caracteritzades per mitjà de les seues funcions de densitat de probabilitat (*PDFs*) , o **fusió hard**, associada a la combinació de les decisions individuals preses en l'etapa de fusió de detectors, on es combinen dades binàries modelatges per mitjà de variables aleatòries discretes, caracteritzades per funcions de massa de probabilitat. Es destaca la **fusió de scores** com un cas particular de fusió soft associada a la fusió de diversos detectors, on les dades a combinar presenten bones propietats discriminatòries de forma aïllada i es troben definits en un mateix rang normalitzat [0,1].

Es realitza una revisió en més profunditat de la tècnica d'estimació de *PDFs* basada en la **teoria de còpules**, la qual pot ser usada en la fusió òptima de dades. Es destaca de forma especial tant per la seua novetat i incipient ús en el camp del processat de senyal, com per la seua adequació en problemes de detecció, permetent-nos modelar de forma aïllada les funcions marginals de les dades i l'estructura de dependència present entre ells, simplificant el problema de modelatge de PDFs de dades heterogènies i dependents.

S'ha proposat una nova tècnica de fusió soft per al cas de la fusió de scores aportats per diferents detectors denominada **integració-α**, basada en una funció de mitja-α, la qual, sense elevar molt la complexitat, aporta un major grau de flexibilitat i d'adaptació, sent capaç de millorar les prestacions que es poden obtindre respecte a la resta de tècniques subòptimes utilitzades comunament en problemes de fusió de scores heterogenis i dependents entre si. S'ha derivat un nou **mètode d'entrenament basat en el criteri de maximització parcial de l'àrea davall la corba ROC**.

S'han utilitzat diverses bases de dades públiques per a poder testear i comprovar el funcionament correcte de les tècniques de fusió proposades en problemes d'autentificació multibiométrica. També s'han aplicat algunes de les tècniques de fusió en la millora d'un sistema de detecció d'esdeveniments acústics. S'ha proposat un **nou tipus de detector** basat en la teoria de còpules denominat ***COCD*** per a torejar amb el problema de la detecció de senyal desconeguda en presència de soroll aleatori dependent i no Gaussiano, centrant-nos en la seua utilització per a una aplicació de detecció d'esdeveniments sonors desconeguts. També es realitza un **estudi de fusió de més d'un canal d'àudio** (utilitzant més d'un micròfon per a captar diferents senyals) com a mètode per a incrementar les prestacions obtingudes.

Abstract

Fusing information from heterogeneous and dependent sources for binary hypotheses testing problems is a challenge. Data fusion techniques are used to combine several sources of information in order to achieve a robust and accurate result in the decision-making process, impossible to reach using each data source in a separate way. This thesis includes a complete review of the state-of-art in data fusion techniques used in detection problems.

Data fusion can be performed at different levels in a detection system: in the pre-detection stage, data from several sensors or different features obtained processing these raw data can be combined at sensor fusion level and feature fusion level; in the post-detection stage, a combination of individual decisions or continuous valuations (scores) from disparate detectors is carried out in the detector fusion level.

Data fusion can be spitted into two categories attending to data characteristics: in soft fusion, different continuous data streams are combined; hard fusion is related to the combination of the individual decision given by several detectors. Score fusion can be considered as a special case of soft fusion in which continuous data (commonly related to the event occurrence probability) given by several detectors are combined.

The use of **statistical theory of copulas** to model the joint distribution of data form heterogeneous and dependent sources in the **soft fusion** has been remarked due to its recently and useful applicability in binary hypotheses testing problems. Copulas are functions that couple multivariate joint distributions to their component marginal distribution functions. It allows one to construct a statistical model by considering, separately, the univariate behavior of the underlying marginal and the dependence structure specified by some copula function. This is well suited for modeling the joint behavior of the heterogeneous and dependent sources in a binary hypotheses testing problem.

As far as the **score fusion** is concerned, a new technique based on α-mean function is presented. We have called it **α-integration**. This technique overperforms other commonly used methods due to its higher grade of adaptation. A new **training method based on the partial area under ROC curve optimization** is proposed.

A multimodal biometric system integrates information from multiple biometric sources to compensate for the limitations in performance of each individual biometric system. We have done some experiments on several multibiometric public databases in order to test these fusion techniques.

This work also focuses on the improvement of a novelties detection system in a monitored acoustic environment by using data fusion methods. A new detector, named as **copula one-class detector *COCD*,** has been proposed to deal with the problem of unknown signal detection in presence of random non-Gaussian noise. Also, we have proposed the **fusion of several microphones** in order to improve the performance of the acoustic detection.

Índice

Capítulo 1: Introducción

"Todo lo que nace proviene necesariamente de una causa; pues sin causa nada puede tener origen."

- Platón -

La presente tesis se centra en la problemática existente a la hora de implementar un sistema de detección o clasificación binaria cuando es necesario combinar, integrar o fusionar diversas fuentes de información que pueden ser dependientes y heterogéneas entre sí.

En este capítulo se introduce inicialmente la teoría general de la detección. Se define el concepto de fusión o integración de información y se destaca su utilización en el desarrollo de un sistema de detección. Se muestra como la integración de la información puede realizarse a diferentes niveles y con diferentes tipos de datos a lo largo del proceso seguido hasta la toma de una decisión final. Se remarca la posible heterogeneidad y dependencia estadística entre las distintas fuentes de información como principal peculiaridad y problemática encontrada a la hora de fusionar información en un sistema de detección. Se destaca la importancia que posee la correcta caracterización de la dependencia estadística para la obtención de buenas prestaciones en detección mediante un pequeño ejemplo.

Una vez presentado el marco de aplicación y la problemática existente, se definen los objetivos que se han planteado conseguir y se introducen las diferentes aplicaciones prácticas incluidas en la segunda parte de la tesis doctoral, en las cuales se han experimentado algunas de las técnicas de fusión de datos presentadas en la primera parte de este trabajo.

1.1. – Teoría de la detección

La teoría de la detección estudia la manera óptima de tomar una decisión entre dos posibles, basándose en el estudio de los datos proporcionados por una o varias fuentes de información (también llamados mediciones u observaciones) o características derivadas del procesado de éstos, para intentar discernir cuándo un determinado evento ha sucedido o no [1].

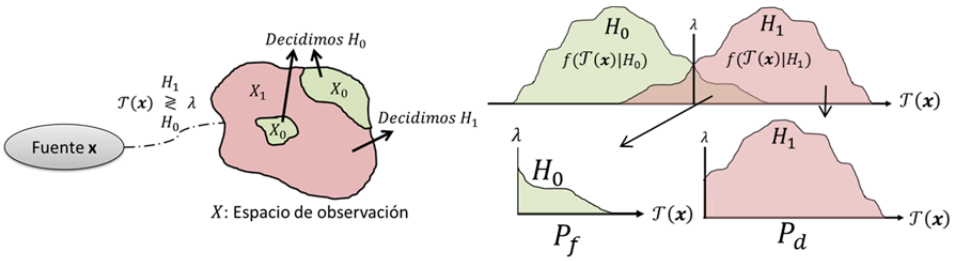


Figura 1.1 – Esquema de un problema de test de hipótesis

La detección pertenece a un tipo de problemas estadísticos tratados como "*test de Hipótesis*". En la implementación de un detector se formula un test de hipótesis binario, en el que la hipótesis H_0 representa la ausencia del evento que se desea

detectar, y la hipótesis H_1 representa la presencia de dicho evento. Asumimos que disponemos de un vector de observaciones $\boldsymbol{x} \triangleq [x_1, x_2, \dots, x_d]^T$ en un espacio *d-dimensional* llamado espacio de observación (X). El detector, mediante una regla de decisión basada en la umbralización de una función o estadístico $\mathcal{T}(\boldsymbol{x})$, decidirá si se ha producido evento o no. Esta función $\mathcal{T}(\boldsymbol{x})$: $\mathbb{R}^d \rightarrow \mathbb{R}$ transforma el espacio de observaciones, mapeándolo en una única dimensión, de forma que las funciones de densidad de probabilidad (*PDFs*) bajo cada hipótesis en el nuevo espacio queden lo más separadas posibles, minimizando el solapamiento entre ellas. Así, la decisión se tomará en función de un valor umbral λ, lo que equivale a dividir el espacio de observación en dos regiones X_0 y X_1. Se decidirá la hipótesis H_0 si las observaciones caen en la región X_0 definida por $\mathcal{T}(\boldsymbol{x}) < \lambda$, y H_1 si lo hacen en la región X_0 definida por $\mathcal{T}(\boldsymbol{x}) \geq \lambda$ [1]. En la figura 1.1 se puede observar un esquema de un problema de test de hipótesis.

Caracterización de las prestaciones de un detector

El funcionamiento de un detector se puede caracterizar mediante dos probabilidades (1.1), la probabilidad de falsa alarma (P_f) y la probabilidad de detección (P_d). Interesa que un detector tenga una alta probabilidad de detección, manteniendo una probabilidad de falsa alarma lo menor posible.

$$P_f = P(H_1|H_0 \; es \; verdadera) = \int_{X_1} f(\boldsymbol{x}|H_0)d\boldsymbol{x} \; = \int_{\lambda}^{\infty} f(\mathcal{T}(\boldsymbol{x})|H_0)d\mathcal{T}(\boldsymbol{x})$$
$$P_d = P(H_1|H_1 \; es \; verdadera) = \int_{X_1} f(\boldsymbol{x}|H_1)d\boldsymbol{x} \; = \int_{\lambda}^{\infty} f(\mathcal{T}(\boldsymbol{x})|H_1)d\mathcal{T}(\boldsymbol{x}) \tag{1.1}$$

Se puede estudiar las prestaciones de un detector por medio de su curva *ROC* ("*Receiver Operating Characteristic*"), la cual representa la P_d en función de la P_f obtenidas en todo el rango de umbrales λ posibles. Las curvas *ROC* pueden considerarse como indicativos de la calidad de un detector, posibilitándonos una comparación visual entre varios detectores. En términos generales, un detector se considera con mejores prestaciones que otros si su curva *ROC* se encuentra por encima del resto en un determinado rango de trabajo; significa que el detector posee la mayor P_d para una determinada P_f o la menor P_f para una determinada P_d. La curva limitante en que $P_d = P_f$ en todo el rango de umbrales λ corresponde al caso de un detector totalmente aleatorio, en el que la decisión es tomada sin tener en cuenta ninguna información referente al evento a detectar.

Uno de los índices extraído de la curva ROC más utilizado en muchos contextos es el área bajo la curva *ROC* [2] (*AUC*, "*Area Under ROC Curve*"):

$$AUC = \int_0^1 P_D(P_F)\, dP_F \tag{1.2}$$

donde $0 \leq AUC \leq 1$, ya que representa una porción de un cuadrado unidad.

En una gran cantidad de aplicaciones se debe diseñar el detector para trabajar en un cierto punto o rango de falsa alarma restringido. Por ejemplo, en ciertas aplicaciones clínicas como la descrita en [3]; ésta se basa en la detección automática de cáncer de pecho en mamografías, donde los radiologistas han limitado el rango de falsas alarmas aceptables entre 0.2 y 0.3, obtenido como compromiso entre la maximización en la detección de casos y una minimización en la carga de trabajo. En aplicaciones de autentificación biométrica [4], los falsos positivos son prácticamente intolerables, limitando el uso de detectores bajo probabilidades de falsa alarma por debajo de 10^{-4}. En una aplicación de detección de fraudes con operaciones de tarjeta, todas las operaciones categorizadas como fraudes deben ser comprobadas por personal de la entidad bancaria, existiendo un límite en el número de operaciones que por día pueden procesar; esto se traduce en que se debe trabajar en un rango de falsa alarma controlado.

En casos como los mencionados, donde se hace conveniente optimizar el rendimiento del detector en una determinada zona de falsa alarma limitada, el índice que mejor caracteriza las prestaciones es el área bajo una determinada porción $\alpha \leq P_F \leq \beta$ de la curva *ROC* ($pAUC_\alpha^\beta$, *"partial Area Under ROC Curve"*):

$$pAUC_\alpha^\beta = \int_\alpha^\beta P_D(P_F)\, dP_F \tag{1.3}$$

Es fácil ver que $0 \leq pAUC_\alpha^\beta \leq \beta - \alpha$, ya que limitamos el cálculo del área en la zona $\alpha \leq P_F \leq \beta$. Con objeto de definir el índice en un rango de valores independientes del tamaño de la zona de falsa alarma escogida se suele presentar esta medida normalizada ($nAUC_\alpha^\beta$):

$$nAUC_\alpha^\beta = \frac{1}{\beta - \alpha}\int_\alpha^\beta P_D(P_F)\, dP_F \in [0,1] \tag{1.4}$$

Diseño e implementación de un detector

La teoría de la detección tiene como objetivo encontrar el detector más eficaz posible mediante el diseño del estadístico $\mathcal{T}(\boldsymbol{x})$ y la selección del umbral λ. Desde un punto de vista estadístico, en teoría de detección se han utilizado dos variantes para tal propósito, encontrándonos con el detector de Bayes [5] y el detector de Neyman-Pearson [6].

Ambos métodos se basan en la utilización de la relación de verosimilitud $\mathcal{T}(\boldsymbol{x}) = \Lambda(\boldsymbol{x})$ (*LR*, *"Likelihood Ratio"*) o una derivación de ella, definida como el cociente entre las *PDFs* $f(\boldsymbol{x}|H_1)$ y $f(\boldsymbol{x}|H_0)$, las cuales contienen toda la información probabilística que define los mecanismos de transición entre la hipótesis acontecida y

las observaciones tomadas. En muchos casos, el estadístico obtenido mediante la relación de verosimilitud se puede simplificar aplicando una función monótona creciente a ambos lados de la inecuación, dando lugar a lo que se conoce como un estadístico suficiente. Por ejemplo, una función muy utilizada para simplificar la relación de verosimilitud cuando ambos lados de la inecuación son positivos es el logaritmo neperiano:

$$ln\big(\Lambda(\boldsymbol{x})\big) \lessgtr_{H_1}^{H_0} ln(\lambda) \tag{1.5}$$

Ambos test difieren en la filosofía de la selección del umbral λ. El test de Bayes define una función riesgo $\Re$ que trata de minimizar en base a unos costes predefinidos C_{ij}, donde el primer subíndice i indica la hipótesis elegida y el segundo subíndice j la hipótesis verdadera. Se asume que el coste de las situaciones erróneas es mayor al de las situaciones de correcto funcionamiento: $C_{10} > C_{00}$ y $C_{01} > C_{11}$. La función riesgo $\Re$, en términos de las probabilidades de transición y las regiones de decisión, se define como:

$$\begin{aligned}\Re = C_{00}P_0 \textstyle\int_{X_0} f(\boldsymbol{x}|H_0)\mathrm{d}\mathbf{x} + C_{10}P_0 \int_{X_1} f(\boldsymbol{x}|H_0)\mathrm{d}\mathbf{x} + C_{11}P_1 \int_{X_1} f(\boldsymbol{x}|H_1)\mathrm{d}\mathbf{x} + \\ + C_{01}P_1 \textstyle\int_{X_0} f(\boldsymbol{x}|H_1)\mathrm{d}\mathbf{x}\end{aligned} \tag{1.6}$$

Mediante la minimización de la función de coste, se puede llegar a la siguiente regla de decisión:

$$\Lambda(\boldsymbol{x}) = \frac{f(\boldsymbol{x}|H_1)}{f(\boldsymbol{x}|H_0)} \lessgtr_{H_1}^{H_0} \frac{P_0(C_{10} - C_{00})}{P_1(C_{01} - C_{11})} \triangleq \lambda \tag{1.7}$$

Bajo la filosofía de Neyman-Pearson, el umbral λ se escoge fijando una restricción en la probabilidad de falsa alarma y utilizando un test que maximice la P_d bajo esa restricción. Se usan para ello multiplicadores de Lagrange; minimizando la función *F* se maximiza la P_d:

$$P_f = \alpha' \leq \alpha \;\rightarrow F = 1 - P_d + \lambda\big[P_f - \alpha'\big] = \lambda(1 - \alpha') + \int_{X_0} f(\boldsymbol{x}|H_1) - \lambda \cdot f(\boldsymbol{x}|H_0)d\boldsymbol{x} \tag{1.8}$$

Del análisis anterior obtenemos el siguiente resultado para el test de Neyman-Pearson, donde el umbral λ se obtiene de fijar la $P_f = \alpha'$:

$$\Lambda(\boldsymbol{x}) = \frac{f(\boldsymbol{x}|H_1)}{f(\boldsymbol{x}|H_0)} \lessgtr_{H_1}^{H_0} \lambda \;\leftrightarrow P_f = \int_{\lambda}^{\infty} f(\boldsymbol{x}|H_0)\mathrm{d}\mathbf{x} = \alpha' \tag{1.9}$$

En teoría de clasificación es común que se seleccione la clase *c* cuya probabilidad a posteriori $P(c|\boldsymbol{x})$ es máxima, lo que se conoce con el criterio *MAP* ("*Maximum a Posteriori*"). En algunos estudios, el problema de la detección es entendido como un

problema de clasificación binaria, donde cada una de las clases representa a una de las hipótesis. Así, el criterio *MAP* aplicado a un problema de detección:

$$P(H_1|\boldsymbol{x}) \lessgtr_{H_1}^{H_0} P(H_0|\boldsymbol{x}) \tag{1.10}$$

La probabilidad a posteriori de cada hipótesis $\left(H_j,\ j = 0{,}1\right)$, puede relacionarse con su *PDF* fácilmente a través del teorema de Bayes:

$$P\left(H_j|\boldsymbol{x}\right) = \frac{f\left(\boldsymbol{x}|H_j\right)P(H_j)}{f(\boldsymbol{x}|H_0)P(H_0) + f(\boldsymbol{x}|H_1)P(H_1)} \tag{1.11}$$

Se puede comprobar que la probabilidad a posteriori es el resultado de una transformación monótona creciente de la relación de verosimilitud, por lo tanto actúa como un estadístico suficiente válido para la implementación de un detector mediante su umbralizado. Observamos como mediante (1.10) y (1.11) se puede relacionar el criterio *MAP* con la *LR*:

$$f(\boldsymbol{x}|H_1)P(H_1) \lessgtr_{H_1}^{H_0} f(\boldsymbol{x}|H_0)P(H_0) \rightarrow \frac{f(\boldsymbol{x}|H_1)P(H_1)}{f(\boldsymbol{x}|H_0)P(H_0)} \lessgtr_{H_1}^{H_0} 1 \rightarrow \frac{P(H_1)}{P(H_0)}\Lambda(\boldsymbol{x}) \lessgtr_{H_1}^{H_0} 1 \tag{1.12}$$

1.2. – Fusión de datos en problemas de detección

Concepto de fusión

La integración de información hace referencia a la unificación de datos con diferente representación contextual y/o conceptual proveniente de distintas fuentes [7]. Dentro del contexto de integración de información, encontramos el campo de fusión de datos, definido de forma general como el uso de técnicas que combinan datos de múltiples fuentes de forma que, usando la congregación de los datos se puedan alcanzar inferencias, deducciones o conclusiones de forma más efectiva y adaptada a la realidad de lo que se puede conseguir usando los datos de las diferentes fuentes de forma aislada [8].

La fusión de datos en problemas de detección se centra en la adquisición, procesado y combinación sinérgica de varias fuentes de información recogidas por uno o varios sensores para suministrar un mejor conocimiento del fenómeno a considerar [9]. Las técnicas de fusión de datos tratan de combinar múltiples fuentes de información para alcanzar la exactitud y precisión en la toma de decisiones que no sería posible conseguir con el uso de una sola fuente de información de forma aislada. La fusión de múltiples fuentes de información, además de añadir una cierta redundancia que aporte fiabilidad y robustez, puede proporcionar información complementaria con la que incrementar las prestaciones y precisión en el proceso de toma de decisiones. En la figura 1.2 se muestra un esquema del proceso de fusión, donde se representa como diversas fuentes de información se fusionan en una única, mejorando las características de separabilidad entre hipótesis.

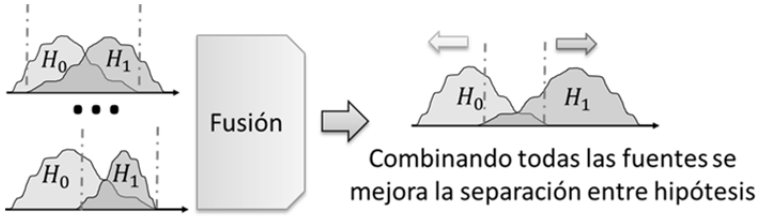


Figura 1.2 – Mejora de las prestaciones de detección mediante la fusión de información

Podemos encontrar áreas o aplicaciones donde es común trabajar con diversas fuentes de información y es necesaria una combinación de información. Por ejemplo, en problemas de detección distribuida [10] se utiliza un sistema multisensor compuesto por un conjunto de sensores iguales colocados con una determinada distribución espacial. Sistemas multimodales [11], donde se utilizan fuentes de información de diversa naturaleza procedente de diferentes sensores y/o fuentes de información no sensada (tales como recursos web, bases de datos, metadatos…) son comunes en sistemas biométricos o problemas de análisis multimedia. Existen también técnicas de detección, como por ejemplo sucede en la combinación de expertos [12], donde se intentan combinar una serie de detectores simples para conseguir resultados que serían muy difíciles de alcanzar mediante un único detector y/o reducir la complejidad y recursos de computación que requeriría la implementación de éste.

Las variables aleatorias que caracterizan los datos de cada fuente de información pueden seguir diferentes distribuciones de probabilidad. Puede ser porque sean datos provenientes de diversos tipos de sensores que captan distintos fenómenos producidos por el mismo evento, o de sensores del mismo tipo, pero cuya posición relativa a la fuente que produce el evento se traduce en diferente distribución de sus observaciones. Muy común es utilizar datos que provienen de la extracción de múltiples características relacionadas con diversos aspectos físicos o de diferente naturaleza de un único flujo de información proporcionado por un sólo sensor. En otros casos, se pretenden combinar diferentes algoritmos o técnicas de procesado con información de salida muy dispar.

El uso de múltiples fuentes, además de información redundante, puede aportar información complementaria, y, mediante la combinación de todas las fuentes se puede incrementar la información discriminatoria sobre la ocurrencia o no del evento a detectar, y así, mejorar tanto las prestaciones globales de detección, como la robustez y la fiabilidad. Todas las posibles fuentes de información generalmente suelen compartir un origen común en el cuál se origina el evento a detectar, por lo que es común encontrar la existencia de dependencia estadística entre ellas. La dependencia estadística entre las fuentes puede introducir información

complementaria de la cual nos podemos beneficiar para mejorar las prestaciones de detección.

Existen en consecuencia diferentes costes, problemas o complejidades añadidas en el proceso de análisis debido a las características de los flujos de información que se pretenden combinar [11]:

- Los flujos de datos capturados por diferentes sensores u obtenidos mediante diferentes técnicas de procesado pueden estar en diferentes formatos, soporte y/o tasas. Además, el tiempo de procesado de cada uno de los flujos puede no ser el mismo. En el proceso de fusión, se tienen que tener en cuenta todas estas posibles asincronías e intentar solventarlas.
- Dada la heterogeneidad de las fuentes de información, las variables aleatorias que las caracterizan pueden poseer diferentes distribuciones de probabilidad.
- El hecho de que las mediciones de los sensores o las características extraídas de éstas provengan de un entorno común o se centren sobre el mismo fenómeno provoca normalmente que las diferentes fuentes de información con las que se trabaja no sean independientes. La dependencia estadística de los datos puede jugar un papel muy importante en las prestaciones obtenidas en el proceso de detección. La consideración de la posible dependencia de los datos puede suponer mayor complejidad en el diseño del sistema, pero puede mejorar de manera notable la eficacia de la detección.

Tipos de información, datos y niveles de fusión en problemas de detección.

Se puede realizar una clasificación de la fusión de datos en problemas de detección atendiendo a diferentes criterios.

Atendiendo a la naturaleza de las fuentes de información de donde provienen los datos podemos clasificar la fusión en unimodal o multimodal.

- **Fusión unimodal**: todos los datos que se combinan, o bien comparten una única fuente común, o bien diferentes fuentes, pero todas de la misma naturaleza. Un caso de fusión unimodal es aquel en el que se combinan diferentes tipos de características, algoritmos de procesado o técnicas de detección extraídas o aplicadas a una única fuente de datos. Otra posibilidad puede ser que los datos sean extraídos de múltiples sensores iguales. Se considera también fusión unimodal el combinar varias instancias capturadas con un mismo sensor.
- **Fusión multimodal**: cuando los datos que se pretenden combinar proceden de fuentes con diferente naturaleza, usualmente provenientes de diferentes tipos de sensores.

Se puede hablar de diferentes niveles de fusión según en la etapa del proceso de detección en que se realiza la integración de la información. Un problema de detección se puede dividir generalmente en cuatro etapas (figura 1.3). La primera de ellas es la etapa de sensado, donde uno o varios sensores se encargan de obtener una serie de mediciones del entorno donde se produce el evento. Estos datos en bruto ($\boldsymbol{y}$) son procesados para extraer ciertas características del evento ($\boldsymbol{x}$), de las cuales unos detectores o algoritmos de clasificación binaria extraerán la información necesaria para poder emitir unas valoraciones ($\boldsymbol{z}$), normalmente relativas a la probabilidad de que se haya producido el evento a detectar, que, tras ser umbralizadas dan lugar a un conjunto de decisiones binarias ($\boldsymbol{u}$) sobre la ocurrencia o no del evento.

Así, se puede hablar de nivel de **fusión de sensores** cuando se combinan los diferentes flujos de muestras proporcionados por estos $(\boldsymbol{y} \rightarrow y_{fus})$, de nivel de **fusión de características** $(\boldsymbol{x} \rightarrow x_{fus})$, de nivel de **fusión de valoraciones** $(\boldsymbol{z} \rightarrow z_{fus})$ y de nivel de **fusión de decisiones** $(\boldsymbol{u} \rightarrow u_{fus})$. Los niveles de fusión de sensores y características también se agrupan en lo que se conoce como fusión temprana o de pre-detección. Los niveles de fusión de valoraciones y de decisiones suelen denominarse también de fusión tardía o de fusión de detectores [7], [13], [14]. Los diferentes niveles de fusión no son excluyentes, de manera que se pueden diseñar sistemas que combinen la fusión a diferentes niveles a partir de la información proporcionada por diferentes fuentes a lo largo de todas las etapas.

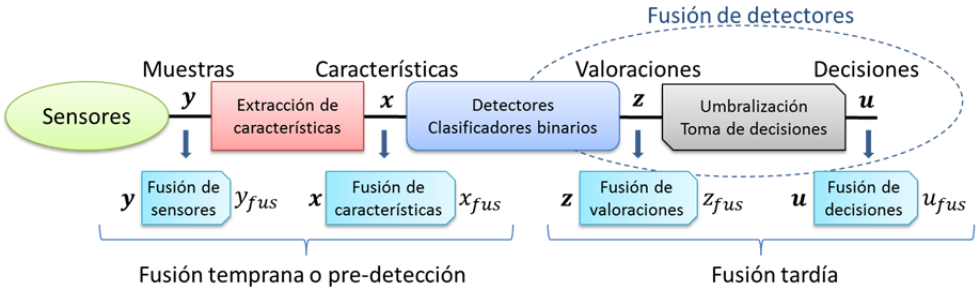


Figura 1.3 – Esquema de un problema de fusión de datos en un sistema de detección.

Atendiendo al tipo de datos a combinar podemos clasificar la fusión en dos grupos, fusión soft o fusión hard:

- La **fusión soft** hace referencia a la combinación de datos continuos, los cuales se pueden modelar mediante variables aleatorias continuas, caracterizadas generalmente mediante las *PDFs* bajo cada una de las hipótesis. Dentro del proceso de detección, generalmente podemos encontrarnos información soft en los datos proporcionados por los sensores ($\boldsymbol{y}$), las características extraídas

de éstos ($\boldsymbol{x}$) y en las valoraciones o scores proporcionados por algunos tipos de detectores ($\boldsymbol{z}$).

- La **fusión hard** combina datos discretos, los cuales se modelan mediante variables aleatorias (*v.a*) discretas caracterizadas por las funciones de masa de probabilidad (*PMFs*, "*Probability Mass Functions*") bajo cada una de las hipótesis. La fusión de información hard es común asociarla únicamente a la etapa de fusión de detectores, donde se utilizan *v.a* y *PMFs* binarias.

Es muy común que ciertos tipos de detectores o clasificadores binarios aporten probabilidades a posteriori como valoraciones z (salida soft). Muchos estudios consideran que la fusión soft de detectores se basa únicamente en la combinación de probabilidades a posteriori. La denominan de forma genérica como fusión de scores, denotando como score s_i a la valoración soft dada por la probabilidad a posteriori $P_i(H_1|x_i) \in [0,1]$ reportada el detector "*i*".

Existen detectores que no necesariamente aportan probabilidades a posteriori como salida soft; por ejemplo pueden aportar como valoración cualquier estadístico $\mathcal{T}(\mathrm{x})$ continuo; por ejemplo la $LR = \Lambda(x) \in [0, +\infty]$ o algún estadístico suficiente derivado de ella, como por ejemplo la $LLR = ln\big(\Lambda(x)\big) \in]-\infty, +\infty[$.

Otros detectores aportan otro tipo de información no estadística, como por ejemplo los basados en máquinas de soporte de vectores (*SVM*, "*Support Vector Machine*"), que proporcionan como valoración la distancia de la observación $\boldsymbol{x}$ a la hipersuperficie que definen como frontera de separación entre hipótesis. En sistemas biométricos de reconocimiento de personas es común la comparación de la medición biométrica del sujeto que se pretende reconocer con respecto a una serie de mediciones almacenadas en una base de datos, por lo que las valoraciones suelen ser diferentes medidas de similitud o disimilitud (figura 1.4).

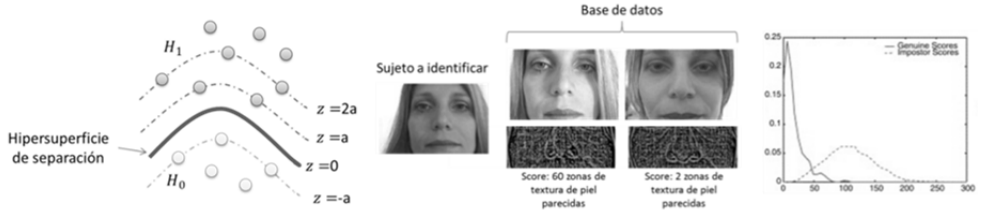


Figura 1.4 – Valoraciones basadas en medidas de distancias a la izquierda. Medida de similitud como valoración en un sistema biométrico a la derecha.

En ciertas técnicas de fusión de datos soft es conveniente que los datos se encuentren definidos en un mismo rango común para su integración. En muchos casos, debido a la heterogeneidad de los datos esto no se cumple y se utilizan diferentes técnicas de normalización y/o calibración, generalmente mapeando los datos en el rango normalizado $[0,1]$ (ver apéndices A y B). En nuestro trabajo

consideraremos que la fusión de scores es un caso particular de fusión soft de diferentes detectores, en donde todos los datos a fusionar se encuentran normalizados en un mismo rango $[0,1]$, sean o no probabilidades a posteriori reportadas por los diversos detectores.

Fusión optima

Sea $\boldsymbol{X} = [X_1, \dots, X_d]^T$ un vector de *v.a* continuas, el cuál agrupa un conjunto de d fuentes de información. La fusión óptima en el contexto de la detección pasa por obtener el estadístico basado en la relación de verosimilitud. El estadístico $\Lambda(\boldsymbol{x})$ realiza la fusión de los datos $\mathbb{R}^d \to \mathbb{R}$, garantizando que las prestaciones en detección son las óptimas:

$$\Lambda(\boldsymbol{x}): \mathbb{R}^d \to \mathbb{R} \implies \Lambda(\boldsymbol{x}) = \frac{f(\boldsymbol{x}|H_1)}{f(\boldsymbol{x}|H_0)} \lessgtr_{H_1}^{H_0} \lambda \tag{1.13}$$

En el caso de datos hard (los asociamos directamente con las decisiones individuales de un grupo de detectores, y por lo tanto son considerados datos binarios) el vector de las *v.a* binarias viene dado por $\boldsymbol{U} = [U_1, \dots, U_d]^T$. La regla de fusión hard óptima $\boldsymbol{u}: \{0,1\}^d \to u_{fus}: \{0,1\}$ pasa por la umbralización de la relación de verosimilitud (*LR*) del vector de decisiones binarias $\boldsymbol{u}$. En este caso, donde los datos a combinar son hard, se define como el cociente entre las funciones de masa de probabilidad (*PMF*, "*Probability Mass Function*") condicionadas las hipótesis H_1 y H_0:

$$\Lambda(\boldsymbol{u}) = \frac{P(\boldsymbol{u}|H_1)}{P(\boldsymbol{u}|H_0)} = \frac{P(u_1, \dots, u_d|H_1)}{P(u_1, \dots, u_d|H_0)} \gtrless_{H_0}^{H_1} \eta \Leftrightarrow u_{fus} = \begin{cases} 1 & si\ \Lambda(\boldsymbol{u}) \geq \eta \\ 0 & si\ \Lambda(\boldsymbol{u}) < \eta \end{cases} \tag{1.14}$$
$$\boldsymbol{u}: \{0,1\}^d \to \Lambda(\boldsymbol{u}): \{a_1, \dots, a_{2^d}\} \subset \mathbb{R}$$

Al ser $\boldsymbol{u}$ un vector de variables aleatorias binarias, existirán 2^d realizaciones diferentes de él, por lo tanto la *LR* en este caso se define en un subconjunto discreto de 2^d posibles valores reales.

En determinadas aplicaciones o situaciones no será posible o viable implementar la fusión óptima de los datos (tanto soft como hard) y se deben acudir a otras técnicas, generalmente subóptimas para la combinación de los datos. Muchas de estas técnicas se basan en la asunción de independencia estadística entre los datos para simplificar la expresión de la *LR* y obtener reglas de fusión más simples. Como veremos en el siguiente punto, la asunción de independencia en problemas de detección puede introducir una importante degradación de prestaciones.

1.3. – Importancia de la dependencia estadística en problemas de detección

A continuación mostramos un sencillo ejemplo para ilustrar la importancia que conlleva considerar la dependencia estadística que existente entre los datos en un problema de fusión [15].

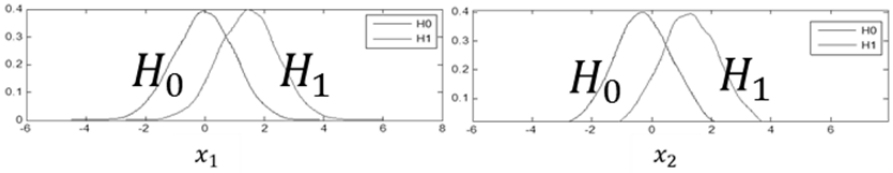


Figura 1.5 – PDFs de los datos soft a fusionar

Consideremos un caso en que disponemos de dos fuentes de información soft, bien sean muestras directamente proporcionadas por sensores, dos características extraídas mediante el procesado de las muestras o valoraciones continuas aportadas por dos detectores. Estos datos están caracterizados por las variables aleatorias x_1 y x_2, ambas con distribuciones Gaussianas bajo cada una de las hipótesis, las cuales son equiprobables $P(H_1) = P(H_0)$ (figura 1.5):

$$\begin{aligned} f(x_1|H_0) &= N_{x_1}(\mu_{0_1} = 0,\ \sigma_{0_1} = 1) \qquad f(x_1|H_1) = N_{x_1}(\mu_{1_1} = 1.5, \sigma_{1_1} = 1) \\ f(x_2|H_0) &= N_{x_2}(\mu_{0_2} = -0.3, \sigma_{0_2} = 1) \quad f(x_2|H_1) = N_{x_2}(\mu_{1_2} = 1.3, \sigma_{1_2} = 1) \end{aligned} \tag{1.15}$$

Pretendemos combinar de forma óptima ambas fuentes de información para mejorar las prestaciones de los detectores individuales que se obtienen con los datos por separado. Para ello las fusionaremos obteniendo la *LR* del vector de datos conjunto $\boldsymbol{x} = [x_1\ x_2]$.

Primero consideramos que ambas variables aleatorias son independientes, por lo que las *PDFs* conjuntas $f(\boldsymbol{x}|H_j),\ j = 0{,}1$ pueden expresarse como el producto de las marginales. Trabajando con la expresión de la *LR* se deriva un estadístico $\mathcal{T}^{Ind}(\boldsymbol{x})$ que fusiona de forma óptima ambas *v.a* bajo la asunción de independencia:

$$\Lambda^{Ind}(\boldsymbol{x} = [x_1\ x_2]) = \frac{f(x_1|H_1)f(x_2|H_1)}{f(x_1|H_0)f(x_2|H_0)} \rightarrow \mathcal{T}^{Ind}(\boldsymbol{x}) = a{x_1}^2 + bx_1 + c{x_2}^2 + dx_2 \tag{1.16}$$

donde a, b, c, d son constantes.

Consideramos ahora que existe dependencia lineal entre ambas variables aleatorias, por lo tanto, el vector $\boldsymbol{x}$ sigue una distribución conjunta Gaussiana multivariante bajo cada una de sus hipótesis H_i, parametrizada por la matriz de covarianza $\boldsymbol{R}_i$ y el vector de medias $\boldsymbol{\mu}_i$:

$$f(\boldsymbol{x}|H_i, \boldsymbol{\mu_i}, \boldsymbol{R_i}) = \frac{1}{(2\pi)^{N/2} \cdot |\boldsymbol{R_i}|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu_i})^{\boldsymbol{T}} \boldsymbol{R_i}^{-1} (\boldsymbol{x}-\boldsymbol{\mu_i})} \tag{1.17}$$

La dependencia lineal bajo cada hipótesis puede ser parametrizada mediante un coeficiente de correlación ρ:

$$\boldsymbol{R}_1 = \begin{pmatrix} {\sigma_{1_1}}^2 & \sigma_{1_{12}} \\ \sigma_{1_{12}} & {\sigma_{1_2}}^2 \end{pmatrix} \quad \boldsymbol{R}_0 = \begin{pmatrix} {\sigma_{0_1}}^2 & \sigma_{0_{12}} \\ \sigma_{0_{12}} & {\sigma_{0_2}}^2 \end{pmatrix} \quad \rho_1 = \frac{\sigma_{1_{12}}}{\sigma_{1_1} \cdot \sigma_{1_2}} \quad \rho_0 = \frac{\sigma_{0_{12}}}{\sigma_{0_1} \cdot \sigma_{0_2}} \tag{1.18}$$

Si obtenemos el logaritmo de la relación de verosimilitud y despreciamos los valores constantes obtenemos en este caso un estadístico $\mathcal{T}^{Dep}(\boldsymbol{x})$:

$$\begin{gathered} \mathcal{T}^{Dep}(\boldsymbol{x}) = (\boldsymbol{x}-\boldsymbol{\mu_0})^{\boldsymbol{T}} {\boldsymbol{R_0}}^{-1} (\boldsymbol{x}-\boldsymbol{\mu_0}) - (\boldsymbol{x}-\boldsymbol{\mu_1})^{\boldsymbol{T}} {\boldsymbol{R_1}}^{-1} (\boldsymbol{x}-\boldsymbol{\mu_1}) \\ \mathcal{T}^{Dep}(\boldsymbol{x}) = a' {x_1}^2 + b' x_1 + c' {x_2}^2 + d' x_2 + e x_1 x_2 \end{gathered} \tag{1.19}$$

Ahora aparece un nuevo término ex_1x_2 dependiente de ambas variables. Observamos cómo, según exista dependencia o no entre las variables aleatorias implicadas, obtenemos reglas de fusión distintas (1.16) y (1.19) respectivamente.

Consideramos ahora el caso de la fusión de decisiones o fusión hard. Aplicando la relación de verosimilitud $\Lambda(x_i)$ a cada variable aleatoria x_i, $i = 1{,}2$ por separado, se puede deducir que el estadístico suficiente para implementar un detector óptimo bajo el criterio de Neyman-Pearson es la propia variable aleatoria:

$$\Lambda(x_i) = \frac{f(x_i|H_1)}{f(x_i|H_0)} \;\rightarrow\; x_i \lessgtr_{H_1}^{H_0} \lambda \tag{1.20}$$

Así, mediante la umbralización de las variables aleatorias obtenemos un vector de decisiones hard $\boldsymbol{u} = [u_1\, u_2]$. Implementamos la regla de fusión óptima $\Lambda(\boldsymbol{u})$, la cual considera la dependencia en las decisiones individuales. En este caso, al tratarse de variables aleatorias binarias, la inclusión de la dependencia entre ellas en la regla de fusión es más sencilla que el caso de variables aleatorias continuas. En el capítulo 4 se puede encontrar un completo análisis de la fusión hard de dos detectores:

$$\boldsymbol{u} = [u_1\, u_2], \quad u_i = \begin{cases} 1\,, & x_i \geq \lambda_i \\ 0\,, & x_i < \lambda_i \end{cases} \rightarrow \quad \Lambda(\boldsymbol{u}) = \frac{P(\boldsymbol{u}|H_1)}{P(\boldsymbol{u}|H_0)} \lessgtr_{H_1}^{H_0} \lambda \tag{1.21}$$

Se han realizado unas simulaciones con datos sintéticos para ilustrar, por un lado, que mediante la fusión o integración de dos fuentes de información se puede mejorar el resultado que se obtiene si se consideran de forma aislada, y por otro, poder comparar y entender el proceso de la fusión soft óptima que tiene en cuenta la posible dependencia, la fusión soft considerando que las fuentes son independientes y la fusión hard.

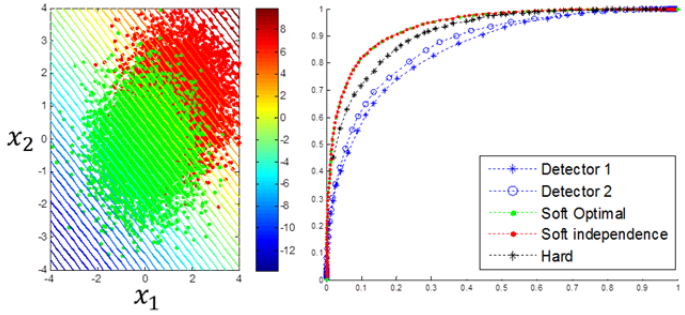


Figura 1.6 – Fusión de dos variables aleatorias Gaussianas independientes

Comenzamos con el caso en el que ambas variables aleatorias son independientes bajo ambas hipótesis ($\rho_1 = 0$ $\rho_0 = 0$). Representamos en la figura 1.6, a la izquierda las fronteras de separación óptima en zonas pertenecientes a H_1 o H_0 según el umbral fijado, y a la derecha la curva ROC que representa el funcionamiento de los detectores individuales, así como las posibilidades de fusión consideradas. En la figura 1.7 se representa la división en zonas que realiza cada técnica de detección, fijando en todas ellas los umbrales correspondientes para obtener una $P_f = 0.1$. Observamos como en este caso, las fronteras de separación óptimas equivalen a rectas que dividen el espacio de las variables aleatorias en dos zonas. Ambos estadísticos (1.16) (1.19) derivan en la misma expresión. Observamos en la curva *ROC* como tanto la fusión soft, como la fusión hard mejoran las prestaciones de los detectores individuales, obteniendo mejores resultados con la fusión soft.

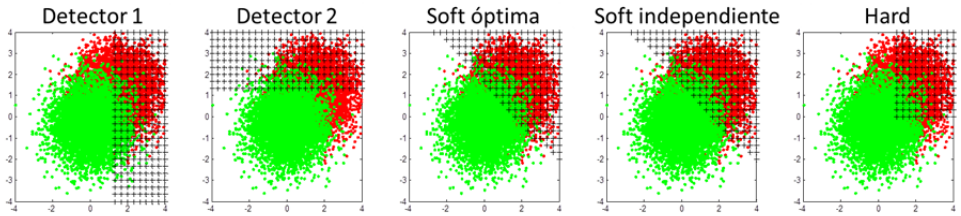


Figura 1.7 – Regiones de decisión para $P_f = 0.1$ Fusión de dos v.a Gaussianas independientes

Consideramos ahora el caso $\rho_1 = -0.3$ $\rho_0 = 0$, en que existe cierta dependencia bajo la hipótesis H_1. Se representa este caso en la figura 1.8. Observamos como la separación óptima entre zonas pertenecientes a distintas hipótesis varía. Pese a esto, comprobamos como las prestaciones obtenidas con la regla de fusión teniendo en cuenta la dependencia no difieren mucho del caso en que asumimos independencia. Observamos como la fusión hard sigue consiguiendo un resultado mejor que cada detector individual por separado, pero sigue siendo peor que la fusión soft tanto de forma óptima como asumiendo independencia.

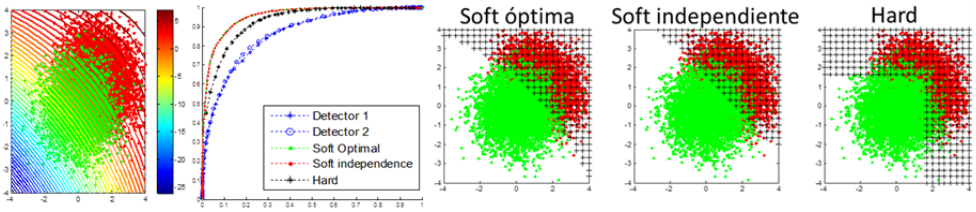


Figura 1.8 – Fusión de dos variables aleatorias Gaussianas dependientes con dependencia caracterizada por los coeficientes de correlación $\rho_1 = -0.3 \quad \rho_0 = 0$

Analizamos ahora los supuestos en los que existe dependencia parametrizada por los coeficientes de correlación $\rho_1 = 0.8 \ \rho_0 = 0$ (figura 1.9) y $\rho_1 = 0.3 \ \rho_0 = 0.8$ (figura 1.10). En estos casos, la separación óptima entre zonas pertenecientes a distintas hipótesis varía de forma considerable al caso en el que exista independencia. Observamos incluso, como la fusión hard incorporando dependencia estadística entre las decisiones es mejor que la fusión soft sin incorporar la información de dependencia.

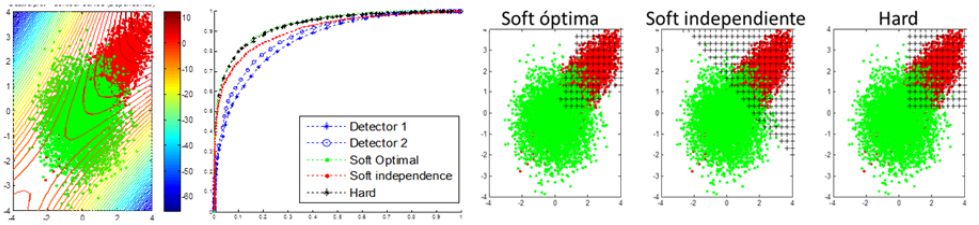


Figura 1.9- Fusión de dos variables aleatorias Gaussianas dependientes con dependencia caracterizada por los coeficientes de correlación $\rho_1 = 0.8 \ \rho_0 = 0$

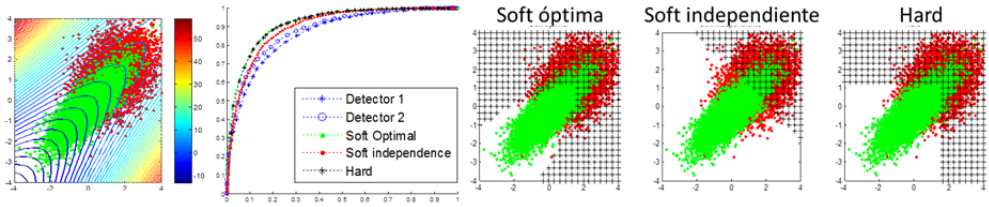


Figura 1.10 - Fusión de dos variables aleatorias Gaussianas dependientes con dependencia caracterizada por los coeficientes de correlación $\rho_1 = 0.3 \ \rho_0 = 0.8$

Mediante este sencillo ejemplo se ha puesto de manifiesto que la existencia de dependencia entre datos en un problema de detección y su correcta caracterización juega un rol muy importante en cuanto a las prestaciones que se pueden obtener. Se ha comprobado como mediante la integración o fusión de varias fuentes de información se pueden mejorar los resultados obtenidos al considerar sólo una fuente

de forma aislada. Además, se ilustra el proceso de fusión optima y la diferencia entre fusión hard y fusión soft.

1.4. – Motivación y objetivos

Esta tesis se enmarca dentro del proyecto: "*Tratamiento integrado de señales multimedia*" 2010-13, del programa *PROMETEO* para grupos de excelencia de la Comunidad Valenciana. En él se contempla el estudio de algoritmos de fusión como vía de mejora de algoritmos clásicos de tratamiento estadístico de señales. En particular la tesis se centra en el área de la detección.

El objeto de la tesis es el estudio y uso de técnicas de fusión de datos para la resolución de problemas de detección cuando se dispone de diferentes fuentes de información que pueden ser heterogéneas y dependientes entre sí. Como se comentó en el punto anterior, debido a que los sensores se encargan de observar diferentes aspectos de un mismo fenómeno, o existen detectores que trabajan sobre los mismos datos recogidos por un único sensor, nos podemos encontrar con la existencia de dependencia en los datos a fusionar. La naturaleza de esta dependencia puede ser no lineal y muy compleja [2].

El problema de fusionar de forma óptima según el criterio Bayesiano o de Neyman-Pearson estos datos dependientes no es sencillo y muchas de las soluciones adoptadas son específicas para un problema en concreto o se usan métodos subóptimos [6] [10]. Por ejemplo, se suelen usar técnicas que obvian la dependencia entre los datos y los consideran independientes entre sí [17], o se hace uso de modelos simples de dependencia, como por ejemplo puede ser el considerar modelos multivariantes Gaussianos, donde la dependencia modelada es lineal y limitan la caracterización de las distribuciones marginales como distribuciones Gaussianas [18].

Así, se ha abordado un estudio sobre la naturaleza y características de los diferentes datos que se pueden combinar, los distintos niveles en los que se puede realizar la integración de información y los métodos o técnicas que pueden ser utilizadas para la fusión de diversos canales de datos. Se introducen los métodos óptimos de fusión de datos y las diferentes técnicas que pueden ser utilizadas para su implementación. Se argumenta cómo en algunas situaciones no será posible la implementación o utilización de estas técnicas óptimas y deberán utilizarse ciertas técnicas subóptimas, donde, algunas de ellas se obtienen o derivan desde la asunción de independencia estadística de los datos. Se mostrará como la asunción de independencia puede conducir a una severa reducción de las prestaciones de detección cuando los datos presentan determinadas estructuras de dependencia estadística.

De forma más específica, los objetivos en la tesis son:

- Revisión del estado del arte en cuanto a técnicas (tanto óptimas como subóptimas) de fusión y combinación de datos aplicadas en problemas de detección, en donde los datos pueden ser heterogéneos y dependientes entre sí.
- Se hará especial hincapié, con un estudio en mayor profundidad de la técnica de estimación de funciones de densidades de probabilidad multivariantes basada en la teoría de cópulas, la cual puede ser usada en la fusión óptima de datos soft. La destacamos de forma especial por su novedad e incipiente uso en el campo del procesado de señal.
- Implementación, estudio y análisis de estas técnicas de fusión de datos.
- Demostrar que, tanto el uso de más de un sensor o fuente de información, como una correcta consideración de las características de dependencia entre ellos, pueden ayudar a mejorar el rendimiento y precisión obtenidas a la hora de diseñar un sistema de detección.
- Proponer modificaciones o variantes de las técnicas y algoritmos existentes, de forma que se mejore su rendimiento incorporando información sobre la dependencia existente o proporcionando una mejor caracterización de ella.
- Aplicar este estudio y técnicas en problemas de detección reales, donde se pretende usar múltiples sensores para mejorar las prestaciones que se obtienen con un único sensor, tal y como se describe en el siguiente apartado.

1.5. – Estructura de la tesis

El presente trabajo se estructura distinguiendo dos partes. En una primera parte nos centramos en el estudio y análisis de las diferentes técnicas que pueden ser usadas para la fusión de información en problemas de detección. Discriminando las técnicas según las características de los datos que se pretenden integrar o fusionar se ha dividido esta primera parte en tres capítulos:

- Fusión de datos soft: aglutina el conjunto de técnicas para combinar datos continuos.
- Fusión de scores: en este capítulo destacamos, como un caso particular de fusión soft, las técnicas de combinación de datos continuos que se encuentran definidos en un mismo rango normalizado $[0,1]$. En muchas aplicaciones este tipo de datos consisten en valoraciones aportadas por un conjunto de detectores representando conceptualmente estimaciones de probabilidades a posteriori $\left(P(H_1|\boldsymbol{x})\right)$.
- Fusión hard: se recogen las técnicas de combinación de datos hard, usualmente en problemas de detección representadas por un conjunto de decisiones binarias aportadas por diversos tipos de detectores.

En esta primera parte, con objeto de realizar un compendio estructurado de las diferentes técnicas de fusión, se entremezcla tanto el estado del arte como las diferentes contribuciones que aporta el presente trabajo de investigación. Estas contribuciones se pueden enumerar en:

- En el apartado de fusión de datos soft se introduce la teoría de las cópulas cómo un método de novedosa e incipiente aplicación en aplicaciones de tratamiento de señal, argumentando la idoneidad de su utilización como método de estimación de funciones de densidad multivariante en problemas de fusión en detección y destacando las ventajas que supone su utilización con respecto al resto de técnicas que se pueden encontrar en la literatura.
- En el apartado de fusión de scores, se propone una nueva técnica de fusión denominada integración-α, capaz de mejorar las prestaciones que proporcionan el resto de técnicas presentadas en ese punto.
- Se propone un nuevo método de entrenamiento basado en el criterio de maximización parcial del área bajo la curva ROC. Se deriva y aplica este método de entrenamiento en la técnica de fusión mediante integración-α.

La segunda parte de la tesis contempla diferentes aplicaciones donde se han usado algunas de las técnicas de fusión comentadas en la primera parte. Estos problemas o aplicaciones no se han elegido de forma arbitraria sino que se relacionan con líneas, proyectos y contratos de investigación del *GTS* del *ITEAM*.

Así, se aplican técnicas de fusión soft en problemas de detección de señales contaminadas con ruido de fondo, más concretamente en la detección de eventos acústicos. El *GTS* ha venido proponiendo nuevas extensiones de detectores conocidos, tales como el detector de energía [19] y el detector adaptado en subespacio [20], apropiados para el caso de que la señal o evento a detectar sean totalmente desconocidos. Dentro del *GTS* la detección de eventos acústicos se aplica en sistemas de vigilancia/monitorización basados en sonido. Proyectos recientes y significativos del *GTS* relacionados con este ámbito son: *Tecnologías para los sistemas de seguridad, video-vigilancia y monitorización remota del futuro (HESPERIA), 2006-08 (*programa *CENIT*), *Monitoring environments from acoustic scene analysis, 2009-10* (Acción Integrada en colaboración con el Instituto de Automática y Robótica de la Universidad de Karlsruhe), *ARtSENSE Augmented RealiTy Supported adaptive and personalized Experience in a museum based oN procesing real-time Sensor Events*, 2011-14, (proyecto *STREP* del Séptimo Programa Marco de la UE).

En este caso, se propone como principal novedad en el problema de la detección de una señal desconocida en ruido un nuevo tipo de detector basado en la teoría de cópulas denominado *COCD*. Así mismo, se ha realizado un estudio sobre la fusión de más de un micrófono como método para mejorar las prestaciones de un sistema de detección de eventos acústicos.

Otra área de aplicación en la que se han utilizado las diferentes técnicas, es en el de fusión de datos en sistemas de autentificación multibiométrica. En el *GTS* se ha trabajado con este tipo de técnicas en la implementación de un sistema de autentificación biométrica basado en electroencefalogramas [21]. En este caso, con objeto de poder testear y comparar las diferentes técnicas, se ha decidido incluir aplicaciones basadas en bases de datos públicas muy usadas en la literatura.

Parte I: Técnicas de fusión

Capítulo 2: Fusión de datos soft

"Todas las cosas por un poder inmortal, cerca o lejos, ocultamente están unidas entre sí, de tal modo que no puedes agitar una flor sin trastornar una estrella"

– Francis Thompson –

En este capítulo se proporciona una visión global del estado del arte en los métodos de fusión o combinación de información soft aplicados en la resolución de un problema de detección. Se han dividido estas técnicas en tres categorías: fusión mediante combinación lineal de los datos, fusión mediante métodos basados en clasificación binaria y fusión mediante estimación de densidades de probabilidad.

Nos centramos con más detenimiento en los métodos de estimación de densidades, ya que la fusión óptima en tareas de detección se consigue mediante un test LR. Se realiza una revisión y estudio en mayor profundidad del método de estimación de PDFs multivariantes mediante la teoría de cópulas por dos motivos principales: primero, su uso en el campo del procesado de señal es novedoso e incipiente, y segundo, el uso de esta técnica permite separar el modelado de las funciones de densidad marginales de cada v.a de la estructura de dependencia estadística existente entre ellas. De esta forma nos permite lidiar de forma más sencilla y con mayor precisión con el modelado de datos heterogéneos y dependientes existentes en una aplicación de detección.

2.1. – Introducción

Supongamos que disponemos de un conjunto compuesto por "*d*" fuentes de información soft que se desean integrar en la resolución de un problema de detección. Cada una de ellas puede ser caracterizada por una variable aleatoria continua $X_i,\ i = 1,..,d$. Tal y como se comentó en el apartado anterior, puede tratarse de datos recogidos directamente de los sensores, características extraídas de éstos o valoraciones proporcionadas por diversos tipos de detectores, clasificadores binarios o algoritmos.

Definimos el vector aleatorio $\boldsymbol{X} = [X_1, \dots, X_d]^T$ que aúna todas las fuentes de información. Muy comúnmente en problemas de detección, se tratará de un vector heterogéneo, es decir, las variables aleatorias marginales X_i que lo componen no estarán igualmente distribuidas. También es probable que exista dependencia estadística entre las variables aleatorias. Tal y como se comentó en el apartado 1.2, el problema de fusión óptimo de estas las fuentes de información pasa por obtener el **estadístico basado en la relación de verosimilitud**. El estadístico $\Lambda(\boldsymbol{x})$ realiza la fusión de los datos $\mathbb{R}^d \to \mathbb{R}$, garantizando que las prestaciones en detección son óptimas:

$$\Lambda(\boldsymbol{x}): \mathbb{R}^d \to \mathbb{R} \implies \Lambda(\boldsymbol{x}) = \frac{f(\boldsymbol{x}|H_1)}{f(\boldsymbol{x}|H_0)} \lessgtr_{H_1}^{H_0} \lambda \tag{2.1}$$

Para la implementación de este estadístico óptimo que integra todas las fuentes de información necesitamos conocer las funciones de densidad de probabilidad

(*PDFs*) conjunta $f(\boldsymbol{x}|H_j)$ bajo cada una de las hipótesis H_j, $j = 0,1$. Generalmente estas funciones de densidad no son conocidas y se debe realizar una estimación de ellas. Para la estimación de las *PDFs* nos valemos de un conjunto de muestras de entrenamiento como representación particular de las variables aleatorias cuya función densidad pretendemos estimar.

La posible heterogeneidad y dependencia estadística de los datos pueden complicar el proceso de estimación de las *PDFs*, necesitando usar técnicas complejas para obtener estimaciones precisas. Generalmente estas técnicas poseen altos requisitos computacionales (almacenamiento de datos y/o capacidad de cálculo elevadas) y requieren de un tiempo elevado para su entrenamiento. En muchos casos se necesitan un número elevado de muestras de entrenamiento para obtener una correcta estimación de las *PDFs*. En determinadas aplicaciones, por tanto, puede no ser factible el uso de estas técnicas de estimación y se deberán usar alternativas.

Una técnica muy usada por su simpleza es la **combinación lineal de los datos**. Esta técnica se basa en aplicar una sencilla función lineal $c(\cdot)$ a los datos multivariantes $\boldsymbol{x}$, combinándolos y proporcionando un sólo canal de datos x_{fus} a la salida:

$$c(\cdot): \mathbb{R}^d \rightarrow \mathbb{R} \;\; \Rightarrow \;\; x_{fus} = \; c(\boldsymbol{x}) \tag{2.2}$$

Otras alternativas usadas en la literatura pasan por enfocar el problema de la fusión de datos desde el punto de vista del aprendizaje automático ("*Machine learning*") o del reconocimiento de patrones ("*pattern recognition*"), aplicando así **técnicas de clasificación binaria** tales como máquinas de soporte de vectores (*SVM*, "*Support Vector Machine*") o redes neuronales entre otras. Algunas de estas técnicas también pueden requerir, al igual que los métodos de estimación de densidades, una etapa de entrenamiento compleja, involucrando altos costes temporales y computacionales; pero por el contrario, es posible pueden conseguir buenos resultados incluso con un escaso número de muestras de entrenamiento.

Esta clasificación de las técnicas se basa en la naturaleza o filosofía de uso de cada método [11], [22]. Por ejemplo, cuando hablamos de *SVM*, uno de los modelos más simples de implementación consiste, inherentemente, en una combinación lineal de los datos; el uso de la técnica general, mediante la utilización de diferentes funciones de kernel, proporciona modelos de combinación más complejos. Al igual ocurre con las redes neuronales, donde la red neuronal denominada como perceptron simple, actúa de igual forma, realizando una combinación lineal de los datos; una mejora de esta red neuronal es la conocida como perceptron multicapa, donde la función de combinación es más compleja, pudiendo ser no lineal. Así, cuando se habla de usar una *SVM* o una red neuronal es porque puede precisarse de un tipo de combinación de datos más complejo; cuando hablamos de técnicas de combinación lineal, de antemano estamos restringidos a ese tipo de modelos más simples.

2.2. – Fusión basada en una combinación lineal de los datos

La combinación lineal de datos soft es una de las técnicas más más utilizadas por su simpleza. Se basa en combinar de forma lineal el conjunto de datos soft $\boldsymbol{x} = [x_1 \dots x_d]^T \in \mathbb{R}^d$, mediante la suma o producto de sus elementos. La combinación mediante el producto estrictamente es una combinación no lineal, que se linealiza mediante una transformación logarítmica. Mantendremos el término lineal por simplificar la presentación.

$$x_{fus} = \sum_{i=1}^{d} x_i \qquad x_{fus} = \prod_{i=1}^{d} x_i \tag{2.3}$$

Es común que unos canales de datos aporten mayor información discriminatoria sobre el evento a detectar que otros. Por ello se suelen utilizar una serie de pesos w_i para ponderar de forma diferente cada uno de los canales de datos a integrar:

$$x_{fus} = \sum_{i=1}^{d} w_i \cdot x_i \qquad x_{fus} = \prod_{i=1}^{d} {x_i}^{w_i} \tag{2.4}$$

En este tipo de técnicas de fusión lineal se considera que los datos bajo ambas hipótesis son separables mediante el hiperplano $w_1 \cdot x_1 + \cdots + w_d \cdot x_d + k = 0$ en el caso de combinación mediante suma o, en el caso de combinación producto, la hipersuperficie ${x_1}^{w_1} \cdot \dots \cdot {x_d}^{w_d} + k = 0$, donde k es una constante arbitraria (ver ejemplos en la figura 2.1).

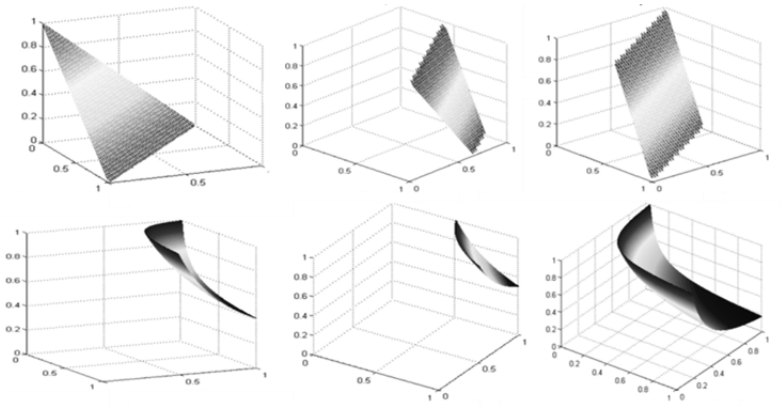


Figura 2.1 – Ejemplos de fronteras de separación entre hipótesis logradas en un espacio de observaciones $\mathbb{R}^3$ mediante una fusión basada en una suma ponderada (fila superior) y en un producto ponderado (fila inferior)

En el caso genérico de heterogeneidad y dependencia entre los diferentes datos, este tipo de técnicas suelen ser subóptimas. Las prestaciones obtenidas dependen del tipo de distribuciones que sigan los datos y las características de dependencia

estadística entre ellos, incluso pudiendo obtener peores resultados tras la fusión, que considerando los detectores de forma aislada.

Ejemplo de fusión mediante combinación lineal

Podemos observarlo de forma clara si retomamos el ejemplo presentado en el apartado 1.3. En él se pretenden fusionar dos variables aleatorias Gaussianas bajo ambas hipótesis. Cuando existe independencia bajo ambas hipótesis entre los datos, utilizando la *LR* se obtiene una regla óptima de fusión dada por:

$$x_{fus} = a{x_1}^2 + bx_1 + c{x_2}^2 + dx_2 \tag{2.5}$$

donde a, b, c y d son constantes. Pese a que esta regla de fusión óptima sea no lineal, se puede demostrar (gráficamente se observa de forma muy clara en la figura 2.2) que es equivalente a una combinación lineal dada por la expresión 2.6. Mediante la umbralización de x_{fus} las diferentes regiones de separación entre hipótesis en el espacio de las observaciones vienen determinadas, al igual que con la expresión 2.5, mediante rectas:

$$x_{fus} = a'x_1 + b'x_2 \tag{2.6}$$

Como ya se comprobó, en el caso que exista dependencia las regiones de separación pueden dejan de ser lineales, por lo que la fusión utilizando una técnica de combinación lineal en esos casos deja de ser óptima.

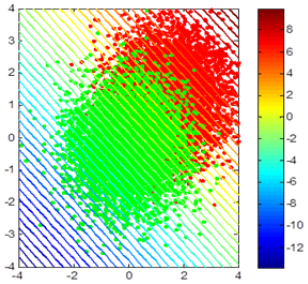


Figura 2.2 – Zonas de separación óptimas en el caso de independencia bajo ambas hipótesis

Normalización de los datos

Debido a la heterogeneidad de los datos usados en problemas de detección, puede ocurrir que los canales de datos estén definidos en diferentes rangos o dominios: $x_i \in [a_i, b_i] \subseteq \mathbb{R}$. Es habitual el normalizar los datos (ver apéndice A), de forma que todos ellos queden definidos en un mismo rango común, habitualmente en el rango $[0,1]$, de forma que, restringiendo las ponderaciones, los datos fusionados queden también definidos en el mismo rango. Sea $\boldsymbol{x}_n = [x_{n1} \dots x_{nd}]^T \in [0,1]^d$ el vector de datos normalizados, se define la combinación lineal como:

$$x_{fus} = \sum_{i=1}^{d} w_i \cdot x_{ni} \quad \text{ó} \quad x_{fus} = \prod_{i=1}^{d} {x_{ni}}^{w_i}, \quad x_{fus} \in [0,1]$$
$$\sum_{i=1}^{d} w_i = 1\,, \qquad 0 \leq w_i \leq 1 \tag{2.7}$$

Este caso especial de combinación, en el que los datos están normalizados en el rango $[0,1]$, se asemeja a lo que cierta corriente de investigación denomina como fusión de scores, entendida como la fusión de información soft a la salida de un conjunto de detectores, consistente en las probabilidades de la clase H_1 condicionadas al vector muestras de entrada de cada uno de los detectores (probabilidades a posteriori). En el capítulo siguiente consideramos este tipo de fusión o combinación soft como un caso especial, y comentaremos con más detalle las técnicas utilizadas, entre las cuales también se encuentran este tipo de combinaciones lineales.

Existen multitud de estudios donde se han aplicado estas técnicas de fusión basadas en combinación lineal de datos. En [11] se realiza un estudio del estado del arte de diferentes técnicas de fusión en aplicaciones multimedia, se puede encontrar recopilados multitud de trabajos donde se hace uso de esta técnica. El uso de estas técnicas de fusión están muy extendidas en aplicaciones de biometría [22]–[26].

2.3. – Fusión mediante técnicas de clasificación binaria

Un problema de detección se puede entender también como un caso particular de un problema de clasificación, en la que sólo se disponen de dos clases, cada una de ellas asociada a una hipótesis. Por lo tanto, se pueden aplicar técnicas de clasificación usadas en ramas como la inteligencia artificial, el aprendizaje automático ("*machine learning*") o el reconocimiento de patrones ("*pattern recognition*").

Así, utilizando técnicas de clasificación binaria también se puede llevar a cabo una fusión o combinación de información soft en detección. Los métodos más usuales usados en esta categoría son las técnicas de análisis discriminante ("*Discriminant Analysis*", bien sea discriminación lineal "*LDA*", cuadrática "*QDA*" o la denominada generalizada, usando diferentes funciones de kernel "*KDA*"), las máquinas de soporte de vectores (*SVM*, "*Support Vector Machine*"), redes Bayesianas dinámicas, redes neuronales, basados en la teoría de Dempster-Shafer y modelos de máxima entropía.

En [11] se puede encontrar una completa descripción de cada uno de estos métodos, así como una extensa colección de referencias a investigaciones en las que se utilizan estas técnicas de fusión de información en aplicaciones de análisis multimedia. En [22], [26] también se recogen una serie de trabajos donde estas

técnicas son usadas para la combinación de información soft en problemas de autenticación biométrica.

Muchas de estas técnicas precisan de una gran carga de trabajo empírico, debiendo probar diversos métodos de entrenamiento, parámetros, configuraciones y/o diferentes kernels para poder obtener la configuración que proporcione mejores resultados. Es complicado generalizar, tanto sobre las prestaciones, como sobre la complejidad que involucra el uso de estas técnicas de fusión de datos en problemas en detección. Por ejemplo, una red neuronal puede ser muy sencilla, como la denominada perceptron simple, por lo que es probable que no se obtengan buenos resultados en casos complejos; por el contrario se pueden encontrar redes neuronales muy complejas, que puedan obtener muy buenas prestaciones en gran parte de problemas, pero requieren de un costoso esfuerzo de implementación, con grandes requerimientos computacionales y temporales para su entrenamiento.

Podemos determinar que estas técnicas poseen como principal inconveniente una falta de generalidad, donde su uso implica una gran carga de trabajo empírico en escenarios donde los datos pueden ser heterogéneos y presentar complejas características de dependencia. Con ellas es posible obtener reglas de combinación, sino completamente óptimas, sí con muy buenas prestaciones, aunque pueden acarrear una elevada complejidad. Por lo tanto, tanto la complejidad que involucra su uso, como las prestaciones obtenidas van a ser muy dependientes de las características del problema concreto en el que se pretenden utilizar y de las configuraciones que de estas técnicas se escojan y prueben.

2.4. – Fusión basada en la estimación de densidades de probabilidad

La relación de verosimilitud $\Lambda(\boldsymbol{x})$ realiza una fusión de los datos $\mathbb{R}^d \to \mathbb{R}$ garantizando que las prestaciones en detección son óptimas. Para poder obtener esta relación de verosimilitud necesitamos conocer las *PDFs* conjuntas bajo cada una de las hipótesis de las *v.a* que caracterizan los diferentes datos. En la mayoría de casos no serán conocidas de antemano, por lo que necesitaremos utilizar alguna técnica de estimación para obtenerlas.

Consideramos que disponemos de un conjunto de entrenamiento $\boldsymbol{\mathcal{X}}$, compuesto por N_j vectores de muestras de dimensión d, obtenidos bajo la hipótesis conocida $H_j, j = 0,1$:

$$\boldsymbol{\mathcal{X}} = \{\boldsymbol{x}^1, \dots, \boldsymbol{x}^{N_j}\},\ \ \boldsymbol{x^i} = \left[x_1^i, \dots, x_d^i\right]^T \epsilon\ H_j\ , i = 1, \dots, N_j \tag{2.8}$$

Se pueden utilizar dos tipos de técnicas para obtener una estimación de las *PDFs* utilizando estos datos de entrenamiento: paramétricas o no paramétricas. Las técnicas paramétricas asumen que las densidades se ajustan a ciertos modelos

estadísticos flexibles conocidos, pudiendo adaptarse al comportamiento de los datos según el valor que tomen un conjunto de parámetros θ:

$$\hat{f}(\boldsymbol{x}|H_j) = f(\boldsymbol{x}|\theta, H_j) \tag{2.9}$$

Estas técnicas tratan de estimar el valor de los parámetros θ de forma que el modelo estadístico se ajuste lo máximo posible al comportamiento que presentan los datos de entrenamiento.

El objetivo de la estimación no paramétrica es, utilizando únicamente la información proporcionada por el conjunto de entrenamiento, estimar la *PDF* sin presuponer ningún modelo concreto:

$$\boldsymbol{X} \implies \hat{f}(\boldsymbol{x}|h_j) \tag{2.10}$$

Estimación de funciones de densidad en problemas de detección

Es muy común la asunción de independencia entre las variables aleatorias para simplificar la estimación de las *PDFs*, pero como ya se argumentó y demostró con un sencillo ejemplo en el capítulo anterior, no contemplar la dependencia estadística presente entre las variables aleatorias puede conllevar una degradación en las prestaciones de detección. Otra posibilidad empleada es la utilización de técnicas *ICA* ("*Independent Component Analysis*") para transformar los datos, situándolos en un nuevo espacio donde sí son independientes.

Un modelo estadístico paramétrico muy usado en la literatura es el modelo de *PDF* multivariante Gaussiano. La utilización de este modelo impone dos limitaciones muy importantes: la asunción de normalidad conjunta limita a las marginales a seguir distribuciones Gaussianas y la matriz de covarianza usada en este modelo sólo puede caracterizar relaciones lineares entre variables. La utilización de este modelo para la integración de fuentes de información heterogéneas y cuya estructura de dependencia puede ser muy compleja, a priori, no es muy adecuada.

Podemos encontrar diversas técnicas más complejas para la estimación de estas *PDFs* conjuntas de una forma más fiel a la realidad, adaptándose tanto a las funciones marginales de cada variable por separado, como a la estructura de dependencia presente entre ellas. Es habitual en casos complejos el utilizar técnicas de estimación de densidades multivariantes no paramétricas, como son los histogramas multidimensionales, la estimación mediante vecinos próximos o las técnicas basadas en funciones núcleo o kernels. En cuanto a técnicas paramétricas no existen demasiadas en la literatura, siendo la mezcla de densidades gaussianas la más popular y ampliamente utilizada.

Recientemente se ha empezado a usar la teoría de cópulas, la cuál ha sido ampliamente utilizada en el campo de las finanzas y la econometría, pero su uso en problemas de procesado de señal es incipiente. Los modelos paramétricos para la

estimación de densidades que se derivan de la teoría de cópulas son idóneos para utilizarlos en problemas de detección, en los que se requiere de una fusión de datos dependientes y heterogéneos, ya que permiten aislar la información marginal de la estructura de dependencia, proporcionando una mayor flexibilidad, sencillez y precisión a la hora de modelar las *PDFs* conjuntas.

2.4.1. – Asunción de independencia

En el caso de vectores heterogéneos, donde se dispone de un conjunto de funciones marginales arbitrarias, es muy habitual el asumir independencia entre las variables aleatorias, de forma que la *PDF* conjunta se modela de una forma sencilla como producto de sus *PDFs* marginales:

$$f_p(\boldsymbol{x}|H_\mathrm{j}) = \prod_{i=1}^{d} f(x_i|H_\mathrm{j}) \tag{2.11}$$

Este modelo $f_p(\cdot)$ se conoce como el modelo del producto o de independencia. Como es evidente, este modelo nos permite incluir diferentes tipos de funciones marginales, pero no tiene en cuenta ningún comportamiento estadístico de dependencia entre las variables aleatorias implicadas.

Mediante este modelo se pueden estimar las funciones marginales de forma aislada usando las técnicas paramétricas o no paramétricas unidimensionales tradicionales, ampliamente conocidas y estudiadas en la literatura. En [27] se puede encontrar un amplio resumen de estas técnicas.

Así, estimando mediante este modelo las *PDFs* conjuntas bajo cada una de las hipótesis, el detector basado en la *LR* que integra o fusiona todas las fuentes de información posee la siguiente expresión:

$$\Lambda(\boldsymbol{x}) = \frac{\prod_{i=1}^{d} f(x_i|H_1)}{\prod_{i=1}^{d} f(x_i|H_0)} = \prod_{i=1}^{d} \Lambda_i(x_i) \lessgtr_{H_1}^{H_0} \lambda \tag{2.12}$$

Observamos en la expresión 2.12 cómo, asumiendo independencia bajo ambas hipótesis, podemos obtener una regla simple para la fusión de detectores basada en la multiplicación de los estadísticos obtenidos mediante la relación de verosimilitud. También es equivalente a la suma de las relaciones de verosimilitud logarítmicas:

$$LLR(\boldsymbol{x}) = ln\big(\Lambda(\boldsymbol{x})\big) = \sum_{i=1}^{d} ln\big(\Lambda_i(x_i)\big) \lessgtr_{H_1}^{H_0} ln(\lambda) \tag{2.13}$$

La asunción de independencia es muy habitual en multitud de trabajos para simplificar la obtención de las *PDFs*, pero como ya se ha comentado, en el caso de que

exista dependencia estadística puede acarrear severas degradaciones en las prestaciones de detección.

2.4.2. – Análisis de componentes independientes

En el caso de que las variables aleatorias no sean independientes entre sí bajo una o ambas hipótesis, la utilización del modelo del producto puede conducir a una degradación en las prestaciones globales de detección que se podrían obtener con una fusión óptima. Se podría obtener una mejora mediante la trasformación de los datos, de forma que se trasladen a un nuevo espacio donde sí se cumpla la independencia entre variables.

El análisis de componentes independientes (*ICA*, "*Independent Component Analysis*") es un método de procesado cuyo objetivo es encontrar una representación linear de datos multivariantes no Gaussianos, de forma que se puedan expresar en función de unas componentes que sean independientes (o lo más independientes posible).

Se considera que cada una de las variables aleatorias que modelan a los datos $X_1, \dots, X_d$ es el resultado de una mezcla lineal de *n* componentes, representadas por las variables aleatorias $S_1, \dots, S_n$. Se conoce como modelo estadístico de variables latentes, y está representado por:

$$x_j = a_{j1} \cdot s_1 + a_{j2} \cdot s_2 + \cdots + a_{jn} \cdot s_n \tag{2.14}$$

Se suele asumir que tanto los datos de entrada como las componentes independientes poseen media cero. Si no es así, siempre se puede centrar las variables x_j restándoles su media. Es conveniente usar una notación vectorial o matricial (2.15) en vez de las sumas de la ecuación (2.14). Denotamos como $\boldsymbol{x}$ al vector aleatorio *d x 1* cuyos elementos son las variables mezclas x_j y como $\boldsymbol{s}$ al vector aleatorio *n x 1* de componentes independientes. La matriz de mezclas $\boldsymbol{A}$ $(d\ x\ n)$ contiene los elementos a_{ji}. El vector $\boldsymbol{a}_i$ hace referencia a cada una de las columnas de la matriz de mezclas.

$$\boldsymbol{x} = \boldsymbol{A} \cdot \boldsymbol{s} = \sum_{i=1}^{n} \boldsymbol{a}_i \cdot s_i \tag{2.15}$$

El modelo *ICA* representado en (2.15) es un modelo generativo, describiendo cómo los datos de las observaciones son generados por un proceso de mezclado de componentes independientes s_i. Se dice que las componentes s_i son variables latentes porque no pueden ser directamente observadas de los datos, sino que se encuentran formando parte de ellos de forma implícita. En el apéndice C se analizan varios principios utilizados por diversos métodos para la estimación del modelo *ICA.*

Fusión de datos en detección usando ICA

Consideramos que disponemos de un par de conjuntos de entrenamiento $\boldsymbol{\mathcal{X}}_j,\ j = 0{,}1$, compuestos por N_j vectores de muestras de dimensión d, habiendo sido obtenido cada uno bajo una de las diferentes hipótesis $H_j,\ j = 0{,}1$:

$$\boldsymbol{\mathcal{X}}_j = \{\boldsymbol{x}^1, \dots, \boldsymbol{x}^{N_j}\},\ \ \boldsymbol{x}^t = [x_1^t, \dots, x_d^t]^{\mathrm{T}} \in\ H_j\ , t = 1, \dots, N_j\ \ j = 0{,}1 \tag{2.16}$$

Así, podemos entrenar dos modelos *ICA* para obtener dos vectores de componentes independientes $\boldsymbol{s}^{H_0}$ y $\boldsymbol{s}^{H_1}$, cada uno asociado a una de las hipótesis.

$$\begin{aligned} \boldsymbol{x} \in H_0 \rightarrow Modelo\ ICA\colon \boldsymbol{x} = \boldsymbol{A_0} \cdot \boldsymbol{s}^{H_0}\ \ \rightarrow \boldsymbol{s}^{H_0} = \boldsymbol{A_0^{-1}} \cdot \boldsymbol{x} = \boldsymbol{W_0} \cdot \boldsymbol{x} \\ \boldsymbol{x} \in H_1 \rightarrow Modelo\ ICA\colon \boldsymbol{x} = \boldsymbol{A_1} \cdot \boldsymbol{s}^{H_1}\ \ \rightarrow \boldsymbol{s}^{H_1} = \boldsymbol{A_1^{-1}} \cdot \boldsymbol{x} = \boldsymbol{W_1} \cdot \boldsymbol{x} \end{aligned} \tag{2.17}$$

Cada vector $\boldsymbol{s}^{H_j}$ estará compuesto por componentes independientes entre sí, con lo cual podremos estimar sus marginales $f\left(s_i|H_j\right)$ de forma aislada y expresar su *PDF* conjunta $f(\boldsymbol{s}^{H_j})$ como producto de ellas.

$$f_{\boldsymbol{s}}\left(\boldsymbol{s}|H_j\right) = f(\boldsymbol{s}^{H_j}) = \prod_{i=1}^{n} f\left(s_i|H_j\right) \rightarrow\ \Lambda(\boldsymbol{x}) = \frac{f_{\boldsymbol{s}}(\boldsymbol{W_1 x}|H_1)}{f_{\boldsymbol{s}}(\boldsymbol{W_0 x}|H_0)} \lessgtr_{H_1}^{H_0}\ \lambda \tag{2.18}$$

2.4.3. – Estimación no paramétrica mediante histogramas multidimensionales

Asumiendo que $f(\boldsymbol{x})$ es continua y que consideramos una región espacial $\Re$ con volumen V lo suficientemente pequeño de manera que $f(\boldsymbol{x})$ no varía significantemente en él, podemos estimar la densidad en los puntos $\boldsymbol{x} \in \Re$ conociendo cuantos, de los N puntos que posee el conjunto de entrenamiento, caen en la región $\Re$ y que denotamos por k:

$$\int_{x' \in \Re} f(\boldsymbol{x}')d\,\boldsymbol{x}' \approx f(\boldsymbol{x})V \rightarrow\ \ \hat{f}(\boldsymbol{x}) = \frac{k/N}{V} \tag{2.19}$$

Por lo tanto, un método muy simple de estimación consiste en realizar un histograma multidimensional, dividiendo el espacio de observaciones en pequeñas regiones, generalmente dividiendo cada dimensión i en segmentos de igual longitud l_i (formando zonas hipercúbicas si en todas las dimensiones se utiliza la misma longitud $l_i = l$ o prismas multidimensionales si en cada dimensión esta longitud es diferente) y estimando la densidad mediante la expresión (2.19). En la figura 2.3 podemos ver un ejemplo de un histograma bidimensional, donde la región de observación se ha dividido en zonas cuadradas mediante una rejilla regular.

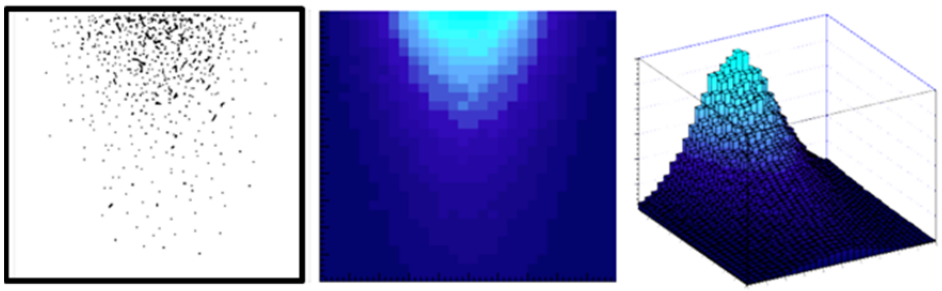


Figura 2.3 –Estimación de densidades de probabilidad mediantes histogramas multidimensionales

Aunque esta técnica de estimación es muy fácil de implementar, no suele ser práctica cuando disponemos de espacios de grandes dimensiones, ya que el número de regiones crece exponencialmente.

El tamaño de las celdas o regiones juega un papel muy importante en la estimación que se logra. Un tamaño de regiones muy pequeño puede conllevar un sobreajuste, tendiendo a proporcionar una estimación de la densidad muy escarpada, con picos centrados en cada una de las muestras del set de entrenamiento. Un tamaño de regiones muy grande produce un excesivo suavizado de la estimación. Existen técnicas de clasificación, como por ejemplo los árboles de decisión, que tratan de lograr una división del espacio en regiones de diferente tamaño, para lograr un histograma que se adapte mejor a las zonas de diferente densidad, pero suelen incrementar mucho la complejidad y requisitos computacionales de la técnica de estimación.

2.4.4. – Estimación no paramétrica de densidades mediante los k vecinos más próximos

Atendiendo a la expresión (2.19), podemos estimar la densidad fijando el volumen V y contando el número de puntos k que caen dentro de ese volumen. Es el proceso que, de forma general, se usa con la estimación mediante histograma. Otra posibilidad consiste en fijar el número de puntos k y determinar los volúmenes mínimos V_i que recogen esa cantidad de puntos en el set de entrenamiento. Esta técnica se usa en el método de estimación de densidades mediante los k vecinos más próximos [28] (*k-NN*, del inglés "*k-Nearest Neighbors*").

En el método *k-NN,* dado un punto $\boldsymbol{x}$, se aumenta progresivamente un volumen, generalmente esférico, a su alrededor hasta que abarque un total de k puntos del conjunto de entrenamiento. Así, la estimación de la densidad se convierte en:

$$\hat{f}(\boldsymbol{x}) = \frac{k/N}{V} = \frac{k}{N \cdot c_d \cdot R_k^d(\boldsymbol{x})} \tag{2.20}$$

donde $R_k^d(\boldsymbol{x})$ hace referencia a la distancia entre el punto $\boldsymbol{x}$ y su *k*-ésimo vecino más cercano en el espacio de dimensión d donde se sitúan los datos. c_d es el volumen de la esfera unidad en un espacio *d*-dimensional:

$$c_d = \begin{cases} \dfrac{\pi^{d/2}}{(d/2)!} & ,si\ d\ es\ par \\ \dfrac{\pi^{(d-1)/2}}{(d/2)!} \cdot 2^n \cdot \left(\dfrac{d-1}{2}\right)! & ,si\ d\ es\ impar \end{cases} \tag{2.21}$$

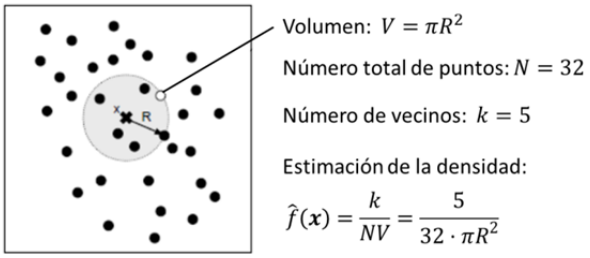


Figura 2.4 – Ejemplo de estimación de densidad mediante k-NN

En la figura 2.4 se puede observar un ejemplo de la estimación de la densidad de probabilidad multivariante asociada a una nueva observación $\boldsymbol{x}$ utilizando el método *k-NN* con 5 vecinos y un set de entrenamiento compuesto por $N = 32$ puntos.

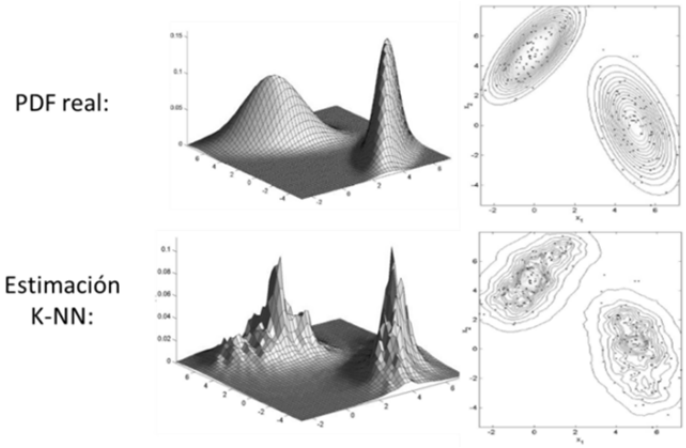


Figura 2.5 – Estimación de densidad multivariante mediante k-NN

En general, el método *k-NN* no obtiene muy buenas estimaciones de las densidades. Es propenso a la inclusión de ruido local produciendo una estimación muy escarpada y suele producir densidades con excesivas colas. Además, la densidad resultante no puede considerarse como una verdadera densidad ya que su integral a lo largo del espacio de observaciones diverge a consecuencia de las discontinuidades

que presenta (debido a que la función $R_k^d(\boldsymbol{x})$ no es diferenciable). En la figura 2.5 se puede observar un ejemplo de estimación de una función densidad compuesta por dos componentes gaussianas bivariantes mediante *k-NN*, con $N = 200$ y $k = 10$. Se aprecia claramente como la estimación es muy escarpada con grandes colas (se observan con gran claridad en las curvas de contorno).

Suele ser más habitual utilizar este método para implementar de forma directa una aproximación del detector, sin estimar de forma separada las densidades bajo cada hipótesis (figura 2.6).

Asumiendo un conjunto de entrenamiento compuesto por $N = N_0 + N_1$ puntos, de los cuales N_0 pertenecen a la clase H_0 y N_1 pertenecen a la clase H_1 y queriendo determinar la hipótesis real bajo la cual se generó una nueva muestra $\boldsymbol{x}_u$, se fija un volumen V alrededor de esta muestra y se contabilizan cuantos de los k puntos del set de entrenamiento que contiene el volumen pertenecen a cada clase ($k_1 \in H_1,\ k_0 \in H_0$). Así, se estima la probabilidad de que la muestra pertenezca a la clase H_j usando el teorema de Bayes como:

$$P(H_j|\boldsymbol{x}_u) = \frac{f(\boldsymbol{x}_u|H_j)}{f(\boldsymbol{x}_u)} P(H_j) = \frac{k_j}{k} \tag{2.22}$$

donde $f(\boldsymbol{x}_u) = \frac{k}{NV}$ es la estimación incondicional de la densidad, $f(\boldsymbol{x}_u|H_j) = \frac{k_j}{N_j V}$ es la estimación de la densidad condicionada a la hipótesis H_j y $P(H_j) = \frac{N_i}{N}$ es la estimación de probabilidad a priori de la hipótesis H_j.

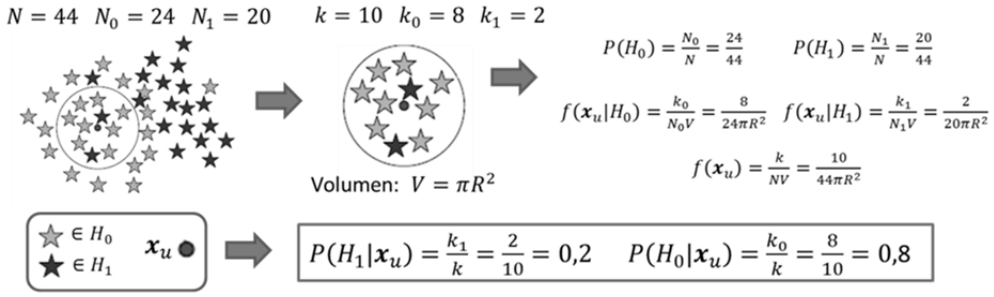


Figura 2.6 – Detección mediante k-NN

El método *k-NN*, tanto usado para la estimación de densidades, como para la implementación directa de un detector que aproxima el test de verosimilitud óptimo, es un método fácilmente tratable analíticamente y que, al usar información local para la estimación, posee un comportamiento altamente adaptativo. Permite una separación entre hipótesis muy flexible, pudiendo ajustarse a zonas complejas. Además, debido a su estructura y naturaleza se puede implementar de forma paralela.

Por el contrario, la búsqueda de vecinos es una tarea compleja que puede requerir gran capacidad de cálculo. También es necesaria una gran capacidad de almacenamiento de información para poder obtener la estima de la densidad o generar un score ante una nueva muestra, ya que requiere del almacenaje de todos los datos de entrenamiento y debe realizar la búsqueda de los puntos más cercanos. Es un método bastante susceptible a la 'maldición de la dimensión' ("*curse of dimensionality*"), también conocida como efecto Hughes, por lo que su uso en espacios de altas dimensiones no es muy recomendable. Además, es muy sensible a datos ruidosos o atípicos ("*outliers*").

2.4.5. – Estimación no paramétrica mediante funciones de núcleo o kernel

El método no paramétrico más ampliamente analizado y utilizado en la literatura para el caso multivariante es el de estimación de densidades mediante funciones de núcleo o Kernel [29], [27], [30] (*KDE*, "*Kernel Density Estimation*"). Se basa en estimar la *PDF* multivariante como una media de otras *PDF* multivariantes llamadas núcleos o kernels $K(\cdot)$ centradas en cada uno de los puntos del espacio $\mathbb{R}^d$ en los que se sitúan los vectores de muestras conocidos $\boldsymbol{x^i}$. Se trata a cada una de las observaciones como una componente diferente de un modelo de mezclas de densidades. La función núcleo $K(\cdot)$ es generalmente una función de densidad multivariante:

$$\int_{\mathbb{R}^d} K(\boldsymbol{x})\, d\boldsymbol{x} = 1, \qquad K(\boldsymbol{x}) \geq 0 \tag{2.23}$$

La expresión general de un estimador KDE es:

$$\hat{f}(\boldsymbol{x}|H_j) = \frac{1}{n}\sum_{i=1}^{n} K_{\boldsymbol{H}}(\boldsymbol{x} - \boldsymbol{x^i}) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{|\boldsymbol{H}|} K\left(\boldsymbol{H}^{-1}(\boldsymbol{x} - \boldsymbol{x^i})\right), \quad \boldsymbol{x^i} \in \boldsymbol{\mathcal{X}}, H_j \tag{2.24}$$

donde $\boldsymbol{H}$ es una matriz simétrica y definida positiva de orden *dxd* siendo denominada como matriz de ancho de banda. Esta matriz controla tanto la extensión como la orientación de la función núcleo, y por consiguiente del "suavizado" obtenido en la estimación. La matriz de ancho de banda $\boldsymbol{H}$ es el principal factor que afecta a la precisión de la estimación.

Observamos como la función estimada mediante esta técnica no paramétrica KDE es también una función de densidad de probabilidad:

$$\int \hat{f}(\boldsymbol{x}|H_j) d\boldsymbol{x} = 1, \qquad \hat{f}(\boldsymbol{x}|H_j) \geq 0 \tag{2.25}$$

Mediante el uso de la matriz de ancho de banda se puede caracterizar dos casos especiales más simples del estimador *KDE*. Un primer caso más simple es el uso del mismo ancho de banda h para todas las componentes; así la matriz de ancho de

banda se puede expresar como $\boldsymbol{H} = h\boldsymbol{I}_d$, siendo la matriz identidad *dxd.* Otra posibilidad es la de usar un ancho de banda $h_j, j = 1, \dots d$ diferente para cada componente, definiéndose la matriz de ancho de banda como $\boldsymbol{H} = diag(h_1, \dots, h_d)$:

$$\boldsymbol{H} = h\boldsymbol{I}_d \;\rightarrow\; \hat{f}(\boldsymbol{x}|H_j) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^d}K\left(\frac{\boldsymbol{x}-\boldsymbol{x^i}}{h}\right),\;\; \boldsymbol{x^i} \epsilon \boldsymbol{\mathcal{X}}, h_j \tag{2.26}$$

$$\boldsymbol{H} = diag(h_1, \dots, h_d) \;\rightarrow\; \hat{f}(\boldsymbol{x}|H_j) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_1 \dots h_d}K\left(\frac{x_1 - x_1^i}{h_1}, \dots, \frac{x_d - x_d^i}{h_d}\right),\;\; \boldsymbol{x^i} \epsilon \boldsymbol{\mathcal{X}}, H_j \tag{2.27}$$

Otra posibilidad que se puede encontrar es la utilización de un kernel multivariante basado en la multiplicación de kernel univariantes independientes en cada dimensión:

$$\hat{f}(\boldsymbol{x}|H_j) = \frac{1}{n}\sum_{i=1}^{n}\left\{\prod_{j=1}^{d}\frac{1}{h_j}K_j\left(\frac{x_j - x_j^i}{h_j}\right)\right\},\;\; \boldsymbol{x^i} \epsilon \boldsymbol{\mathcal{X}}, H_j \tag{2.28}$$

Uno de los kernel más utilizado es el kernel multivariante Gaussiano $K(\cdot) \sim N(0, \boldsymbol{R}_\sigma)$, donde el papel de la matriz de anchos de banda lo realiza la matriz de correlación $\boldsymbol{R}_\sigma$:

$$\hat{f}(\boldsymbol{x}|H_j) = \frac{1}{n(2\pi)^{d/2}|\boldsymbol{R}_\sigma|^{1/2}}\sum_{i=1}^{n} exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{x^i})^T {\boldsymbol{R}_\sigma}^{-1}(\boldsymbol{x}-\boldsymbol{x^i})\right) \tag{2.29}$$

En la figura 2.7 podemos ver un ejemplo de estimación de densidades mediante kernels Gaussianos. A la izquierda se representa diferentes núcleos dependiendo de la matriz de covarianza $\boldsymbol{R}_\sigma$ tomada (el primer caso con una matriz identidad multiplicada por un escalar positivo, el segundo con una matriz diagonal con los elementos de la diagonal principal positivos y el tercero con una matriz simétrica definida positiva). A la derecha podemos observar el proceso de estimación de la densidad.

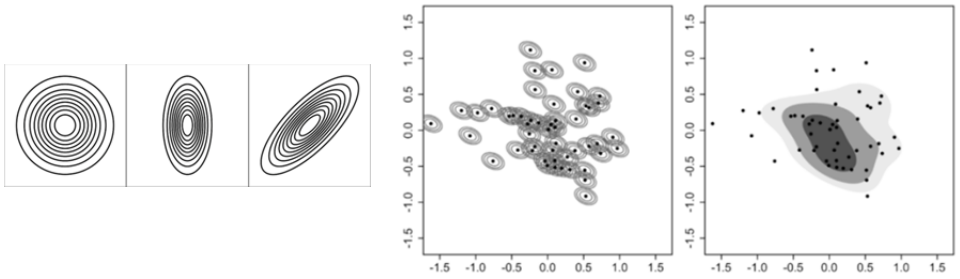


Figura 2.7 – Estimación de densidades mediante funciones de núcleo Gaussianas[31]

Estimación eficiente de la matriz de ancho de banda

Muchos investigadores han llegado a la conclusión de que el modelo que usa el mismo ancho de banda h para todas las componentes es menos eficaz que el modelo general usando anchos de banda diferentes para cada componente. Existen multitud de estudios de como estimar de forma eficiente la matriz de ancho de banda de cada componente (sean [32]–[36] una pequeña muestra de todos los que podemos encontrar).

El criterio óptimo más usado en muchos estudios para la elección de una matriz de ancho de banda es la minimización del error cuadrático integrado medio (*MISE*, "*Mean Integrated Squared Error*"):

$$MISE(\boldsymbol{H}) = E\left[\int \left(\hat{f}\left(\boldsymbol{x}|H_j\right) - f\left(\boldsymbol{x}|H_j\right)\right)^2 d\boldsymbol{x}\right] \tag{2.30}$$

Como podemos observar en el cálculo del *MISE* se supone conocida la función real que pretendemos estimar $f\left(\boldsymbol{x}|h_j\right)$. Es muy usual utilizar una aproximación asintótica de esta función (*AMISE*), categorizada como una buena aproximación cuando se dispone de suficientes vectores de muestras ($n \to \infty$) [31]:

$$AMISE(\boldsymbol{H}) = n^{-1}|\boldsymbol{H}|^{-\mathbf{1/2}}R(K) + \frac{1}{4}m_2(K)^2(vec^T\boldsymbol{H})\boldsymbol{\Psi}_4(vec^T\boldsymbol{H}) \tag{2.31}$$

donde,

$$\begin{gathered} R(K) = \int_{\Re^d} K(\boldsymbol{x})^2\, d\boldsymbol{x} \quad \int \boldsymbol{x}\boldsymbol{x}^T K(\boldsymbol{x})^2\, d\boldsymbol{x} = m_2(K)\boldsymbol{I}_d \\ \boldsymbol{\Psi}_4 = \int \left(vec\, D^2 f(\boldsymbol{x})\right)\left(vec^{\,T} D^2 f(\boldsymbol{x})\right) d\boldsymbol{x} \end{gathered} \tag{2.32}$$

siendo $\boldsymbol{I}_d$ la matriz identidad *dxd,* $D^2 f(\cdot)$ es la matriz Hessiana *dxd* de las derivadas parciales de segundo orden de $f(\cdot)$, $\boldsymbol{\Psi}_4$ es una matriz d^2xd^2, y $vec(\cdot)$ se define como un operador vectorial que distribuye las columnas de una matriz en un único vector, por ejemplo:

$$vec\left(\begin{bmatrix} a & c \\ b & d \end{bmatrix}\right) = [a\ b\ c\ d]^T \tag{2.33}$$

Así, muchos de los estudios para la estimación de la matriz de anchos de banda parten de la expresión general del *AMISE* como selector óptimo, la cual no puede ser usada directamente, ya que aún es función de la densidad desconocida $f(\boldsymbol{x})$. En la literatura se usan distintas técnicas para obtener diferentes estimadores del *AMISE,* dando lugar a diferentes métodos de estimación de la matriz de anchos de banda.

Siguiendo la intuición, no es difícil entender que puede ser beneficioso disponer de anchos de banda pequeños en regiones del espacio de características que estén muy compactas y pobladas, mientras que usar amplios ancho de banda será apropiado en

regiones cuya población esté más dispersa. Se han propuesto variantes de estimadores no estacionarios, los cuales tratan de utilizar matrices de ancho de banda diferentes según la región [36]–[38].

Uno de los principales problemas de la utilización de *KDE* es que su complejidad, relacionada con el número de componentes, se incrementa de forma lineal con el número de observaciones de las que se dispone. Para mitigar este incremento se han propuesto algunos métodos para comprimir o reducir el número de componentes, ya sea optimizando algún criterio relativo a los datos [39], [40] o para ajustando el número a un cierto valor pre-establecido [41], [42].

En [30] se propone una variante basada en *KDE* para la estimación de funciones de densidad de probabilidad de forma continua. Este método se utiliza para mantener y actualizar un modelo *KDE* de forma continua, realizando una reestimación de la matriz de ancho de banda y utilizando un esquema de compresión de forma que se mantenga la complejidad baja.

2.4.6. – Modelo de mezclas Gaussianas

Un modelo de mezcla Gaussiana (*GMM,* "*Gaussian Mixture Model"*) es una función de densidad de probabilidad paramétrica representada por una suma ponderada de componentes con densidades de probabilidad Gaussianas multivariantes.

Para un conjunto de d variables aleatorias continuas $\boldsymbol{X} = [X_1, \dots, X_d]$ modelamos su función de densidad de probabilidad conjunta mediante un *GMM* con K componentes como:

$$f(x_1, \dots, x_d) \sim f_{GMM}(\boldsymbol{x}|\boldsymbol{\theta_1} \dots \boldsymbol{\theta_K}, \boldsymbol{\omega}) = \sum_{i=1}^{K} \omega_i \cdot g(\boldsymbol{x}|\boldsymbol{\theta_i}) = \sum_{i=1}^{K} \omega_i \cdot g(\boldsymbol{x}|\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}) \qquad (2.34)$$

donde ω_i, corresponden a los pesos de la mezcla, los cuales cumplen con la restricción $\sum_{i=1}^{K} \omega_i = 1$, y $g(\boldsymbol{x}|\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}), i = 1, .., K$ son las componentes de densidad de probabilidad *d*-dimensional Gaussiana, con media $\boldsymbol{\mu_i}$ y matriz de covarianza $\boldsymbol{\Sigma_i}$:

$$g(\boldsymbol{x}|\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}) = \frac{1}{(2\pi)^{N/2} \cdot |\boldsymbol{\Sigma_i}|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu_i})^T \boldsymbol{\Sigma_i}^{-1} (\boldsymbol{x}-\boldsymbol{\mu_i})} \qquad (2.35)$$

El modelo *GMM* es parametrizado completamente por el conjunto de parámetros $\vartheta = \{\omega_i, \boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}\},\ i = 1, \dots, K$. Así, definimos el modelo *GMM* como:

$$f(x_1, \dots, x_d) \sim f_{GMM}(\boldsymbol{x}|\vartheta) \qquad (2.36)$$

Se pueden utilizar diferentes variantes de este modelo. Las matrices de covarianza, en vez de utilizarlas completas (*full rank*) se pueden restringir a ser diagonales. Pueden predefinirse algunos parámetros, como por ejemplo utilizar la misma ponderación para todas las componentes, o definir parámetros compartidos, como por ejemplo utilizar una matriz de covarianza común para todas las componentes. El

objeto de estas simplificaciones en el modelo suele ser el aliviar la carga computacional requerida para estimar sus parámetros.

Es importante resaltar que, dado que las componentes Gaussianas están actuando de forma conjunta para modelar la densidad de probabilidad global, las matrices de covarianza completas pueden no ser necesarias aunque no exista independencia entre los distintos elementos de las observaciones, ya que, al combinar de forma lineal componentes Gaussianas con matrices de covarianza diagonales, se es capaz de modelar correlaciones arbitrarias entre los diferentes elementos del vector de observaciones. En el apéndice D se habla del método usado por excelencia para la estimación de los parámetros del modelo *GMM*, el cual está basado en el algoritmo *EM*. También se habla del método más popular de selección del número de componentes Gaussianas basado en criterios de mínima longitud. En [43] (*http://www.lx.it.pt/~mtf/mixturecode2.zip*) podemos encontrar un algoritmo de entrenamiento no supervisado del modelo *GMM* implementado en *MATLAB* con las técnicas expuestas.

2.5. – Revisión de la teoría de cópulas

2.5.1. – Introducción

El término “cópula” fue usado por primera vez en el trabajo de Sklar [44], derivado de la palabra en latín “copulae”, cuyo significado es conectar o unir. Este nombre se debe a que las cópulas se pueden entender como funciones que unen o conectan funciones de probabilidad multidimensionales con sus funciones marginales de menor orden. Las cópulas, en términos probabilísticos, representan funciones de distribución de probabilidad (*CDF*) multivariantes, cuyas funciones univariantes marginales son uniformes en el intervalo $[0,1]$.

Las funciones de cópula pueden ser usadas para describir la dependencia entre variables aleatorias. La función de distribución de probabilidad de un vector aleatorio puede ser escrita en términos de las funciones de distribución marginales de cada componente y una función de cópula. Las funciones de distribución marginales describen el comportamiento estadístico de cada componente del vector por separado y la función de cópula describe la estructura de dependencia existente entre esas componentes.

Por lo tanto, las cópulas son herramientas que nos permiten modelar dependencia entre diferentes variables aleatorias. Nos permitirán modelar y estimar de forma simple la distribución de probabilidad de un vector de variables aleatorias heterogéneas y dependientes.

Existen diversos tipos y familias de cópulas parametrizables en la literatura, donde cada una de ellas puede caracterizar diferentes tipos de comportamiento de dependencia entre las variables aleatorias [45].

Las cópulas han sido usadas de forma extensiva dentro del ámbito de la econometría y las finanzas [46]–[49], sin embargo el uso de estas herramientas en aplicaciones y problemas relacionadas con el tratamiento de señal es incipiente [2], [18], [50].

Teorema de Sklar: Función de cópula

La función de distribución de probabilidad conjunta $F(z_1, z_2, \dots, z_d)$ de una serie de variables aleatorias continuas (*v.a*) $Z_1, Z_2, \dots, Z_d$ puede ser expresada en función de las distribuciones de probabilidad marginales asociadas a cada una de las variables aleatorias $F_i(z_i),\ i = 1, \dots, d$ según,

$$F(z_1, z_2, \dots, z_d) = C\big(F_1(z_1), F_2(z_2), \dots, F_d(z_d)\big) \tag{2.37}$$

La variable aleatoria U_i definida como $u_i = F_i(z_i)$, siendo F_i la función de distribución de probabilidad de otra variable aleatoria arbitraria Z_i posee una distribución uniforme.

La función C es la llamada como función de cópula, la cual no es más que una función de distribución de probabilidad conjunta de variables aleatorias uniformemente distribuidas en el intervalo $[0,1]$. Por lo tanto la función de cópula C se define en un hipercubo unitario,

$$C: U(0\ 1)^d \to (0\ 1) \tag{2.38}$$

Una demostración detallada del teorema de Sklar, junto con una serie de propiedades adicionales de las funciones de cópula pueden encontrarse en [46], [51].

Función densidad de probabilidad y relación con la función de cópula

Si la función de cópula C y las distribuciones marginales F_i son lo suficientemente diferenciables, entonces la función de densidad de probabilidad (*PDF*) conjunta se puede obtener como el producto de las funciones de densidad marginales f_i y la función densidad de cópula c derivada de función de cópula C. Para distribuciones continuas, la función de densidad de probabilidad conjunta se obtiene diferenciando ambos factores de la ecuación (2.37):

$$f(z_1, z_2, \dots, z_d) = f_1(z_1) \cdot f_2(z_2) \cdot \dots \cdot f_d(z_d) \cdot c\big(F_1(z_1), F_2(z_2), \dots, F_d(z_d)\big) \tag{2.39}$$

donde,

$$c(\boldsymbol{u}) = \frac{\partial^d \big(C(u_1, u_2, \dots, u_d)\big)}{\partial u_1 \partial u_2, \dots, \partial u_d}, \qquad u_i = F_i(z_i) \tag{2.40}$$

La función cópula de densidad $c(\cdot)$, en este caso, es una función de densidad de probabilidad de variables aleatorias uniformemente distribuidas. Podemos expresar la función de densidad de probabilidad multivariante separando la información que aporta cada una de las distribuciones marginales de la información de dependencia

existente entre ellas. A la distribución de densidad de probabilidad obtenida considerando independencia, o lo que es lo mismo, la obtenida mediante el producto de las distribuciones marginales, se le incorpora la información de dependencia mediante la multiplicación por la función de cópula.

$$f(z_1, z_2, \dots, z_d) = \underbrace{f_1(z_1) \cdot f_2(z_2) \cdot \cdots \cdot f_d(z_d)}_{\text{Distribuciones marginales}} \cdot \underbrace{c(F_1(z_1), F_2(z_2), \dots, F_d(z_d))}_{\text{Dependencia estadística}} \tag{2.41}$$

Si las variables aleatorias fueran independientes, la funcion densidad de probabilidad conjunta sería un producto de las funciones de densidad marginales:

$$f(z_1, \dots, z_d)_{Independientes} = \prod_{i=1}^{d} f_i(z_i) = f_p(\boldsymbol{z}) \tag{2.42}$$

A la hora de parametrizar una distribución de densidad conjunta podemos encontrar dos conjuntos de parámetros, los asociados a las distribuciones marginales y los asociados a la función de cópula, o lo que es lo mismo, los parámetros que modelan la dependencia:

$$f(z_1, z_2, \dots, z_d; \Phi) = \left(\prod_{i=1}^{d} f_i(z_i; \lambda_i) \right) \cdot c(F_1(z_1; \lambda_1), \dots, F_d(z_d; \lambda_d); \theta) \tag{2.43}$$

Parámetros de la distribución $\Phi = \{\lambda: marginales,\ \ \theta: dependencia\}$

Función densidad de cópula asociada a una determinada distribución

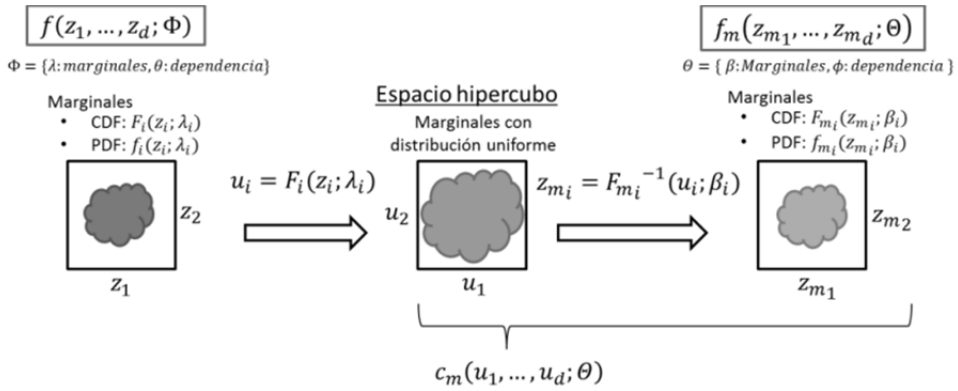


Figura 2.8 – Uso de la función densidad de cópula derivada de una PDF conocida para modelar otra PDF desconocida

Sea $f_m(z_{m_1}, \dots, z_{m_d}; \Theta)$ una expresión conocida de una determinada distribución conjunta de probabilidad multivariante, donde Θ son los parámetros con los que se define el modelo. Usando una función de cópula la podemos expresar como:

$$f_m(z_{m_1},\dots,z_{m_d};\Theta)=\left(\prod_{i=1}^{d} f_{m_i}(z_{m_i};\beta_i)\right)\cdot c_m\left(F_{m_1}(z_{m_1};\beta_1),\dots,F_{m_d}(z_{m_d};\beta_d);\phi\right) \tag{2.44}$$

Parámetros del modelo $\Theta=\{\beta: Marginales,\ \phi: dependencia\}$

La función de probabilidad multivariante queda parametrizada por el conjunto de parámetros Θ, los cuales podemos dividir en dos subconjuntos, uno asociado a las funciones marginales β y otro asociado a los parámetros de la cópula, y por tanto, los que modelan la estructura de dependencia ϕ. Podemos encontrar la expresión de la función de cópula asociada al modelo fácilmente sabiendo que:

$$z_{m_i}={F_{m_i}}^{-1}(u_i;\beta_i) \tag{2.45}$$

Obtenemos la siguiente expresión para la función de cópula derivada de la distribución multivariante $f_m(z_{m_1},\dots,z_{m_d};\Theta)$:

$$c_m(u_1,\dots,u_d;\Theta)=\frac{f_m\left({F_{m_1}}^{-1}(u_1;\beta_1),\dots,{F_{m_d}}^{-1}(u_d;\beta_d);\Theta\right)}{\left(\prod_{i=1}^{d} f_{m_i}\left({F_{m_i}}^{-1}(u_i;\beta_i);\beta_i\right)\right)} \tag{2.46}$$

Consideremos ahora un conjunto de variables aleatorias $Z_1,Z_2,\dots,Z_d$, con *PDF* conjunta $f(z_1,z_2,\dots,z_d)$ desconocida. Se pretende obtener una estimación de esta *PDF* conjunta usando la teoría de cópulas para ello. Asumimos conocidas las funciones de probabilidad marginales asociadas a cada una de las variables aleatorias, y que posee las mismas propiedades de dependencia que el modelo $f_m(z_{m_1},\dots,z_{m_d};\Theta)$ cuya función densidad de cópula es conocida $c_m(u_1,\dots,u_d;\Theta)$. Llamamos $\hat{f}(\boldsymbol{z};\widehat{\Phi})$ a la expresión de PDF estimada usando la función densidad de cópula $c_m(\cdot)$ asociada a otro modelo $f_m(\boldsymbol{z_m};\Theta)$, el cuál comparte las mismas características de dependencia (figura 2.8):

$$f(\boldsymbol{z};\Phi)\backsim\hat{f}(\boldsymbol{z};\widehat{\Phi})=\left(\prod_{i=1}^{d} f_i(z_i;\lambda_i)\right)\cdot c_m\left(F_1(z_1;\lambda_1),\dots,F_d(z_d;\lambda_d);\hat{\theta}\right) \tag{2.47}$$

$$\widehat{\Phi}=\left\{\lambda,\hat{\theta}=\Theta=\{\beta,\phi\}\right\}$$

Si observamos la figura 2.9 podemos contemplar un ejemplo en el que existen dos variables aleatorias Z_1,Z_2 cuyas marginales son conocidas y diferentes, existiendo una correlación lineal entre ellas. La PDF conjunta es desconocida. Si sólo consideramos la información marginal, es decir, las consideramos independientes observamos como la PDF obtenida no se adecua a la realidad. Utilizando una distribución Gaussiana bivariante, donde las variables poseen la misma estructura de dependencia lineal o correlación, podemos obtener la función densidad de cópula que contendrá la información referente a esta dependencia lineal.

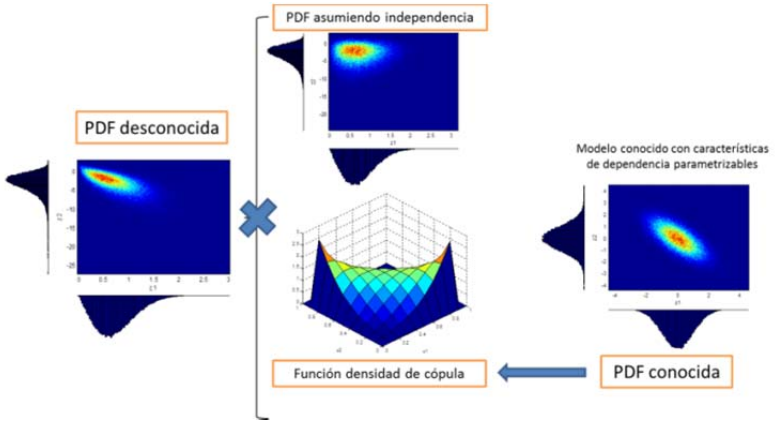


Figura 2.9 – Ejemplo de modelado del PDF usando una función densidad de cópula derivada de un modelo multivariante conocido.

2.5.2. – Uso de las funciones de cópula en teoría de detección

La regla de decisión óptima cuando tenemos "d" fuentes de información representadas por el conjunto de variables aleatorias continuas $Z_1, Z_2, \dots, Z_d$ pasa por un test basado en la relación de verosimilitud (*LR, "Likelihood Ratio"*):

$$LR = \Lambda_{Opt}(z_1, \dots, z_d) = \frac{f(z_1, \dots, z_d|H_1)}{f(z_1, \dots, z_d|H_0)} \lessgtr_{H_1}^{H_0} \eta \tag{2.48}$$

Podemos modelar cada una de esas PDF conjuntas mediante el uso de funciones de densidad de cópula:

$$LR = \Lambda_{\mathrm{Opt}}(\mathrm{z}_1, \dots, z_d) = \frac{f_p(\boldsymbol{z}|H_1) \cdot c(F_1(z_1|H_1), \dots, F_d(z_d|H_1)|H_1)}{f_p(\boldsymbol{z}|H_0) \cdot c(F_1(z_1|H_0), \dots, F_d(z_d|H_0)|H_0)} \lessgtr_{H_1}^{H_0} \eta \tag{2.49}$$

Tomando el logaritmo del test de verosimilitud podemos dividir el test en dos partes, una relacionada con la información marginal y otra relacionada con la información que nos reporta la dependencia entre las variables aleatorias implicadas:

$$LLR(\boldsymbol{z}) = ln\left(\frac{f_p(\boldsymbol{z}|H_1)}{f_p(\boldsymbol{z}|H_0)}\right) + ln\left(\frac{c_{H_1}\left(F_1(z_1|H_1), \dots, F_d(z_d|H_1)\right)}{c_{H_0}\left(F_1(z_1|H_0), \dots, F_d(z_d|H_0)\right)}\right) \lessgtr_{H_1}^{H_0} ln(\eta) \tag{2.50}$$

Es necesario conocer las funciones marginales y la función densidad de cópula bajo cada una de las hipótesis para poder realizar una fusión óptima de la información que nos reportan las variables aleatorias $Z_1, Z_2, \dots, Z_d$ en detección. Al no conocer la expresión de la distribución multivariante real que se adapta al conjunto $Z_1, Z_2, \dots, Z_d$ no podemos conocer la expresión de la densidad de cópula real $c(\cdot)$. Tendremos que escoger una función densidad de cópula $\hat{c}(\cdot)$, que codifique razonablemente la misma estructura de dependencia entre las variables aleatorias.

Por lo tanto, la resolución de un problema de detección usando la teoría de cópulas implica, por una parte, la selección u obtención de las PDF marginales de cada variable aleatoria bajo cada hipótesis, y por otra parte, la selección de una función densidad de cópula $\hat{c}(\cdot)$ para cada hipótesis y su correctamente parametrizada, de forma que se codifique la información de dependencia presente entre las variables aleatorias de forma más fidedigna a la realidad.

Técnicas de estimación de parámetros

Desde un punto de vista estadístico, una función de cópula simplemente es una expresión simple para un modelo multivariante, y como para la mayoría de modelos estadísticos multivariantes, no todos los métodos clásicos de inferencia estadística pueden ser aplicados. La técnica que más ha sido utilizada es la estimación por máxima verosimilitud (*MLE, "Máximum Likelihood Estimation"*). Existen otras técnicas de estimación basadas en una mezcla de conceptos de inferencia estadística no paramétrica y técnicas de simulación, que se han propuesto para aliviar la carga computacional que supone encontrar el estimador *MLE* óptimo.

Empezaremos describiendo el estimador de máxima verosimilitud *MLE* y veremos algunas alternativas a él. Los métodos de estimación que vamos a ver van a requerir de técnicas de optimización numéricas de una función objetivo, ya que una cópula es intrínsecamente un modelo multivariante y su función de verosimilitud implica derivadas parciales mixtas.

- *Estimación por máxima verosimilitud*

Recordemos la representación canónica de una función densidad de probabilidad multivariante expresada en términos de una función de cópula:

$$f(z_1, \dots, z_d; \Phi) = \left(\prod_{i=1}^{d} f_i(z_i; \lambda_i) \right) \cdot c(F_1(z_1; \lambda_1), \dots, F_d(z_d; \lambda_d); \theta)$$
$$c(\boldsymbol{u}) = \frac{\partial^d \left(C\left(F_1(z_1; \lambda_1), \dots, F_d(z_d; \lambda_d)\right)\right)}{\partial F_1(z_1; \lambda_1), \dots, \partial F_d(z_d; \lambda_d)}, \qquad u_i = F_i(z_i; \lambda_i) \tag{2.51}$$

Esta representación canónica para la función de densidad multivariante nos permite descomponer el problema de modelado estadístico mediante cópulas en dos pasos:

1. Identificación de las distribuciones marginales
2. Definición de la función de cópula más apropiada.

Sean *N* vectores de muestras $\boldsymbol{\mathcal{Z}} = [\boldsymbol{z_1}, \dots., \boldsymbol{z_N}]$, con $\boldsymbol{z_n} = \left[{z_1}^{(n)}, \dots, {z_d}^{(n)}\right]$, $n = 1 \dots N$. Una vez se han definido las distribuciones marginales y la función de cópula, la expresión de la función de verosimilitud logarítmica se puede expresar como:

$$log\big(L(\boldsymbol{\mathcal{Z}};\Phi)\big) = \sum_{n=1}^{N} log\left(c\big(F_1\big(z_1{}^{(n)};\lambda_1\big),\dots,F_d\big(z_d{}^{(n)};\lambda_d\big);\theta\big)\right) + \sum_{n=1}^{N}\sum_{i=1}^{d} log\left(f_i\big(z_i{}^{(n)};\lambda_i\big)\right) \tag{2.52}$$

donde, Φ es el conjunto de parámetros compuesto por los parámetros que definen a las funciones marginales $\{\lambda\}$ y los de la función de cópula $\{\theta\}$.

El estimador de máxima verosimilitud $\widehat{\Phi}_{MLE}$ del conjunto de parámetros Φ, viene determinado por:

$$\widehat{\Phi}_{MLE} = \underset{\Phi}{\operatorname{argmax}}\big(L(\boldsymbol{\mathcal{Z}};\Phi)\big) \tag{2.53}$$

Se asume que se cumplen las condiciones de regularidad de la teoría de máxima verosimilitud asintótica [52] tanto para la función densidad de cópula como para las densidades marginales. Bajo estas condiciones de regularidad, el estimador de máxima verosimilitud existe y es consistente y asintóticamente eficiente. También se puede demostrar que cumple la propiedad de ser asintóticamente normal [53] :

$$\sqrt{N}\big(\widehat{\Phi}_{MLE} - \Phi_0\big) \to \mathcal{N}\big(0, \Im^{-1}(\Phi_0)\big) \tag{2.54}$$

donde $\Im(\Phi_0)$ denota a la matriz de información de Fisher y Φ_0 denota al valor real óptimo.

- *Inferencia por funciones marginales*

El método de estimación por máxima verosimilitud puede requerir de mucha carga computacional en el caso de tener un gran número de variables aleatorias implicadas, ya que es necesario estimar de forma conjunta los parámetros de las distribuciones marginales y los parámetros de la estructura de dependencia representada por la función densidad de cópula.

Observamos que la función de verosimilitud logarítmica se puede descomponer en dos términos positivos, uno involucrando a la función densidad de cópula y sus parámetros, y otro involucrando a las densidades marginales con sus propios parámetros. Por esta razón, Joe y Hu [54] propusieron que estos conjuntos de parámetros podían ser estimados en dos pasos:

1. Se estiman los conjuntos de parámetros de cada una de las funciones marginales por separado:

$$\hat{\lambda} = \{\hat{\lambda}_1, \dots, \hat{\lambda}_d\}, \qquad \hat{\lambda}_i = \underset{\lambda_i}{\operatorname{argmax}} \left(\sum_{n=1}^{N} log\left(f_i\left({z_i}^{(n)}; \lambda_i \right) \right) \right), i = 1, .. d \tag{2.55}$$

2. Se realiza la estimación de los parámetros de la cópula $\hat{\theta}$, previo paso de uniformizar cada una de las componentes usando el conjunto estimado $\hat{\lambda}$:

$$\begin{aligned} {u_i}^{(n)} &= F_i\left({z_i}^{(n)}; \hat{\lambda}_i \right), \quad i = 1, \dots, d, \quad n = 1, \dots, N \\ \hat{\theta} &= \underset{\theta}{\operatorname{argmax}} \left(\sum_{n=1}^{N} log\left(c\left({u_1}^{(n)}, \dots, {u_d}^{(n)}; \theta \right) \right) \right) \end{aligned} \tag{2.56}$$

Este método es el llamado inferencia por funciones marginales (*IFM, "Inference Functions for Margins"*). El estimador *IFM* se define como:

$$\hat{\theta}_{IFM} = \left(\hat{\lambda}, \hat{\theta} \right) \tag{2.57}$$

Sea L la ecuación de la verosimilitud logarítmica global, L_i la verosimilitud logarítmica de la densidad marginal i-ésima y L_c la verosimilitud de la función densidad de cópula. El estimador *IFM* es la solución de:

$$\left(\frac{\partial L_1}{\partial \lambda_1}, \dots, \frac{\partial L_d}{\partial \lambda_d}, \ \frac{\partial L_c}{\partial \theta} \right) = \mathbf{0} \tag{2.58}$$

mientras que el estimador MLE se obtiene resolviendo:

$$\left(\frac{\partial L}{\partial \lambda_1}, \dots, \frac{\partial L}{\partial \lambda_d}, \ \frac{\partial L}{\partial \theta} \right) = \mathbf{0} \tag{2.59}$$

por lo tanto, la equivalencia entre estos dos estimadores por lo general no se cumple.

El estimador *IFM* es mucho más eficiente en lo que a tiempo de cómputo se refiere, pero por lo general es una estimación subóptima. Puede ser utilizado como método para encontrar un buen punto de comienzo para obtener el estimador óptimo *MLE*.

- *Máxima verosimilitud canónica*

Dado que los parámetros de la cópula pueden ser estimados sin necesidad de especificar las funciones marginales, otro método de estimación que podemos utilizar se basa en utilizar una estimación no paramétrica de las funciones de distribución marginales
con las que poder uniformizar las variables aleatorias. Este método se denomina como máxima verosimilitud canónica (*CML,* "*Canonical Maximum Likelihood*"):

1. Primero estimamos las funciones de distribución de forma no paramétrica y uniformizamos cada componente:

$$Estimamos\colon\ \hat{F}_i\left(z_i{}^{(n)}\right) \rightarrow Uniformizamos\colon\ u_i{}^{(n)} = \hat{F}_i\left(z_i{}^{(n)}\right) \tag{2.60}$$

2. Estimamos mediante *MLE* los parámetros de la cópula:

$$\hat{\theta}_{CMD} = \underset{\theta}{\operatorname{argmax}}\left(\sum_{n=1}^{N} log\left(c\left(\hat{F}_1\left(z_1{}^{(n)}\right), \dots, \hat{F}_d\left(z_d{}^{(n)}\right); \theta\right)\right)\right) \tag{2.61}$$

Técnicas de selección de cópulas

La definición de la función de densidad de cópula más apropiada puede no ser trivial, ya que en muchos casos desconocemos a priori las características de dependencia existentes entre las variables aleatorias. En este caso, se deben utilizar un conjunto de funciones de densidad de cópula diferentes y escoger de entre todas la más idónea. Introducimos algunas técnicas para seleccionar de forma automática la función densidad de cópula más apropiada.

- *Selección mediante teoría de longitud de descripción mínima*

Uno de los criterios de selección de modelo óptimo muy utilizado en teoría de información y teoría del aprendizaje es el criterio basado en la longitud de descripción mínima [55] (*MDL , "Minimum Description Length"*). Las técnicas de selección de modelo *MDL* se basan en el principio que argumenta que, de entre todos los modelos que tenemos como alternativas, el que alcanza la mayor compresión, es el modelo que mejor describe las observaciones.

Sean *N* vectores de muestras $\boldsymbol{\mathcal{Z}} = [\boldsymbol{z_1}, \dots., \boldsymbol{z_N}]$, con $\boldsymbol{z_n} = \left[z_1{}^{(n)}, \dots, z_d{}^{(n)}\right]$, $n = 1 \dots N$. Cada uno de estos vectores es considerado como una realización de un conjunto de *d* variables aleatorias. Sea $\varsigma = \{c_1(\cdot), \dots, c_K(\cdot)\}$, $k = 1, \dots, K$ el conjunto de funciones densidad de cópula, del cual deseamos escoger el modelo más adecuado. Definimos $L_k\left(\boldsymbol{\mathcal{Z}}; \widehat{\Phi}\right)$ como la función de verosimilitud evaluada utilizando el estimador de máxima verosimilitud de los parámetros de la función densidad de cópula $c_k(\cdot)$:

$$L_k\left(\boldsymbol{\mathcal{Z}}; \widehat{\Phi}\right) = \max_{\Phi}\left(\prod_{n=1}^{N} c_k\left(F_1\left(z_1{}^{(n)}; \lambda_1\right), \dots, F_d\left(z_d{}^{(n)}; \lambda_d\right); \Phi\right)\right) \tag{2.62}$$

Los criterios de selección basados en *MDL* calculan un índice relativo a las prestaciones del modelo en contraposición al incremento de complejidad que supone su uso. En la expresión (2.63), q_k hace referencia a un término de penalización

proporcional a la complejidad del modelo $c_k(\cdot)$, diferente según el criterio basado en *MDL* se utilice. Se escoge la función densidad de cópula que minimice:

$$c(\cdot) = c_k(\cdot) \leftrightarrow \min_k \left(I_k = -2 \cdot log\left(L_k\left(\boldsymbol{\mathcal{Z}}; \hat{\Phi}\right)\right) + q_k\right) \tag{2.63}$$

En el marco de selección *MDL* podemos encontrar distintos criterios de selección, como pueden ser *AIC* "*Akaike Information Criterion*", *BIC* "*Bayesian Information Criterion*", *SIC* "*Stochastic Information Complexity*" o *NML* "*Normalized Maximum Likelihood*".

$$\begin{aligned} AIC &= -2 \cdot log\left(L_k\left(\boldsymbol{\mathcal{Z}}; \hat{\Phi}\right)\right) + n_p \\ BIC &= -2 \cdot log\left(L_k\left(\boldsymbol{\mathcal{Z}}; \hat{\Phi}\right)\right) + n_p \cdot log(N) \\ SIC &= -2 \cdot log\left(L_k\left(\boldsymbol{\mathcal{Z}}; \hat{\Phi}\right)\right) + log\left(\left|\hat{\Sigma}\right|\right) \\ NML &= -2 \cdot log\left(L_k\left(\boldsymbol{\mathcal{Z}}; \hat{\Phi}\right)\right) + n_p \cdot log\left(\frac{N}{2\pi}\right) + log\left(\int \sqrt{|I(\phi)|} d\phi\right) \end{aligned} \tag{2.64}$$

Donde n_p hace referencia al número de parámetros dentro del conjunto Φ que son necesarios optimizar para encontrar el máximo de la función de verosimilitud, T es el número de vectores de observación utilizados en el entrenamiento del modelo, $\left|\hat{\Sigma}\right|$ denota el determinante de la matriz Hessiana de $log\left(L_k\left(\boldsymbol{\mathcal{Z}}; \hat{\Phi}\right)\right)$ y $|I(\phi)|$ es el determinante de la matriz de información de Fisher evaluada sobre $c_k\left(F_1\left({z_1}^{(n)}; \lambda_1\right), \ldots, F_d\left({z_d}^{(n)}; \lambda_d\right); \phi\right)$.

- *Selección mediante criterios de rendimiento en detección*

Las técnicas de selección anteriormente mencionadas, se centran en escoger las funciones de densidad de cópula de forma que la *PDF* conjunta bajo cada una de las hipótesis se aproxime lo máximo posible a la realidad. Dado que se propone utilizar la teoría de cópulas para la aplicación específica de fusión de información en problemas de detección, donde los resultados dependen de la umbralización del estadístico $\Lambda(\boldsymbol{z})$ obtenido del cociente de ambas *PDF* y por tanto, de un hiperplano que divide el espacio de las observaciones en dos regiones, es posible que aunque las estimaciones de las *PDF* no sean muy precisas, se puedan obtener buenos resultados en cuanto a prestaciones de detección se refiere. Así pues, se pueden utilizar reglas de selección de cópulas basadas en escoger aquellas con las que mejores prestaciones de detección se alcancen.

Por ejemplo, podemos utilizar el criterio de rendimiento del área bajo la curva ROC (AUC), o bajo una determinada porción de esta $\left(pAUC_\alpha^\beta\right)$, ya introducidos en el apartado 1.1:

$$AUC = \int_0^1 P_D(P_F)\, dP_F \qquad pAUC_\alpha^\beta = \int_\alpha^\beta P_D(P_F)\, dP_F \tag{2.65}$$

Podemos simplificar el proceso de selección de la pareja de funciones densidad de cópula que maximice el *AUC*, bajo la asunción de Gaussianidad de las densidades de probabilidad del test de hipótesis $\Lambda(\boldsymbol{z})$ bajo cada una de las hipótesis H_0 y H_1. Se puede demostrar que [2] Escogiendo las dos funciones densidad de cópula c_{H_1} y c_{H_0} que nos proporcionen mayor d_a, nos proporcionaran un valor más elevado de *AUC:*

$$\Lambda(\boldsymbol{z}|H_i) \sim \mathcal{N}\left(\mu_{H_i}, \sigma_{H_i}^2\right) \;\leftrightarrow\; AUC = \Phi^{-1}\left(\frac{d_a}{\sqrt{2}}\right),\;\; d_a = \frac{\mu_{H_1} - \mu_{H_0}}{\sqrt{\frac{\sigma_{H_1}^2 + \sigma_{H_0}^2}{2}}} \tag{2.66}$$

En el caso en que no se cumpla la asunción de Gaussianidad del test de hipótesis $\Lambda(\boldsymbol{z})$ bajo alguna, o ambas hipótesis, se puede usar el estimador Wilcoxon-Mann-Whitney (*WMW*), una forma no paramétrica de estimar el *AUC*. Sean $\left\{{l_1}^{(a)}\right\}_{a=1}^{N_a}$ y $\left\{{l_0}^{(b)}\right\}_{b=1}^{N_b}$ los valores del *LR* dado por (2.62) bajo H_1 y H_0 respectivamente. La expresión del estadístico *WMW*:

$$WMW = \frac{\sum_{a=1}^{N_a}\sum_{b=1}^{N_b} \mathbb{I}({l_1}^{(a)} > {l_0}^{(b)})}{N_a N_b} \tag{2.67}$$

donde $\mathbb{I}(\cdot)$ hace referencia a una función lógica que devuelve 1 cuando se cumple la relación y 0 cuando no; N_a, N_b son el número total de elementos que de cada hipótesis que se contemplan.

2.5.3. – Funciones de cópulas y densidades de cópula

Densidades de cópula derivadas de una PDF multivariante conocida.

En el apartado anterior ya comentamos como utilizando la expresión (2.46) se puede derivar una función de cópula partiendo de una *PDF* conocida. Dentro de este grupo de cópulas obtenidas a través de este método podemos encontrar las cópulas elípticas y la cópula derivada de una mezcla de Gaussianas.

Cópulas elípticas

Se definen como las cópulas asociadas a las distribuciones elípticas. Su rasgo más característico es que representan relaciones de dependencia simétricas sin importar que se analice la cola izquierda o derecha de las distribuciones implicadas. Las curvas de nivel de las variables aleatorias con este tipo de cópulas forman elipses. Las

cópulas Gaussianas y de Studen-T son cópulas elípticas, simétricas y de utilización relativamente simple, ya que se conocen bien las distribuciones que están asociadas.

❖ *Cópula Gaussiana*

La distribución normal (o Gaussiana) univariante, con media μ y varianza σ^2, se define como:

$$PDF \rightarrow f(z;\, \mu, \sigma) = N_z(\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1(z-\mu)^2}{2\ \sigma^2}}$$
$$CDF \rightarrow F(z;\, \mu, \sigma) = \Phi\left(\frac{z-\mu}{\sigma}\right) = \frac{1}{2}\left[1 + erf\left(\frac{z-\mu}{\sigma\sqrt{2}}\right)\right] \quad \text{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2} dt \tag{2.68}$$

El caso $\mu = 0$ y $\sigma^2 = 1$ se denomina distribución normal estándar:

$$N_z(\mu = 0, \sigma = 1): \quad PDF \rightarrow f(z) = \varphi(z) \qquad CDF \rightarrow F(z) = \Phi(z) \tag{2.69}$$

En el caso multivariante en el que se modela el conjunto de variables aleatorias $Z_1, Z_2, \dots, Z_d$, la expresión de la *PDF* pasa a ser:

$$f(\boldsymbol{z} = [z_1 \dots z_d]^T) = N_z(\boldsymbol{\mu}, \boldsymbol{R_z}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{R_z}|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T \boldsymbol{R_z}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})} \tag{2.70}$$

$\boldsymbol{\mu} = [\mu_1 \dots \mu_n]^T$: Vector de medias $(dx1)$, $\boldsymbol{R_z}$: Matriz de covarianza (dxd)

La matriz de covarianza es una matriz real definida positiva, cuya entrada (i, j) se corresponde con la covarianza de las variables aleatorias z_i y z_j:

$$\boldsymbol{R}_{\text{ij}} = \sigma_{z_i z_j} = E\left[(z_i - \mu_i)(z_j - \mu_j)\right] \tag{2.71}$$

observar que los elementos de la diagonal $(i = j)$ se corresponderán con las varianzas de cada una de las componentes. El coeficiente de correlación de Pearson es un índice que mide la relación lineal entre dos variables aleatorias cuantitativas, donde a diferencia de la covarianza, es independiente de la escala de medida de las variables. Tanto la covarianza como el coeficiente de correlación son índices que indican la fuerza y la dirección de una relación lineal y proporcionalidad entre dos variables estadísticas.

$$\rho_{z_i z_j} = \frac{\sigma_{z_i z_j}}{\sigma_{z_i}\sigma_{z_j}}, \qquad i \neq j \tag{2.72}$$

Si tomamos como referencia un modelo Gaussiano multivariante, cuyas funciones marginales son estándar, es decir con medias nulas y varianzas unidad y aplicamos la expresión 2.44 podemos obtener la función de densidad de cópula gaussiana:

$$c(\boldsymbol{u} = [u_1 \dots u_d]^T; \Theta) = \frac{1}{|\Sigma_\rho|^{1/2}} \cdot \exp\left(\frac{-\boldsymbol{z_m}^T \cdot (\Sigma_\rho{}^{-1} - \boldsymbol{I}) \cdot \boldsymbol{z_m}}{2}\right)$$

$$\boldsymbol{z_m} = \left[z_{m_1} = \Phi^{-1}(u_1), \dots, z_{m_d} = \Phi^{-1}(u_d)\right]^T \tag{2.73}$$

$$\boldsymbol{I}: Matriz\ identidad, \quad Matriz\ de\ correlación\ \Sigma_\rho = \begin{cases} 1 & ,si\ i = j \\ \rho_{u_i u_j} & ,si\ i \neq j \end{cases}$$

La función densidad de cópula gaussiana modela dependencia lineal de las variables aleatorias parametrizada a través de los coeficientes de correlación $\rho_{z_i z_j}$:

$$\Theta = \{\rho_{ij}\ ,\ \ i \neq j,\ \ i = 1, \dots, d,\ \ j = 1, \dots, d\} \tag{2.74}$$

La función cópula gaussiana se define como:

$$C(\boldsymbol{u} = [u_1 \dots u_d]^T; \Sigma_\rho) = \Phi(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d);\ \Sigma_\rho)$$

Φ_Σ: Función de distribución (CDF) Gaussiana multivariante con matriz de correlación Σ_ρ

Φ^{-1}: Función de distribución (CDF) Gaussiana univariante (2.75)

En la figura 2.10 podemos observar la función densidad de cópula gaussiana bivariante para varios valores del parámetro de correlación y la PDF del modelo del cuál se deriva:

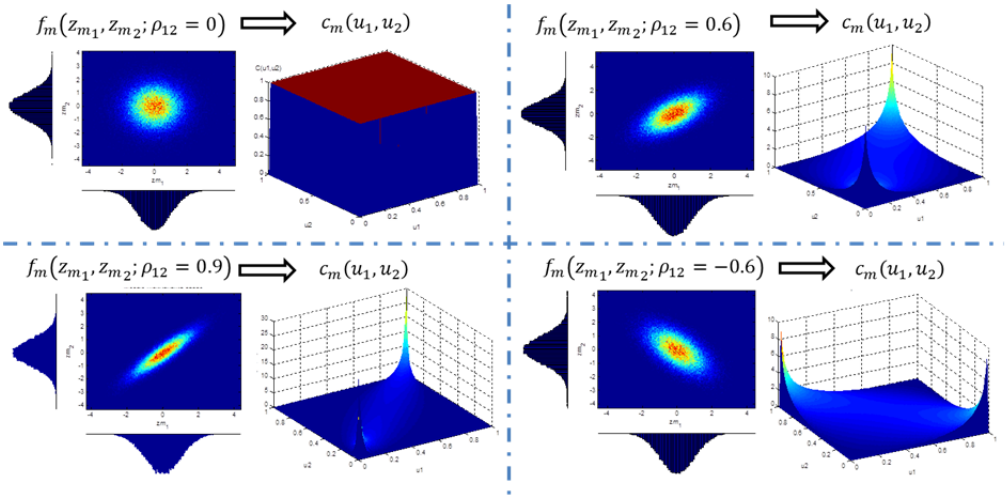


Figura 2.10 – Varios ejemplos de una densidad de cópula Gaussiana para distintos valores del parámetro ρ

❖ *Cópula Student-t*

La distribución Student-T es una distribución de probabilidad que surge del problema de estimar la media de una población normalmente distribuida cuando el tamaño de la muestra es pequeño. La distribución Student-T es la distribución de probabilidad del cociente:

$$\frac{Z}{\sqrt{V/\upsilon}}, \quad \begin{cases} Z: distribución\ normal\ standar \\ V: distribución\ Chi\ cuadrado\ con\ \upsilon\ grados\ de\ libertad \\ Z\ y\ V\ son\ independientes \end{cases} \tag{2.76}$$

La función de densidad de probabilidad univariante se define como:

$$f(z;\ \upsilon) = t_z(\upsilon) = \frac{\Gamma\left(\frac{\upsilon+1}{2}\right)}{\sqrt{\pi\upsilon}\cdot\Gamma\left(\frac{\upsilon}{2}\right)}\cdot\left(1+\frac{z^2}{\upsilon}\right)^{-\frac{(\upsilon+1)}{2}}$$
$$\Gamma: Función\ Gamma\ \ \Gamma(\mathrm{x}) = \int_0^\infty e^{-t}t^{x-1}dt \tag{2.77}$$

La extensión al caso multivariante:

$$f(\boldsymbol{z} = [z_1 \dots z_d]^T) = t_z(\upsilon, \Sigma) = \frac{\left(1+\frac{1}{\upsilon}(\boldsymbol{z_m}^T\cdot\Sigma\cdot\boldsymbol{z_m})\right)^{-(\upsilon+d)/2}\cdot\Gamma\left(\frac{\upsilon+d}{2}\right)}{\sqrt{(\pi\upsilon)^d}\cdot|\Sigma|^{1/2}\cdot\Gamma\left(\frac{\upsilon}{2}\right)} \tag{2.78}$$

Utilizando un modelo de distribución Student-T multivariante donde las varianzas de las componentes poseen un valor unidad, es decir definida por la matriz de correlación Σ_ρ, y aplicando la expresión 2.44 podemos obtener la expresión de la densidad de cópula t de Student:

$$c(\boldsymbol{u} = [u_1 \dots u_d]^T;\ \upsilon, \Sigma_\rho) = \frac{1}{|\Sigma_\rho|^{\frac{1}{2}}}\cdot\frac{\Gamma\left(\frac{\upsilon+d}{2}\right)}{\Gamma\left(\frac{\upsilon}{2}\right)}\cdot\left[\frac{\Gamma\left(\frac{\upsilon}{2}\right)}{\Gamma\left(\frac{\upsilon+1}{2}\right)}\right]^d\cdot\frac{\left(1+\frac{-\boldsymbol{y}^T{\Sigma_\rho}^{-1}\boldsymbol{y}}{\upsilon}\right)^{-\frac{(\upsilon+d)}{2}}}{\prod_{i=1}^d\left(1+\frac{y_i^2}{\upsilon}\right)^{-\frac{(\upsilon+1)}{2}}} \tag{2.79}$$
$$\boldsymbol{y} = [y_1 = t_\upsilon^{-1}(u_1), \dots, y_d = t_\upsilon^{-1}(u_d)]^T, \qquad \upsilon > 2$$
$$\Sigma_\rho = \begin{cases} 1\ , si\ i = j \\ \rho_{ij}\ , si\ i \neq j \end{cases}$$

La función cópula Student-T se define como:

$$C(\boldsymbol{u} = [u_1 \dots u_d]^T;\ \upsilon, \Sigma_\rho) = t_{\upsilon,\Sigma_\rho}(t_\upsilon^{-1}(u_1), \dots, t_\upsilon^{-1}(u_d)), \qquad \upsilon > 2 \tag{2.80}$$

$t_{\upsilon,\Sigma}$: Función de distribución (*CDF*) Student-T multivariante con matriz de correlación Σ y υ grados de libertad

t_υ^{-1}: Función de distribución (*CDF*) Student-t univariante con υ grados de libertad

En las figura 2.11 y 2.12 podemos observar ejemplos de la función densidad de cópula Student-T bivariante para diversos valores de sus parámetros. Cuando el parámetro que representa los grados de libertad υ de la cópula Student-T tiene un valor muy elevado, la cópula tiende a ser una cópula Gaussiana. Lo podemos comprobar en la figura 2.11, donde la densidad de cópula posee un coeficiente de correlación nulo, y pese a ello, para un valor bajo del parámetro ν todavía existe una estructura de dependencia entre ambas variables aleatorias. Cuando ν es muy elevado, la densidad de cópula tiende a ser gaussiana y por lo tanto las variables aleatorias tienden a ser independientes ($\rho = 0$).

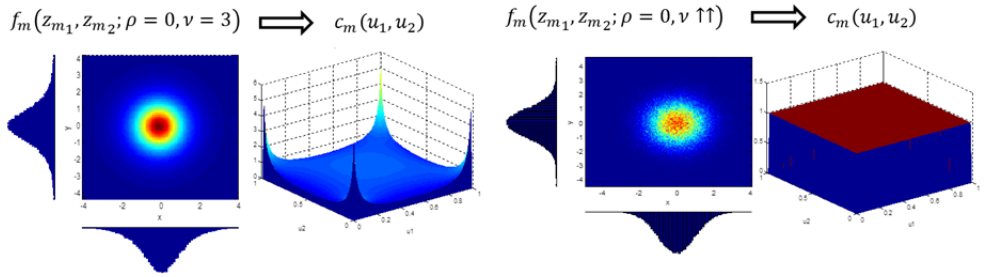


Figura 2.11 –Ejemplos de densidades de cópula Student-t bivariante con $\rho = 0$

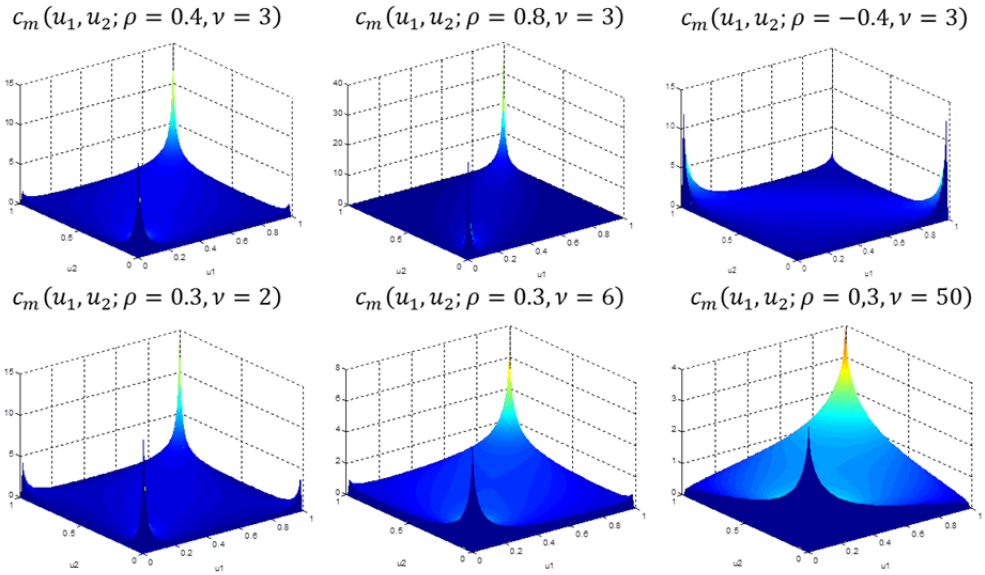


Figura 2.12 – Varios ejemplos de una densidad de cópula Student-T bivariante para distintos valores de los parámetros ρ y ν

Cópula basada en mezcla de Gaussianas (GMCM)

Tewari y Raghunathan [56] obtienen la función densidad de cópula asociada a un modelo multivariante basado en una mezcla de componentes Gaussianas. Como ya se vio en el apartado 2.4.6 la función de densidad de probabilidad multivariante *GMM* se define como:

$$f_{GMM}(\boldsymbol{z_m}|\Theta) = \sum_{i=1}^{K} \omega_i \cdot g(\boldsymbol{z_m}|\theta_i) = \sum_{i=1}^{K} \omega_i \cdot g(\boldsymbol{z_m}|\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}) \ , \qquad \boldsymbol{z_m} = [z_{m_1}, \dots, z_{m_d}] \tag{2.81}$$

donde $g(\boldsymbol{z_m}|\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i})$ es una *PDF* Gaussiana multivariante.

La función es parametrizada por el conjunto Θ, el cuál contiene el vector de medias y matriz de covarianza, así como el peso con que se pondera cada una de las componentes Gaussianas multivariantes.

$$\Theta = \{\theta_1 \dots \theta_K, \omega_1 \dots \omega_K\}, \qquad \omega_i \geq 0 \ \sum \omega_i = 1 \ , \ \ \theta_i = \{\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}\} \tag{2.82}$$

La proyección de la mezcla de componentes Gaussianas multivariantes sobre la dimensión j, es decir, la PDF marginal asociada a z_{m_j}, es una mezcla de distribuciones gaussianas univariantes dada por la PDF marginal:

$$f_{m_j}\left(z_{m_j}; \beta_j\right) = \sum_{i=1}^{K} \omega_i \cdot g(z_{m_j}|\mu_j^{(i)}, {\sigma^2}_j^{(i)}) \ , \qquad \beta_j = \left\{\mu_j^{(i)}, {\sigma^2}_j^{(i)}\right\} \tag{2.83}$$

Conocida la expresión de la *PDF* marginal no es difícil obtener las funciones *CDF* y *CDF* inversas de forma empírica:

$$f_{m_j}\left(z_{m_j}; \beta_j\right) \rightarrow \ F_{m_j}\left(z_{m_j}; \beta_j\right) = u_j \qquad F_{m_j}^{-1}\left(u_j; \beta_j\right) = \ z_{m_j} \tag{2.84}$$

La función de densidad de cópula derivada del modelo de mezclas de componentes Gaussianas se obtendrá, como ya se comentó en el punto 5.5.2, mediante la expresión:

$$c_{GMCM}(u_1, \dots, u_d; \Theta) = \frac{f_{GMM}\left({F_{m_1}}^{-1}(u_1; \beta_1), \dots, {F_{m_d}}^{-1}(u_d; \beta_d); \Theta\right)}{\left(\prod_{j=1}^{d} f_{m_j}\left({F_{m_j}}^{-1}(u_j; \beta_j); \beta_j\right)\right)} \tag{2.85}$$

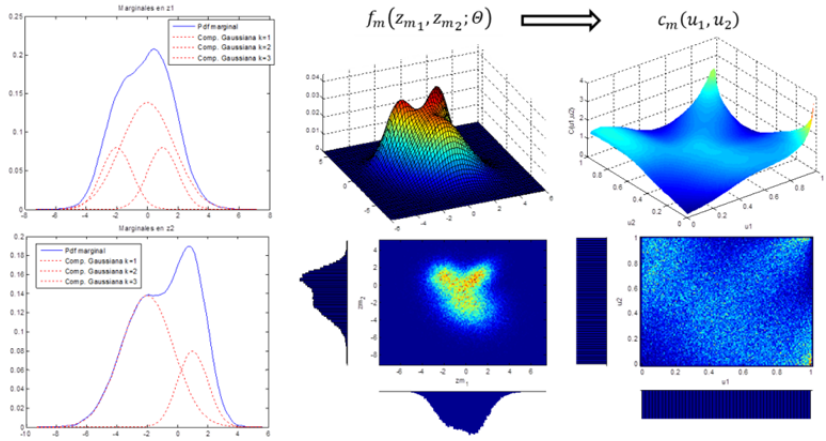
Las cópulas elípticas imponen la restricción de que la naturaleza de la dependencia es simétrica y de carácter lineal. La cópula basada en mezcla de Gaussianas permite relajar las condiciones de simetría y de linealidad impuestas por las cópulas elípticas, y permiten modelar de una forma mucho más flexible cualquier estructura de dependencia que puedan tener las variables aleatorias.

En la figura 2.13 se muestra un ejemplo de dos cópulas bivariantes derivadas de dos modelos con una mezcla de tres componentes Gaussianas con diferente parametrización. Podemos observar la versatilidad que nos proporciona este tipo de

cópula, donde su gran diversidad de parametrización nos permite modelar estructuras de dependencia muy complejas.

Parametrización 1: $\Theta = \{\theta_1\ ,\ \theta_2\ ,\theta_3,\ \omega_1 = 0.6,\ \omega_2 = 0.2,\ \omega_3 = 0.2\}$

$$\theta_1 = \left\{\mu_1 = [0,-2], \Sigma_1 = \begin{bmatrix} 3 & -0.8 \\ -0.8 & 3 \end{bmatrix}\right\}\ \theta_2 = \left\{\mu_2 = [1,1], \Sigma_2 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}\right\}\ \theta_3 = \left\{\mu_3 = [-2,1], \Sigma_3 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right\}$$



Parametrización 2: $\Theta = \{\theta_1\ ,\ \theta_2\ ,\theta_3,\ \omega_1 = 0.3,\ \omega_2 = 0.4,\ \omega_3 = 0.3\}$

$$\theta_1 = \left\{\mu_1 = [0,-2],\ \boldsymbol{\Sigma_1} = \begin{bmatrix} 3 & 0.6 \\ 0.6 & 2 \end{bmatrix}\right\}\ \theta_2 = \left\{\mu_2 = [1,1],\ \boldsymbol{\Sigma_2} = \begin{bmatrix} 2 & -0.5 \\ 0.5 & 1 \end{bmatrix}\right\}$$

$$\theta_3 = \left\{\mu_3 = [-2,-3],\ \boldsymbol{\Sigma_3} = \begin{bmatrix} 3 & -0.6 \\ -0.6 & 1.5 \end{bmatrix}\right\}$$

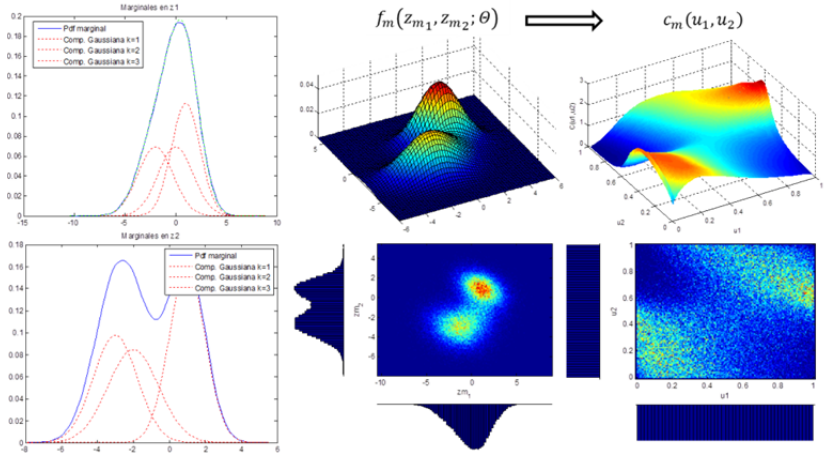


Figura 2.13 –Ejemplos de una PDF basada en una mezcla de tres componentes Gaussianas para distintos. A la izquierda se muestran las PDFs marginales, a la derecha se muestran el modelo GMM y su función densidad de cópula

- *Estimación de los parámetros de la GMCM*

Para la estimación de los parámetros de la función de cópula, Tewari y Raghunathan [56] han propuesto la aplicación de dos algoritmos concatenados. Un primer algoritmo es una versión del algoritmo *Expectation-Maximization* usado para la estimación de los parámetros de un modelo de PDF multivariante basada en mezcla de componentes Gaussianas, modificado convenientemente para adaptarse a la peculiaridad de que los datos con los que se estima el modelo no son fijos, sino que también varían según el valor de los parámetros a estimar. Dado que este primer algoritmo no garantiza un valor óptimo de los parámetros obtenidos, ya que en la etapa de maximización no garantiza que se alcanza un máximo local, sí que aporta una excelente estimación inicial con la que partir desde un segundo algoritmo basado en optimización por gradiente, con el que poder garantizar la convergencia a un máximo local.

Algoritmo "Expectation-Maximization"

Sean *N* vectores de muestras $\boldsymbol{U} = [\boldsymbol{u_1}, \dots., \boldsymbol{u_N}]$, con $\boldsymbol{u_n} = \left[{u_1}^{(n)}, \dots, {u_d}^{(n)}\right],\ n = 1 \dots N$ correspondientes al espacio hipercúbico con marginales uniformes, al cual se deben trasladar las muestras cuya función densidad conjunta de probabilidad pretendemos modelar mediante el uso de la densidad de cópula *GMCM*. Los parámetros Θ de esta función de densidad de cópula serán aquellos que maximicen la función de verosimilitud logarítmica $L(\boldsymbol{U}; \Theta)$:

$$\Theta = arg \max_{\Theta} \left(L(\boldsymbol{U}; \Theta) = \sum_{n=1}^{N} log \left(c_{GMCM}\left({u_1}^{(n)}, \dots, {u_d}^{(n)}; \Theta\right)\right)\right) \tag{2.86}$$

Observamos que el conjunto de parámetros a optimizar de la cópula (Θ), es el mismo que el del modelo de mezcla Gaussiana (*GMM*) del cual se deriva (ver apéndice D):

$$f_{GMM}\left({F_{m_1}}^{-1}(u_1; \beta_1), \dots, {F_{m_d}}^{-1}(u_d; \beta_d); \Theta\right) \tag{2.87}$$

A la hora de aplicar el algoritmo *EM* para estimar los parámetros de una *GMM*, las entradas del modelo, llamémosles $(x_1, \dots, x_d)$ son fijas, mientras que en nuestro caso, las entradas del modelo *GMM* varían en función de los parámetros $\left({F_{m_1}}^{-1}(u_1; \beta_1), \dots, {F_{m_d}}^{-1}(u_d; \beta_d)\right)$. En [56] se ha modificado el algoritmo *EM* usado en la estimación de parámetros de una *GMM* para incluir esta peculiaridad y poder ser usado en la estimación de los parámetros de la densidad de cópula *GMCM*.

Algoritmo de optimización basado en gradiente

Una forma alternativa para estimar los parámetros en un modelo de mezclas es mediante un algoritmo de optimización basado en gradiente. En *GMM* tenemos un caso único de algoritmo de descenso por gradiente donde el gradiente está

regularizado por la multiplicación por una matriz conocida definida positivamente. Se puede aplicar la factorización de Cholesky sobre la matriz de covarianzas Σ_k, ya que en una distribución normal multivariante es una matriz real simétrica:

$$\Sigma_k = \boldsymbol{V}_k \cdot {\boldsymbol{V}_k}^T, \;\; \boldsymbol{V}_k: matriz\ triangular\ inferior \tag{2.88}$$

Si restringimos que los elementos de la diagonal de las matrices triangulares inferiores $\boldsymbol{V}_k$ sean positivos, se garantiza que las matrices de covarianza Σ_k estén definidas positivamente.

$$(\boldsymbol{V}_k)_{j,j} > 0 \tag{2.89}$$

La función a optimizar, por lo tanto puede ser expresada como:

$$\underset{\omega_k,\mu_k,\Sigma_k}{\operatorname{argmax}} \left\{ \sum_{n=1}^{N} \left(ln\left(\sum_{k=1}^{K} \frac{\omega_k}{\sqrt{\boldsymbol{V}_k \cdot {\boldsymbol{V}_k}^T}} exp\left(-\frac{{y_k}^{(n)^T} \cdot \left(\boldsymbol{V}_k \cdot {\boldsymbol{V}_k}^T\right)^{-1} \cdot {y_k}^{(n)}}{2} \right) \right) - \sum_{j=1}^{d} ln\left(\sum_{k=1}^{K} \frac{\omega_k}{\sqrt{\left(\boldsymbol{V}_k \cdot {\boldsymbol{V}_k}^T\right)_{j,j}}} exp\left(\frac{-\left({y_k}^{(n)}\right)^2}{2 \cdot \left(\boldsymbol{V}_k \cdot {\boldsymbol{V}_k}^T\right)_{j,j}} \right) \right) \right) \right\} \tag{2.90}$$

donde las muestras de entrada varían con los parámetros:

$${y_k}^{(n)} = \left({x_k}^{(n)} - \mu_k\right) \quad {x_k}^{(n)} = {F_{m_1}}^{-1}\left({u_1}^{(n)}; \Theta(t)\right) \tag{2.91}$$

y sujeta a unas ciertas restricciones:

$$(\boldsymbol{V}_k)_{j,j} > 0\,, \qquad \omega_k \geq 0\,, \;\; \sum \omega_k = 1\,, \quad k = 1, \dots, K \tag{2.92}$$

Cópulas Arquimedianas

Es una de las familias más representativas y estudiadas en el ámbito teórico y aplicado. Su popularidad se encuentra asociada a su sencilla definición, la cual ha permitido construir un numeroso grupo de funciones que pertenecen a esta familia [57]. Son una clase de cópulas asociativas. Son populares porque permiten modelar la dependencia en grandes dimensiones con sólo un parámetro gobernado la fuerza de la dependencia.

Las funciones de densidad de cópula Arquimedianas no están derivadas de ningún modelo conocido. Se obtienen a través de la función de cópula que caracteriza a esta familia. Una función de cópula se define como Arquimediana si admite la representación [58]:

$$C(u_1, \dots, u_d) = \psi^{-1}\left(\psi(u_1) + \dots + \psi(u_d)\right) \tag{2.93}$$

donde, a la función $\psi(u)$ se le conoce como función generadora, definida en el rango $[0,1]$, es monotónicamente decreciente y $\psi(0) = \infty \;\; \psi(1) = 0$.

La condición necesaria y suficiente para que la función generadora constituya una cópula d-dimensional válida es que la función generadora inversa ψ^{-1} debe ser d-monótona en $[0, \infty)$, esto es, las derivadas k-ésimas de ψ deben cumplir:

$$(-1)^k \psi^{-1^{(k)}}(x) \geq 0, \qquad x \geq 0 \;\; k = 0,1, \dots, d \tag{2.94}$$

La función densidad de cópula de una cópula Arquimediana multivariante puede ser expresada como:

$$c(u_1, \dots, u_d) = \psi^{-1^{(k)}}\big(\psi(u_1) + \dots + \psi(u_d)\big) \prod_{i=1}^{d} \psi'(u_i) \tag{2.95}$$

Existen multitud de cópulas Arquimedianas bivariantes. En [51] se definen 22 tipos distintos de cópulas bivariantes, las cuales se recogen en la tabla 2.1. No todas ellas cumplen las condiciones necesarias para poder extenderse a un número arbitrario de dimensiones. Tres tipos de cópulas Arquimedianas son usadas con asiduidad para modelar distribuciones de densidad de probabilidad multivariantes: las cópulas de Clayton, Frank y de Gumbel.

	$C(u,v;\theta)$	$\psi_\theta(u)$	$\theta \in$
1	$\left[\max\left(u^{-\theta}+v^{-\theta}-1,0\right)\right]^{-1/\theta}$	$\frac{1}{\theta}\left(t^{-\theta}-1\right)$	$[-1,\infty)\backslash\{0\}$
2	$\max\left(1-\left[(1-u)^\theta+(1-v)^\theta\right]^{1/\theta},0\right)$	$(1-t)^\theta$	$[1,\infty)$
3	$\frac{uv}{1-\theta(1-u)(1-v)}$	$\ln\frac{1-\theta(1-t)}{t}$	$[-1,1)$
4	$\exp\left(-\left[(-\ln u)^\theta+(-\ln v)^\theta\right]^{1/\theta}\right)$	$(-\ln t)^\theta$	$[1,\infty)$
5	$-\frac{1}{\theta}\ln\left(1+\frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}\right)$	$-\ln\frac{e^{-\theta t}-1}{e^{-\theta}-1}$	$(-\infty,\infty)\backslash\{0\}$
6	$1-\left[(1-u)^\theta+(1-v)^\theta-(1-u)^\theta(1-v)^\theta\right]^{1/\theta}$	$-\ln\left[1-(1-t)^\theta\right]$	$[1,\infty)$
7	$\max\left(\theta uv+(1-\theta)(u+v-1),0\right)$	$-\ln\left[\theta t+(1-\theta)\right]$	$(0,1]$
8	$\max\left(\frac{\theta^2uv-(1-u)(1-v)}{\theta^2-(\theta-1)^2(1-u)(1-v)},0\right)$	$\frac{1-t}{1+(\theta-1)t}$	$[1,\infty)$
9	$uv\exp(-\theta\ln u\ln v)$	$\ln(1-\theta\ln t)$	$(0,1]$
10	$uv\Big/\left[1+(1-u^\theta)(1-v^\theta)\right]^{1/\theta}$	$\ln\left(2t^{-\theta}-1\right)$	$(0,1]$
11	$\left[\max\left(u^\theta v^\theta-2(1-u^\theta)(1-v^\theta),0\right)\right]^{1/\theta}$	$\ln\left(2-t^\theta\right)$	$(0,1/2]$

	$C(u,v;\theta)$	$\psi_\theta(u)$	$\theta \in$
12	$\left(1+\left[(u^{-1}-1)^\theta+(v^{-1}-1)^\theta\right]^{1/\theta}\right)^{-1}$	$\left(\frac{1}{t}-1\right)^\theta$	$[1,\infty)$
13	$\exp\left(1-\left[(1-\ln u)^\theta+(1-\ln v)^\theta-1\right]^{1/\theta}\right)$	$(1-\ln t)^\theta-1$	$(0,\infty)$
14	$\left(1+\left[(u^{-1/\theta}-1)^\theta+(v^{-1/\theta}-1)^\theta\right]^{1/\theta}\right)^{-\theta}$	$(t^{-1/\theta}-1)^\theta$	$[1,\infty)$
15	$\left\{\max\left(1-\left[(1-u^{1/\theta})^\theta+(1-v^{1/\theta})^\theta\right]^{1/\theta},0\right)\right\}^\theta$	$\left(1-t^{1/\theta}\right)^\theta$	$[1,\infty)$
16	$\frac{1}{2}\left(S+\sqrt{S^2+4\theta}\right),\ S=u+v-1-\theta\left(\frac{1}{u}+\frac{1}{v}-1\right)$	$\left(\frac{\theta}{t}+1\right)(1-t)$	$[0,\infty)$
17	$\left(1+\frac{[(1+u)^{-\theta}-1][(1+v)^{-\theta}-1]}{2^{-\theta}-1}\right)^{-1/\theta}-1$	$-\ln\frac{(1+t)^{-\theta}-1}{2^{-\theta}-1}$	$(-\infty,\infty)\backslash\{0\}$
18	$\max\left(1+\theta\Big/\ln\left[e^{\theta/(u-1)}+e^{\theta/(v-1)}\right],0\right)$	$e^{\theta/(t-1)}$	$[2,\infty)$
19	$\theta\Big/\ln\left(e^{\theta/u}+e^{\theta/v}-e^\theta\right)$	$e^{\theta/t}-e^\theta$	$(0,\infty)$
20	$\left[\ln\left(\exp(u^{-\theta})+\exp(v^{-\theta})-e\right)\right]^{-1/\theta}$	$\exp\left(t^{-\theta}\right)-e$	$(0,\infty)$
21	$1-(1-\{\max([1-(1-u)^\theta]^{1/\theta}+[1-(1-v)^\theta]^{1/\theta}-1,0)\}^\theta)^{1/\theta}$	$1-\left[1-(1-t)^\theta\right]^{1/\theta}$	$[1,\infty)$
22	$\max\left(\left[1-(1-u^\theta)\sqrt{1-(1-v^\theta)^2}-(1-v^\theta)\sqrt{1-(1-u^\theta)^2}\right]^{1/\theta},0\right)$	$\arcsin\left(1-t^\theta\right)$	$(0,1]$

Tabla 2.1 –Funciones de cópula Arquimedianas bivariantes recogida por Nelsen en [51].

Cópula de Clayton

Este tipo de cópula Arquimediana fue definida por Clayton [59] y también fue extensamente estudiada por Cook y Johnson [60]. La función generadora viene definida por:

$$\psi(u) = \frac{1}{\theta}\left(u^{-\theta} - 1\right) \tag{2.96}$$

La inversa de la función generadora es:

$$\psi^{-1}(y) = (\theta \cdot y + 1)^{-1/\theta} \tag{2.97}$$

La función de cópula multivariante posee la siguiente expresión:

$$C(u_1, \ldots, u_d;\ \theta) = \left(\sum_{i=1}^{d} u_i^{-\theta-1} - d + 1\right)^{-1/\theta}, \qquad \theta > 0 \tag{2.98}$$

La función densidad de cópula es fácil de obtener aplicándola expresión 2.40, donde la sucesiva derivación ofrece una fórmula recursiva [61], con una expresión cerrada simple de expresar:

$$c(u_1, \ldots, u_d;\ \theta) = \theta^d \frac{\Gamma\left(\frac{1}{\theta} + d\right)}{\Gamma\left(\frac{1}{\theta}\right)} \left(\prod_{i=1}^{d} u_i^{-\theta-1}\right)\left(\sum_{i=1}^{d} u_i^{-\theta} - d + 1\right)^{-\frac{1}{\theta}-d} \tag{2.99}$$

En la figura 2.14 se muestra la función densidad de cópula para distintos valores del parámetro θ. Para $\theta \sim 0$ se tiene el caso de independencia de las variables aleatorias, y conforme aumenta el valor del parámetro θ aumenta la fuerza de la dependencia que modela esta cópula.

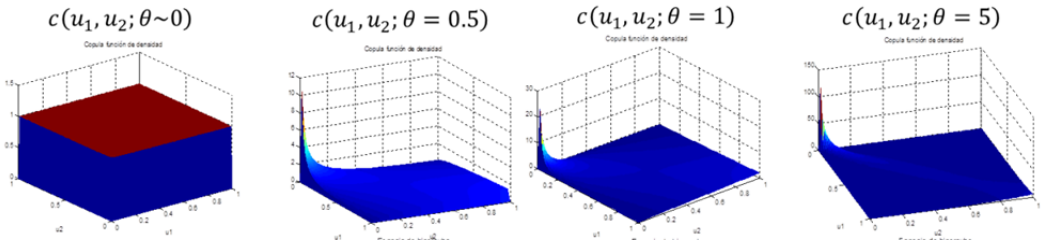


Figura 2.14 – Ejemplos bivariantes de la densidad de cópula de Clayton para diferentes valores del parámetro θ

Cópula de Frank

Este tipo de cópula Arquimediana ha sido definida y estudiada por Frank [62], [63]. La función generadora viene definida por:

$$\psi(u) = ln\left(\frac{e^{-\theta u} - 1}{e^{-\theta} - 1}\right) \tag{2.100}$$

La inversa de la función generadora es:

$$\psi^{-1}(y) = -\frac{1}{\theta} ln\left(1 + e^{y}\left(e^{-\theta} - 1\right)\right) \tag{2.101}$$

La función de cópula multivariante posee la siguiente expresión:

$$C(u_1, \dots, u_d;\, \theta) = -\frac{1}{\theta} ln\left(1 + \frac{\prod_{i=1}^{d}\left(e^{-\theta u_i} - 1\right)}{(e^{-\theta} - 1)^{d-1}}\right) \tag{2.102}$$

Si $d = 2 \rightarrow \theta \in \mathbb{R}\backslash\{0\}$. En el caso $d \geq 3 \rightarrow \theta > 0$

La expresión cerrada de la densidad de cópula de Frank no es conocida. Se puede obtener mediante la derivación de la función de cópula según (2.40). La hemos obtenido analíticamente hasta cuatro variables:

$$c(u_1, u_2) = -\theta(w_1 + 1)(w_2 + 1) \cdot \frac{k}{(k + w_1 w_2)^2} \qquad k = e^{-\theta} - 1$$

$$c(u_1, u_2, u_3) = \theta^2(w_1 + 1)(w_2 + 1)(w_3 + 1)k \cdot \frac{k - w_1 w_2 w_3}{(k + w_1 w_2 w_3)^3} \qquad k = \left(e^{-\theta} - 1\right)^2 \tag{2.103}$$

$$c(u_1, \dots, u_4) = -\theta^3(w_1 + 1)(w_2 + 1)(w_3 + 1)(w_4 + 1)k \cdot \frac{k^2 - 4k w_1 w_2 w_3 w_4 + (w_1 w_2 w_3 w_4)^2}{(k + w_1 w_2 w_3 w_4)^4} \qquad k = \left(e^{-\theta} - 1\right)^3$$

Para más de cuatro variables la derivación es muy tediosa. Se recomienda utilizar algún programa de cálculo simbólico. En el apéndice E incluimos una función de MATLAB en la que usamos el cálculo simbólico para obtener la expresión de la densidad de cópula para más de cuatro variables.

En la figura 2.15 se muestra la función densidad de cópula para distintos valores del parámetro θ. Para $\theta \sim 0$ se tiene el caso de independencia de las variables aleatorias, y conforme aumenta el valor del parámetro θ aumenta la fuerza de la dependencia que modela esta cópula.

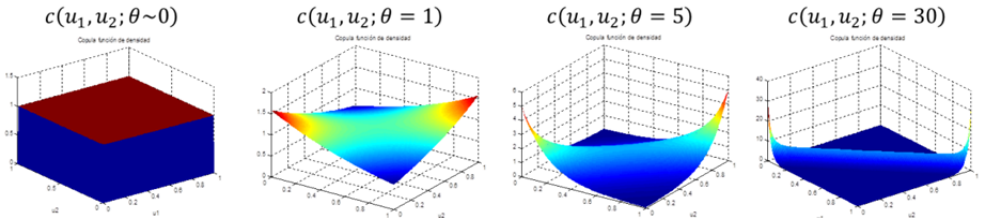


Figura 2.15 – Ejemplos bivariantes de la densidad de cópula de Frank para diferentes valores del parámetro θ

Cópula de Gumbel

Este tipo de cópula Arquimediana fue definida por Gumbel [64]. También fue estudiada por Hougaard [65], por eso podemos encontrárnosla también definida en la literatura por el nombre de cópula de Gumbel-Hougaard. En algunas publicaciones se le denomina también como cópula de valor extremo. La función generadora $\psi(u)$ y su inversa $\psi^{-1}(y)$ se definen como:

$$\psi(u) = \left(-ln(u)\right)^{\theta} \qquad \psi^{-1}(y) = e^{-y^{1/\theta}} \tag{2.104}$$

La función de cópula multivariante posee la siguiente expresión:

$$C(u_1, \dots, u_d;\, \theta) = exp\left(-\left(\sum_{i=1}^{d}(-ln\, u_i)^{\theta}\right)^{1/\theta}\right), \qquad \theta > 1 \tag{2.105}$$

La función densidad de cópula bivariante es:

$$C(u_1, u_2;\, \theta) = \frac{2\left[\left((-ln\, u_1)^{\theta} + (-ln\, u_2)^{\theta}\right)^{1/\theta}\right] + \theta - 1}{u_1\, u_2} exp\left(-\left((-ln\, u_1)^{\theta} + (-ln\, u_2)^{\theta}\right)^{1/\theta}\right) \cdot \tag{2.106}$$

$$\cdot \left((-ln\, u_1)^{\theta} + (-ln\, u_2)^{\theta}\right)^{-2+1/\theta} (-ln\, u_1)^{-1+\theta} (-ln\, u_2)^{-1+\theta}$$

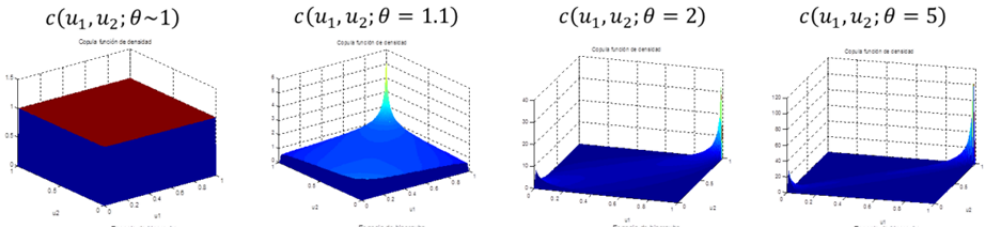


Figura 2.16 – Ejemplos bivariantes de la densidad de cópula de Gumbel para diferentes valores del parámetro θ

Para más de dos variables se puede obtener fácilmente aplicando la expresión 2.40, pero la expresión resultante es muy engorrosa, siendo no cerrada para un número arbitrario de d dimensiones. Al igual que en el caso de la cópula de Frank, puede obtenerse mediante algún programa de cálculo simbólico.

En la figura 2.16 se muestra la función densidad de cópula para distintos valores del parámetro θ. Para $\theta \sim 1$ se tiene el caso de independencia de las variables aleatorias, y conforme aumenta el valor del parámetro θ aumenta la fuerza de la dependencia que modela esta cópula.

Cópula de Farlie-Gumbel-Morgenstern (FGM)

La familia de cópulas de Farlie-Gumbel-Morgenstern (donde Morgenstern [66], Farlie [67] y Gumbel [64] discutieron sobre los principios que definen esta cópula, pero no es hasta el trabajo de H. Eyraud [68] donde se encuentra una referencia clara a la cópula *FGM* como distribución de marginales uniformes), al igual que las cópulas Arquimedianas modelan un cierto tipo de estructura fija de dependencia, pero a diferencia de las Arquimedianas la estructura de dependencia se codifica con más de un parámetro. En las cópulas Arquimedianas, donde sólo se utiliza un parámetro, la estructura de correlación es igual para todas las parejas de variables aleatorias que forman la distribución. Las cópulas *FGM*, en este sentido, proporcionan mayor versatilidad a la hora de modelar una estructura de dependencia, ya que asignan un parámetro de dependencia distinto, no solo a cada pareja, sino también a cada grupo de variables aleatorias que forman la distribución [69]. La función de Cópula viene dada por:

$$C(u_1, \dots, u_d;\, \theta) = \left[1 + \sum_{s=2}^{d} \sum_{1 \le j_1 < \cdots < j_s \le d} \alpha_{j_1,\dots,j_s} \prod_{i=1}^{s} \left(1 - u_{j_i}\right)\right] \cdot \prod_{i=1}^{d} u_i \qquad (2.107)$$

Viene parametrizada por el conjunto de parámetros θ:

$$\theta = \left\{\alpha_{j_1,\dots,j_s}\right\}, \qquad |\theta| = 2^d - d - 1 \qquad (2.108)$$

Los parámetros $\alpha_{j_1,\dots,j_s}$deben cumplir las siguientes condiciones (para una demostración en profundidad de las restricciones de los parámetros ver [45]):

$$1 + \sum_{s=2}^{m} \sum_{1 \le j_1 < \cdots < j_s \le d} \left(\alpha_{j_1,\dots,j_s} \prod_{i=1}^{s} \zeta_{j_i}\right) \ge 0 \qquad (2.109)$$

donde ζ_{j_i} puede tomar el valor +1 o -1, sucesivamente para $m = 2, \dots, d$.

Una representación interesante de la función de cópula puede realizarse definiendo el término polinomial:

$$P_d(\boldsymbol{u}) = 1 + \sum_{s=2}^{d} \sum_{1 \le j_1 < \cdots < j_s \le d} \alpha_{j_1,\ldots,j_s} \prod_{i=1}^{s} (1 - u_{j_i}) \tag{2.110}$$

por lo tanto, la podemos expresar como:

$$C(u_1, \ldots, u_d;\ \theta) = P_d(\boldsymbol{u}) \cdot \prod_{i=1}^{d} u_i \tag{2.111}$$

Podemos encontrar la expresión de la densidad de cópula de manera fácil mediante la derivación de la función de cópula (2.44). De esta manera obtenemos:

$$c(u_1, \ldots, u_d;\ \theta) = 1 + \sum_{s=2}^{d} \sum_{1 \le j_1 < \cdots < j_s \le d} \alpha_{j_1,\ldots,j_s} \prod_{i=1}^{s} (1 - 2u_{j_i}) = P_d(2 \cdot \boldsymbol{u}) \tag{2.112}$$

Por ejemplo, en el caso de tres variables tenemos, que la función de cópula FGM es:

$$\begin{aligned} C(u_1, u_2, u_3;\ \theta) = u_1 u_2 u_3 [1 &+ \alpha_{12}(1 - u_1)(1 - u_2) \\ &+ \alpha_{13}(1 - u_1)(1 - u_3) + \alpha_{23}(1 - u_2)(1 - u_3) \\ &+ \alpha_{123}(1 - u_1)(1 - u_2)(1 - u_3)] \end{aligned} \tag{2.113}$$

donde los parámetros deben cumplir las siguientes restricciones:

$$\begin{gathered} 1 + \alpha_{12}\zeta_1\zeta_2 + \alpha_{13}\zeta_1\zeta_3 + \alpha_{23}\zeta_2\zeta_3 \ge 0 \\ 1 + \alpha_{12}\zeta_1\zeta_2 + \alpha_{13}\zeta_1\zeta_3 + \alpha_{23}\zeta_2\zeta_3 + \alpha_{123}\zeta_1\zeta_2\zeta_3 \ge 0, \quad \zeta_k = 1 \text{ ó } \zeta_k = -1, k = 1,2,3 \end{gathered} \tag{2.114}$$

La función densidad de cópula, según lo visto en (2.40) será:

$$\begin{aligned} c(u_1, u_2, u_3;\ \theta) = 1 &+ \alpha_{12}(1 - 2u_1)(1 - 2u_2) + \alpha_{13}(1 - 2u_1)(1 - 2u_3) \\ &+ \alpha_{23}(1 - 2u_2)(1 - 2u_3) + \alpha_{123}(1 - 2u_1)(1 - 2u_2)(1 - 2u_3) \end{aligned} \tag{2.115}$$

En la figura 2.17 se muestra la función densidad de cópula *FGM* bivariante para distintos valores del unico parámetro θ. Para el caso bivariante, las ecuaciones que restringen el valor del parámetro son $1 + \theta \ge 0$ y $1 - \theta \ge 0$, de lo que se deriva que $\theta \in [-1,1]$. Para $\theta = 0$ se tiene el caso de independencia de las variables aleatorias, y conforme aumenta el valor del parámetro θ acercandose a $|\theta| = 1$ aumenta la fuerza de la dependencia que modela esta cópula.

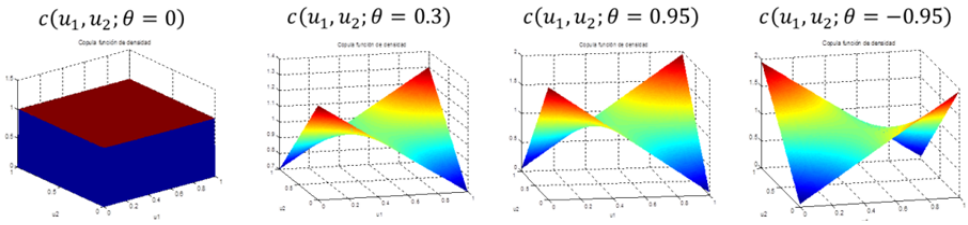


Figura 2.17 – Ejemplos bivariantes de la densidad de cópula FGM para diferentes valores del parámetro θ

Árboles de pares de cópulas (Vines)

Si repasamos la literatura veremos que existen muchísimas cópulas únicamente definidas para el caso bivariante. En cuanto a cópulas para modelos multivariantes, existe un menor conjunto con el que poder trabajar.

Hemos hablado de las cópulas elípticas multivariantes, las cuales son poco generalistas, ya que sólo modelan estructuras de dependencia lineal con la restricción de simetría. Las cópulas Arquimedianas multivariantes, al igual que las elípticas, modelan estructuras con unas características de dependencia muy concretas. Con la cópula *FGM* multivariante, se gana en generalidad a la hora de modelar una distribución, puesto que posee más grados de parametrización que las anteriores, pero sigue sin poder adaptarse a cualquier estructura de dependencia arbitraria.

La cópula derivada de una distribución basada en mezcla de Gaussianas es de carácter mucho más generalista, capaz de modelar correctamente muchas estructuras de dependencia arbitrarias, que no cumplen una serie de características prestablecidas.

Otro método más general para modelar compleja dependencia multivariante es usar un esquema jerárquico, donde se va modelando dependencia en parejas de dos variables. Bedford and Cooke [70] introdujeron en su trabajo una forma de construir distribuciones de probabilidad multivariantes basadas en bloques simples a los que llamaban “pair-cópulae”. Mediante esta técnica se pueden construir modelos multivariantes con dependencia compleja.

El esquema de modelado se basa en una descomposición de una densidad multivariante en un árbol de cópulas dos a dos, aplicadas sobre las variables aleatorias originales y sobre sus distribuciones incondicionales y condicionales. La descomposición mediante una jerarquía de cópulas en parejas de dos representa una manera más potente, flexible, adaptativa e intuitiva de extender la gran cantidad de cópulas bivariantes a modelos con muchas más dimensiones.

Descomposición en pares de cópulas de una distribución multivariante

Consideremos el conjunto de variables aleatorias dado por $\boldsymbol{Z} = \{Z_1, \dots, Z_d\}$, con función de densidad conjunta $f(z_1, \dots, z_d)$. La función de densidad puede ser factorizada como:

$$f(z_1, \dots, z_d) = f(z_d) f(z_{d-1}|z_d) f(z_{d-2}|z_{d-1} z_d) \cdot \dots \cdot f(z_1|z_2, \dots, z_d) \tag{2.116}$$

Esta descomposición es única para cada una de las posibles reordenaciones de las variables aleatorias. Cada función de densidad conjunta contiene de forma implícita la información marginal o individual del comportamiento de cada una de las variables, así como una descripción sobre la estructura de dependencia existente entre ellas. Mediante el uso de las cópulas podemos separar o aislar cada una de las fuentes de información marginal de la información sobre la estructura de dependencia. Por ejemplo, en el caso bivariante:

$$f(z_1, z_2) = c_{12}\big(F_1(z_1), F_2(z_2)\big) \cdot f_1(z_1) \cdot f_2(z_2) \tag{2.117}$$

donde $c_{12}(\cdot)$ es la cópula asociada al par de variables transformadas $F_1(z_1)$ y $F_2(z_2)$.

La función de densidad condicionada $f(z_1|z_2)$ la podemos expresar por tanto como:

$$f(z_1|z_2) = c_{12}\big(F_1(z_1), F_2(z_2)\big) \cdot f_1(z_1) \tag{2.118}$$

Para una densidad condicional de tres variables podemos encontrar varias descomposiciones:

$$f(z_1|z_2 z_3) = c_{12|3}\left(F_{1|3}(z_1|z_3), F_{2|3}(z_2|z_3)\right) \cdot f(z_1|z_3) \tag{2.119}$$

$$f(z_1|z_2 z_3) = c_{13|2}\left(F_{1|2}(z_1|z_2), F_{3|2}(z_3|z_2)\right) \cdot f(z_1|z_2) \tag{2.120}$$

donde las cópulas $c_{12|3}(\cdot)$ y $c_{13|2}(\cdot)$ son diferentes. Podemos seguir descomponiendo en parejas de cópulas, por ejemplo continuando en (2.120) tenemos una descomposición en dos densidades de cópula bivariantes:

$$f(z_1|z_2 z_3) = c_{13|2}\left(F_{1|2}(z_1|z_2), F_{3|2}(z_3|z_2)\right) \cdot c_{12}\big(F_1(z_1), F_2(z_2)\big) \cdot f_1(z_1) \tag{2.121}$$

Se puede aplicar sobre cada uno de los términos de densidad condicional marginal de (2.121) el mismo procedimiento de descomposición en funciones de densidad de cópula bivariantes. La fórmula general usada en la descomposición puede ser expresada como:

$$f(x|\boldsymbol{v}) = c_{x v_j|\boldsymbol{v}_{-j}}\left(F\big(x|\boldsymbol{v}_{-j}\big), F\big(v_j|\boldsymbol{v}_{-j}\big)\right) \cdot f\big(x|\boldsymbol{v}_{-j}\big) \tag{2.122}$$

donde $\boldsymbol{v}$ hace referencia a un vector de variables aleatorias n-dimensional, v_j es una componente de ese vector arbitrariamente escogida y el vector $\boldsymbol{v}_{-j}$ es el resultante de extraer la componente v_j del vector $\boldsymbol{v}$.

Como conclusión podemos determinar que una densidad de cópula multivariante, bajo determinadas condiciones de regularidad, puede ser expresada como un producto de cópulas bivariantes [71], cada una de ellas actuando sobre diferentes distribuciones de probabilidad condicional. La construcción de estas cópulas bivariantes involucran las distribuciones condicionales del tipo $F(x|\boldsymbol{v})$. Joe mostró en [72] que, para cada j:

$$F(x|\boldsymbol{v}) = \frac{\partial C_{x,v_j|\boldsymbol{v}_{-j}}\left(F(x|\boldsymbol{v}_{-j}), F(v_j|\boldsymbol{v}_{-j})\right)}{\partial F(v_j|\boldsymbol{v}_{-j})} \tag{2.123}$$

En el caso en que $\boldsymbol{v}$ sólo posea una componente v tenemos:

$$F(x|v) = \frac{\partial C_{x,v}\left(F_x(x), F_v(v)\right)}{\partial F_v(v)} \tag{2.124}$$

Definimos la función $h(x, v; \theta)$ como la función de distribución condicional $F(x|v)$ cuando las v.a x y v son uniformes, es decir, $f(x) = f(v) = 1$ y $F(x) = x \;\; F(v) = v$.

$$h(x, v; \theta) = F(x|v) = \frac{\partial C_{x,v}(x, v; \theta)}{\partial v} \tag{2.125}$$

Vines

Para distribuciones de grandes dimensiones, existen un número muy elevado de posibles descomposiciones en productos de cópulas bivariantes. Para ayudar a organizar las posibles descomposiciones, Bedford and Cooke [73] introdujeron un modelo gráfico de construcción de densidades de cópula usando "vines" (más concretamente vines regulares).

Un vine es un conjunto de árboles anidados, donde los extremos de un árbol *k-ésimo* son los nodos del árbol *(k+1)-ésimo,* estando cada uno de los árboles limitado por un número máximo de nodos. $\mathcal{V}$ es un vine de K elementos si [18]:

1. $\mathcal{V} = (T_1, \dots, T_{K-1})$
2. T_1 es un árbol que conecta los nodos $N_1 = \{1, \dots, K\}$ con los extremos E_1; para $k = 2, \dots, K-1, T_k$ es un árbol que conecta N_k y E_{k-1}.

$\mathcal{V}$ es un vine regular de K elementos si además cumple que para $k = 2, \dots, K-1$ si $a = \{a_1, a_2\}$ y $b = \{b_1, b_2\}$ son nodos conectados en el árbol T_k, donde $a = \{a_1, a_2\}$ y $b = \{b_1, b_2\}$ también pertenecen al árbol T_{k-1}, entonces uno de los elementos a_1 *o* a_2 es el mismo que uno de los elementos b_1 *o* b_2.

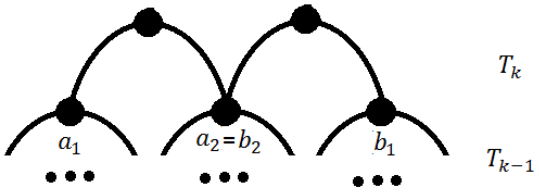


Figura 2.18 –Representación gráfica de un vine regular

El número de descomposiciones que admite la estructura de los vines regulares aún es muy general y abarca un gran número de posibilidades para distribuciones de grandes dimensiones. Existen dos subtipos más específicos de regular vines que dan lugar a menos posibles descomposiciones: los D-vines ("*Drawable-vines*") y los C-vines ("*Canonical-vines*"). Cada uno de los modelos describe una forma específica de descomponer la función densidad de cópula multivariante en producto de densidades de cópula bivariantes [70]:

- D-Vine:

$$c(\boldsymbol{z}) = \prod_{j=1}^{d} \prod_{i=1}^{d-j} c_{i,i+j|i+1,\dots,i+j-1}\left(F\left(z_i|z_{i+1},\dots,z_{i+j-1}\right), F\left(z_{i+j}|z_{i+1},\dots,z_{i+j-1}\right)\right) \quad (2.126)$$

- C-Vine:

$$c(\boldsymbol{z}) = \prod_{j=1}^{d} \prod_{i=1}^{d-j} c_{j,j+i|1,\dots,j-1}\left(F\left(z_j|z_1,\dots,z_{j-1}\right), F\left(z_{i+j}|z_1,\dots,z_{j-1}\right)\right) \quad (2.127)$$

donde el índice j identifica a cada árbol y el índice i varia entre los extremos de cada árbol.

Cada posible reordenación de las variables dará lugar a una factorización distinta de la función densidad de cópula.

Los D-vines son más flexibles que los C-vines. En la estructura C-vines, las relaciones de dependencia se realizan entre una variable de referencia o piloto y el resto, así como las relaciones de dependencia condicionadas siempre incluyen esa variable de referencia. Los C-vines suelen ser utilizados cuando se sabe que un determinado detector juega un papel principal gobernando la dependencia con el resto.

Por ejemplo, la descomposición de una función de densidad de cuatro dimensiones usando una cópula basada en un árbol con estructura:

- D-vine:

$$f(z_1,z_2,z_3,z_4) = f(z_1)f(z_2)f(z_3)f(z_4)\cdot \\ \cdot\ c_{12}\big(F(z_1),F(z_2)\big)\cdot c_{23}\big(F(z_2),F(z_3)\big)\cdot c_{34}\big(F(z_3),F(z_4)\big)\cdot \\ \cdot\ c_{13|2}\big(F(z_1|z_2),F(z_3|z_2)\big)\cdot c_{24|3}\big(F(z_2|z_3),F(z_4|z_3)\big)\cdot \\ \cdot\ c_{14|23}\big(F(z_1|z_2,z_3),F(z_4|z_2,z_3)\big) \tag{2.128}$$

- C-vine:

$$\mathrm{f}(z_1,z_2,z_3,z_4) = \mathrm{f}(z_1)\mathrm{f}(z_2)\mathrm{f}(z_3)\mathrm{f}(z_4)\cdot \\ \cdot\ \mathrm{c}_{12}\big(\mathrm{F}(z_1),\mathrm{F}(z_2)\big)\cdot \mathrm{c}_{13}\big(\mathrm{F}(z_1),\mathrm{F}(z_3)\big)\cdot \mathrm{c}_{14}\big(\mathrm{F}(z_1),\mathrm{F}(z_4)\big)\cdot \\ \cdot\ \mathrm{c}_{23|1}\big(\mathrm{F}(z_2|z_1),\mathrm{F}(z_3|z_1)\big)\cdot \mathrm{c}_{24|1}\big(\mathrm{F}(z_2|z_1),\mathrm{F}(z_4|z_1)\big)\cdot \\ \cdot\ \mathrm{c}_{34|12}\big(\mathrm{F}(z_3|z_1,z_2),\mathrm{F}(z_4|z_1,z_2)\big) \tag{2.129}$$

A la hora de realizar una inferencia de una función de densidad de cópula basada en arboles de densidades de cópulas bivariantes tenemos tres niveles de libertad:

1. Selección de una estructura y una de sus posibles factorizaciones.
2. Selección de cada una de las cópulas bivariantes usadas (no tiene por qué pertenecer todas las cópulas bivariantes a la misma familia).
3. Selección del conjunto de parámetros asociados a todas las cópulas bivariantes.

Para una determinada estructura y factorización, escogiendo la familia de la cópula con la parametrización que se adapta perfectamente a cada pareja de variables con las que se trabaja en los árboles de la factorización, se garantiza que el modelo de densidad de cópula obtenido se adapta también perfectamente a las condiciones de dependencia conjunta entre todas las variables.

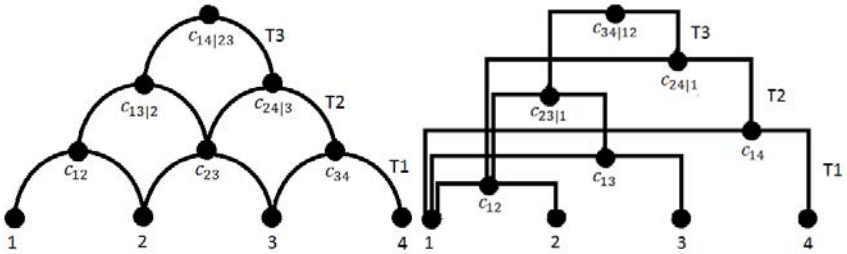


Figura 2.19 –Esquema de descomposición de PDF de cuatro dimensiones usando cópulas en arboles basadas en D-vines a la izquierda y C-vines a la derecha

2.6. - Conclusiones

El problema de fusión soft óptima en detección se plantea como un test de hipótesis basado en la relación de verosimilitud, por lo tanto, involucra la estimación de las funciones de densidad de probabilidad multivariante de las variables a fusionar.

La heterogeneidad y la dependencia estadística de los datos a fusionar pueden resultar en funciones de densidad de probabilidad multivariantes muy complejas, las cuales no pueden ser modeladas de una forma precisa mediante la asunción de independencia, la cual deriva en un sencillo modelo basado en un producto de las funciones de probabilidad marginales.

Para una correcta caracterización del comportamiento conjunto de los datos se deberá hacer uso de técnicas de estimación más complejas. Por ello hemos revisado los principales métodos paramétricos y no paramétricos existentes. Técnicas como el análisis de componentes independientes para transformar los datos a un dominio donde sean independientes o la estimación directa mediante funciones de núcleo o mediante mezclas de Gaussianas se han presentado como diferentes alternativas válidas. El principal problema que plantean estas técnicas es la alta complejidad y costes computacionales que involucran, siendo poco recomendable su uso cuando el número de datos a fusionar es elevado.

Se ha realizado una completa revisión del estado del arte de la teoría de cópulas por su novedad e incipiente uso en el ámbito de procesado de señal. Se ha mostrado como una función de densidad de probabilidad puede ser factorizada utilizando el modelo de independencia multiplicado por una función densidad de cópula, la cual se encarga de modelar la información de dependencia. Así, mediante la utilización de la teoría de cópulas puede simplificarse el proceso de estimación de *PDFs* multivariantes, pudiendo modelar de una forma sencilla y precisa complejas distribuciones de datos dependientes y heterogéneos, incluso si el número de datos a fusionar es muy elevado.

También se ha argumentado la posibilidad de que no sea posible o viable en todos los escenarios una fusión de información basada en la relación de verosimilitud, ante lo que se han propuesto y comentado brevemente otra serie de técnicas para llevar a cabo la fusión de datos, tales como las combinaciones lineales y los clasificadores binarios.

Algunos de estos métodos serán aplicados en los sucesivos capítulos pertenecientes a la segunda parte de la presente tesis doctoral para la fusión en problemas de detección de eventos acústicos y problemas de autentificación biométrica.

Capítulo 3:
Fusión de scores

"La época de la individualidad llegó a su conclusión, y es el deber de los reformistas iniciar la época de la asociación"

- Giuseppe Mazzini -

Este capítulo se centra en la fusión de información soft proporcionada por diversos detectores, donde todos los datos que se pretenden combinar se encuentran definidos en un mismo rango normalizado $[0,1]$ *y se presupone que poseen buenas propiedades discriminatorias de forma aislada. Así, denominamos scores a los datos que cumplen con este supuesto. Usualmente en la literatura se liga la fusión de scores a la combinación de diversos detectores a través de la fusión de las probabilidades a posteriori que proporcionan a su salida; las técnicas usadas en este escenario también son válidas para el caso en que se pretende combinar cualquier tipo de información soft normalizada en este rango, sin necesidad de que estos datos constituyan conceptualmente probabilidades a posteriori. Por ejemplo, en el punto 2.2 ya introdujimos una técnica de combinación lineal de datos para la fusión de información soft, donde es usual la normalización previa de los datos; esta técnica también es usada para la fusión de detectores a través de sus probabilidades a posteriori.*

Inicialmente se realizará una pequeña revisión centrándonos en el contexto de fusión de detectores a través de sus probabilidades a posteriori, para observar que, únicamente en el caso en que los detectores trabajen con diferentes entradas independientes, el hecho de que la información soft sea considerada conceptualmente como probabilidades a posteriori conduce a un tipo especial de regla de fusión. En cualquier otro caso, las técnicas que se exponen a continuación pueden ser aplicadas tanto para la fusión de probabilidades a posteriori o de cualquier otro tipo de información normalizada.

3.1. – Introducción

El diseño de un detector óptimo pasa por la umbralización un estadístico basado en la relación de verosimilitud $\Lambda(\boldsymbol{x})$. Existen casos en los que la implementación directa de este detector óptimo no será posible. En la implementación de ciertos sistemas, puede ser difícil el tener acceso a todas las componentes de este vector $\boldsymbol{x}$; por ejemplo, imaginemos un sistema de detección distribuida donde los sensores están separados espacialmente y donde aglutinar toda la información recogida por los sensores en un único centro de procesado puede ser inviable. En otros casos será complicado tener toda la información sincronizada temporalmente, en el mismo soporte y/o tasa en una etapa tan prematura del sistema de detección; por ejemplo, en la fusión de audio con video, cada fuente de información puede estar muestreada a diferente tasa de datos, o en la fusión de varias fuentes de video, cada una de ellas puede estar en un formato diferente. Aunque se tenga acceso a toda la información, bajo un mismo soporte, una tasa de datos común y correctamente sincronizada, el detector óptimo puede resultar muy complejo de implementar o su utilización no ser viable en el sistema: puede necesitar una gran cantidad de muestras o un elevado

tiempo para su entrenamiento, precisar de una gran capacidad de cómputo para su funcionamiento, una gran capacidad de almacenamiento…

En el caso que no pueda realizarse una implementación directa, una solución puede pasar por la utilización de una serie de detectores subóptimos, que trabajen con sólo un subconjunto de las fuentes de información y/o con técnicas con menores requerimientos de computación y almacenaje. Así, mediante la fusión de detectores subóptimos (figura 3.1) se pretende alcanzar las prestaciones obtenidas con el detector óptimo, salvando todas las dificultades y complicaciones que puede entrañar la implementación y utilización en un determinado sistema de éste.

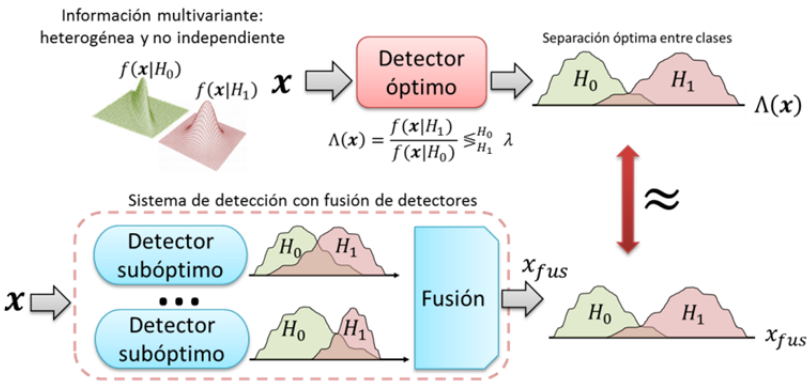


Figura 3.1 – Fusión de detectores subóptimos como forma de aproximarse a las prestaciones del detector óptimo

Dentro de los campos de la inteligencia artificial, el análisis y reconocimiento de patrones, y el aprendizaje automático, existe una corriente de investigación centrada en la combinación de diferentes clasificadores [74], [17], [75], [76]. El estudio de la fusión de detectores, entendido como un caso de clasificador binaria podría encajarse dentro de estos campos.

Un clasificador, dado un patrón de datos $\boldsymbol{x}$, trata de discernir a qué clase pertenece de entre d posibles $\{c_1, \dots, c_d\}$. Para la implementación de un clasificador, desde el punto de vista de la teoría de probabilidad (de una u otra forma, la gran mayoría de clasificadores poseen una naturaleza derivada de esta teoría o puede relacionarse de alguna forma con ella), cada clase c_i se debe modelar en el espacio de las observaciones $\boldsymbol{x}$ mediante la función densidad de probabilidad $f(\boldsymbol{x}|c_i)$ y su probabilidad a priori de ocurrencia $P(c_i)$. Se selecciona la clase que, ante un patrón $\boldsymbol{x}$, proporciona la máxima probabilidad a posteriori (criterio *MAP*, "*Maximum A Posteriori*"):

$$\boldsymbol{x} \in c_j \Leftrightarrow \max_j \left(P(c_j|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|c_j)P(c_j)}{f(\boldsymbol{x})} \right), \quad f(\boldsymbol{x}) = \sum_{i=1}^{d} f(\boldsymbol{x}|c_i)P(c_i) \tag{3.1}$$

Así, es común que los clasificadores aporten como salida soft algún tipo de estimación de estas probabilidades a posteriori $\hat{P}(c_j|\boldsymbol{x}) \in [0,1]$ (incluso aquellos basados en conceptos no puramente estadísticos, como puede ser medidas de distancia normalizadas en clasificadores como ocurre con los *SVM*). A este tipo de salidas soft dadas por estimaciones de las probabilidades a posteriori se les denomina scores.

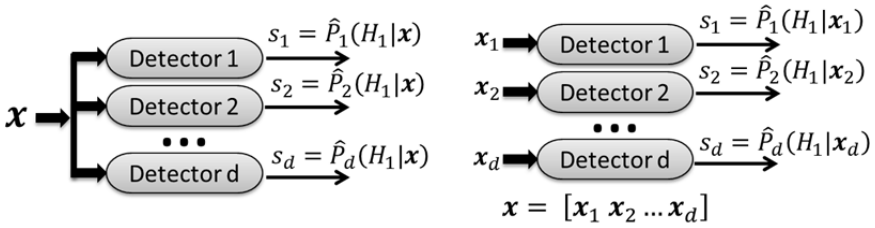


Figura 3.2 –Fusión de detectores, o bien todos los detectores trabajando con la misma entrada (izquierda), o cada detector trabajando con diferentes entradas (derecha).

Por lo tanto, cuando se habla de fusión de scores en estos campos, se están refiriendo a la combinación de d clasificadores mediante la integración de las diferentes probabilidades a posteriori para cada clase c_i aportadas por éstos (figura 3.2). Puede ocurrir que cada uno esté trabajando con un conjunto de observaciones $\boldsymbol{x}_j,\ j = 1 \ldots d$ diferente,

$$\left[\hat{P}_1(c_i|\boldsymbol{x}_1)\ \ldots\ \hat{P}_K(c_i|\boldsymbol{x}_d)\right] \in [0,1]^d \rightarrow \hat{P}_{fus}(c_i|\boldsymbol{x}) \in [0,1], \qquad \boldsymbol{x} = [\boldsymbol{x}_1 \ldots \boldsymbol{x}_d] \qquad (3.2)$$

donde $\hat{P}_j(c_i|\boldsymbol{x}_j)$ hace referencia a la estimación de la probabilidad a posteriori de la clase c_i aportada como información soft a la salida del clasificado k-ésimo, en base a los datos de entrada con los que trabaja $\boldsymbol{x}_j$. Mediante la fusión de estos scores se pretende obtener una estimación de la probabilidad de la clase c_i condicionada al conjunto global de entradas $\boldsymbol{x} = [\boldsymbol{x}_1 \ldots \boldsymbol{x}_d]$. También puede ocurrir que todos los clasificadores trabajen con las mismas observaciones como entrada $\boldsymbol{x}$, y lo que se busque es que la estimación obtenida con la fusión $\hat{P}_{fus}(c_i|\boldsymbol{x})$ sea más precisa que la de cualquier clasificador en particular:

$$\left[\hat{P}_1(c_i|\boldsymbol{x})\ \ldots\ \hat{P}_K(c_i|\boldsymbol{x})\right] \in [0,1]^K \rightarrow \hat{P}_{fus}(c_i|\boldsymbol{x}) \in [0,1] \qquad (3.3)$$

Otro campo donde se pueden encontrar diferentes estudios de fusión de datos en tareas de detección es la biometría. Por ejemplo, existen sistemas multibiométricos que utilizan más de un canal de información para la identificación de personas. En estos sistemas es habitual hablar de "*matchers*" para referirse a los diferentes algoritmos de detección usados, denominando como score a cualquier tipo de valoración soft aportada por éstos. Los scores de salida de los algoritmos de detección en estas aplicaciones pueden ser de diversa naturaleza y estar definidos en rangos

muy diversos. Por lo tanto, cuando se habla de fusión de scores en este campo, se habla de combinación de información soft genérica aportada por diferentes algoritmos de detección, sin asociarla a una probabilidad a posteriori. Es muy común el uso de técnicas de normalización previas a ciertas técnicas de fusión para conseguir que todas las valoraciones aportadas por los matchers estén definidas en el mismo rango.

Por lo tanto, para aunar todos estos los campos, podríamos asociar la fusión de scores a la fusión de cualquier tipo de valoración soft aportada por un conjunto de detectores, bien sea en forma de probabilidades a posteriori o cualquier otro tipo de información continua. Pero como estamos realizando una división de las técnicas de fusión según las características de los propios datos a combinar, sin atender a su origen o nivel de integración, consideramos la fusión de scores como un caso particular de técnicas de fusión soft, en el que todos los datos a fusionar se encuentran definidos en el rango normalizado $[0,1]$, no necesariamente representando exclusivamente probabilidades a posteriori, ni siendo datos aportados únicamente por diversos detectores o "*matchers*". Denotamos a los scores aportados por el canal de datos "*i*" como s_i. Utilizando técnicas de normalización como las que incluimos en el apéndice A se podrá transformar cualquier tipo de información soft con objeto de representarla en el rango $[0,1]$, y mediante técnicas de calibración (apéndice B) se le podrá dar un sentido de probabilidad a posteriori, y así, poder aplicar cualquiera de las técnicas de fusión que en este capítulo se incluyen.

Inicialmente nos centramos en el caso particular de la fusión de detectores a través de probabilidades a posteriori, para comprobar que, en el caso de utilizar diferentes entradas, y que éstas sean independientes, podemos encontrar una regla óptima de fusión a través de los scores. Se muestra cómo esta regla deja de ser óptima en el caso de que exista dependencia estadística entre las entradas, y se deben usar otras técnicas de combinación. A continuación introducimos las diferentes técnicas que pueden ser usadas para combinar scores, pudiendo usarse tanto para combinación de detectores que comparten la misma entrada, como para cualquier tipo de información soft normalizada en el rango $[0,1]$.

3.2. – Fusión de detectores a través de probabilidades a posteriori

En diversas investigaciones, al hablar de fusión de scores se están refiriendo a la fusión de probabilidades a posteriori aportadas como salida soft por varios detectores. Analizaremos de forma separada el caso de que todos los detectores trabajan con el mismo vector de observaciones, del caso en que cada detector trabaja con un vector de observaciones diferente al resto. Veremos que, en este último caso, bajo el supuesto de que exista independencia entre sus entradas, el hecho de que los datos se definan como probabilidades a posteriori nos permitirá obtener una regla de

fusión óptima. Argumentaremos como, en el caso de que exista dependencia, el hecho de que los datos representen probabilidades a posteriori no supondrá ninguna ventaja con respecto a cualquier otro tipo de información normalizada. Por lo tanto concluirnos que, tanto si se desean fusionar detectores a través de información soft dada por probabilidades a posteriori, como si se desea combinar cualquier otro tipo de información soft normalizada rango $[0,1]$, pueden emplearse el conjunto de técnicas que presentaremos en este capítulo.

Diferentes detectores trabajando con el mismo vector de observación como entrada

Nos situamos inicialmente en el escenario donde se pretende combinar una serie de detectores que utilizan una misma entrada dada por el vector observaciones $\boldsymbol{x}$. Se pueden combinar diferentes tipos de detectores, o también se puede utilizar un mismo tipo de detector, pero utilizando diferentes parámetros en su implementación (por ejemplo, un detector basado en k-vecinos pero utilizando diferentes números de vecinos) o diferentes técnicas de entrenamiento (por ejemplo, redes neuronales con estructura fija pueden tener diferentes conjuntos de pesos según el método usado para su entrenamiento [77]).

En este caso, se considera que cada uno de los detectores $i = 1, \dots, d$ produce una salida soft dada por una estimación diferente de la misma probabilidad a posteriori $P(H_1|\boldsymbol{x}) \approx \hat{P}_i(H_1|\boldsymbol{x})$. Es razonable pensar en algún tipo de valoración media como regla de combinación de todas ellas, ya que todas las valoraciones son relativas a un mismo evento, obtenidas usando los mismos datos y bajo las mismas condiciones.

Así, podemos encontrar diversos estudios y aplicaciones donde se han utilizado la media aritmética, geométrica o harmónica como función de combinación de los scores $s_i = \hat{P}_i(H_1|\boldsymbol{x})$ [76]. Con el uso de estas medias como funciones para conseguir un consenso, se valora de igual forma los scores de cada uno de los detectores. Se consideran todos los detectores igual de buenos o fiables. Con objeto de combinar detectores con diferente nivel de fiabilidad o prestaciones se utilizan reglas como la suma o producto ponderado, asignándole diferentes pesos a cada uno de los scores. Una versión más compleja de la media aritmética ponderada, en donde el peso de cada componente es fijo, es la que se propone en la arquitectura de mezcla de expertos, donde los pesos w_i dependen del valor de las entradas $\boldsymbol{x}$: $w_i = w_i(\boldsymbol{x})$; no sólo se tiene en cuenta que un detector es más fiable que otro, sino que también se discrimina en qué zonas del espacio de observación lo es. Comentaremos estas técnicas en siguientes apartados.

Detectores con diferentes entradas

Nos centramos ahora en el caso de fusión de detectores en el que cada uno utiliza como entrada un conjunto diferente de observaciones o características $\boldsymbol{x}_i$,

proporcionando una salida *soft* dada por una estimación de la probabilidad a posteriori $P(H_1|\boldsymbol{x}_i)$.

Bajo este escenario, en [17] se realiza un estudio de la combinación óptima de clasificadores considerando que los vectores de observación de entrada $\boldsymbol{x}_i$ son independientes. Denominan como regla del producto a la regla de combinación optima obtenida. Mediante una serie de estrictas asunciones y aproximaciones derivan o justifican el uso de otras reglas de combinación como son la regla de la media, las reglas del máximo o del mínimo, la mediana o una regla de fusión hard como es la regla de la votación por mayoría. En [75] se continúa el estudio, justificando el uso de una regla de fusión dada por una suma ponderada en este escenario.

A continuación particularizaremos el estudio realizado en [17] para el caso de la detección, el cual se puede considerar como un problema de clasificación en el que se disponen de dos clases, cada una asociada a cada una de las dos posibles hipótesis, H_1 o H_0. Posteriormente analizaremos el caso en el que las entradas $\boldsymbol{x}_i$ no sean independientes entre sí y observaremos la influencia que tiene la dependencia en la combinación de los scores. Se mostrará cómo, a causa de la dependencia estadística entre las entradas, reglas de fusión como la media pueden llegar a ser mejores que la regla del producto. Así, rebatimos los enrevesados razonamientos que se hacen en [17] para justificar este hecho, a causa de asumir la regla de la media como una regla derivada de la del producto bajo estrictas condiciones. Dependiendo de las *PDFs* y la dependencia estadística que presenten los datos, las zonas óptimas de separación entre hipótesis, tanto en el espacio de las observaciones como en el de los scores, pueden ser muy variadas; así, cualquier tipo de función que proporcione una separación próxima a la óptima puede ser una regla de fusión valida, sin intentar buscarle un sentido de combinación estrictamente probabilístico a la función. La validez de las técnicas depende de lo que en algunos trabajos denominan como "geometría de los datos".

- *Detectores con diferentes entradas independientes*

Consideramos la fusión de un conjunto de detectores, donde cada uno trabaja con un vector de observaciones $\boldsymbol{x}_i$ diferente e independiente bajo ambas hipótesis del resto. El score óptimo que combina todos los vectores de observaciones viene dado por la probabilidad a posteriori del vector de observaciones conjunto $\boldsymbol{x} = [\boldsymbol{x}_1 \dots \boldsymbol{x}_d]$ bajo la hipótesis H_1. Podemos relacionar este score con la relación de verosimilitud $\Lambda(\boldsymbol{x})$:

$$s_{fus} = P(H_1|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|H_1)P_{H_1}}{f(\boldsymbol{x}|H_1)P_{H_1} + f(\boldsymbol{x}|H_0)P_{H_0}} = \frac{\frac{f(\boldsymbol{x}|H_1)}{f(\boldsymbol{x}|H_0)}k}{\frac{f(\boldsymbol{x}|H_1)}{f(\boldsymbol{x}|H_0)}k + 1} = \frac{k \cdot \Lambda(\boldsymbol{x})}{1 + k \cdot \Lambda(\boldsymbol{x})} \qquad (3.4)$$

donde se define la constante k como la relación entre las probabilidades a priori $k = \frac{P_{H_1}}{P_{H_0}}$.

Los scores individuales s_i de cada uno de los detectores también pueden relacionarse de igual forma con su relación de verosimilitud $\Lambda(\boldsymbol{x_i})$:

$$s_i = P(H_1|\boldsymbol{x_i}) = \frac{k \cdot \Lambda(\boldsymbol{x_i})}{1 + k \cdot \Lambda(\boldsymbol{x_i})} \rightarrow \Lambda(\boldsymbol{x_i}) = \frac{s_i}{k \cdot (1 - s_i)} \tag{3.5}$$

Dado que existe independencia entre los vectores $\boldsymbol{x_i}$ bajo ambas hipótesis se puede expresar de forma fácil la relación de verosimilitud de la fusión óptima de los vectores de muestras $\Lambda(\boldsymbol{x})$ en base a las relaciones de verosimilitud individuales de los detectores $\Lambda(\boldsymbol{x_i})$:

$$\Lambda(\boldsymbol{x}) = \frac{f(\boldsymbol{x}|H_1)}{f(\boldsymbol{x}|H_0)} = \frac{\prod_{i=1}^{d} f(\boldsymbol{x_i}|H_1)}{\prod_{i=1}^{d} f(\boldsymbol{x_i}|H_0)} = \prod_{i=1}^{d} \Lambda(\boldsymbol{x_i}) = k^{-d} \cdot \prod_{i=1}^{d} \frac{s_i}{1 - s_i} \tag{3.6}$$

Mediante 3.4 y 3.6 podemos encontrar la expresión que relaciona el score de fusión óptimo bajo la asunción de independencia de los vectores de muestras $\boldsymbol{x_i}$ con los scores individuales s_i que aportan los detectores individuales:

$$s_{fus} = \frac{k^{-(d-1)} \cdot \prod_{i=1}^{d} s_i}{\prod_{i=1}^{d}(1 - s_i) + k^{-(d-1)} \cdot \prod_{i=1}^{d} s_i}, \qquad k = \frac{P_{H_1}}{P_{H_0}} \tag{3.7}$$

Consideremos el caso de dos detectores y obtengamos la región de separación entre hipótesis en el espacio de los scores. Para ello, fijamos el valor del umbral λ:

$$\frac{k^{-1} \cdot s_1 s_2}{(1 - s_1)(1 - s_2) + k^{-1} \cdot s_1 s_2} = \lambda \leftrightarrow \; s_2 = \frac{\lambda(1 - s_1)}{\lambda + s_1\big(k - \lambda(1 + k)\big)} \tag{3.8}$$

Representamos en la figura 3.3 las fronteras de separación para distintos valores del umbral λ (en este caso λ = 0.99, 0.95, 0.9, 0.8, 0.6, 0.4 y 0.3), con hipótesis equiprobables ($P(H_1) = P(H_0) = 0.5$). Se observa como las regiones de separación entre hipótesis siguen el razonamiento lógico que a priori podríamos tener si pensamos en detectores independientes: se escoge como hipótesis verdadera regiones donde ambos detectores aportan scores elevados, donde si uno de ellos aporta valores cada vez menores (esta menos seguro), el otro debe aportarlos cada vez mayores (compensar la incerteza del otro detector teniendo más seguridad en su decisión). En la figura 3.4 podemos ver como varían las regiones de separación asociadas a un valor en concreto del umbral λ conforme las probabilidades a priori de cada hipótesis cambian ($P(H_1) = 0.9$ y $P(H_1) = 0.1$).

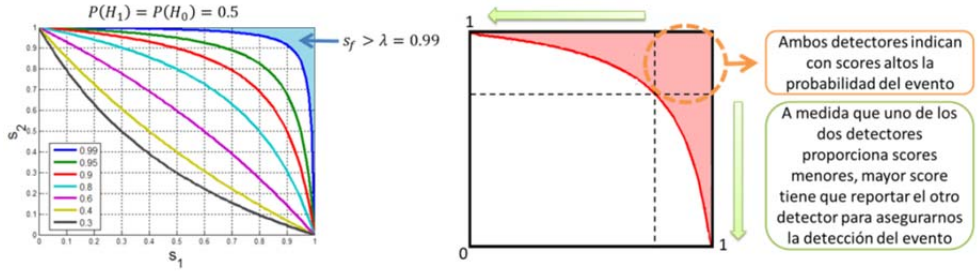


Figura 3.3 – Regiones de separación en el espacio de los scores mediante la combinación de dos scores dados por la probabilidad a posteriori de 2 vectores de observación diferentes considerados independientes

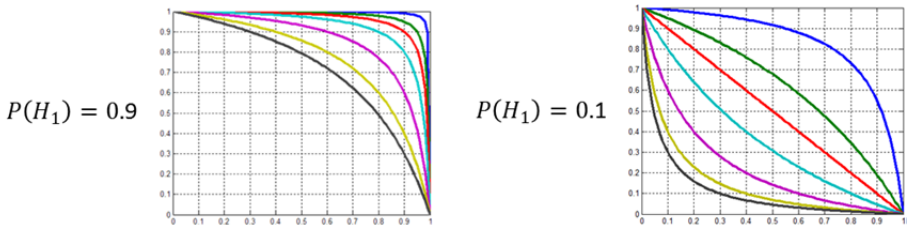


Figura 3.4 – Modificación de las regiones de decisión asociadas a cada valor del umbral λ con respecto a los priors $P(H_0)$ y $P(H_1)$

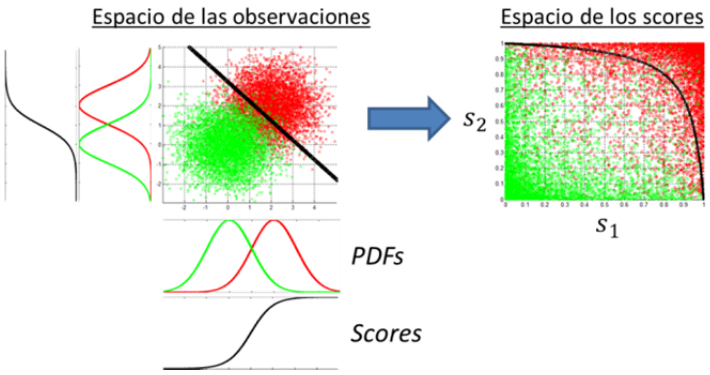


Figura 3.5 – Ejemplo de combinación de scores provenientes de distribuciones Gaussianas independientes bajo ambas hipótesis

En la figura 3.5 se muestra el mismo ejemplo que se presentó en el apartado 1.3, en el que tenemos un vector de observaciones $\boldsymbol{x}$ bivariante, con distribuciones marginales Gaussianas independientes entre sí bajo cada una de las hipótesis. Observamos el proceso de transformación de la región de separación que impone la regla de fusión óptima en el espacio de las observaciones al pasar al espacio de los

scores dados por las probabilidades a posteriori. En ambos dominios las reglas de fusión obtenidas son óptimas.

$$x_{fus} = a \cdot x_1 + b \cdot x_2 \quad \Leftrightarrow \quad s_{fus} = \frac{k^{-1} \cdot s_1 s_2}{(1-s_1)(1-s_2) + k^{-1} \cdot s_1 s_2} \tag{3.9}$$

❖ *Detectores con diferentes entradas dependientes*

Supongamos ahora que existe dependencia estadística bajo una o ambas hipótesis entre las entradas de los detectores. Atendiendo al teorema de Sklar's [44], se pueden expresar las *PDFs* conjuntas $f(\boldsymbol{x}|H_j)$ como el producto de la distribución considerando independencia por una función de densidad de copula $c(\boldsymbol{x}|H_1)$; así:

$$\Lambda_{dep}(\boldsymbol{x} = [\boldsymbol{x}_1 \dots \boldsymbol{x}_d]) = \frac{\left(\prod_{i=1}^{d} f(\boldsymbol{x}_i|H_1)\right) c(\boldsymbol{x}|H_1) P(H_1)}{\left(\prod_{i=1}^{d} f(\boldsymbol{x}_i|H_0)\right) c(\boldsymbol{x}|H_0) P(H_0)} \tag{3.10}$$

Si obtenemos el score óptimo observamos como ahora ya no sólo depende de los scores individuales de cada detector de forma aislada, sino que existe un término $k(\boldsymbol{x})$, que representa la información de interdependencia y que modifica las regiones de separación del caso independiente.

$$s_f = \frac{\Lambda_{dep}(\boldsymbol{x})}{1 + \Lambda_{dep}(\boldsymbol{x})} = \frac{k(\boldsymbol{x}) \cdot \prod_{i=1}^{d} s_i}{\prod_{i=1}^{d}(1-s_i) + k(\boldsymbol{x}) \cdot \prod_{i=1}^{d} s_i} \lessgtr_{H_1}^{H_0} \lambda$$
$$k(\boldsymbol{x}) = \frac{c(\boldsymbol{x}|H_1)}{c(\boldsymbol{x}|H_0)} \left(\frac{P(H_0)}{P(H_1)}\right)^{d-1} \tag{3.11}$$

En la figura 3.6 continuamos con el mismo ejemplo que el apartado anterior, pero en este caso mostramos dos casos con diferente dependencia bajo la hipótesis H_1, caracterizada por el factor de correlacion ρ_{H_1}. En cada uno de los ejemplos representamos, a la izquierda la región de separación óptima entre hipótesis en el espacio de observaciones y su equivalente en el espacio de scores.

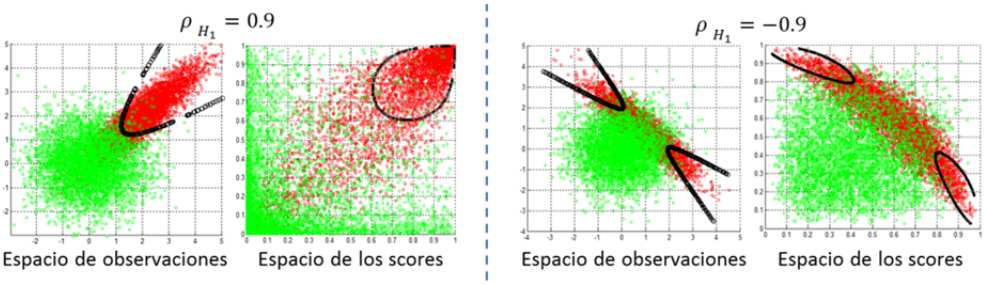


Figura 3.6 – Diferentes zonas de separación óptima entre hipótesis en el espacio de observaciones y de scores

Observamos como al modificarse el comportamiento que tienen los detectores bajo el caso de independencia, la forma de la región de separación óptima cambia. Podemos observar como en el primer caso, al existir una gran dependencia positiva bajo la hipótesis H_1, provoca que ambos detectores tiendan a aportar scores elevados de forma conjunta bajo esta hipótesis. Justamente lo contrario pasa en el ejemplo de dependencia bajo H_1 negativa, donde, si uno de los scores es muy elevado bajo H_1 el otro tiende a ser muy pequeño.

La dependencia estadística puede complicar mucho la obtención de una regla de fusión óptima que presente tales regiones de separación entre hipótesis. Existen otras reglas de fusión subóptimas, con las que se pueden obtener muy buenas prestaciones, como puede ser por ejemplo la fusión mediente la media de scores en el primer caso, o la regla del máximo en el segundo (figura 3.7).

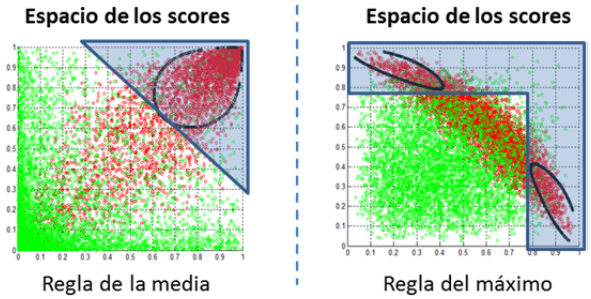


Figura 3.7 –Comparación entre zonas de separación óptima (negro) y subóptimas (azul) entre hipótesis en el espacio de scores

3.3. – Técnicas simples de fusión de scores

Pasamos ahora a comentar una serie de técnicas básicas de fusión de scores, muy extendidas y utilizadas en multitud de trabajos. Como veremos podremos aglutinar todas estas técnicas básicas en una sola, utilizando para ello una función denominada media-α y que describiremos en el siguiente apartado.

3.3.1. – Media aritmética, geométrica y harmónica

En el caso de fusión de varios detectores con una misma entrada común, donde cada uno de ellos aporta una estimación diferente de una misma probabilidad a posteriori $P(H_1|\boldsymbol{x}) \approx \hat{P}_i(H_1|\boldsymbol{x})$, en algunos estudios se justifica su fusión mediante una función que realice una media de estas valoraciones. Se suelen utilizar la media aritmética, geométrica y harmónica:

$$s_{fus} = \frac{1}{d}\sum_{i=1}^{d} s_i \qquad s_{fus} = \left(\prod_{i=1}^{d} s_i\right)^{1/d} \qquad s_{fus} = \frac{d}{\sum_{i=1}^{d}\frac{1}{s_i}} \quad , \qquad s_i,\ s_{fus}, \in [0,1] \tag{3.12}$$

Al utilizar este tipo de medias se considera que los datos bajo ambas hipótesis son separables mediante unas hipersuperficies establecidas. Por ejemplo, en el caso de la media la separación lograda mediante la fusión equivale a una división de las hipótesis mediante un hiperplano que corta cada uno de los dejes del espacio multidimensional ortogonal en el que se pueden situar los scores $\boldsymbol{s} = [s_1 \dots s_d]$ a 45^0. En la figura 3.8 podemos ver la separación entre hipótesis lograda en el caso bidimensional:

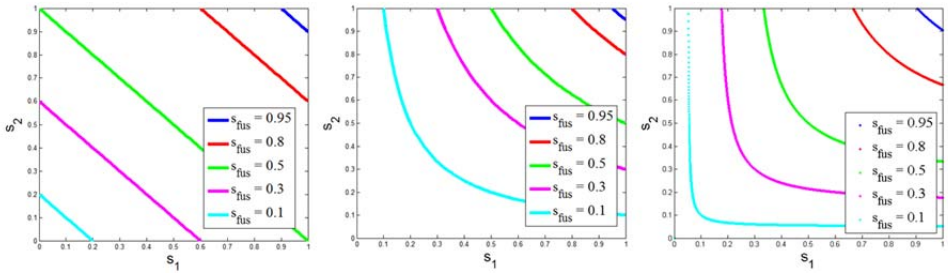


Figura 3.8 – Regiones de separación entre hipótesis en el espacio de los scores obtenidas mediante la media aritmética, geométrica y harmónica de izquierda a derecha respectivamente

3.3.2. – Reglas de combinación: Mínimo, máximo

Otro tipo de reglas de combinación de scores son las reglas del mínimo y del máximo, aunque realmente poseen una esencia hard, pues podemos ver que se tratan de reglas *AND* (se considera que se ha producido el evento cuando todos los detectores deciden H_1) *y OR* (se considera que se ha producido el evento cuando al menos un detector decide H_1) tras umbralizar cada uno de los scores utilizando un mismo valor umbral:

$$s_{fus} = \max(s_i) \quad s_{fus} = \min(s_i) \tag{3.13}$$

En la figura 3.9 se representa el caso bivariante, donde se puede apreciar con gran claridad la naturaleza hard de estas reglas de combinación.

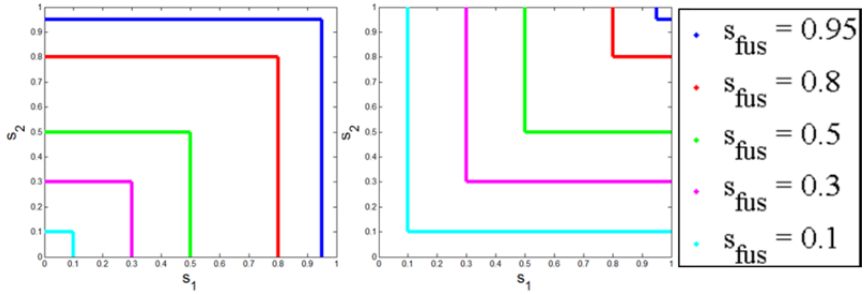


Figura 3.9 – Regiones de separación entre hipótesis en el espacio de los scores obtenidas mediante las reglas del máximo (izquierda) y del mínimo (derecha)

3.3.3. – Combinación lineal: suma y producto ponderados

En el apartado 2.2 ya introdujimos las técnicas de combinación lineal para la fusión de información soft. Como ya se comentó es común la normalización de los datos antes de la aplicación de algunas de estas técnicas. El hecho de poseer la información normalizada el rango común $[0,1]$ nos proporciona una serie de ventajas a la hora de poder utilizar diversas técnicas de entrenamiento para adaptar de forma automática los pesos de la ponderación al conjunto de datos que estemos utilizando.

$$s_{fus} = \sum_{i=1}^{d} w_i \cdot s_i = \boldsymbol{w}^T \boldsymbol{s} \qquad s_{fus} = \prod_{i=1}^{d} {s_i}^{w_i}$$
$$s_i,\ s_{fus}, \in [0,1] \quad \Leftrightarrow \quad \sum_{i=1}^{d} w_i = 1\,, \qquad 0 \leq w_i \leq 1 \tag{3.14}$$

Se justifica la utilización de una serie de pesos w_i que ponderan los datos a integrar, cuando se considera que existen canales de datos que aportan mayor información para discriminar el evento a detectar que otros.

Dependiendo de las ponderaciones la forma de las regiones de separación que se consiguen mediante la fusión de datos variarán, por lo tanto las ponderaciones aportan un grado de libertad, pudiendo ajustarse a diferentes problemas para mejorar las prestaciones obtenidas. Por lo tanto se necesitará una etapa de entrenamiento para fijar el valor de las ponderaciones. En muchas aplicaciones se escoge el valor de las ponderaciones de forma empírica, mediante ensayo y error. También podemos encontrar trabajos en los que se utilizan ciertas técnicas para poder fijar de forma automática el valor de estos parámetros.

Lo más extendido es el uso de la suma ponderada, por ello nos centraremos en ella para ilustrar las principales técnicas utilizadas en la literatura para el entrenamiento de estas ponderaciones. Lo realizamos a modo de introducción, para dar una visión del estado del arte, y poner así en contexto los beneficios que supondrá el uso de la técnica de fusión mediante una función de integración f_α y el entrenamiento que de ella proponemos. Esta técnica, que constituye la principal novedad introducida en la presente tesis en aplicaciones de fusión de scores para aplicaciones de detección, será introducida en el siguiente apartado.

Suma ponderada

Para la obtención de las ponderaciones $\boldsymbol{w}$ nos valdremos de un conjunto etiquetado de N vectores de scores de entrenamiento $S = \left\{\boldsymbol{s}^{(n)}, h^{(n)}\right\}_{n=1}^{N}$, donde $h^{(n)}$ es la etiqueta asignada a cada muestra en el instante n, siendo $h^{(n)} = 1$ si la muestra pertenece a hipótesis H_1 y $h^{(n)} = 0$ si pertenece a H_0. La minimización del error de

mínimos cuadrados y la maximización de la curva ROC son dos métodos que se han utilizado en la literatura para el entrenamiento del modelo.

- Entrenamiento mediante minimización del error de mínimos cuadrados

Una posibilidad para el aprendizaje de los pesos $\boldsymbol{w}$ pasa por el uso del criterio de mínimos cuadrados (*LSE*, "*Least Mean Squares*") [78]. Un score $s_i \in [0,1]$ con buena capacidad de discriminación entre las hipótesis se presupone con una *PDF* bajo H_1 orientada hacia valores próximos a la unidad, al contrario que bajo H_0, donde lo hará hacia valores cercanos a cero.

Con ayuda de los datos de entrenamiento $\mathcal{T} = \{\boldsymbol{s}^{(n)}, h^{(n)}\}_{n=1}^{N}$, se define la función error que mide el error cuadrático de los scores obtenidos mediante la fusión con una suma ponderada. Así, bajo la hipótesis H_1, $h^{(n)} = 1$ y el error introducido por $s_{fus} = \boldsymbol{w}^T\boldsymbol{s}$ será menor cuando más cercano esté del valor unidad. Bajo H_0, $h^{(n)} = 0$ y error será mayor cuanto más alejado de cero se encuentre $s_{fus} = \boldsymbol{w}^T\boldsymbol{s}$:

$$J = \frac{1}{2}\sum_{n=1}^{N}\left(h^{(n)} - \boldsymbol{w}^T\boldsymbol{s}\right)^2 + \frac{\eta}{2}\boldsymbol{w}^T\boldsymbol{w} = \frac{1}{2}\|\boldsymbol{h} - \boldsymbol{S}\cdot\boldsymbol{w}\|^2 + \frac{\eta}{2}\|\boldsymbol{w}\|^2$$
$$\boldsymbol{S}_{Nxd} = \begin{bmatrix}\boldsymbol{s}^{(1)}\\ \dots \\ \boldsymbol{s}^{(N)}\end{bmatrix} \tag{3.15}$$

donde el término $\frac{\eta}{2}\|\boldsymbol{w}\|^2$ se introduce para regularizar la expresión y estabilizar la solución cerrada que se obtiene del problema de minimización del error $\min_{\boldsymbol{w}}\left(J(\boldsymbol{w})\right)$ (observamos como conduce a la adición de un pequeño valor η a la diagonal de la matriz $\boldsymbol{S^TS}$, la cual debe invertirse):

$$\boldsymbol{w}_{LSE} = (\boldsymbol{S^TS} + \eta\boldsymbol{I})^{-1}\boldsymbol{S^Th} \tag{3.16}$$

Así, utilizando el conjunto de pesos $\boldsymbol{w}_{LSE}$ que minimiza el error cuadrático se intenta que el score fusionado mediante la suma ponderada presente unas distribuciones bajo cada hipótesis lo más separadas posibles. El principal inconveniente es que la minimización del error cuadrático no garantiza que el resultado obtenido sea el mejor que se puede conseguir mediante esta técnica; puede existir un score fusionado con una determinada ponderación $\boldsymbol{w} \neq \boldsymbol{w}_{LSE}$ que, pese a tener un valor de *LSE* muchísimo mayor, presente mejores prestaciones en detección. En la figura 3.10 se muestra un ejemplo que ilustra este hecho.

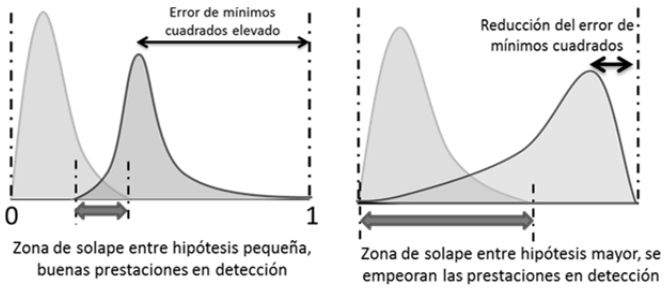


Figura 3.10 – Ejemplo ilustrativo de que el criterio de minimizar el error cuadrático no siempre tiende a obtener las mejores prestaciones en detección

- Entrenamiento mediante la maximización del área bajo la curva *ROC*

Otro criterio usado para el entrenamiento del modelo de suma ponderada de scores es el de maximización del área bajo la curva ROC (*AUC*) [25]. El *AUC* suele ser una figura de mérito muy utilizada para evaluar y comparar el rendimiento de clasificadores binarios. Para obtener la *AUC* es necesaria la computación de una integral, pero se puede utilizar el estadístico Wilcoxon-Mann-Whitney (*WMW*) como forma no paramétrica de estimar el *AUC*. Dividimos el conjunto de scores de entrenamiento $\mathcal{S}$ en dos subconjuntos $\mathcal{S}_{H_1}$ y $\mathcal{S}_{H_0}$ que recogen los valores de los scores bajo H_1 y H_0 respectivamente:

$$\begin{gathered}\mathcal{S} = \left\{\boldsymbol{s}^{(n)}, {s_f}^{(n)}\right\}_{n=1}^{N} = \mathcal{S}_{H_1} \cup \mathcal{S}_{H_0} \\ \mathcal{S}_{H_1} = \left\{\boldsymbol{s}^{(n_1)}, {s_f}^{(n_1)}\right\}_{n_1=1}^{N_1} \quad \mathcal{S}_{H_0} = \left\{\boldsymbol{s}^{(n_0)}, {s_f}^{(n_0)}\right\}_{n_0=1}^{N_0}\end{gathered} \tag{3.17}$$

El estadístico *WMW* está dado por:

$$\begin{gathered}AUC \approx \frac{1}{N_1 N_0}\sum_{n_1=1}^{N_1}\sum_{n_0=1}^{N_0}\mathbb{I}\left({s_f}^{(n_1)} > {s_f}^{(n_0)}\right) = \frac{1}{N_1 N_0}\sum_{n_1=1}^{N_1}\sum_{n_0=1}^{N_0}\mathcal{U}\left(\mathcal{E}_{n_1 n_0}\right) \\ \mathbb{I}(Expresion) = \begin{cases}1 & Expresion\ es\ verdadera \\ 0 & Expresion\ es\ falsa\end{cases} \\ \mathcal{E}_{n_1 n_0} = {s_f}^{(n_1)} - {s_f}^{(n_0)} \qquad \mathcal{U}(\mathcal{E}) = \begin{cases}1 & ,\mathcal{E} > 0 \\ 0 & ,\mathcal{E} \leq 0\end{cases}\end{gathered} \tag{3.18}$$

El entrenamiento se basa en la maximización da la *AUC*:

$$\underset{\boldsymbol{w}}{argmax}(AUC) \approx \underset{\boldsymbol{w}}{argmax}\left(\frac{1}{N_1 N_0}\sum_{n_1=1}^{N_1}\sum_{n_0=1}^{N_0}\mathcal{U}\left(\mathcal{E}_{n_1 n_0}\right)\right) \tag{3.19}$$

Para poder obtener una solución analítica al problema de maximización, ya que la función escalón $\mathcal{U}(\cdot)$ no es diferenciable, se utilizan funciones continuas $\phi(\cdot)$ para aproximarla. En [25] se propone el uso de una función cuadrática:

$$\phi(\mathcal{E}) = (\mathcal{E} + \delta)^2, \qquad \delta \in \mathbb{R} \tag{3.20}$$

Su inclusión en la estimación del *AUC* deriva en una solución para las ponderaciones con una expresión cerrada:

$$\boldsymbol{w} = \left(\eta \boldsymbol{I} + \frac{1}{N_1 N_0} \sum_{n_1=1}^{N_1} \sum_{n_0=1}^{N_0} \boldsymbol{\mathcal{E}}^T \boldsymbol{\mathcal{E}}\right)^{-1} \left(-\frac{\delta}{N_1 N_0} \sum_{n_1=1}^{N_1} \sum_{n_0=1}^{N_0} \boldsymbol{\mathcal{E}}^T\right), \quad \boldsymbol{\mathcal{E}} = \boldsymbol{s}^{(n_0)} - \boldsymbol{s}^{(n_1)} \tag{3.21}$$

En una gran cantidad de aplicaciones se debe diseñar el detector para trabajar en un cierto punto o rango de falsa alarma limitado o restringido. Por ejemplo, en aplicaciones de autentificación biométrica, los falsos positivos son prácticamente intolerables, limitando el uso de detectores bajo probabilidades de falsa alarma por debajo de 10^{-4}. En la aplicación de detección de fraudes con operaciones de tarjeta en la que ha trabajado el *GTS*, todas las operaciones categorizadas como fraudes deben ser comprobadas por un limitado número de personas que trabajan en el banco, por lo que existe un límite en el número de operaciones que por día pueden ser procesadas; esto se traduce en que se debe trabajar en un rango de falsa alarma controlado. En estos casos una ponderación $\boldsymbol{w}_{AUC}$ que maximice la curva ROC completa puede que no se traduzca en la obtención de los mejores resultados posibles en la zona de interés (figura 3.11).

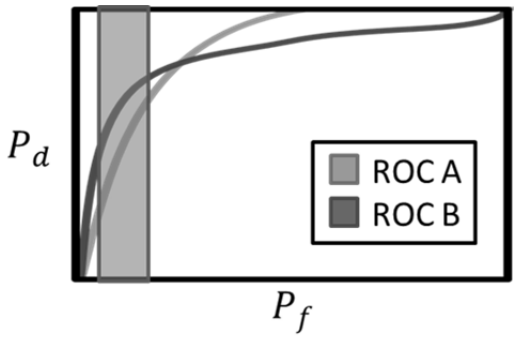


Figura 3.11 – Ejemplo donde, pese que la curva ROC A posea un AUC mayor que la curva ROC B, no se traduce en unas mejores prestaciones en un cierto área de interés

3.4. – Función de integración α en fusión de scores

Todas las técnicas de fusión vistas hasta ahora consideran que los datos son separables mediante un tipo predefinido de hipersuperficie continua. Las zonas de separación entre hipótesis son fijas, tanto en el caso de las medias aritmética, geométrica y armónica, como en el caso de las reglas del mínimo y del máximo. Usando las técnicas de suma ponderada y el producto ponderado existe un cierto grado de libertad mediante la variación de los pesos de las ponderaciones.

Consideramos todas estas técnicas de forma general como subóptimas, ya que sus prestaciones dependen de la "geometría de los datos", es decir, de cómo se distribuyan los datos en el espacio de los scores (por lo tanto, dependen de las *PDFs* marginales y de la dependencia estadística que presenten los datos). En el caso de la suma y el producto ponderado, al poder parametrizar la fusión mediante los pesos de la ponderación, podremos entrenar la fusión para que se adapte mejor a las características de los datos y así obtener mejores prestaciones. Aun así, las regiones de separación entre hipótesis resultantes, aun considerando todo el rango de ponderaciones posible, siguen siendo muy rígidas; además, como ya introdujimos, las técnicas de entrenamiento utilizadas pueden conducir a ponderaciones que no sean capaces de conseguir los mejores resultados posibles en el contexto de la detección. Otro inconveniente es el componente empírico que supone su utilización, ya que debemos implementar y testear cada una de estas técnicas para poder escoger, ante un determinado problema, la que mejores resultados nos aporte.

En el presente trabajo introducimos una novedosa técnica de fusión que hemos denominado integración-α, la cual aporta un mayor grado de flexibilidad y de adaptación, siendo capaz de mejorar las prestaciones que se pueden obtener con respecto al resto de técnicas de las que hemos hablado.

Todas las técnicas anteriormente mencionadas pueden ser consideradas como casos particulares de la integración-α, por lo tanto nos permite salvar el problema de tener que testear todas ellas en busca de la que mejor prestaciones proporcione. Se definirá inicialmente la técnica de fusión de scores mediante la integración α. Después introduciremos dos posibles métodos para su entrenamiento. La primera está basada en el criterio de minimización del error cuadrático. Como otra alternativa proponemos un novedoso método de entrenamiento basado en el criterio de maximización parcial del área bajo la curva *ROC.* Este nuevo criterio de entrenamiento es idóneo en el contexto del diseño de aplicaciones de detección como las tratadas en la presente tesis doctoral, donde el rango de trabajo está limitado entre ciertos valores de falsa alarma.

Definición de integración α

La integración o fusión α se basa en combinar d fuentes de información m_i para obtener un único valor m_α utilizando la denominada media-α [79]. La media-α es una función capaz de recoger toda una familia de diferentes medias según el valor un parámetro $\alpha \in]-\infty, +\infty[$. Así, con ponderaciones fijas $w_i = 1/d$, podemos ver como la media-α recoge las funciones y reglas mencionadas como casos especiales: para $\alpha = -1, 1, 3, +\infty$ ó $-\infty$ la media-α se convierte en la media aritmética, la media geométrica, la media harmónica, la regla del mínimo o la regla del máximo respectivamente. Observamos como la suma y el producto ponderado también son casos particulares de la media-α:

$$m_\alpha = f_\alpha^{-1}\left(\sum_{i=1}^{d} w_i \cdot f_\alpha(m_i)\right), \qquad f_\alpha(x) = \begin{cases} x^{\frac{1-\alpha}{2}} & , \alpha \neq 1 \\ log(x) & , \alpha = 1 \end{cases} \qquad m_i \geq 0 \tag{3.22}$$

donde m_α es el resultado de la integración de d fuentes de información m_i, cada una de ellas ponderadas por un coeficiente w_i, que cumple:

$$0 \leq w_i \leq 1, \qquad \sum_{i=1}^{d} w_i = 1 \tag{3.23}$$

Se ha demostrado en [79] que si asociamos m_i y m_α respectivamente con las *PDFs* $m_\alpha(x)$ y $m_i(x)$ de una determinada variable aleatoria *x*, entonces $m_\alpha(x)$ es la *PDF* que minimiza la función de coste $\mathbb{C}\big(m_\alpha(x)\big)$:

$$\mathbb{C}\big(m_\alpha(x)\big) = \sum_{i=1}^{d} w_i \cdot D\langle m_i(x) | m_\alpha(x)\rangle \tag{3.24}$$

donde $D\langle m_i(x) | m_\alpha(x)\rangle$ es la divergencia α [80] entre ambas PDFs.

En nuestro caso pretendemos utilizar la integración α para la fusión de scores $0 \leq s_i \leq 1$, i$= 1, \dots, d$ y así obtener un único score al que denotamos por s_α:

$$s_\alpha(\boldsymbol{s} = [s_1 \dots s_d]) = \begin{cases} \left(\sum_i w_i \cdot s_i^{\frac{1-\alpha}{2}}\right)^{\frac{2}{1-\alpha}} & , \alpha \neq 1 \\ exp\left(\sum_i w_i \cdot log(s_i)\right) & , \alpha = 1 \end{cases} \tag{3.25}$$

Entrenamiento de integración α para fusión de scores

Para la obtención de las ponderaciones $\boldsymbol{w}$ y el parámetro α nos valdremos de un conjunto etiquetado de N vectores de scores de entrenamiento $S = \left\{\boldsymbol{s}^{(n)}, h^{(n)}\right\}_{n=1}^{N}$,

donde $h^{(n)}$ es la etiqueta asignada a cada muestra en el instante n, siendo $h^{(n)} = 1$ si la muestra pertenece a hipótesis H_1 y $h^{(n)} = 0$ si pertenece a H_0.

- Entrenamiento mediante minimización del error cuadrático

En [81] se propone un método de aprendizaje de los parámetros α y $\boldsymbol{w}$ basado en el criterio de minimización de error cuadrático (*LSE*). La función objetivo a minimizar que se propone en [81] quedaría de esta forma:

$$\mathcal{J}(\alpha, \boldsymbol{w}) = \frac{1}{N}\sum_{n=1}^{N}\left(h^{(n)} - s_\alpha\left(\boldsymbol{s}^{(n)}\right)\right)^2 \in [0,1], \quad e^{(n)} = \left(h^{(n)} - s_\alpha\left(\boldsymbol{s}^{(n)}\right)\right)^2 \in [0,1] \qquad (3.26)$$

Observamos que bajo la hipótesis H_0 interesará un score lo más próximo a cero posible y bajo la hipótesis H_1 un score lo más cercano a uno. Así, podemos ver que la función objetivo introduce una medida del error $e^{(n)}$ que aporta cada score con respecto a la hipótesis correspondiente. El máximo error que introduce cada muestra es de 1, que se corresponde con el caso de que, o bajo H_0 ($h^{(n)} = 0$) proporcionemos un score $s_\alpha = 1$, o bajo H_1 ($h^{(n)} = 1$) proporcionemos un score $s_\alpha = 0$.

En muchos escenarios podemos encontrar un desbalance significativo entre el tamaño de los subconjuntos de entrenamiento pertenecientes a una u otra hipótesis (H_1 ó H_0). Para normalizar la contribución de cada hipótesis a la función de coste, independientemente del tamaño de los subconjuntos de entrenamiento, proponemos la siguiente modificación:

$$\mathcal{J} = \frac{1}{2}\left(\frac{1}{N_1}\sum_{n=1}^{N}\left(h^{(n)} - s_\alpha\left(\boldsymbol{s}^{(n)}\right)\right)^2 h^{(n)} + \frac{1}{N_0}\sum_{n=1}^{N}\left(h^{(n)} - s_\alpha\left(\boldsymbol{s}^{(n)}\right)\right)^2 \left(1 - h^{(n)}\right)\right) \qquad (3.27)$$

donde N_1 y N_0 hacen referencia al tamaño de los subconjuntos correspondientes a H_1 y H_0 respectivamente ($N = N_1 + N_0$). Así, se obtienen los errores cuadráticos de cada hipótesis por separado, minimizándose su media. Podemos considerar también la posibilidad de dar más o menos importancia a cada tipo de error: decidir H_0 cuando la hipótesis verdadera es H_1 (perdida de detección) o decidir H_1 cuando H_0 es la verdadera (falsa alarma). Así, mediante una simple modificación de la expresión 3.27 introducimos un parámetro $0 \leq \beta \leq 1$ para controlar la importancia que se le otorga a cada tipo de error:

$$\mathcal{J}(\alpha, \beta, \boldsymbol{w}) = \sum_{n=1}^{N}\left(h^{(n)} - s_\alpha\left(\boldsymbol{s}^{(n)}\right)\right)^2 \cdot c_\beta^n, \quad c_\beta^n = \left(\beta\frac{h^{(n)}}{N_1} + (1-\beta)\frac{\left(1 - h^{(n)}\right)}{N_0}\right) \qquad (3.28)$$

Observamos como para un valor de $\beta = 0.5$ ambos tipos de errores serán igualmente ponderados. Para valores $\beta > 0.5$ se pondera con mayor peso el error

cometido cuando la hipótesis verdadera es H_1, mientras que para $\beta < 0.5$ ponderamos más el error bajo la hipótesis verdadera H_0.

En este caso, es complicado encontrar una expresión cerrada para la obtención directa de los parámetros, tal y como pasaba en el caso de entrenamiento de suma ponderada mediante el mismo criterio. Se plantea en este caso un problema de optimización para minimizar el *LSE* mediante un algoritmo de descenso por gradiente. Para ello, una vez definido unos valores iniciales de los parámetros $\alpha = \alpha(0)$ y $\boldsymbol{w} = \boldsymbol{w}(0)$, se redefinen de forma iterativa con objeto de alcanzar el punto que minimiza el criterio *LSE*:

$$\begin{aligned} \alpha(l) &= \alpha(l-1) - \eta_\alpha \frac{\partial \mathcal{J}(\alpha, \boldsymbol{w})}{\partial \alpha}(l-1) \\ \boldsymbol{w}(l) &= \boldsymbol{w}(l-1) - \eta_w \frac{\partial \mathcal{J}(\alpha, \boldsymbol{w})}{\partial \boldsymbol{w}}(l-1) \end{aligned} \tag{3.29}$$

donde l hace referencia a la iteración del algoritmo en que nos encontramos y los valores η_α y η_w al salto que se realiza en la dirección del gradiente en cada parámetro.

Las componentes del gradiente $\frac{\partial \mathcal{J}}{\partial \alpha}(l-1)$ y $\frac{\partial \mathcal{J}}{\partial \boldsymbol{w}}(l-1)$ se obtienen de las expresiones 3.30 y 3.31, sustituyendo α por $\alpha(l-1)$ y $\boldsymbol{w}$ por $\boldsymbol{w}(l-1)$ donde sea necesario:

$$\begin{aligned} \frac{\partial \mathcal{J}(\alpha, \boldsymbol{w})}{\partial \alpha} &= -2 \sum_{n=1}^{N} \left(h^{(n)} - s_\alpha\left(\boldsymbol{s}^{(n)}\right) \right) \frac{\partial s_\alpha\left(\boldsymbol{s}^{(n)}\right)}{\partial \alpha} c_\beta^n \\ \frac{\partial \mathcal{J}(\alpha, \boldsymbol{w})}{\partial w_i} &= -2 \sum_{n=1}^{T} \left(h^{(n)} - s_\alpha\left(\boldsymbol{s}^{(n)}\right) \right) \frac{\partial s_\alpha\left(\boldsymbol{s}^{(n)}\right)}{\partial w_i} c_\beta^n \end{aligned} \tag{3.30}$$

Las expresiones de las derivadas parciales del score fusionado $\frac{\partial s_\alpha}{\partial \alpha}$ y $\frac{\partial s_\alpha}{\partial w_i}$ vienen dada por:

$$\begin{aligned} \frac{\partial s_\alpha(\boldsymbol{s})}{\partial \alpha} &= \frac{2 s_\alpha}{1-\alpha} \left(\frac{log(\sum_i w_i \cdot f_\alpha(s_i))}{1-\alpha} + \frac{\sum_i w_i \cdot \frac{\partial f_\alpha(s_i)}{\partial \alpha}}{\sum_i w_i \cdot f_\alpha(s_i)} \right) \\ \frac{\partial s_\alpha\left(\boldsymbol{s}^j\right)}{\partial w_i} &= \begin{cases} \frac{2}{1-\alpha} \left(\frac{s_\alpha \cdot f_\alpha(s_i)}{\sum_i w_i \cdot f_\alpha(s_i)} \right) & , \alpha \neq 1 \\ s_\alpha \cdot log(s_i) & , \ \alpha = 1 \end{cases} \end{aligned} \tag{3.31}$$

donde:

$$f_\alpha(s_i) = \begin{cases} s_i^{\frac{1-\alpha}{2}} & ,\alpha \neq 1 \\ log(s_i) & ,\alpha = 1 \end{cases} \qquad \frac{\partial f_\alpha(s_i)}{\partial \alpha} = -\frac{1}{2} log(s_i) \cdot s_i^{\frac{1-\alpha}{2}} \tag{3.32}$$

- Entrenamiento basado en la maximización del área bajo la curva ROC parcial

En problemas de detección podemos encontrarnos con ciertas peculiaridades a la hora de diseñar el sistema. En muchas aplicaciones existen ciertas limitaciones que obligan a trabajar en un cierto punto o rango de falsa alarma restringido. También podemos encontrar con una gran descompensación en las probabilidades a priori de aparición cada una de las hipótesis, y por tanto del número de datos de entrenamiento disponibles para cada una de las hipótesis.

El entrenamiento bajo el criterio LSE propuesto las contempla mediante la normalización y la posibilidad de ponderación del error introducido entre clases. Aun con las modificaciones introducidas, como ya se comentó en el apartado 3.3.3, el principal inconveniente es que la minimización del error cuadrático no garantiza que el resultado obtenido sea el óptimo en lo referente a las prestaciones que puede proporcionar la técnica de integración. Proponemos un novedoso método de entrenamiento que, teniendo en cuenta todas las peculiaridades del diseño del sistema de detección, tienda a encontrar el mejor resultado posible que la técnica de integración-α proporciona (y por ende, todas las técnicas anteriores, ya que ésta engloba a todas).

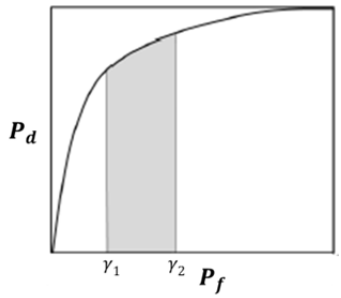


Figura 3.12 – AUC parcial en el rango de tasa de falsa alarma dado por $\gamma_1 \leq P_f \leq \gamma_2$

Se pretende encontrar el conjunto de parámetros $\upsilon^* = \{\alpha,\ \boldsymbol{w} = [w_1 \dots w_d]\}$ que maximiza el área bajo la curva *ROC* en un determinado rango de trabajo dado por $\gamma_1 \leq P_f \leq \gamma_2$ (ver figura 3.12),

$$\upsilon^* = \underset{\upsilon=\{\alpha,w\}}{\operatorname{argmax}}\left(\mathrm{AUC}_{\gamma_1}^{\gamma_2}\right) \tag{3.33}$$

sujeto a las restricciones:

$$0 \leq \mathrm{w_i} \leq 1\,, \quad \sum_{\mathrm{i}=1}^{\mathrm{d}} \mathrm{w_i} = 1 \tag{3.34}$$

Para llevar a cabo el proceso de optimización con restricciones nos valdremos de un novedoso estimador empírico y no paramétrico del área parcial bajo la curva ROC normalizada [82], presentado como una mejora del estimador que se define en [83].

Para su obtención, separamos inicialmente S en dos subconjuntos discriminando los datos según se hayan obtenido bajo una u otra hipótesis, $S = \{S_1, S_0\}$:

$$\begin{aligned} S_1 &= \left\{ \{ \boldsymbol{s}_i \rightarrow s_{\alpha_i} \} \in H_1, \quad i = 1, .., N_1 \right\} \\ S_0 &= \left\{ \{ \boldsymbol{s}_j \rightarrow s_{\alpha_j} \} \in H_0, \quad j = 1, .., N_0 \right\} \end{aligned} \tag{3.35}$$

Después, ordenamos el conjunto S_0 de mayor a menor valor del score fusionado s_α:

$$S_0^* = \left\{ s_{\alpha_j}^* > s_{\alpha_{j+1}}^* \in H_0, j = 1, .., N_0 \right\} \tag{3.36}$$

Dependiendo del número de datos de entrenamiento que bajo la hipótesis disponemos , N_0, el estimador se define como:

- Si $N_0 < \frac{1}{\gamma_2 - \gamma_1}$:

$$\widehat{nAUC}_{\gamma_1}^{\gamma_2} = \frac{1}{N_1 N_0 \cdot (\gamma_2 - \gamma_1)} \cdot \sum_{i=1}^{N_1} N_0 (\gamma_2 - \gamma_1) \cdot \mathbb{I}\left(s_{\alpha_i}^{H_1} > s_{\alpha_{j_{\gamma_1}}}^{* H_0} \right) \tag{3.37}$$

- Si $\mathrm{N_0} \geq \frac{1}{\gamma_2 - \gamma_1}$:

$$\widehat{nAUC}_{\gamma_1}^{\gamma_2} = \frac{1}{N_1 N_0 \cdot (\gamma_2 - \gamma_1)} \cdot \sum_{i=1}^{N_1} (a_1 + a_2)$$

$$a_1 = \left(j_{\gamma_1} - N_0 \gamma_1 \right) \cdot \mathbb{I}\left(s_{\alpha_i}^{H_1} > s_{\alpha_{j_{\gamma_1}}}^{* H_0} \right) + \left(N_0 \gamma_2 - j_{\gamma_2} \right) \cdot \mathbb{I}\left(s_{\alpha_i}^{H_1} > s_{\alpha_{j_{\gamma_2}+1}}^{* H_0} \right) \tag{3.38}$$

$$a_2 = \sum_{j=j_{\gamma_1}+1}^{j_{\gamma_2}} \mathbb{I}\left(s_{\alpha_i}^{H_1} > s_{\alpha_j}^{* H_0} \right)$$

donde se define el valor j_{γ_1} como el entero inmediatamente superior al valor $N_0\gamma_1$: $j_{\gamma_1} = \lceil N_0 \gamma_1 \rceil$. Se define el valor j_{γ_2} como el entero inferior a $N_0\gamma_2$: $j_{\gamma_2} = \lfloor N_0 \gamma_2 \rfloor$.

En las expresiones 3.37 y 3.38 $\mathbb{I}(\cdot)$ hace referencia a una función lógica que devuelve 1 cuando se cumple la relación y 0 cuando no. Se puede utilizar en su lugar una función escalón $\mathcal{U}(\cdot)$, definiendo para ello la variable $\varepsilon_{ij} = s_{\alpha_i}^{H_1} - s_{\alpha_j}^{*H_0} \in [0,1]$:

$$\mathbb{I}\left(s_{\alpha_i}^{H_1} > s_{\alpha_j}^{*H_0}\right) = \mathcal{U}\left(\varepsilon_{ij} = s_{\alpha_i}^{H_1} - s_{\alpha_j}^{*H_0}\right), \qquad \mathcal{U}(x) = \begin{cases} 1 & , x \geq 0 \\ 0 & , x < 0 \end{cases} \tag{3.39}$$

Para poder formular un problema de optimización, necesitamos utilizar una función continua y derivable que aproxime la función escalón. Una de las funciones más empleadas en cualquier tipo de algoritmos para la aproximación de la función escalón es una función sigmoidea [84], [85]. Como se puede observar en la figura 3.13, la función sigmoide, con un valor elevado del parámetro δ, aproxima bastante bien a la función escalón:

$$\mathbb{I}\left(s_{\alpha_i}^{H_1} > s_{\alpha_j}^{*H_0}\right) = \mathcal{U}\left(\varepsilon_{ij} = s_{\alpha_i}^{H_1} - s_{\alpha_j}^{*H_0}\right) \approx \theta(\varepsilon_{ij}) = \frac{1}{1 + e^{-\delta \cdot \varepsilon_{ij}}} \tag{3.40}$$

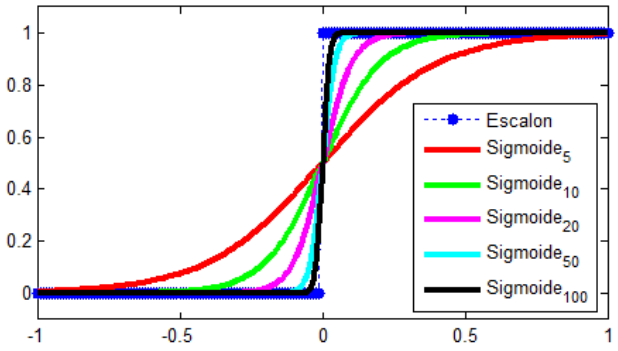


Figura 3.13 – Aproximación de la función escalón mediante una función sigmoidea con diferentes parámetros δ

Utilizando esta función sigmoidea, planteamos el problema de minimización no lineal con restricciones:

$$v^* = \underset{v=\{\alpha,\boldsymbol{w}\}}{\operatorname{argmin}}\left(g(\alpha,\boldsymbol{w}) = 1 - n\widehat{AUC}_{\gamma_1}^{\gamma_2}\right), \qquad 0 \leq w_i \leq 1\,, \quad \sum_{i=1}^{d} w_i = 1 \tag{3.41}$$

Para la resolución del problema de optimización se ha empleado el "*Optimization Toolbox*" del software de cómputo matemático MATLAB. Se ha utilizado un algoritmo de punto interior para su resolución, basado en los trabajos [86], [87].

Utilizando la función sigmoide 3.40 para aproximar la función escalón, sustituyéndola en las expresiones 3.37 y 3.38 y derivando las expresiones $\widehat{\mathrm{nAUC}}_{\gamma_1}^{\gamma_2}$ obtenidas con respecto al parámetro genérico v:

- Si $N_0 < \frac{1}{\gamma_2 - \gamma_1}$:

$$\frac{\partial}{\partial v}\left(n\widehat{AUC}_{\gamma_1}^{\gamma_2}\right) = \frac{1}{N_1 N_0 \cdot (\gamma_2 - \gamma_1)} \cdot \sum_{i=1}^{N_1} N_0(\gamma_2 - \gamma_1) \cdot \frac{\partial \theta}{\partial v}\left(s_{\alpha_i}^{H_1} - s_{\alpha_{j_{\gamma_1}}}^{* H_0}\right) \tag{3.42}$$

- Si $N_0 \geq \frac{1}{\gamma_2 - \gamma_1}$:

$$\frac{\partial}{\partial v}\left(n\widehat{AUC}_{\gamma_1}^{\gamma_2}\right) = \frac{1}{N_1 N_0 \cdot (\gamma_2 - \gamma_1)} \cdot \sum_{i=1}^{N_1} \left(\frac{\partial a_1}{\partial v} + \frac{\partial a_2}{\partial v}\right)$$

$$\frac{\partial a_1}{\partial v} = \left(\left(j_{\gamma_1} - N_0\gamma_1\right)\frac{\partial \theta}{\partial v}\left(s_{\alpha_i}^{H_1} - s_{\alpha_{j_{\gamma_1}}}^{* H_0}\right) + \left(N_0\gamma_2 - j_{\gamma_2}\right)\frac{\partial \theta}{\partial v}\left(s_{\alpha_i}^{H_1} - s_{\alpha_{j_{\gamma_2}+1}}^{* H_0}\right)\right) \tag{3.43}$$

$$\frac{\partial a_2}{\partial v} = \sum_{j=j_{\gamma_1}+1}^{j_{\gamma_2}} \frac{\partial \theta}{\partial v}\left(s_\alpha - s_{\alpha_j}^{* H_0}\right)$$

Continuando con la cadena de derivación, obtenemos la derivada parcial de la función aproximación de escalón con respecto al parámetro genérico v:

$$\frac{\partial \theta\left(\varepsilon_{ij}\right)}{\partial v} = \delta \frac{e^{-\delta \cdot \varepsilon_{ij}}}{\left(1 + e^{-\delta \cdot \varepsilon_{ij}}\right)^2} \frac{\partial \varepsilon_{ij}}{\partial v} \tag{3.44}$$

Seguimos con la derivada de la diferencia de scores fusionados ε_{ij}. Observamos que depende de la derivada parcial del score fusionado, la cual ya obtuvimos (expresiones 3.31 y 3.32):

$$\frac{\partial \varepsilon_{ij}}{\partial v} = \frac{\partial s_\alpha}{\partial v}\left(\boldsymbol{s}_i^{H_1}\right) - \frac{\partial s_\alpha}{\partial v}\left(\boldsymbol{s}_j^{* H_0}\right) \tag{3.45}$$

Por lo tanto, se puede mejorar las prestaciones del algoritmo de punto interior usado, ya que el gradiente $\Delta g(\alpha, \boldsymbol{w}) = \left(\frac{\partial}{\partial \alpha} g(\alpha, \boldsymbol{w}),\ \frac{\partial}{\partial \boldsymbol{w}} g(\alpha, \boldsymbol{w})\right)$ es conocido. Tan sólo se debe sustituir el parámetro genérico v por cada uno de los parámetros α y w_i en la cadena de derivación compuesta por las expresiones 3.42 ó 3.43, 3.44, 3.45, y 3.31 o 3.32.

Ejemplo de fusión de scores mediante la integración α

A continuación mostramos un pequeño ejemplo muy ilustrativo donde se muestra la fusión de dos detectores que aportan scores caracterizados por variables aleatorias con distribuciones uniformes bajo cada una de sus hipótesis. Con objeto de mostrar la

función de cada uno de los parámetros que caracterizan la integración α realizamos inicialmente una serie de simulaciones utilizando como método de entrenamiento la minimización del error cuadrático, fijando alguno de los parámetros. Posteriormente utilizando el nuevo entrenamiento que proponemos mostramos cómo es posible entrenar esta técnica para obtener buenas prestaciones en una determinada zona de trabajo.

Entrenamiento mediante minimización del error cuadrático

Iniciamos las simulaciones situándonos en el caso de disponer de scores aportados por dos detectores con las mismas prestaciones, por lo tanto, definimos las *PDFs* bajo cada hipótesis de ambos de igual forma: bajo H_0 siguen distribuciones uniformes entre los valores 0 y 0.8 y bajo H_1 entre los valores 0.2 y 1. Como ambos detectores poseen las mismas prestaciones fijamos de antemano las ponderaciones, $\mathrm{w}_1 = \mathrm{w}_2 = 0.5$. Consideramos que ambas hipótesis son equiprobables $P_0 = P_1 = 0.5$ y ponderamos el error de cada una de ellas de igual forma $\beta = 0.5$. Mostramos el resultado de las simulaciones mediante 6 gráficas (figura 3.14): en la primera fila se muestran un scatterplot de los datos, el valor de parámetro α con respecto a las iteraciones del algoritmo de descenso de gradiente y las curvas *ROC* de los scores individuales, y en la segunda fila ve observan las diferentes regiones de separación entre hipótesis logradas con la fusión, las distribuciones marginales de cada uno de los scores y la distribución final del score fusionado.

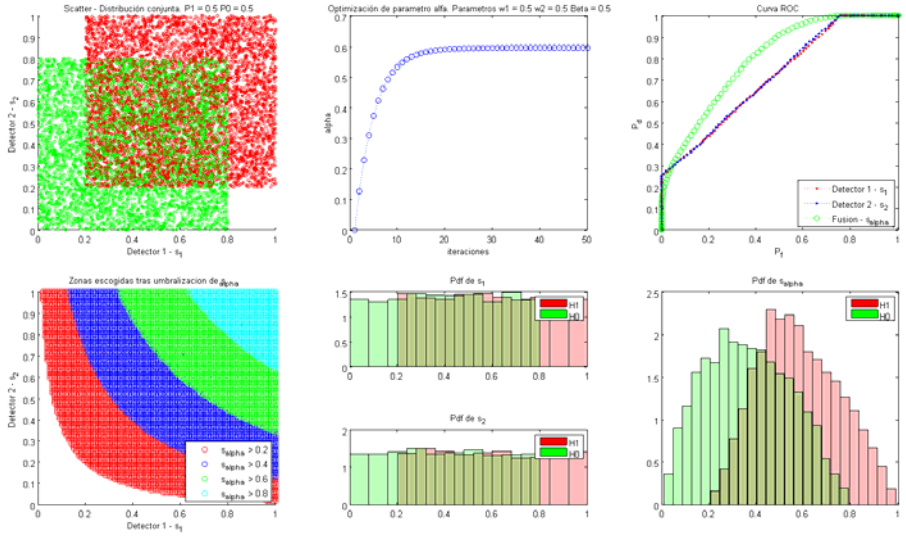


Figura 3.14 – Resultado de la simulación: Ambos detectores iguales, con $P_0 = P_1 = 0.5\,, \beta = 0.5\ y\ w_1 = w_2 = 0.5$

Observamos en la figura 3.14 como en este caso minimizando el error cuadrático mejoramos las prestaciones de los detectores, logrando una mejor separación global

entre clases; se observa perfectamente en las *PDFs* del score fusionado y en cómo se mejora la curva ROC en todo el rango de falsa alarma. Hemos comprobado que el resultado de las simulaciones obtenidas variando las probabilidades a priori P_0 y P_1 es el mismo, ya que como ya comentamos, hemos introducido una normalización en la función de error cuadrático para que el resultado no se vea afectado por la cantidad de datos que disponemos de ambas hipótesis.

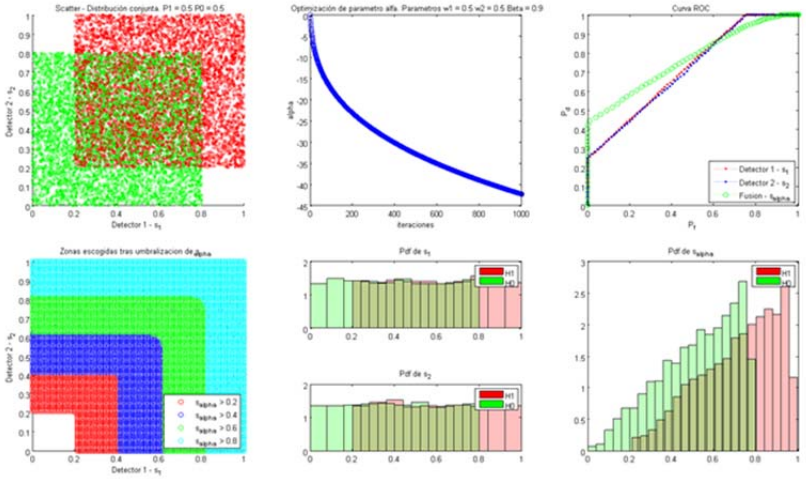


Figura 3.15 – Resultado de la simulación: Ambos detectores iguales, con $P_0 = P_1 = 0.5\,, \beta = 0.9\ y\ w_1 = w_2 = 0.5$.

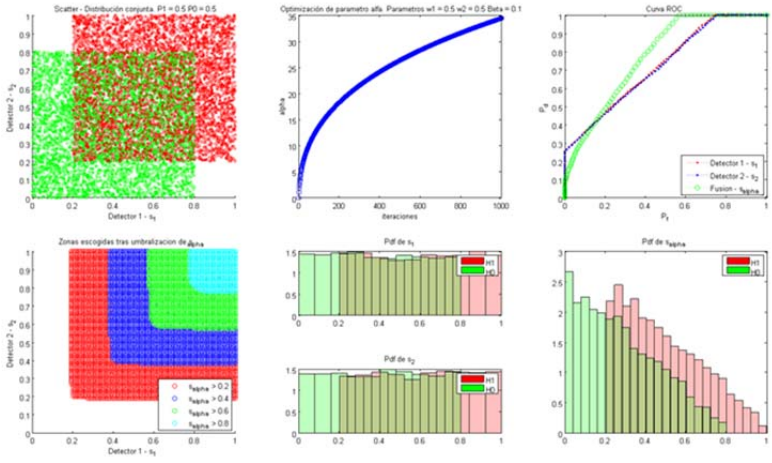


Figura 3.16 – Resultado de la simulación: Ambos detectores iguales, con $P_0 = P_1 = 0.5\,, \beta = 0.1\ y\ w_1 = w_2 = 0.5$.

Modificamos ahora el parámetro de ponderación del error que le damos a cada una de las hipótesis β. Lo ajustamos inicialmente a $\beta = 0.9$ (figura 3.15), dándole más

importancia al error cometido cuando bajo la hipótesis real H_1 se reportan valores de scores bajos que pueden caer en la zona de escoger H_0; intentamos minimizar así el error de pérdida de detección. Después lo ajustamos a $\beta = 0.1$ (figura 3.16), con objeto de ver el caso opuesto, en el que le damos más importancia al caso de error por falsa alarma.

Observamos como el parámetro α controla la "forma" de las superficies de separación entre hipótesis. Según el valor del parámetro β que controla la ponderación de cada tipo de error, el parámetro α varía la forma de las regiones de separación, pudiendo obtener así mejoras en ciertas partes de la curva *ROC*.

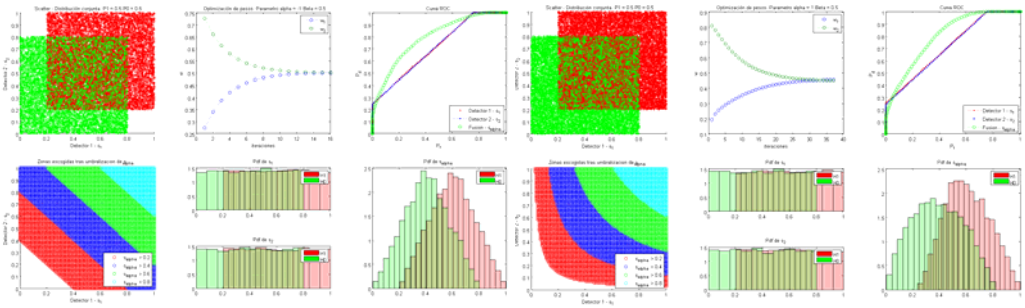


Figura 3.17 – Resultado de la simulaciones: Ambos detectores iguales, con $P_0 = P_1 = 0.5$ *,* $\beta = 0.5$*. A la izquierda simulación con* $\alpha = -1$ *(media ponderada) y a la derecha simulación con* $\alpha = 1$ *(producto ponderado)*

Ahora con objeto de ver el efecto de las ponderaciones, fijamos el parámetro α al los valor -1 para obtener una suma ponderada y a 1 para obtener un producto ponderado. Consideramos que ambas hipótesis son equiprobables $\mathrm{P_0 = P_1 = 0.5}$ y ponderamos el error de cada una de ellas de igual forma $\beta = 0.5$. Observamos en figura 3.17 como las ponderaciones convergen hacia un mismo valor $\mathrm{w_1 = w_2 = 0.5}$ en ambos casos, ya que los detectores poseen las mismas prestaciones.

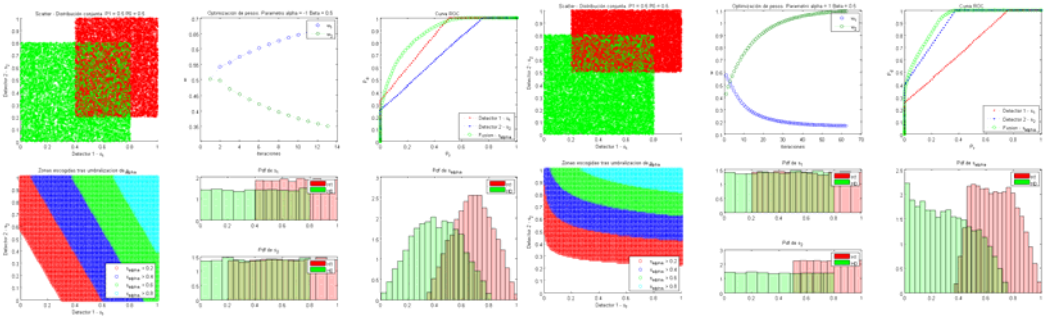


Figura 3.18 – Resultado de la simulaciones: $P_0 = P_1 = 0.5$ *,* $\beta = 0.5$*. A la izquierda simulación con* $\alpha = -1$ *(media ponderada) donde el detector 1 es mejor que el 2. A la derecha simulación con* $\alpha = 1$ *(producto ponderado) donde el detector 2 es mejor.*

Realizamos ahora la simulación modificando la distribución marginal H_1 de uno de los detectores para mejorar sus prestaciones con respecto al otro. En la figura 3.18 podemos ver como en el caso de la media ponderada (izquierda) se mejoran las prestaciones del detector 1, y por lo tanto se tiende a obtener una ponderación $w_1 > w_2$. En el caso del producto ponderado (derecha) se han mejorado las prestaciones del detector 2, por lo que se obtiene una ponderación donde $w_1 < w_2$.

Entrenamiento mediante la maximización del AUC parcial

En este caso utilizaremos el nuevo método de entrenamiento que planteamos. Empezamos igual que en el caso anterior, con detectores de iguales prestaciones y pretendemos optimizar los parámetros de la integración α de forma que se maximice el área bajo toda la curva ROC $P_f \in [0,1]$. En la figura 3.19 podemos ver que el resultado muy parecido al que se obtiene mediante la minimización del error cuadrático; observar como las distribuciones bajo H_0 y H_1 que hemos escogido son simétricas, por lo que la minimización del error cuadrático desplaza por igual ambas *PDFs* y logra una mejora global. En el caso de que las distribuciones no sean simétricas, como ya comentamos en el apartado 3.3.3 la minimización del error cuadrático puede conducir incluso a empeorar las prestaciones.

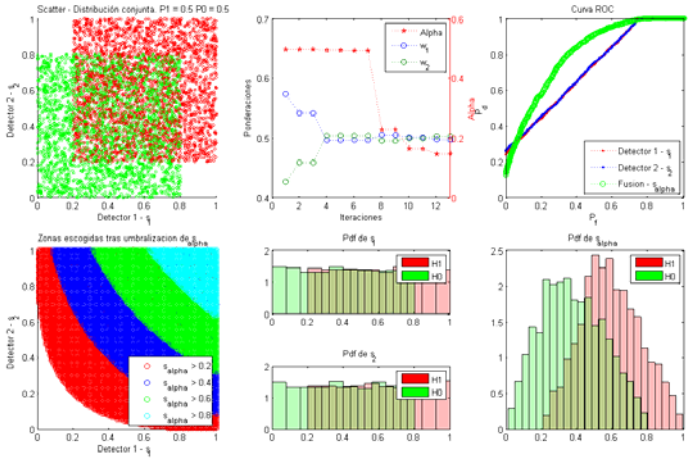


Figura 3.19 – Resultado de la simulaciones: Detectores iguales y $P_0 = P_1 = 0.5$. *Entrenamos los parámetros maximizando el área bajo la curva ROC completa*

Ahora realizamos una simulación intentando mejorar los resultados obtenidos en los rangos de trabajo $P_f \in [0,0.3]$ y $P_f \in [0.6,1]$

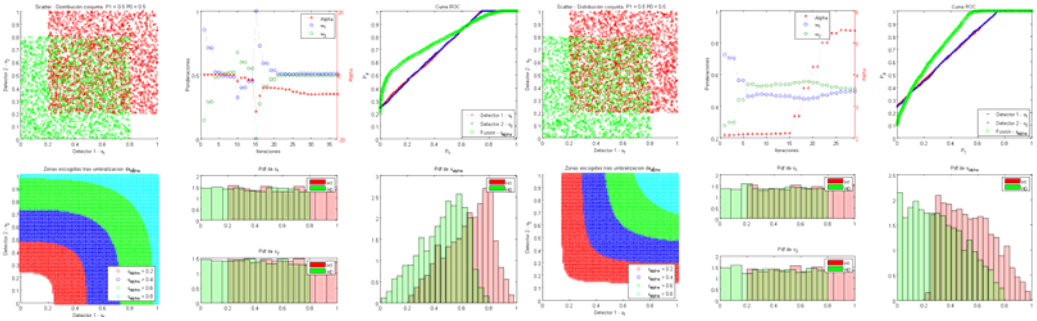


Figura 3.20 – Resultado de la simulaciones: Detectores iguales y $P_0 = P_1 = 0.5$. *Entrenamos los parámetros maximizando el área bajo la curva ROC parcial en los rangos* $P_f \in [0,0.3]$ *(izquierda) y* $P_f \in [0.6,1]$ *(derecha)*

Si intentamos maximizar el área bajo la curva *ROC* en una zona en la que los detectores individuales ya son óptimos, como es el caso de la zona con $P_f \in [0.8,1]$, podemos observar en la figura 3.21 como el proceso de optimización converge adecuadamente y escoge el primer detector asignando ponderaciones $w_1 = 0$ $w_2 = 1$.

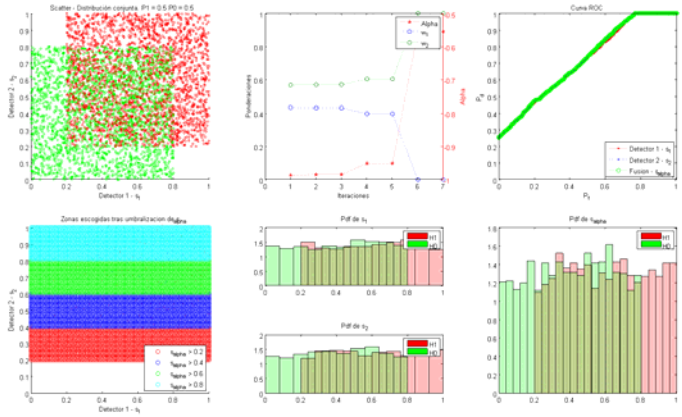


Figura 3.21 – Resultado de la simulaciones: Detectores iguales y $P_0 = P_1 = 0.5$. *Entrenamos los parámetros maximizando el área bajo la curva ROC parcial en el rango* $P_f \in [0.8,1]$

Por último obtenemos en la figura 3.22 los resultados del caso en que uno de los dos detectores es mejor que el otro, entrenado para todo el rango completo de *ROC* $P_f \in [0,1]$, y para el rango restringido $P_f \in [0,0.3]$.

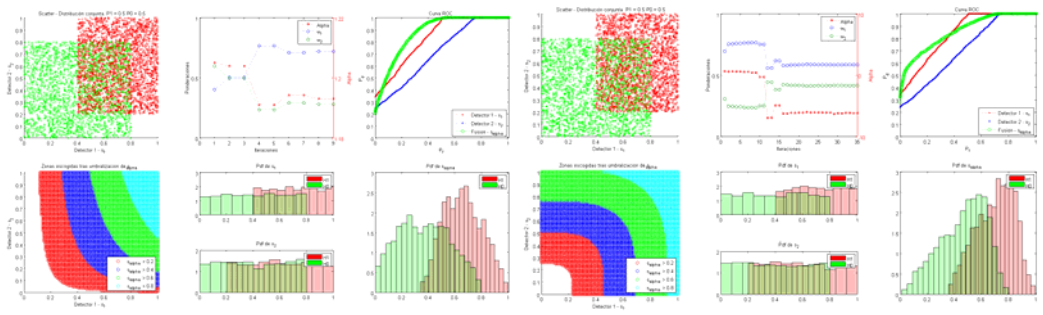


Figura 3.22 – Resultado de la simulaciones: Detector 1 mejor y $P_0 = P_1 = 0.5$*. Entrenamos los parámetros maximizando el área bajo la curva ROC completa (izquierda) y en el rango* $P_f \in [0,0.3]$ *(derecha)*

Con este sencillo ejemplo se comprueba la flexibilidad que proporciona la integración α en la fusión de scores, permitiendo adaptarse a la diferente geometría que puedan presentar los datos. Así mismo, hemos visto como el método de entrenamiento propuesto nos permite entrenar esta ponderación de una forma sencilla para obtener los mejores resultados posibles en problemas donde la tasa de falsa alarma está restringida.

3.5. – Modelo de mezcla de expertos para la fusión de scores

El modelo de mezcla de expertos (*ME, "Mixture of Experts"*) ha sido investigado, desarrollado y utilizado desde su introducción en 1991 en numerosas aplicaciones de regresión, clasificación y fusión de información. En [88] se puede encontrar un amplio estudio en el que se recopilan veinte años de investigación relacionada con esta técnica. El modelo original de mezcla de expertos fue introducido por Robert A. Jacobs en [89]:

$$P(H_j|\boldsymbol{x},\Theta) = \sum_{k=1}^{K} P(H_j,k|\boldsymbol{x},\Theta) = \sum_{k=1}^{K} g_k(\boldsymbol{x},\Theta_g)\, P(H_j|k,\boldsymbol{x},\Theta_e) \tag{3.46}$$

donde K es el número de expertos, $g_k(\boldsymbol{x},\Theta_g) = P(k|\boldsymbol{x},\Theta_g)$ representa la probabilidad de que la opinión del experto i, dada por la probabilidad a posteriori $P(H_j|k,\boldsymbol{x},\Theta_e)$, sea correcta.

El modelo de mezcla de expertos (3.46) posee una arquitectura compuesta por tres elementos (figura 3.23):

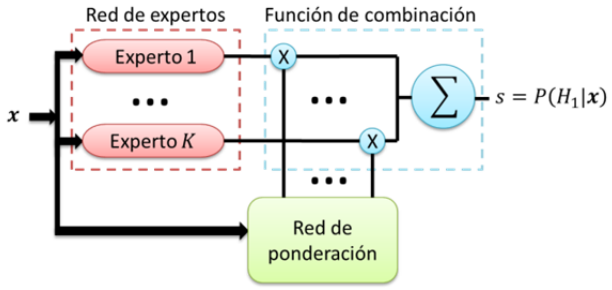


Figura 3.23. – Modelo de mezcla de expertos

- Una serie de K expertos: en el caso que nos ocupa, los expertos serán diversos detectores.
- Una red de ponderación ("Gating network") compuesta por puertas K ("*gates*"): cada una de estas puertas realiza una partición soft del espacio de observaciones, de forma que se definen diferentes regiones en base a la fiabilidad de las opiniones de cada uno de los expertos.
- Función de combinación: modelo probabilístico que combina los expertos y las puertas de ponderación, en este caso, una suma ponderada de los expertos, donde los pesos dependen de las entradas y están determinados por las puertas.

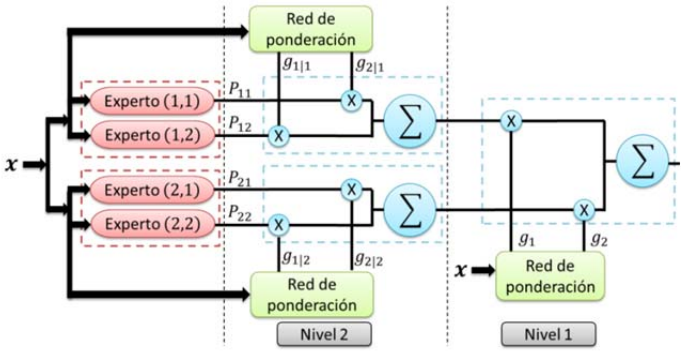


Figura 3.24. – Estructura de una mezcla de expertos jerárquica de dos niveles

La arquitectura de mezcla de expertos puede ser extendida de forma modular, dando lugar a lo que se conoce como mezcla de expertos jerárquica (*HME, "Hierarchical Mixture of Experts"*) [90], [91]. Para el caso de dos niveles jerárquicos mostrado en la figura 3.24, el modelo de mezcla de expertos es:

$$P(H_j|\boldsymbol{x},\Theta) = \sum_{k=1}^{K} g_k\left(\boldsymbol{x},\Theta_{g_k}\right)\sum_{i=1}^{I_k} g_{i|k}\left(\boldsymbol{x},\Theta_{g_{i|k}}\right)P_{ki}(H_j,\Theta_e) \tag{3.47}$$

donde, $K = 2$ es el número de nodos conectados a la red de ponderación de nivel uno, y $I_k = 2$ es el número de nodos conectados en las redes de ponderación de nivel dos, g_k es la salida k de la red de ponderación en la capa superior y $g_{i|k}$ son las salidas de las redes de ponderación del nivel dos.

En esencia, el modelo de mezcla de expertos no es más que otra forma de estimar una función de densidad de probabilidad (*PDF*) desconocida mediante una mezcla de otras *PDFs*, sólo que en este caso se realiza la mezcla en un espacio de probabilidades asociadas a estas *PDFs*. Así se define un tipo de *PDFs* para las puertas $f_g(\boldsymbol{x}, \theta_k)$ y otro para los expertos $f_e(\boldsymbol{x}, \varphi_k)$ donde θ_k y φ_k son los conjuntos de parámetros asociados a la componente *k-ésima* de la mezcla, por lo tanto $\Theta_g = \{\theta_k\}_1^K$ y $\Theta_e = \{\varphi_k\}_1^K$. Así:

$$g_k(\boldsymbol{x}, \Theta_g) = \frac{f_g(\boldsymbol{x}, \theta_k)}{\sum_{i=1}^{K} f_g(\boldsymbol{x}, \theta_i)} \quad P_k(\boldsymbol{x}, \Theta_e) = \frac{f_e(\boldsymbol{x}, \varphi_k)}{\sum_{i=1}^{K} f_e(\boldsymbol{x}, \varphi_i)} \tag{3.48}$$

Dos tipos de *PDFs* son habituales en el modelo de mezclas, una basada en la función soft-max con $\beta(\boldsymbol{x}, \theta_k)$ lineal:

$$f_g(\boldsymbol{x}, \theta_k) = e^{\beta(\boldsymbol{x}, \theta_k)}, \quad \beta(\boldsymbol{x}, \theta_k) = \boldsymbol{v}_k^T[\boldsymbol{x}, 1] \tag{3.49}$$

donde $\boldsymbol{v}_k^T$ es un vector de ponderaciones de dimensión $K + 1$. Otro tipo de *PDF* usada es la Gaussiana multivariante:

$$f_g(\boldsymbol{x}, \theta_k) = \frac{1}{(2\pi)^{N/2} \cdot |\boldsymbol{\Sigma}_k|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m}_k)^T \cdot \boldsymbol{\Sigma}_k^{-1} \cdot (\boldsymbol{x}-\boldsymbol{m}_k)}, \quad \theta_k = \{\boldsymbol{\Sigma}_k, \boldsymbol{m}_k\} \tag{3.50}$$

En el caso de fusión de información en problemas de detección que nos ocupa en el presente estudio, la técnica de mezcla de expertos puede utilizarse, tanto para la fusión de información soft, debiéndose entrenar tanto los expertos como las redes de ponderación, como para la fusión de scores (figura 3.25), en donde solamente se deben entrenar las redes de ponderación. Observamos cómo es un caso muy parecido al del modelo *GMM*, pero más complejo, añadiendo las redes de puertas.

Consideramos interesante el uso de esta arquitectura para la fusión de scores aportados por diversos detectores, donde actuarán como expertos y por lo tanto sólo deberemos entrenar las redes de ponderación. En este caso sí debemos considerar los scores como valoración aportada por un conjunto de detectores o algoritmos, puesto que la red de ponderación opera con el vector de datos soft global $\boldsymbol{x}$. Lo podemos ver como un paso más de la regla de fusión dada por la media ponderada, donde ahora las ponderaciones no son fijas, sino que dependen del vector de observaciones; así se pondera cada detector no sólo por sus prestaciones globales, sino por sus prestaciones relativas en cada zona del espacio de observaciones, lo que puede permitir mejorar mucho las prestaciones globales tras la fusión de los detectores.

$$s_{fus} = \sum_{i=1}^{d} w_i \cdot s_i \;\Rightarrow\; s_{fus} = \sum_{i=1}^{d} w_i(\boldsymbol{x}) \cdot s_i \tag{3.51}$$

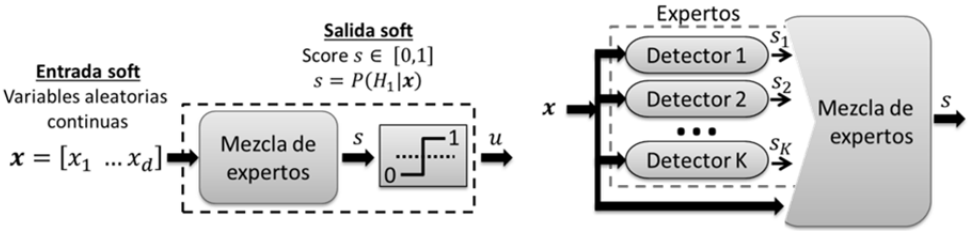


Figura 3.25 – Esquema de uso de la técnica de mezcla de expertos para la fusión de información soft y fusión de scores

Mezcla de expertos en problemas de detección

La mezcla de expertos se puede usar en problemas de regresión y de clasificación, utilizándose con una arquitectura diferente en cada uno de ellos. En la figura 3.26 se representa la arquitectura del modelo de mezcla de expertos en un problema de clasificación con C clases. La salida deseada es un vector $\boldsymbol{y} = [y_1 \dots y_c \dots y_C]$ con valor $y_c = 1$ cuando el vector de entrada $\boldsymbol{x}$ pertenezca a la clase c y el resto de valores nulos. En el caso de un problema de detección tendremos únicamente dos clases, cada una asociada a una de las hipótesis H_0 o H_1. En el apéndice E podemos encontrar el método de entrenamiento de este modelo basado en el algoritmo "Expectation-Maximization".

$$\begin{aligned} y_0 &= P(H_0|\boldsymbol{x},\Theta_0) = \sum_{k=1}^{K} g_k\left(\boldsymbol{x},\Theta_{g_0}\right) P\left(H_0|k,\boldsymbol{x},\Theta_{e_0}\right) = \sum_{k=1}^{K} g_k\left(\boldsymbol{x},\Theta_{g_0}\right) y_{k0} \\ y_1 &= P(H_1|\boldsymbol{x},\Theta_1) = \sum_{k=1}^{K} g_k\left(\boldsymbol{x},\Theta_{g_1}\right) P\left(H_1|k,\boldsymbol{x},\Theta_{e_1}\right) = \sum_{k=1}^{K} g_k\left(\boldsymbol{x},\Theta_{g_1}\right) y_{k1} \end{aligned} \tag{3.52}$$

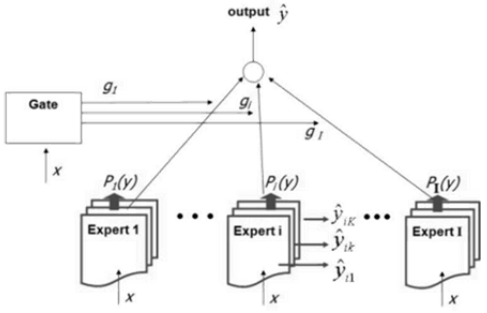


Figura 3.26. – Mezcla de expertos en clasificación

En los casos prácticos de clasificación se escoge la clase $c = arg\max_c(y_c \, \epsilon \, \boldsymbol{y})$. En detección podemos combinar ambos resultados:

$$\boldsymbol{y} = [y_0 \; y_1] \rightarrow P(H_1|\boldsymbol{x}) = \frac{P(H_1|\boldsymbol{x}, \Theta_1)}{P(H_1|\boldsymbol{x}, \Theta_1) + P(H_0|\boldsymbol{x}, \Theta_0)} = \frac{y_1}{y_1 + y_0} \tag{3.53}$$

Ejemplo

Proponemos ahora un sencillo ejemplo para ilustrar el funcionamiento de la técnica de fusión mediante una arquitectura de mezcla de expertos, con el que además se demuestra el correcto funcionamiento de nuestra implementación de esta técnica. Consideremos dos variables aleatorias $X = \{X_1, X_2\}$, cuya función densidad conjunta (figura 3.27) bajo cada una de las hipótesis H_0 y H_1 se corresponde con una mezcla de dos componentes Gaussianas bivariantes $\boldsymbol{N}_{c_1}(\cdot)$ y $\boldsymbol{N}_{c_2}(\cdot)$:

$$\begin{aligned} f(\boldsymbol{x}|H_1) &= \alpha_{H_1c_1} \cdot \boldsymbol{N}_{c_1}(\boldsymbol{\mu}_{H_1c_1}, \boldsymbol{R}_{H_1c_1}) + \alpha_{H_1c_2} \cdot \boldsymbol{N}_{c_2}(\boldsymbol{\mu}_{H_1c_2}, \boldsymbol{R}_{H_1c_2}) \\ f(\boldsymbol{x}|H_0) &= \alpha_{H_0c_1} \cdot \boldsymbol{N}_{c_1}(\boldsymbol{\mu}_{H_0c_1}, \boldsymbol{R}_{H_0c_1}) + \alpha_{H_0c_2} \cdot \boldsymbol{N}_{c_2}(\boldsymbol{\mu}_{H_0c_2}, \boldsymbol{R}_{H_0c_2}) \end{aligned} \tag{3.54}$$

donde $\alpha_{H_jc_k} > 0$ es la ponderación de la componente Gaussiana c_k bajo la hipótesis H_j (cumplen $\sum_{k=1}^{2} \alpha_{H_jc_k} = 1$); $\boldsymbol{\mu}_{H_jc_k} = \left[\mu_{1_{H_jc_k}} \; \mu_{2_{H_jc_k}}\right]$ denota el vector de medias y $\boldsymbol{R}_{H_jc_k}$ la matriz de covarianza de la componente Gaussiana c_k bajo la hipótesis H_j. En el caso presentado en este ejemplo, las matrices de covarianza son matrices diagonales, cuya diagonal principal está dada por el vector $\left[\sigma^2_{1\,H_jc_k} \; \sigma^2_{2\,H_jc_k}\right]$.

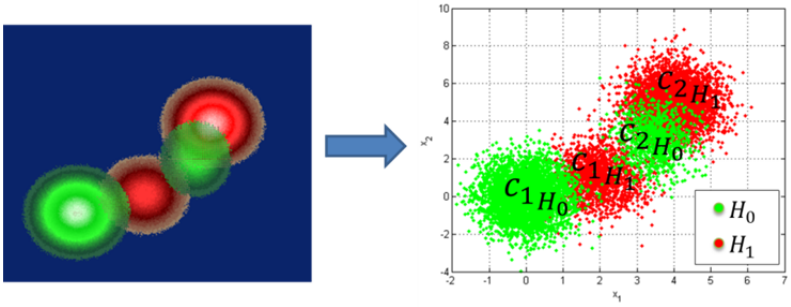


Figura 3.27.- Distribución conjunta (izquierda) y scatterplot (derecha) de los datos $\boldsymbol{x} = [x_1 \; x_2]$ usados en el ejemplo

Diseñamos dos detectores, uno por cada variable aleatoria x_i; para ello estimamos sus *PDFs* bajo cada una de las hipótesis y obtenemos un score $s_i = P(H_1|x_i)$ por cada detector:

$$s_i = \frac{f(x_i|H_1)\mathrm{P}(H_1)}{f(x_i|H_1)\mathrm{P}(H_0) + f(x_i|H_1)\mathrm{P}(H_0)} \tag{3.55}$$

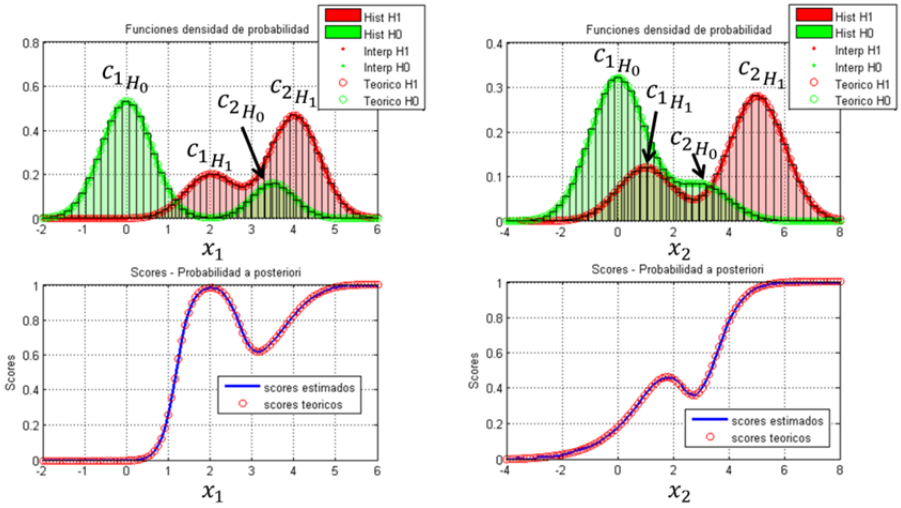


Figura 3.28.- PDFs marginales de cada variable aleatoria (arriba) y scores asignados dependiendo del valor de esta (abajo).

En la figura 3.28 se muestran estas *PDFs* marginales, obtenidas tanto de forma práctica mediante un histograma y una posterior interpolación, como de forma teórica, donde las *PDFs* marginales bajo cada hipótesis corresponden a una mezcla de Gaussianas univariantes $N_{c_1}(\cdot)$ y $N_{c_2}(\cdot)$. También se muestra el score asignado a cada posible valor de las variables x_i.

$$f(x_i|H_j) = \alpha_{H_j c_1} \cdot N_{c_1}\left(\mu_{i_{H_j c_1}}, \sigma_{i}^{2}{}_{H_j c_1}\right) + \alpha_{H_j c_2} \cdot N_{c_2}\left(\mu_{i_{H_j c_2}}, \sigma_{i}^{2}{}_{H_j c_2}\right) \tag{3.56}$$

Como conocemos la función de densidad conjunta $f(\boldsymbol{x}|H_j)$ bajo cada una de las hipótesis podemos obtener también el score que se obtendría con la fusión óptima:

$$s_{opt} = \frac{f(\boldsymbol{x}|H_1)\mathrm{P}(H_1)}{f(\boldsymbol{x}|H_1)\mathrm{P}(H_0) + f(\boldsymbol{x}|H_1)\mathrm{P}(H_0)} \tag{3.57}$$

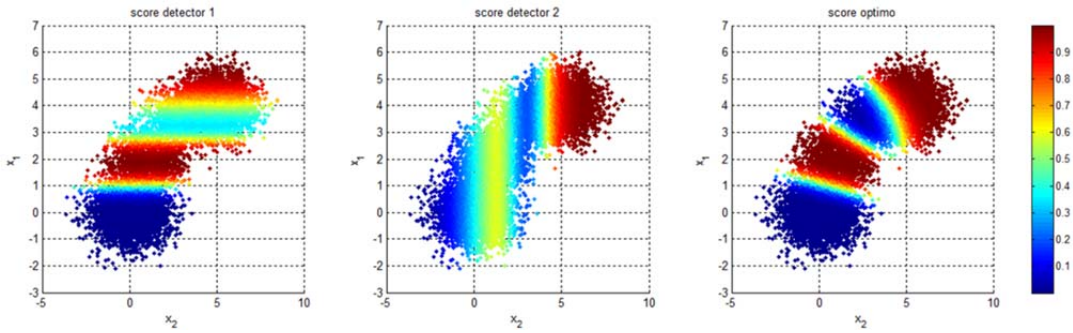


Figura 3.29 - Representación de los scores óptimos y aportados por cada detector con respecto a las muestras soft de entrada

En la figura 3.29 podemos ver una representación de las muestras soft de entrada $\boldsymbol{x}$ asociadas, tanto su score óptimo, como los asignados por cada uno de los detectores individuales. Observamos como por ejemplo, el detector 1 es capaz de discernir con claridad gran parte de las zonas asociadas a $c_{1_{H_0}}$ y a $c_{1_{H_1}}$, mientras que el detector 2 lo hace con gran parte de la zona asociada a $c_{2_{H_1}}$.

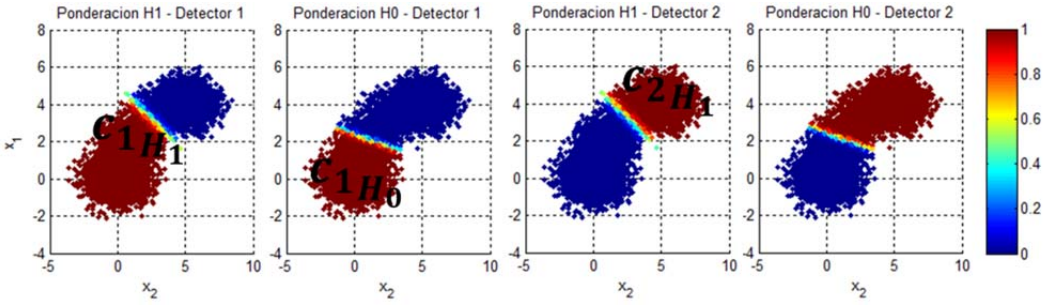


Figura 3.30 – Red de ponderación bajo cada una de las hipótesis de la fusión de scores utilizando una arquitectura de mezcla de expertos

Utilizamos una fusión de los scores s_1 y s_2 basada en la mezcla de expertos. En la figura 3.30 se pueden observar las redes de ponderación para cada una de las hipótesis obtenidas. Se observa claramente como bajo la hipótesis H_1 se escoge el detector 1 para detectar la componente $c_{1_{H_1}}$ y el detector 2 para la componente $c_{2_{H_1}}$, y para la hipótesis H_1 se escoge el detector 1 para detectar $c_{1_{H_0}}$.

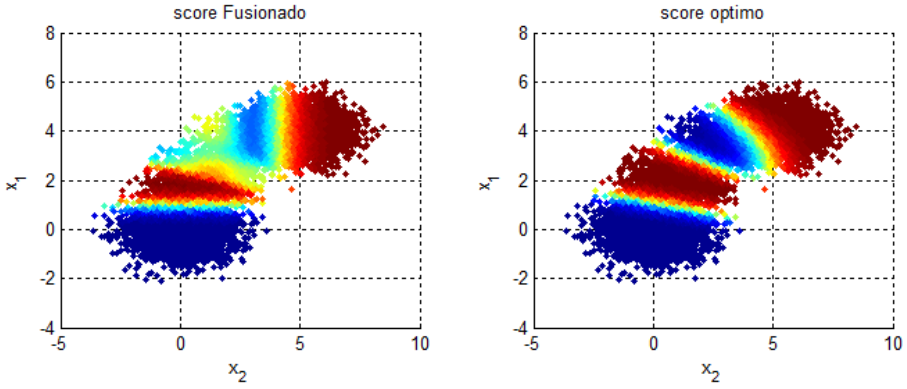


Figura 3.31 – Representación de los scores del caso óptimo y del obtenido tras la fusión de los dos detectores individuales en el espacio de las características

En la figura 3.31 podemos ver, tanto el score obtenido tras la fusión de los scores s_1 y s_2, como el score óptimo que se obtendría con la fusión soft óptima de las variables x_1 y x_2. Observamos como fusionando ambos scores mejoramos las

prestaciones obtenidas. En la figura 3.32 se muestra una curva *ROC* con el funcionamiento de ambos detectores individuales, la fusión óptima y la fusión utilizando *ME*.

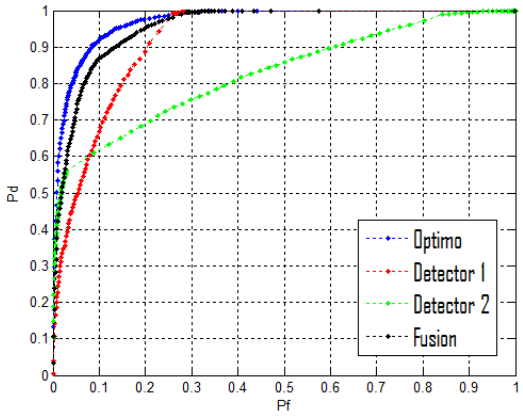


Figura 3.32 – Curvas ROC con las prestaciones de los detectores individuales, de la fusión óptima y de la fusión mediante ME

3.5.1. – Mezcla de expertos con integración α como función de combinación

En el modelo de mezcla de expertos para la fusión de scores de detectores, la función de integración probabilística usada es una función suma ponderada. Las puertas se encargan de obtener la ponderación de cada uno de los scores individuales s_k bajo la hipótesis H_j:

$$s_{f_{H_j}} = \sum_{k=1}^{K} g_k\left(\boldsymbol{x}, \Theta_{g_j}\right) s_k(\boldsymbol{x}) \tag{3.58}$$

Proponemos un nuevo modelo de mezcla de expertos, basado en la integración α:

$$s_{f_{H_j}} = \left(\sum_{k=1}^{K} g_k\left(\boldsymbol{x}, \Theta_{g_j}\right) s_k(\boldsymbol{x})^{\frac{1-\alpha}{2}}\right)^{\frac{2}{1-\alpha}} \tag{3.59}$$

Si reescribimos la ecuación (3.59) pasando al score fusionado el término que eleva a todo el sumatorio $\frac{2}{1-\alpha}$, podemos observar que nos volvemos a encontrar un problema similar al del caso de la mezcla de expertos con integración probabilística mediante la función suma, pero ahora con los scores individuales s_k modificados:

$$s'_{f_{H_j}} = \left(s_{f_{H_j}}\right)^{\frac{1-\alpha}{2}} = \sum_{k=1}^{K} g_k\left(\boldsymbol{x}, \Theta_{g_j}\right) s_k(\boldsymbol{x})^{\frac{1-\alpha}{2}} \tag{3.60}$$

Podemos observar cómo, según el valor del parámetro α la relación $x^{\frac{1-\alpha}{2}}$ es monótona creciente $(\alpha < 1)$ o decreciente $(\alpha > 1)$. Para $\alpha = 1$ nos encontramos con el caso particular del modelo de integración mediante el producto. Definimos el nuevo score $s'_{f_{H_j}}$ de forma que la relación con el score fusionado sea siempre monótona creciente:

$$s'_{f_{H_j}} = \begin{cases} \left(s_{f_{H_j}}\right)^{\frac{1-\alpha}{2}} = \sum_{k=1}^{K} g_k\left(\boldsymbol{x}, \Theta_{g_j}\right) s_k(\boldsymbol{x})^{\frac{1-\alpha}{2}} & \alpha < 1 \\ ln\left(s_{f_{H_j}}\right) = \sum_{k=1}^{K} g_k\left(\boldsymbol{x}, \Theta_{g_j}\right) ln\left(s_k(\boldsymbol{x})\right) & \alpha = 1 \\ -\left(s_{f_{H_j}}\right)^{\frac{1-\alpha}{2}} = \sum_{k=1}^{K} g_k\left(\boldsymbol{x}, \Theta_{g_j}\right)\left(-s_k(\boldsymbol{x})^{\frac{1-\alpha}{2}}\right) & \alpha > 1 \end{cases} \quad (3.61)$$

Observamos que podemos obtener el score fusionado mediante la mezcla de expertos con función α como modelo de integración probabilística, realizando primero una transformación en los scores individuales, entrenando la mezcla de expertos tal y como se define para la integración mediante la función suma para obtener unos scores modificados $s'_{f_{H_j}}$ para, posteriormente realizar la transformación inversa (3.62):

$$s_{f_{H_j}} = \begin{cases} \left(s'_{f_{H_j}}\right)^{\frac{2}{1-\alpha}} & \alpha < 1 \\ exp\left(s'_{f_{H_j}}\right) & \alpha = 1 \\ \left(-s'_{f_{H_j}}\right)^{\frac{2}{1-\alpha}} & \alpha > 1 \end{cases} \quad (3.62)$$

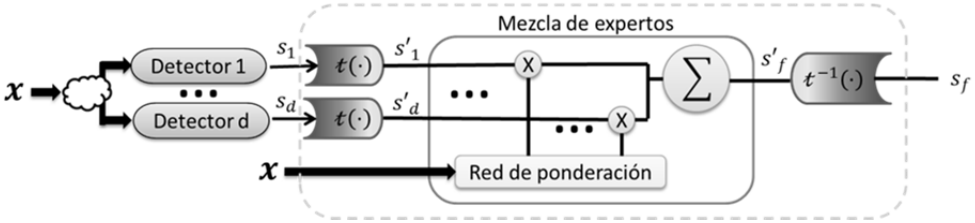


Figura 3.33 – Mezcla de expertos con función α como modelo probabilístico de integración de información

Mediante esta modificación del modelo de mezcla de expertos, basada en la utilización de la integración α como nueva función de integración probabilística, el parámetro α debe ser fijado de antemano. El entrenamiento completo del modelo de

mezcla de expertos con integración α propuesto, dado por la expresión 3.57, será objeto de futuros estudios.

3.6. – Conclusiones

Se ha planteado el problema de la fusión de scores como un caso especial de fusión soft en el que todos los datos a combinar se definen en un mismo rango normalizado $[0,1]$. Bajo este supuesto se ha realizado una presentación de las principales técnicas que pueden ser aplicadas.

Inicialmente se ha revisado la teoría de fusión de detectores a través de probabilidades a posteriori como salida soft, para observar que, únicamente en el caso en que los detectores trabajen con diferentes entradas independientes, el hecho de que la información soft sea considerada conceptualmente como probabilidades a posteriori conduce a un tipo especial de regla de fusión. En cualquier otro caso, las técnicas generales descritas en este capítulo se pueden aplicar en la fusión de detectores a través de sus probabilidades a posteriori.

Se ha realizado una revisión del estado del arte de las técnicas de combinación de scores, resaltando sus pros y contras a la hora de utilizarse en un problema de detección, en donde, debido a que los datos pueden ser heterogéneos y dependientes, pueden presentar geometrías complejas en cuanto a zonas de separación. Se ha comprobado como el principal problema de las técnicas que usualmente se utilizan es la falta de grados de libertad y flexibilidad a la hora de poder adaptarse a diferentes tipos de datos.

Se ha propuesto una novedosa técnica de fusión denominada integración α, la cual aporta un mayor grado de flexibilidad y de adaptación, siendo capaz de mejorar las prestaciones que se pueden obtener con respecto al resto de técnicas introducidas. Se ha derivado un novedoso método de entrenamiento basado en el criterio de maximización parcial del área bajo la curva *ROC.* Este nuevo criterio de entrenamiento es idóneo en el contexto del diseño de aplicaciones de detección como las tratadas en la presente tesis doctoral, donde el rango de trabajo está limitado entre ciertos valores de falsa alarma.

También se ha propuesto un modelo basado en la popular técnica de mezcla de expertos para realizar la fusión de scores en detección. Se ha introducido el modelo general de mezcla de expertos para posteriormente particularizarlo y adaptarlo en el caso particular de combinación de scores. Se ha propuesto como objeto de futuros estudios una variación de este método, introduciendo la integración α como función de combinación probabilística.

Capítulo 4:
Fusión de datos hard

"Una flecha sola, puede ser rota fácilmente, pero, muchas flechas son indestructibles"

– Gengis Kan –

Este capítulo se centra en las técnicas de fusión de información hard (discreta). En este caso, la fusión hard en un problema de detección se asocia directamente a la combinación de las decisiones individuales $u_i \in [0,1]$ tomadas por un conjunto de detectores. Así, el comportamiento estadístico de los datos se puede modelar mediante variables aleatorias binarias, completamente caracterizadas por sus funciones de masa de probabilidad.

Inicialmente se presenta la regla de fusión óptima en el caso de independencia, la cual deriva en una sencilla regla de combinación lineal de las decisiones. Posteriormente se estudia el caso en el que existe dependencia estadística entre las observaciones; se muestra como la regla de decisión óptima se torna más compleja, requiriendo, de una u otra forma, la caracterización de la correlación existente entre todos los posibles conjuntos de distinto tamaño (parejas, ternas, cuartetos…) que se pueden formar con las decisiones individuales.

Así, la implementación de la regla de fusión hard óptima puede ser inviable en ciertas aplicaciones; por ejemplo, puede requerir una cantidad de memoria de almacenamiento inadmisible. Presentamos diversas técnicas y reglas de fusión subóptimas que, aunque no consideran toda la estructura completa de dependencia presente, pueden proporcionar buenas prestaciones en determinados casos aliviando los costes computacionales de la regla de decisión óptima.

4.1. – Introducción

Tal y como se introdujo en el capítulo 1, la fusión de información en un sistema de detección puede realizarse en diferentes niveles. Es preferible en el diseño de un sistema de detección realizar la integración de información en la etapa denominada como fusión temprana, ya que a este nivel los datos presentan la mayor riqueza en cuanto información estadística sobre el evento a detectar. Así, es preferible combinar observaciones directamente recogidas de los sensores o ciertas características extraídas del procesado de estos.

No siempre será posible integrar la información en esos niveles; en determinados casos no se podrá tener acceso de forma conjunta a todos los datos o no se dispondrá de la capacidad de cómputo necesaria para la implementación de una fusión óptima de este tipo de datos soft. Por ejemplo, en el caso de las redes distribuidas de sensores, donde se debe ahorrar en potencia, ancho de banda de transmisión y requerimientos de procesado, puede ser inviable el transmitir toda la información a un único centro de procesado, el cuál suele poseer severas limitaciones en cuanto a capacidad de cómputo.

Mediante la fusión de las decisiones aportadas por diferentes detectores intermedios se reduce significantemente el grado de complejidad y costes

computacionales que implica la fusión temprana. Así, la fusión de decisiones hard puede convertirse en la elección ideal en el caso en el que no se pueda tener acceso a toda la información soft del sistema o cuando es inviable la implementación de una fusión soft por los diferentes requisitos computacionales y temporales que involucra su entrenamiento e implementación.

La regla de fusión hard óptima $\boldsymbol{u}: \{0,1\}^d \rightarrow u_{fus}: \{0,1\}$ pasa por la umbralización de la relación de verosimilitud (*LR*) del vector de decisiones binarias $\boldsymbol{u}$. En este caso, donde los datos a combinar son hard, se define como el cociente entre las funciones de masa de probabilidad (*PMF*, "*Probability Mass Function*") condicionadas a cada una de las hipótesis H_1 y H_0:

$$\Lambda(\boldsymbol{u}) = \frac{P(\boldsymbol{u}|H_1)}{P(\boldsymbol{u}|H_0)} = \frac{P(u_1, \dots, u_d|H_1)}{P(u_1, \dots, u_d|H_0)} \gtrless_{H_0}^{H_1} \eta \Leftrightarrow u_{fus} = \begin{cases} 1 & si\ \Lambda(\boldsymbol{u}) \geq \eta \\ 0 & si\ \Lambda(\boldsymbol{u}) < \eta \end{cases} \quad (4.1)$$

$$\boldsymbol{u}: \{0,1\}^d \rightarrow \Lambda(\boldsymbol{u}): \{a_1, \dots, a_{2^d}\} \subset \mathbb{R}$$

Al ser $\boldsymbol{u}$ un vector de variables aleatorias binarias, existirán 2^d posibles realizaciones de él, por lo tanto la *LR* en este caso se define en un subconjunto discreto de 2^d posibles valores reales.

Obtendremos ahora las reglas de fusión óptimas que se derivan partiendo de la expresión 4.1 para el caso de datos independientes y dependientes. Posteriormente introducimos una serie de reglas subóptimas, más sencillas de obtener e implementar que las reglas óptimas. Con estas reglas subóptimas se pueden obtener buenas prestaciones en ciertos casos, manteniendo una baja complejidad y requisitos computacionales en su implementación.

4.2. – Fusión hard óptima bajo independencia estadística de los datos

En el caso de considerar independencia estadística entre las decisiones individuales u_i, las *PMFs* bajo cada una de las hipótesis H_j, $P(\boldsymbol{u}|H_j)$, pueden ser expresadas de forma simple como el producto de las *PMFs* marginales $P(u_i|H_j)$:

$$P(\boldsymbol{u}|H_j) = \prod_{i=i}^{d} P(u_i|H_j) \quad (4.2)$$

Observamos que la probabilidad $P(u_i = 1|H_1)$ corresponde a la probabilidad de detección P_{d_i} que posee cada canal $i = 1, \dots, d$ (dedicir H_1 cuando la hipótesis verdadera es H_1). Así mismo, la probabilidad $P(u_i = 1|H_0)$ corresponde a la probabilidad de falsa alarma P_{f_i} (dedicir H_1 cuando la hipótesis verdadera es H_0). Es fácil expresar las *PMFs* en función de estas probabilidades de detección y falsa alarma asociadas a cada canal de datos:

$$P_{ind}(\boldsymbol{u}|H_j) = \begin{cases} \prod_{i=1}^{d} (P_{d_i})^{u_i} \cdot (1-P_{d_i})^{1-u_i} & bajo\ H_1 \\ \prod_{i=1}^{d} (P_{f_i})^{u_i} \cdot (1-P_{f_i})^{1-u_i} & bajo\ H_0 \end{cases} \tag{4.3}$$

Implementando el test de relación de verosimilitud logarítmica, y despreciando los términos constantes se consigue el estadístico de fusión óptima obtenido por Chair-Varshney [92], el cual se basa en una suma ponderada de las decisiones individuales (figura 4.1):

$$\begin{gathered} ln\,\Lambda(\boldsymbol{u}) = ln\left(\prod_{i=1}^{d} \frac{(P_{d_i})^{u_i} \cdot (1-P_{d_i})^{1-u_i}}{(P_{f_i})^{u_i} \cdot (1-P_{f_i})^{1-u_i}}\right) \Rightarrow \mathcal{T}(\boldsymbol{u}) = \sum_{i=1}^{d} u_i \cdot \mathrm{c}_i \\ \mathrm{c}_i = \log\left(\frac{P_{d_i} \cdot \left(1-P_{f_i}\right)}{P_{f_i} \cdot \left(1-P_{d_i}\right)}\right) \qquad u_{fus} = \begin{cases} 1 & si\ \mathcal{T}(\boldsymbol{u}) \geq \eta \\ 0 & si\ \mathcal{T}(\boldsymbol{u}) < \eta \end{cases} \end{gathered} \tag{4.4}$$

En el caso de que exista dependencia estadística entre las decisiones bajo alguna hipótesis, este estadístico de fusión hard será subóptimo. Las prestaciones obtenidas estarán fuertemente relacionadas con las características de dependencia que presenten los datos. En el apartado 4.5, ilustramos este hecho mediante un completo estudio de dos detectores.

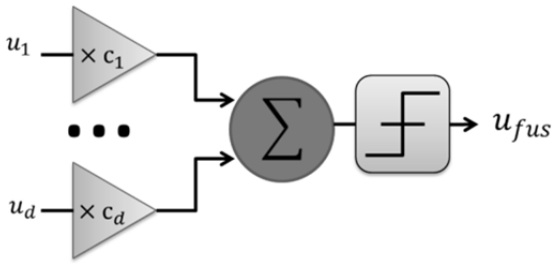


Figura 4.1 – Fusión de información hard mediante la regla de Chair-Varshney [92]

4.3. – Fusión hard óptima bajo dependencia estadística de los datos

En el caso de que exista dependencia estadística entre los datos bajo alguna de las hipótesis, las *PMFs* ya no podrán expresarse de forma precisa mediante (4.2), y por lo tanto la regla de fusión (4.4) puede convertirse en subóptima. Para la implementación de la regla de fusión óptima, se deben estimar las *PMFs* de forma que se modele

fielmente, tanto el comportamiento marginal de los datos, como las características de dependencia estadística que presentan.

Estimación de PMFs directa

Puesto que la *PMF* del vector aleatorio $\boldsymbol{u} = [u_1, \dots, u_d]$ se compone de 2^d probabilidades asociadas a cada una de las posibles realizaciones de este vector (desde $[u_1 = 0, \dots, u_d = 0] \equiv \boldsymbol{u}^0$ hasta $[u_1 = 1, \dots, u_d = 1] \equiv \boldsymbol{u}^{2^d-1}$**)**, podemos estimarlas de forma empírica mediante el conteo del número de apariciones de cada una de estas realizaciones:

$$P(\boldsymbol{u}^i|H_j) = \frac{Nº\ de\ apariciones\ de\ \boldsymbol{u}^i\ bajo\ H_j}{Nº\ total\ de\ realizaciones\ de\ \boldsymbol{u}\ bajo\ H_j} \tag{4.5}$$

Así, debemos de estimar $2^d - 1$ parámetros por cada una de las hipótesis (sabiendo que la suma de probabilidades es siempre la unidad, podemos ahorrarnos la estimación de una de ellas).

Estimación de PMFs mediante desarrollo de Drakopoulos

En [93] se obtiene una expresión para obtener las *PMFs* cuando las decisiones locales no son independientes:

$$P(\boldsymbol{u}|H_j) = P_j(u_1, \dots, u_d) = \sum_{I \subseteq A_0} (-1)^{|I|} E_j\left[\prod_{i \in A_1 \cup I} u_i\right] \tag{4.6}$$

$A_\mu = \{i : u_i = \mu\} \qquad 1 \le i \le d, \quad \mu = 0,1. \quad E_j[\prod_{i\in A} u_i] = 1\ si\ A = \phi$

Por lo tanto, observamos como existe un conjunto de coeficientes C que contienen toda la información tanto de las *PMFs* marginales como de las características de dependencia estadística:

$$C = \left\{E_j\left[\prod_{i \epsilon I} u_i\right] : I \subseteq \{1, \dots, d\}, \qquad I \neq \phi\right\} \tag{4.7}$$

Dado que cada decisión individual u_i es binaria, podemos desarrollar la expresión (4.6) para cada subconjunto $I = \{i_1, \dots, i_k\} \subseteq \{1, \dots, d\}$:

$$E_j\left[\prod_{i \epsilon I} u_i\right] = E_j[u_{i_1} u_{i_2} \dots u_{i_k}] = P_j(u_{i_1} = 1, \dots, u_{i_k} = 1) \tag{4.8}$$

Como observamos en la expresión (4.8) necesitamos conocer, además de las *PMF* marginales de cada canal, las probabilidades conjuntas de todas las posibles

combinaciones de 2 hasta d decisiones individuales escogiendo la hipótesis H_1, cuando la hipótesis real es $H_j, j = 0,1$.

Al igual que en el caso anterior, el número total de parámetros a estimar es de $2^d - 1$ por cada una de las hipótesis.

Estimación de PMFs mediante expansión de Bahadur-Lazarsfeld

Otro desarrollo diferente de la expresión de la *PMF* es llevado a cabo en [94], donde se utiliza la denominada expansión de Bahadur-Lazarsfeld. Se introducen unas nuevas variables normalizadas w_k a partir de las decisiones individuales, transformándolas de modo que posean media cero y varianza unidad:

$$w_k = \frac{u_k - p_k}{\sqrt{p_k(1-p_k)}} \quad k = 1,2,\dots,d \qquad \text{donde } p_k = P(u_k = 1) \tag{4.9}$$

Mediante estas variables y considerando los polinomios de Bahadur-Lazarsfeld se demuestra que la *PMF* bajo la hipótesis H_j se puede expresar como sigue:

$$P(\boldsymbol{u}|H_j) = P_{ind}(\boldsymbol{u}|H_j)\left[1 + \sum_{k<l}\gamma_{kl}^j w_k w_l + \sum_{k<l<m}\gamma_{klm}^j w_k w_l w_m + \dots + \gamma_{12\dots d}^j w_1 w_2 \dots w_d\right] \tag{4.10}$$

$\gamma_{kl}^j = \sum_{\boldsymbol{u}} w_k\, w_l P(\boldsymbol{u}|H_j),\ \gamma_{klm}^j = \sum_{\boldsymbol{u}} w_k\, w_l w_m P(\boldsymbol{u}|H_j), \dots,\ \gamma_{12\dots n}^j = \sum_{\boldsymbol{u}} w_1 \dots w_n P(\boldsymbol{u}|H_j)$

dónde $P_{ind}(\boldsymbol{u}|H_j)$ hace referencia a la *PMF* obtenida considerando independencia (4.3).

Podemos observar como $P_{ind}(\boldsymbol{u}|H_j)$ corresponde a la función de densidad del caso en que las decisiones individuales de los detectores fueran independientes (4.2), la cual es multiplicada por un factor de corrección debido a la dependencia entre los detectores. Este factor de corrección viene determinado por $2^d - d - 1$ coeficientes de correlación ($\gamma_{kl}^j,\ \gamma_{klm}^j, \dots,\ \gamma_{12\dots d}^j$), obtenidos tras "uniformizar los datos".

Nótese la similitud existente entre la técnica de estimación de *PDFs* mediante la teoría de las cópulas expuesta en el capítulo anterior, y esta técnica de estimación de *PMF*s. Aquí, la *PMF* considerando independencia se ve multiplicada por lo que podríamos denominar función de masa de cópula, la cual contiene toda la información de dependencia estadística entre los datos.

Podemos concluir que la *PMF* conjunta de las decisiones individuales bajo la hipótesis H_j, en el caso en el que no son independientes entre sí, depende, además de las *d* probabilidades marginales asociadas a cada uno de los detectores bajo la hipótesis H_j, de $2^d - d - 1$ coeficientes asociados a la estructura de dependencia existente entre los datos.

Ejemplo

Presentamos un ejemplo en el que disponemos de tres canales de datos binarios, recogidos en el vector $\boldsymbol{u} = [u_1\, u_2\, u_3]$. Los canales presentan dependencia estadística bajo ambas hipótesis. Si obtenemos la matriz de correlación de las decisiones individuales bajo ambas hipótesis, podemos comprobar que existe dependencia entre las decisiones de los detectores:

	ρ Hipótesis H_0			ρ Hipótesis H_1		
	Detector 1	Detector 2	Detector 3	Detector 1	Detector 2	Detector 3
Detector 1	1	0.1227	0.5646	1	0.4681	0.1968
Detector 2		1	-0.0621		1	0.1094
Detector 3			1			1

Tabla 4.1 – Matrices de correlación de los datos binarios bajo cada hipótesis

Desarrollando la expresión (4.6) de la *PMF* para cada hipótesis:

$$\begin{aligned}
&\mathrm{P_j}(\mathbf{u} = [0,0,0]) = 1 - \mathrm{P_j}(\mathrm{u}_1 = 1) - \mathrm{P_j}(\mathrm{u}_2 = 1) - \mathrm{P_j}(\mathrm{u}_3 = 1) + \mathrm{P_j}(\mathrm{u}_1 = 1, \mathrm{u}_2 = 1) + \mathrm{P_j}(\mathrm{u}_1 = 1, \mathrm{u}_3 = 1) \\
&\qquad + \mathrm{P_j}(\mathrm{u}_2 = 1, \mathrm{u}_3 = 1) - \mathrm{P_j}(\mathrm{u}_1 = 1, \mathrm{u}_2 = 1, \mathrm{u}_3 = 1) \\
&P_j(\boldsymbol{u} = [0,0,1]) = P_j(u_3 = 1) - P_j(u_1 = 1, u_3 = 1) - P_j(u_2 = 1, u_3 = 1) + P_j(u_1 = 1, u_2 = 1, u_3 = 1) \\
&P_j(\boldsymbol{u} = [0,1,0]) = P_j(u_2 = 1) - P_j(u_1 = 1, u_2 = 1) - P_j(u_2 = 1, u_3 = 1) + P_j(u_1 = 1, u_2 = 1, u_3 = 1) \\
&P_j(\boldsymbol{u} = [0,1,1]) = P_j(u_2 = 1, u_3 = 1) - P_j(u_1 = 1, u_2 = 1, u_3 = 1) \\
&P_j(\boldsymbol{u} = [1,0,0]) = P_j(u_1 = 1) - P_j(u_1 = 1, u_3 = 1) - P_j(u_1 = 1, u_2 = 1) + P_j(u_1 = 1, u_2 = 1, u_3 = 1) \\
&P_j(\boldsymbol{u} = [1,0,1]) = P_j(u_1 = 1, u_3 = 1) - P_j(u_1 = 1, u_2 = 1, u_3 = 1) \\
&P_j(\boldsymbol{u} = [1,1,0]) = P_j(u_1 = 1, u_2 = 1) - P_j(u_1 = 1, u_2 = 1, u_3 = 1) \\
&P_j(\boldsymbol{u} = [1,1,1]) = P_j(u_1 = 1, u_2 = 1, u_3 = 1)
\end{aligned} \tag{4.11}$$

Desarrollando la expresión (4.10) de la *PMF* para cada hipótesis:

$$\begin{aligned}
P_j(\mathbf{u}) = &\left(\prod_{i=1}^{n} P_j(u_i = 1)^{u_i} \cdot (1 - P_j(u_i = 1))^{1-u_i}\right)\left[1 + \gamma_{12}^{j} w_1 w_2 + \gamma_{13}^{j} w_1 w_3 + \gamma_{23}^{j} w_2 w_3\right. \\
&\left. + \gamma_{123}^{j} w_1 w_2 w_3\right]
\end{aligned} \tag{4.12}$$

Obtenemos los parámetros que definen la *PMF* conjunta bajo cada hipótesis en la tabla 4.2 en cada uno de los desarrollos.

Podemos estimar de forma sencilla la *PMF* obviando el hecho de que existe dependencia entre las decisiones y las consideramos independientes entre sí, utilizando la expresión 4.2. Representamos en la figura 4.2 las *PMF* bajo cada hipótesis, tanto en el caso en el que hemos tenido en cuenta la dependencia existente, como en el caso en el que hemos considerando independencia. Representamos también la relación de verosimilitud de ambos casos. Observamos

como las *PFMs* no son exactamente iguales, y por lo tanto tampoco lo es la relación de verosimilitud.

	Hipótesis H_0		Hipótesis H_1	
$I = \{i_1, \dots, i_k\}$	$P_0(u_{i_1} = 1, \dots, u_{i_k} = 1)$	$\gamma^0_{i_1 \dots i_k}$	$P_1(u_{i_1} = 1, \dots, u_{i_k} = 1)$	$\gamma^1_{i_1 \dots i_k}$
{1}	0.2215		0.8133	
{2}	0.2694		0.7932	
{3}	0.1887		0.7846	
{1, 2}	0.0823	0.1227	0.7190	0.4681
{1, 3}	0.1335	0.5646	0.6697	0.1968
{2, 3}	0.0401	-0.0621	0.6405	0.1094
{1, 2, 3}	0.0367	-0.0157	0.5961	-0.1207

Tabla 4.2 – Parámetros en los que se descomponen las PMFs

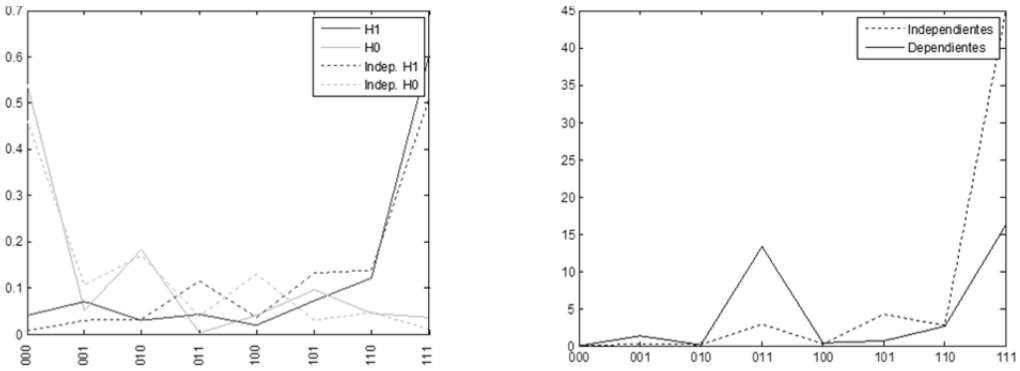


Figura 4.2 – PMFs estimadas en el caso óptimo de considerar la dependencia y en el caso de despreciarla a la izquierda. Relación de verosimilitud obtenida con cada una de las estimaciones a la derecha

Una vez conocida la relación de verosimilitud podemos implementar la regla de fusión óptima dada por (4.1). Realizando un barrido del parámetro η y obteniendo las prestaciones en cuanto a probabilidad de falsa alarma y de detección podemos representar la curva ROC del proceso de fusión (figura 4.3). Obtenemos también la curva ROC del caso en que hemos considerado independencia en las decisiones y observamos que se obtiene peores resultados tras la fusión que en el caso óptimo:

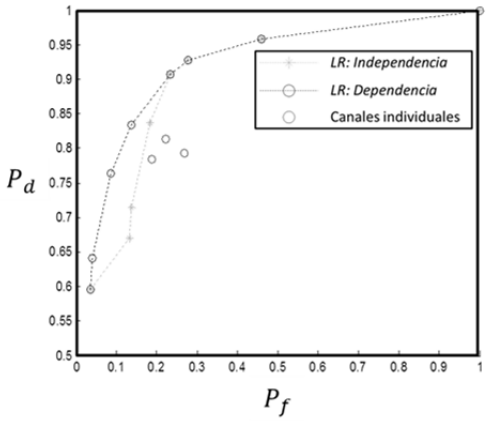


Figura 4.3 – Curva ROC donde se representa las prestaciones de los canales individuales y las obtenidas mediante su fusión basada en la LR considerando la dependencia estadística de los datos y despreciándola

4.4. – Técnicas de fusión hard subóptimas para datos dependientes

La presencia de dependencia estadística entre los datos hard puede implicar unos elevados costes de entrenamiento e implementación cuando se desea implementar una regla de fusión óptima basada en la *LR*. Como ya hemos visto, utilicemos uno u otro método para estimar la *PMF* se necesitan conocer $2^d - 1$ valores. En el caso de un número elevado de detectores en ciertas aplicaciones puede resultar inviable la utilización de estas técnicas de estimación, sobre todo por la cantidad de memoria que implica almacenar la gran cantidad de valores que precisa la *PMF*.

Se pueden utilizar una serie de técnicas, a priori subóptimas, para aliviar estos problemas de coste. Dependiendo de las características de dependencia existentes, algunas de estas reglas subóptimas, pueden conseguir prestaciones similares a las del caso óptimo.

En el apartado 4.5, realizamos un análisis del caso de fusión de dos detectores a través de sus decisiones hard, donde se puede observar la importante degradación que puede causar el no considerar las características de dependencia, incluso en un caso tan simple como el de fusión de dos detectores.

4.4.1. – Asunción de independencia

Como ya hemos comentado, podemos despreciar la posible dependencia estadística entre los datos y utilizar una fusión basada en la expresión (4.4). Así, utilizamos únicamente la información marginal que proporciona cada canal de los datos. Bajo determinadas condiciones de dependencia estadística, esto puede suponer una importante degradación de las prestaciones.

4.4.2. – Estimación subóptima de *PMFs*

Otra posibilidad, es la utilización de métodos de estimación subóptimos de PMFs:

- Truncamiento de la expansión de Bahadur-Lazarsfeld

Consiste en considerar nulos ciertos factores γ de la expansión de Bahadur-Lazarsfeld utilizada para estimar las *PMFs* (4.9) bajo cada hipótesis, con objeto de implementar una regla de fusión basada en la *LR* [95].

Truncando la expansión de Bahadur-Lazarsfeld reducimos la exactitud del factor de corrección (lo que podemos denominar como función de masa de cópula), pudiendo incluso resultar en valores inapropiados para la función densidad de probabilidad, como valores negativos o superiores a la unidad.

- Aproximar las PMFs conjuntas mediante el producto de sus distribuciones de menor orden

La teoría de la probabilidad nos proporciona una regla de cadena básica para descomponer una distribución conjunta de probabilidad:

$$P(u_1, \dots, u_d) = \prod_{i=1}^{d} P\left(u_{m_i} | u_{m_{i-1}}, \dots, u_{m_1}\right) \tag{4.13}$$

$$\{m_1, \dots, m_n\} \ \ cualquier\ permutacion\ del\ conjunto\ \{1, \dots, n\}$$

La permutación elegida para ordenar las variables en la cadena no es importante, cualquiera da lugar a una descomposición válida. Aunque si puede ser interesante ordenar las variables u_i bajo una forma determinada para obtener provecho de determinadas independencias condicionales si estas existen. La expresión (4.13) no proporciona una reducción en los costes de implementación, pero a raíz de este tipo de representación, mediante ciertas asunciones, se pueden obtener buenas aproximaciones reduciendo la cantidad de información necesaria.

Así, podemos encontrar una amplia colección de trabajos donde se proponen aproximaciones para estimar una *PMF* reduciendo la cantidad de información necesaria para su modelado. Lewis y Brown [96], [97] propusieron un tipo de aproximación para estimar la *PMF* utilizando únicamente ciertos términos con dependencias de primer orden. En [98] se propuso un método eficiente para estimar y construir este tipo de aproximaciones, basada en la creación de los llamados árboles de dependencia; así se conoce como la aproximación mediante un árbol de Chow-Liu. Existen multitud de generalizaciones de estos trabajos, donde se buscan "árboles" modelados con componentes de orden mayor, sin restringirse a componentes con dependencias de primer orden, por ejemplo como los incluidos en [99]–[101] entre otros.

4.4.3. – Regla del conteo

Una de las reglas de fusión más ampliamente usadas por su sencillez es la regla del conteo. También es conocida por el nombre de votación por mayoría ("*majority voting*"). Ésta se basa simplemente en una umbralización del conteo de decisiones individuales a favor de la hipótesis H_1:

$$m = \sum_{i=1}^{d} u_i \lessgtr_{H_1}^{H_0} k \tag{4.14}$$

Esta regla del conteo ha sido estudiada aplicada en diferentes escenarios [102], [103]. Destacar entre estos estudios el desarrollado en el *GTS* [104]. Se puede concluir que es una regla de fusión óptima en el caso de que los diferentes datos sean independientes y posean las mismas características marginales; es una regla subóptima en el caso de datos dependientes y/o distintos puntos de trabajo (diferentes características marginales).

4.4.4. – Reglas AND, OR y XOR

En diversos trabajos [102], [105] podemos encontrar otro tipo de reglas subóptimas de fusión de decisiones muy básicas como lo son la regla *AND*, consistente en decidir globalmente la hipótesis $H_1(u_{fus} = 1)$, únicamente cuando todas las decisiones individuales la escogen de forma conjunta $(u_i = 1 \; \forall \; i)$, la regla *OR* basada en decidir H_1 cuando al menos una de las decisiones individuales la reporte $u_i = 1$ y la regla *XOR* en la que se escoge H_1 únicamente cuando sólo una decisión individual está a favor de ella.

4.5. – Análisis de la fusión hard de dos detectores dependientes

Con objeto de ilustrar el proceso de la obtención de la regla de fusión óptima y comprobar cómo la dependencia estadística juega un papel fundamental en ella, proponemos un sencillo ejemplo en el que se pretende fusionar dos detectores a través de las decisiones hard que aportan a su salida $\boldsymbol{u} = [u_1, u_2]$. Utilizando el desarrollo propuesto por Drakopoulos mediante la expresión 4.6 obtenemos las expresiones de las *PMFs* conjuntas bajo ambas hipótesis (denotamos como detector 1 al de mayor probabilidad de detección):

$$\begin{aligned}
&P(\boldsymbol{u} = [0,0]|H_j) = 1 \; + \; (-1) \cdot E_1[u_1] \; + \; (-1) \cdot E_1[u_2] \; + \; 1 \cdot E_1[u_1 u_2] \\
&P(\boldsymbol{u} = [1,0]|H_j) = E_1[u_1] \; + \; (-1) \cdot E_1[u_1 u_2] \\
&P(\boldsymbol{u} = [0,1]|H_j) = E_1[u_2] \; + \; (-1) \cdot E_1[u_1 u_2] \\
&P(\boldsymbol{u} = [1,1]|H_j) = E_1[u_1 u_2]
\end{aligned} \tag{4.15}$$

$$E_1[u_1] = P_{d1} \qquad E_1[u_2] = P_{d2} \qquad E_0[u_1] = P_{f1} \qquad E_0[u_2] = P_{f2}$$

$$E_1[u_1u_2] = P_1(u_1 = 1, u_2 = 1) \qquad E_0[u_1u_2] = P_0(u_1 = 1, u_2 = 1)$$

$$P_d = P_{d1} \quad P_f = P_{f1} \quad R_d = \frac{P_{d2}}{P_{d1}} \leq 1 \quad R_f = \frac{P_{f2}}{P_{f1}}$$

Definimos dos parámetros para caracterizar el grado de dependencia entre ambos detectores. Estos parámetros nos permiten simplificar las expresiones obtenidas y cuantificar la dependencia en relación a las probabilidades de detección y falsa alarma de ambos detectores:

$$\begin{aligned} \beta_0 &= \frac{P_0\left(u_1 = \frac{1}{u_2} = 1\right)}{P_{f1}} = \frac{P_0(u_1 = 1, u_2 = 1)}{R_f \cdot {P_f}^2} \\ \beta_1 &= \frac{P_1(u_1 = 1/u_2 = 1)}{P_{d1}} = \frac{P_1(u_1 = 1, u_2 = 1)}{R_d \cdot {P_d}^2} \end{aligned} \tag{4.16}$$

Podemos evaluar la relación de verosimilitud $\Lambda(\boldsymbol{u})$ para cada uno de los casos de fusión en función de las probabilidades de falsa alarma y detección individuales (asociadas al comportamiento marginal individual) y los parámetros que miden el grado de correlación. Se observa claramente como en la relación de verosimilitud, y por tanto la regla de fusión óptima, la dependencia estadística posee un gran peso:

$$\begin{aligned} \Lambda([0,0]) &= \frac{(1 - P_d \cdot (1 + R_d) + R_d \cdot \beta_1 \cdot {P_d}^2)}{(1 - P_f \cdot (1 + R_f) + R_f \cdot \beta_0 \cdot {P_f}^2)} & \Lambda([1,0]) &= \frac{P_d - R_d \cdot \beta_1 \cdot {P_d}^2}{P_f - R_f \cdot \beta_0 \cdot {P_f}^2} \\ \Lambda([0,1]) &= \frac{P_d \cdot R_d - R_d \cdot \beta_1 \cdot {P_d}^2}{P_f \cdot R_f - R_f \cdot \beta_0 \cdot {P_f}^2} & \Lambda([1,1]) &= \frac{R_d \cdot \beta_1 \cdot {P_d}^2}{R_f \cdot \beta_0 \cdot {P_f}^2} \end{aligned} \tag{4.17}$$

Sean *X* e *Y* dos variables aleatorias binarias, las cuales cumplen $P(X = 1) = p$ y $P(Y = 1) = q$. Se define la covarianza de ambas variables como: $C = Cov(X, Y) = P(X = 1, Y = 1) - pq$. No es complicado obtener las siguientes relaciones:

$$\begin{aligned} &\mathrm{C} \geq -\mathrm{pq} && P(X = 1, Y = 1) \leq min(p, q) \\ &C \leq min\,(p(1 - q), q(1 - p)) \\ &0 \leq P(X = 0, Y = 0) = 1 - p - q + P(X = 1, Y = 1) = 1 - p - q + C - pq \\ &P(X = 1, Y = 1) \geq p + q - 1 && C \geq -(1 - p)(1 - q). \end{aligned} \tag{4.18}$$

Basándonos en esas relaciones podemos obtener de forma sencilla los límites de los parámetros β utilizados para caracterizar el grado de dependencia:

$$\begin{aligned} max(0, \frac{P_f(1 + R_f) - 1}{R_f \cdot {P_f}^2}) \leq \ \beta_0 \ \leq min(\frac{1}{P_f}, \frac{1}{R_f \cdot P_f}) \\ max\left(0, \frac{P_d(1 + R_d) - 1}{R_d \cdot {P_d}^2}\right) \leq \beta_1 \ \leq min\left(\frac{1}{P_d}, \frac{1}{R_d \cdot P_d}\right) = \{R_d \leq 1\} = \frac{1}{P_d} \end{aligned} \tag{4.19}$$

Se pueden relacionar los parámetros β_j con los coeficientes de correlación. Observamos como un valor $\beta = 1$ denota independencia, mientras que $\beta < 1$ implica correlación negativa y $\beta > 1$ correlación positiva:

$$\beta_0 = \rho_0 \cdot \sqrt{\frac{(1-P_f)\cdot(1-R_f\cdot P_f)}{R_f\cdot {P_f}^2}} + 1 \qquad \rho_0 = (\beta_0 - 1)\cdot\sqrt{\frac{R_f\cdot {P_f}^2}{(1-P_f)\cdot(1-R_f\cdot P_f)}}$$
$$\beta_1 = \rho_1 \cdot \sqrt{\frac{(1-P_d)\cdot(1-R_d\cdot P_d)}{R_d\cdot {P_d}^2}} + 1 \qquad \rho_1 = (\beta_1 - 1)\cdot\sqrt{\frac{R_d\cdot {P_d}^2}{(1-P_d)\cdot(1-R_d\cdot P_d)}} \tag{4.20}$$

Según el valor del umbral en la regla $\Lambda(\boldsymbol{u})$, se escogerá la hipótesis H_1 cuando se den uno o varios posibles valores de $\boldsymbol{u}$**.** Podemos obtener fácilmente de 4.17 las *PFA* y *PD* tras la fusión de cada una de las posibles reglas resultantes:

1- $\underline{\boldsymbol{u} = (1,1)}$
$PFA = R_f \cdot \beta_0 \cdot {P_f}^2$
$PD = R_d \cdot \beta_1 \cdot {P_d}^2$

2- $\underline{\boldsymbol{u} = (0,1)}$
$PFA = P_f \cdot R_f - R_f \cdot \beta_0 \cdot {P_f}^2$
$PD = P_d \cdot R_d - R_d \cdot \beta_1 \cdot {P_d}^2$

3- $\underline{\boldsymbol{u} = (1,0)}$
$PFA = P_f - R_f \cdot \beta_0 \cdot {P_f}^2$
$PD = P_d - R_d \cdot \beta_1 \cdot {P_d}^2$

4- $\underline{\boldsymbol{u} = (1,1),\ \boldsymbol{u} = (1,0)}$
$PFA = P_f$
$PD = P_d$

5- $\underline{\boldsymbol{u} = (1,1),\ \boldsymbol{u} = (0,1)}$
$PFA = R_f \cdot P_f$
$PD = R_d \cdot P_d$

6- $\underline{\boldsymbol{u} = (0,1)\ ,\boldsymbol{u} = (1,0)}$
$PFA = P_f \cdot (1 + R_f) - 2 \cdot R_f \cdot \beta_0 \cdot {P_f}^2$
$PD = P_d \cdot (1 + R_d) - 2 \cdot R_d \cdot \beta_1 \cdot {P_d}^2$

7- $\underline{\boldsymbol{u} = (0,0), \boldsymbol{u} = (1,1)}$
$PFA = 1 - P_f \cdot (1 + R_f) + 2 \cdot R_f \cdot \beta_0 \cdot {P_f}^2$
$PD = 1 - P_d \cdot (1 + R_d) + 2 \cdot R_d \cdot \beta_1 \cdot {P_d}^2$

8- $\underline{\boldsymbol{u} = (0,1), \boldsymbol{u} = (1,1), \boldsymbol{u} = (1,0)}$
$PFA = P_f \cdot (1 + R_f) - R_f \cdot \beta_0 \cdot {P_f}^2$
$PD = P_d \cdot (1 + R_d) - R_d \cdot \beta_1 \cdot {P_d}^2$

9- $\underline{\boldsymbol{u} = (0,0), \boldsymbol{u} = (0,1), \boldsymbol{u} = (1,1)}$
$PFA = 1 - P_f + R_f \cdot \beta_0 \cdot {P_f}^2$
$PD = 1 - P_d + R_d \cdot \beta_1 \cdot {P_d}^2$

10- $\underline{\boldsymbol{u} = (0,0), \boldsymbol{u} = (1,0), \boldsymbol{u} = (1,1)}$
$PFA = 1 - R_f \cdot P_f + R_f \cdot \beta_0 \cdot {P_f}^2$
$PD = 1 - R_d \cdot P_d + R_d \cdot \beta_1 \cdot {P_d}^2$ (4.21)

En la tesis final de master realizada en la etapa de formación pre-doctoral [106], se realiza un profundo estudio teórico del comportamiento de la regla óptima 4.17, partiendo del caso de independencia y estudiando su variación con respecto a las características de dependencia. Mostramos en las siguientes figuras extraídas de ese trabajo un ejemplo del mejor caso, un caso intermedio y el peor caso según la dependencia existente en reglas subóptimas como son la *AND*, *OR* y *XOR*. Se muestran tanto el punto de operación de los dos detectores (círculos) como el resultado de la detección (estrella) en diversos escenarios con diferentes características de dependencia. Mostramos con este sencillo ejemplo como la regla óptima de fusión debe tener en cuenta siempre las características de dependencia entre los datos. Observamos como la idoneidad en el uso de diferentes técnicas subóptimas dependerá de las características de los datos, tanto en sus prestaciones marginales, como en la dependencia estadística que presenten. Podemos ver como en caso de existencia de dependencia estadística, el escoger una técnica de fusión no idónea puede derivar incluso en una pérdida de prestaciones con respecto a cada fuente de información aislada.

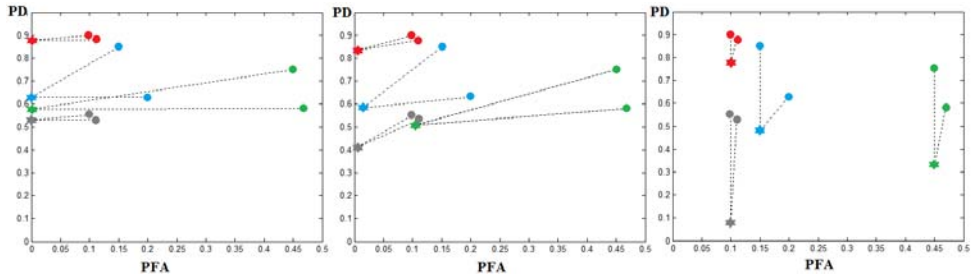


Figura 4.4 – Gráficas ROC de fusión de regla AND

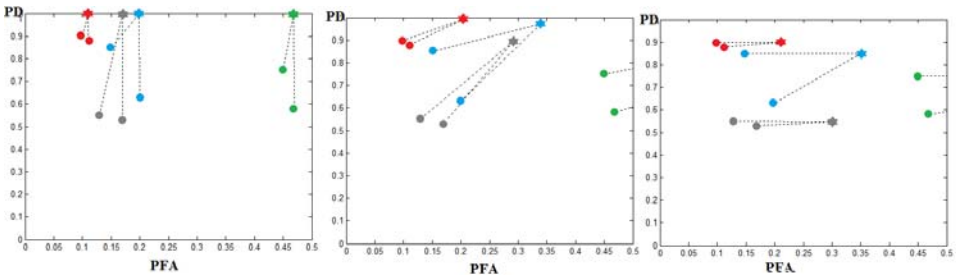


Figura 4.5 – Gráficas ROC de fusión de regla OR

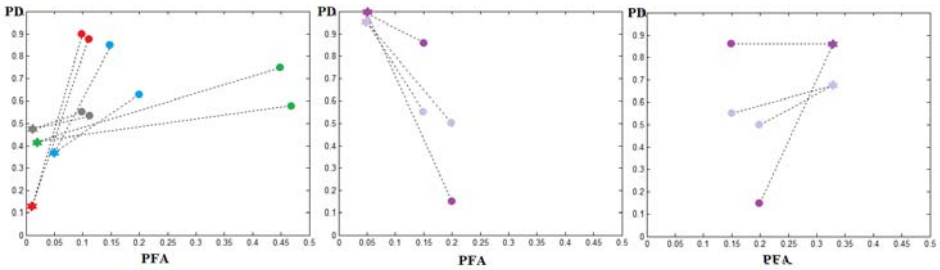


Figura 4.6 – Gráficas ROC de fusión de regla XOR

4.6. – Conclusiones

En este capítulo se ha realizado un repaso del estado del arte de las técnicas de fusión de información hard, situándonos en el contexto en que se realiza una integración de las decisiones individuales aportadas por diferentes detectores.

Se ha definido la regla de fusión óptima como la obtenida mediante la derivación de la relación de verosimilitud, dada por el cociente de las *PMFs* de los datos bajo las hipótesis H_1 y H_0. Se ha obtenido la regla de fusión óptima bajo los casos en el que exista independencia estadística o no entre los datos. Posteriormente se han presentado otra serie de reglas de fusión subóptimas, más simples y con menores requerimientos.

Se ha argumentado y mostrado con sencillos ejemplos como, el no tener en cuenta las características de dependencia, puede conllevar a una degradación en las prestaciones obtenidas en el sistema de detección, incluso siendo peores que las conseguidas considerando los datos de forma aislada. Se ha mostrado como ciertas técnicas de fusión subóptimas, pueden aprovechar cierta información de dependencia de forma implícita, y en determinados escenarios y bajo determinadas condiciones, pueden obtener buenas prestaciones. Por ejemplo, una regla muy usada como es la regla del conteo, pese a ser subóptima en el caso de dependencia estadística, puede conducir a mejores prestaciones que la implementación de la regla óptima considerando decisiones independientes.

También podemos concluir (se demostró en el apartado 1.3), que en ciertos casos es posible obtener mejores prestaciones con una fusión hard, incluso subóptima, que tenga en cuenta cierto grado de dependencia estadística existente entre las decisiones individuales, con respecto al caso que se implemente una fusión soft de los datos de los cuales se derivan estas decisiones (en principio más ricos en cuanto información sobre el evento) sin tener en cuenta las características de dependencia entre ellos [15].

Parte II: Aplicaciones

Capítulo 5:
Fusión en sistemas de autenticación multi-biométrica

"Mi voz es mi pasaporte, verifícame."

- Película: Sneakers –

En la era de la tecnología de la información en la que vivimos, existe la necesidad de implementar técnicas de autenticación para mantener la seguridad en el acceso a determinados recursos. Las técnicas biométricas se basan en los rasgos fisiológicos y/o comportamientos de una persona para conseguir su autenticación. Los sistemas uni-biométricos, en los cuales un único algoritmo trabaja con una instancia de un determinado rasgo tomado por un solo sensor, presentan diversos problemas; es común para evitar estos problemas el uso de sistemas muli-biométricos, en los que se trabaja con varias instancias, rasgos y/o algoritmos. En los sistemas multi-biométricos, donde se utiliza más de un rasgo, muestra o algoritmo, es común el uso de diferentes técnicas de fusión de datos.

En este capítulo se realiza inicialmente un estudio sobre el estado del arte de las diferentes técnicas de fusión aplicadas en problemas de detección multi-biométrica. El GTS ha trabajado con este tipo de técnicas en la implementación de un sistema de autentificación biométrica basado en electrocardiogramas [21]. En este caso, con objeto de poder testear y comparar el correcto funcionamiento de las diferentes técnicas consideradas novedosas en este campo, se ha decidido incluir en esta sección aplicaciones basadas en una serie de bases de datos públicas, las cuales han sido usadas en multitud de investigaciones, convirtiéndose en un estándar para la verificación de los algoritmos.

Se han probado y testeado los diferentes métodos de fusión mediante estimación de PDFs utilizando la teoría de cópulas, cuya aplicación, como ya hemos comentado, supone una novedad en el campo del procesado de señal. Así mismo, también se ha utilizado el novedoso método que se propuso en capítulo 3 de fusión mediante integración α.

5.1. – Introducción

En la era de la tecnología de la información en la que vivimos, existe la necesidad de implementar técnicas de autenticación para mantener la seguridad en el acceso a determinados recursos [26]. Existen varias formas de verificación para ser autenticado. La mayoría de métodos están basados en el uso de algún testigo o token, como puede ser una llave, tarjeta, una contraseña… En estas técnicas un cierto grado de autoridad recae sobre el propio token usado y por lo tanto no la posee de forma íntegra la propia persona [25], lo cual no es deseable.

Las técnicas biométricas se basan en los rasgos fisiológicos y/o comportamientos de una persona para conseguir la autenticación [107]. Por ejemplo se usan las huellas dactilares, huellas palmares, la geometría de la mano, la cara, la voz, el iris, la retina, la forma de andar… Por lo tanto, en la biometría toda la autoridad recae sobre la propia persona, haciéndolo un método más seguro y fiable.

Un sistema biométrico se suele componer de cuatro módulos o etapas. Una etapa de sensado, donde uno o varios sensores adquieren los datos biométricos de un usuario; una etapa donde se procesan y extraen una serie de características o índices representativos de los datos biométricos; un módulo de verificación, que compara las características extraídas con las almacenadas en una base de datos y, mediante algún determinado clasificador o algoritmo de verificación, se genera una valoración sobre la verificación de la persona; una etapa final de toma de decisiones, donde en base a las valoraciones, se considera al usuario como un usuario autenticado en el sistema (denominado como usuario genuino) o no autenticado (denominado como usuario impostor). Dependiendo del número de rasgos usados para la autenticación o autorización los sistemas biométricos se dividen en dos tipos de sistemas, los sistemas uni-biométricos y los sistemas multi-biométricos.

En los sistemas uni-biométricos solamente se utiliza una única muestra de un determinado rasgo tomada por un solo sensor para la autenticación y habitualmente suelen sufrir una serie de problemas: los datos del sensor son ruidosos, el rasgo usado carece de universalidad, falta de poder de distinción, tasas de error inaceptables o la susceptibilidad a ataques de falsificación de la identidad [108]. Todos estos problemas se pueden superar con el uso de un sistema multi-biométrico, el cual utiliza más de un rasgo, muestra o algoritmo. La autenticación multi-biométrica puede conseguirse de diferentes formas [26]: los sistemas multi-algoritmo utilizan el mismo tipo de dato biométrico, pero procesado con diferentes algoritmos; los sistemas multi-muestra usan un único sensor que toma más de una muestra de un mismo tipo de dato biométrico; en los sistemas multi-modales se utilizan diferentes sensores para captar diferentes datos biométricos.

Este tipo de problemas multibiométricos suponen un reto tecnológico, ya que se debe trabajar con un gran volumen de datos, los cuales se caracterizan por un gran desbalance entre el número de datos de usuarios genuinos y de usuarios impostores. Además, en el diseño de un sistema de autenticación, estamos sujetos a probabilidades de falsa alarma muy bajas ($P_f < 10^{-3}$); en este área es usual denominar a la P_f como tasa de falsa aceptación (*FAR,* "*False Acceptance Rate*"), mientras que la probabilidad de detección P_d suele ser denominada como probabilidad de aceptación genuina (*GAR*, "*Genuine Acceptance Rate*").

Realizamos inicialmente un estudio sobre el estado del arte de las diferentes técnicas de fusión aplicadas en problemas de detección multi-biométrica. Con objeto de poder testear y comparar el correcto funcionamiento de las diferentes técnicas consideradas novedosas, se ha decidido incluir en esta sección aplicaciones basadas en una serie de bases de datos públicas, las cuales han sido usadas en multitud de investigaciones, convirtiéndose casi en un estándar para la verificación de los algoritmos. Así, se han probado y testeado los diferentes métodos de fusión mediante estimación de *PDFs* utilizando la teoría de cópulas, cuya aplicación, como ya hemos

comentado, supone una novedad en el campo del procesado de señal. Así mismo, también se ha utilizado el novedoso método de fusión mediante integración α, entrenado con la técnica que se ha propuesto en el presente trabajo consistente en maximización del área bajo la curva *ROC* parcial, para mostrar su adecuación y las buenas prestaciones obtenidas en problemas donde, el rango de trabajo en cuanto a falsa alarma está limitado.

5.2. – Fusión en sistemas de autenticación biométrica. Estado del arte.

En este tipo de sistemas multibiométricos se debe llevar a cabo una fusión de información (figura 5.1). La fusión de información en un sistema multibiométrico se puede encontrar en dos etapas diferenciadas, antes de la verificación, fusionando la información de los sensores o de las características extraídas de éstos, o después de la etapa verificación, fusionando las valoraciones o decisiones tomadas por los detectores, algoritmos o técnicas de verificación que componen el sistema [109].

En los sistemas multibiométricos se suele preferir realizar la fusión después de la etapa de verificación, combinando las valoraciones o scores de verificación de los detectores, algoritmos o técnicas usadas (en la literatura se les conoce por el término inglés 'matchers'), ya que es la solución de compromiso óptima entre contenido de información y facilidad de fusión. La fusión en nivel de verificación tiene por objetivo identificar mediante las diferentes valoraciones cuándo un usuario es genuino (identificado y por tanto autorizado) o es un impostor (una persona no identificada o no autorizada). Se trata de un sistema de detección con dos hipótesis: H_1, donde el usuario es genuino y H_0, donde el usuario es impostor, debiendo el sistema decidir cuál es de las dos hipótesis la correcta.

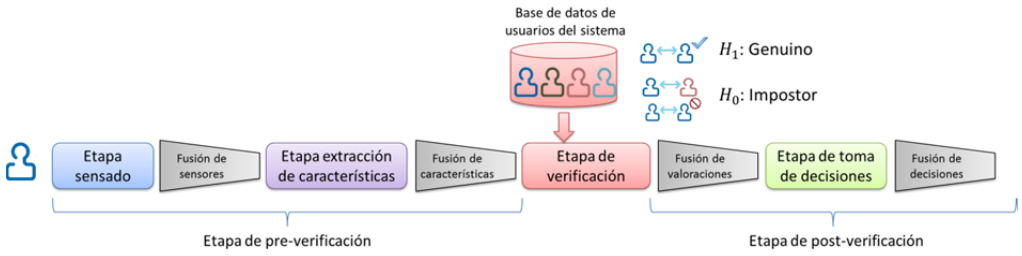


Figura 5.1 – Esquema general de un sistema de detección multi-biométrico

Los métodos de fusión de las valoraciones aportadas por los distintos "*matchers*" pueden dividirse en dos categorías. En la primera de ellas, las valoraciones son inicialmente normalizadas hacia un dominio común, donde son fusionadas mediante una determinada función de combinación. En la segunda categoría encontramos las técnicas que consideran las distintas valoraciones como características de entrada de un nuevo detector, ya pueda ser implementado, bien mediante un test estadístico

basado en la relación de verosimilitud (se debe realizar una estimación de las *PDFs* conjuntas de las valoraciones, pudiendo utilizar cualquier técnica de estimación incluidas en el capítulo 2), o bien mediante alguna técnica de clasificación binaria como puede ser una máquina de soporte de vectores, una red neuronal, una red Bayesiana..

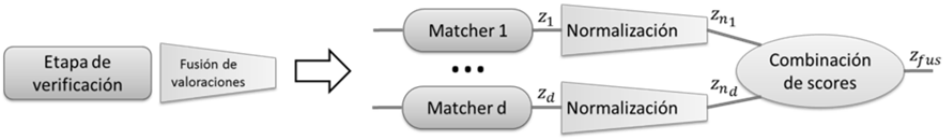


Figura 5.2 – Fusión de valoraciones mediante normalización a un espacio común y combinándolas mediante alguna función o regla.

Fusión de datos mediante combinación de scores normalizados		
Técnica de combinación	**Modalidades**	**Trabajo**
Normalización (Min-Max, decimal scaling, z-score, mediana, MAD, doble sigmoide, tanh, estimador biweight) + suma, regla del máximo y del mínimo.	1 facial, 1 huella dactilar y 1 geometría palmar.	Jain et al. [22].
Normalizaciónes (Min-Max, Tangente inversa + Min-Max) + suma.	Varias combinaciones [(1)].	Nandakuman et al. [110]
Normalizaciones (*Min-Max y Tanh*) + Suma o suma ponderada, entrenada: - Mínimos cuadrados (*LSE*). - Maximizando área bajo la curva ROC	2 facial y 1 geometría palmar.	Kar-Ann Toh et al. [25].
Normalización (probabilística, Min-Max, tanh, z-score, RHE) + suma.	• 2 facial. • 2 facial. • 2 huella dactilar, 2 facial. • 1 huella dactilar, 1 venas de los dedos.	Mingxing He et al. [24].
Normalización (min-max, z-score, tanh, median-MAD, double-sigmoid, and piecewise-linear).	• 1 facial y 1 geometría palmar. • 4 huella dactilar y 4 geometría del dedo. • 5 huella dactilar y 5 huella palmar	Ribaric y Fratic [111].
Normalización + suma y producto ponderado.	1 huella dactilar y 1 facial.	Roli y Marcialis [23].

Tabla 5.1- Estado del arte en fusión biométrica mediante normalización y combinación de las valoraciones

En los últimos años se han llevado a cabo multitud de trabajos de investigación dedicados al estudio de la fusión o integración de diversos matchers biométricos. A continuación recogemos alguno de estos trabajos, distinguiéndolos según la técnica de fusión empleada. Comenzamos recopilando algunos de los trabajos cuya técnica de

fusión está basada en la normalización de las diferentes valoraciones y su combinación para obtener un solo valor escalar (figura 5.2), el cual será umbralizado para obtener una decisión final: $\boldsymbol{z} \rightarrow \boldsymbol{z}_n \rightarrow z_{fus} = \not{f}(\boldsymbol{z}_n)$ (tabla 5.1).

Técnicas de clasificación binaria		
Clasificadores binarios	**Modos de reconocimiento**	**Trabajo**
• Discriminante lineal (*LDA*). • Red neuronal de base radial (*RBFN*). • Suma y suma ponderada.	1 facial, y 1 Iris.	Wang et al. [112].
• Vecinos próximos (*k-NN*) con cuantización. • Árbol de decisiones. • Regresión logística.	2 facial y 1 voz.	Verlinde y Chollet [113].
• Técnicas de clustering: – Fuzzy k-means (*FKM*). – Fuzzy vector quantization (FVQ). • Red neuronal de base radial (*RBFN*).	5 facial y 5 voz.	Chatzis et al. [114].
• Máquinas de soporte de vectores (*SVM*). • Clasificadores resistentes al ruido: – Clasificador Bayesiano modificado. – Clasificador lineal definido a trozos (Piece-wise Lineal, *PL*).	1 facial y 1 voz.	Sanderson y Paliwal [109].
• Árbol de decisiones. • Discriminante lineal (*LDA*).	1 facial, 1 huella dactilar y 1 geometría palmar.	Ross y Jain [115].
• Máquinas de soporte de vectores (*SVM*) lineales y con kernel radial.	(1) Varias combinaciones: • 2 facial. • 2 huellas dactilares. • 2 facial y 2 huella dactilar. • 5 facial, 2 voz. • 1 huella dactilar, 1 Iris.	Nandakuman et al. [110]
• Particle Swarm Optimization (*PSO*). • Sistema adaptativo Neuro Fuzzy (*ANFIS*). • Algoritmo genético (*GA*). • Máquinas de soporte de vectores (*SVM*).	1 facial y 1 voz.	Mazouni y Rahmoun [116].

Tabla 5.2.- Estado del arte en fusión biométrica usando técnicas de clasificación binaria

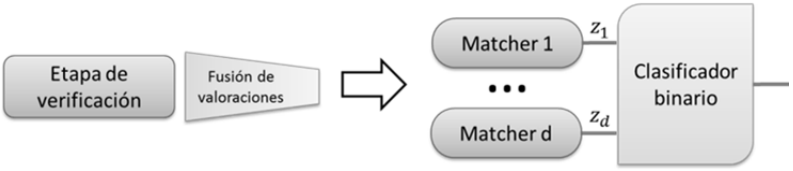


Figura 5.3 – Fusión de valoraciones utilizando una técnica de clasificación binaria

Otro criterio de fusión consiste en considerar las valoraciones de los diferentes matchers como nuevas características e implementar un nuevo detector o clasificador binario (figura 5.3). En los siguientes casos se ha utilizado alguna técnica general de clasificación binaria como detector. En este caso, dependiendo del tipo de clasificador usado la salida será directamente la decisión final o un valor escalar, el cual habrá que umbralizar para obtenerla (tabla 5.2).

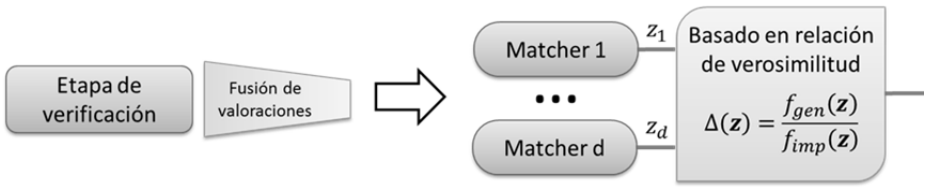


Figura 5.4 – Fusión de valoraciones mediante estimación de densidades multidimensionales

Relación de verosimilitud → Estimación de densidades multivariantes		
Técnica de estimación	**Modalidades**	**Trabajo**
• No paramétrica mediante función de kernel Gaussiano (*KDE*), ancho de banda fijado de forma empírica.	4 huellas dactilares.	Prabhakar y Jain [117].
• Paramétrica mediante modelo de mezcla de Gaussianas (*GMM*).	Varias combinaciones [(1)].	Nandakuman et al. [110].
• Modelo del producto asumiendo independencia. • Estimación paramétrica de densidad usando cópula Gaussiana.	• 2 facial y 2 huellas dactilares. • 1 facial, 1 huella dactilar y 1 geometría palmar.	Dass et al. [118].
• Modelo del producto asumiendo independencia. • Estimación paramétrica de densidad usando cópula Gaussiana, Clayton, Frank y Gumbel.	2 facial.	Iyengar et al. [119].

Tabla 5.3.- Estado del arte en fusión biométrica usando técnicas de clasificación binaria

Dentro del criterio de utilización de un nuevo detector para fusionar los datos, encontramos los trabajos en los que se ha tratado de implementar un detector óptimo basado en la relación de verosimilitud (figura 5.4) mediante la estimación de las densidades de probabilidad multivariantes de los sujetos genuinos e impostores (tabla 5.3).

5.3. – Experimentos y pruebas prácticas

En este apartado realizamos inicialmente una descripción de las bases de datos multi-biométricas utilizadas. Una vez presentadas las bases de datos, destacando sus características y peculiaridades, pasamos a realizar diferentes pruebas con las técnicas de fusión que se incluyen como novedad en el presente trabajo, consistentes en la fusión basada en la relación de verosimilitud, donde se utiliza la teoría de las cópulas en la estimación de *PDFs*, y la fusión de scores mediante la denominada integración α.

5.3.1. – Descripción de las bases de datos multi-biométricas utilizadas

Biometric DS2

La base de datos *Biometric DS2* [120] se compone de un conjunto de muestras tomadas a 333 personas utilizando 8 medidas biométricas: una imagen facial, una imagen del iris y 6 huellas dactilares de diferentes dedos. Se realizan 4 medidas de cada rasgo biométrico por cada persona. La base de datos se compone de dos sesiones *S1* y *S2.* Trabajaremos con los datos extraídos de la sesión de evaluación *S1* de esta base de datos multimodal. Utilizamos los datos del matcher facial, del iris y de la huella dactilar del dedo pulgar de la mano derecha. En este caso disponemos de 156 valoraciones de usuarios genuinos y 129368 scores de usuarios impostores, lo que supone una probabilidad a priori entre clases muy desbalanceada $\left(P_{gen} = 1.2 \cdot 10^{-3}\right)$. En la figura 5.5 podemos encontrar la representación de las valoraciones de los usuarios genuinos e impostores en un espacio ortogonal de tres dimensiones formado por los tres matchers.

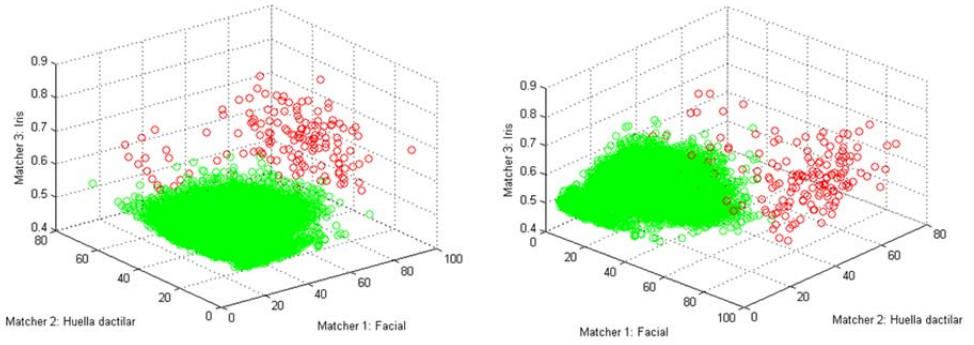


Figura 5.5 – Scatterplot de las valoraciones aportadas por los matchers asociados a la huella dactilar, a la imagen facial y a la imagen del iris

NIST-BSSR1

La base de datos *Biometric Scores Set - Release 1 (BSSR1)* [121] es un conjunto de valoraciones de similitud aportados por diferentes matchers en tres sistemas de reconocimiento multibiométrico: un sistema de reconocimiento facial, otro sistema de reconocimiento mediante huella dactilar y un sistema multimodal que combina ambos tipos de reconocimiento.

- Sistema de reconocimiento facial

En este sistema se emplean dos algoritmos (*matchers*) de reconocimiento facial diferentes. La base de datos consta de 3000 sujetos, donde se han tomado 2 mediciones por cada sujeto. Así, el número total de valoraciones total de la base de datos será de 2 *matchers*, por 3000 sujetos, por 3000 valoraciones de similitud por cada usuario (cada usuario genera una valoración por sí mismo y otra por la comparación con cada uno de los demás) y por 2 mediciones por usuario: $2 \cdot 3000 \cdot 3000 \cdot 2 = 2 \cdot 18000000$. Tendremos 18 millones de valoraciones por cada *matcher*, de los cuales 6000 corresponderán a usuarios genuinos (H_1) y el resto a usuarios impostores (H_0). Existen valoraciones con valor -1, indicando que el matcher ha tenido un error intentando procesar la imagen. Despreciamos estos scores considerando que, ante una situación de error, el sistema pedirá al usuario una nueva toma de datos biométricos [119]. En la figura 5.6 podemos encontrar una representación de las diferentes valoraciones aportadas por cada uno de los matchers para usuarios genuinos e impostores. En este caso tenemos un problema de fusión de datos desbalanceado, donde las probabilidades a priori de las hipótesis son $P_{gen} = 0.000333$ y $P_{imp} = 0.999667$.

- Sistema de reconocimiento mediante huella dactilar

En este sistema se utiliza un único matcher para dar dos valoraciones, utilizando las huellas dactilares de los dedos índices de ambas manos para la autenticación de los individuos. Existen un total de 6000 individuos en el sistema, donde se han tomado 1 muestras por cada uno de los dedos índices. Así, el número total de valoraciones por cada uno de los matchers 6000*6000 = 36000000 de los cuales 6000 corresponderán a usuarios genuinos (H_1) y el resto a usuarios impostores (H_0). Tenemos un problema de fusión de datos desbalanceado, donde las probabilidades a priori de la hipótesis H_1 es muy baja $P_{gen} = 1.6669 \cdot 10^{-4}$. En la figura 5.7 podemos encontrar una representación de las diferentes valoraciones aportadas por cada uno de los matchers para usuarios genuinos e impostores.

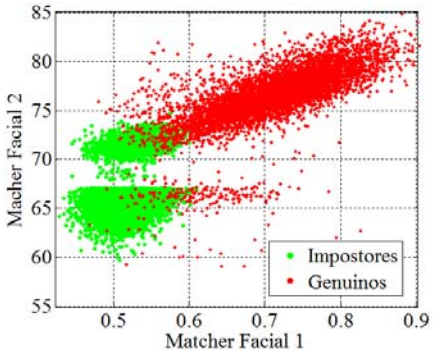


Figura 5.6 – Scatterplot de las valoraciones de los dos matchers faciales

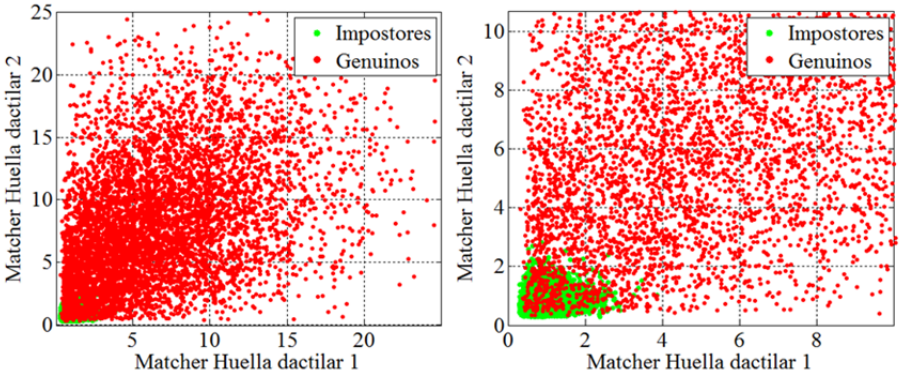


Figura 5.7 – Scatterplot de los scores de los 2 matchers de huella dactilar, donde en la figura de la derecha representa un zoom del scatterplot de la izquierda

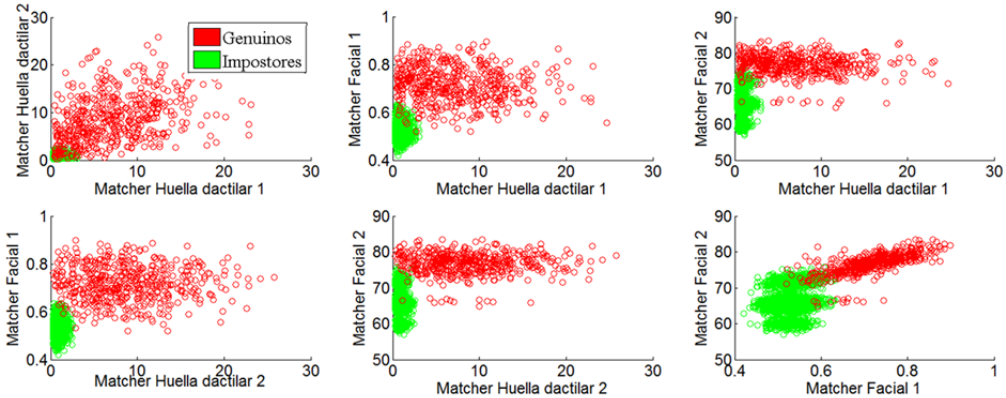


Figura 5.8 – Scatterplot de los scores del sistema de reconocimiento multimodal compuesto por dos matcher faciales y dos mediante huella dactilar.

- Sistema de reconocimiento multimodal mediante imagen facial y huella dactilar.

En este sistema se utilizan dos diferentes matchers faciales y un mismo matcher utilizado con el dedo índice de cada mano (en total 4 valoraciones por sujeto), para autenticar un grupo de 517 individuos. Así, disponemos de un conjunto de 517*517 = 267289 valoraciones por cada matcher, de las cuales 517 corresponderán a usuarios genuinos (H_1) y el resto a usuarios impostores (H_0). Así, tenemos un problema de fusión de datos desbalanceado, donde las probabilidades a priori de la hipótesis H_1 es muy baja $P_{gen} = 1.93 \cdot 10^{-3}$. En la figura 5.8 se muestran las distribuciones de estas valoraciones para los usuarios genuinos y los impostores.

5.3.2. – Fusión de datos mediante estimación de *PDFs* utilizando la teoría de cópulas

En este apartado se realizarán diferentes pruebas intentando realizar una fusión soft óptima, la cual, como ya se introdujo en el capítulo 2, viene determinada por la relación de verosimilitud. Se utiliza la teoría de cópulas para la estimación de las *PDFs,* probando diferentes densidades de cópula con objeto de seleccionar la que mejor caracterice el modelo de dependencia que siguen los datos:

$$\Delta(\boldsymbol{z}) = \frac{f_{gen}(\boldsymbol{z})}{f_{imp}(\boldsymbol{z})} \lessgtr_{H_1}^{H_0} \eta \tag{5.1}$$

Prueba 1 – Criterios de selección de cópulas

En esta primera prueba hemos utilizado la base de datos *Biometric DS2*. Se ha considerado para modelar cada una de las distribuciones conjuntas $f_{gen}(\boldsymbol{z})$ y $f_{imp}(\boldsymbol{z})$, o bien asumiendo independencia entre los scores de los matchers y utilizando el producto de funciones marginales, o bien considerando que existe dependencia y que puede ser parametrizada usando una función de cópula. Utilizaremos la cópula Gaussiana, la Student-T, Clayton, Frank, Gumbel y la basada en mezcla de Gaussianas (*cGMM*) como posibles alternativas. Así, tenemos 7 posibles modelos de probabilidad para caracterizar a cada *PDFs* conjunta bajo cada hipótesis, lo que supone un total de 49 posibles combinaciones.

Se han utilizado las cuatro técnicas de selección diferentes que se introdujeron en el apartado 2.5.2 para escoger los dos modelos de probabilidad para caracterizar las *PDFs*. La primera de ellas está basada en la técnica de selección de mínima longitud de descripción *BIC*, por lo tanto, se obtendrán 7 parámetros por cada una de las hipótesis. Las otras técnicas se basan en la medida de las prestaciones de detección a través de la curva *ROC*: se escoge la pareja que obtenga la máxima área bajo la curva *ROC* (*AUC*), o bien mediante la integración de la curva *ROC* completa, o bien

estimándola mediante el estimador *WMW*. También podemos escoger la pareja con la que se maximice el *AUC* en un el área de trabajo de interés mediante la integración de la curva *ROC* entre dos límites; hemos escogido un rango comprendido entre 0.009 % y 0.2 % de *FAR*. En cualquiera de los tres casos se debe escoger los modelos por parejas, por lo que tendremos 7 x 7 = 49 parámetros de selección.

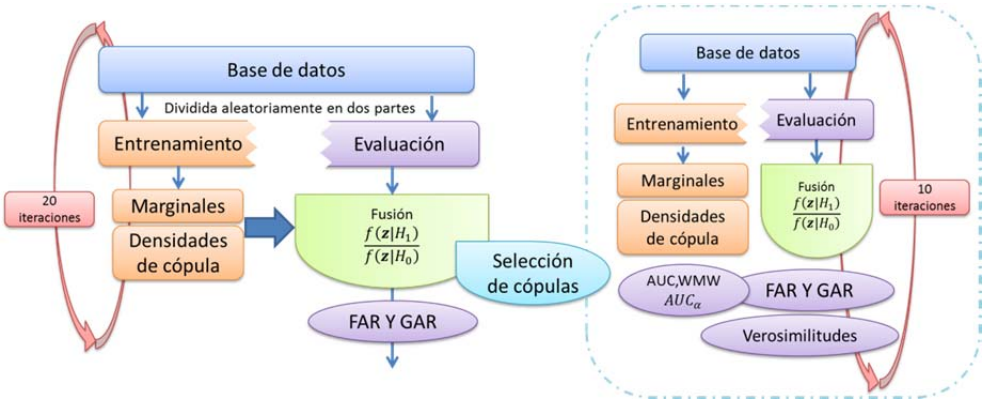


Figura 5.9 – Diagrama de flujo de las pruebas realizadas

Se han usado los parámetros de selección medios recogidos durante 10 pruebas. Se ha escogido la mitad de muestras de la base de datos original, seleccionándolas de forma aleatoria, como datos de entrenamiento en cada iteración y la otra mitad como datos de evaluación para obtener los diferentes parámetros. En el apéndice *G* se recogen las tablas con todos los parámetros obtenidos. En la figura 5.9 se puede observar un diagrama de flujo del proceso seguido.

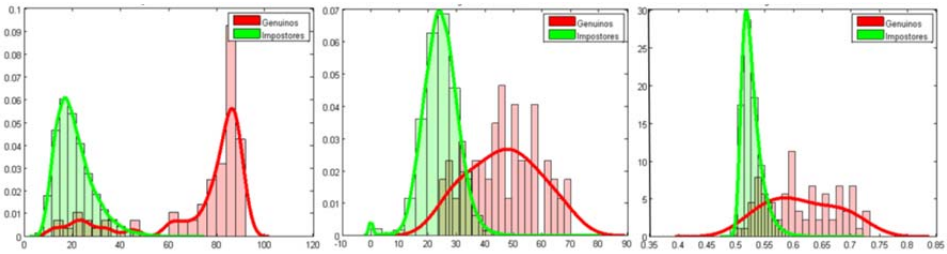


Figura 5.10 – PDFs marginales de los matchers facial, huella dactilar e iris respectivamente de izquierda a derecha

Tanto mediante la integración de la curva *ROC*, como usando el estimador *WMW*, la pareja que maximiza el *AUC* es la compuesta por el modelo basado en cópula Gaussiana para H_1 (genuino) y en el modelo del producto para H_0 (impostor). En el caso de selección maximizando el *AUC* restringida al rango de trabajo de entre 0.009 % y 0.2 % de *FAR,* la pareja escogida se basa en una cópula Gaussiana para H_1 (genuino) y en cópula *cGMM* para H_0 (impostor). El modelo escogido para H_1

(genuino) según el criterio *BIC* es el modelo del producto, mientras que para H_0 (impostor) es el basado en cópula *cGMM.*

En la figura 5.10 se pueden observar las distribuciones marginales de cada una de las valoraciones proporcionadas por los diferentes matchers. En la figura 5.11 se muestran las curvas *ROC* donde se representan tanto los matchers individuales, como los diferentes resultados de la fusión. Se puede apreciar como mediante la fusión de los tres sistemas de detección se incrementa notablemente las prestaciones obtenidas en detección. También podemos comprobar cómo, con la técnica de selección basada en una maximización del área parcial bajo la curva ROC, obtenemos mejores resultados que en el caso de maximizarla en todo el rango. En este caso, mediante el criterio de selección *BIC* también se obtienen muy buenos resultados. En la tabla 5.4 se recogen los resultados en forma de la *GAR* media obtenida para los valores de FAR de 0.01 % y 0.1 %.

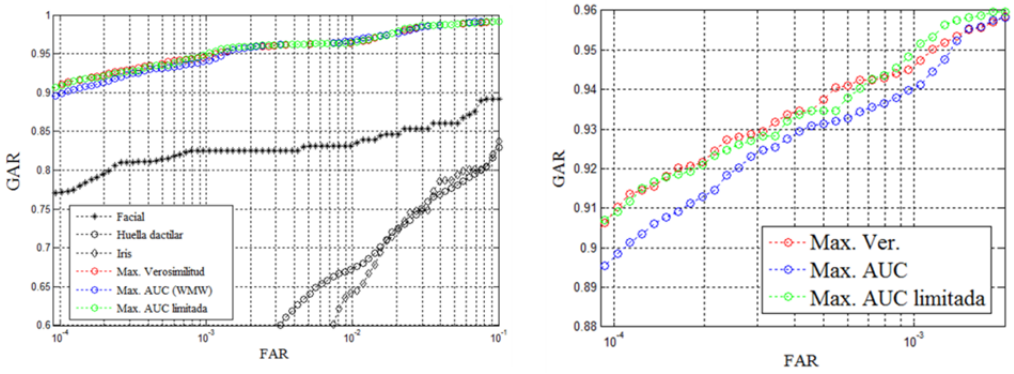


Figura 5.11 – Curvas ROC obtenidas como promedio de 20 iteraciones

	Copulas	GAR media(%) en 0.01 % FAR	GAR media(%) en 0.1 % FAR
Criterio *BIC*	H_1: Independiente H_0: GMM	**90.94**	94.62
AUC: Área bajo la curva ROC (y estadístico WMW)	H_1: Gaussiana H_0: Independiente	89.77	94.04
AUC: Área bajo la curva ROC limitada entre 0.009 y 0.02 %	H_1: Gaussiana H_0: GMM	90.87	**94.98**

Tabla 5.4 – Resultados de la fusión utilizando técnicas de selección de cópulas

Prueba 2 – Comparativa de resultados con trabajos previos usando la nueva cópula GMM

Se utiliza la base de datos *NIST-BSSR1* de verificación de caras. En este caso se realiza el mismo estudio que Iyengar et al. [119]. En este estudio se utilizan las cópulas Gaussiana, Clayton, Frank y Gumbel, así como el modelo basado en la asunción de independencia. Una de las extensiones que proponen para mejorar las prestaciones en este trabajo es la utilización de algún modelo de cópulas más potente, capaz de mejorar la caracterización de la dependencia presente. Hemos incorporado la cópula basada en la mezcla de densidades Gaussianas como modelo más complejo y potente para caracterizar complejas estructuras de dependencia.

En la figura 5.12 se muestra un diagrama de flujo de los pasos seguidos para realizar las pruebas. En este caso se ha utilizado el mismo modelo de cópula para estimar las *PDFs* bajo cada una de las hipótesis. La base de datos de las valoraciones de los matchers es dividida en cada iteración en dos partes, una para realizar el entrenamiento de las *PDFs* marginales y los parámetros de las densidades de cópula, y otra parte para evaluar los resultados de la fusión. En este caso obtenemos los resultados medios obtenidos durante 30 iteraciones.

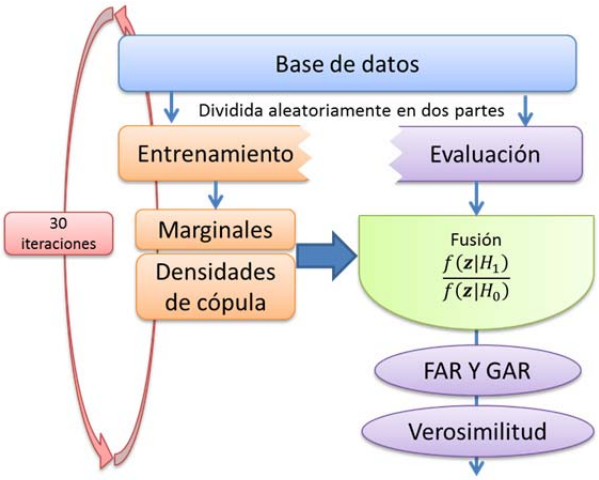


Figura 5.12 – Diagrama de flujo de las pruebas realizadas

Para el entrenamiento de los parámetros de las cópulas se ha utilizado el método de máxima verosimilitud canónica (*CML,* "*Canonical Maximum Likelihood*"), realizando primero una estimación no paramétrica de las funciones de distribución marginales, uniformizando las variables aleatorias y realizando una estimación *MLE* de los parámetros de la cópula.

Para la estimación de las funciones marginales se ha utilizado un método de estimación no paramétrica por kernel suavizado [122] (*KDE*, "*Kernel smoothing*

density estimate"). Se puede apreciar en la figura 5.13 el carácter heterogéneo en las valoraciones que proporcionan ambos matchers, tanto en el dominio o rango en el que se definen (Matcher 1: [0.4 1] Matcher 2: [45 90]), como en las diferentes *PDFs* marginales que poseen.

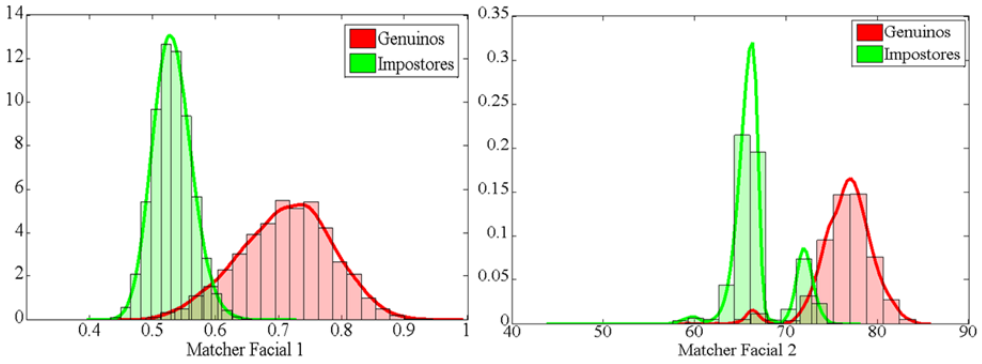


Figura 5.13 – Distribuciones de densidad de probabilidad marginales de los 2 matchers faciales

Dadas las valoraciones z_1 y z_2 de ambos matchers, la regla de fusión considerando independencia entre ellas bajo ambas hipótesis consiste en:

$$\frac{f_{gen}(z_1)\cdot f_{gen}(z_2)}{f_{imp}(z_1)\cdot f_{imp}(z_2)} \lessgtr_{H_1}^{H_0} \eta \tag{5.2}$$

Considerando que existe una dependencia bajo cada hipótesis tal que se puede caracterizar mediante una determinada función densidad de cópula $c(\cdot)$, la regla de fusión pasa a ser:

$$\frac{f_{gen}(z_1)\cdot f_{gen}(z_2)\cdot c_{gen}\left(F_{gen}(z_1),F_{gen}(z_2)\right)}{f_{imp}(z_1)\cdot f_{imp}(z_2)\cdot c_{imp}\left(F_{imp}(z_1),F_{imp}(z_2)\right)} \lessgtr_{H_1}^{H_0} \eta \tag{5.3}$$

Mostramos los resultados obtenidos en la tabla 5.5, representando las curvas *ROC* en la figura 5.14. En este caso se ha utilizado el mismo tipo de cópula para modelar ambas hipótesis. Aunque podemos ver en la tabla de resultados que atendiendo a un criterio de selección del modelo que maximiza la función de verosimilitud bajo cada hipótesis la cópula basada en una mezcla de Gaussianas sería seleccionada para modelar ambas hipótesis. Además es la que mejores resultados aporta.

Se comprueba cómo esta cópula es muy flexible, capaz de caracterizar complejas estructuras de dependencia. Mediante un correcto modelado de la información marginal y de las características de dependencia de los datos se consigue mejorar las prestaciones obtenidas en un sistema que precisa de una fusión de datos.

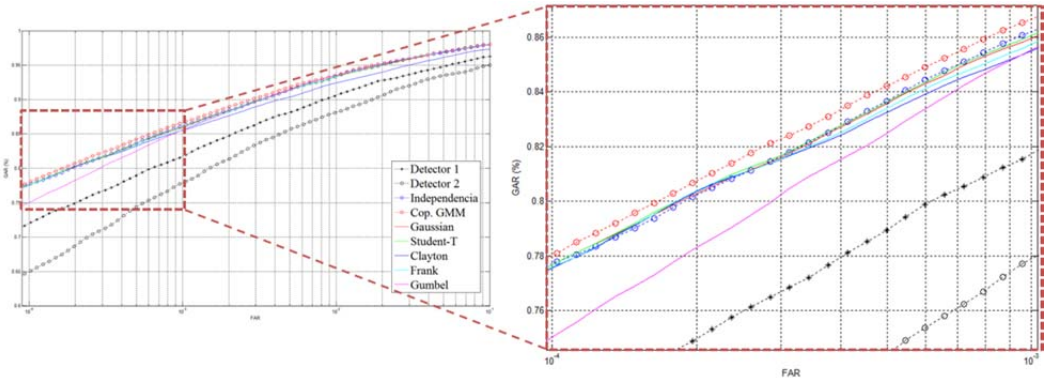


Figura 5.14 – Curvas ROC mostrando los resultados de la fusión

Matcher/Fusión	GAR media(%) en 0.01 % FAR	GAR media(%) en 0.1 % FAR	Verosimilitud logarítmica media	
			Genuinos	Impostores ($\times 10^{-2}$)
Matcher 1	72.02	81.73	-	-
Matcher 2	64.98	77.91	-	-
Independencia	77.73	86.21	-1.205	2.71
Cópula GMM	**78.01**	**86.69**	**-0.733**	**8.17**
Cópula Gaussiana	77.68	85.95	-0.777	7.20
Cópula Student-T	77.68	86.06	-0.752	7.22
Cópula Clayton	77.56	85.53	-0.891	4.57
Cópula Frank	77.57	85.76	-0.742	6.59
Cópula Gumbel	75.01	85.55	-0.783	7.78

Tabla 5.5 – Resultados de la fusión

5.3.3. – Fusión de scores mediante la integración α

En esta prueba se pretende mostrar las ventajas del método de fusión de scores dado por la integración α, entrenado con el método propuesto en el presente trabajo, consistente en buscar los parámetros que maximicen el *AUC* parcial. En este caso se ha buscado maximizar el *AUC* parcial en el rango de *FAR* $\in [0, 10^{-4}]$.

Como ya hemos visto, las valoraciones aportadas por los matchers no son heterogéneas, encontrándose cada una definida en un rango diferente. En este caso hemos utilizado tres técnicas de normalización (ver apéndice A) para transformarlas todas al dominio normalizado [0,1] (a información soft normalizada en este rango es lo que hemos denominado como score): mediante la estimación de probabilidades a posteriori $\left(s = P(z|H_1) = P_{gen}(z)\right)$, mediante la regla Min-Max y mediante una función sigmoide doble.

Una vez normalizadas las valoraciones se han fusionado, aparte de mediante la integración α, utilizando las reglas simples de la media, el producto, el mínimo y el máximo. Se ha realizado la fusión de información en los tres sistemas que componen la base de datos *NIST-BSSR1*. Comenzamos analizando los resultados obtenidos en el sistema de reconocimiento multimodal compuesto por cuatro matchers.

- Sistema de reconocimiento multimodal mediante imagen facial y huella dactilar.

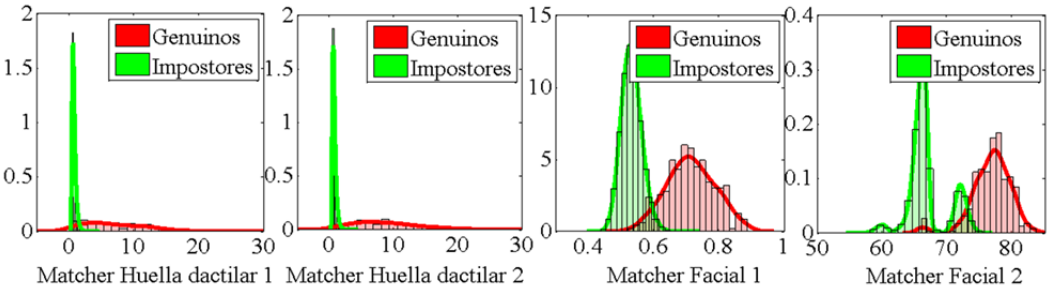


Figura 5.15 – PDFs marginales del sistema de reconocimiento multimodal compuesto por dos matcher faciales y dos mediante huella dactilar

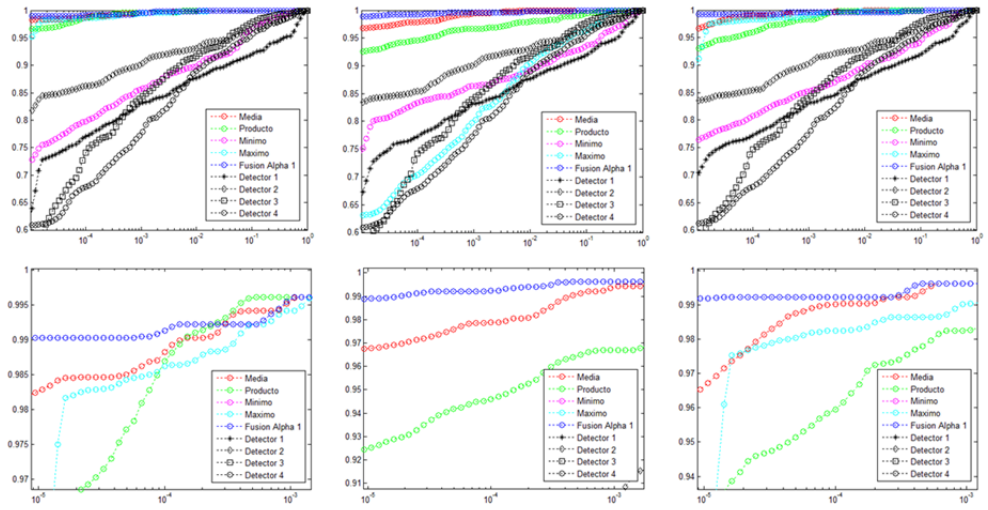


Figura 5.16 – Curvas ROC completas en la fila superior y zoom de la zona de interés en la fila inferior, para diferentes normalizaciones (Probabilidad posteriori a la izquierda, Min-Max en el medio y sigmoide doble a la derecha)

En la figura 5.15 se pueden ver las *PDFs* marginales de cada uno de los diferentes matchers que componen el sistema. Se puede observar como cada uno de ellos se encuentra definido en un rango diferente de valores. En la tabla 5.6 se muestran los resultados obtenidos por cada método de fusión con cada una de las normalizaciones,

dados por la *GAR* media obtenida a diferentes valores de *FAR*. En la figura 5.16 se muestran las curvas *ROC* obtenidas por los diferentes métodos de fusión para cada una de las normalizaciones aplicadas.

	GAR media(%) en 0.001 % FAR			GAR media(%) en 0.01 % FAR			GAR media(%) en 0.1 % FAR		
	$P_{gen}(z)$	Min-Max	Double sigmoid	$P_{gen}(z)$	Min-Max	Double sigmoid	$P_{gen}(z)$	Min-Max	Double sigmoid
Media	*97.859*	*96.768*	*96.622*	*98.823*	*97.883*	*99.013*	*99.510*	*99.378*	***99.611***
Producto	*96.229*	*92.479*	*92.897*	*98.691*	*94.591*	*95.947*	*99.609*	*96.699*	*98.256*
Min	*72.305*	*74.282*	*76.342*	*79.816*	*83.343*	*80.768*	*85.724*	*86.379*	*85.264*
Max	*97.424*	*63.030*	*90.071*	*98.622*	*70.371*	*98.252*	*99.426*	*79.996*	*99.019*
Integración α	***98.851***	***98.892***	***99.183***	***99.135***	***99.224***	***99.223***	***99.601***	***99.611***	***99.611***

Tabla 5.6 – Valores de GAR para diferentes puntos concretos de FAR y normalizaciones

Podemos observar como para todos los tipos de normalización la fusión mediante la integración α obtiene las mejores prestaciones. Como ya comentamos, tanto las funciones de media y producto, como las reglas del mínimo y del máximo, pueden ser obtenidas como caso particulares de la integración α; mediante el método de entrenamiento que se ha propuesto para la integración α, se pueden buscar el conjunto de parámetros que optimice las prestaciones de detección en el rango de trabajo deseado, sea cual sea el tipo de normalización utilizada. Observamos como la integración α proporciona una gran flexibilidad a la hora de adaptarse a diferentes datos, sea cual sean sus características marginales y de dependencia. Mediante el método de entrenamiento propuesto podemos optimizar las prestaciones obtenidas en el rango de trabajo deseado.

Realizamos ahora un entrenamiento de la integración α mediante el criterio de minimización de error cuadrático (3.28) para ilustrar como este criterio no garantiza que la integración α proporcione las mejores prestaciones posibles en problemas donde el rango de trabajo está limitado. Utilizando la normalización mediante la obtención de probabilidades a posteriori y utilizando una misma ponderación para ambos tipos de errores en el entrenamiento de la integración α ($\beta = 0.5$), hemos obtenido los valores del error *LSE* que proporcionan cada una de las técnicas de fusión (tabla 5.7). Se puede comprobar como la integración α posee el error cuadrático más pequeño. Como ya se comentó, minimizar el error cuadrático no siempre implica mejorar las prestaciones en detección, pudiendo observarlo en las curvas ROC mostradas en la figura 5.17, donde se observa como la fusión mediante la integración α en este caso obtiene peores prestaciones que el resto de técnicas.

	Media	Producto	Mínimo	Máximo	Integración α
Error cuadrático medio	31.533	145.407	140.670	37.019	**21.633**

Tabla 5.7 – Valores del error cuadrático medio tras la aplicación de cada una de las técnicas de fusión

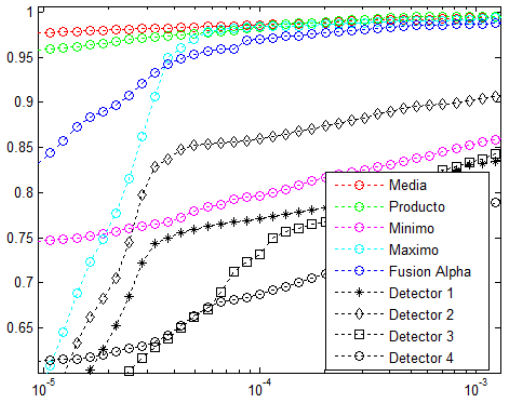


Figura 5.17 – Curvas ROC con normalización mediante la obtención de probabilidades a posteriori; integración α entrenada con criterio LSE

En la tabla 5.8 podemos comprobar como los resultados obtenidos utilizando el entrenamiento desarrollado en el presente trabajo son mucho mejores que los obtenidos por el criterio de minimización del error cuadrático.

Integración α	GAR media(%) en 0.001 % FAR	GAR media(%) en 0.01 % FAR	GAR media(%) en 0.1 % FAR
Criterio LSE	*83.767*	*97.019*	*98.693*
Criterio max. AUC parcial	***98.851***	***99.224***	***99.611***

Tabla 5.8 – Resultados obtenidos con las diferentes técnicas de entrenamiento de la integración α

En la figura 5.18 mostramos las *PDFs* de los scores fusionados mediante el producto y mediante la integración α con ambos tipos de entrenamiento, además de una representación de los datos alineados temporalmente con respecto a la hipótesis correcta para ver mejor el comportamiento de los datos fusionados. Observamos como mediante la fusión producto, la gran mayoría de scores de usuarios impostores se distribuyen con valores muy cercanos a cero, al igual que una gran cantidad de genuinos, lo cual implica que el error cuadrático de este tipo de distribución será elevado; aun así en el entorno de scores tan cercanos a cero se discriminan ambas clases con facilidad. Mediante la integración α minimizando el error cuadrático observamos cómo ahora las clases tienden a separarse, bajando mucho el error cuadrático, distribuyéndose los usuarios genuinos con valores cercanos a la unidad y

los impostores con valores cercanos a cero; pero observamos como la distribución de usuarios impostores ahora posee pequeñas densidades de valores que se alejan de cero, los cuales suponen un porcentaje de *FAR* mayor al permitido. Por último se muestra el caso de la integración α entrenada con el criterio de maximización de la *AUC* parcial, donde se observa claramente que consigue una clara separación de ambas clases, concentrando todos los scores de impostores a valores muy cercanos a cero y los scores de los usuarios genuinos esparcidos con una distribución creciente hacia valores unidad.

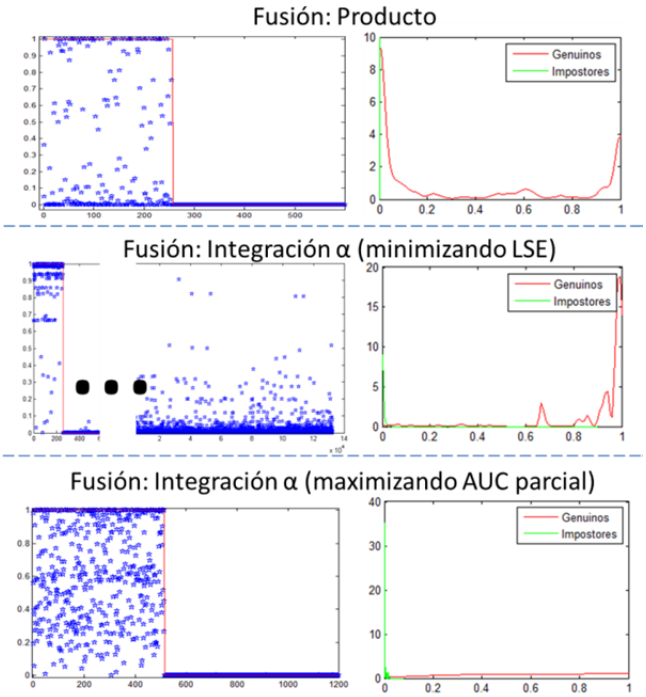


Figura 5.18 – Curvas ROC con normalización mediante la obtención de probabilidades a posteriori; integración α entrenada con criterio LSE

Mediante el entrenamiento minimizando el error cuadrático se limita el uso de la integración α a la fusión de scores, ya que precisa que los datos se encuentren necesariamente en el rango [0,1]. Mediante el entrenamiento de maximización de la *AUC* podemos utilizarla también para realizar fusión de datos soft sin restricciones. En la tabla 5.9 mostramos los resultados de la fusión de los datos sin normaliza. Podemos comprobar como, por el hecho de que cada tipo de datos se encuentre en diferente dominio, los resultados obtenidos mediante las técnicas simples empeoran. Observamos como en este caso, mediante la integración α y el entrenamiento propuesto se obtienen muy buenas prestaciones.

	GAR media(%) en 0.001 % FAR	GAR media(%) en 0.01 % FAR	GAR media(%) en 0.1 % FAR
	Sin normalizar	Sin normalizar	Sin normalizar
Media	*92.990*	*93.901*	*96.172*
Producto	*90.799*	*92.864*	*95.404*
Min	*57.969*	*73.896*	*84.135*
Max	*87.161*	*90.223*	*93.436*
Integración α	***98.093***	***99.417***	***99.611***

Tabla 5.9 – Valores de GAR para diferentes puntos concretos de FAR sin normalizar

Con objeto de comprobar la versatilidad y las buenas prestaciones que proporciona la técnica de integración α ante cualquier tipo de datos, se ha repetido los experimentos de fusión con el resto de sistemas de reconocimiento que incluye la base de datos NIST-BSSR1. A continuación se presentan los resultados obtenidos.

- Sistema de reconocimiento mediante huella dactilar

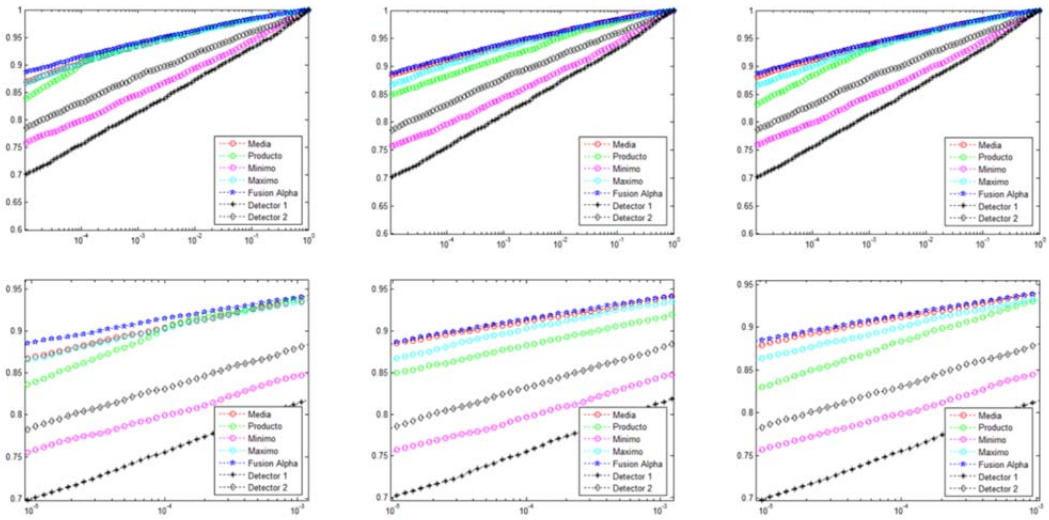


Figura 5.19 – Curvas ROC completas en la fila superior y zoom de la zona de interés en la fila inferior.

	GAR media(%) en 0.001 % FAR			GAR media(%) en 0.01 % FAR			GAR media(%) en 0.1 % FAR		
	$P_{gen}(z)$	Min-Max	Double sigmoid	$P_{gen}(z)$	Min-Max	Double sigmoid	$P_{gen}(z)$	Min-Max	Double sigmoid
Media	*86.836*	*88.400*	*88.045*	*90.402*	*91.140*	*91.158*	*93.413*	*93.883*	***93.931***
Producto	*83.772*	*84.856*	*83.131*	*90.287*	*88.288*	*88.394*	*93.775*	*91.583*	*93.097*
Min	*75.667*	*75.595*	*75.816*	*79.916*	*79.627*	*79.862*	*84.581*	*84.414*	*84.653*
Max	*86.629*	*86.601*	*86.500*	*90.216*	*90.314*	*90.018*	*93.355*	*93.323*	*93.324*
Integración α	***88.590***	***88.581***	***88.603***	***91.497***	***91.377***	***91.460***	***93.942***	***93.978***	*93.924*

Tabla 5.10 – Valores de GAR para diferentes puntos concretos de FAR y normalizaciones.

	GAR media(%) en 0.001 % FAR	GAR media(%) en 0.01 % FAR	GAR media(%) en 0.1 % FAR
	Sin normalizar	Sin normalizar	Sin normalizar
Media	*88.393*	*91.170*	*93.895*
Producto	*85.410*	*89.007*	*92.304*
Min	*75.546*	*79.740*	*84.425*
Max	*86.570*	*90.298*	*93.311*
Integración α	***88.542***	***91.409***	***94.011***

Tabla 5.11 – Valores de GAR para diferentes puntos concretos de FAR sin normalizar

- <u>Sistema de reconocimiento facial</u>

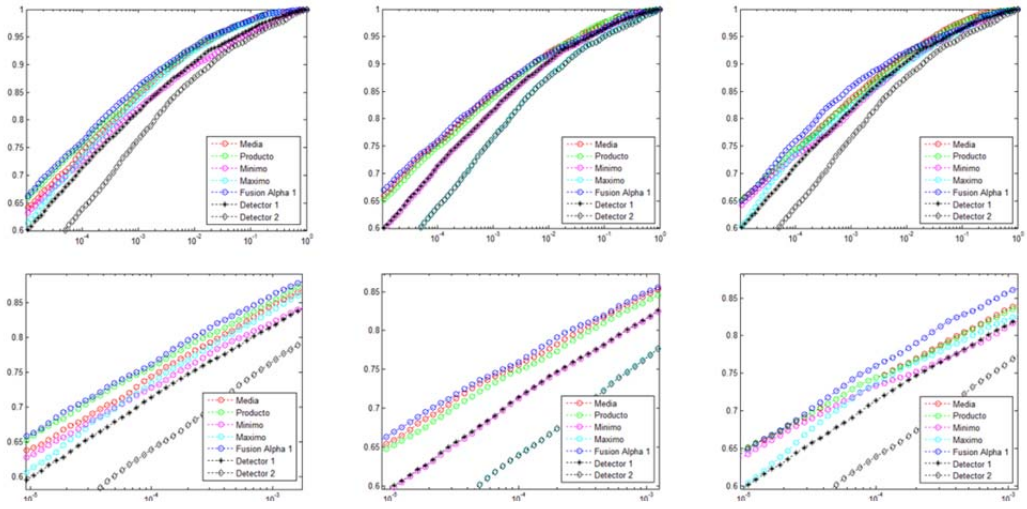


Figura 5.20 – Curvas ROC completas en la fila superior y zoom de la zona de interés en la fila inferior.

	GAR media(%) en 0.001 % FAR			GAR media(%) en 0.01 % FAR			GAR media(%) en 0.1 % FAR		
	$P_{gen}(z)$	Min-Max	Double sigmoid	$P_{gen}(z)$	Min-Max	Double sigmoid	$P_{gen}(z)$	Min-Max	Double sigmoid
Media	*63.942*	*65.598*	***64.805***	*74.336*	*75.565*	*74.423*	*84.597*	*84.534*	*83.620*
Producto	*65.551*	*64.981*	*64.892*	*75.618*	*74.874*	*74.389*	*85.112*	*83.768*	*83.230*
Min	*62.988*	*59.452*	*63.907*	*72.742*	*71.248*	*73.340*	*82.086*	*81.419*	*81.288*
Max	*60.974*	*49.004*	*59.914*	*73.415*	*63.964*	*73.612*	*83.839*	*76.501*	*82.210*
Integración α	***66.035***	***66.599***	*64.538*	***76.134***	***75.921***	***75.992***	***86.023***	***84.898***	***85.774***

Tabla 5.12 – Valores de GAR para diferentes puntos concretos de FAR y normalizaciones

	GAR media(%) en 0.001 % FAR	GAR media(%) en 0.01 % FAR	GAR media(%) en 0.1 % FAR
	Sin normalizar	Sin normalizar	Sin normalizar
Media	*50.752*	*65.018*	*77.320*
Producto	*65.135*	*74.904*	*83.998*
Min	*59.807*	*71.365*	*81.538*
Max	*49.176*	*63.914*	*76.416*
Integración α	***66.799***	***75.971***	***84.995***

Tabla 5.13 – Valores de GAR para diferentes puntos concretos de FAR y sin normalizar.

5.4. – Conclusiones

En este capítulo se ha realizado un completo estudio sobre el estado del arte de las diferentes técnicas de fusión aplicadas en problemas de detección multi-biométrica. El *GTS* ha trabajado con este tipo de técnicas en la implementación de un sistema de autentificación biométrica basado en electrocardiogramas [21]. En este caso, con objeto de poder testear y comparar el correcto funcionamiento de las diferentes técnicas consideradas novedosas en este campo, se ha decidido incluir en esta sección aplicaciones basadas en una serie de bases de datos públicas, las cuales han sido usadas en multitud de investigaciones, convirtiéndose en un estándar para la verificación de los algoritmos.

Así, se han probado y testeado los diferentes métodos de fusión mediante estimación de *PDFs* utilizando la teoría de cópulas, cuya aplicación, como ya hemos comentado, supone una novedad en el campo del procesado de señal. Así mismo, también se ha utilizado el novedoso método que se propuso en capítulo 3 de fusión mediante integración α.

Hemos comprobado que la utilización de la teoría de las cópulas para la estimación de *PDFs* multivariantes, al permitirnos separar el modelado de la información marginal del modelado de la estructura de dependencia que poseen los datos, simplifica el proceso de caracterización de los datos, proporcionándonos una herramienta muy potente para incluir la información de dependencia entre los distintos datos heterogéneos que se pretenden combinar o fusionar en un problema de detección. Así, hemos visto como la correcta caracterización de la dependencia nos ayuda a mejorar las prestaciones obtenidas mediante la fusión de diversas fuentes de información.

Se ha aplicado la fusión mediante la integración α en diferentes sistemas, con diversos escenarios de normalización de datos, contrastando sus prestaciones con respecto a las técnicas simples de fusión. Se ha comprobado cómo, utilizando el método de entrenamiento propuesto, esta técnica permite adaptarse a las distintas características que posean los datos y proporcionar una mejora en las prestaciones de detección mediante la fusión de datos. Se han comparado los métodos de entrenamiento basados en el criterio de la minimización del error cuadrático medio y en el criterio de maximización de la *AUC* parcial, demostrando como este segundo método proporciona mejores resultados.

Así, se ha demostrado que tanto la técnica de fusión soft basada en la *LR,* utilizando funciones de cópula en la estimación de las *PDFs*, como la técnica de fusión soft basada en la integración α (pudiendo utilizarse tanto para realizar una fusión soft genérica, como para la fusión de scores) pueden mejorar las prestaciones obtenidas en un sistema de detección con diversas fuentes de información, mediante una correcta caracterización o adaptación a los diversos datos heterogéneos y dependientes.

Capítulo 6: Detección de señal desconocida en presencia de ruido aleatorio

"Cualquier poder, si no se basa en la unión, es débil"
- Jean de La Fontaine –

La detección de señales en presencia de ruido aleatorio de fondo es un problema clásico en teoría de detección. Para poder llevar a cabo esta tarea de una forma óptima es necesario conocer determinadas características, tanto del evento que se pretende detectar, como del ruido en el cual se encuentra inmerso. Es posible encontrarse con el caso en el que se deben detectar eventos asociados a señales totalmente desconocidas, produciéndose en presencia de ruidos aleatorios perfectamente conocidos y caracterizados por su PDF. En este tipo de situación se centra el presente estudio.

Inicialmente se realiza una revisión del estado del arte de las técnicas de detección comúnmente usadas en estos escenarios. Posteriormente se plantea el problema de la detección de señal desconocida en presencia de ruido conocido como un problema de detección One-Class y se deriva una nueva técnica de detección, donde se hace uso de la teoría de cópulas para incorporar información estadística sobre dependencia entre las muestras.

Se utilizará esta técnica de detección basada en funciones de cópulas en la detección de eventos sonoros como ejemplo de área dónde nos encontramos con el problema de detección de señal desconocida en presencia de ruido aleatorio. Así mismo, se realiza un estudio de fusión de más de un canal de audio aportado por más de un micrófono como método para incrementar las prestaciones de un sistema de detección de eventos acústicos.

6.1. – Introducción

La detección de señales en presencia de ruido aleatorio de fondo es un problema clásico en teoría de detección. La teoría de detección se basa en tomar una decisión entre dos posibles utilizando información extraída de un conjunto de medidas. En este caso, los algoritmos de detección tratan de discernir cuando existe solamente "ruido" o cuando existe una "señal (evento) enmascarada con ruido". En muchos sistemas, a la hora de tomar determinadas decisiones o para la extracción de determinada información, nos encontramos con este tipo de problema. Ejemplos de estos sistemas pueden ser sistemas de comunicación, radar, biomedicina, procesamiento de imagen, etc. En todos ellos se precisa la capacidad de poder discernir cuando un determinado evento considerado de interés ocurre, existiendo siempre un ruido de fondo contaminando las medidas [123].

Para poder llevar a cabo esta tarea de una forma óptima es necesario conocer determinadas características, tanto del evento que se pretende detectar, como del ruido en el cual se encuentra inmerso. Así pues, el grado de dificultad que entraña la implementación de un detector está inversamente relacionado con el conocimiento sobre la señal y ruido del que disponemos. Este conocimiento, en teoría de detección,

se recoge en términos del correcto modelado de las funciones de densidad de probabilidad (*PDF*) involucradas. El caso ideal ocurre cuando conocemos las *PDFs* correspondientes a las variables aleatorias asociadas a los eventos y al ruido, situación que nos permite la obtención del detector óptimo [5]. Cuando no se tiene un total conocimiento sobre dichas *PDFs*, otras posibles soluciones de detección pueden ser adoptadas, y aunque probablemente no sean óptimas, pueden ser usadas para implementar un detector adecuado para la resolución del problema. Muchas de estas soluciones se basan en la presunción de una determinada *PDF* de forma que se caractericen los eventos y/o el ruido de una forma más o menos correcta aunque no se ajuste fielmente a la realidad.

Muy común en este tipo de problemas es encontrarse con el caso en el que se deben detectar eventos asociados a señales totalmente desconocidas, produciéndose en presencia de ruidos aleatorios perfectamente conocidos y caracterizados por su *PDF*. En este tipo de situación se centra el presente estudio.

Estado del arte

Un problema de detección de señal desconocida ($\boldsymbol{s}$) inmersa en un ruido conocido ($\boldsymbol{w}$) deriva en un test de hipótesis aplicado a un vector de observaciones ($\boldsymbol{y}$). La hipótesis denominada H_0 hace referencia a la presencia de sólo ruido y la hipótesis H_1 indica la presencia de señal contaminada con ruido. La regla de decisión se basa en la umbralización de un estadístico $\Delta(\boldsymbol{y})$ extraído del vector de observaciones:

$$\begin{aligned} &H_0\colon\ \boldsymbol{y} = \boldsymbol{w} \\ &H_1\colon\ \boldsymbol{y} = \boldsymbol{s} + \boldsymbol{w} \end{aligned} \qquad \Delta(\boldsymbol{y}) \underset{H_0}{\overset{H_1}{\gtrless}} \lambda \tag{6.1}$$

Para la detección de un determinado evento en presencia de ruido aleatorio de fondo, dejando de lado el popular filtro adaptado [124], el cual requiere de un perfecto conocimiento de la forma de onda de la señal a detectar, diferentes métodos existen para lidiar con el problema práctico del desconocimiento total o parcial de la señal a detectar. Así, el denominado detector de energía ("*Enegy Detector, ED*") y el filtro adaptado al subespacio ("*Matched subspace filter, MSF*") se han utilizado ampliamente en diversas aplicaciones cuando no se tiene un total conocimiento de la señal que se pretende detectar (detección de novedad) [125].

El detector de energía [126], [127] es óptimo cuando, tanto señal como ruido de fondo son procesos aleatorios Gaussianos independientes e incorrelados entre sí. En el caso de señales aleatorias completamente desconocidas es, al menos, un test *GLRT* [5].

$$\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R_w} = \sigma_w^2 \boldsymbol{I}) \qquad E_{t_n} = \frac{\boldsymbol{y}^T \boldsymbol{y}}{\sigma_w^2} \underset{H_1}{\overset{H_0}{\lessgtr}} \lambda \tag{6.2}$$

El detector de energía se ha aplicado en el caso de que el ruido de fondo es Gaussiano, pero no independiente (coloreado), utilizando una transformación de pre-blanqueado [128]. A esta extensión se le ha denominado detector de energía con preprocesado ("*Preprocessed-Energy Detector, PED*"):

$$\boldsymbol{w}\sim\mathcal{N}(\mathbf{0},\boldsymbol{R_w}\neq\sigma_w^2\boldsymbol{I}) \qquad \begin{matrix}\boldsymbol{y}^T\boldsymbol{R_w^{-1}}\boldsymbol{y}\lessgtr_{H_1}^{H_0}\lambda \\ \boldsymbol{y}_p=\boldsymbol{R_w}^{-1/2}\boldsymbol{y}\ \rightarrow\sigma_{w_p}^2=1,\ \ \boldsymbol{y}_p{}^T\boldsymbol{y}_p\lessgtr_{H_1}^{H_0}\lambda\end{matrix} \tag{6.3}$$

El filtro adaptado al subespacio [129] también es óptimo en el caso en el que el ruido sea Gaussiano e independiente, pero en este caso se supone que la señal a detectar recae en un determinado subespacio modelado por una matriz $\boldsymbol{H}$.

$$\begin{matrix}\boldsymbol{w}\sim\mathcal{N}(\mathbf{0},\boldsymbol{R_w}=\sigma_w^2\boldsymbol{I}) \\ \boldsymbol{s}\colon\boldsymbol{H\theta}\end{matrix} \qquad \frac{\boldsymbol{y}^T\boldsymbol{H}(\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T\boldsymbol{y}}{\sigma_w^2}\ \lessgtr_{H_1}^{H_0}\ \lambda \tag{6.4}$$

Un gran esfuerzo en la investigación en este campo ha estado enfocado en obtener generalizaciones del detector de energía cuando el ruido de fondo es no Gaussiano y/o no independiente. Muchas de las soluciones obtenidas se basan en blanquear el ruido primero, seguido de una función no lineal escalar aplicada a todos los componentes del vector de observación. Diversas alternativas en funciones no lineales se han propuesto [130][131], dando lugar a detectores comúnmente denominados como "*Generalized Energy Detectors" (GED).*

En el caso de ruido no Gaussiano independiente e idénticamente distribuido (*i.i.d),* cabe destacar la extensión del detector de energía [19] que realiza un preprocesado de las muestras para conseguir que el ruido sea Gaussiano e *i.i.d,* contexto donde el detector de energía es al menos un test *GLRT*. Utiliza una función no lineal $g(\cdot)$ encargada de convertir la variable aleatoria que caracteriza al ruido *w* con distribución de probabilidad arbitraria $f_w(\cdot)$, en una variable aleatoria Gaussiana estándar (media nula y varianza unidad).

$$\begin{matrix}\boldsymbol{w}\sim f_w(\cdot)\neq\mathcal{N}(\cdot) \\ \text{Muestras } \textit{i.i.d}.\end{matrix} \qquad \begin{matrix}u=g(w)=\Phi^{-1}\big(F_w(w)\big) \\ g(\boldsymbol{y})^T g(\boldsymbol{y})\lessgtr_{H_1}^{H_0}\lambda\end{matrix} \tag{6.5}$$

El test de Rao [5] también hace uso de una función no lineal $g'(\cdot)$ para implementar una mejora del filtro adaptado al subespacio cuando el de ruido es aleatorio no Gaussiano *i.i.d*:

$$\begin{matrix}\boldsymbol{w}\sim f_w(\cdot)\neq\mathcal{N}(\cdot) \\ \boldsymbol{s}\colon\boldsymbol{H\theta} \\ \text{Muestras } \textit{i.i.d}.\end{matrix} \qquad \begin{matrix}\dfrac{g'(\boldsymbol{y})^T\boldsymbol{H}(\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T g'(\boldsymbol{y})}{P_{g'}(w)}\ \lessgtr_{H_1}^{H_0}\ \lambda \\ P_{g'}(w)=E[g'^2(w)]=\displaystyle\int_{-\infty}^{\infty}g'^2(w)\,f_w(w)\,dw\end{matrix} \tag{6.6}$$

La dependencia estadística presente en el caso de ruido no Gaussiano ha sido problemática a la hora de derivar detectores que lidien con ella de forma eficaz. Así, en la literatura encontramos una escasez de esfuerzos directamente enfocados a resolver este problema. En contraste, ha tenido una gran atención la investigación centrada en encontrar transformaciones lineales que a partir de vectores cuyas componentes son dependientes, obtienen vectores con componentes independientes [132], [133]. De hecho, el análisis de componentes independientes se utiliza actualmente en multitud de aplicaciones, como por ejemplo la separación ciega de fuentes (*Blind Source Separation, BSS*).

Esta técnica *ICA* se ha aplicado para obtener una extensión del detector de energía [19] pre-procesando el vector de observaciones dependientes, para obtener vectores con componentes independientes. Para ello se utiliza una transformación lineal $\boldsymbol{U}$ del vector de observaciones $\boldsymbol{y}$. Una vez independizadas las muestras, se utiliza una transformación no lineal $g(\cdot)$ para Gaussianizarlas:

$$\begin{array}{c} \boldsymbol{w} \sim f_w(\cdot) \neq \mathcal{N}(\cdot) \\ \text{Muestras no independientes} \end{array} \qquad \begin{array}{c} g\left(\boldsymbol{Q}\boldsymbol{R}_{\boldsymbol{w}}^{-1/2}\boldsymbol{y}\right)^T g\left(\boldsymbol{Q}\boldsymbol{R}_{\boldsymbol{w}}^{-1/2}\boldsymbol{y}\right) \lessgtr_{H_1}^{H_0} \lambda \\ \boldsymbol{U} = \boldsymbol{Q}\boldsymbol{R}_{\boldsymbol{w}}^{-1/2} \end{array} \tag{6.7}$$

El mismo proceso de independizar primero las muestras mediante una transformación lineal, y después utilizar una transformación no lineal para Gaussianizarlas, se ha utilizado en [20] como extensión del filtro adaptado al subespacio cuando el ruido de fondo es no Gaussiano y no independiente:

$$\begin{array}{c} \boldsymbol{w} \sim f_w(\cdot) \neq \mathcal{N}(\cdot) \\ \boldsymbol{s}\text{: } \boldsymbol{H}\boldsymbol{\theta} \\ \text{Muestras no independientes} \end{array} \qquad \begin{array}{c} \boldsymbol{u} = \boldsymbol{U}\boldsymbol{w} = \boldsymbol{Q}\boldsymbol{R}_{\boldsymbol{w}}^{-1/2}\boldsymbol{w} \qquad \boldsymbol{H}_U = \boldsymbol{U}\boldsymbol{H} \\ \boldsymbol{P}_U = \boldsymbol{H}_U(\boldsymbol{H}_U^T\boldsymbol{H}_U)^{-\mathbf{1}}\boldsymbol{H}_U^T \quad P_g(u) = E[g^2(u)] \\ \frac{g(\boldsymbol{U}\boldsymbol{y})^T \cdot \boldsymbol{P}_U \cdot g(\boldsymbol{U}\boldsymbol{y})}{P_g(u)} \lessgtr_{H_1}^{H_0} \lambda \end{array} \tag{6.8}$$

Se observa como las extensiones del detector de energía y del filtro adaptado al subespacio se basan en pre-procesar la señal con objeto de transformar las muestras de ruido para que estas sean Gaussianas e *i.i.d*, cumpliendo así la premisa con la que se derivan ambos detectores como soluciones óptimas.

Motivación y objetivos

Este capítulo de la tesis se centra en el estudio y mejora de detectores para aplicarlos en problemas de detección de señales desconocidas en presencia de ruido aleatorio no Gaussiano y no independiente. Más concretamente, en el seno del grupo de investigación en el que se ha desarrollado la presente tesis, se está interesado en utilizarlos en sistemas de detección de eventos acústicos.

Como ya hemos comentado, la principal problemática encontrada en la investigación en este área radica en considerar la dependencia estadística presente en el caso de ruido no Gaussiano. Utilizar las extensiones propuestas en la literatura que se basan en la independización de muestras mediante una técnica *ICA* en aplicaciones donde se considera una alta dimensionalidad en el vector de observación, se pueden tornar muy complejas, requiriendo altos costes computacionales. Se ha estudiado la posibilidad que brinda la teoría de cópulas de poder caracterizar de una forma simple las características de dependencia como posible solución al problema. Se ha pretendido obtener una alternativa a las extensiones del detector de energía clásico en el contexto de ruido de fondo no Gaussiano y no independiente, donde se tenga en cuenta directamente la información de dependencia de una forma sencilla, sin pasar por un proceso que involucre una técnica *ICA*.

Otro de los objetivos que se ha propuesto en esta investigación es la de estudiar la aplicación de los detectores en un sistema multicanal, más concretamente en nuestro caso, centrándose en un sistema donde se utilice más de un micrófono para la detección de eventos acústicos.

Organización del capitulo

En el primer apartado se define el concepto de detección *One-Class*. Se plantea el problema de la detección de señal desconocida en presencia de ruido conocido como un problema de detección *One-Class* y se deriva una nueva técnica de detección, donde se hace uso de la teoría de cópulas para incorporar información estadística sobre dependencia entre las muestras. Posteriormente se demuestra que, tanto el detector de energía, como alguna de sus extensiones son casos particulares de esta nueva técnica de detección.

En un segundo apartado se introduce la detección de eventos sonoros como ejemplo de área donde nos encontramos con el problema de detección de señal desconocida en presencia de ruido aleatorio. Así mismo, se comentan distintos sistemas y aplicaciones en las que se trabaja en el GTS y en las que es necesaria una detección de eventos acústicos. Se define el marco experimental de detección de eventos sonoros para testear las prestaciones del detector propuesto, contrastando su bondad con respecto a los detectores de energía clásicos. Se argumenta el uso de los detectores en un sistema multicanal, comprobado que combinando toda la información que aportan todos los canales se puede obtener una mejora en las prestaciones con respecto al caso monocanal. Adicionalmente se proponen unas simplificaciones en los detectores, las cuales dan lugar a la posibilidad de implementar un sistema de detección multicanal basado en la fusión de detectores individuales en cada canal.

6.2. – Detector *One-Class* basado en función de cópula

6.2.1. – Detección One-Class

Un enfoque diferente al seguido por los detectores de energía consiste en reconocer que el problema de detección de señal desconocida en ruido conocido es, conceptualmente, un problema de clasificación *One-Class* [134], donde la clase "ruido" puede ser aprendida o entrenada (tanto en el escenario donde el ruido sea o no Gaussiano, o bien exista independencia o dependencia entre las muestras), mientras que no existe ninguna opción de aprender la clase "señal". Este tipo de enfoque también puede considerarse dentro del problema denominado como detección de novedades [125]. Ambos enfoques pueden relacionarse partiendo del marco en el que se utiliza un test óptimo basado en la relación de verosimilitud [124]:

$$\Delta(\boldsymbol{y}) = \frac{f(\boldsymbol{y}|H_1)}{f(\boldsymbol{y}|H_0)} \lessgtr_{H_1}^{H_0} \lambda \tag{6.9}$$

Sin embargo, al considerar la señal totalmente desconocida, una opción para simplificar el problema es asumir que la distribución $f(\boldsymbol{y}|H_1)$ es constante [134], derivando en un test *One-Class*:

$$\Delta'(\boldsymbol{y}) = f(\boldsymbol{y}|H_0)^{-1} \lessgtr_{H_1}^{H_0} \lambda \quad \rightarrow \quad LLR = -\ln\big(f(\boldsymbol{y}|H_0)\big) \lessgtr_{H_1}^{H_0} \ln(\lambda) \tag{6.10}$$

Notar que $f(\boldsymbol{y}|H_0)^{-1}$ es una medida del grado de desviación que posee el vector de observaciones $\boldsymbol{y}$ con respecto a la clase aprendida "ruido", recogida en la función $f(\boldsymbol{w}|H_0)$. Al existir una señal (pese a ser esta desconocida) presente junto al ruido, es razonable pensar que el vector de observaciones se perturbará de forma que se aleje del rango de valores que con mayor probabilidad se dan en el caso de que sólo exista ruido.

6.2.2. – Uso de funciones de cópula en el detector One-Class: detector COCD.

Centrándonos en implementar un detector basado en el enfoque *One-Class*, el principal problema pasa a ser la estimación de la *PDF* multidimensional de la clase "ruido" $f(\boldsymbol{w})$. Si asumiéramos independencia entre las muestras del vector de observación $\boldsymbol{y}$, $f(\boldsymbol{w})$ se obtendría simplemente como el producto de la *PDF* marginal del ruido $f_w(\cdot)$ aplicada a todas sus muestras. Multitud de métodos unidimensionales, tanto paramétricos como no paramétricos, existen para la estimación de esta *PDF* marginal. Sin embargo, la inclusión de la posible dependencia que pueda existir entre las muestras supone un problema. Existen algunos métodos multidimensionales no paramétricos para la estimación de la función $f(\boldsymbol{w})$ [135], pero no es tan obvio encontrar extensiones paramétricas. Una posibilidad muy usada es la utilización de un modelo Gaussiano multivariante, donde una matriz de correlación parametriza la

dependencia. Otra posibilidad también ampliamente usada para modelar *PDFs* multidimensionales es la utilización de un modelo basado en mezcla de Gaussianas, aunque la elevada dimensionalidad del vector de observaciones utilizado hace que el modelo sea muy complejo y poco práctico de utilizar.

Una posibilidad más flexible es el uso de las funciones de cópula para modelar y parametrizar la dependencia. Hemos denominado a este tipo de detectores como "*Cópula-Based One Class Detectors, COCD*". Así, podemos utilizar técnicas paramétricas o no paramétricas para estimar la *PDF* marginal del ruido de forma aislada, y combinarla con una función de densidad de cópula que parametriza la dependencia entre muestras, para obtener un modelo estadístico para $f(\boldsymbol{w}|H_0)$:

$$\boldsymbol{w} = [w_1 \dots w_N] \sim f(\boldsymbol{w}) = \left(\prod_{i=1}^{N} f_w(w_i)\right) \cdot c\big(F_w(w_1), \dots, F_w(w_N)\big) \tag{6.11}$$

De entre todas las posibles funciones de densidad de cópula paramétricas que podemos utilizar (capitulo 2), nos centraremos en la cópula Gaussiana (2.73), debido a su simpleza y su conexión directa con los detectores de energía clásicos. Así, la función $f(\boldsymbol{w})$ se puede expresar como:

$$f(\boldsymbol{w}) = \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \cdot \exp\left(-\frac{g(\boldsymbol{w})^T \cdot (\boldsymbol{\Sigma}^{-1} - \boldsymbol{I}) \cdot g(\boldsymbol{w})}{2}\right) \cdot \prod_{i=1}^{N} f_w(w_i) \tag{6.12}$$

donde, $g(\boldsymbol{w}) = [g(w_1) \dots g(w_N)]^T$ y $g(w_1) = \Phi^{-1}\big(F_w(w_i)\big)$. $\Phi(\cdot)$ es la *CDF* de una variable aleatoria Gaussiana estándar. Así pues, $g(\cdot)$ es una función no lineal que transforma las componentes originales en variables aleatorias Gaussianas estándar. La información de dependencia queda parametrizada en la matriz de correlación estándar:

$$\boldsymbol{\Sigma}(n,m) = \begin{cases} 1 & ,n = m \\ \rho_{nm} = \rho_{mn} < 1 & ,n \neq m \end{cases} \tag{6.13}$$

Aplicando 6.12 en 6.10, derivamos el nuevo detector propuesto, el cual hemos denominado como "*Gaussian Copula-based One-Class Detector (Gaussian COCD)*":

$$g(\boldsymbol{y})^T(\boldsymbol{\Sigma}^{-1} - \boldsymbol{I})g(\boldsymbol{y}) - 2 \cdot \sum_{i=1}^{N} \ln\big(f_w(y_i)\big) \lessgtr_{H_1}^{H_0} ln\left(\frac{\lambda^2}{|\boldsymbol{\Sigma}|}\right) \tag{6.14}$$

Se puede observar como el detector COCD propuesto, separa de una forma clara la información marginal de la información de dependencia estadística que posee el ruido. Notar que en el caso en el que exista independencia $\boldsymbol{\Sigma} = \boldsymbol{I}$, y sólo estará presente el segundo término, el cual equivale al producto de las marginales. El primer término captura la posible dependencia lineal entre las muestras de ruido y estará presente sólo si $\boldsymbol{\Sigma} \neq \boldsymbol{I}$.

6.2.3. – Comparación del detector COCD con los detectores de energía clásicos

Se puede demostrar de una forma directa que los detectores de energía clásicos son un caso particular del detector *COCD* Gaussiano propuesto en (6.14) cuando se asumen las mismas premisas en su derivación.

Empecemos considerando que el ruido es Gaussiano y no independiente, con matriz de covarianza $\boldsymbol{R}$, la cual está relacionada con la matriz de correlación anteriormente definida a través de: $\boldsymbol{R} = \sigma^2\boldsymbol{\Sigma}$. Caracterizamos la *PDF* del vector de observaciones del ruido $f(\boldsymbol{w})$ por una distribución Gaussiana multivariante, con marginales $f_w(w)$ idénticamente distribuidas, Gaussianas de media nula y varianza σ^2:

$$f(\boldsymbol{w}) = \frac{1}{\sqrt{|\boldsymbol{\Sigma}|(2\pi\sigma)^N}} e^{-\frac{\boldsymbol{w}^T\cdot\boldsymbol{\Sigma}^{-1}\cdot\boldsymbol{w}}{2\cdot\sigma^2}} \quad \rightarrow \quad f_w(w) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\mathrm{w}^2}{2\cdot\sigma^2}} \tag{6.15}$$

En este caso la función para transformar las muestras de ruido en variables aleatorias Gaussianas estándar viene dada por un simple escalado de la varianza $g(w) = \frac{w}{\sigma}$.

Si sustituimos 6.15 en 6.14 para obtener el detector COCD propuesto, considerando un vector de observación $\boldsymbol{y}$, llegamos a la siguiente expresión:

$$\boldsymbol{y}^T \cdot \boldsymbol{\Sigma}^{-1} \cdot \boldsymbol{y} \lessgtr_{H_1}^{H_0} \sigma^2 \ln\left(\frac{\lambda^2}{|\boldsymbol{\Sigma}|(2\pi\sigma^2)^N}\right) \tag{6.16}$$

la cual se comprueba que es equivalente al test de hipótesis utilizado por el detector *PED* (6.3), extensión del detector de energía clásico. En el caso de que el ruido Gaussiano sea independiente, $\boldsymbol{\Sigma} = \boldsymbol{I}$, y obtenemos la expresión del detector de energía clásico:

$$\boldsymbol{y}^T\boldsymbol{y} \lessgtr_{H_1}^{H_0} \sigma^2 \ln\left(\frac{\lambda^2}{(2\pi\sigma^2)^N}\right) \tag{6.17}$$

6.3. – Aplicación en sistemas de detección de eventos sonoros

6.3.1. – Contexto: Análisis de la escena acústica.

Existen multitud de áreas donde la detección de eventos desconocidos es necesaria. Uno de los campos de investigación más interesantes es el del análisis de la escena acústica, donde las señales captadas por un grupo de micrófonos son procesadas para extraer tanta información sobre el entorno como sea posible. En estas aplicaciones de detección acústica, las fuentes de sonido suelen ser muy diversas e incontroladas, por tanto no pueden ser completamente conocidas y

caracterizadas, y es común encontrar ruido de fondo que degrada las prestaciones de la detección de los eventos acústicos.

El análisis acústico de un determinado escenario se ha considerado durante los últimos años como una fuente de información muy valiosa en diferentes sistemas de vigilancia o monitorización. Los sistemas de vigilancia que incorporan análisis de audio se están incrementando, ya que el sonido ofrece capacidades de vigilancia alternativas en el caso donde las cámaras de video son ciegas (gente escondida, condiciones de poca luz, zonas ciegas…) y en las situaciones donde las imágenes son aparentemente normales pero los sonidos pueden ser anormales [136]. Existen otros tipos de aplicaciones donde la monitorización acústica resulta de interés para conocer propiedades del entorno y actuar en consecuencia. Por ejemplo, en [137] se considera un sistema automático de ayuda en el hogar para gente anciana y en [138] el problema radica en la detección de sonidos anormales, tales como lloros o explosiones, en espacios públicos. Otra importante área donde se requiere el análisis de la escena acústica es en la interacción entre hombre y máquina, como en el caso de [139], donde un humano colabora con un robot y/o es asistido por uno.

Dentro del "Grupo de Tratamiento de Señal (*GTS*)" la detección de eventos acústicos se aplica en sistemas de vigilancia/monitorización basados en sonido. Proyectos recientes y significativos del GTS relacionados con este ámbito son: *Tecnologías para los sistemas de seguridad, video-vigilancia y monitorización remota del futuro (HESPERIA), 2006-08 (*programa CENIT*), Monitoring environments from acoustic scene analysis, 2009-10* (Acción Integrada en colaboración con el Instituto de Automática y Robótica de la Universidad de Karlsruhe), *ARtSENSE Augmented RealiTy Supported adaptive and personalized Experience in a museum based oN procesing real-time Sensor Events*, 2011-14, (proyecto STREP del Séptimo Programa Marco de la UE).

El proyecto Hesperia busca como objetivo el desarrollo de tecnologías que permitan la creación de sistemas de seguridad innovadores, video vigilancia y control de operaciones en edificios privados y lugares públicos. El *GTS* se encargó de la parte de monitorización acústica con el objetivo de prevenir situaciones potencialmente peligrosas mediante la detección, clasificación y localización de eventos sospechosos que pueden estar enmascarados por ruido de fondo.

El *GTS*, colaborando con el grupo de análisis de la escena acústica de la "Universität Karlsruhe", participó en el trabajo llevado a cabo en [140] para el desarrollo de un robot humanoide, el cual precisaba de un análisis completo de la escena acústica para localizar y detectar todos los tipos de eventos sonoros que pudieran ocurrir en sus proximidades.

Actualmente se está trabajando en un proyecto de investigación europeo denominado ARtSENSE, que involucra a nueve instituciones europeas, desde centros tecnológicos y universidades hasta compañías privadas e instituciones culturales. El

principal objetivo del proyecto es el de proporcionar una experiencia personalizada para todos los visitantes en un museo, adaptando la información proporcionada por un sistema de realidad aumentada tomando en cuenta el estado psicológico del visitante. El rol del *GTS* dentro del proyecto es el desarrollo del sistema de análisis acústico de la escena que rodea al visitante en cada instante y la generación de contenido de audio en el sistema de realidad aumentada.

6.3.2. – Marco experimental

Para probar, testear y comparar los diferentes detectores utilizados para resolver el problema de detección de señal desconocida en presencia de ruido de fondo conocido se han grabado una serie de diferentes eventos acústicos reales de diferentes características y se han combinado con diferentes tipos de ruido de fondo, tanto reales como sintéticos. De esta forma, se realiza una verificación híbrida entre simulación y realidad, permitiendo la simulación un mayor control de los experimentos.

Consideramos inicialmente el caso monocanal, en el que disponemos únicamente de la información obtenida por un solo micrófono. Posteriormente analizamos el caso multicanal, en el que se utiliza más de un micrófono para obtener la información, con objeto de poder mejorar las prestaciones obtenidas en el caso de un único canal.

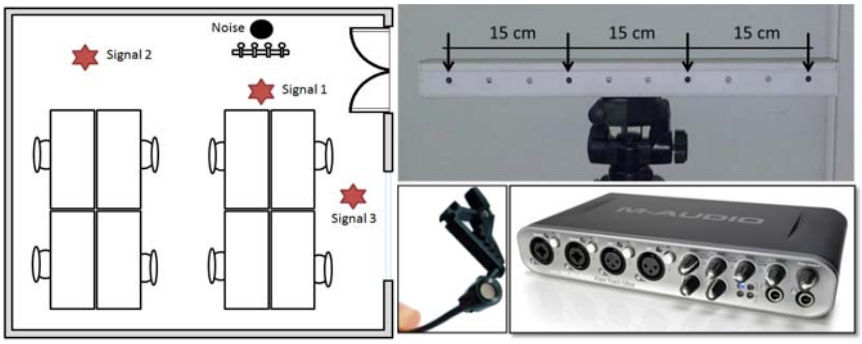


Figura 6.1- Descripción de la sala e instrumentos empleados en la campaña de medida de eventos acústicos.

Para la realización de la campaña de medidas (figura 6.1) se ha utilizado un array lineal de cuatro micrófonos capacitivos, separados 15 cm entre sí. Se ha utilizado para capturar el audio una tarjeta de adquisición de cuatro canales, con una frecuencia de muestreo de 44.1 KHz.

Se han considerado cuatro tipos de eventos acústicos: el sonido de un cristal roto, un grito, una palmada y el sonido de un walkie-talkie (figura 6.2). Se han generado cinco eventos consecutivos de cada clase, con algunos segundos de diferencia entre ellos, en tres posiciones distintas del laboratorio donde se trabaja, con objeto de tener diversas condiciones de dependencia espacial entre los canales. Los sesenta

eventos (veinte por cada posición) se han agrupado de forma conjunta, formando cuatro secuencias de $1.455 \text{x} 10^7$ muestras (~ 330 segundos), una por cada canal.

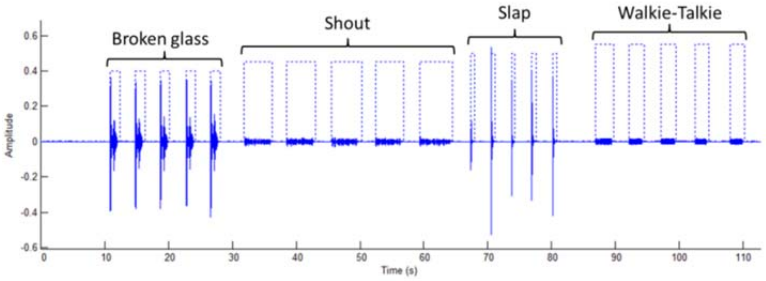


Figura 6.2. Secuencia de eventos acústicos reales generados en cada una de las diferentes posiciones.

Se ha captado, de forma aislada a los eventos acústicos, el sonido del sistema de aire acondicionado del laboratorio como ruido de fondo, el cual posee una distribución Gaussiana. Se ha comprobado que existe dependencia tanto temporal entre muestras consecutivas del mismo canal, como espacial entre los diferentes $N_C = 4$ canales (figura 6.3). En este caso la dependencia es lineal y puede ser recogida en la matriz de correlación de las muestras.

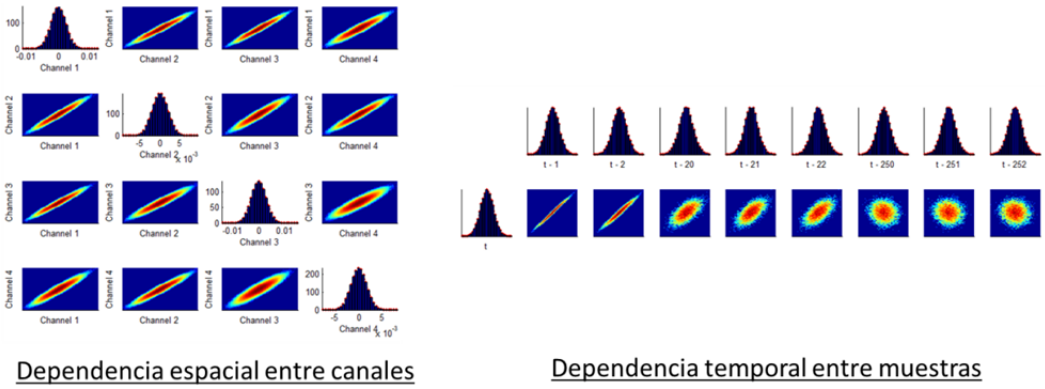


Figura 6.3- Dependencia espacial entre canales (izquierda) y entre diferentes muestras temporales a la (derecha).

Para generar los distintos ruidos no Gaussianos manteniendo las características de dependencia, tanto temporal como espacial, nos hemos basado en la teoría de cópulas. Partiendo de un cierto modelo multivariante con unas determinadas funciones marginales y características de dependencia, podemos, mediante una serie de transformaciones, generar un modelo que mantiene las características de dependencia del original, pero con distintas funciones marginales (figura 6.5).

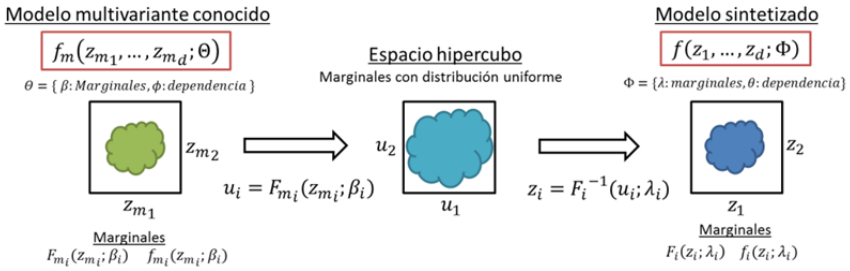


Figura 6.4 - Síntesis de una determinada distribución multivariante de probabilidad con diferentes funciones marginales y mismas características de dependencia que una distribución multivariante conocida

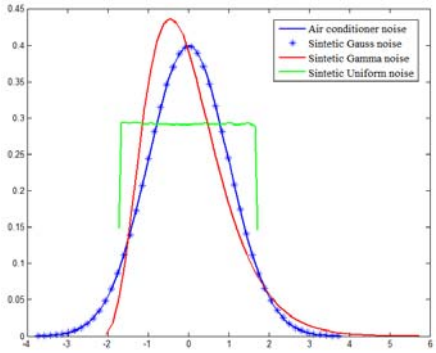


Figura 6.5 - Síntesis de ruidos no gaussianos dependientes: funciones marginales

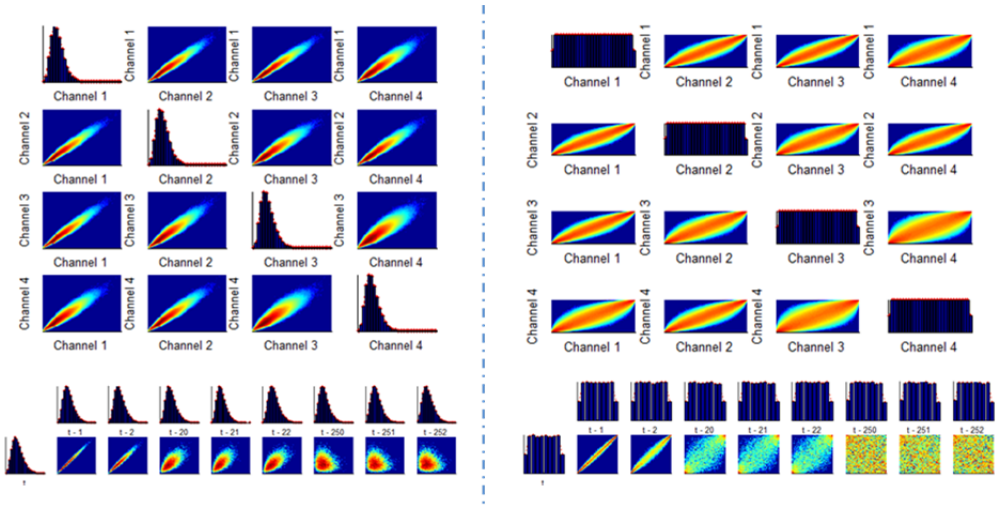


Figura 6.6 - Síntesis de ruidos no gaussianos dependientes: a la izquierda ruido Gamma, a la derecha ruido uniforme; arriba dependencia entre canales, abajo dependencia temporal.

Se han generado dos tipos de ruidos no Gaussianos con las mismas características de dependencia que el ruido real obtenido, parametrizados de tal forma que poseen la misma varianza (σ_w^2) que el ruido Gaussiano y con media nula. El primero de ellos con distribución Gamma como ejemplo de función que no se aleja sobremanera de una distribución Gaussiana cuando ambos se parametrizan con la misma varianza, y el segundo de ellos, con una distribución uniforme como ejemplo de un comportamiento altamente no Gaussiano (figura 6.5 y 6.6).

Una ventana deslizante de $N_T = 256$ muestras ha sido utilizada en todas las pruebas para ir definiendo de forma secuencial el vector de observaciones (***y)***, compuesto por la sobreimposición de las secuencias de eventos acústicos (***s***) con los diferentes ruidos (***w***). Utilizando como referencia la varianza de la señal obtenida al agrupar todos los eventos acústicos de forma consecutiva (σ_s^2), se ha ecualizado ($w_{eq} \sim \sigma_{eq}^2$) la varianza de las señales de ruido ($w \sim \sigma_w^2$) para representar diferentes relaciones de señal a ruido *SNR:*

$$SNR = 10 \cdot \log\left(\frac{\sigma_s^2}{\sigma_{eq}^2}\right) \rightarrow \quad w_{eq} = \sqrt{\frac{\sigma_{eq}^2}{\sigma_w^2}} \cdot w \tag{6.18}$$

Cada detector aplicará su test de hipótesis en cada uno de los desplazamientos de la ventana y mediante un barrido de umbrales, al conocer de antemano la localización de los eventos acústicos, podremos determinar cuándo se produce una detección correcta y cuando se produce una falsa alarma, pudiendo reflejar el funcionamiento de cada detector mediante una curva *ROC* ("*Receiver Operating Characteristics"*).

6.3.3. – Sistema de detección monocanal

En este apartado hemos evaluado y comparado las prestaciones del detector *COCD* propuesto con respecto al detector de energía clásico (6.2) y su extensión (6.3), donde se utiliza un blanqueamiento de la señal en el caso de ruido no independiente. Nos referiremos a ellos como *ED* en el caso del detector de energía clásico y como *PED* en el caso en el que se utiliza un preprocesado para blanquear el ruido.

Se han considerado las mediciones acústicas de ruidos y eventos sonoros recogidos por un solo micrófono. Comenzamos con ruido de fondo Gaussiano, analizando tanto el caso en el que es independiente como el caso en que no. Posteriormente pasamos a analizar los dos ruidos no Gaussianos propuestos, distinguiendo también el caso en el que existe independencia y el caso en que no.

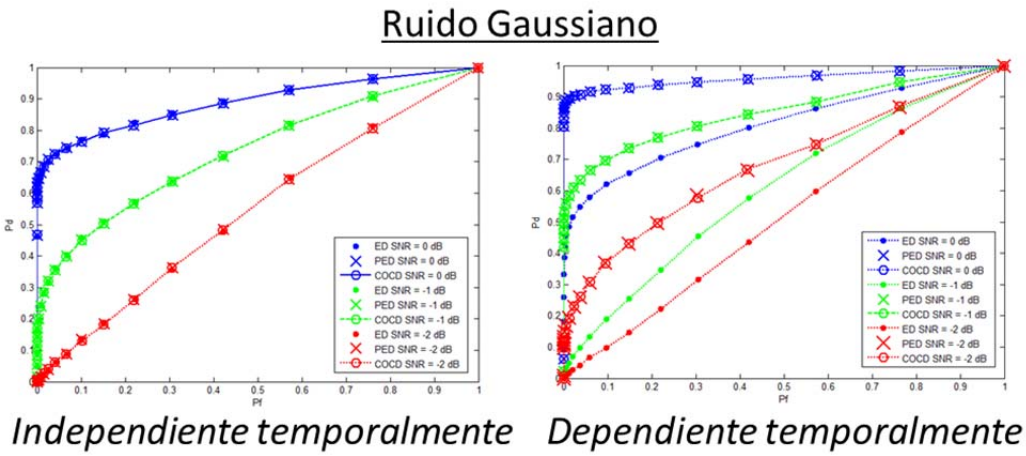


Figura 6.7 - Curvas ROC: Ruido Gaussiano independiente (izquierda) y no independiente (derecha).

En la figura 6.7 se muestra a la izquierda las curvas *ROC* obtenidas cuando el ruido de fondo es Gaussiano e independiente. Se puede apreciar como los tres detectores *ED*, *PED* y *COCD* Gaussiano son equivalentes. A la derecha se muestra el caso en el que existe dependencia temporal entre las muestra de ruido. Se observa como los detectores *PED* y *COCD* son equivalentes, y muestran unas mejores prestaciones que el detector de energía clásico, ya que son capaces de incorporar la información de dependencia.

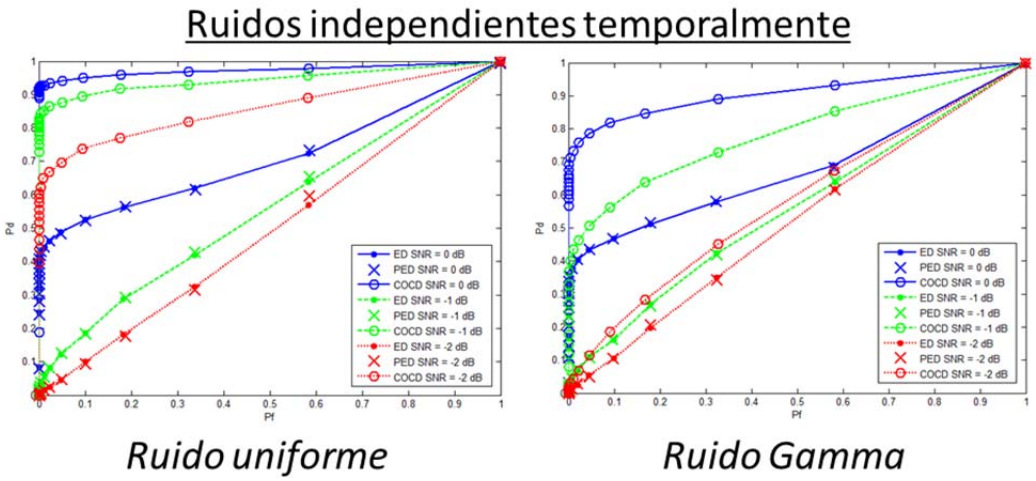


Figura 6.8 - Curvas ROC: Ruido uniforme (izquierda) y Gamma (derecha) independientes temporalmente.

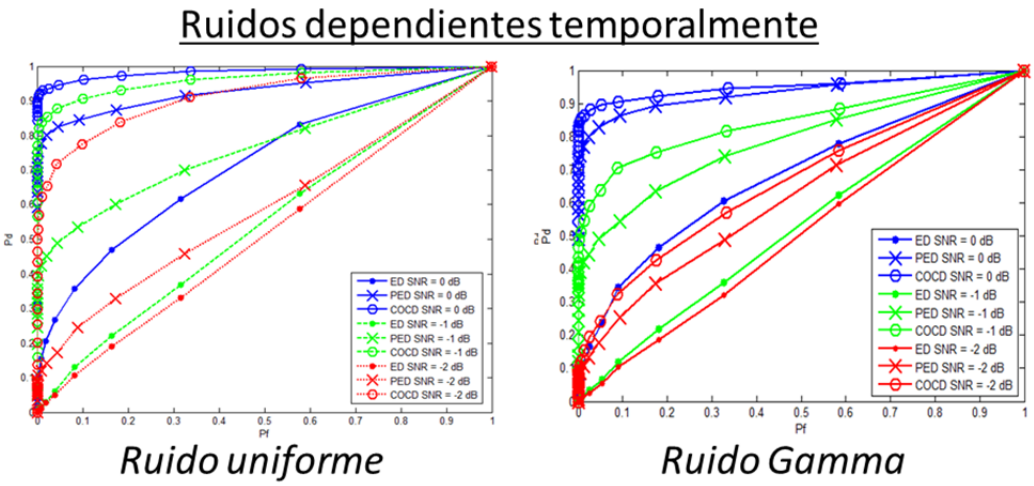


Figura 6.9 - Curvas ROC: Ruido uniforme (izquierda) y Gamma (derecha) independientes temporalmente.

En la figura 6.8, mostramos el caso de los dos ruidos no Gaussianos e independientes. Observamos como los detectores *PED* y *ED* son iguales (existe independencia), pero el detector *COCD* mejora las prestaciones de ambos, ya que no presupone Gaussianidad y modela de una forma más precisa las funciones marginales del ruido. Observamos que la mejora en prestaciones obtenidas con respecto a *PED* y *ED* se incrementa conforme la distribución del ruido se asemeja menos a una Gaussiana (reflejada en la distribución uniforme frente a la Gamma).

En la figura 6.9, se muestra el caso de ruidos no Gaussianos y no independientes. Otra vez el detector *COCD* muestra los mejores resultados, pero ahora el *PED* es mejor al *ED* debido a las características de dependencia temporal.

6.3.4. – Sistema de detección multicanal

En este apartado nos centramos en el apartado de detección de señal desconocida en presencia de ruido conocido en el caso de que dispongamos de más de un canal de información del cuál obtener las observaciones para resolver el problema de detección. En la aplicación propuesta se refleja en el uso de más de un micrófono captando señal sonora.

Así, se pretende mejorar las prestaciones de un sistema de detección de eventos acústicos mediante la integración de la información recogida por más de un micrófono. De esta forma, es preciso combinar o fusionar la información de diversos canales de datos.

Definición de datos en un detector multicanal

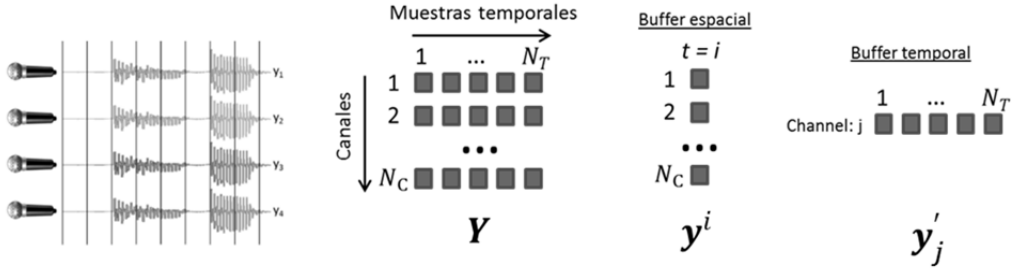


Figura 6.10 – Conjunto de muestras con las que implementar un detector multicanal

En el caso de la detección multicanal dispondremos de N_C canales con las muestras de audio capturadas por diferentes micrófonos para la resolución del problema de señal desconocida en presencia de ruido conocido. Así, en un determinado intervalo temporal T dispondremos de una matriz de muestras $\boldsymbol{Y}$ de tamaño $N_C x N_T$, donde N_T hace referencia al tamaño de muestras temporales capturadas en dicho intervalo (figura 6.10):

$$\boldsymbol{Y} = \begin{bmatrix} y_1^1 & \dots & y_1^i & \dots & y_1^{N_T} \\ y_2^1 & \dots & y_2^i & \dots & y_2^{N_T} \\ \dots & & \dots & & \dots \\ y_{N_C}^1 & \dots & y_{N_C}^i & \dots & y_{N_C}^{N_T} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}_1' \\ \boldsymbol{y}_2' \\ \boldsymbol{y}_3' \\ \boldsymbol{y}_4' \end{bmatrix} \tag{6.19}$$

Definimos como $\boldsymbol{y}^i$ al vector espacial con N_C muestras pertenecientes al instante temporal '*i*'. Definimos como $\boldsymbol{y}_j$ al vector con N_T muestras temporales pertenecientes al canal '*j*'. Denotamos como $\boldsymbol{y}_v$ a cualquier representación vectorial de todos los elementos de la matriz $\boldsymbol{Y}$.

El problema de detección en este escenario se basa en discernir cuando, en un determinado intervalo temporal T, el conjunto de muestras capturadas $\boldsymbol{Y}$ pertenece al caso en el que se ha producido algún tipo evento acústico $\boldsymbol{S}$ en la escena, el cual estará contaminado por un determinado ruido de fondo $\boldsymbol{W}$ (hipótesis H_1), del caso en el que solamente esté presente el ruido de fondo (hipótesis H_0).

$$\begin{aligned} H_0&: \ \boldsymbol{Y} = \boldsymbol{W} && \text{(Noise)} \\ H_1&: \ \boldsymbol{Y} = \boldsymbol{S} + \boldsymbol{W} && \text{(Signal + Noise)} \end{aligned} \tag{6.20}$$

Considerando que el ruido de fondo es Gaussiano correlado y que no disponemos de información del posible evento acústico que puede generarse en la escena,

utilizamos, al igual que se hizo en el caso monocanal, una técnica "one-class" para obtener el detector basado en la relación de verosimilitud:

$$\Lambda(\boldsymbol{y}) = \frac{1}{f(\boldsymbol{y}|H_0)}\ (monocanal) \quad \Lambda(\boldsymbol{Y}) = \frac{1}{f(\boldsymbol{Y}|H_0)}\ (Multicanal) \tag{6.21}$$

Observamos como en este caso la fusión de información entre canales se está realizando **siguiendo un método basado en la estimación de densidades de probabilidad,** ya que consideramos el modelado estadístico conjunto de todos los datos presentes, siendo realizada la integración a **nivel de fusión de muestras**.

Implementación de un detector multicanal

Todos los canales estarán contaminados con el mismo tipo de ruido de fondo, existiendo como única diferencia un posible desbalance en las ganancias con las que se capta este en cada canal. Con objeto de considerar las muestras de ruido idénticamente distribuidas, independientemente del canal en que se capten, se puede aplicar un preprocesado para ecualizar los canales. Sea $\sigma^2_{w\,j}$ la varianza del ruido captado por cada canal '*j*', se aplica un pre-procesado a los canales de forma que todos posean la misma varianza de ruido σ^2_w:

$$\boldsymbol{w_{eq_j}} = \sqrt{\frac{\sigma^2_w}{\sigma^2_{w\,j}}} \cdot \boldsymbol{w_j} \quad , j = 1 \dots N_C \tag{6.22}$$

Considerando ya todas las muestras de ruido captadas idénticamente distribuidas, podemos implementar un único detector *COCD* (6.14) que trabaje con toda la información multicanal disponible. Así, en la implementación del detector se utilizará un vector de observaciones $\boldsymbol{y_v}$ que aúne o combine las N_C x N_T muestras de la matriz de observación (figura 6.11):

$$g(\boldsymbol{y_v})^T(\boldsymbol{\Sigma_W}^{-1} - \boldsymbol{I})g(\boldsymbol{y_v}) - 2 \cdot \sum_{i=1}^{N_C \cdot N_T} \ln\left(f_w(y_{\boldsymbol{v}_i})\right) \lessgtr_{H_1}^{H_0} \lambda \tag{6.23}$$

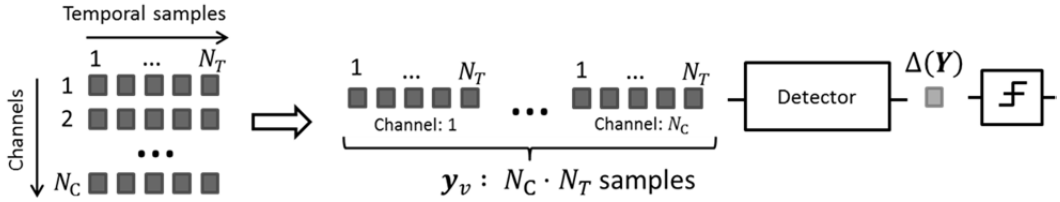


Figura 6.11 – Combinación de muestras en un problema de detección multicanal

En este esquema de detección multicanal, además de existir una posible dependencia temporal entre muestras consecutivas en cada canal, se puede encontrar una dependencia espacial entre muestras de distintos canales en el mismo instante de tiempo y/o espacio-temporal, entre muestras de distintos canales e instantes de tiempo.

Así, utilizando más de un micrófono para la detección de eventos acústicos, además de conseguir más muestras de señal en un mismo periodo de observación, lo cual puede mejorar las características del detector, posibilita la incorporación de más información de dependencia con la que caracterizar el ruido de fondo. En la figura 6.12 se puede observar una representación de la matriz de correlación $\boldsymbol{\Sigma_W}$ del conjunto de $N_C \text{ x } N_T = 1024$ muestras del ruido del aire acondicionado. Se ha conformado concatenando los vectores de observaciones $\boldsymbol{w}_j$ de los $N_C = 4$ canales, con $N_T = 256$ muestras cada vector. Se puede observar como existe dependencia temporal entre muestras consecutivas de cada canal, dependencia espacial entre muestras entre canales para los mismos instantes temporales y una combinación de ambas entre muestras de canales e instantes temporales diferentes.

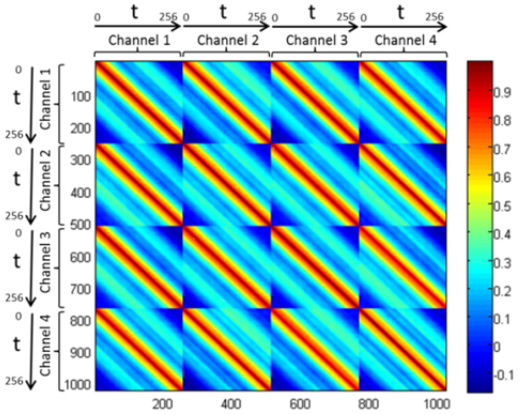


Figura 6.12. – Dependencia espacio-temporal en el ruido de aire acondicionado captado por cuatro micrófonos recogida en la matriz de correlación

Fusión de detectores

En un contexto práctico, podemos encontrarnos que no siempre se podrá implementar un único detector que tenga acceso a todas las muestras captadas por los distintos canales para realizar su fusión. Por ejemplo, podemos encontrarnos con un esquema en el que los micrófonos se encuentren separados del sistema de procesado y, con objeto de reducir la cantidad de información que deben enviar, sea conveniente implementar un detector local más simple por cada canal que fusione en un primer nivel toda la información monocanal y la transmita al sistema de

procesado, donde se combine con la del resto de canales en un segundo nivel de fusión.

El principal problema que plantea el uso del detector que implementa la fusión óptima de canales es que incrementa considerablemente el tamaño de la matriz de correlación que caracteriza al ruido: $Monocanal: \boldsymbol{\Sigma_w},\ N_T\ x\ N_T \rightarrow Multicanal: \boldsymbol{\Sigma_W},\ (N_C \cdot N_T)x(N_C \cdot N_T)$. Esto puede plantear problemas en la etapa de entrenamiento del ruido, puesto que para una correcta estimación de esta nueva matriz, además del coste computacional que añade, también se requiere de un mayor periodo temporal de muestras. Por ejemplo, en un sistema en el que las características del ruido de fondo no sean continuas a lo largo del tiempo puede ser necesario una monitorización, calibración y estimación en tiempo real de este. La incorporación de matrices de correlación de tan elevada dimensionalidad puede ser incompatible con el requerimiento de procesado en tiempo real.

Una posibilidad para solventar este problema en el caso que la aplicación de la fusión de todas las muestras sea inviable, pasa por considerar que los canales de datos son independientes entre sí, con lo que la expresión de la densidad del ruido se simplifica:

$$f(\boldsymbol{Y}|H_0) = f(\boldsymbol{W}) = f(\boldsymbol{w}_1) \cdot \ldots \cdot f\left(\boldsymbol{w}_{N_c}\right) \tag{6.24}$$

Podemos obtener una simplificación del detector 'one-class' *COCD* asumiendo independencia entre canales, lo cual nos conduce a un detector cuyo estadístico es equivalente a la suma de los estadísticos de los detectores *COCD* implementados en cada uno de los canales individuales.

$$\Delta(\boldsymbol{Y}) \approx \sum_{j=1}^{N_c} \Delta_j(\boldsymbol{y}_j) = \sum_{j=1}^{N_c} \left(g(\boldsymbol{y}_j)^T \left({\boldsymbol{\Sigma}_{\boldsymbol{w}_j}}^{-1} - \boldsymbol{I} \right) g(\boldsymbol{y}_j) - 2 \cdot \sum_{i=1}^{N_T} \ln\left(f_w(y_j^i) \right) \right) \tag{6.25}$$

Si lo particularizamos para el caso Gaussiano, es fácil ver que esta regla de fusión equivale a una suma de las energías de los detectores individuales en cada canal (figura 6.13).

$$\Lambda(\boldsymbol{Y}) = \boldsymbol{y'}_{\boldsymbol{v}} \cdot \boldsymbol{\Sigma}_{\mathbf{W}} \cdot \boldsymbol{y}_{\boldsymbol{v}} \approx \sum_{j=1}^{N_c} \boldsymbol{y}'_j \cdot \boldsymbol{\Sigma}_{\boldsymbol{w}_j} \cdot \boldsymbol{y}_j \lessgtr_{H_1}^{H_0} \lambda \tag{6.26}$$

En el caso del detector *ED* la fusión mediante un único detector que considere todas las muestras $\boldsymbol{y_v}$, o considerar detectores *ED* individuales por cada canal y luego sumar sus energías es equivalente:

$$\Lambda(\boldsymbol{Y}) = \frac{\boldsymbol{y'}_{\boldsymbol{v}} \boldsymbol{y}_{\boldsymbol{v}}}{\sigma_w^2} = \sum_{j=1}^{N_c} \frac{\boldsymbol{y}'_j \cdot \boldsymbol{y}_j}{\sigma_w^2} \tag{6.27}$$

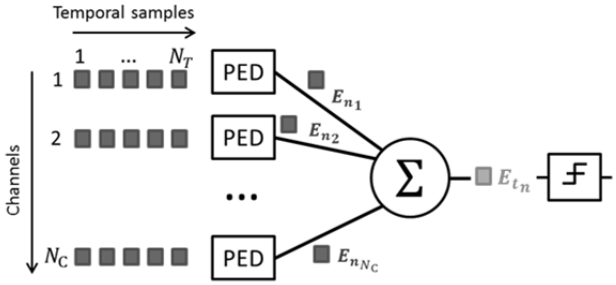


Figura 6.13. – Esquema de fusión de diversos canales a través de la suma de las energías normalizadas de detectores individuales por cada canal

Observamos cómo, mediante la asunción de independencia entre canales realizada, hemos obtenido una regla de fusión subóptima consistente en el sumatorio de las valoraciones aportadas por diversos detectores *COCD* aplicados en cada canal individual. Si lo particularizamos para el caso Gaussiano, la fusión se realiza mediante la suma de energías normalizadas. Observamos que esta técnica equivale a una **fusión a nivel de valoraciones** de diversos detectores mediante una **combinación lineal ponderada** (en este caso utilizando pesos unidad para todos los canales).

Pruebas multicanal

Como se comentó se utilizaron cuatro micrófonos para la captación tanto del ruido, como de los eventos acústicos. Presentamos ahora las pruebas realizadas con los cuatro canales, implementando tanto el detector one-class que fusiona las muestras de todos los canales, como mediante la fusión de detectores individuales por cada canal (figura 6.14).

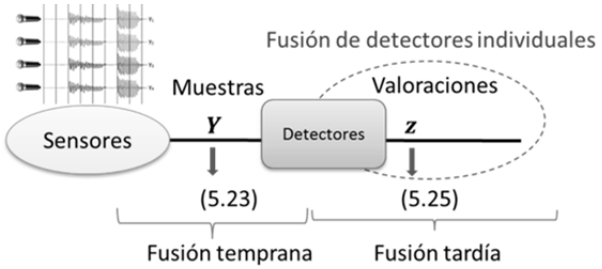


Figura 6.14. – Esquema de posibles fusiones planteadas para combinar diferentes micrófonos en un problema de detección de eventos acústicos

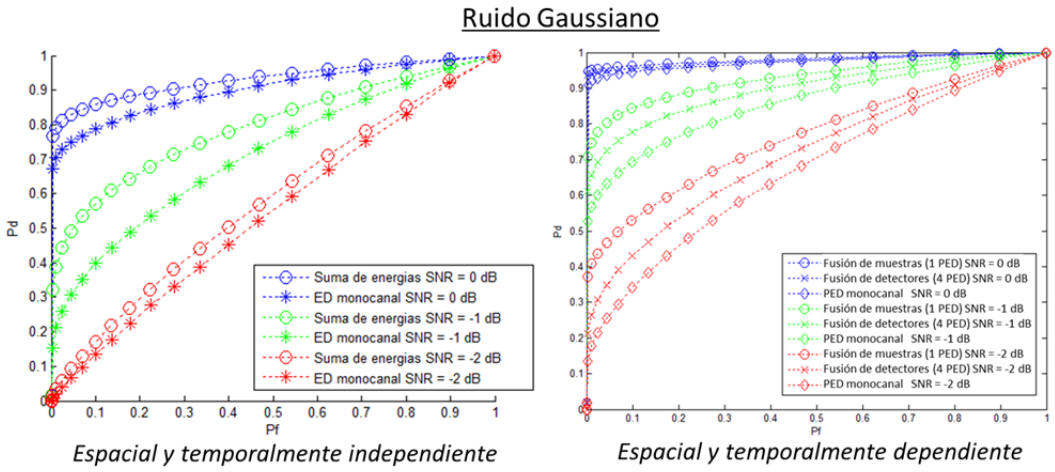


Figura 6.15 – Detección multicanal con ruido Gaussiano independiente (izquierda) y dependiente (derecha) temporal y espacialmente. Comparación con el caso monocanal

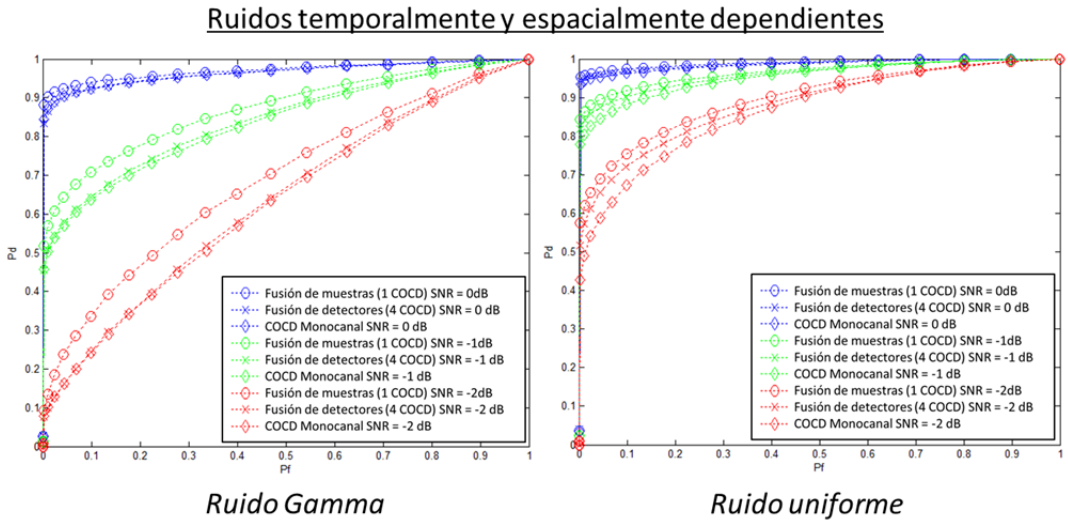


Figura 6.16 –Comparación la detección multicanal (combinando todas las muestras o fusionando detectores individuales en cada canal) con la detección monocanal. Ruido Gamma (izquierda) y uniforme (derecha) dependiente temporal y espacialmente

En la figura 6.15 se obtienen los resultados con ruido Gaussiano, tanto independiente espacial y temporalmente, como dependiente. En el caso de ruido Gaussiano independiente (izquierda), como ya se demostró, los tres detectores son equivalentes. Se puede comprobar como la utilización de los cuatro micrófonos

mejora las prestaciones en detección de los eventos acústicos en ambos casos. Se comprueba que cuando existe dependencia, lo mejor es la fusión de todas las muestras, puesto que como ya hemos comentado, es donde se encuentra la riqueza de información estadística. Observamos como mediante la fusión de los detectores (*PED*), al considerar independencia espacial se produce una degradación en las prestaciones, pero aún se mejora al caso de un único canal.

En la figura 6.16 se muestran los resultados de la detección multicanal con los dos tipos de ruido no Gaussiano considerados. Observamos como mediante la fusión de las muestras de los cuatro micrófonos realizada por el detector COCD se mejoran las prestaciones en la detección de los eventos acústicos con respecto al caso multicanal. La fusión de los detectores, al suponer independencia entre los canales, introduce una degradación, observando que, por ejemplo en el caso del ruido Gamma la mejora de prestaciones de esta fusión con respecto al caso monocanal es ínfima. En la figura 6.17 se utilizan los mismos ruidos no Gaussianos para comparar la fusión óptima de las muestras multicanal mediante el detector *PED* y el *COCD.* Se comprueba como el detector derivado de la teoría de la cópula Gaussiana (*COCD*), debido a que es capaz de modelar mejor la distribución un ruido arbitrario no Gaussiano, modelando tanto la información marginal como la información de dependencia, obtiene mejores prestaciones también en este caso multicanal.

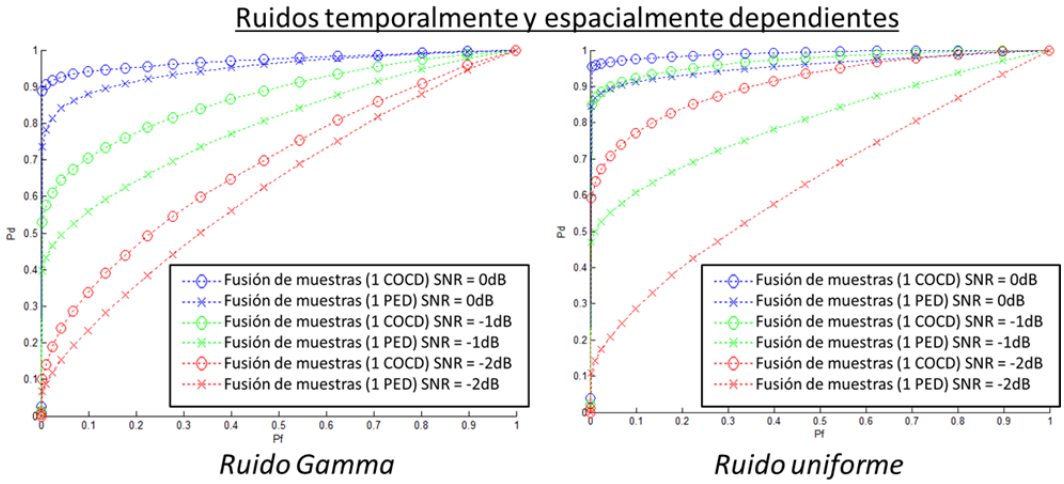


Figura 6.17– Comparación de detección multicanal PED y COCD combinando todas las muestras. Ruido Gamma (izquierda) y uniforme (derecha) dependiente temporal y espacialmente

6.4. – Conclusiones

En este capítulo se ha realizado una revisión del estado del arte de las técnicas o detectores usados comúnmente en problemas de detección de señal desconocida en

presencia de ruido aleatorio. Así hemos comprobado los detectores más usados en estos problemas son los detectores de energía, tanto el detector clásico denominado *ED* que asume ruido Gaussiano independiente, como su variante denominada *PED,* el cual asume que el ruido Gaussiano puede estar correlado de forma lineal. En muchas áreas o aplicaciones el ruido de fondo que se puede encontrar no es Gaussiano como en ciertas aplicaciones musicales o de sonido, en dispositivos ópticos, datos económicos, en aplicaciones de detección del habla… (ver [141], [142] y las referencias adjuntas en ellos). Un gran esfuerzo en la investigación en este campo se realiza para buscar técnicas para lidiar con el caso de ruido de fondo no Gaussiano y coloreado.

Hemos propuesto un nuevo tipo de detector denominado *COCD* basado en la teoría de cópulas para lidiar con este problema. Se obtiene el detector utilizando una técnica de detección *One-Class*, caracterizando la *PDF* multivariante de un cierto intervalo de ruido usando una función de densidad de cópula para caracterizar la dependencia y la PDF marginal. Partiendo de una cópula Gaussiana, la cual modela dependencia lineal entre variables aleatorias, se ha derivado un detector capaz de modelar ruido no Gaussiano coloreado (con dependencia lineal modelada por una matriz de covarianza) sin añadir mucha complejidad, ni poseer altos requerimientos temporales y/o computacionales para su entrenamiento y evaluación, como las alternativas existentes a los detectores de energía. Se ha demostrado que los detectores de energía pueden ser obtenidos como caso particular del *Gaussian-COCD* propuesto.

Se ha comprobado sus prestaciones en una aplicación de detección de eventos sonoros en presencia de ruido de fondo. Para ello se ha realizado una serie de pruebas con diferentes eventos acústicos y ruidos de fondo. Se ha utilizado eventos acústicos reales; el ruido de fondo Gaussiano ha sido grabado de un aparato de aire acondicionado. Con objeto de mantener las características de dependencia del ruido, se han generado dos ruidos sintéticos con diferentes distribuciones, pero el mismo tipo de dependencia. Se ha demostrado como el detector *Gaussian-COCD* mejora las prestaciones obtenidas por los detectores de energía clásicos en el caso de ruido de fondo no Gaussiano, debido a la capacidad para caracterizar tanto la información marginal, como la información de dependencia del ruido.

También se ha realizado un estudio sobre la detección multicanal, añadiendo más micrófonos y por lo tanto canales de datos para mejorar las prestaciones de detección. Ante este escenario, se ha planteado la combinación óptima de los diferentes canales mediante la fusión del conjunto de muestras llevada a cabo por un único detector. En este caso, aparte de la introducción de más muestras en un mismo intervalo temporal, se incluye la posibilidad de caracterizar el ruido también por su dependencia espacial. Así, teniendo en cuenta toda la información espacio-temporal de dependencia se consigue modelar mejor el ruido de fondo y se consigue discriminar mejor cualquier tipo de evento acústico. También se ha planteado un caso

subóptimo en el que se desprecia la dependencia estadística entre canales para poder combinar diferentes detectores individuales por cada canal. Se ha demostrado mediante unos casos prácticos como, mediante el uso de más de un micrófono y aplicando técnicas de fusión de datos, se pueden mejorar las prestaciones de detección. También se ha comprobado como el no considerar la dependencia estadística presente en los datos conlleva a una degradación en las prestaciones obtenidas.

Conclusiones y líneas futuras

Capítulo 7:
Discusión y líneas futuras de trabajo

La presente tesis se ha centrado en la problemática existente a la hora de implementar un **sistema de detección** o clasificación binaria cuando es necesario combinar, integrar o **fusionar diversas fuentes de información**. Se ha discurrido sobre la naturaleza y características de los diferentes datos que se pueden combinar, los distintos niveles en los que se puede realizar la integración de información y los métodos o técnicas que pueden ser utilizadas para la fusión de diversos canales de datos.

Se ha observado como el principal problema que se plantea en la fusión de datos en un sistema de detección es lidiar con la posible **heterogeneidad y dependencia estadística** entre los diferentes datos. Así, a lo largo del presente estudio se ha podido comprobar, mediante diversos ejemplos y demostraciones, como la dependencia estadística entre los datos juega un papel muy importante en cuanto a las prestaciones que se pueden obtener en un sistema de detección. La asunción de independencia o una mala caracterización de la dependencia estadística entre los datos puede acarrear una severa reducción o degradación de las prestaciones de detección, pudiendo incluso llevar a la obtención de peores resultados tras la fusión de varias fuentes de información, que si se consideraran estas de forma aislada. Incluso se ha podido demostrar que en ciertos casos es posible obtener mejores prestaciones con una fusión hard de decisiones individuales, incluso subóptima, que tenga en cuenta cierto grado de dependencia estadística existente entre las diferentes decisiones, con respecto al caso que se implemente una fusión soft de los datos de los cuales se derivan estas decisiones (en principio más ricos en cuanto información sobre el evento) sin tener en cuenta las características de dependencia entre ellos.

En el presente trabajo se ha realizado una completa revisión del estado del arte en cuanto a técnicas de fusión y combinación de datos, tanto soft como hard, aplicadas en problemas de detección donde los datos pueden ser heterogéneos y dependientes entre sí. Se han introducido los **métodos óptimos de fusión** de datos y las diferentes técnicas que pueden ser utilizadas para su implementación.

Así, se ha planteado el problema de fusión soft óptima en detección como un problema de test de hipótesis basado en la relación de verosimilitud, y por lo tanto un problema de estimación de las funciones de densidad de probabilidad multivariante de las variables a fusionar. Se han revisado los principales métodos paramétricos y no paramétricos existentes. Se realiza una revisión en mayor profundidad de la técnica de estimación de funciones de densidades de probabilidad (*PDFs*) basada en la teoría de cópulas, la cual puede ser usada en la fusión óptima de datos soft. Se destaca de forma especial tanto por su novedad e incipiente uso en el campo del procesado de

señal, como por su adecuación en problemas de detección, permitiéndonos modelar de forma aislada las funciones marginales de los datos y la estructura de dependencia presente entre ellos, simplificando el problema de modelado de *PDFs* de datos heterogéneos y dependientes.

El problema de fusión hard óptima en detección se ha planteado como el consiste en la fusión de diversas decisiones individuales proporcionadas por un conjunto de diferentes detectores. Así, la regla de fusión óptima es obtenida mediante la derivación de la relación de verosimilitud dada por el cociente de las funciones de masa de probabilidad de los datos binarios bajo las hipótesis H_1 y H_0.

Se ha argumentado cómo en algunas situaciones no será posible la implementación o utilización de estas técnicas óptimas y deberán utilizarse ciertas **técnicas subóptimas de fusión**. Se han presentado otra serie de reglas de fusión subóptimas, normalmente más simples y con menores requerimientos. Se ha mostrado como algunas otras técnicas de fusión subóptimas pueden aprovechar cierta información de dependencia, y en determinados escenarios y bajo determinadas condiciones pueden conseguir buenas prestaciones.

Con objeto de poder testear y comparar el correcto funcionamiento de las diferentes técnicas consideradas novedosas, se ha decidido incluir un conjunto de aplicaciones de **autentificación multibiométrica** basadas en una serie de bases de datos públicas, las cuales han sido usadas en multitud de investigaciones, convirtiéndose en un estándar para la verificación de muchos algoritmos de fusión. Así, se han probado y testeado los diferentes métodos de fusión mediante estimación de *PDFs* utilizando la teoría de cópulas, cuya aplicación, como ya hemos comentado, supone una novedad en el campo del procesado de señal. Así mismo, también se ha utilizado el novedoso método de fusión mediante integración α propuesto en el presente trabajo como mejora de las técnicas subóptimas habituales usadas en fusión.

En estos ejemplos hemos comprobado que la utilización de la **teoría de las cópulas** para la estimación de *PDFs* multivariantes, al permitirnos separar el modelado de la información marginal del modelado de la estructura de dependencia que poseen los datos, simplifica el proceso de caracterización de los datos, proporcionándonos una herramienta muy potente para incluir la información de dependencia entre los distintos datos heterogéneos que se pretenden combinar o fusionar en un problema de detección. Así, hemos visto como la correcta caracterización de la dependencia nos ayuda a mejorar las prestaciones obtenidas mediante la fusión de diversas fuentes de información.

Se ha aplicado la **fusión mediante la integración α** en diferentes sistemas, con diversos escenarios de normalización de datos, contrastando sus prestaciones con respecto a otras técnicas simples de fusión utilizadas en multitud de estudios. Se ha comprobado cómo, utilizando el método de entrenamiento propuesto, esta técnica

permite adaptarse a las distintas características que posean los datos y proporcionar una mejora en las prestaciones de detección mediante la fusión de datos. Se han comparado los métodos de entrenamiento basados en el criterio de la minimización del error cuadrático medio y en el criterio de maximización de la *AUC* parcial, demostrando como este segundo método proporciona mejores resultados.

Así, se ha demostrado que tanto la técnica de fusión soft basada en la relación de verosimilitud*,* utilizando funciones de cópula en la estimación de las *PDFs*, como la técnica de fusión soft basada en la integración α (pudiendo utilizarse tanto para realizar una fusión soft genérica, como para la fusión de scores) pueden mejorar las prestaciones obtenidas en un sistema de detección con diversas fuentes de información mediante una correcta caracterización y adaptación a los diversos datos heterogéneos y dependientes.

También se han aplicado algunas de las técnicas de fusión en la mejora de un sistema de detección de eventos acústicos. Se ha realizado una revisión del estado del arte de las técnicas o detectores usados comúnmente en problemas de detección de señal desconocida en presencia de ruido aleatorio, problema donde se enmarca la detección de eventos acústicos. Así hemos comprobado como los detectores más usados en estos problemas son los detectores de energía, tanto el detector clásico denominado *ED* que asume ruido Gaussiano independiente, como su variante denominada *PED,* el cual asume que el ruido Gaussiano puede estar correlado de forma lineal. En muchas áreas o aplicaciones el ruido de fondo que se puede encontrar no es Gaussiano como en ciertas aplicaciones musicales o de sonido, en dispositivos ópticos, datos económicos, en aplicaciones de detección del habla… Un gran esfuerzo en la investigación en este campo se realiza para buscar técnicas para lidiar con el caso de ruido de fondo no Gaussiano y coloreado.

Así, se ha propuesto un nuevo tipo de detector denominado ***COCD*** basado en la teoría de cópulas para lidiar con este problema. Se obtiene el detector utilizando una técnica de detección *One-Class*, caracterizando la *PDF* multivariante de un cierto intervalo de ruido usando una función de densidad de cópula para caracterizar la dependencia y la PDF marginal. Partiendo de una cópula Gaussiana, la cual modela dependencia lineal entre variables aleatorias, se ha derivado un detector capaz de modelar ruido no Gaussiano coloreado (con dependencia lineal modelada por una matriz de covarianza) sin añadir mucha complejidad, ni poseer altos requerimientos temporales y/o computacionales para su entrenamiento y evaluación, como las alternativas existentes a los detectores de energía. Se ha demostrado que los detectores de energía pueden ser obtenidos como caso particular del *Gaussian-COCD* propuesto.

Se ha comprobado sus prestaciones en una aplicación de detección de eventos sonoros en presencia de ruido de fondo. Se ha demostrado como el detector *Gaussian-COCD* mejora las prestaciones obtenidas por los detectores de energía

clásicos en el caso de ruido de fondo no Gaussiano, debido a la capacidad para caracterizar tanto la información marginal, como la información de dependencia del ruido.

También se ha realizado un estudio sobre la detección multicanal, añadiendo más micrófonos y por lo tanto canales de datos para mejorar las prestaciones de detección. Ante este escenario, se ha planteado la combinación óptima de los diferentes canales mediante la fusión del conjunto de muestras llevada a cabo por un único detector. En este caso, aparte de la introducción de más muestras en un mismo intervalo temporal, se incluye la posibilidad de caracterizar el ruido también por su dependencia espacial. Así, teniendo en cuenta toda la información espacio-temporal de dependencia se consigue modelar mejor el ruido de fondo y se consigue discriminar mejor cualquier tipo de evento acústico. Se ha planteado un caso subóptimo en el que se desprecia la dependencia estadística entre canales para poder combinar diferentes detectores individuales por cada canal. Se ha demostrado mediante unos casos prácticos como, mediante el uso de más de un micrófono y aplicando técnicas de fusión de datos, se pueden mejorar las prestaciones de detección. También se ha comprobado como el no considerar la dependencia estadística presente en los datos conlleva a una degradación en las prestaciones obtenidas.

Podemos comprobar cómo se han alcanzado los diferentes objetivos que se han propuesto con la realización del presente trabajo de investigación:

- Se ha realizado una revisión del estado del arte en cuanto a técnicas (tanto óptimas como subóptimas) de fusión y combinación de datos aplicadas en problemas de detección, en donde los datos pueden ser heterogéneos y dependientes entre sí.
- Se ha llevado a cabo un completo estudio de la técnica de estimación de funciones de densidades de probabilidad multivariantes basada en la teoría de cópulas, la cual es usada en la fusión óptima de datos soft. La destacamos de forma especial por su novedad e incipiente uso en el campo del procesado de señal.
- Se han analizado e implementado las diferentes técnicas de fusión de datos.
- Se ha demostrado que, tanto el uso de más de un sensor o fuente de información, como una correcta consideración de las características de dependencia entre los datos, pueden ayudar a mejorar el rendimiento y precisión obtenidos a la hora de diseñar un sistema de detección.
- Se han propuesto modificaciones o variantes de las técnicas y algoritmos existentes, de forma que se mejore su rendimiento incorporando información

sobre la dependencia existente o proporcionando una mejor caracterización de ella.

- Se ha aplicado el estudio sobre las técnicas de fusión en problemas de detección reales, donde se pretende usar múltiples sensores para mejorar las prestaciones que se obtienen con un único sensor.

En las tablas 7.1, 7.2 y 7.3 se recogen una comparativas entre las diferentes técnicas de fusión incluidas en el presente documento.

Con respecto al trabajo futuro que se pretende realizar, se puede dividir en varias líneas. Una de ellas se centra en la obtención de nuevas funciones de densidad de cópula capaces de modelar estructuras de dependencia muy complejas; para ello se pretende derivar las densidades de cópula a partir de modelos de *PDFs* basadas en mezclas de distribuciones, como son el modelo de mezcla de expertos y un modelo de mezcla de componentes no Gaussianas (mezcla de componentes independientes [143]). Otro posible objetivo es el estudio de nuevos métodos de selección de cópulas más simples, de forma que no se precise entrenar cada una de las posibles funciones de cópula para escoger la que mejor se adapte a los datos.

Otra de las líneas en las que se pretende seguir investigando, ya se comentó en el apartado 3.5.1, es en la utilización de la función de media-α (como alternativa a la suma ponderada) como forma de integración probabilística en un modelo de mezcla de expertos.

También se pretende seguir investigando en el campo de la detección de señal desconocida en presencia de ruido conocido. Se seguirá utilizando la estructura propuesta en el detector *COCD,* incorporando diferentes tipos de cópulas diferentes a la Gaussiana, con objeto de poder mejorar su comportamiento ante datos con dependencia muy alejada de la lineal.

Existe un gran abanico de posibles aplicaciones prácticas en las que poder utilizar las técnicas expuestas en este trabajo. Las técnicas de fusión soft se pretenden utilizar en aplicaciones como la detección de fraudes bancarios, en problemas de tratamiento de señales biomédicas (*EEGs, ECGs* y *fMRI*) para el análisis de patologías como arritmias cardiacas, desordenes del sueño y epilepsia, en una aplicación de identificación biométrica basada en *EEGs* y en aplicaciones de detección de eventos acústicos submarinos. La fusión hard se utilizará para la mejora de un sistema de detección precoz de incendios basado en el procesado de señales infrarrojas.

Fusión soft: $\boldsymbol{x} = [x_1 \dots x_d]^T : \mathbb{R}^d \rightarrow x_{fus} : \mathbb{R}$

Fusión basada en la relación de verosimilitud: $\Lambda(\boldsymbol{x}) : \mathbb{R}^d \rightarrow \mathbb{R} \implies \Lambda(\boldsymbol{x}) = \frac{f(\boldsymbol{x}|H_1)}{f(\boldsymbol{x}|H_0)} \lessgtr_{H_1}^{H_0} \lambda$

Estimar las *PDFs* bajo cada hipótesis $(H_j,\ j = 0,1)$ usando un conjunto de datos de entrenamiento $\mathcal{X}_j = \{\boldsymbol{x}^1, \dots, \boldsymbol{x}^{N_j}\},\ \boldsymbol{x}^i = [x_1^i, \dots, x_d^i]^T \in H_j\ , i = 1, \dots, N_j$

Relativa a la independencia estadística	Independencia entre datos Se asume independencia estadística entre las v.a	$\hat{f}_p(\boldsymbol{x}\|H_j) = \prod_{i=1}^{d} \hat{f}(x_i\|H_j)$	Modelo simple. Se estima las PDFs marginales de forma aislada usando técnicas unidimensionales muy conocidas y estudiadas Si existe dependencia estadística puede acarrear severas degradaciones en las prestaciones de detección.
	Pre-procesado de los datos mediante *ICA* Preprocesado $(\boldsymbol{s}^{H_j} = \boldsymbol{A_j^{-1}} \cdot \boldsymbol{x} = \boldsymbol{W_j} \cdot \boldsymbol{x},\ \boldsymbol{s}^{H_j} \in \mathbb{R}^n)$ para conseguir independencia	$\hat{f}_s(\boldsymbol{s}\|H_j) = \hat{f}(\boldsymbol{s}^{H_j}) = \prod_{i=1}^{d} \hat{f}(s_i\|H_j)$	Mediante *ICA* se pueden simplificar complejas distribuciones, con marginales heterogéneas y complejas estructuras de dependencia. En general *ICA* es una técnica compleja, involucrando grandes costes computacionales y temporales para su entrenamiento.
Estimación no paramétrica de densidades	Discretización del espacio de datos División de cada dimensión en L_i segmentos: se divide el espacio $\mathcal{X}$ en $L = \prod L_i$ prismas multidimensionales $\mathcal{X} \equiv \Re_1 \cup \Re_2 \dots \cup \Re_L$	$\boldsymbol{x} \in \Re_i \rightarrow \hat{f}_{Hist}(\boldsymbol{x}\|H_j) = \frac{k_j^i / N_j}{V_{\Re_i}}$ k_j^i: nº datos entrenamiento $\in \Re_i, H_j$ $V_{\Re_i}$: Volumen prisma $\Re_i$	Método muy sencillo, bajos costes computacionales y temporales. Estimación discreta de la PDF, pero está normalizada. La calidad de la estimación muy dependiente del tamaño y posición de las zonas en que se divide el espacio. Estimación muy pobre con baja cantidad de muestras o para elevado número de variables.
	Uso de los k vecinos más próximos(k-NN) Dado un vector $\boldsymbol{x}$, se aumenta un volumen esférico a su alrededor hasta que abarque un total de k puntos del conjunto de entrenamiento.	$\hat{f}_{kNN}(\boldsymbol{x}\|H_j) = \frac{k_j}{N_j \cdot c_d \cdot R_k^d(\boldsymbol{x})}$ $R_k^d(x)$: distancia entre x y su k-ésimo vecino más cercano en $\mathbb{R}^d$ c_d: volumen de esfera unidad en $\mathbb{R}^d$	k-NN no obtiene muy buenas estimaciones de las densidades. Búsqueda de vecinos es una tarea compleja que puede requerir gran capacidad de cálculo. Es necesaria una gran capacidad de almacenamiento ya que requiere del almacenaje de todos los datos de entrenamiento. Habitual utilizar este método para implementar de forma directa una aproximación del detector
	Uso funciones de núcleo (*KDE*) Media de *PDFs* $K(\cdot) \in \mathbb{R}^d$ centradas en puntos del espacio $\mathbb{R}^d$ en que se encuentran los vectores de entrenamiento $\boldsymbol{x}^i \in \mathcal{X}_j$	$\hat{f}_{KDE}(\boldsymbol{x}\|H_j) = \frac{1}{N_j} \sum_{i=1}^{N_j} K_{\boldsymbol{H}}(\boldsymbol{x} - \boldsymbol{x}^i)$ $K(\cdot)$: *PDF* multivariante conocida H: matriz de ancho de banda	Se pueden obtener buenas estimaciones, aun cuando los datos presenten complejas estructuras de dependencia y heterogeneidad. Calidad de la estimación depende del ancho de banda elegido. Complejos algoritmos de selección de ancho de banda óptimo. Involucra grandes costes computacionales. Necesaria gran capacidad de almacenamiento ya que requiere del almacenaje de todos los datos de entrenamiento
Estimación paramétrica de densidades	Mezcla de *PDFs* Suma ponderada de un conjunto de K *PDFs*. El modelo más utilizado es el de mezclas Gaussianas *GMM* ($g(\cdot)$: *PDF* Gaussiana multivariante)	$\hat{f}_{GMM}(\boldsymbol{x}\|H_j) = \sum_{i=1}^{K} \omega_i \cdot g_i(\boldsymbol{x}\|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ ω_i: ponderación de cada componente	Se pueden obtener estimaciones continuas y robustas con buenas propiedades asintóticas. Se pueden modelar complejas estructuras de dependencia y heterogeneidad en los datos. Se requiere de grandes cantidades de muestras para obtener buena precisión. Modelos muy complejos pueden necesitar una gran cantidad de componentes, conllevando un gran coste computacional el entrenamiento. Importantes degradaciones pueden presentarse cuando el número de variables es grande.
	Modelo de mezcla de expertos Mezcla de *PDFs* $(f_e(\cdot))$, pero en este caso las ponderaciones también dependen de las observaciones $\boldsymbol{x}$ (a través de $f_g(\cdot)$). Común el uso de las funciones Soft-Max y Gaussianas.	$\hat{f}_{ME}(\boldsymbol{x}\|H_j) = \sum_{k=1}^{K} g_k(\boldsymbol{x}\|\Theta_g^k)\, f_e(\boldsymbol{x}\|\Theta_e^k)$ $g_k(\boldsymbol{x}) = \frac{f_g(\boldsymbol{x}, \theta_k)}{\sum_{i=1}^{K} f_g(\boldsymbol{x}, \theta_i)}$	

Tabla 7.1 – Comparativa de técnicas de fusión soft

Teoría de cópulas

Teorema de Sklar

La *CDF* $F(z_1, , \dots, z_d)$ de una serie de *v.a* continuas $Z_1, \dots, Z_d$ puede ser expresada en función de las CDF marginales $F_i(\cdot)$:

$$F(z_1, \dots, z_d) = C\big(F_1(z_1), F_2(z_2), \dots, F_d(z_d)\big), \quad C: U(0\,1)^d \to (0\,1), \quad u_i = F_i(z_i)$$

Si la función de cópula C y las distribuciones marginales $F_i(\cdot)$: son lo suficientemente diferenciables:

$$f(z_1, \dots, z_d) = f_1(z_1) \cdot f_2(z_2) \cdot \dots \cdot f_d(z_d) \cdot c\big(F_1(z_1), \dots, F_d(z_d)\big)$$

Esta estructura permite separar el modelado de las distribuciones marginales y de la estructura de dependencia.

Definida la función de cópula $C(u_1, \dots, u_d)$, se obtiene la densidad mediante:

$$c(\boldsymbol{u}) = \frac{\partial^d\big(C(u_1, u_2, \dots, u_d)\big)}{\partial u_1 \partial u_2, \dots, \partial u_d}$$

Sea $f_m(\boldsymbol{z_m})$ una *PDF* multivariante conocida, con marginales $f_{m_i}(\cdot)$ y $F_{m_i}(\cdot)$, se define su densidad de cópula como:

$$c_m(\boldsymbol{u}) = \frac{f_m\left({F_{m_1}}^{-1}(u_1), \dots, {F_{m_d}}^{-1}(u_d)\right)}{\left(\prod_{i=1}^d f_{m_i}\left({F_{m_i}}^{-1}(u_i)\right)\right)}$$

Funciones densidad de cópula obtenidas de una PDF multivariante conocida

Densidad de cópula Gaussiana

$$c_{Gauss}(\boldsymbol{u}) = \frac{1}{|\Sigma_\rho|^{1/2}} \cdot \exp\left(\frac{-\boldsymbol{z_m}^T \cdot (\Sigma_\rho{}^{-1} - \boldsymbol{I}) \cdot \boldsymbol{z_m}}{2}\right)$$

$$\boldsymbol{z_m} = \left[z_{m_1} = \Phi^{-1}(u_1), \dots, z_{m_d} = \Phi^{-1}(u_d)\right]^T$$

Densidad de cópula Student-T

$$c_T(\boldsymbol{u}) = \frac{1}{|\Sigma_\rho|^{\frac{1}{2}}} \cdot \frac{\Gamma\left(\frac{v+d}{2}\right)}{\Gamma\left(\frac{v}{2}\right)} \cdot \left[\frac{\Gamma\left(\frac{v}{2}\right)}{\Gamma\left(\frac{v+1}{2}\right)}\right]^d \cdot \frac{\left(1 + \frac{-\boldsymbol{y}^T \Sigma_\rho{}^{-1} \boldsymbol{y}}{v}\right)^{-\frac{(v+d)}{2}}}{\prod_{i=1}^d \left(1 + \frac{y_i^2}{v}\right)^{-\frac{(v+1)}{2}}}$$

$$\boldsymbol{y} = [y_1 = t_v^{-1}(u_1), \dots, y_d = t_v^{-1}(u_d)]^T, \qquad v > 2$$

Densidad de cópula mezcla de Gaussianas

PDF multivariante: $f_{GMM}(\boldsymbol{z_m}) = \sum_{i=1}^K \omega_i \cdot g(\boldsymbol{z_m}|\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i})$

Marginales: $f_{m_j}\left(z_{m_j}\right) = \sum_{i=1}^K \omega_i \cdot g(z_{m_j}|\mu_j^{(i)}, \sigma_j^{2(i)})$

F_{m_j} y $F_{m_j}^{-1}$ obtenidas empíricamente

$$c_{GMCM}(\boldsymbol{u}) = \frac{f_{GMM}\left({F_{m_1}}^{-1}(u_1), \dots, {F_{m_d}}^{-1}(u_d)\right)}{\left(\prod_{j=1}^d f_{m_j}\left({F_{m_j}}^{-1}(u_j)\right)\right)}$$

Funciones de cópula Arquimedianas: Familia con función de cópula $C(u_1, \dots, u_d) = \psi^{-1}\big(\psi(u_1) + \dots + \psi(u_d)\big)$, $\psi(u)$: función generadora

Cópula de Clayton

$$\psi(u) = \tfrac{1}{\theta}\left(u^{-\theta} - 1\right) \quad \psi^{-1}(y) = (\theta \cdot y + 1)^{-1/\theta}$$

$$C(\boldsymbol{u}) = \left(\sum_{i=1}^d u_i{}^{-\theta-1} - d + 1\right)^{-1/\theta}, \qquad \theta > 0$$

Cópula de Frank

$$\psi(u) = ln\left(\frac{e^{-\theta u} - 1}{e^{-\theta} - 1}\right) \quad \psi^{-1}(y) = -\frac{1}{\theta} ln\left(1 + e^y\left(e^{-\theta} - 1\right)\right)$$

$$C(\boldsymbol{u}) = -\frac{1}{\theta} ln\left(1 + \frac{\prod_{i=1}^d \left(e^{-\theta u_i} - 1\right)}{(e^{-\theta} - 1)^{d-1}}\right)$$

Cópula de Gumbel

$$\psi(u) = \left(-ln(u)\right)^\theta \quad \psi^{-1}(y) = e^{-y^{1/\theta}}$$

$$C(\boldsymbol{u}) = exp\left(-\left(\sum_{i=1}^d (-ln\, u_i)^\theta\right)^{1/\theta}\right), \qquad \theta > 1$$

Cópula de Farlie-Gumbel-Morgenstern (FGM)

$$c(u_1, \dots, u_d;\ \theta) = 1 + \sum_{s=2}^d \sum_{1 \le j_1 < \dots < j_s \le d} \alpha_{j_1 \dots j_s} \prod_{i=1}^s \left(1 - 2u_{j_i}\right), \quad 1 + \sum_{r=2}^m \sum_{k_1 \le j_1 < \dots < j_s \le k_m} \left(\alpha_{j_1, \dots, j_s} \prod_{i=1}^s \zeta_{j_i}\right) \ge 0$$

Modelado de función densidad de cópula mediante árboles de pares

Se obtiene la densidad de cópula mediante un esquema jerárquico en parejas de dos variables. Para distribuciones de grandes dimensiones, existen un número muy elevado de posibles descomposiciones en productos de cópulas bivariantes.

Es común usar un modelo de construcción de densidades de cópula usando "vines" (más concretamente vines regulares)

C-Vines:

$$c(z_1, \dots, z_d) = \prod_{j=1}^d \prod_{i=1}^{d-j} c_{j,j+i|1,\dots,j-1}\left(F\left(z_j|z_1, \dots, z_{j-1}\right), F\left(z_{i+j}|z_1, \dots, z_{j-1}\right)\right)$$

D-Vines:

$$c(z_1, \dots, z_d) = \prod_{j=1}^d \prod_{i=1}^{d-j} c_{i,i+j|i+1,\dots,i+j-1}\left(F\left(z_i|z_{i+1}, \dots, z_{i+j-1}\right), F\left(z_{i+j}|z_{i+1}, \dots, z_{i+j-1}\right)\right)$$

Tabla 7.1 – Teoría de cópulas

Fusión de scores $\boldsymbol{s} = [s_1 \dots s_d]^T : [0,1]^d \rightarrow s_{fus} : [0,1]$			
Fusión basada en la relación de verosimilitud		$\Lambda(\boldsymbol{s}) = \frac{f(\boldsymbol{s}\|H_1)}{f(\boldsymbol{s}\|H_0)} \lessgtr_{H_1}^{H_0} \lambda$ Estimar las *PDFs* bajo cada hipótesis $(H_j,\ j = 0,1)$ usando un conjunto de datos de entrenamiento $S_j = \{\boldsymbol{s}^1, \dots, \boldsymbol{s}^{N_j}\},\ \boldsymbol{s}^i = [s_1^i, \dots, s_d^i]^T \in H_j\ , i = 1, \dots, N_j$	En la mayoría de casos los scores se obtienen a la salida de diferentes detectores, clasificadores o algoritmos de detección, presentando buenas características de discriminación entre hipótesis. Se puede presuponer una fusión más simple mediante una separación entre hipótesis en el espacio de los scores mediante una hipersuperficie continua, pudiendo utilizar diversas técnicas más simples.
Fusión de detectores a través de probabilidades a posteriori		$s_{fus} = \frac{k^{-(d-1)} \cdot \prod_{i=1}^{d} s_i}{\prod_{i=1}^{d}(1-s_i) + k^{-(d-1)} \cdot \prod_{i=1}^{d} s_i},\ k = \frac{P_{H_1}}{P_{H_0}}$	Óptima en el caso en que los detectores trabajen con diferentes entradas independientes estadísticamente entre sí.
Funciones y reglas simples de fusión	Funciones de media	$s_{fus} = \frac{1}{d}\sum_{i=1}^{d} s_i \quad s_{fus} = \left(\prod_{i=1}^{d} s_i\right)^{1/d} \quad s_{fus} = \frac{d}{\sum_{i=1}^{d} \frac{1}{s_i}}$	Técnicas muy simples que no requieren de entrenamiento y que bajo determinadas distribuciones de los datos pueden proporcionar buenos resultados. Separación entre hipótesis muy rígidas, sin capacidad de adaptación.
	Reglas del máximo y mínimo	$s_{fus} = \max(s_i) \quad s_{fus} = \min(s_i)$	
	Combinación lineal: Suma y productos ponderados	$s_{fus} = \sum_{i=1}^{d} w_i \cdot s_i = \boldsymbol{w}^T \boldsymbol{s} \qquad s_{fus} = \prod_{i=1}^{d} s_i^{w_i}$ $s_i,\ s_{fus}, \in [0,1] \ \Leftrightarrow\ \sum_{i=1}^{d} w_i = 1, \qquad 0 \le w_i \le 1$	Se mejoran las prestaciones añadiendo la capacidad de adaptarse a las diferentes características de los datos mediante los coeficientes de ponderación. Requiere de entrenamiento.
Integración α		$s_\alpha(\boldsymbol{s}) = \begin{cases} \left(\sum_i w_i \cdot s_i^{\frac{1-\alpha}{2}}\right)^{\frac{2}{1-\alpha}} & , \alpha \neq 1 \\ exp\left(\sum_i w_i \cdot log(s_i)\right) & , \alpha = 1 \end{cases}$	Las reglas y funciones simples de combinación pueden obtenerse como un caso particular de esta técnica. Aporta un mayor grado de flexibilidad y de adaptación a las distribuciones que presenten los datos. Requiere de entrenamiento. Se ha propuesto un método de entrenamiento orientado a obtener las mejores prestaciones posibles en detección mediante la maximización del área bajo la curva *ROC* en el intervalo que se desee.
Arquitectura de mezcla de expertos		$y_j = \sum_{i=1}^{d} g_i\left(\boldsymbol{x}, \Theta_{g_j}\right) s_i, \quad j = 0,1 \rightarrow s_{fus} = \frac{y_1}{y_1 + y_0}$ $g_i(\boldsymbol{x}, \Theta_i) = \frac{f_g(\boldsymbol{x}, \theta_i)}{\sum_{k=1}^{d} f_g(\boldsymbol{x}, \theta_k)}$	Orientado sobre todo a la fusión de diferentes detectores, permitiendo combinarlos con ponderaciones dependientes de las observaciones $\boldsymbol{x}$, discriminando así la zona del espacio de observaciones donde es mejor cada uno de ellos.

Tabla 7.2 – Comparativa de técnicas de fusión de scores

<table>
<tr><th colspan="4">Fusión hard $\quad \boldsymbol{u} = [u_1 \dots u_d]^T : \{0,1\}^d \rightarrow u_{fus} : \{0,1\}$</th></tr>
<tr><td colspan="4">$$\Lambda(\boldsymbol{u}) = \frac{P(\boldsymbol{u}|H_1)}{P(\boldsymbol{u}|H_0)} = \frac{P(u_1, \dots, u_d|H_1)}{P(u_1, \dots, u_d|H_0)} \gtrless_{H_0}^{H_1} \eta \Leftrightarrow u_{fus} = \begin{cases} 1 & si\ \Lambda(\boldsymbol{u}) \geq \eta \\ 0 & si\ \Lambda(\boldsymbol{u}) < \eta \end{cases} \quad \boldsymbol{u}: \{0,1\}^d \rightarrow \Lambda(\boldsymbol{u}): \{a_1, \dots, a_{2^d}\} \subset \mathbb{R}$$</td></tr>
<tr><td>Asunción de independencia</td><td colspan="2">$$P_{ind}(\boldsymbol{u}|H_j) = \prod_{i=1}^{d} P(u_i|H_j)$$
$$\mathcal{T}(\boldsymbol{u}) = \sum_{i=1}^{d} u_i \cdot c_i \quad u_{fus} = \begin{cases} 1 & si\ \mathcal{T}(\boldsymbol{u}) \geq \eta \\ 0 & si\ \mathcal{T}(\boldsymbol{u}) < \eta \end{cases} \quad c_i = \log\left(\frac{P_{d_i} \cdot (1 - P_{f_i})}{P_{f_i} \cdot (1 - P_{d_i})}\right)$$</td><td>Regla simple de fusión.
En caso de que exista dependencia estadística puede conllevar a una degradación en las prestaciones obtenidas en el sistema de detección.</td></tr>
<tr><td rowspan="6">Estimación de PMFs</td><td colspan="2">Estimación directa</td><td rowspan="6">Utilizando un número elevado de vectores de observación se obtienen estimaciones altamente precisas, incorporando toda la información de dependencia que presenta la estructura de datos.
En el caso de un número elevado de variables se requiere de una gran cantidad de memoria para almacenar la gran cantidad de valores que precisa la PMF.</td></tr>
<tr><td colspan="2">$$P(\boldsymbol{u}^i|H_j) = \frac{Nº\ de\ apariciones\ de\ \boldsymbol{u}^i\ bajo\ H_j}{Nº\ total\ de\ realizaciones\ de\ \boldsymbol{u}\ bajo\ H_j}$$</td></tr>
<tr><td colspan="2">Desarrollo de Drakopoulos</td></tr>
<tr><td colspan="2">$$P(\boldsymbol{u}|H_j) = P_j(u_1, \dots, u_d) = \sum_{I \subseteq A_0} (-1)^{|I|} E_j\left[\prod_{i \in A_1 \cup I} u_i\right]$$
$$A_\mu = \{i : u_i = \mu\} \quad 1 \leq i \leq d, \quad \mu = 0,1. \quad E_j[\textstyle\prod_{i \in A} u_i] = 1\ si\ A = \phi$$</td></tr>
<tr><td colspan="2">Expansión de Bahadur-Lazarsfeld</td></tr>
<tr><td colspan="2">$$P(\boldsymbol{u}|H_j) = P_{ind}(\boldsymbol{u}|H_j)\left[1 + \sum_{k<l} \gamma_{kl}^j w_k w_l + \sum_{k<l<m} \gamma_{klm}^j w_k w_l w_m + \dots + \gamma_{12\dots d}^j w_1 w_2 \dots w_d\right]$$
$$\gamma_{kl}^j = \sum_{\boldsymbol{u}} w_k\, w_l P(\boldsymbol{u}|H_j), \dots, \quad \gamma_{12\dots n}^j = \sum_{\boldsymbol{u}} w_1 \dots w_n P(\boldsymbol{u}|H_j)$$</td></tr>
<tr><td rowspan="4">Métodos subóptimos</td><td rowspan="2">Estimación subóptima de PMFs</td><td>Truncamiento de la expansión de Bahadur-Lazarsfeld</td><td rowspan="2">Se pueden obviar dependencias de órdenes más elevados para obtener estimaciones con menos cantidad de parámetros.
Dependiendo de la estructura de dependencia global puede suponer cierta degradación.</td></tr>
<tr><td>Aproximar las PMFs conjuntas mediante el producto de sus distribuciones de menor orden</td></tr>
<tr><td>Regla del conteo (Majority voting):</td><td>$$m = \sum_{i=1}^{d} u_i \lesssim_{H_1}^{H_0} k,\ k: nº\ de\ detectores\ a\ favor\ de\ H_1$$</td><td rowspan="2">Reglas muy simples de fusión. Bajo determinadas condiciones de dependencia pueden proporcionar buenos resultados. La regla del conteo se ha demostrado que es óptima en el caso de detectores iguales e independientes.
Dependiendo de la estructura de dependencia se pueden producir importantes degradaciones en las prestaciones de detección.</td></tr>
<tr><td colspan="2">Reglas AND, OR y XOR</td></tr>
</table>

Tabla 7.3 – Comparativa de técnicas de fusión hard

Lista de publicaciones

Publicaciones en Revistas Internacionales

- A. Soriano, L. Vergara, J. Moragues, R. Miralles, “Unknown signal detection by one-class detector based on Gaussian copula”, Signal Processing, Volume 96, Part B, March 2014, Pages 315-320, ISSN 0165-1684.
- A. Soriano, L. Vergara, G. Safont, and A. Salazar, "On the fusion of two equally-operated non-independent detectors", Information Fusion, Elsevier, submitted to *Information Fusion*, April. 2013.
- J. Moragues, A. Soriano, A. Serrano and L. Vergara, "Acoustic detection by fusion of multiple audio channels", submitted to *Journal of the Acoustical Society of America*, Jun. 2013.
- A. Soriano, L. Vergara: “Fusion of scores based on α-integration in a detection context”, Submitted to *Information Fusion*. November 2013.

Publicaciones en congresos internacionales

- G. Safont, A. Salazar, A. Soriano, and L. Vergara, “Automatic Credit Card Fraud Detection based on Non-linear Signal Processing,” in *Proceedings of the 46th IEEE International Carnahan Conference on Security Technology*, pp. 207-212, Boston (USA), 2012.
- G. Safont, A. Salazar, A. Soriano, and L. Vergara, “Combination of Multiple Detectors for EEG based Biometric Identification/Authentication,” in *Proceedings of the 46th IEEE International Carnahan Conference on Security Technology*, pp. 230-236, Boston (USA), 2012.
- A. Soriano, L. Vergara, G. Safont, and A. Salazar, “On Comparing Hard and Soft Fusion of Dependent Detectors,” in *Proceedings of the 2012 IEEE International Workshop on Machine Learning for Signal Processing*, in print, Santander (Spain), 2012.

Publicaciones en congresos nacionales

- A. Soriano, L. Vergara, G. Safont, and A. Salazar, “Hard versus soft fusion of dependent data”, in *URSI*: *xxvii simposium nacional de la unión científica internacional de radio,Elche (Spain), 2012.*
- G. Safont, A. Salazar, A. Soriano and L. Vergara, “Análisis de componentes independientes aplicado a la recuperación de señales GPR”, in URSI: xxvii simposium nacional de la unión científica internacional de radio,Elche (Spain), 2012.

Apéndices

Apéndice A:
Normalización de datos soft

Cuando se pretenden combinar distintas fluentes de datos soft, con objeto de usar un único dominio común, se suelen utilizar funciones de normalización $\mathfrak{n}(\cdot)\colon \mathbb{R} \to \mathbb{R}$ para transformar las valoraciones $z \in [a,b] \neq [0,1]$ y ubicarlas en el rango normalizado $[0,1]$. Denominamos de forma general como 'score', denotado por $s = \mathfrak{n}(z) \in [0,1]$, a cualquier información normalizada entre cero y uno.

En [22] se recopilan, explican y analizan varias técnicas y funciones de normalización. Realizamos un breve resumen de ellas:

- Obtención de la probabilidad a posteriori de la valoración:

Se deben estimar tanto las probabilidades a priori de cada una de las hipótesis, como las *PDFs* de la valoración z bajo cada una de ellas. Así, obtenemos la probabilidad a posteriori como:

$$s = P(H_1|z) = \frac{f(z|H_1)P(H_1)}{f(z|H_1)P(H_1) + f(z|H_0)P(H_0)} \tag{A.1}$$

Un detector se dice bien calibrado si la probabilidad a posteriori de su score a la salida $s \in [0,1]$ tiende a ser igual a s: $P(H_1|s) = s$. Si se cumple $P(H_1|s) = P(H_1|\boldsymbol{x})$ entonces el score es óptimo. En el apéndice B se incluyen técnicas para calibrar scores.

- Técnica Min-Max:

Se necesita conocer los valores máximo y mínimo que toma la valoración $z \in [min(z), max(z)]$:

$$s = \frac{z - min(z)}{max(z) - min(z)} \tag{A.2}$$

- Función sigmoide doble:

Mediante esta normalización se trata de realizar una transformación lineal en la región de solape entre hipótesis, mientras que los valores fuera de esta región se transforman de forma no lineal. Se necesitan fijar los parámetros t, r_1 y r_2. El parámetro t se fija a un valor dentro de la zona de solape entre hipótesis; los parámetros r_1 y r_2 se escogen de forma que abarquen toda la zona de solape, fijando r_1 a la izquierda de t y r_2 a la derecha (figura A.1)

$$s = \begin{cases} \dfrac{1}{1 + exp\left(-2\dfrac{z-t}{r_1}\right)} & \text{si } z < t \\ \dfrac{1}{1 + exp\left(-2\dfrac{z-t}{r_2}\right)} & \text{resto} \end{cases} \tag{A.3}$$

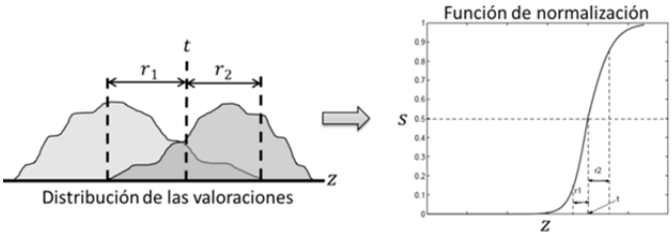


Figura A.1 – Normalización mediante la función sigmoide doble

- Normalización mediante tangente hiperbólica:

La función de normalización viene dada por:

$$s = \frac{1}{2}\left[\tanh\left(0.01\left(\frac{z - \mu_{r_{H_1}}}{\sigma_{r_{H_1}}}\right)\right) + 1\right] \tag{A.4}$$

Donde $\mu_{r_{H_1}}$ y $\sigma_{r_{H_1}}$ representan la media y la desviación estándar de la distribución asociada a H_1, usualmente obtenidas mediante una estimación robusta que evite *outliers*. Por ejemplo, en [22] se propone el uso de los estimadores de Hampel (en [144] podemos encontrar una descripción más detallada).

Apéndice B:
Calibración de scores: transformación en estimadores de probabilidad

Denotemos mediante $\mathfrak{s}(\cdot)$: $\mathbb{R}^n \rightarrow \mathbb{R} \in [0,1]$ a la función o proceso seguido por un determinado detector para aportar un score $s = \mathfrak{s}(\boldsymbol{x}) \in [0,1]$ a su salida. Cualquier detector busca obtener una salida en la que, dadas dos observaciones $\boldsymbol{x}$ e $\boldsymbol{y}$ en las que se cumple $\mathfrak{s}(\boldsymbol{x}) > \mathfrak{s}(\boldsymbol{y})$, siempre implique que $P(H_1|\boldsymbol{x}) > P(H_1|\boldsymbol{y})$, es decir, el score actúa como un ranking asociado a la probabilidad a posteriori.

En determinados casos, como por ejemplo si se desea combinar con otros scores mediante algún criterio estadístico, será conveniente que estos scores actúen como un estimador real de una probabilidad a posteriori. Un detector se dice bien calibrado si la probabilidad a posteriori de su score a la salida $s \in [0,1]$ tiende a ser igual a s: $P(H_1|s) = s$. En el caso de un detector bien calibrado, su score de salida puede interpretarse como una estimación de la probabilidad a posteriori $P(H_1|s) = s \approx P(H_1|\boldsymbol{x})$. Es decir, si un detector nos proporciona un score de 0.8, debe significar, que el 80 % de las veces que proporciona ese valor se corresponde con la hipótesis H_1. Si se cumple $P(H_1|s) = P(H_1|\boldsymbol{x})$, entonces el score es óptimo.

Mediante un diagrama de fiabilidad (figura B.1) se puede visualizar si un detector está bien o mal calibrado, el cual se representa la probabilidad $P(H_1|s)$ con respecto al score s. Cuando el detector está bien calibrado la representación deber ser una recta $x = y$.

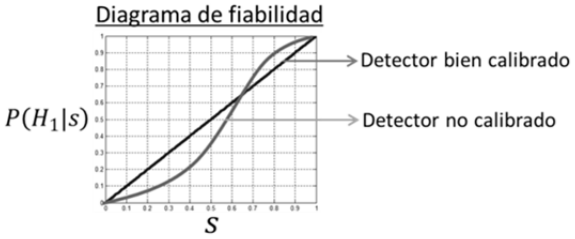


Figura B.1 – Diagrama de fiabilidad de un detector

Si conocemos la función $\mathfrak{c}(\cdot)$: $[0,1] \rightarrow [0,1]$ que mapea un score s hacia la probabilidad a posteriori $P(H_1|s)$, podemos obtener el score calibrado $s_c = \mathfrak{c}(s) \rightarrow P(H_1|s_c) = s_c$.

La función de mapeo es generalmente desconocida y habrá que estimarla. En [145] se proponen tres métodos para ello: mediante una función sigmoide, mediante una estimación no paramétrica mediante histogramas o mediante una regresión isotónica (basándose en la suposición de que esta función es monótona creciente) calculada mediante una técnica *PAV* (del inglés, "*Pair-Adjacent Violators*"). El uso de la función sigmoide está motivado por el hecho empírico de que en muchos estudios con máquinas de soporte de vectores (*SVM*, del inglés "*Support Vector Machines*") parece que la relación entre los scores de las *SVM* y las probabilidades $P(H_1|s)$ se adecúa a

esta función, pero en [145] se muestra un caso de detector basado en naive Bayes donde la función sigmoide no se adapta bien. Por lo tanto nos centramos en los otros dos métodos:

- Estimación no paramétrica mediante histogramas:

Se basa en calcular la probabilidad empírica $P(H_1|s)$ en un conjunto de regiones determinadas por un troceado del rango de los scores en bines de igual tamaño. Considerando un bin $\mathfrak{B}_k = \{s|\ s_a \leq s < s_b\}$, calculamos la probabilidad empírica mediante el conteo del número de muestras $n_{1_{\mathfrak{B}_k}}$ con scores pertenecientes a H_1, dividido entre el número total de muestras $n_{\mathfrak{B}_k}$ con scores que recaen en el bin $\mathfrak{B}_k$.

Los principales problemas del uso de este método son, primero, que para escoger un número de bines óptimo debe realizarse mediante validación cruzada (dividir el conjunto de entrenamiento en dos partes, una para obtener la estimación del mapeo mediante diferentes números de bines y la otra para evaluar cuál de las estimaciones es mejor) y que la división en bines equiespaciados puede no ser la adecuada para obtener buenas estimaciones de la probabilidad en zonas con diferentes densidades de puntos.

- Mapeo estimado mediante una regresión isotónica:

Motivados por el hecho de que si un detector es capaz de asignar scores s proporcionales a la probabilidad a posteriori óptima $P(H_1|\boldsymbol{x})$, es fácil deducir que el mapeo del diagrama de fiabilidad es no decreciente y por lo tanto se puede usar una regresión isotónica para entrenar el mapeo. Así, en [145] proponen el uso del algoritmo *PAV* [146] para el entrenamiento de una función isotónica en forma de escalera que mejor se adapta a los datos de acuerdo con un criterio de minimización del error cuadrático.

En la figura B.2 se puede ver un ejemplo de calibración de un cierto score. En el diagrama de fiabilidad de la izquierda se observa que el score no se encuentra calibrado. Una vez calibrado mediante ambas técnicas se muestra el diagrama de fiabilidad a la derecha.

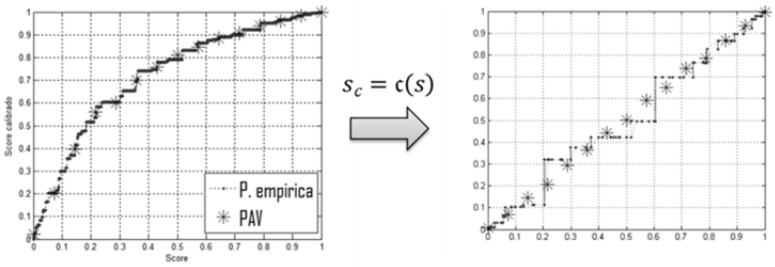


Figura B.2 – Ejemplo de calibración de un score mediante ambos métodos

Apéndice C:
Principios de la estimación de componentes independientes en ICA

El análisis de componentes independientes (*ICA*, "*Independent Component Analysis*") es un método de procesado cuyo objetivo es encontrar una representación linear de datos multivariantes no Gaussianos, de forma que se puedan expresar en función de unas componentes que sean independientes (o lo más independientes posible). Se considera que cada una de las variables aleatorias que modelan a los datos $X_1, \dots, X_d$ es el resultado de una mezcla lineal de *n* componentes, representadas por las variables aleatorias $S_1, \dots, S_n$. Se conoce como modelo estadístico de variables latentes representado por:

$$\boldsymbol{x} = \boldsymbol{A} \cdot \boldsymbol{s} = \sum_{i=1}^{n} \boldsymbol{a}_i \cdot s_i \qquad \text{(C.1)}$$

Tanto la matriz de mezclas, como el vector de componentes $\boldsymbol{s}$, son desconocidos y deben ser estimados a partir del vector de observaciones $\boldsymbol{x}$. El punto de partida de *ICA* es la asunción de independencia entre las componentes s_i. Sabiendo que la independencia implica incorrelación, muchos algoritmos de *ICA* restringen el proceso de estimación, forzando también la incorrelación de las componentes independientes, reduciendo de esta forma el número de grados de libertad en la estimación y simplificando el problema.

Otra de las asunciones que debe asumirse es que las componentes independientes deben de poseer distribuciones no Gaussianas, ya que la matriz de mezclas $\boldsymbol{A}$ no puede ser identificable en el caso de que todas las componentes sean Gaussianas independientes. Esta conclusión se deriva del hecho de que aplicando cualquier transformación ortogonal a una función Gaussiana multivariante se obtiene la misma distribución en la que sus componentes son independientes (ver demostración en [147]).

En el modelo general de *ICA* se suponen las distribuciones de las componentes desconocidas (otras variantes las suponen conocidas, de forma que el problema se simplifica de manera considerable). Se estima inicialmente la matriz de mezclas $\boldsymbol{A}$, para después calcular su inversa $\boldsymbol{W} = \boldsymbol{A}^{-1}$ y poder obtener de forma simple las componentes independientes mediante:

$$\boldsymbol{s} = \boldsymbol{W} \cdot \boldsymbol{x} \qquad \text{(C.2)}$$

Modelos de estimación ICA

El método *ICA* se basa en estimar las componentes maximizando de alguna forma la independencia estadística entre ellas. Podemos encontrar diferentes técnicas para

definir esta independencia, las cuales gobiernan el funcionamiento de los diferentes algoritmos *ICA* que podemos encontrar en la literatura. Las más populares son la maximización de la no gaussianidad, la minimización de la información mutua y la estimación por máxima verosimilitud:

❖ Maximización de la no gaussianidad

Aparte del hecho ya comentado sobre el problema de tener componentes independientes gaussianas, se puede encontrar la demostración de que maximizando la no gaussianidad entre componentes se alcanza la independencia entre ellas. Esta demostración se deriva del teorema del límite central; se puede encontrar de forma detallada en [147] .

Los algoritmos basados en esta técnica, tratan de maximizar una medida cuantitativa de la no gaussianidad de una variable aleatoria y. Dos medidas de la no gaussianidad son utilizadas: la kurtosis o la entropía negativa:

- Kurtosis

$$kurt(y) = E[y^4] - 3(E[y^2])^2 \quad \text{(C.3)}$$

La kurtosis es cero para una variable aleatoria gaussiana. Para muchas (no es generalizable para todas) variables aleatorias no gaussianas, la kurtosis es distinta de cero. La kurtosis puede ser tanto positiva como negativa. Las variables aleatorias que poseen kurtosis negativa son llamadas subgaussianas, y aquellas que poseen kurtosis positiva son llamadas supergaussianas.

- Entropía negativa

La entropía es un concepto básico en la teoría de la información. La entropía de una variable aleatoria se puede interpretar como el grado de información que una observación de la variable nos proporciona. Cuanto más aleatoria, impredecible y desestructurada sea una variable, más grande será su entropía. La expresión de la entropía $H(\cdot)$ de un vector de variables aleatorias $\boldsymbol{y}$ con *PDF* $f(\boldsymbol{y})$ es:

$$H(\boldsymbol{y}) = -\int f(\boldsymbol{y})\, log\big(f(\boldsymbol{y})\big) \quad \text{(C.4)}$$

Un resultado fundamental de la teoría de la información es que una variable gaussiana posee la mayor entropía de entre todas las variables a igualdad de varianzas. Por lo tanto la medida de la entropía puede ser usada para medir no gaussianidad. Para obtener una medida de no gaussianidad que sea cero para variables gaussianas y siempre positiva para variables no gaussianas se utiliza una versión modificada de la medida de la entropía, llamada entropía negativa, y definida como $J(\cdot)$:

$$J(\boldsymbol{y}) = H\big(\boldsymbol{y}_{\boldsymbol{gauss}}\big) - H(\boldsymbol{y}) \quad \text{(C.5)}$$

donde $\boldsymbol{y_{gauss}}$ hace referencia a un vector aleatorio gaussiano con la misma matriz de covarianza que el vector aleatorio $\boldsymbol{y}$.

❖ Minimización de la información mutua

Otra técnica usada en la estimación *ICA*, inspirada en la teoría de la información, es la minimización de la información mutua. Se define la información mutua $I(\cdot)$ entre un conjunto de m variables aleatorias Y_i como:

$$I(y_1, \dots, y_m) = \sum_{i=1}^{m} H(y_i) - H(\boldsymbol{y}) \tag{C.6}$$

La información mutua es una medida natural de la dependencia entre variables aleatorias. Es equivalente a la divergencia de Kullback-Leibler entre la densidad conjunta $f(\boldsymbol{y})$ y el producto de sus densidades marginales, una medida de independencia muy utilizada en la literatura. Es siempre positiva, y cero si y sólo si las variables implicadas son estadísticamente independientes. Así, la información mutua toma en consideración toda posible estructura de dependencia entre las variables, no solo la proporcionada por la matriz de covarianza como en otros métodos.

❖ Estimación por máxima verosimilitud

Una de las técnicas más populares en la estimación del modelo *ICA* es la estimación por máxima verosimilitud. Se puede demostrar que esencialmente es equivalente a la minimización de la información mutua [147]. El modelo de estimación trata de maximizar la función de verosimilitud L dada por la expresión:

$$L = \sum_{t=1}^{T} \sum_{i=1}^{n} ln\left(f_i\left(\boldsymbol{w}_i^T \boldsymbol{x}(t)\right)\right) + T \cdot ln\left(|\boldsymbol{W}|\right) \tag{C.7}$$

donde $\boldsymbol{W} = [\boldsymbol{w}_1, \dots, \boldsymbol{w}_n]^T$ denota la inversa de la matriz de mezclas $\boldsymbol{A}$, $f_i(\cdot)$ son las *PDF* de las componentes s_i y $\boldsymbol{x}(t)$ son las realizaciones del vector de observaciones a lo largo de un tiempo discreto T.

Preprocesado ICA

Antes de aplicar cualquier algoritmo *ICA* suele ser habitual y muy recomendable realizar un pequeño preprocesado de los datos, para que la estimación mediante *ICA* sea más simple y esté mejor condicionada. Dos tipos de preprocesado suelen ser habituales, primero un centrado de los datos y a continuación un blanqueo de éstos.

El centrado se basa en centrar las observaciones mediante la sustracción de la media.

$$x' = x - E[x] = x - m \tag{C.8}$$

Esto implica que las componentes $\boldsymbol{s}$ también serán de media nula. Una vez aplicado *ICA* se puede añadir las medias otra vez al vector de componentes:

$$s' = s + m_s = s + A^{-1}m \tag{C.9}$$

El blanqueo de los datos hace referencia a una transformación lineal del vector $\boldsymbol{x}'$ para incorrelar sus componentes y normalizar sus varianzas a la unidad, lo que equivale a decir que la matriz de covarianza de los nuevos datos $\tilde{\boldsymbol{x}}$ es la identidad $E[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^T] = \boldsymbol{I}$. La operación de blanqueo es siempre posible. Un método muy popular para llevarla a cabo es usando la descomposición en valores singulares (*EVD*, del inglés "*Eigen-Value Decomposition*") de la matriz de covarianza $E[\boldsymbol{x}'\boldsymbol{x}'^T] = \boldsymbol{R} = \boldsymbol{EDE^T}$, donde $\boldsymbol{E}$ es la matriz ortogonal de vectores singulares de $\boldsymbol{R}$ y $\boldsymbol{D} = diag(d_1,..,d_n)$ es la matriz diagonal de sus valores singulares. Así, mediante el pre-blanqueo se obtiene una matriz de mezclas $\tilde{\boldsymbol{A}}$ que es ortogonal:

$$\tilde{\boldsymbol{x}} = \boldsymbol{R^{-1/2}} \cdot \boldsymbol{As} = \boldsymbol{ED^{-1/2}E^T} \cdot \boldsymbol{As} = \tilde{\boldsymbol{A}}\boldsymbol{s} \rightarrow E[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^T] = \tilde{\boldsymbol{A}}E[\boldsymbol{ss}^T]\tilde{\boldsymbol{A}}^T = \tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}^T = \boldsymbol{I} \tag{C.10}$$

Puede ser también recomendable en algunas aplicaciones reducir la dimensión de los datos a la misma vez que se aplica el blanqueo. Así, se buscan los valores singulares de la matriz de covarianza de menor valor y se descartan (al igual que se realiza en otras técnicas estadísticas como por ejemplo *PCA, "Principal Component Analysis"*).

Apéndice D:
Estimación de parámetros y selección del número de componentes en la estimación de PDFs mediante el modelo de mezcla de Gaussianas

Un modelo de mezcla Gaussiana (*GMM,* "*Gaussian Mixture Model"*) es una función de densidad de probabilidad paramétrica representada por una suma ponderada de componentes con densidades de probabilidad Gaussianas multivariantes. Para un conjunto de d variables aleatorias continuas $\boldsymbol{X} = [X_1, \dots, X_d]$ modelamos su función de densidad de probabilidad conjunta mediante un *GMM* con K componentes como:

$$f(x_1, \dots, x_d) \sim f_{GMM}(\boldsymbol{x}|\boldsymbol{\theta_1} \dots \boldsymbol{\theta_K}, \boldsymbol{\omega}) = \sum_{i=1}^{K} \omega_i \cdot g(\boldsymbol{x}|\boldsymbol{\theta_i}) = \sum_{i=1}^{K} \omega_i \cdot g(\boldsymbol{x}|\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}) \quad \text{(D.1)}$$

donde ω_i, corresponden a los pesos de la mezcla, los cuales cumplen con la restricción $\sum_{i=1}^{K} \omega_i = 1$, y $g(\boldsymbol{x}|\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}), i = 1, .., K$ son las componentes de densidad de probabilidad *d*-dimensional Gaussiana, con media $\boldsymbol{\mu_i}$ y matriz de covarianza $\boldsymbol{\Sigma_i}$:

$$g(\boldsymbol{x}|\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}) = \frac{1}{(2\pi)^{N/2} \cdot |\boldsymbol{\Sigma_i}|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu_i})^T \boldsymbol{\Sigma_i}^{-1} (\boldsymbol{x}-\boldsymbol{\mu_i})} \quad \text{(D.2)}$$

El modelo *GMM* es parametrizado completamente por el conjunto de parámetros $\vartheta = \{\omega_i, \boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}\},\ i = 1, \dots, K$. Existen diferentes técnicas para estimar los parámetros de un modelo *GMM*, pero el más consolidado y utilizado es el método de estimación de máxima verosimilitud (*MLE*, *"Maximum Likelihood Estimator"*). Dada una secuencia de *T* vectores de entrenamiento independientes entre sí $X_T = \{\boldsymbol{x_1}, \dots, \boldsymbol{x_T}\}$, con $\boldsymbol{x_j} = [x_1{}^j, \dots, x_d{}^j]$, el entrenamiento de *GMM* se basa en encontrar el conjunto de parámetros ϑ^* que maximiza la función de verosimilitud:

$$\vartheta^* = \underset{\vartheta}{\operatorname{argmax}} \left(L(\vartheta; X_T) = f(X_T|\vartheta) = \prod_{j=1}^{T} f(\boldsymbol{x_j}|\vartheta) \right) \quad \text{(D.3)}$$

Esta expresión es una función no lineal con respecto a los parámetros ϑ y una maximización directa no es posible. Existen diversos métodos para encontrar el *MLE*, tales como el algoritmo de descenso por gradiente, el método de gradiente conjugado o variaciones del método Gauss-Newton. De entre todos los métodos para maximizar esta función, el método más popular y usado por su rapidez, relativa sencillez y resultados es el algoritmo llamado "*Expectation Maximization*" (*EM*). A diferencia de estos métodos, el algoritmo *EM* no necesita la evaluación de la primera y segunda derivada de la función de verosimilitud.

Algoritmo "Expectation Maximization"

El algoritmo "*Expectation Maximization*" es un método iterativo para encontrar la estimación de máxima verosimilitud de una serie de parámetros en un modelo estadístico, donde el modelo depende de variables ocultas ("*latent variables"*). Cada iteración del algoritmo alterna entre dos pasos, el primero llamado "*Expectation"* (*E*), donde se calcula la esperanza de la verosimilitud usando la estimación en ese punto de los parámetros, y un segundo paso llamado "*Maximization*" (*M*), donde se recalcula la estimación de los parámetros maximizando la esperanza de la verosimilitud encontrada en el paso *E*.

Se introduce una serie de variables ocultas de tal forma que su conocimiento pueda simplificar la maximización de la verosimilitud $f(X_T|\vartheta)$. En el paso *E*, se estima la distribución de las variables ocultas dadas las observaciones y el valor que toman en ese paso los parámetros, y en el paso *M* se modifica el valor de los parámetros para maximizar la distribución conjunta de las observaciones y las variables ocultas.

Dado un modelo estadístico consistente en un conjunto X_T de observaciones, un conjunto de datos ocultos $Y_T = \{\boldsymbol{y_1}, \dots, \boldsymbol{y_T}\}$ (donde $\boldsymbol{y_j} = [{y_1}^j, \dots, {y_M}^j]$, $j = 1..T$ es una realización de las variables aleatorias ocultas $\boldsymbol{Y} = [Y_1, \dots, Y_M]$ en el instante j), y un conjunto de parámetros desconocidos λ, con una función de verosimilitud dada por $L(\vartheta; X_T, Y_T) = f(X_T, Y_T|\vartheta)$, el estimador de máxima verosimilitud del conjunto de parámetros desconocidos ϑ es determinado por la verosimilitud marginal de las observaciones:

$$L(\vartheta; X_T) = f(X_T|\vartheta) = \sum_{Y_T} f(X_T, Y_T|\vartheta) \tag{D.4}$$

El algoritmo *EM* trata de buscar el *MLE* de la verosimilitud marginal aplicando iterativamente los siguientes dos pasos:

- Paso 1: Expectation (*E*) - Se calcula la esperanza de la función de verosimilitud logarítmica, con respecto a la distribución de Y_t dado X_T bajo la estimación en esta iteración de los parámetros $\vartheta^{(t)}$:

$$Q\left(\vartheta\middle|\vartheta^{(t)}\right) = E_{Y_t|X_T,\lambda^{(t)}}\left[ln\left(L(\vartheta; X_T, Y_T)\right)\right] \tag{D.5}$$

- Paso 2: Maximization (*M*) - Recalculamos la estimación de los parámetros

$$\vartheta^{(t+1)} = \underset{\lambda}{arg\max}\, Q\left(\vartheta\middle|\vartheta^{(t)}\right) \tag{D.6}$$

Aplicación del Algoritmo EM en estimación de parámetros de modelo GMM

En la aplicación del algoritmo *EM* en la estimación de los parámetros de un modelo GMM, sólo encontramos una variable aleatoria oculta, $M = 1 \rightarrow \boldsymbol{Y} = [Y]$, que describe qué componente Gaussiana ha generado cada uno de los vectores de observación, por lo tanto Y será una v.a discreta que podrá tomar valores del conjunto $\{1,2, \dots, K\}$. La función de verosimilitud, será por tanto:

$$\begin{aligned} L(\vartheta; X_T) = f(X_T|\vartheta) &= \prod_{j=1}^{T} \sum_{y} f(\boldsymbol{x_j}, y^j|\vartheta) \\ L(\vartheta; X_T) = f(X_T|\vartheta) &= \prod_{j=1}^{T} \sum_{i=1}^{K} \omega_i \cdot g(\boldsymbol{x_j}|\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}) \end{aligned} \tag{D.7}$$

Aplicando el algoritmo *EM* se obtienen las siguientes expresiones para la actualización iterativa de los parámetros:

- La probabilidad a posteriori de la componente Gaussiana $i,\ i = 1, .., K$ para el instante j:

$$P(i|\boldsymbol{x_j}, \lambda) = \frac{\omega_i \cdot g(\boldsymbol{x_j}|\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i})}{\sum_{k=1}^{K} \omega_k \cdot g(\boldsymbol{x_j}|\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})} \tag{D.8}$$

- Los pesos de la mezcla, las medias y las matrices de covarianza

$$\begin{aligned} \widehat{\omega}_i = \frac{1}{T}\sum_{j=1}^{T} P(i|\boldsymbol{x_j}, \lambda) \qquad \widehat{\boldsymbol{\mu}}_{\boldsymbol{i}} &= \frac{\sum_{j=1}^{T} P(i|\boldsymbol{x_j}, \lambda)\boldsymbol{x_j}}{\sum_{j=1}^{T} P(i|\boldsymbol{x_j}, \lambda)} \\ \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{i}} &= \frac{\sum_{j=1}^{T} P(i|\boldsymbol{x_j}, \lambda)(\boldsymbol{x_j} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{i}})^T(\boldsymbol{x_j} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{i}})}{\sum_{j=1}^{T} P(i|\boldsymbol{x_j}, \lambda)} \end{aligned} \tag{D.9}$$

Se garantiza que el algoritmo converge a una solución local óptima, es decir, siempre convergerá hacia un máximo de la función de verosimilitud, aunque no se puede garantizar que sea el máximo global. Las características de convergencia del algoritmo, así como la solución obtenida dependen del valor de los parámetros usados en la inicialización del algoritmo.

Selección del número de componentes Gaussianas

Otro de los parámetros del modelo *GMM* es el número de componentes Gaussianas utilizadas (K). Para la estimación de este parámetro podemos entrenar distintos modelos *GMM* de la densidad de probabilidad con diferentes valores del número de componentes Gaussianas y nos apoyamos en un método de selección de modelo óptimo basado en la longitud de descripción mínima (*MDL*, *"Minimum*

Description Length"). Se incluye una revisión de estas técnicas el apartado 2.5.3, al utilizarse también la estimación mediante teoría de cópulas.

El número de parámetros a estimar para un modelo con K componentes Gaussianas y con N modalidades en los vectores de observaciones, con matrices de covarianzas completas es (fíjese en que, debido a que la suma de pesos está restringida a la unidad, sólo se deben optimizar $K-1$ de los pesos, ya que el último será fijado por dicha restricción):

$$n_{param} = \underbrace{K \cdot N \cdot \left(\frac{N-1}{2}\right)}_{Covarianzas} + \underbrace{N \cdot K}_{Medias} + \underbrace{(K-1)}_{Pesos} \tag{D.10}$$

Apéndice E:
Uso de cálculo simbólico para derivar la expresión de la copula de Frank

Derivar la función de cópula de Frank para obtener la función de densidad de cópula a través de la expresión 2.40 para un número de más de 4 variables es un trabajo tedioso, con expresiones muy farragosas. Hemos utilizado un programa de cálculo simbólico para ayudarnos a conseguir las densidades de cópula. La función E.1 es un ejemplo de cómo obtener la expresión de la densidad de cópula de Frank en el caso de 5 variables a través del software matemático MATLAB.

```
% 5 variables   % Expresión simbolica
    % Hacemos un cambio de variables
       syms w1 w2 w3 w4 w5 alpha
       a = (exp(-alpha)-1).^4 ;
       CDF = -(1/alpha).*log( 1 + w1*w2*w3*w4*w5/a ) ;
       PDF = diff(diff( diff( diff( diff( CDF,w1 ),w2 ),w3 ),w4 ),w5) ;
    %  Añadimos la parte que falta de derivar de la regla de cadena tras
elcambio de variables
PDF = (-alpha.*( w1 + 1 ))*(-alpha.*( w2 + 1 ))*(-alpha.*( w3 + 1 ))*…
               …(-alpha.*( w4 + 1 )) *(-alpha.*( w5 + 1 ))  *PDF  ;
    % La expresamos como handle de matlab
        pdf = matlabFunction(PDF,'file','frankpdf_5v');
```

Función E.1 – Obtención de la densidad de cópula de Frank para 5 variables mediante cálculo simbólico

La función E.2 es el resultado obtenido. Se puede ver como la expresión resultante es bastante compleja.

```
function PDF = frankpdf_5v(alpha,w1,w2,w3,w4,w5)
%    This function was generated by the Symbolic Math Toolbox version 5.6.
t29 = alpha.^2;                      t30 = t29.^2;
t31 = exp(-alpha);                   t32 = t31-1.0;
t33 = 1.0./t32.^4;                   t34 = 1.0./alpha;
t35 = t33.*w1.*w2.*w3.*w4.*w5;       t36 = t35+1.0;
t37 = w1.^2;                         t38 = w2.^2;
t39 = w3.^2;                         t40 = w4.^2;
t41 = w5.^2;

PDF =
alpha.*t30.*(w1+1.0).*(w2+1.0).*(w3+1.0).*(w4+1.0).*(w5+1.0).*((t33.*t34).
/t36+1.0./t32.^20.*t34.*1.0./t36.^5.*t37.^2.*t38.^2.*t39.^2.*t40.^2.*t41.^
2.*2.4e1+1.0./t32.^12.*t34.*1.0./t36.^3.*t37.*t38.*t39.*t40.*t41.*5.0e1-
1.0./t32.^8.*t34.*1.0./t36.^2.*w1.*w2.*w3.*w4.*w5.*1.5e1-
1.0./t32.^16.*t34.*1.0./t36.^4.*t37.*t38.*t39.*t40.*t41.*w1.*w2.*w3.*w4.*w
5.*6.0e1);
```

Función E.2 – Función de cálculo de la densidad de cópula de Frank para 5 variables

Apéndice F:
Entrenamiento de mezcla de expertos mediante Algoritmo EM

El algoritmo de esperanza-maximización (*EM,* "*Expectation-Maximization"*) [148] es un método iterativo usado en estadística para encontrar estimadores de máxima verosimilitud (*ML*, "*Maximum Likelihood"*) de parámetros en modelos probabilísticos que dependen de variables no observables u ocultas. Para la mezcla de expertos se introduce una serie de variables aleatorias $Z = \{\{z_k^n\}_{n=1}^N\}_{k=1}^K$ para poder entrenar el modelo mediante el algoritmo *EM*.

$$D = \{X, Y\} \quad X = \{\boldsymbol{x}^n\}_{n=1}^N \quad Y = \left\{y^n = \begin{Bmatrix} 1 \; si \; \boldsymbol{x}^n \epsilon \; H_1 \\ 0 \; si \; \boldsymbol{x}^n \epsilon \; H_0 \end{Bmatrix}_{n=1}^N \right.$$
$$D = \{D_1, D_0\}, \qquad D_j = \left\{\boldsymbol{x}^n \epsilon \; H_j\right\}_{n=1}^{N_j} \quad Z = \{Z_1, Z_0\} \tag{F.1}$$

La función de verosimilitud logarítmica completa usando estas variables aleatorias ocultas se expresa como:

$$l\left(\Theta_j; D_j; Z_j\right) = \sum_{n=1}^{N_j} \sum_{k=1}^{K} z_k^n \cdot \left[ln\left(g_k(\boldsymbol{x}^n, \boldsymbol{v})\right) + ln\left(P\left(H_j|k, \boldsymbol{x}^n, \boldsymbol{w}\right)\right)\right] \tag{F.2}$$

El algoritmo EM se emplea para promediar sobre z_k y maximizar el valor esperado de la verosimilitud logarítmica de las observaciones $E_Z\left[l\left(\Theta_j; D_j; Z_j\right)\right]$, lo que resulta en:

$$Q\left(\Theta_j, \Theta_j^{(p)}\right) = \sum_{n=1}^{N_j} \sum_{k=1}^{K} h_k^n \cdot \left[ln\left(g_k(\boldsymbol{x}^n, \boldsymbol{v})\right) + ln\left(P\left(H_j|k, \boldsymbol{x}^n, \boldsymbol{w}\right)\right)\right] = \sum_{k=1}^{K} Q_k^g + Q_k^e$$
$$h_k^n = E\left[z_k^n | D_j\right] \quad Q_k^g = \sum_{n=1}^{N_j} h_k^n \cdot ln\left(g_k(\boldsymbol{x}^n, \boldsymbol{v})\right) \quad Q_k^e = \sum_{n=1}^{N_j} h_k^n \cdot ln\left(P\left(H_j|k, \boldsymbol{x}^n, \boldsymbol{w}\right)\right) \tag{F.3}$$

donde el superíndice (p) indica la iteración en la que se encuentra el algoritmo.

El conjunto de parámetros Θ_j se estima iterativamente mediante dos pasos en cada iteración:

1. Paso *E*: Se calcula h_k^n, la esperanza de las variables aleatoria ocultas.
2. Paso *M*: Se busca una nueva estimación de los parámetros

$$\boldsymbol{v}_k^{(p+1)} = \underset{\boldsymbol{v}_k}{\operatorname{argmax}}\left(Q_k^g\right) \quad \boldsymbol{w}_k^{(p+1)} = \underset{\boldsymbol{w}_k}{\operatorname{argmax}}\left(Q_k^e\right) \tag{F.4}$$

Para obtener los argumentos $\text{argmax}_{\boldsymbol{v}_k}\left(Q_k^g\right)$ y $\text{argmax}_{\boldsymbol{w}_k}(Q_k^e)$, se puede resolver $\partial Q_k^g/\partial \boldsymbol{v}_k$ y/o $\partial Q_k^e/\partial \boldsymbol{w}_k$ en el caso que las expresiones de las puertas y de los expertos permitan resolverlo analíticamente.

Sin embargo, a veces es muy difícil obtener una solución analítica por este método, como ocurre, por ejemplo, con la función soft-max debido a su carácter no lineal. Para obtener los argumentos en estos casos se propuso la técnica iterativa de mínimos cuadrados recursivos (*IRLS*, del inglés '*Iterative Recursive Least Squares*') [90] para modelos lineales de puertas y expertos, y una versión extendida de este [91] algoritmo para expertos y puertas no lineales. Cuando se utilizan cualquiera de estas dos técnicas se le suele denominar como técnica *EM* de doble bucle (en inglés '*double-loop EM*'). Otra posibilidad es la de usar lo que se conoce como algoritmo *EM* generalizado (*GEM*, del inglés '*Generalized Expectation-Maximization*') [149]. Simplemente se trata de buscar una nueva estimación de los parámetros en cada iteración (p) que conduzca a:

$$Q_k^g\left(\boldsymbol{v}_k^{(p+1)}\right) \geq Q_k^g\left(\boldsymbol{v}_k^{(p)}\right) \quad Q_k^e\left(\boldsymbol{w}_k^{(p+1)}\right) \geq Q_k^e\left(\boldsymbol{w}_k^{(p)}\right) \tag{F.5}$$

- Algoritmo EM single-Loop: Nuevo modelo para la red de puertas

Para evitar la utilización del '*double-loop EM*' en [149] se propone un modelo alternativo para la mezcla de expertos, donde el modelo de puertas utilizado está basado en una mezcla de densidades Gaussianas:

$$g_k(\boldsymbol{x},\boldsymbol{v}) = \frac{\alpha_k \cdot f(\boldsymbol{x}|\boldsymbol{v}_k)}{\sum_i \alpha_i \cdot f(\boldsymbol{x}|\boldsymbol{v}_i)} = \frac{\alpha_k \cdot f(\boldsymbol{x}|\boldsymbol{v}_k)}{P(\boldsymbol{x},\boldsymbol{v})}\ , \qquad \sum_i \alpha_i = 1, \quad \alpha_k \geq 0 \tag{F.6}$$

donde $f(\boldsymbol{x}|\boldsymbol{v}_k)$ corresponde a la expresión de una densidad Gaussiana multivariante:

$$f(\boldsymbol{x}|\boldsymbol{v}_k) = \frac{1}{(2\pi)^{N/2} \cdot |\boldsymbol{\Sigma}_k|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m}_k)^T \cdot \boldsymbol{\Sigma}_k^{-1} \cdot (\boldsymbol{x}-\boldsymbol{m}_k)} \tag{F.7}$$

Con la nueva puerta propuesta, el modelo de mezcla de expertos queda como:

$$P\left(H_j|\boldsymbol{x},\Theta\right) = \sum_{k=1}^{K} \frac{\alpha_k \cdot f(\boldsymbol{x}|\boldsymbol{v}_k)}{P(\boldsymbol{x},\boldsymbol{v})} P\left(H_j|k,\boldsymbol{x},\Theta_e\right) \tag{F.8}$$

Si intentamos derivar el algoritmo *EM* sobre la estimación de máxima verosimilitud de los parámetros de $P\left(H_j|\boldsymbol{x},\Theta\right)$ volvemos a encontrarnos con el problema de que la maximización $\max_{\boldsymbol{v}_k}\left(Q_k^g\right)$ no puede ser resuelta de forma analítica. Para evitar este problema, se expresa (X) como:

$$P(H_j, \boldsymbol{x}) = P(H_j|\boldsymbol{x}, \Theta)P(\boldsymbol{x}, \boldsymbol{v}) = \sum_{k=1}^{K} \alpha_k \cdot f(\boldsymbol{x}|\boldsymbol{v}_k)\, P(H_j|k, \boldsymbol{x}, \Theta_e) \tag{F.9}$$

y ahora se define una nueva función de verosimilitud:

$$l' = \sum_n ln\left(P(H_j, \boldsymbol{x}^n)\right) \tag{F.10}$$

Utilizando ahora el algoritmo *EM* con la nueva función de verosimilitud obtenemos:

1.- Paso E:

$$h_k^{(p)}(H_j|\boldsymbol{x}^n) = \frac{{\alpha_k}^{(p)} f(\boldsymbol{x}^n|{\boldsymbol{v}_k}^{(p)}) \cdot P_k\left(H_j\middle|\boldsymbol{x}^n, \Theta_e^{(p)}\right)}{\sum_i {\alpha_i}^{(p)} f(\boldsymbol{x}^n|{\boldsymbol{v_i}}^{(p)}) \cdot P_i\left(H_j\middle|\boldsymbol{x}^n, \Theta_e^{(p)}\right)} \tag{F.11}$$

Las funciones objetivo son:

$$Q_k^e = \sum_{n=1}^{N_j} h_k^{(p)}(H_j|\boldsymbol{x}^n) \cdot ln\left(P_k\left(H_j|\boldsymbol{x}^n, \Theta_e^{(p)}\right)\right) \quad Q_k^g = \sum_{n=1}^{N_j} h_k^{(p)}(H_j|\boldsymbol{x}^n) \cdot ln\left(f(\boldsymbol{x}^n|{\boldsymbol{v}_k}^{(p)})\right)$$
$$Q^{\boldsymbol{\alpha}} = \sum_{n=1}^{N_j}\sum_{k=1}^{K} h_k^{(p)}(H_j|\boldsymbol{x}^n) \cdot ln(\alpha_k), \qquad \boldsymbol{\alpha} = [\alpha_1 \dots \alpha_K] \tag{F.12}$$

2.- Paso M:

Para encontrar una nueva estimación de los parámetros:

$$\Theta_e^{(p+1)} = \underset{\Theta_e}{argmax}\left(Q_k^e\left(\Theta_e^{(p)}\right)\right) \quad \boldsymbol{v}_k^{(p+1)} = \underset{\boldsymbol{v}_k}{argmax}\left(Q_k^g\left(\boldsymbol{v}_k^{(p)}\right)\right)$$
$$\boldsymbol{\alpha}^{(p+1)} = \underset{\boldsymbol{v}_k}{argmax}\left(Q_k^g\left(\boldsymbol{v}_k^{(p)}\right)\right) \tag{F.13}$$

Observamos como la maximización de la red de expertos no se modifica, mientras que ahora la maximización de la red de ponderación puede realizarse de forma analítica:

$${\alpha_k}^{(p+1)} = \frac{1}{N}\sum_t h_k^{(p)}(H_j|\boldsymbol{x}^n) \qquad {\boldsymbol{m}_k}^{(p+1)} = \frac{1}{\sum_n h_k^{(p)}(H_j|\boldsymbol{x}^n)}\sum_t h_k^{(p)}(H_j|\boldsymbol{x}^n) \cdot \boldsymbol{x}^n$$
$${\boldsymbol{\Sigma}_k}^{(p+1)} = \frac{1}{\sum_n h_k^{(p)}(H_j|\boldsymbol{x}^n)}\sum_t h_k^{(p)}(H_j|\boldsymbol{x}^n)\left[\boldsymbol{x}^n - {\boldsymbol{m}_k}^{(p)}\right]\left[\boldsymbol{x}^n - {\boldsymbol{m}_k}^{(p)}\right]^T \tag{F.14}$$

Apéndice G:
Tablas de resultados en la selección del modelo de estimación de *PDFs*

Criterio *BIC*							
	Modelo indep.	**Cópula GMM**	**Cópula Gaussiana**	**Cópula Student-T**	**Cópula Clayton**	**Cópula Frank**	**Cópula Gumbel**
Genuinos	**-524.47**	-539.78	-525.88	-528.64	-525.27	-524.56	-524.76
Impostores ($\times 10^5$)	-2.477	**-2.475**	- 2.476	-2.476	-2.477	-2.477	-2.477

Criterio: Máxima área bajo la curva ROC, AUC (%)							
H_0 / H_1	**Modelo indep.**	**Cópula GMM**	**Cópula Gaussiana**	**Cópula Student-T**	**Cópula Clayton**	**Cópula Frank**	**Cópula Gumbel**
Independencia	99,7227	99,7119	99,7140	99,7140	99,7221	99,7226	99,7224
Copula GMM	99,6680	99,6544	99,6603	99,6603	99,6675	99,6680	99,6672
Cop. Gaussiana	**99,7396**	99,7278	99,7378	99,7378	99,7394	99,7396	99,7392
Copula Student-T	99,7315	99,7196	99,7219	99,7219	99,7309	99,7315	99,7309
Copula Clayton	99,7223	99,7111	99,7137	99,7137	99,7214	99,7220	99,7213
Copula Frank	99,7228	99,7119	99,7139	99,7139	99,7220	99,7226	99,7223
Copula Gumbel	99,7226	99,7117	99,7137	99,7137	99,7218	99,7226	99,7223
Criterio: Máxima área bajo la curva ROC con estimador WMW, AUC (%)							
H_0 / H_1	**Modelo indep.**	**Cópula GMM**	**Cópula Gaussiana**	**Cópula Student-T**	**Cópula Clayton**	**Cópula Frank**	**Cópula Gumbel**
Independencia	99,7247	99,7123	99,7157	99,7157	99,7247	99,7247	99,7243
Copula GMM	99,6691	99,6548	99,6602	99,6602	99,6687	99,6690	99,6685
Cop. Gaussiana	**99,7419**	99,7305	99,7337	99,7337	99,7418	99,7418	99,7414
Copula Student-T	99,7309	99,7171	99,7210	99,7210	99,7309	99,7309	99,7306
Copula Clayton	99,7233	99,7106	99,7140	99,7140	99,7233	99,7232	99,7228
Copula Frank	99,7246	99,7122	99,7156	99,7156	99,7246	99,7246	99,7242
Copula Gumbel	99,7242	99,7117	99,7151	99,7150	99,7241	99,7241	99,7238
Criterio: Máxima área bajo la curva ROC limitada entre 0.00009 y 0.002 ($\times 10^{-3}$)							
H_0 / H_1	**Modelo indep.**	**Cópula GMM**	**Cópula Gaussiana**	**Cópula Student-T**	**Cópula Clayton**	**Cópula Frank**	**Cópula Gumbel**
Independencia	1,7573	1,7793	1,7570	1,7435	1,7628	1,7617	1,7411
Copula GMM	1,7310	1,7617	1,7636	1,7755	1,7519	1,7292	1,7455
Cop. Gaussiana	1,7689	**1,7870**	1,7448	1,7433	1,7779	1,7689	1,7497
Copula Student-T	1,7548	1,7788	1,7584	1,7570	1,7320	1,7564	1,7351
Copula Clayton	1,7481	1,7641	1,7462	1,7432	1,7705	1,7617	1,7562
Copula Frank	1,7573	1,7793	1,7569	1,7420	1,7508	1,7601	1,7291
Copula Gumbel	1,7590	1,7751	1,7540	1,7510	1,7644	1,7602	1,7198

Bibliografía

[1] R. D. Hippenstiel, *Detection Theory: Applications and Digital Signal Processing*, 1.ª ed. CRC Press, 2001.

[2] S. G. Iyengar, P. K. Varshney, y T. Damarla, «A Parametric Copula-Based Framework for Hypothesis Testing Using Heterogeneous Data», *IEEE Transactions on Signal Processing*, vol. 59, n.º 5, pp. 2308-2319, may 2011.

[3] R. B. Rao, O. Yakhnenko, y B. Krishnapuram, «KDD cup 2008 and the workshop on mining medical data», *SIGKDD Explor. Newsl.*, vol. 10, n.º 2, pp. 34–38, dic. 2008.

[4] A. K. Jain y A. Ross, «Multibiometric systems», *Commun. ACM*, vol. 47, n.º 1, pp. 34–40, ene. 2004.

[5] S. Kay, *Fundamentals of Statistical Signal Processing, Volume II: Detection Theory*, 1.ª ed. Prentice Hall, 1998.

[6] J. Neyman y E. S. Pearson, «On the Problem of the Most Efficient Tests of Statistical Hypotheses», *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 231, pp. 289-337, ene. 1933.

[7] D. L. Hall y J. Llinas, *Handbook of multisensor data fusion*. CRC Press, 2001.

[8] E. Blasch, I. Kadar, J. Salerno, M. M. Kokar, S. Das, G. M. Powell, D. D. Corkill, y E. H. Ruspini, «Issues and challenges of knowledge representation and reasoning methods in situation assessment (Level 2 Fusion)», *Proceedings of SPIE*, vol. 6235, n.º 1, pp. 623510-623510-14, may 2006.

[9] I. Bosch Roig, «Algoritmos de detección distribuida en sistemas monosensor». [En línea]. Disponible en: http://riunet.upv.es/handle/10251/1898. [Accedido: 04-abr-2011].

[10] E. Drakopoulos y C. C. Lee, «Optimum fusion of correlated local decisions», en *Decision and Control, 1988., Proceedings of the 27th IEEE Conference on*, 1988, pp. 2489-2494 vol.3.

[11] P. Atrey, M. Hossain, A. El Saddik, y M. Kankanhalli, «Multimodal fusion for multimedia analysis: a survey», *Multimedia Systems*, abr. 2010.

[12] S. E. Yuksel, J. N. Wilson, y P. D. Gader, «Twenty Years of Mixture of Experts», *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, n.º 8, pp. 1177-1193, 2012.

[13] W. H. Adams, G. Iyengar, C.-Y. Lin, M. R. Naphade, C. Neti, H. J. Nock, y J. R. Smith, «Semantic Indexing of Multimedia Content Using Visual, Audio, and Text Cues», *EURASIP Journal on Advances in Signal Processing*, vol. 2003, n.º 2, pp. 170-185, 2003.

[14] H. Sridharan, H. Sundaram, y T. Rikakis, «Computational models for experiences in the arts, and multimedia», en *Proceedings of the 2003 ACM SIGMM workshop on Experiential telepresence*, New York, NY, USA, 2003, pp. 31–44.

[15] A. Soriano, L. Vergara, G. Safont, y A. Salazar, «On comparing hard and soft fusion of dependent detectors», en *2012 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2012, pp. 1-6.

[16] Z. Zhu y T. S. Huang, Eds., *Multimodal Surveillance: Sensors, Algorithms, and Systems*. Artech House Publishers, 2007.

[17] J. Kittler, M. Hatef, R. P. . Duin, y J. Matas, «On combining classifiers», *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, n.º 3, pp. 226-239, mar. 1998.

[18] A. Subramanian, A. Sundaresan, y P. K. Varshney, «Fusion for the detection of dependent signals using multivariate copulas», en *2011 Proceedings of the 14th International Conference on Information Fusion (FUSION)*, 2011, pp. 1-8.

[19] J. Moragues, L. Vergara, J. Gosálbez, y I. Bosch, «An extended energy detector for non-Gaussian and non-independent noise», *Signal Processing*, vol. 89, n.º 4, pp. 656-661, abr. 2009.

[20] J. Moragues, L. Vergara, y J. Gosalbez, «Generalized Matched Subspace Filter for Nonindependent Noise Based on ICA», *IEEE Transactions on Signal Processing*, vol. 59, n.º 7, pp. 3430 -3434, jul. 2011.

[21] G. Safont, A. Salazar, A. Soriano, y L. Vergara, «Combination of multiple detectors for EEG based biometric identification/authentication», en *2012 IEEE International Carnahan Conference on Security Technology (ICCST)*, 2012, pp. 230-236.

[22] A. Jain, K. Nandakumar, y A. Ross, «Score normalization in multimodal biometric systems», *Pattern Recognition*, vol. 38, n.º 12, pp. 2270-2285, dic. 2005.

[23] G. L. Marcialis y F. Roli, «Score-level fusion of fingerprint and face matchers for personal verification under "stress" conditions», 2007. .

[24] M. He, S.-J. Horng, P. Fan, R.-S. Run, R.-J. Chen, J.-L. Lai, M. K. Khan, y K. O. Sentosa, «Performance evaluation of score level fusion in multimodal biometric systems», *Pattern Recognition*, vol. 43, n.º 5, pp. 1789-1800, may 2010.

[25]K.-A. Toh, J. Kim, y S. Lee, «Maximizing area under ROC curve for biometric scores fusion», *Pattern Recognition*, vol. 41, n.º 11, pp. 3373-3392, nov. 2008.

[26]C. K. K. Divyakant T Meva, «Comparative Study of Different Fusion Techniques in Multimodal Biometric Authentication», 2013.

[27]D. W. Scott y S. R. Sain, «Multidimensional Density Estimation», en *Handbook of Statistics*, vol. Volume 24, E. J. W. and J. L. S. C.R. Rao, Ed. Elsevier, 2005, pp. 229-261.

[28]B. W. Silverman y M. C. Jones, «E. Fix and J.L. Hodges (1951): An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation: Commentary on Fix and Hodges (1951)», *International Statistical Review / Revue Internationale de Statistique*, vol. 57, n.º 3, p. 233, dic. 1989.

[29]E. Parzen, «On Estimation of a Probability Density Function and Mode», *The Annals of Mathematical Statistics*, vol. 33, n.º 3, pp. 1065-1076, sep. 1962.

[30]M. Kristan, A. Leonardis, y D. Skočaj, «Multivariate online kernel density estimation with Gaussian kernels», *Pattern Recognition*, vol. 44, n.º 10-11, pp. 2630-2642, oct. 2011.

[31]«Multivariate kernel density estimation», *Wikipedia, the free encyclopedia*. 29-mar-2013.

[32]*Density estimation for statistics and data analysis*. Chapman and Hall, 1986.

[33]P. Hall, S. J. Sheather, M. C. Jones, y J. S. Marron, «On optimal data-based bandwidth selection in kernel density estimation», *Biometrika*, vol. 78, n.º 2, pp. 263-269, ene. 1991.

[34]Y. Hamamoto, Y. Fujimoto, y S. Tomita, «On the estimation of a covariance matrix in designing Parzen classifiers», *Pattern Recognition*, vol. 29, n.º 10, pp. 1751-1759, oct. 1996.

[35]M. P. Wand y M. C. Jones, *Kernel Smoothing*. Chapman & Hall/CRC, 1994.

[36]J. M. Leiva-Murillo y A. Artés-Rodríguez, «Algorithms for maximum-likelihood bandwidth selection in kernel density estimators», *Pattern Recognition Letters*, vol. 33, n.º 13, pp. 1717-1724, oct. 2012.

[37]J. Cwik, J. Koronacki, y H. R, *A Combined Adaptive-Mixtures/Plug-In Estimator of Multivariate Probability Densities*. 1996.

[38]P. Vincent y Y. Bengio, «Manifold Parzen windows», en *Advances in Neural Information Processing Systems 15*, 2003, pp. 825–832.

[39]M. Girolami y C. He, «Probability density estimation from optimally condensed data samples», *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, n.º 10, pp. 1253-1264, 2003.

[40]Z. Deng, F.-L. Chung, y S. Wang, «FRSDE: Fast reduced set density estimator using minimal enclosing ball approximation», *Pattern Recognition*, vol. 41, n.º 4, pp. 1363-1372, abr. 2008.

[41]J. Goldberger y S. Roweis, «Hierarchical clustering of a mixture model», en *In NIPS*, 2005, pp. 505–512.

[42]E. López-Rubio y J. M. Ortiz-de-Lazcano-Lobato, «Soft clustering for nonparametric probability density function estimation», *Pattern Recognition Letters*, vol. 29, n.º 16, pp. 2085-2091, dic. 2008.
[43]M. A. T. Figueiredo y A. K. Jain, «Unsupervised learning of finite mixture models», *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, n.º 3, pp. 381-396, 2002.
[44]A. Sklar, «Fonctions de répartition à n dimensions et leurs marges», *Publ. Inst. Statist. Univ. Paris*, vol. 8, pp. 229–231, 1959.
[45]D. D. Mari y S. Kotz, *Correlation and Dependence*. World Scientific, 2001.
[46]Throsten Schimidt, «Coping with Copulas. Forthcoming in Risk Books "Copulas - From Theory to Applications in Finance"». .
[47]J. Rank, *Copulas: From theory to application in finance*, 1.ª ed. Risk Books, 2006.
[48]R. T. Clemen y T. Reilly, «Correlations and Copulas for Decision and Risk Analysis», *Management Science*, vol. 45, n.º 2, pp. 208-224, ene. 1999.
[49]U. Cherubini, E. Luciano, y W. Vecchiato, *Copula Methods in Finance*, 1.ª ed. Wiley, 2004.
[50]A. Sundaresan, P. K. Varshney, y N. S. . Rao, «Copula-Based Fusion of Correlated Decisions», *IEEE Transactions on Aerospace and Electronic Systems*, vol. 47, n.º 1, pp. 454-471, ene. 2011.
[51]R. B. Nelsen, *An Introduction to Copulas*. Springer, 2006.
[52]E. Bouyé, V. Durrleman, A. Nikeghbali, G. Riboulet, y T. Roncalli, «Copulas for Finance - A Reading Guide and Some Applications», *SSRN eLibrary*, mar. 2000.
[53]R. Davidson y J. G. MacKinnon, *Estimation and Inference in Econometrics*. Oxford University Press, 1993.
[54]H. Joe y T. Hu, «Multivariate Distributions from Mixtures of Max-Infinitely Divisible Distributions», *Journal of Multivariate Analysis*, vol. 57, n.º 2, pp. 240-265, 1996.
[55]M. H. Hansen y B. Yu, «Model Selection and the Principle of Minimum Description Length», *Journal of the American Statistical Association*, vol. 96, pp. 746–774, 1998.
[56]A. Tewari, M. J. Giering, y A. Raghunathan, «Parametric Characterization of Multimodal Distributions with Non-gaussian Modes», en *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, 2011, pp. 286 -292.
[57]Roger Nelsen, «Dependence Modeling with Archimedean Copulas». Department of Mathematical Sciences. Lewis and Clark College.
[58]C. H. Kimberling, «A probabilistic interpretation of complete monotonicity», *Aequationes Mathematicae*, vol. 10, n.º 2, pp. 152-164, 1974.
[59]D. Clayton y J. Cuzick, «Multivariate Generalizations of the Proportional Hazards Model», *Journal of the Royal Statistical Society. Series A (General)*, vol. 148, n.º 2, pp. 82-117, ene. 1985.
[60]R. D. Cook y M. E. Johnson, «A Family of Distributions for Modelling Non-Elliptically Symmetric Multivariate Data», *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 43, n.º 2, pp. 210-218, ene. 1981.
[61]Etienne Cuvelier, Monique Noirhomme-Fraiture, «Décomposition de mélange avec la copule de Clayton», *Société Française de Statistique*, 2003.
[62]M. Frank, «On the simultaneous associativity of "F(x,y)" and "x+y-F(x,y)"», *Aequationes Mathematicae*, vol. 19, n.º 1, pp. 194-226, 1979.
[63]C. Genest, «Frank's Family of Bivariate Distributions», *Biometrika*, vol. 74, n.º 3, pp. 549-555, 1987.
[64]E. J. Gumbel, «Bivariate Exponential Distributions», *Journal of the American Statistical Association*, vol. 55, n.º 292, pp. 698-707, dic. 1960.
[65]P. Hougaard, «A Class of Multivanate Failure Time Distributions», *Biometrika*, vol. 73, n.º 3, pp. 671-678, ene. 1986.
[66]Morgenstern, «Einfache Beispiele zweidimensionaler Verteilungen.», *Mitteislingsblatt fur Mathematische Statistik*, vol. 8, pp. 234-235, 1956.

[67]D. J. G. Farlie, «The Performance of Some Correlation Coefficients for a General Bivariate Distribution», *Biometrika*, vol. 47, n.º 3/4, pp. 307-323, dic. 1960.
[68]H. Eyraud, «Les principes de la mesure des correlations», *Ann. Univ. Lyon, Sect.*, vol. A1, pp. 30-47, 1936.
[69]Y. K. Leong y E. A. Valdez, «Claims Prediction with Dependence using Copula Models», *Group*, 2005.
[70]T. Bedford y R. Cooke, «Probability Density Decomposition for Conditionally Dependent Random Variables Modeled by Vines», *Annals of Mathematics and Artificial Intelligence*, vol. 32, n.º 1, pp. 245-268, 2001.
[71]E. C. Brechmann y U. Schepsmeier, «Modeling Dependence with C- and D-Vine Copulas: The R Package CDVine», *Journal of Statistical Software*, vol. 52, n.º 3, pp. 1–27, 2013.
[72]H. Joe, *Multivariate Models and Multivariate Dependence Concepts*, 1.ª ed. Chapman and Hall/CRC, 1997.
[73]T. Bedford y R. M. Cooke, «Vines: A New Graphical Model for Dependent Random Variables», *The Annals of Statistics*, vol. 30, n.º 4, pp. 1031-1068, 2002.
[74]L. Xu, A. Krzyzak, y C. Y. Suen, «Methods of combining multiple classifiers and their applications to handwriting recognition», *IEEE Transactions on Systems, Man and Cybernetics*, vol. 22, n.º 3, pp. 418-435, 1992.
[75]J. Kittler y S. A. Hojjatoleslami, «A weighted combination of classifiers employing shared and distinct representations», en *1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1998. Proceedings*, 1998, pp. 924-929.
[76]Xu, L., & Amari, S. (2009), «Combining Classifiers and Learning Mixture-of-Experts. In J. Rabuñal Dopico, J. Dorado, & A. Pazos (Eds.), Encyclopedia of Artificial Intelligence (pp. 318-326). Hershey, PA: Information Science Reference. doi:10.4018/978-1-59904-849-9.ch049», .
[77]S. Hashem y B. Schmeiser, «Improving Model Accuracy using Optimal Linear Combinations of Trained Neural Networks», *IEEE Transactions on Neural Networks*, vol. 6, pp. 792–794, 1992.
[78]K.-A. Toh, Q.-L. Tran, y D. Srinivasan, «Benchmarking a reduced multivariate polynomial pattern classifier», *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, n.º 6, pp. 740-755, 2004.
[79]S. Amari, «Integration of stochastic models by minimizing alpha-divergence», *Neural Comput*, vol. 19, n.º 10, pp. 2780-2796, oct. 2007.
[80]S. Amari, «Differential Geometry of Statistical Models», en *Differential-Geometrical Methods in Statistics*, Springer New York, 1985, pp. 11-65.
[81]H. Choi, S. Choi, A. Katake, y Y. Choe, «Learning alpha-integration with partially-labeled data», en *2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2010, pp. 2058-2061.
[82]H. Narasimhan y S. Agarwal, «A Structural {SVM} Based Approach for Optimizing Partial AUC», presentado en Proceedings of the 30th International Conference on Machine Learning (ICML-13), 2013, pp. 516-524.
[83]L. E. Dodd y M. S. Pepe, «Partial AUC estimation and regression», *Biometrics*, vol. 59, n.º 3, pp. 614-623, sep. 2003.
[84]A. Herschtal y B. Raskutti, «Optimising area under the ROC curve using gradient descent», en *Proceedings of the twenty-first international conference on Machine learning*, New York, NY, USA, 2004, p. 49–.
[85]K.-A. Toh, «Learning from Target Knowledge Approximation», en *2006 1ST IEEE Conference on Industrial Electronics and Applications*, 2006, pp. 1-8.
[86]R. H. Byrd, M. E. H. Y, y J. N. Z, «An interior point algorithm for large scale nonlinear programming», *SIAM Journal on Optimization*, 1999.

[87]R. A. Waltz, J. L. Morales, J. Nocedal, y D. Orban, «An interior algorithm for nonlinear optimization that combines line search and trust region steps», *Math. Program.*, vol. 107, n.º 3, pp. 391-408, jul. 2006.

[88]S. E. Yuksel, J. N. Wilson, y P. D. Gader, «Twenty Years of Mixture of Experts», *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, n.º 8, pp. 1177-1193, 2012.

[89]R. A. Jacobs, M. I. Jordan, S. J. Nowlan, y G. E. Hinton, «Adaptive mixtures of local experts», *Neural Comput.*, vol. 3, n.º 1, pp. 79–87, mar. 1991.

[90]M. I. Jordan, «Hierarchical mixtures of experts and the EM algorithm», *Neural Computation*, vol. 6, pp. 181–214, 1994.

[91]M. I. Jordan y L. Xu, *Convergence results for the EM Approach to Mixtures of Experts Architectures*. 1995.

[92]Z. Chair y P. K. Varshney, «Optimal Data Fusion in Multiple Sensor Detection Systems», *Aerospace and Electronic Systems, IEEE Transactions on DOI - 10.1109/TAES.1986.310699*, vol. AES-22, n.º 1, pp. 98-101, 1986.

[93]E. Drakopoulos y C. C. Lee, «Optimum multisensor fusion of correlated local decisions», *IEEE Transactions on Aerospace and Electronic Systems*, vol. 27, n.º 4, pp. 593-606, 1991.

[94]M. Kam, Q. Zhu, y W. S. Gray, «Optimal data fusion of correlated local decisions in multiple sensor detection systems», *Aerospace and Electronic Systems, IEEE Transactions on DOI - 10.1109/7.256317*, vol. 28, n.º 3, pp. 916-920, 1992.

[95]R. M. Losee y Jr, «Term Dependence: Truncating the Bahadur Lazarsfeld Expansion», en *Information Processing and Management*, 1994, pp. 293–303.

[96]P. M. Lewis II, «Approximating probability distributions to reduce storage requirements», *Information and Control*, vol. 2, n.º 3, pp. 214-225, sep. 1959.

[97]D. T. Brown, «A note on approximations to discrete probability distributions», *Information and Control*, vol. 2, n.º 4, pp. 386-392, dic. 1959.

[98]C. Chow y C. Liu, «Approximating discrete probability distributions with dependence trees», *Information Theory, IEEE Transactions on*, vol. 14, n.º 3, pp. 462-467, may 1968.

[99]H.-J. Kang, K. Kim, y J. H. Kim, «Approximating optimally discrete probability distribution with kth-order dependency for combining multiple decisions», *Information Processing Letters*, vol. 62, n.º 2, pp. 67-75, abr. 1997.

[100]K. Huang, I. King, y M. R. Lyu, «Constructing a large node Chow-Liu tree based on frequent itemsets», en *Proceedings of the 9th International Conference on Neural Information Processing, 2002. ICONIP '02*, 2002, vol. 1, pp. 498-502 vol.1.

[101]E. Kovács y T. Szántai, «On the Approximation of a Discrete Multivariate Probability Distribution Using the New Concept of t-Cherry Junction Tree», en *Coping with Uncertainty*, K. Marti, Y. Ermoliev, y M. Makowski, Eds. Springer Berlin Heidelberg, 2010, pp. 39-56.

[102] R. Viswanathan y V. Aalo, «On counting rules in distributed detection», *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, n.º 5, pp. 772-775, 1989.

[103]L. Lam y S. Y. Suen, «Application of majority voting to pattern recognition: an analysis of its behavior and performance», *Trans. Sys. Man Cyber. Part A*, vol. 27, n.º 5, pp. 553–568, sep. 1997.

[104]L. Vergara, «On the equivalence between likelihood ratio tests and counting rules in distributed detection with correlated sensors», *Signal Processing*, vol. 87, n.º 7, pp. 1808-1815, jul. 2007.

[105]V. Aalo y R. Viswanathou, «On distributed detection with correlated sensors: two examples», *IEEE Transactions on Aerospace and Electronic Systems*, vol. 25, n.º 3, pp. 414-421, may 1989.

[106]A. Soriano Tolosa, «Fusión de decisiones en problemas de detección con dos detectores arbitrarios no independientes», *Vergara Domínguez, L. dir. 40 p. http://riunet.upv.es/handle/10251/30950*, jul. 2013.

[107]A. K. Jain, A. Ross, y S. Prabhakar, «An introduction to biometric recognition», *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, n.º 1, pp. 4-20, 2004.

[108]A. K. Jain y A. Ross, *Multibiometric Systems*. 2004.

[109]C. Sanderson y K. Paliwal, *Information Fusion and Person Verification Using Speech & Face Information*. 2002.

[110]K. Nandakumar, Y. Chen, S. C. Dass, y A. K. Jain, «Likelihood Ratio-Based Biometric Score Fusion», *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, n.º 2, pp. 342-347, 2008.

[111]S. Ribaric y I. Fratric, «Experimental Evaluation of Matching-Score Normalization Techniques on Different Multimodal Biometric Systems», en *Electrotechnical Conference, 2006. MELECON 2006. IEEE Mediterranean*, 2006, pp. 498-501.

[112]Y. Wang, T. Tan, y A. K. Jain, «Combining face and iris biometrics for identity verification», en *Proceedings of the 4th international conference on Audio- and video-based biometric person authentication*, Berlin, Heidelberg, 2003, pp. 805–813.

[113]P. Verlinde y G. Chollet, *Comparing decision fusion paradigms using k-NN based classifiers, decision trees and logistic regression in a multi-modal identity verification application*. 1999.

[114]V. Chatzis, A. G. Bors, y I. Pitas, «Multimodal decision-level fusion for person authentication», *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, vol. 29, n.º 6, pp. 674-680, 1999.

[115]A. Ross y A. Jain, «Information fusion in biometrics», *Pattern Recognition Letters*, vol. 24, pp. 2115–2125, 2003.

[116]R. Mazouni y A. Rahmoun, «On Comparing Verification Performances of Multimodal Biometrics Fusion Techniques», 2011.

[117]S. Prabhakar y A. K. Jain, «Decision-level Fusion in Fingerprint Verification», *PATTERN RECOGNITION*, vol. 35, pp. 861–874, 2001.

[118]S. C. Dass, K. N, y A. K. Jain, «A principled approach to score level fusion in multimodal biometric systems», 2005, pp. 1049–1058.

[119]S. G. Iyengar, P. K. Varshney, y T. Damarla, «Biometric Authentication: A Copula-Based Approach», en *Multibiometrics for Human Identification*, Cambridge University Press, 2011.

[120]N. Poh, T. Bourlai, y J. Kittler, «A multimodal biometric test bed for quality-dependent, cost-sensitive and client-specific score-level fusion algorithms», *Pattern Recogn.*, vol. 43, n.º 3, pp. 1094–1105, mar. 2010.

[121]N. US Department of Commerce, «Biometric Scores Set». [En línea]. Disponible en: http://www.nist.gov/itl/iad/ig/biometricscores.cfm. [Accedido: 24-jun-2013].

[122]A. W. Bowman y A. Azzalini, *Applied Smoothing Techniques for Data Analysis : The Kernel Approach with S-Plus Illustrations: The Kernel Approach with S-Plus Illustrations*. Oxford University Press, 1997.

[123]J. Moragues Escrivá, «New energy detector extensions with application in sound based surveillance systems», *Riunet*, 13-sep-2011. [En línea]. Disponible en: http://www.tdx.cat/handle/10803/36356. [Accedido: 27-feb-2013].

[124]L. Scharf, *Statistical Signal Processing*, 1.ª ed. Prentice Hall, 1990.

[125]M. Markou y S. Singh, «Novelty Detection: A Review - Part 1: Statistical Approaches», *Signal Processing*, vol. 83, p. 2003, 2003.

[126]V. I. Kostylev, «Energy detection of a signal with random amplitude», en *IEEE International Conference on Communications, 2002. ICC 2002*, 2002, vol. 3, pp. 1606 - 1610 vol.3.

[127]H. Urkowitz, «Energy detection of unknown deterministic signals», *Proceedings of the IEEE*, vol. 55, n.º 4, pp. 523 - 531, abr. 1967.

[128]R. D. Hippenstiel, *Detection Theory: Applications and Digital Signal Processing*. CRC Press, 2010.

[129]L. L. Scharf y B. Friedlander, «Matched subspace detectors», *IEEE Transactions on Signal Processing*, vol. 42, n.º 8, pp. 2146-2157, Aug.

[130]S. A. Kassam, *Signal Detection in Non-Gaussian Noise*, Softcover reprint of the original 1st ed. 1988. Springer, 2011.

[131]P. M. Schultheiss y L. C. Godara, «Detection of weak stochastic signals in non-Gaussian noise: a general result», en *, 1994 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1994. ICASSP-94*, 1994, vol. iv, p. IV/361 -IV/364 vol.4.

[132]T.-W. Lee, *Independent Component Analysis - Theory and Applications*, 1st ed. Springer, 1998.

[133]A. Hyvärinen y E. Oja, «Independent component analysis: algorithms and applications», *Neural Networks*, vol. 13, pp. 411–430, 2000.

[134]O. Mazhelis, «One-class classifiers : a review and analysis of suitability in the context of mobile-masquerader detection», *South African Computer Journal*, vol. 36, pp. 29-48, 2006.

[135]D. W. Scott y S. R. Sain, «Multidimensional Density Estimation», en *Handbook of Statistics*, vol. Volume 24, E. J. W. and J. L. S. C.R. Rao, Ed. Elsevier, 2005, pp. 229-261.

[136]J. Moragues, A. Serrano, L. Vergara, y J. Gosalbez, «Acoustic detection and classification using temporal and frequency multiple energy detector features», en *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 1940 -1943.

[137]X. Wu, H. Gong, P. Chen, Z. Zhong, y Y. Xu, «Surveillance Robot Utilizing Video and Audio Information», *J Intell Robot Syst*, vol. 55, n.º 4-5, pp. 403-421, ago. 2009.

[138]S. Ntalampiras, I. Potamitis, y N. Fakotakis, «An Adaptive Framework for Acoustic Monitoring of Potential Hazards», *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2009, n.º 1, p. 594103, oct. 2009.

[139]T. Asfour, K. Regenstein, P. Azad, J. Schroder, A. Bierbaum, N. Vahrenkamp, y R. Dillmann, «ARMAR-III: An Integrated Humanoid Platform for Sensory-Motor Control», en *2006 6th IEEE-RAS International Conference on Humanoid Robots*, Dec., pp. 169-175.

[140]«ROBUST IMPULSIVE SOUND SOURCE LOCALIZATION BY MEANS OF AN ENERGY DETECTOR FOR TEMPORAL ALIGNMENT AND».

[141]M. S. Keshner, «1/f noise», *Proceedings of the IEEE*, vol. 70, n.º 3, pp. 212-218, 1982.

[142]L. Zão y R. Coelho, «Generation of coloured acoustic noise samples with non-Gaussian distributions», *IET Signal Processing*, vol. 6, n.º 7, pp. 684-688, 2012.

[143]A. Salazar, L. Vergara, A. Serrano, y J. Igual, «A general procedure for learning mixtures of independent component analyzers», *Pattern Recognition*, vol. 43, n.º 1, pp. 69-85, ene. 2010.

[144]P. J. Huber y E. M. Ronchetti, *Robust Statistics*, 2.ª ed. Wiley, 2009.

[145]B. Zadrozny y C. Elkan, «Transforming classifier scores into accurate multiclass probability estimates», en *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, 2002, pp. 694–699.

[146]M. Ayer, H. D. Brunk, G. M. Ewing, W. T. Reid, y E. Silverman, «An Empirical Distribution Function for Sampling with Incomplete Information», *The Annals of Mathematical Statistics*, vol. 26, n.º 4, pp. 641-647, dic. 1955.

[147]A. Hyvärinen y E. Oja, «Independent component analysis: algorithms and applications», *Neural Networks*, vol. 13, pp. 411–430, 2000.

[148]A. P. Dempster, N. M. Laird, y D. B. Rubin, «Maximum likelihood from incomplete data via the EM algorithm», *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, vol. 39, n.º 1, pp. 1–38, 1977.

[149]L. Xu, M. I. Jordan, y G. E. Hinton, *An Alternative Model for Mixtures of Experts*. 1995.