UNIVERSIDAD
DE SEVILLA

UNIVERSIDAD DE SEVILLA
ESCUELA SUPERIOR DE INGENIEROS
Dpto. de INGENIERÍA ELECTRÓNICA
Área de TEORÍA DE LA SEÑAL Y COMUNICACIONES

TESIS DOCTORAL

ECUACIONES LINEALES PARA EL PROBLEMA DE LA SEPARACIÓN CIEGA DE FUENTES

AUTOR: RUBÉN MARTÍN CLEMENTE
DIRECTOR: Prof. Dr. JOSÉ I. ACHA CATALINA
Sevilla, 2000

A mis padres.

A mi padre,

en su memoria.

Resumen

Esta Tesis Doctoral es un trabajo sobre ***separación ciega de fuentes***. La Teoría de la Señal estudia generalmente problemas en los que se considera que un único *emisor*, conocido, ha generado todo el flujo de información útil. Ahora bien, con frecuencia nos encontramos en la práctica con señales que están compuestas, en realidad, de la superposición de muchas otras, bien diferenciadas, y a las que en esta Tesis llamaremos ***fuentes***. Resulta evidente que se obtendría una gran ganancia de ser capaces de separar las distintas componentes (*fuentes*) de la señal, para su uso y estudio posterior.

Tradicionalmente, algunas variantes de este problema -típico en las Telecomunicaciones- han sido objeto de una atención considerable: así, entre otros, se puede considerar *separación de fuentes* al problema de estimar la señal de interés en una comunicación ruidosa. Incluso puede que la mezcla de señales reporte beneficios -y haya sido provocada a propósito-, caso que se da en las multicanalizaciones de frecuencia.

Ahora bien, cuando lo desconocemos todo, tanto sobre la naturaleza de las señales *fuente* como de la manera en la que se han unido, ya hablamos propiamente de un problema de *separación ciega*.

Así, entre otros, separar los discursos de personas que hablan simultáneamente o distinguir entre comunicaciones transmitidas por el mismo canal, son problemas de *separación ciega de fuentes*. También se utilizan estas técnicas para descomponer señales y poner de manifiesto su estructura interna; como, por ejemplo, se hace en la actualidad con el electroencefalograma, entre otras señales de naturaleza biomédica. Cualquier estudio sobre la materia es también de aplicación inmediata en, por ejemplo, la reducción de la diafonía

-*crosstalk*- entre comunicaciones transmitidas por líneas próximas. La medida de la contaminación acústica en las ciudades, separando los ruidos y clasificándolos según su origen, también se beneficiaría de lo aprendido investigando este problema. En general, la lista de aplicaciones se llega a hacer atractivamente inacabable.

En esta Tesis estudiaremos el problema de las mezclas de señales que son *lineales e invariantes en el tiempo*. Probaremos que la Separación es posible bajo las hipótesis de que hay tantos sensores como señales y que las fuentes son *estadísticamente independientes* unas de otras, siendo, a lo más, una de distribución gaussiana.

La Tesis se desarrollará en dos partes: en primer lugar, presentaremos las bases matemáticas que sustentan la separación ciega señales y los algoritmos que, hoy por hoy, se han desarrollado alrededor de la hipótesis de independencia estadística entre las fuentes. Básicamente, todos utilizan estimaciones de máxima verosimilitud de los parámetros de la mezcla y, en este sentido, son *eficientes* aunque sólo bajo hipótesis restrictivas; en particular, que la distribución estadística de las fuentes es conocida de antemano o bien estimada junto con las señales, al coste de incrementar el esfuerzo computacional. Cuando estas hipótesis no se satisfacen, es dudoso que se obtenga la verdadera separación de las fuentes.

Alternativamente, en esta Tesis presentaremos un conjunto de condiciones *necesarias* y *suficientes* para obtener la separación y que llevan a ecuaciones *lineales* para encontrar los parámetros de la mezcla y a un algoritmo (SEVILLA) para resolverlas de forma eficiente. Las simulaciones muestran que la exactitud de las soluciones es comparable a las obtenidas por el estimador de máxima verosimilitud. Más interesante resulta el que la carga computacional de SEVILLA sea varias veces menor que la de los algoritmos más conocidos.

Summary

This Doctoral Thesis is a study on **blind separation of sources**. The Signal Theory usually investigates those problems in which a well-known, single transmitter is considered to have generated all the flow of useful information. However, in practice, studies are frequently carried out with signals that are actually made up of the addition of many others, each of which with its own entity and which in this Thesis we shall call *sources*. In this case, it is evident the gain that would be obtained on separating the different signal components for their later use and study.

Traditionally, variants of this problem, typical in Telecommunications, have been the object of considerable attention: thus, the elimination of noise can be considered a problem of source separation. The mixture of signals can even afford benefits as in the case of multiplexations of frequency.

Now, when we ignore both the nature of the source signals and the way in which they have been mixed, then we can speak of blind source separation.

This problem arises in a great many situations. In the case of microphones that register the voices of several people at the same time and also, in one of the most interesting situations, when an aerial simultaneously receives signals in the same frequency band. This problem also appears when a target signal or an image have to be decomposed into others that allow us to better understand their structure: for example, these techniques are applied to the study of the electroencephalogram and other biomedical signals. Any study on source separation has immediate application to reduce crosstalk and to measure the acoustic pollution in the cities as the noises can be classified according to their

origin. The list of possible applications, some studied others imagined, becomes endless.

In this Thesis, we shall study the signal mixtures that are linear and time invariant. We shall prove that the separation is possible if we assume that there are as many sensors as sources, that the sources are statistically independent and that at most one of them is gaussian distributed.

The Thesis will be exposed in two parts. Firstly, we present the algorithms and methods that have been developed, as to date, using the assumption of statistical independence among the sources. These methods are based on maximum likelihood estimations of parameters of the mixtures and, for this reason, they are efficient assuming that the statistical distribution of the sources is known beforehand or has been estimated, though it increases the computational cost. In any case, when this condition is not fulfilled, the source separation may not be achieved. However, there are algorithms that estimate the statistics of the observed signals in order to calculate the statistical distribution of the sources, but they require a considerable computational effort.

Secondly, a set of necessary and sufficient conditions to obtain the source separation is presented, from which we obtain a set of polynomic equations. Next, we present an algorithm (Sevilla) to solve them efficiently. We shall show that the separation can be achieved by solving linear equations and that the accuracy can be increased as far as the data allow. Finally, we obtain very low computational cost algorithms.

Agradecimientos

Esta Tesis Doctoral no hubiera sido terminada sin el apoyo y la supervisión de tres personas, a saber, mi tutor D. José I. Acha y mis padres. A ellos dedico mi trabajo. Muy especialmente, deseo que esta Tesis honre, de forma póstuma, la memoria de mi padre, cuya mayor ilusión fue siempre la de ver llegar este día.

Quiero también recordar a mis compañeros y amigos del Área de Teoría de la Señal y Comunicaciones, entre los cuales he desarrollado este trabajo, para expresarles mis mejores deseos. De hecho, debo hacer extensivo este agradecimiento a todos mis amigos, de dentro y fuera de la Universidad.

Por último, no quiero dejar de reconocer a mi antiguo profesor del bachillerato, D. José Miguel Pino, q.e.p.d., por haberme transmitido su gran amor por las Matemáticas, condicionando así toda mi carrera.

ÍNDICE

Parte I: Introducción

1. La Separación de Fuentes

1.1 Introducción

Cuando he tratado de describir qué pretende esta Tesis Doctoral, a menudo he utilizado el siguiente ejemplo: "supón que conversamos frente a un par de micrófonos. Salvo que hablemos por turnos, los micrófonos recogerán la *superposición* de nuestras voces. Pues bien, tomemos las grabaciones y tratemos de separar tu voz de la mía". Esto es, a grandes rasgos, lo que se conoce como problema de "Separación Ciega de Fuentes".

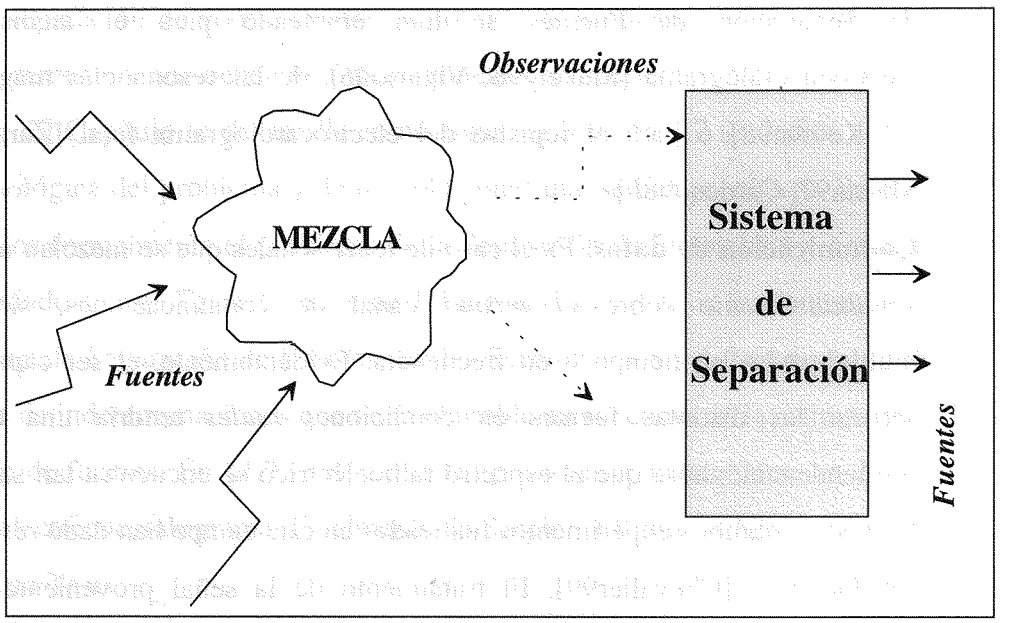

Figura 1.1. *Planteamiento del Problema.* Señales *fuente* desconocidas se *mezclan* dando lugar a las señales que llamamos *observaciones*. El propósito es la estimación de dichas *fuentes* a partir de las *observaciones.*

Como muestra la Figura 1.1, el problema se plantea en los siguientes términos: un determinado proceso combina, superpone o, genéricamente, "mezcla" señales a las

que se conoce como "fuentes". Las señales mezcladas reciben el nombre de "observaciones". El adjetivo "ciega" enfatiza el hecho de que, por hipótesis, tanto las *fuentes* como el proceso de *mezcla* son, en cada situación particular, desconocidos. En estas condiciones, se desea diseñar un sistema capaz de invertir todo el proceso de la mezcla y recuperar las *fuentes*.

La Separación se debe basar tanto en el conocimiento de las *observaciones* como en algunas hipótesis genéricas sobre la naturaleza de las fuentes y de la mezcla.

En concreto, en esta Tesis se trabajará sobre la suposición de que las fuentes son *estadísticamente independientes* y la mezcla es *lineal* e *invariante en el tiempo*. Con estas hipótesis, se han identificado los siguientes campos de aplicación:

- **Estudio de señales de naturaleza biomédica.** En particular, los algoritmos de Separación de Fuentes se han empleado para el análisis del electroencefalograma [Makeig96, Vigario96], de las resonancias magnéticas [McKeown98] o para el registro del electrocardiograma fetal [Zarzoso98, Martín97, Cardoso98*b*].
- **Comunicación de datos.** Es el caso de tener señales que se mezclan al viajar *simultáneamente* sobre *el mismo* canal de comunicaciones, sin estar multiplexadas en tiempo o en frecuencia. Evidentemente, el ser capaces de separar las distintas fuentes en condiciones *reales* tendría una enorme trascendencia, ahora que el espectro radioeléctrico se encuentra tan saturado. De hecho, algunos experimentos realizados en este campo han dado resultados satisfactorios [Chevalier99]. El tratamiento de la señal proveniente de un sistema ('array') de antenas también se beneficia de la investigación en el campo de la Separación de Fuentes [Cardoso93].
- **Codificación y extracción de características de las señales.** Mediante las técnicas de Separación de Fuentes, se trata de descomponer la señal de interés en otras, independientes entre sí, que pongan mejor de manifiesto la información. Esta aplicación se considera una extensión del clásico Análisis de

Componentes Principales [Haykin94*b*, pág. 363 y *ss.*]. Merece la pena destacar la investigación en el campo del tratamiento de imágenes [Bell97] y sus aplicaciones, por ejemplo, al reconocimiento de rasgos faciales [Barlett97] y la lectura de labios [Gray98].

- **Tratamiento de señales de audio.** Evidentemente, la separación de voces (problema del 'cocktail party') parece ser una aplicación natural de los algoritmos de Separación de Fuentes. Puede ser útil en videoconferencias, conciertos y, en general, en cualquier situación en la que se desee destacar una señal sobre el nivel del ruido o de las interferencias [Te-Won98, pág. 94].

Se debe decir que la investigación, en su mayor parte, se ha desarrollado en condiciones de *laboratorio* y, por el momento, no hay productos suficientemente desarrollados que tengan interés comercial. No obstante, al menos en lo que se refiere a la separación de voces, existen prototipos de pequeño tamaño, realizados electrónicamente y que tienen un funcionamiento, en apariencia, satisfactorio.

El Capítulo se desarrolla como sigue: la Sección 1.2 se dedica a la revisión cronológica del problema y de las soluciones que se han propuesto. En §1.3 planteamos formalmente el problema de la Separación de Fuentes tal y como va a ser abordado en esta Tesis. La Sección 1.4 presenta la hipótesis fundamental en la que se basa la Separación de Fuentes. En las Secciones 1.5 y 1.6 se discute sobre una serie de indeterminaciones que tiene el problema así como de la estimación del número de fuentes. En §1.7 se describe la aplicación a un problema *real*, el análisis del electroencefalograma. Finalmente, la Sección 1.8 fija las hipótesis y el alcance de esta Tesis.

1.2 Un poco de Historia.

El primer artículo del que tenemos noticia sobre *Separación Ciega de Fuentes* fue publicado por dos profesores de la Universidad de Grenoble, J. Herault y C.

Jutten, en las Actas del Congreso sobre "Redes Neuronales y Computación" celebrado en Snowbird (EE.UU) en Abril de 1986 [Herault86]. En él presentaban una red neuronal que separaba una mezcla de dos fuentes, lo que corroboraban mediante experimentos. En palabras del propio Jutten, la motivación para su trabajo habría nacido tras una discusión informal, de cafetería, sobre los mecanismos que tiene el cerebro para extraer información útil de entre la amalgama de estímulos que recibe.

Sin embargo, en aquel momento el algoritmo de Herault y Jutten pasó casi desapercibido. Las causas hay que buscarlas en que, por una parte, la fundamentación del algoritmo era prácticamente heurística, con numerosas lagunas, y, por otra parte, su aparición prácticamente coincidió en el tiempo con la de la red de Hopfield (presentada en 1982) [Haykin94*b*] y la del algoritmo de retropropagación (publicado en 1986) [Haykin94*b*], que acapararon todas las atenciones.

La verdadera presentación en sociedad del algoritmo de Herault y Jutten llega en 1991 de la mano de tres artículos [Jutten91, Comon91, Sorouchyari91]. En ellos se formaliza la estructura de la red neuronal y su regla de aprendizaje; pero, sobre todo, se estudia con rigor la convergencia del algoritmo. Desde entonces, el algoritmo de Herault y Jutten ha sido objeto de una gran atención [Deville96], [Macchi97].

En 1993, J.F Cardoso y A. Souloumiac dan a conocer el algoritmo JADE [Cardoso93], que hoy se considera uno de los más potentes y seguros. Como anécdota, se cuenta que Cardoso se interesó por el problema de la Separación después de ver cómo una implementación electrónica de la red de Herault y Jutten separaba un par de señales durante un Congreso. Finalmente, en 1994 P. Comon sienta definitivamente los presupuestos y el alcance del problema de la Separación de Fuentes [Comon94].

De igual forma, queremos destacar la aportación en estos años de los españoles C. Puntonet y A. Prieto, que aportan un punto de vista fresco y muy original con lo que ellos llaman "Procedimientos Geométricos" [Puntonet95, Prieto97].

Finalmente, a partir de un trabajo preliminar del francés J.P. Nadal y el español N. Parga [Nadal94], los norteamericanos A. Bell y T. Sejnowski [Bell95] proponen en 1995 el algoritmo Infomax, en el que se ha fundado buena parte de la investigación posterior. En los últimos años, cabe citar la importantísima aportación del Profesor S.I. Amari [Amari97*a*, Amari98*a*].

En Enero de 1999 se ha celebrado en Aussois (Francia) ICA'99[1], el primer Congreso monográfico dedicado a la Separación de Fuentes. Su continuación se espera este año 2000 en Helsinki.

1.3 Planteamiento del Problema.

Pasamos ahora a formalizar matemáticamente el problema de la Separación de Fuentes. En §1.3.1 se precisa la definición de las señales que intervienen en el problema y en §1.3.2 se hace una clasificación de *los procesos de mezcla.*

1.3.1 Fuentes, Observaciones y Señales de Salida

En la literatura especializada parece haberse aceptado la siguiente notación, que será mantenida a lo largo de la Tesis:

- Las fuentes serán representadas por la letra s (del inglés 'source'). Dadas M fuentes distintas, las distinguiremos nombrándolas como $s_1(t)$, ..., $s_M(t)$. Para manejarlas con comodidad, se agruparán en el vector $\mathbf{s}(t)$ que se define como $\mathbf{s}(t) = [\ s_1(t), ..., s_M(t)\ ]^T$. En lo que sigue, supondremos que las *fuentes* son realizaciones de procesos estocásticos, de media cero y *estadísticamente independientes entre sí.* Que la media sea cero no es en absoluto restrictivo. En cambio, la hipótesis de *independencia* entre las fuentes puede ser discutible

[1]ICA es acrónimo de 'Independent Component Analysis'

según qué aplicaciones; en todo caso, parece que se puede asumir cuando las señales tienen un origen físico distinto.

- Las observaciones serán representadas por la letra x. Dadas N observaciones, generalmente la salida de N sensores, se denotarán como $x_1(t)$, ..., $x_N(t)$ y serán agrupadas en el vector $\mathbf{x}(t) = [\, x_1(t), ..., x_N(t) \,]^T$.
- El sistema de separación nos devolverá M señales de *salida* $y_1(t)$, ..., $y_M(t)$ que deben reproducir fielmente a las *fuentes*. Ahora bien, no se considera que soluciones como la siguiente sean admisibles:

$$y_1(t) = \begin{cases} s_1(t) & t < T \\ s_2(t) & t \geq T \end{cases} \qquad y_2(t) = \begin{cases} s_2(t) & t < T \\ s_1(t) & t \geq T \end{cases}$$

Se define el vector de señales de *salida* como $\mathbf{y}(t) = [\, y_1(t), ..., y_M(t) \,]^T$.

En una aplicación real, cabe esperar que el número de fuentes sea *desconocido* y, lo que es aún peor, varíe con el tiempo. No obstante, estos problemas serán tratados sólo de forma marginal en esta Tesis.

Finalmente, una pequeña precisión: es difícil distinguir entre fuente y *ruido*. En general, se considera que *fuente* es toda señal *deseada* que aparece en el registro de, al menos, dos sensores.

1.3.2 Clasificación de la mezcla

Se considera que la *mezcla* es el proceso que transforma unas señales, las *fuentes*, en otras, las *observaciones*. En esta Tesis se considerará que la mezcla es *lineal e invariante* en el tiempo, que, a su vez, se dividen en *instantáneas* y *convolutivas*:

A) Las Mezclas Lineales e Instantáneas.

Sea el vector $\mathbf{s}(t) = [\, s_1(t), \ldots, s_M(t) \,]^T$ que contiene a las M señales *fuente*, el vector $\mathbf{x}(t) = [\, x_1(t), \ldots, x_N(t) \,]^T$, con las N *observaciones* e $\mathbf{y}(t) = [\, y_1(t), \ldots, y_M(t) \,]^T$ el vector que contiene las M salidas. Se dice que la *mezcla* es *lineal* e *instantánea* si

$$\mathbf{x}(t) = \mathbf{A}\, \mathbf{s}(t) \tag{1.1}$$

donde **A** es una matriz N×M que recibe el nombre de *matriz de mezcla*. Nótese que la transformación es *lineal e invariante en el tiempo*. Como se dijo antes, identificaremos

N ⇔ N°. de *sensores*
M ⇔ N°. de *fuentes*

Por hipótesis, tanto **A** como $\mathbf{s}(t)$ *son desconocidos* y sólo se cuenta con las *observaciones y la hipótesis de independencia estadística* entre las fuentes.

Las salidas se determinan como

$$\mathbf{y}(t) = \mathbf{B}\, \mathbf{x}(t) \tag{1.2}$$

siendo **B** una matriz M×N, que recibe el nombre de *matriz de separación*. La relación que liga *fuentes* y *salidas* es, por lo tanto,

$$\begin{aligned} \mathbf{y}(t) &= \mathbf{B}\, \mathbf{x}(t) = \\ &= \mathbf{B}\, \mathbf{A}\, \mathbf{s}(t) = \mathbf{G}\, \mathbf{s}(t) \end{aligned} \tag{1.3}$$

donde **G** es una matriz M×M a la que se llama *matriz global de transferencia*. La Figura 1.2 muestra la relación entre las distintas señales.

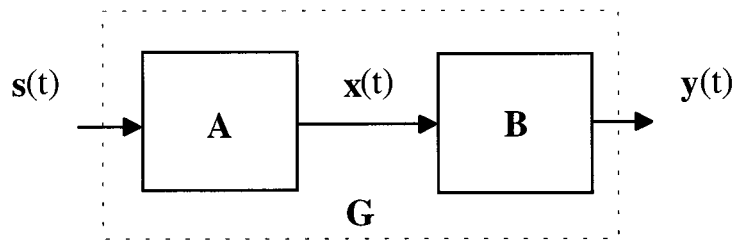

Figura 1.2. Modelo de la mezcla lineal e instantánea.

Por último, hagamos una observación que es interesante: si $\mathbf{s}(t)=[s_1(t),...,s_N(t)]^T$ es un vector cuyas componentes no son *fuentes diferentes* sino *muestras* de una *misma* señal y **A** es una matriz de Toeplitz triangular inferior, entonces (1.1) y (1.2) representan justamente un problema de *Igualación Ciega de Canal* [Comon94, Haykin94*a*]. De esta forma, la Separación de Fuentes y el problema de Igualación son, en cierta forma, problemas equivalentes. De hecho, buena parte de los algoritmos de Separación de Fuentes son en parte deudores de los métodos de Igualación Ciega [Comon94, Delfosse95].

Ejemplo 1.1. Veamos un pequeño ejemplo para ilustrar el problema. Supongamos tener las dos fuentes que se muestran en la Figura 1.3 (*a*), donde $s_1(t)$ es un tren de pulsos rectangulares y $s_2(t)$ un ruido de distribución uniforme. La matriz de mezcla se toma igual a

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$$

Las observaciones se muestran en la Figura 1.3 (*b*), siendo las fuentes irreconocibles. El algoritmo SEVILLA de Separación de Fuentes, tomando como entrada *únicamente* las observaciones, devuelve las señales de la Figura 1.3 (*c*), muy parecidas a las fuentes originales.

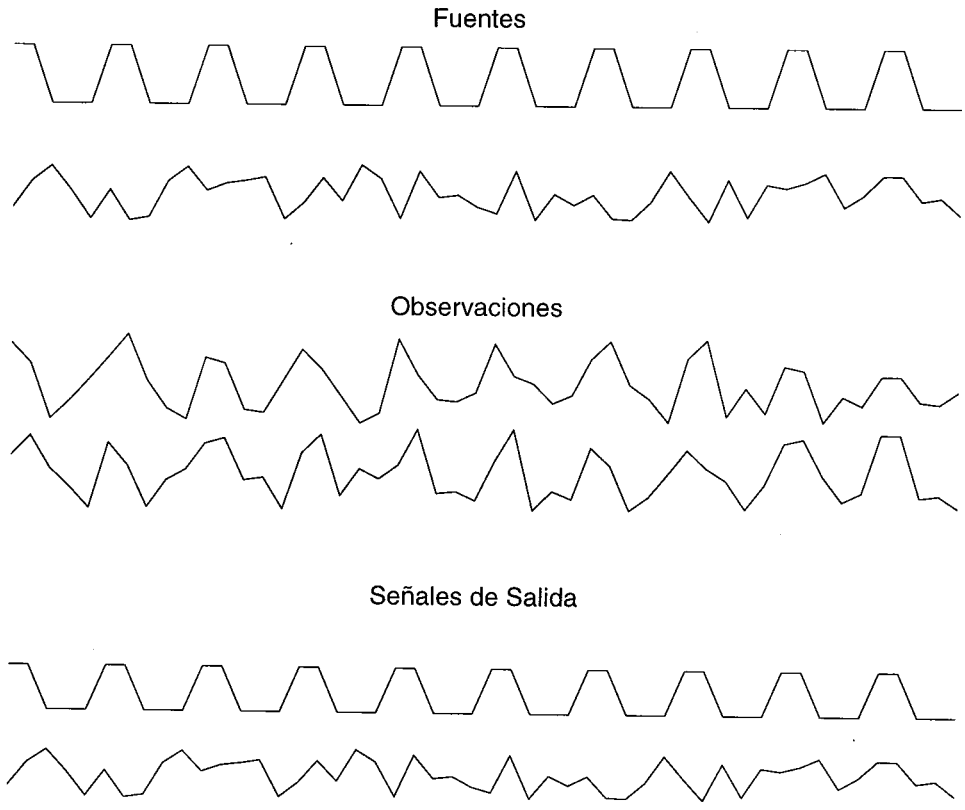

Figura 1.3. De arriba a abajo: (*a*) fuentes (*b*) observaciones (*c*) fuentes tal y como las estima el algoritmo.

B) Las Mezclas Lineales y Convolutivas

El modelo de *mezcla lineal* e *instantánea* se generaliza suponiendo que las *fuentes* son *filtradas* por el medio antes de llegar a los sensores. Es decir, la i-ésima observación ahora se construye como

$$x_i(t) = \sum_{j=1}^{N} h_{ij}(t) * s_j(t) \tag{1.4}$$

donde '$*$' denota la operación de convolución y $h_{ij}(t)$ es la *respuesta al impulso* del medio por el que se propaga la j-ésima fuente hasta llegar al i-ésimo sensor.

El modelo de *mezcla convolutiva* recoge básicamente dos fenómenos:

- Que las *fuentes* no alcancen *simultáneamente* todos los *sensores*. En este caso, el medio de propagación tan sólo retrasa las señales.
- Que las *fuentes* sean además distorsionadas por el canal.

La Transformada de Fourier permite convertir (1.4) en una relación algebraica:

$$X_i(f) = \sum_{i=1}^{M} H_{ij}(f)\, S_i(f) \tag{1.5}$$

por lo que, *para cada frecuencia,* (1.4) nos define una *mezcla lineal* e *instantánea*. Por esta razón, nos centraremos en esta Tesis en el estudio de *las mezclas lineales e instantáneas*, que, aunque constituyen el problema más simple, son la clave para estudiar los casos más generales.

1.4 Principios de Separación

Se admite que las fuentes son *estadísticamente independientes* entre sí. Recordemos que las señales $s_1(t)$, $s_2(t)$, ..., $s_N(t)$ son *independientes* si su *función de densidad de probabilidad* (*f.d.p*) *conjunta* se puede factorizar en el producto de las *funciones de densidad marginal* de cada una de las señales.

De igual forma, diremos que las mismas señales son *independientes dos a dos* si la *f.d.p conjunta* de cada *pareja* de señales es el producto de las *f.d.p* marginales de ambas.

En general, la *independencia dos a dos* no implica la *independencia* de las señales cuando se las considerada en su conjunto [Papoulis91, pág. 184 y problema 8.2, pág. 237).

La hipótesis de independencia resulta útil porque da pié a toda una serie de resultados interesantes. Empezamos por un teorema dado a conocer de forma independiente por Darmois y Skitovich alrededor de 1950 [Comon94, Cao96]:

Teorema 1.1 (*Teorema de Darmois-Skitovich*). Sea $\mathbf{s} = [s_1(t), s_2(t), ..., s_N(t)]^T$ un vector de N señales estadísticamente *independientes*, siendo N mayor que 1, y definamos

$$y_1(t) = a_{11}\, s_1(t) + a_{12}\, s_2(t) + ... + a_{1N}\, s_N(t)$$
$$y_2(t) = a_{21}\, s_1(t) + a_{22}\, s_2(t) + ... + a_{2N}\, s_N(t)$$

Supongamos que $y_1(t)$ e $y_2(t)$ son *independientes*. Entonces, todas las variables $s_i(t)$ para las que $a_{1i}\, a_{2i} \neq 0$ son de *distribución gaussiana*.

El Teorema 1.1 garantiza que $y_1(t)$ e $y_2(t)$ no están compuestas de las mismas fuentes (lo que es un primer paso hacia la Separación) cuando $y_1(t)$ e $y_2(t)$ son independientes, supuesto que a lo más una fuente tiene distribución *gaussiana.*

A partir de este Teorema, Comon [Comon94] prueba otro que es fundamental en el problema de Separación de Fuentes:

Teorema 1.2. (*Teorema de Comon*). Sea $\mathbf{s}(t)$ un vector *aleatorio* N×1 de componentes *independientes*, de las cuales como mucho una es *gaussiana*. Sea **G** una matriz *ortogonal* N×N e $\mathbf{y}(t) = \mathbf{G}\ \mathbf{s}(t)$. Entonces, las siguientes proposiciones son *equivalentes*:

1. -Las componentes de $\mathbf{y}(t)$ son *independientes.*

2.- Las componentes de $\mathbf{y}(t)$ son *independientes* dos a dos.

3.- La matriz **G** tiene *un sólo elemento no nulo por fila y columna.*

Es decir, cuando se impone que las componentes de $\mathbf{y}(t)$ sean *independientes* o, incluso, la condición más débil de *independencia dos a dos*, entonces se consigue deshacer la mezcla de las fuentes. Este Teorema proporciona una base muy sólida sobre la que construir los algoritmos de Separación y hace que sea importante y útil la hipótesis de independencia estadística entre las fuentes. Nótese que el Teorema es aplicable *sólo cuando el número de fuentes es igual que el número de sensores* (N = M en (1.1)).

De todas formas, merece la pena hacer comentar que la hipótesis de *independencia estadística* no es útil si la mezcla *no* es lineal: Darmois [Taleb99] también ha demostrado que hay infinitas transformaciones *no lineales* capaces de transformar un vector de *componentes independientes* en otro cuyas componentes

también lo son. Por ejemplo, supongamos una mezcla en la que las observaciones se generan como $x_1(t) = s_1(t)$ y $x_2(t) = s_1^2(t) + s_2(t)$. Tanto

$$y_1(t) = x_1(t) = s_1(t) \qquad y_2(t) = x_2(t) - x_1^2(t) = s_2(t)$$

como

$$y_1(t) = x_1^2(t) = s_1^2(t) \qquad y_2(t) = x_2(t) - x_1^2(t) = s_2(t)$$

son pares de salidas *estadísticamente independientes* y, sin embargo, sólo en el primer caso la solución es satisfactoria. Por lo tanto, la hipótesis de *independencia* entre las fuentes no es, en general, *suficiente* cuando hay *no linealidades* en el proceso de mezcla o en el de separación. Taleb y Jutten [Taleb99*a*, Taleb99*b*] han estudiado ciertos casos; pero, en general, el problema de las *mezclas no lineales* es poco conocido.

Por otra parte, puede haber situaciones en las que la hipótesis de *independencia estadística* de las fuentes no sea razonable. Merece la pena notar la existencia de algoritmos que sustituyen esta hipótesis por otras relativas a las características temporales de las fuentes, su caracter discreto, etc. [Puntonet95, Prieto97, Veen98].

1.5 Indeterminaciones del problema

Hay un par de *indeterminaciones* en este problema:

- En el modelo (1.1) podemos multiplicar cualquier columna de **A** por una constante y dividir la correspondiente fuente por la misma constante, sin que ello tenga *ningún* efecto sobre las observaciones. En conclusión, la *escala* o *potencia* de las fuentes no puede ser determinada.

- Si permutamos el *orden* de las fuentes y, de igual manera, el *orden* de las columnas de **A**, no alteramos las observaciones. Por lo tanto, no se puede identificar ninguna relación de orden entre las fuentes.

No es posible eliminar estas indeterminaciones haciendo uso tan sólo de la hipótesis de independencia estadística entre las fuentes (de hecho, ambas están presentes en la formulación del Teorema 1.2, cuando no se afirma que **G** sea la matriz identidad). Por ello, se admite que la *separación es completa* si **G** = **P D**, donde **P** es una *matriz de permutación*, es decir, que se obtiene permutando el orden de las filas de la matriz identidad, y **D** es una *matriz diagonal*, pero no necesariamente la matriz *identidad*, es decir, **D** modela los cambios en la escala de las fuentes. En suma,

$$\mathbf{G} = \mathbf{P}\,\mathbf{D}$$

es una matriz en la que hay un elemento (y sólo uno) distinto de cero por fila y columna, como aparece en el Teorema 1.2.

Por otra parte, como ya se mencionó, el Teorema 1.2 implícitamente asume que hay tantos sensores como fuentes. En general, la matriz de mezcla **A** será una matriz N×M, siendo N el número de sensores y M el número de fuentes. Con referencia a (1.3):

- Si M = N y **A** es invertible, está claro que siempre podemos separar las fuentes: basta con que **B** sea la matriz *inversa* de **A** o, en general, que la matriz **B** sea un producto de la forma $\mathbf{P}\,\mathbf{D}\,\mathbf{A}^{-1}$.
- Si N > M, es decir, *hay más sensores que fuentes* y el rango de **A** es justamente M, reducimos este caso al anterior sin más que desechar la información de N – M de los sensores.

- Si N < M, es decir, *hay más fuentes que sensores* o bien el rango de **A** es menor que el número de fuentes, no es posible encontrar una matriz **B** que consiga que la *separación* sea *completa* [Cao96]: por ejemplo, supongamos tener N = 2 sensores y M = 3 fuentes con la matriz de mezcla:

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & 1 \end{bmatrix}$$

No existe la matriz **B**, de dimensiones 3×2, tal que **G** = **B** **A** = **I**, la matriz *identidad* (por extensión, no se puede conseguir que **G** sólo tenga un elemento distinto de cero por fila y columna). No obstante, consideremos el caso en el que

$$\mathbf{B} = \frac{1}{2}\begin{bmatrix} 1 & -1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \quad \Rightarrow \quad \mathbf{G} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

A la vista de **G**, hemos *extraído* perfectamente la *primera* fuente. Este ejemplo muestra que resulta posible recuperar *alguna* fuente, incluso cuando el número de señales es mayor que el de sensores. De todas formas, no podemos asegurar que siempre vaya a existir tal matriz **B** [Cao96]. Aún en estos casos, se han propuesto algoritmos que separan las *fuentes*, aunque para ello la hipótesis de independencia no es suficiente y necesitamos otras adicionales, como el conocimiento de la función de densidad de probabilidad de las fuentes [Te-Won99, Zhang99].

Ejemplo 1.2 (*Ruido en la mezcla*) Supongamos el siguiente modelo:

$$\mathbf{x}(t) = \mathbf{A}\,\mathbf{s}(t) + \mathbf{r}(t)$$

donde $\mathbf{r}(t)$ es un vector de *ruidos*. Supongamos que $\mathbf{s}(t)$ y $\mathbf{r}(t)$ son vectores de dimensiones N × 1 y **A** es una matriz N × N. Estas ecuaciones se pueden transformar en

$$\mathbf{x}(t) = [\ \mathbf{A} \mid \mathbf{I}\ ]\ \mathbf{u}(t)$$

donde $\mathbf{u}(t)$ es el vector de dimensiones 2 N × 1 que se construye concatenando $\mathbf{s}(t)$ y $\mathbf{r}(t)$, mientras que $[\ \mathbf{A} \mid \mathbf{I}\ ]$ es una matriz N × 2 N, siendo $\mathbf{I}$ la matriz identidad. En esta situación hay más señales que sensores, por lo que no se puede esperar separarlas todas completamente. Incluso aunque la inversa de la matriz de mezcla fuese conocida sólo obtendríamos

$$\mathbf{y}(t) = \mathbf{A}^{-1}\mathbf{x}(t) = \mathbf{s}(t) + \mathbf{A}^{-1}\mathbf{n}(t)$$

de forma que el ruido seguiría presente en la salida y no se puede eliminar con estos métodos.

1.6 Determinación del número de Fuentes

La determinación del número de fuentes es un problema poco tratado en este contexto. Habitualmente, se aprovechan resultados que son ya clásicos del tratamiento de la señal en sistemas ('arrays') de antenas y que se describen a continuación. La matriz de mezcla $\mathbf{A}$, M × N, siempre puede ser descompuesta como

$$\mathbf{A} = \mathbf{V}\,\mathbf{\Sigma}\,\mathbf{U}^{H} \tag{1.6}$$

donde $\mathbf{V}$ es M × M y unitaria, $\mathbf{U}$ es N × N y unitaria y $\mathbf{\Sigma}$ es M × N y *diagonal*, en el sentido de que $[\ \mathbf{\Sigma}\ ]_{ij} = 0$ a menos que $i = j$. En particular, resulta que el *rango* de la matriz $\mathbf{\Sigma}$ es precisamente *min*(M, N). Ésta es la conocida *descomposición en valores singulares* de la matriz [Noble89, pág. 375].

En adelante supondremos que el número de *sensores* es *mayor* o *igual* que el *número de fuentes* ($M \geq N$), de tal forma que el *rango de* Σ (y, por extensión, de $\mathbf{A}$) sea precisamente el número de *fuentes* presentes en la mezcla.

Debido a la indeterminación en la potencia o varianza de las fuentes, podemos suponer sin pérdida de generalidad que $E[\mathbf{s}(t)\mathbf{s}(t)^H] = \mathbf{I}$. Entonces, dado el modelo de mezcla con ruido

$$\mathbf{x}(t) = \mathbf{A}\,\mathbf{s}(t) + \mathbf{r}(t) \tag{1.7}$$

resulta que, de (1.6) y (1.7)

$$\begin{aligned} E[\mathbf{x}(t)\mathbf{x}(t)^H] &= \mathbf{A}\,\mathbf{A}^H + E[\mathbf{r}(t)\mathbf{r}(t)^H] = \\ &= \mathbf{V}\,\Sigma\Sigma^H\,\mathbf{V}^H + E[\mathbf{r}(t)\mathbf{r}(t)^H] \end{aligned} \tag{1.8}$$

donde se ha supuesto que el ruido es *estadísticamente independiente* de las fuentes. Es sencillo verificar que $\Sigma\Sigma^H$ es una matriz $M \times M$ *diagonal* que sólo tiene N elementos distintos de cero en su diagonal principal. Por simplicidad, se asume que

$$E[\mathbf{r}(t)\mathbf{r}(t)^H] = \sigma^2\,\mathbf{I}$$

aunque esta hipótesis podría ser relajada. Entonces, resulta que

$$E[\mathbf{x}(t)\mathbf{x}(t)^H] = \mathbf{V}\,(\,\Sigma\Sigma^H + \sigma^2\,\mathbf{I}\,)\,\mathbf{V}^H \tag{1.9}$$

Como $\mathbf{V}$ es unitaria, $E[\mathbf{x}(t)\mathbf{x}(t)^H]$ y $\Sigma\Sigma^H + \sigma^2\,\mathbf{I}$ son *semejantes*, esto es, tienen los *mismos* autovalores y, ya que $\Sigma\Sigma^H + \sigma^2\,\mathbf{I}$ es una matriz diagonal, se puede comprobar fácilmente que esta matriz

- Tiene exactamente M – N autovalores iguales a σ^2. Los autovectores de $E[\mathbf{x}(t)\mathbf{x}(t)^H]$ asociados generan el llamado "subespacio de ruido" [Nikias93, pág.341].
- Tiene tantos autovalores mayores que σ^2 como *fuentes hay en la mezcla*. Los autovectores de $E[\mathbf{x}(t)\mathbf{x}(t)^H]$ asociados generan el llamado "subespacio de señal" [Nikias93, pág.341].

Por lo tanto, conocidos los autovalores de $E[\mathbf{x}(t)\mathbf{x}(t)^H]$ se puede *estimar* con facilidad *el número de fuentes*. En la práctica, habida cuenta de que σ^2 es desconocido y la matriz $E[\mathbf{x}(t)\mathbf{x}(t)^H]$ no se puede determinar con exactitud, se emplean criterios sofisticados para estimar el número de fuentes [Nikias93]. Estos criterios se deben aplicar con precaución, ya que las expresiones que habitualmente se encuentran en los libros [Nikias93] presuponen que la distribución de las fuentes es gaussiana, lo que, como se ha visto, no resulta apropiado en el problema de Separación de Fuentes.

Además, la determinación del *subespacio de señal* permite mejorar la calidad de la Separación: proyectando las observaciones $\mathbf{x}(t)$ en este subespacio se reduce apreciablemente el ruido, especialmente cuando el número de sensores es bastante mayor que el de fuentes [Nikias93].

1.7 Separación de Fuentes: aplicación al estudio del EEG

Después de esta definición del problema, cabe preguntarse hasta qué punto tiene aplicación práctica. Por ejemplo, parece natural que los discursos de hablantes diferentes sean estadísticamente independientes entre sí. De igual manera, se admite bien que la mezcla de voces sea, normalmente, convolutiva. Sin embargo, hay problemas en los que el ajuste de los modelos a la realidad es la propia materia de

investigación. Nos parece especialmente llamativo el análisis del electroencefalograma (EEG) mediante algoritmos de Separación de Fuentes y, por ello, lo describiremos a continuación. El contenido de esta Sección está tomado de las referencias [Makeig96, Makeig97, Jung98].

El EEG está compuesto de 17 señales registradas simultáneamente. En primera aproximación, despreciando los efectos de la propagación a través del cráneo, se puede considerar que estas señales son la mezcla *lineal e instantánea* de otras generadas en el cerebro, cuya naturaleza es desconocida. Por otra parte, cuesta aceptar que los diferentes generadores cerebrales de las señales actúen de forma *independiente*. De igual manera, no se conoce el número de fuentes.

El análisis del EEG se basa en la identificación de una serie de patrones, que se conocen como ondas alfa, beta, delta y theta. Las ondas alfa están presentes en el EEG de cualquier adulto relajado. Las ondas beta aparecen cuando la mente se concentra en algún pensamiento. Las ondas delta y theta son propias del sueño y su presencia en el EEG de un adulto en vigilia denota la existencia de alguna patología. Estas cuatro ondas aparecen repartidas por los 17 registros. Por otra parte, el EEG es una señal muy débil y, por ello, a menudo está enmascarado por todo tipo de interferencias (ruido de la red eléctrica, actividad eléctrica de los músculos faciales, etc.).

Los algoritmos tradicionales de Separación de Fuentes aplicados al EEG obtienen 17 señales que son independientes entre sí y cuyo significado no se sabe aún interpretar. Probablemente, las fuentes que obtenemos ni siquiera han sido generadas en una región cerebral bien definida, sino que tienen su origen en redes de neuronas distribuidas por todo el cerebro.

Ahora bien, los algoritmos tienden a separar la señal EEG propiamente dicha de las interferencias. En opinión de los autores [Makeig96, Makeig97], la eliminación que se consigue de la interferencia de la red de potencia eléctrica es más que notable, en comparación con lo que consiguen otros algoritmos tradicionales. Por otra parte, las ondas delta y theta tienden a quedar concentradas en unas pocas de las señales de salida, facilitando así su localización.

En suma, en esta aplicación es discutible que realmente se lleve a cabo una separación de fuentes *reales*. En su lugar, hemos de entender que se realiza una transformación sobre las señales que facilita la recogida de la información, lo que, en sí, es una nueva aplicación de las técnicas de Separación de Fuentes. De hecho, en este caso se prefiere hablar de "Análisis de Componentes Independientes" (ICA, 'Independent Component Analysis') antes que de Separación de Fuentes [Hyvärinen98]. El "Análisis de Componentes Independientes" a menudo se presenta como una generalización del clásico "Análisis de Componentes Principales" [Haykin94*b*, pág. 363].

1.8 Planteamiento y Estructura de la Tesis

La presente Tesis se va a dedicar al estudio del problema de Separación Ciega de Fuentes en las siguientes condiciones:

- La mezcla es *lineal e instantánea.*
- Las fuentes son realizaciones de procesos *estacionarios*, *reales*, de *media cero* y *estadísticamente independientes.*
- Se considera que hay *tantos sensores como fuentes en la mezcla.* Específicamente, *esto excluye la presencia de ruido.*

A este fin, se analizan las prestaciones de los algoritmos de Separación ya existentes, poniéndose especial cuidado en destacar tanto sus puntos fuertes como sus limitaciones. De igual manera, intentaremos poner de manifiesto las relaciones que existen entre los distintos algoritmos.

Después, desarrollaremos un algoritmo propio de Separación de Fuentes, SEVILLA (SEparación por la VIa de una fórmuLa LineAl), que compensa ciertas deficiencias de las restantes aproximaciones al problema. En particular,

presentaremos condiciones que garantizan la Separación de las Fuentes en cualquier situación y métodos computacionalmente simples de lograrlo.

En todo caso, la Tesis se centra en el campo de la investigación teórica, poniendo poco énfasis en las Aplicaciones.

Se desarrollará como sigue: en el Capítulo 2 se repasan brevemente las aportaciones fundamentales de otros autores. En el Capítulo 3 se estudiará el estimador de máxima verosimilitud de la matriz de mezcla y, después, en el Capítulo 4 presentaremos una selección de los algoritmos de Separación de Fuentes.

Nuestra aportación se desarrolla en los Capítulos 4 y 5. El primero de ellos elabora la teoría en la que nos basaremos para separar las fuentes. En el segundo, presentaremos el algoritmo SEVILLA.

Finalmente, el Capítulo 7 se dedica a las simulaciones y comparativa entre los distintos algoritmos.

El Capítulo 8 se dejará para las Conclusiones y líneas futuras de investigación.

Parte II: Revisión de anteriores aproximaciones

2. Fundamentos Estadísticos de la Separación de Fuentes

2.1 Introducción

Recordemos el modelo de mezcla de las fuentes, en ausencia de ruido:

$$\mathbf{x}(t) = \mathbf{A}\,\mathbf{s}(t)$$

siendo $\mathbf{s}(t) = [s_1(t), ..., s_N(t)]^T$ un vector de N fuentes *estadísticamente independientes* de media cero, $\mathbf{A}$ una matriz invertible $N \times N$ (*matriz de mezcla*) y, por último, $\mathbf{x}(t) = [x_1(t), ..., x_N(t)]^T$ el vector de *observaciones*, único dato disponible. Para separar las fuentes, tenemos primero que estimar *la matriz de mezcla* para después multiplicar $\mathbf{x}(t)$ por la matriz inversa $\mathbf{A}^{-1}$. Por ello, se dedica este Capítulo a revisar los métodos de estimación de la matriz $\mathbf{A}$.

El Capítulo se desarrolla como sigue: comenzaremos definiendo en la Sección 2.2 el concepto de *función de estimación* de la matriz de mezcla y discutiendo algunas de sus propiedades.

En la Sección 2.3, se tratan los métodos basados en la optimización de un determinado criterio, haciendo hincapié en las particularidades propias del problema de Separación de Fuentes.

En la Sección 2.4 se estudia el *sesgo* y la *varianza* de los estimadores de la matriz de mezcla. El resultado fundamental de esta Sección y, por ende, del Capítulo entero, será la obtención de las *funciones de estimación* más *eficientes* o de *varianza mínima*.

Por último la Sección 2.5 se dedica a las conclusiones.

Recordamos de nuevo que las principales hipótesis que utilizaremos a lo largo del Capítulo van a ser las siguientes:

- La matriz de mezcla **A** es *invertible*.
- Las señales fuente son realizaciones de procesos estocásticos *reales* y *estacionarios*, siendo *estadísticamente independientes* entre sí y de media *cero*.

Cualquier otra suposición será especificada en su momento oportuno.

En cualquier caso, ninguna introducción al Capítulo es mejor que la escrita por Sir Arthur Conan Doyle en su relato "*The Adventure of the Blue Carbuncle*":

– " I can see nothing ", said I, handing it back to my friend

– " On the contrary, Watson, you can see everything. You fail, however, to reason from what you see. You are too timid in drawing your inferences " [Casella90].

2.2 Las Funciones de Estimación

En el Apartado 2.2.1 introducimos el concepto de *función de estimación* de la matriz de mezcla. En el Apartado 2.2.2 se evalúa la posibilidad de conseguir la Separación utilizando sólo estadísticos de segundo orden de las señales. La propiedad de *equivarianza* de los estimadores se presenta en el Apartado 2.2.3. El Apartado 2.2.4 trata de las llamadas *funciones contraste*, que son aquellos estimadores que se basan en la optimización de un determinado criterio.

2.2.1 Funciones de Estimación de la Matriz de Mezcla

Dada una señal aleatoria $\mathbf{x}(t) = [x_1(t), ..., x_N(t)]^T$ tal que la función de densidad de probabilidad (*f.d.p*) conjunta de sus componentes depende del parámetro determinista $\theta \in \Theta$, se llama *función de estimación* a cualquier función $\mathbf{F}$ (generalmente matricial) tal que

$$E[\mathbf{F}(\mathbf{x}(t); \theta)] = \mathbf{0} \tag{2.1}$$

donde la esperanza matemática se toma respecto de la variable aleatoria vectorial $\mathbf{x}(t)$. En cuanto a la notación, tenemos que hacer un comentario: en la literatura especializada (ver por ejemplo [Papoulis91]), es habitual reservar la letra negrita para las variables aleatorias mientras que la cursiva se asocia a cantidades deterministas. Sin embargo, no adoptaremos esta convención. Por el contrario, $x_i(t)$ denotará a la i-ésima observación, que, evidentemente, es la realización de un proceso estocástico. Las observaciones se agruparán en un vector, $\mathbf{x}(t)$, que se escribe en letra negrita. Para cada instante t, $x_i(t)$ y $\mathbf{x}(t)$ son variables aleatorias, escalar y vectorial, respectivamente. De hecho, fuentes, observaciones y salidas del sistema serán las únicas variables aleatorias que se manejen en nuestra exposición. Por lo demás, se emplea la cursiva para las cantidades escalares, la negrita minúscula para los vectores y la negrita mayúscula para las matrices.

Nótese que (2.1) es una ecuación para θ. Ahora bien, no todas las soluciones de (2.1) tienen por qué ser *aceptables*, es decir, (2.1) es una condición *necesaria* pero *no suficiente*.

Suponiendo que los procesos son *ergódicos*, podemos sustituir la esperanza matemática en (2.1) por un promedio temporal: dadas T muestras vectoriales $\mathbf{x}(1),..., \mathbf{x}(T)$, llamaremos θ_T al valor del parámetro que satisface la siguiente *ecuación de estimación*:

$$\frac{1}{T} \sum_{t=1}^{T} \mathbf{F}(\mathbf{x}(t); \theta_T) = \mathbf{0} \tag{2.2}$$

En general, $\theta \neq \theta_T$; pero se espera que $\theta \approx \theta_T$ si T es suficientemente grande.

En el problema de la Separación de Fuentes, $\mathbf{x}(t)$ es precisamente el *vector de observaciones* y Θ es el conjunto formado por *la matriz de mezcla* (o su inversa, según los casos) y todas aquellas otras matrices derivadas de ella mediante la permutación y/o escalado de sus columnas (respectivamente, sus filas).

Ejemplo 2.1. Supongamos que las *fuentes* sólo toman los valores ± 1 y sea $\mathbf{y}(t) = \mathbf{B}\ \mathbf{x}(t)$ el vector N×1 de salidas del sistema. Definamos [Cardoso98*a*]

$$F(\mathbf{x}(t); \mathbf{B}) = \sum_{i=1}^{N} (y_i^2(t) - 1)^2$$

Se cumple evidentemente que $E[\ F(\mathbf{x}(t); \mathbf{B})\ ] = 0$ si $\mathbf{B} = \mathbf{A}^{-1}$, por lo tanto $F(\mathbf{x}(t); \mathbf{B})$ es una *función de estimación* según (2.1). Podemos considerar esta función de estimación como la extensión natural del bien conocido criterio de "módulo constante" que se utiliza en deconvolución ciega [Haykin94*a*, pág. 34].

Ejemplo 2.2. Si el valor de los *estadísticos* de las fuentes es conocido, podemos utilizar una variante del conocido *método de los momentos* [Borovkov88, pág. 90] para estimar la matriz de mezcla. Por ejemplo, supongamos que las fuentes tienen media cero, varianza unidad y denotemos por κ_{si} la curtosis de la *i*-ésima fuente. Como los cumulantes cruzados de cualquier orden de variables estadísticas *independientes* valen cero (ver Apéndice B), definamos

$$c_2(\mathbf{x}(t); \mathbf{B}) = \sum_{i,j} \left| cum(y_i(t), y_j(t)) - \delta_{ij} \right|^2$$

y

$$c_4(\mathbf{x}(t);\mathbf{B}) = \sum_{i,j,k,l} \left| cum(y_i(t), y_j(t), y_k(t), y_l(t)) - \kappa_{si}\delta_{ijkl} \right|^2$$

donde δ_{ij} y δ_{ijkl} son, respectivamente, los símbolos de Kronecker de 2° y 4° orden. Resultará que $c_2(\mathbf{x}(t);\mathbf{B}) = c_4(\mathbf{x}(t);\mathbf{B}) = 0$ cuando $\mathbf{B}$ es una matriz de separación correcta. Al determinar el gradiente de estas cantidades $c_2(\mathbf{x}(t);\mathbf{B})$ y $c_4(\mathbf{x}(t);\mathbf{B})$ e igualarlo a cero se obtiene la ecuación matricial [Cardoso97*b*] (se omite la dependencia con t)

$$E[\,\mathbf{y}\mathbf{y}^T - \mathbf{I} + \varphi(\mathbf{y})\,\mathbf{y}^T - \mathbf{y}\,\varphi(\mathbf{y})^T] = \mathbf{0}$$

siendo $\varphi(\mathbf{y}) = [\,-\kappa_{s1}y_1^3, -\kappa_{s2}y_2^3, \ldots, -\kappa_{sN}y_{sN}^3\,]^T$. Así, asociada a este criterio se encuentra la función de estimación

$$\mathbf{F}(\mathbf{x};\mathbf{B}) = \mathbf{y}\mathbf{y}^T - \mathbf{I} + \varphi(\mathbf{y})\,\mathbf{y}^T - \mathbf{y}\,\varphi(\mathbf{y})^T$$

Los estadísticos de las fuentes son conocidos cuando se conoce la naturaleza de estas señales. En la referencia [Cardoso96*b*] encontramos la aplicación de este método para mezclas de señales 16-QAM.

Junto al *método de los momentos*, el estimador de *máxima verosimilitud* es el más utilizado en los problemas de inferencia estadística clásica. Estudiaremos este último en el Capítulo siguiente.

Ejemplo 2.3. Supongamos que todas las fuentes tienen media cero y varianza unidad, $E[\,s_i^2(t)\,] = 1$ para todo i, siendo $\mathbf{y}(t) = \mathbf{B}\,\mathbf{x}(t)$ el vector N×1 de salidas del sistema. Definamos

$$\mathbf{F}(\mathbf{x}(t);\mathbf{B}) = \mathbf{y}(t)\,\mathbf{y}^T(t) - \mathbf{I}$$

siendo $\mathbf{I}$ la matriz identidad. Se verifica que E[$\mathbf{F}(\mathbf{x}(t); \mathbf{B})$] = $\mathbf{0}$ si la matriz $\mathbf{B} = \mathbf{A}^{-1}$ por lo que $\mathbf{F}(\mathbf{x}(t); \mathbf{B})$ es otra función de estimación. Ahora bien, definiendo $\mathbf{G} = \mathbf{B}\,\mathbf{A}$, esto es,

$$\mathbf{y}(t) = \mathbf{G}\,\mathbf{s}(t)$$

basta que $\mathbf{G}$ sea cualquier matriz *ortogonal* para que E[$\mathbf{y}(t)\mathbf{y}(t)^T$] = $\mathbf{I}$ y, por lo tanto, E[$\mathbf{F}(\mathbf{x}(t); \mathbf{B})$] = $\mathbf{0}$ sin que por ello hayamos separado las fuentes. Este ejemplo ilustra que, en general, las ecuaciones para la matriz de mezcla que se obtienen de E[$\mathbf{F}(\mathbf{x}(t); \mathbf{B})$] = $\mathbf{0}$ pueden tener soluciones no deseadas. De hecho, conseguir que E[$\mathbf{y}(t)\mathbf{y}^T(t)$] = $\mathbf{I}$ ya es posible por medio del clásico Análisis de Componentes Principales [Haykin94*b*, pág. 363], cuyo objetivo no es separar componentes independientes.

2.2.2 Separación utilizando Estadísticos de Segundo Orden

La matriz de mezcla $\mathbf{A}$ puede ser descompuesta como $\mathbf{A} = \mathbf{V}\,\mathbf{\Sigma}\,\mathbf{U}^H$, donde $\mathbf{V}$ y $\mathbf{U}$ son matrices unitarias de dimensiones $N \times N$ y $\mathbf{\Sigma}$ es una matriz diagonal, lo que se conoce como *descomposición en valores singulares* de la matriz [Noble89].

Como

$$E[\,\mathbf{x}(t)\mathbf{x}^H(t)\,] = \mathbf{A}\,\mathbf{A}^H = \mathbf{V}\,\mathbf{\Sigma}\mathbf{\Sigma}^H\,\mathbf{V}^H$$

donde la esperanza matemática se toma respecto a la variable vectorial $\mathbf{x}(t)$, las matrices $\mathbf{V}$ y $\mathbf{\Sigma}$ *pueden ser identificadas a partir de los estadísticos de segundo orden de las observaciones*. Sin embargo, *no contienen ninguna información acerca de la matriz* $\mathbf{U}$. En conclusión, los estadísticos de segundo orden no son, en general, *suficientes* para resolver el problema de la Separación de Fuentes. No

obstante, siempre hay excepciones, aunque la hipótesis de independencia estadística entre las fuentes debe ser completada y/o substituida por otras. Por ejemplo,

- Si las fuentes están *incorreladas* (no es necesario que sean independientes) y la amplitud de una de ellas es *mucho mayor* que la de las restantes, entonces esta fuente se puede estimar bien con el clásico Análisis de Componentes Principales [Haykin94*b*, pág. 363], incluso *cuando sólo se dispone de un sensor*. Nótese que así obtenemos la fuente de mayor potencia; pero no la matriz de mezcla. Por ejemplo, en [Martín97] se aplica esta idea para la estimación del electrocardiograma fetal.
- Si las fuentes están incorreladas y su densidad espectral de potencia es distinta (pero desconocida, de manera que no sea posible discriminarlas mediante filtros lineales), entonces es posible separarlas con la información contenida en las matrices $E[\mathbf{x}(t)\mathbf{x}^H(t-\tau)]$ para distintos valores de τ [Tong91, Belouchrani97, VanGerven95].

En todo caso, el conocimiento de *los estadísticos de segundo orden* de las observaciones siempre permite decorrelarlas, como sigue [Comon93, Cardoso94]:

Algoritmo de Decorrelación de las Observaciones

- Estimar la matriz $\mathbf{R}_x = E[\ \mathbf{x}(t)\mathbf{x}^T(t)\ ]$
- Encontrar las matrices $\mathbf{V}$ y Λ tales que $\mathbf{R}_x = \mathbf{V}\,\Lambda\mathbf{V}^T$. Evidentemente, $\mathbf{V}$ es una matriz ortogonal cuyas columnas son los autovectores de $\mathbf{R}_x$ y Λ es una matriz diagonal que contiene los correspondientes autovalores.
- Formar la matriz $\mathbf{W} = \Lambda^{-1/2}\,\mathbf{V}^T$
- Sea $\mathbf{z}(t) = \mathbf{W}\,\mathbf{x}(t)$. Por construcción, $E[\ \mathbf{z}(t)\mathbf{z}^T(t)\ ] = \mathbf{I}$, es decir, las componentes de $\mathbf{z}(t)$ están incorreladas.
- Sustituir $\mathbf{x}(t)$ por $\mathbf{z}(t)$, esto es, $\mathbf{x}(t) = \mathbf{z}(t)$.

Tras aplicar este algoritmo, las componentes de $\mathbf{x}(t)$ en el instante t estarán incorreladas. En este caso, se puede demostrar que la matriz de *mezcla* correspondiente es *ortogonal*:

Ya que la escala de las fuentes no puede ser determinada, supondremos que su varianza es la unidad, es decir, $E[\ \mathbf{s}(t)\mathbf{s}^T(t)\ ] = \mathbf{I}$. Dado que

$$\mathbf{x}(t) = \mathbf{A}\ \mathbf{s}(t)$$

resulta que $E[\ \mathbf{x}(t)\mathbf{x}(t)^T\ ] = \mathbf{A}\ E[\ \mathbf{s}(t)\mathbf{s}(t)^T\ ]\ \mathbf{A}^T = \mathbf{A}\ \mathbf{A}^T$, de donde

$$\mathbf{A}\ \mathbf{A}^T = \mathbf{I}$$

En conclusión, **A** es una matriz ortogonal y, por lo tanto, *goza de excelentes propiedades*. Esto también implica que la matriz de separación **B** debe ser ortogonal.

2.2.3 Equivarianza

Notemos una característica de las observaciones que no por obvia deja de ser interesante (en adelante omitiremos la dependencia explícita de las señales con el tiempo t en aquellos casos en los que no pueda haber confusión):

"Dado que la *función de densidad de probabilidad* (*f.d.p*) de $\mathbf{x} = \mathbf{A}\ \mathbf{s}$ está parametrizada por **A**, la *f.d.p* del producto $\mathbf{x'} = \mathbf{M}\ \mathbf{x} = (\ \mathbf{M}\ \mathbf{A}\ )\ \mathbf{s}$ depende del parámetro $\mathbf{M}\ \mathbf{A}$, sea cual sea la matriz **M**".

Debido a esta propiedad, se dice que la distribución de **x** es *equivariante* (o invariante) respecto al grupo de transformaciones lineales [Borovkov88, pág. 185], [Cardoso96*a*]. Es decir, el parámetro **A** de la distribución se transforma de la

misma manera que **x**. Resulta entonces muy natural y deseable que los estimadores de **A** compartan esta propiedad: formalmente, dado un conjunto de T observaciones vectoriales $\mathbf{X}_T = \{\mathbf{x}(1), ..., \mathbf{x}(T)\}$ y un estimador cualquiera $\mathfrak{A}$ de la matriz de mezcla **A**, se dice que $\mathfrak{A}$ es *equivariante* si

$$\mathfrak{A}(\mathbf{M}\, \mathbf{X}_T) = \mathbf{M}\, \mathfrak{A}(\mathbf{X}_T) \tag{2.3}$$

donde $\mathbf{M}\, \mathbf{X}_T$ denota al conjunto $\{\mathbf{M}\, \mathbf{x}(1), ..., \mathbf{M}\, \mathbf{x}(T)\}$, siendo **M** cualquier matriz *invertible*. Con otras palabras, si $\mathfrak{A}$ es un estimador equivariante, entonces las observaciones $\mathbf{X}_T$ y $\mathbf{M}\, \mathbf{X}_T$ son equivalentes pues, en virtud de la definición anterior, las deducciones acerca de **A** en el primer caso pueden convertirse en deducciones sobre **M A** en el segundo.

Sea $\mathbf{S}_T = \{\mathbf{s}(1), ..., \mathbf{s}(T)\}$ *una realización particular de las fuentes* y $\mathfrak{A}$ un estimador equivariante, entonces:

$$\hat{\mathbf{A}} = \mathfrak{A}(\mathbf{X}_T) = \mathfrak{A}(\mathbf{A}\, \mathbf{S}_T) = \mathbf{A}\, \mathfrak{A}(\mathbf{S}_T) \tag{2.4}$$

Según (2.4), las fuentes estimadas tienen la expresión:

$$\begin{aligned} \mathbf{s}_{est}(t) = \hat{\mathbf{A}}^{-1}\, \mathbf{x}(t) = \mathfrak{A}(\mathbf{S}_T)^{-1}\, \mathbf{A}^{-1}\, \mathbf{A}\, \mathbf{s}(t) = \\ = \mathfrak{A}(\mathbf{S}_T)^{-1}\, \mathbf{s}(t) \end{aligned} \tag{2.5}$$

De esta manera, *la Separación de Fuentes mediante estimadores equivariantes sólo depende de la realización particular* $\mathbf{S}_T$ pero *no* de la *matriz de mezcla*. Por eso no cabe esperar que la Separación sea mejor si **A** es "casi" diagonal; pero, sobre todo, tampoco empeorará cuando **A** sea "defectuosa" (casi singular).

En general es más sencillo verificar la relación (2.5) que (2.3). Por esta razón, a menudo se adopta (2.5) como definición de *equivarianza* [Cardoso95].

Es muy interesante notar que *el propio modelo del problema condiciona la existencia de estimadores equivariantes*: recordemos que la varianza de las fuentes es desconocida y, por ello, la escala de la matriz de mezcla está indeterminada. Para eliminar esta indeterminación, es muy corriente suponer que la varianza de las fuentes vale uno. Alternativamente, se puede convenir en que todos los elementos de la diagonal de **B** son iguales a la unidad, como se encuentra en la referencia [Jutten91*a*]; de esta forma, además, se reduce el número de coeficientes de **B** que hay que estimar con lo que se simplifica el problema.

Sin embargo, este último enfoque tiene un inconveniente, que explicaremos mejor mediante un ejemplo tomado de [Cardoso95]: sea una matriz de mezcla tan sencilla como

$$\mathbf{A} = \begin{bmatrix} \varepsilon & 1 \\ 1 & \varepsilon \end{bmatrix}$$

Hay dos posibles matrices de separación que contienen unos en su diagonal:

$$\mathbf{B}_1 = \begin{bmatrix} 1 & -1/\varepsilon \\ -1/\varepsilon & 1 \end{bmatrix} \quad \text{y} \quad \mathbf{B}_2 = \begin{bmatrix} 1 & -\varepsilon \\ -\varepsilon & 1 \end{bmatrix}$$

La relación entre las fuentes originales y sus estimaciones está fijada, en el primer caso, por la matriz de transferencia $\mathbf{C}_1 = \mathbf{B}_1 \mathbf{A} = (\varepsilon - 1/\varepsilon)\, \mathbf{I}$. Si $\mathbf{B} \approx \mathbf{B}_1$ y ε es muy pequeño, la salida $\mathbf{y}(t) = \mathbf{C}_1\, \mathbf{s}(t)$ se verá muy amplificada. Esto puede condicionar el comportamiento de los algoritmos, especialmente si son *adaptativos* [Cardoso95].

2.2.4 Funciones "contraste"

Se debe a Comon [Comon94] la introducción del concepto de *función contraste* en el problema de la Separación de Fuentes. En la literatura española especializada, se suele entender que un "contraste" es una *regla de decisión* que permite elegir entre diferentes hipótesis. Sin embargo, en nuestro contexto el significado de la palabra "contraste" es distinto: denotemos las observaciones con la letra $\mathbf{x}(t)$, como es habitual. Entonces, se dice que la función *escalar* $\psi(\mathbf{x}(t))$ es un *contraste* si verifica [Comon94]:

- $\psi(\mathbf{x}(t))$ sólo depende de la distribución estadística de $\mathbf{x}(t)$.
- $\psi(\mathbf{x}(t)) = \psi(\mathbf{P}\,\mathbf{x}(t))$ si es $\mathbf{P}$ una matriz *diagonal* o de *permutación.*
- $\psi(\mathbf{B}\,\mathbf{x}(t)) \geq \psi(\mathbf{M}\,\mathbf{x}(t))$ si las componentes de $\mathbf{B}\,\mathbf{x}(t)$ son independientes, esto es, $\mathbf{B}$ es una *auténtica matriz de separación*, para cualquier matriz $\mathbf{M}$.

Entonces, queda claro que un *contraste* es una función objetivo cuya optimización lleva a la Separación de las Fuentes. Veamos ahora un par de definiciones:

- Se dice que el contraste es *discriminante* [Comon94] si $\psi(\mathbf{B}\mathbf{x}(t))$ alcanza un máximo global sólo cuando $\mathbf{B}$ es una matriz que separa las fuentes. Es decir, *todos* los máximos globales de un contraste *discriminante* son auténticas matrices de separación. Sin embargo, no afirmamos nada sobre el carácter de los máximos *locales.*

- Se dice que el contraste es *ortogonal* [Comon94] si, como hipótesis de partida, supone que las *observaciones* están *incorreladas*, es decir,

$$E[\,\mathbf{x}(t)\mathbf{x}^{T}(t)\,] = \mathbf{I}$$

Recordemos que, sin pérdida de generalidad, esto equivale a suponer que tanto la matriz de mezcla como la de separación son *ortogonales*. Decorrelar las observaciones también es útil para incrementar la relación señal a ruido de la mezcla, como se vio en el Capítulo anterior (ver la Sección 1.6).

Por último, veamos la relación que existe entre las *funciones contraste* y las *funciones de estimación* previamente definidas. Suponiendo que $\psi(\mathbf{Bx}(t))$ es diferenciable, entonces de

$$\nabla_{\theta}\, \psi(\mathbf{Bx}(t)) = \mathbf{0} \tag{2.6}$$

se puede obtener una *función de estimación*, siendo $\nabla_{\theta}\, \psi(\mathbf{Bx}(t))$ la matriz cuya componente en la posición (i, j) vale $\frac{\partial}{\partial b_{ij}} \psi(\mathbf{Bx}(t))$. Sin embargo, no es cierto que toda *función de estimación* se obtenga como la derivada de una función *contraste*.

Ejemplo 2.4. El mismo Comon [Comon94], demuestra que

$$\psi(\mathbf{y}(t)) = \psi(\mathbf{Bx}(t)) = \sum_{i=1}^{N} (\kappa_{yi}^{r})^2$$

es un contraste *ortogonal* para todo $r \geq 2$, siendo κ_{yi}^{r} el *cumulante* de orden r de la salida y_i; es decir, cuando $r = 2$ tenemos la *varianza*; para $r = 3$, el *coeficiente de asimetría*; con $r = 4$, la *curtosis*, etc. Es más, cuando $r \geq 3$ el contraste es *discriminante* siempre que, como mucho, sólo el *cumulante* de orden r de *una* de las fuentes se anule (por lo que, a lo más, *una* fuente puede ser de distribución gaussiana). Esto da a entender que los estadísticos de orden mayor o igual que tres contienen información suficiente para separar las señales, aunque los cumulantes de orden *impar* son poco prácticos porque se anulan

cuando las *f.d.p* son simétricas. En la práctica es corriente trabajar con los cumulantes de orden *cuatro* [Comon94, Cardoso93].

Probar que $\psi(\mathbf{y}(t))$ es un *contraste* es prolijo y omitiremos los detalles. En su defecto, repasaremos las líneas generales de la demostración, remitiéndonos siempre a la referencia original [Comon94]:

La prueba sólo hace uso de las propiedades algebraicas de los *cumulantes*. Sea $\mathbf{y}(t) = \mathbf{B}\,\mathbf{x}(t)$, donde $E[\,\mathbf{x}(t)\mathbf{x}(t)^T\,] = \mathbf{I}$ y la matriz $\mathbf{B}$ es *ortogonal*. Bajo estas condiciones, Comon demuestra que la *suma* del cuadrado de *todos* los *cumulantes* de orden r de $\mathbf{y}(t)$ es *constante* y *no depende de* $\mathbf{B}$. Por *cumulantes* de $\mathbf{y}(t)$ entendemos todos aquellos cumulantes que involucran a las componentes del vector $\mathbf{y}(t)$, tanto los cumulantes *cruzados* como los ya definidos cumulantes κ_{yi}^r. Por lo tanto, maximizar $\psi(\mathbf{y}(t))$ equivale a minimizar la suma de los cumulantes *cruzados* (al cuadrado) de las señales $y_i(t)$. Cuando $\mathbf{B}$ es una matriz de separación, las componentes del vector $\mathbf{y}(t)$ son *estadísticamente independientes* y todos los cumulantes *cruzados se anulan*, con lo que $\psi(\mathbf{y}(t))$ alcanza, en efecto, su valor máximo.

De igual forma, omitimos la prueba de que $\psi(\mathbf{y}(t))$ es un contraste *discriminante*.

Por último, diremos que en el Capítulo 4 se presenta el algoritmo que ha propuesto Comon [Comon94] para la optimización de su *función contraste*.

Sin embargo, se puede esperar que la distribución estadística de las fuentes propicie la aparición en los contrastes de máximos *locales* que *no sean matrices de separación*. Comon no puede probar lo contrario pero, basándose en simulaciones, sostiene que es muy improbable que tales óptimos locales consigan que su algoritmo fracase [Comon94].

Ejemplo 2.5. Extendiendo el trabajo de Comon, Moreau y Macchi [Moreau96] prueban que

$$\psi(\mathbf{y}(t)) = \psi(\mathbf{B}\mathbf{x}(t)) = \sum_{i=1}^{N} \left|\kappa_{yi}^{r}\right|$$

es un contraste ortogonal para todo $r \geq 2$. Como caso particular, cuando todas las fuentes tienen curtosis de signo negativo, se obtiene el contraste ortogonal

$$\psi(\mathbf{y}(t)) = -\sum_{i=1}^{N} \mathrm{E}\left[y_i^4\right]$$

Igualando las derivadas del contraste a cero, Cardoso [Cardoso97*b*] muestra que este contraste es equivalente a la *función de estimación*

$$\mathbf{F}(\mathbf{x}(t); \mathbf{B}) = \mathbf{y}(t)\,\mathbf{y}^{\mathrm{T}}(t) - \mathbf{I} + \mathbf{g}(t)\,\mathbf{y}^{\mathrm{T}}(t) - \mathbf{y}(t)\,\mathbf{g}^{\mathrm{T}}(t)$$

siendo $\mathbf{g}(t) = [\, y_1^3(t), \ldots, y_N^3(t) \,]^{\mathrm{T}}$, en el sentido de que las raíces de la ecuación $\mathrm{E}[\mathbf{F}(\mathbf{x}(t); \mathbf{B})] = \mathbf{0}$ coinciden con los máximos del contraste. Esta observación será de interés para el análisis del algoritmo EASI [Cardoso96*a*] en el Capítulo 4. Además, esta expresión es muy parecida a la de la función de estimación vista en el Ejemplo 2.2.

Ejemplo 2.6. Supongamos que las matrices **A** y **B** son ortogonales, o bien, que $\mathrm{E}[\,\mathbf{x}(t)\mathbf{x}(t)^{\mathrm{T}}\,] = \mathbf{I}$. Además, ninguna de las fuentes tiene, por hipótesis, una distribución gaussiana. Sea $\mathbf{b}_i^{\mathrm{T}}$ la i-ésima fila de **B** y definamos

$$y_i(t) = \mathbf{b}_i^{\mathrm{T}}\,\mathbf{x}(t)$$

Sea κ_{yi}^4 la curtosis de $y_i(t)$. Evidentemente, κ_{yi}^4 puede ser optimizada respecto a $\mathbf{b}_i^T$, con la restricción de que $\| \mathbf{b}_i \|_2 = 1$. Pues bien, Delfosse y Loubaton [Delfosse95] han probado que:

1.- Si al menos una fuente es super-gaussiana (curtosis positiva) y otra es sub-gaussiana (curtosis negativa), entonces todos los *máximos y mínimos locales* de κ_{yi}^4 se alcanzan cuando **B** es una matriz de separación correcta.

2.- Si todas las fuentes son sub-gaussianas, entonces los *mínimos locales* de la curtosis se alcanzan cuando **B** es una matriz de separación. Los *máximos locales* no separan las fuentes.

3.- Si todas las fuentes son super-gaussianas, entonces los *máximos locales* de la curtosis se alcanzan cuando **B** es una matriz de separación. Los *mínimos locales* no separan las fuentes.

Gracias a este Teorema, Delfosse y Loubaton [Delfosse95] proponen el contraste

$$\psi(y_i(t)) = \left(\kappa_{yi}^4\right)^2$$

cuya optimización les permite recuperar una de las filas de **B**. Las restantes filas se obtienen mediante la aplicación repetida del algoritmo.

Como han propuesto, entre otros, Shalvi y Weinstein [Shalvi90], se puede identificar un sistema de fase no mínima optimizando un criterio basado en la curtosis de la señal de salida del sistema. No podemos dejar de notar el parecido de esta técnica con las que hemos presentado en los sucesivos ejemplos y es que parte de

los métodos de Separación se derivan de algoritmos de Igualación Ciega de Canal y viceversa [Comon96]. De hecho, ya hicimos notar en el Capítulo 1 la relación entre ambos problemas.

Los Ejemplos 2.1 a 2.6 se han escogido porque ofrecen una visión panorámica de los fundamentos de buena parte de los algoritmos de Separación de Fuentes. Esta panorámica será completada en los dos Capítulos que siguen.

2.3 El Gradiente Natural

Los *contrastes* $\psi(\mathbf{y}(t)) = \psi(\mathbf{B}\mathbf{x}(t))$ son, propiamente, funciones de una *matriz* $\mathbf{B}$ de N^2 elementos, no funciones de N^2 variables escalares. Partiendo de esta observación, Amari [Amari98*a*] ha presentado unos resultados muy interesantes sobre la optimización de las funciones *contraste.*

En el Apartado 2.3.1 introducimos el concepto de *distancia entre matrices* que servirá en el Apartado 2.3.2 para dotar de una *métrica* al espacio de las matrices. Conocida la *métrica* es posible determinar con exactitud la expresión del *gradiente* de las funciones contraste, como se verá en el Apartado 2.3.3, induciendo además un algoritmo para su optimización.

2.3.1 Distancia entre matrices

Este Apartado examina los conceptos de *matriz* y de *norma* de una matriz, que, como veremos, serán necesarios para entender los desarrollos posteriores.

Empezaremos notando que el conjunto de todas las matrices *invertibles* N×N tiene la estructura de un *espacio vectorial*, esto es, se verifica que la suma de matrices y el producto de matrices por escalares tienen propiedades bien conocidas (asociativa, conmutativa, etc.) [Noble, pág. 201].

Es razonable preguntarse por la *dimensión* de este espacio. Sobre el conjunto de todas las matrices N×N *invertibles* está bien definida la operación *producto* entre matrices, que tiene la propiedad *asociativa*. Esta operación es "suave" (diferenciable al menos dos veces) porque cada elemento de la matriz producto **A B** es un polinomio de segundo grado en los coeficientes de **A** y **B**. Además, cada matriz **A** tiene un *elemento inverso* $\mathbf{A}^{-1}$ respecto a la operación *producto*, siendo la matriz identidad **I** el *elemento neutro*. Un conjunto con las propiedades descritas recibe el nombre de *grupo de Lie*. Por lo tanto, el conjunto de las matrices N×N invertibles forma también un *grupo de Lie*. En base a esta observación, se demuestra que la dimensión de este grupo y, por extensión, del espacio de las matrices es precisamente N^2 [Mishchenko88, pág. 217].

Definamos, en principio, el *producto escalar* o *interno* entre dos matrices **A** y **B** de dimensiones N×N como:

$$< \mathbf{A} \mid \mathbf{B} > = \text{Traza}(\mathbf{A}\, \mathbf{B}^{T}) \tag{2.7}$$

Sea $\mathbf{E}_{ij}$ la matriz N×N cuyos elementos son todos iguales a cero, exceptuando al que se encuentra en la intersección de la i-ésima fila y la j-ésima columna, que vale *uno*. Puesto que podemos escribir

$$\mathbf{A} = \Sigma_{i,j}\, a_{ij}\, \mathbf{E}_{ij}\,, \qquad\qquad \mathbf{B} = \Sigma_{i,j}\, b_{ij}\, \mathbf{E}_{ij}$$

resulta que el conjunto de las matrices $\mathbf{E}_{ij}$ es una base del espacio, a la que llamaremos $\mathfrak{E}$, esto es, $\mathfrak{E} = \{ \mathbf{E}_{ij} \}$. De hecho, utilizando la definición (2.7) de producto *interno*, $\mathfrak{E}$ es una base ortonormal. Como se verifica que

$$< \mathbf{A} \mid \mathbf{B} > = \text{Traza}(\mathbf{A}\, \mathbf{B}^{T}) = \Sigma_{i,j}\, a_{ij}\, b_{ij}\,,$$

reconocemos de inmediato que (2.7) es el *producto escalar* de las matrices cuando sus coordenadas se expresan en la base ortonormal $\mathbf{E}_{ij}$. Además, cuando

identificamos la matriz **A** con el vector $\mathbf{A} = \Sigma_{i,j}\ a_{ij}\ \mathbf{E}_{ij}$, podemos asociar a cada matriz la *norma euclídea* del correspondiente vector: $< \mathbf{A} \mid \mathbf{A} > = \| \mathbf{A} \|^2_{Fro}$, siendo precisamente $\| \mathbf{A} \|^2_{Fro}$ la *norma de Frobenius* de **A**.

Dadas dos matrices **A** y **B** diremos están separadas por una *distancia* igual a la norma de su diferencia **A** − **B**, que vale: $< \mathbf{A} - \mathbf{B} \mid \mathbf{A} - \mathbf{B} > = \Sigma_{i,j}(a_{ij} - b_{ij})^2$.

Definida la *distancia*, el concepto de *métrica* aparece de forma muy natural: sea $\mathfrak{G} = \{ \mathbf{G}_{ij} \}$ una base del espacio distinta de $\mathfrak{E}$.. El hecho de que $\mathfrak{G}$ sea una base permite escribir

$$\mathbf{A} = \Sigma_{i,j}\ \alpha_{ij}\ \mathbf{G}_{ij}\ , \qquad\qquad \mathbf{B} = \Sigma_{i,j}\ \beta_{ij}\ \mathbf{G}_{ij}$$

Ahora bien, en general, la distancia entre **A** y **B** medida en la base $\mathfrak{E}$, es decir, la cantidad $\Sigma_{i,j}(a_{ij} - b_{ij})^2$, no parece coincidir con la distancia en la base $\mathfrak{G}$, que sería, por definición, $\Sigma_{i,j}(\alpha_{ij} - \beta_{ij})^2$. Sin embargo, la *distancia* es *invariante* en un grupo de Lie [Amari98*a*], es decir, no depende de la base o sistema de referencia utilizado. Por todo esto, la definición de *producto escalar* dada en (2.7) no es completamente apropiada. En su lugar, vamos a definir una nueva operación en la base $\mathfrak{G}$

$$< \mathbf{A} \mid \mathbf{B} >_{\mathfrak{G}} = \Sigma_{i,j,l,m}\ g_{ij,lm}\ \alpha_{ij}\ \beta_{lm}$$

siendo las cantidades $g_{ij,lm}$ tales que $< \mathbf{A} \mid \mathbf{B} >_{\mathfrak{G}}$ es *invariante*, de tal forma que

$$< \mathbf{A} \mid \mathbf{B} >_{\mathfrak{G}} = < \mathbf{A} \mid \mathbf{B} >_{\mathfrak{E}}$$

siendo $< \mathbf{A} \mid \mathbf{B} >_{\mathfrak{E}} = \text{Traza}(\ \mathbf{A}\ \mathbf{B}^T\ ) = \Sigma_{i,j}\ a_{ij}\ b_{ij}$, como en (2.7). Por todo ello, se dice que la *distancia* entre las matrices **A** y **B**, expresadas en la base $\mathfrak{G}$, es igual a $< \mathbf{A} - \mathbf{B} \mid \mathbf{A} - \mathbf{B} >_{\mathfrak{G}}$ Evidentemente, los coeficientes $\{\ g_{ij,lm}\ \}$ dependen de $\mathfrak{G}$.

Se dice que, de esta forma, hemos inducido una *métrica* en el espacio de las matrices [Amari98*a*].

2.3.2 Determinación de la métrica

Los coeficientes $g_{ij,lm}$ nos permiten medir distancias en $\mathfrak{G}$. Ahora bien, es necesario determinar su valor. El siguiente desarrollo se debe a Amari [Amari98*a*]:

Sea $d\mathbf{X}$ una ligera perturbación de la matriz identidad $\mathbf{I}$. Suponiendo sin perder generalidad que $d\mathbf{X}$ está expresado en la base $\mathfrak{E}$, la distancia que separa $\mathbf{I}$ de la matriz perturbada $\mathbf{I} + d\mathbf{X}$ es

$$< d\mathbf{X} \mid d\mathbf{X} >_{\mathfrak{E}} = \Sigma_{i,j}(d\mathbf{X}_{ij})^2 = \text{Traza}(d\mathbf{X}\, d\mathbf{X}^{\mathrm{T}}) \tag{2.8}$$

La siguiente transformación convierte a $d\mathbf{X}$ en una perturbación que puede ser aplicada a cualquier matriz $\mathbf{B}$:

$$(\mathbf{I} + d\mathbf{X})\mathbf{B} = \mathbf{B} + d\mathbf{B} \tag{2.9}$$

Ahora bien, consideramos que multiplicar por $\mathbf{B}$ hace que cambie la base respecto a la que se expresa la perturbación. Es decir, $d\mathbf{B} = d\mathbf{X}\, \mathbf{B}$ son las coordenadas de $d\mathbf{X}$ en otra base a la que llamaremos $\mathfrak{B}$, por ser "natural" a $\mathbf{B}$.

Cualquier cambio de base debe conservar las *distancias*, por lo que

$$< d\mathbf{B} \mid d\mathbf{B} >_{\mathfrak{B}} = < d\mathbf{X} \mid d\mathbf{X} >_{\mathfrak{E}} \tag{2.10}$$

Como $d\mathbf{X} = d\mathbf{B}\, \mathbf{B}^{-1}$, resulta que, de (2.8),

$$< d\mathbf{X} \mid d\mathbf{X} >_{\mathfrak{E}} = \text{Traza}(d\mathbf{B}\, \mathbf{B}^{-1} (\mathbf{B}^{-1})^{\mathrm{T}}\, d\mathbf{B}^{\mathrm{T}}) \tag{2.11}$$

pero, por otra parte, $< d\mathbf{B} \mid d\mathbf{B} >_{\mathfrak{B}} = \Sigma_{i,j,l,m}\, g_{ij,lm}\, d\mathbf{B}_{ij}\, d\mathbf{B}_{lm}$ por lo que, comparando esta expresión con (2.11) se deduce que [Amari98*a*]:

$$g_{ij,kl}(\mathfrak{B}) = \Sigma_m \delta_{ik} \mathbf{B}^{-1}{}_{jm} \mathbf{B}^{-1}{}_{lm} \tag{2.12}$$

siendo δ_{ik} la delta de Kronecker.

2.3.3 El gradiente natural

Sea $\psi(\mathbf{B})$ una función *contraste*. Supondremos que $\psi(\mathbf{B})$ es una función real y diferenciable de $\mathbf{B}$. Entonces, dadas dos matrices $\mathbf{B}_o$ y $\mathbf{B}_o'$ muy próximas, tales que $\mathbf{B}_o' = \mathbf{B}_o + \Delta\mathbf{B} \approx \mathbf{B}_o$, podemos admitir el siguiente desarrollo de Taylor:

$$\psi(\mathbf{B}_o + \Delta\mathbf{B}) = \psi(\mathbf{B}_o) + \langle \nabla_{\mathbf{B}}\psi \mid \Delta\mathbf{B} \rangle \tag{2.13}$$

siendo $\langle \nabla_{\mathbf{B}}\psi \mid \Delta\mathbf{B} \rangle = \text{Traza}(\nabla_{\mathbf{B}}\psi \ \Delta\mathbf{B}^T)$. Por su parte, $\nabla_{\mathbf{B}}\psi$ fue definida en (2.6) como la matriz cuya componente (i, j) vale $\frac{\partial}{\partial b_{ij}}\psi(\mathbf{B})$

Si $\Delta\mathbf{B} = \mu \nabla_{\mathbf{B}}\psi$, donde μ es positivo y pequeño, a fin de que se siga verificando que $\mathbf{B}_o' \approx \mathbf{B}_o$, resulta al sustituir en (2.13)

$$\Delta\psi = \psi(\mathbf{B}_o + \Delta\mathbf{B}) - \psi(\mathbf{B}_o) = \mu \| \nabla_{\mathbf{B}}\psi \|^2_{\text{Fro}} \geq 0 \tag{2.14}$$

siendo $\| \nabla_{\mathbf{B}}\psi \|^2_{\text{Fro}} = \langle \nabla_{\mathbf{B}}\psi \mid \nabla_{\mathbf{B}}\psi \rangle$ la norma de *Frobenius* de $\nabla_{\mathbf{B}}\psi$. Por lo tanto, se deduce de (2.14) que $\psi(\mathbf{B}_o + \Delta\mathbf{B}) \geq \psi(\mathbf{B}_o)$ y el *contraste* crece. En realidad, la elección $\Delta\mathbf{B} = \mu \nabla_{\mathbf{B}}\psi$ no es sino, en apariencia, la expresión que tiene el tradicional algoritmo del gradiente. Sin embargo, *el incremento de* $\mathbf{B}_o$ *no se ha llevado a cabo, necesariamente, en la dirección de mayor crecimiento de* ψ [Amari98*a*] porque, a fin de cuentas, $\| \nabla_{\mathbf{B}}\psi \|^2_{\text{Fro}}$ es una *longitud*, la del vector gradiente, para cuya medida no se está utilizando la *métrica* propia de la matriz $\mathbf{B}_o$.

Ejemplo 2.7. Sea $f(x, y)$ una función de dos variables expresada en las coordenadas cartesianas x e y. El gradiente de $f(x, y)$ es el vector:

$$\nabla f(x,y) = \frac{\partial f(x,y)}{\partial x}\mathbf{i} + \frac{\partial f(x,y)}{\partial y}\mathbf{j}$$

y, como es bien sabido, este gradiente apunta en la dirección de mayor cambio de $f(x, y)$. Hagamos un cambio de las coordenadas, pasando de las cartesianas x e y a coordenadas polares. En estas nuevas coordenadas, el *gradiente*

$$\nabla f(r,\theta) \neq \frac{\partial f(r,\theta)}{\partial r}\mathbf{r} + \frac{\partial f(r,\theta)}{\partial \theta}\theta$$

sino que, por el contrario,

$$\nabla f(r,\theta) = \frac{\partial f(r,\theta)}{\partial r}\mathbf{r} + \frac{1}{r}\frac{\partial f(r,\theta)}{\partial \theta}\theta$$

siendo $\mathbf{r}$ y θ los vectores unitarios en el nuevo sistema de referencia. Resulta que la expresión correcta del gradiente en coordenadas polares *sí* tiene en cuenta la *métrica*. Pues bien, esta misma discusión pretendemos plantearla en el campo de las *funciones contraste*.

Amari llama *natural* al gradiente que se calcula de acuerdo con la métrica (aunque, técnicamente, se prefiere decir que el *gradiente* está expresado *en coordenadas contravariantes* [Mishchenko88]).

Como ya hemos hecho notar, es un resultado bien establecido que el *gradiente natural* apunta en la dirección de mayor cambio de la función, independientemente del sistema de coordenadas que se elija.

Conocidos los coeficientes $g_{ij,kl}(\mathfrak{B})$ es posible determinar el gradiente *natural* sin ninguna dificultad. El desarrollo se puede encontrar en cualquier texto de *geometría diferencial* ([Mishchenko88]):

Teorema 2.1 [Amari98*a*] El *gradiente natural* calculado en $\mathbf{B}_0$ vale

$$\nabla_{\mathbf{B}}^{\text{nat}} \psi = (\nabla_{\mathbf{B}} \psi) \mathbf{B}_0^{\mathrm{T}} \mathbf{B}_0$$

siendo $\nabla_{\mathbf{B}} \psi$ la matriz $N \times N$ cuyo elemento (i, j) es $\dfrac{\partial}{\partial b_{ij}} \psi(\mathbf{B})$.

Demostración. El siguiente razonamiento se debe a Te-Won Lee [Te-Won98] y ha sido seleccionado por su sencillez. No obstante, también remitimos al lector al escrito original de Amari [Amari98*a*]. El desarrollo de Taylor de la función *contraste* nos lleva a la siguiente aproximación:

$$\psi(\mathbf{B}_0 + \Delta\mathbf{B}) = \psi(\mathbf{B}_0) + \langle \nabla_{\mathbf{B}} \psi \mid \Delta\mathbf{B} \rangle_{\mathfrak{E}} \qquad \text{(T2.1.1)}$$

siendo $\langle \nabla_{\mathbf{B}} \psi \mid \Delta\mathbf{B} \rangle_{\mathfrak{E}} = \langle \nabla_{\mathbf{B}} \psi \mid \Delta\mathbf{B} \rangle$ como en (2.7). Impongamos que

$$\langle \nabla_{\mathbf{B}}^{\text{nat}} \psi \mid \Delta\mathbf{B} \rangle_{\mathfrak{B}} = \langle \nabla_{\mathbf{B}} \psi \mid \Delta\mathbf{B} \rangle_{\mathfrak{E}} \qquad \text{(T2.1.2)}$$

Para cambiar las coordenadas de $\nabla_{\mathbf{B}}^{\text{nat}} \psi$ y $\Delta\mathbf{B}$ desde la base $\mathfrak{E}$ a la base $\mathfrak{B}$, basta con multiplicar ambas matrices por $\mathbf{B}^{-1}$ desde la derecha. Entonces,

$$\langle \nabla_{\mathbf{B}}^{\text{nat}} \psi \mid \Delta\mathbf{B} \rangle_{\mathfrak{B}} = \langle \nabla_{\mathbf{B}}^{\text{nat}} \psi \, \mathbf{B}^{-1} \mid \Delta\mathbf{B} \, \mathbf{B}^{-1} \rangle_{\mathfrak{E}} \qquad \text{(T2.1.3)}$$

Por lo tanto, comparando (T2.1.2) y (T2.1.3) resulta que

$$<\nabla_{\mathbf{B}}^{nat}\psi\,\mathbf{B}^{-1}\,|\,\Delta\mathbf{B}\,\mathbf{B}^{-1}>_{\mathfrak{E}} = <\nabla_{\mathbf{B}}\psi\,|\,\Delta\mathbf{B}>_{\mathfrak{E}}$$

De donde, tras varias operaciones, se obtiene el resultado final:

$$\nabla_{\mathbf{B}}^{nat}\,\psi = (\,\nabla_{\mathbf{B}}\psi\,)\,\mathbf{B}_o^{T}\,\mathbf{B}_o.$$

Corolario 2.1. El *algoritmo del gradiente* para las *funciones contraste* ψ tiene la expresión:

$$\Delta\mathbf{B} = \mu\,\nabla_{\mathbf{B}}^{nat}\,\psi \qquad (2.15)$$

siendo μ pequeño y positivo, si se pretende maximizar ψ.

El algoritmo (2.15) utiliza el gradiente correcto, adecuado a la matriz **B**.

La regla de adaptación (2.15) también ha sido presentada, de manera independiente, por Cardoso y Laheld [Cardoso96*a*]. No obstante, Cardoso y Laheld no llegan a ella mediante un razonamiento tan formal como el de Amari, aunque su punto de vista es muy interesante. Ellos proponen el algoritmo

$$\Delta\mathbf{B} = -\,\mu\,\mathbf{F}(\,\mathbf{x}(t);\,\mathbf{B}\,)\,\mathbf{B} \qquad (2.16)$$

siendo $\mathbf{F}(\mathbf{x}(t); \mathbf{B})$ una función de estimación. La *regla* (2.16) y el algoritmo del gradiente natural (2.15) son equivalentes si se toma $\mathbf{F}(\mathbf{x}(t); \mathbf{B}) = -(\nabla_{\mathbf{B}}\psi)\,\mathbf{B}^{\mathrm{T}}$. La matriz $(\nabla_{\mathbf{B}}\psi)\,\mathbf{B}^{\mathrm{T}}$ recibe el nombre de "gradiente relativo" en el trabajo de Cardoso y Laheld [Cardoso96*a*].

Si $\mathbf{F}$ se puede expresar como una función exclusivamente de las variables de salida $\mathbf{y}(t) = \mathbf{B}\,\mathbf{x}(t) = \mathbf{G}\,\mathbf{s}(t)$, siendo esta matriz $\mathbf{G} = \mathbf{B}\,\mathbf{A}$ *la matriz global de transferencia del sistema*, es decir,

$$\mathbf{F}(\mathbf{x}(t); \mathbf{B}) = \mathbf{F}(\mathbf{y}(t)) = \mathbf{F}(\mathbf{G}\,\mathbf{s}(t))$$

entonces, multiplicando (2.16) por la matriz $\mathbf{A}$, se obtiene la regla equivalente:

$$\Delta\mathbf{G} = -\mu\,\mathbf{F}(\mathbf{G}\,\mathbf{s}(t))\,\mathbf{G} \tag{2.17}$$

La ley de adaptación (2.17) implica que la evolución de la matriz $\mathbf{G}$ es *independiente* de la matriz de mezcla. En efecto, las matrices de mezcla $\mathbf{A}$ y $\mathbf{A}'$ darán lugar a la misma matriz $\mathbf{G}$ si, en algún momento, se ven multiplicadas por matrices $\mathbf{B}$ y $\mathbf{B}'$ tales que $\mathbf{B}\,\mathbf{A} = \mathbf{B}'\mathbf{A}'$. Por ello, se dice que la *adaptación* (2.16) (y, por ende, la (2.15)) presenta una *convergencia uniforme* para todas las matrices de mezcla [Cardoso96*a*]. Esto da pié al siguiente resultado [Cardoso96*a*]:

> **Corolario 2.2.** La propiedad de *convergencia uniforme* es la extensión natural del concepto de *equivarianza* a un algoritmo adaptativo.

2.4 Eficiencia de los estimadores

Nuestra aproximación al problema de la Separación de Fuentes está basada en las llamadas *funciones de estimación*. Dada la *función de estimación* $\mathbf{F}(\mathbf{x}(t), \mathbf{B})$, se

cumple siempre, por definición, que $E[\mathbf{F}(\mathbf{x}(t), \mathbf{B})] = \mathbf{0}$ si $\mathbf{B}$ es una matriz que separa las fuentes, donde la esperanza matemática se calcula respecto de $\mathbf{x}(t)$. De aquí se deduce que la estimación de $\mathbf{B}$ no es *sesgada*. Complementariamente, se desea conocer la *varianza* del error de *estimación* y a ello dedicaremos la presente Sección.

En el Apartado 2.4.1 determinaremos genéricamente el orden de magnitud que tiene la varianza de las estimaciones de $\mathbf{B}$. Este estudio servirá para caracterizar en el Apartado 2.4.2 a las funciones de estimación más eficientes, esto es, de menor *varianza*. Finalmente, en los Apartados 2.4.3 y 2.4.4 se presentarán condiciones para que dicha varianza sea muy pequeña en, respectivamente, estimadores de bloque y adaptativos.

Además de las hipótesis que se hicieron en la introducción del Capítulo, se supone que:

- Las variables aleatorias $s_i(t)$ tienen varianza unidad.
- Además, $s_i(t)$ y $s_i(t')$ serán variables aleatorias *independientes* e *idénticamente distribuidas* (*i.i.d*) si $t \neq t'$.

La segunda hipótesis es muy restrictiva; pero necesaria para calcular con facilidad las propiedades asintóticas de la *funciones de estimación*.

2.4.1 Determinación de la varianza de las estimaciones

Sea $\mathbf{X}_T = \{\mathbf{x}(1), ..., \mathbf{x}(T)\}$ la muestra disponible de T observaciones vectoriales. De las hipótesis se deduce que $\mathbf{x}(1)$, ...y $\mathbf{x}(T)$ son muestras de variables aleatorias vectoriales *independientes* e *idénticamente distribuidas*, como se pretendía.

Sea $\mathbf{B}_T$ el estimador de $\mathbf{B}$, la inversa de la matriz de mezcla, que se obtiene a partir de $\mathbf{X}_T$ y $\mathbf{F}(\mathbf{x}(t), \mathbf{B})$, la función de estimación. Definimos el *error relativo* de la estimación como [Amari98*b*]:

$$\Delta\mathbf{B}^{rel} = \Delta\mathbf{B}_T \ \mathbf{B}^{-1} \tag{2.18}$$

donde $\Delta\mathbf{B}_T = \mathbf{B}_T - \mathbf{B}$ es un error absoluto. Por definición, la matriz $\mathbf{B}_T$ es aquella que satisface la ecuación de estimación (ver la ecuación (2.2)):

$$\sum_{t=1}^{T} \mathbf{F}(\mathbf{x}(t), \mathbf{B}_T) = \sum_{t=1}^{T} \mathbf{F}(\mathbf{x}(t), \mathbf{B} + \Delta\mathbf{B}^{rel}\mathbf{B}) = \mathbf{0} \tag{2.19}$$

siendo $\mathbf{B} + \Delta\mathbf{B}^{rel}\mathbf{B} = \mathbf{B}_T$. Si T es suficientemente grande, admitiremos que, en efecto, $\mathbf{B}_T \approx \mathbf{B} = \mathbf{A}^{-1}$, obviando con ello que las ecuaciones de estimación pueden tener otras soluciones, quizás no deseadas. Por ello, es posible hacer el siguiente desarrollo en serie de Taylor de (2.19):

$$\Sigma_t \mathbf{F}(\mathbf{x}(t), \mathbf{B}) + \Delta\mathbf{F}(\mathbf{x}(t), \mathbf{B}) = \mathbf{0} \tag{2.20}$$

donde $\Delta\mathbf{F}$ (omitimos la dependencia explícita con $\mathbf{x}(t)$ y $\mathbf{B}$ por simplicidad) es la matriz cuyo elemento (a, b) vale

$$[\,\Delta\mathbf{F}\,]_{ab} = \sum_{c,\,d} \frac{\partial f_{ab}}{\partial b_{cd}} [\,\Delta\mathbf{B}\,]_{cd} \tag{2.21}$$

siendo $f_{ab} = [\,\mathbf{F}\,]_{ab}$ y $b_{cd} = [\,\mathbf{B}\,]_{cd}$ (recordemos que $[\,\cdot\,]_{ij}$ denota al elemento (i, j) de la matriz entre corchetes). Ahora vamos a tratar de escribir $\Delta\mathbf{F}$ en función de $\Delta\mathbf{B}^{rel}$. Sea

$$[\,\mathbf{M}\,]_{ab,\,cd} = \frac{\partial f_{ab}}{\partial b_{cd}} \tag{2.22}$$

que, estrictamente, no denota las componentes de ninguna matriz; pero conservaremos esta notación. Como $\Delta\mathbf{B} = \Delta\mathbf{B}^{\text{rel}}\,\mathbf{B}$ implica, componente a componente, que:

$$[\,\Delta\mathbf{B}\,]_{cd} = \Sigma_i\,[\,\Delta\mathbf{B}^{\text{rel}}\,]_{ci}\,b_{id} \tag{2.23}$$

resulta, llevando (2.22) y (2.23) a (2.21),

$$[\,\Delta\mathbf{F}\,]_{ab} = \sum_{c,\,d}[\,\mathbf{M}\,]_{ab,\,cd}\,[\,\Delta\mathbf{B}\,]_{cd} = \sum_{c,\,d,\,i}[\,\mathbf{M}\,]_{ab,\,cd}\,b_{id}\,[\,\Delta\mathbf{B}^{\text{rel}}\,]_{ci} =$$

$$= \sum_{c,\,i}[\,\mathbf{D}\,]_{ab,\,ci}\,[\,\Delta\mathbf{B}^{\text{rel}}\,]_{ci} \tag{2.24}$$

que es una expresión que relaciona explícitamente $\Delta\mathbf{F}$ y $\Delta\mathbf{B}^{\text{rel}}$, como se pretendía, siendo

$$[\,\mathbf{D}\,]_{ab,\,ci} = \Sigma_d\,[\,\mathbf{M}\,]_{ab,\,cd}\,b_{id} \tag{2.25}$$

La expresión (2.24) que liga $\Delta\mathbf{F}$ y $\Delta\mathbf{B}^{\text{rel}}$ se escribe de manera simbólica como [Amari98*b*]

$$\Delta\mathbf{F} = \mathfrak{D}\,\Delta\mathbf{B}^{\text{rel}} \tag{2.26}$$

donde $\mathfrak{D}$ es un operador lineal que actúa sobre el espacio de las matrices. Finalmente, llevando (2.26) a (2.20) se obtiene:

$$\Sigma_t\,\mathbf{F}(\,\mathbf{x}(t),\mathbf{B}\,) + \mathfrak{D}\,\Delta\mathbf{B}^{\text{rel}} = \mathbf{0} \tag{2.27}$$

A partir de (2.27) vamos a derivar las propiedades asintóticas de nuestro estimador $\mathbf{F}(\,\mathbf{x}(t),\mathbf{B}\,)$. Para ello, reescribimos (2.27) como sigue

$$\frac{1}{T}\sum_t \mathfrak{D}\,\Delta\mathbf{B}^{rel} = -\frac{1}{\sqrt{T}}\frac{1}{\sqrt{T}}\sum_t \mathbf{F} \qquad (2.28)$$

Como las muestras vectoriales $\mathbf{x}(1)$, ..., $\mathbf{x}(T)$ son independientes entre sí e idénticamente distribuidas por hipótesis, se sabe que para T suficientemente grande [Papoulis91, pág.213], [Amari98*b*]

$$\frac{1}{T}\sum_t \mathfrak{D} = E[\ \mathfrak{D}\ ] + O(\frac{1}{\sqrt{T}}) \qquad (2.29)$$

Por otra parte, *el teorema central del límite* [Papoulis91, pág. 214] nos permite asegurar que

$$\frac{1}{\sqrt{T}}\sum_t \mathbf{F}(\mathbf{x}(t), \mathbf{B})$$

converge en distribución hacia una *matriz* de variables aleatorias de distribución normal, media cero (pues $E[\mathbf{F}] = \mathbf{0}$) y covarianzas dadas por:

$$E[F_{ab}(\mathbf{x}(t), \mathbf{B})\, F_{cd}(\mathbf{x}(t), \mathbf{B})] \qquad (2.30)$$

Es interesante hacer notar que la hipótesis, muy restrictiva, de que las muestras vectoriales $\mathbf{x}(1)$, ..., $\mathbf{x}(T)$ son *independientes* también es necesaria para aplicar el Teorema Central del Límite.

Combinando (2.28) y (2.29) se obtiene [Amari98*b*]:

$$\Delta\mathbf{B}^{rel} = -\frac{1}{\sqrt{T}}E[\ \mathfrak{D}\ ]^{-1}\frac{1}{\sqrt{T}}\sum_t \mathbf{F}(\mathbf{x}(t), \mathbf{B}) \qquad (2.31)$$

siendo $E[\ \mathfrak{D}\ ]^{-1}$ el operador inverso de $E[\ \mathfrak{D}\ ]$. Entonces, $\Delta\mathbf{B}^{rel}$ *es también una matriz de variables aleatorias normales de media cero*, igual que $\sum_t \mathbf{F}(\mathbf{x}(t), \mathbf{B})$

(pues al aplicar cualquier transformación lineal a una matriz Σ **F** de variables aleatorias *gaussianas* su distribución estadística no cambia, siendo esto último característico de los procesos gaussianos [Papoulis91, pág.309]).

Definimos la correlación entre las componentes de $\Delta\mathbf{B}^{rel}$ del mismo modo que en (2.30) se definió la correlación entre las componentes de $\sum_t$ **F**. Esta correlación viene dada por el siguiente resultado [Amari98*b*]:

Lema 2.1 (*Varianza del Error en la Matriz de Mezcla*) Sean

$$\Delta b_{ab}^{rel} \quad \text{y} \quad \Delta b_{cd}^{rel},$$

respectivamente, los elementos (a, b) y (c, d) de la matriz $\Delta\mathbf{B}^{rel}$. Entonces

$$\mathrm{E}[\, \Delta b_{ab}^{rel} \Delta b_{cd}^{rel} \,] = \frac{1}{\mathrm{T}} \sigma_{ab,cd}^2 + \mathrm{O}(\frac{1}{\mathrm{T}^2})$$

siendo $\sigma_{ab,cd}^2$ un parámetro que depende de la distribución estadística de las observaciones y de la función de estimación; *pero no del número de muestras* T.

Es decir, aunque la varianza depende del propio estimador a través de $\sigma_{ab,cd}^2$ en general siempre podemos esperar que sea inversamente proporcional al número T de muestras. La *prueba* del Lema 2.1 es inmediata a partir de (2.31) [Amari98*b*].

La salida del sistema es igual a

$$\mathbf{y}(t) = \mathbf{B}_{\mathrm{T}}\, \mathbf{x}(t) = (\, \mathbf{B} + \Delta\mathbf{B}_{\mathrm{T}} \,)\, \mathbf{x}(t) \qquad (2.32)$$

Como $\mathbf{B} = \mathbf{A}^{-1}$ y $\Delta\mathbf{B}_T = \Delta\mathbf{B}^{rel}\,\mathbf{B}$, resulta que

$$\mathbf{y}(t) = \mathbf{s}(t) + \Delta\mathbf{s}(t) \tag{2.33}$$

siendo $\Delta\mathbf{s}(t) = \Delta\mathbf{B}^{rel}\,\mathbf{s}(t)$. Se puede demostrar [Amari98*b*] lo siguiente:

Lema 2.2. (*Correlación cruzada de las salidas*) Para $i \neq j$ se verifica que:

$$\mathrm{E}[\,\Delta s_i(t)\Delta s_j(t)] = \sum_k \mathrm{E}[\,\Delta b_{ik}^{rel}\,\Delta b_{jk}^{rel}\,]\,\mathrm{E}[s_k^2(t)]$$

Si $i \neq j$, no es difícil probar que $\mathrm{E}[\,\Delta s_i(t)\,\Delta s_j(t)\,] = \mathrm{E}[\,y_i(t)y_j(t)]$, la correlación entre las señales de salida, que, idealmente, debería ser cero. De acuerdo al Lema 2.1, resulta que

$$\mathrm{E}[\,\Delta s_i(t)\Delta s_j(t)] = \mathrm{O}(\frac{1}{\mathrm{T}})$$

Estos resultados son clásicos; pero poco operativos. Sin embargo, ayudan a entender propiedades fundamentales del problema de Separación de Fuentes.

2.4.2 Funciones de Estimación Eficientes

El *estimador de máxima verosimilitud* (MLE) es asintóticamente *eficiente*, esto es, alcanza la cota de Cramér-Rao (ver Apéndice C). Sin embargo, para formular

las ecuaciones del MLE es necesario conocer la función de densidad de probabilidad de las señales fuente. A ello se refiere el siguiente Teorema:

Teorema 2.2 (*El Estimador de Máxima Verosimilitud*) Sea q_i la *función de densidad de probabilidad marginal* de la i-ésima fuente $s_i(t)$. Definamos $\varphi_i(y_i(t)) = \frac{\partial}{\partial y_i} \log q_i(y_i(t))$. Entonces

$$\mathbf{F}(\mathbf{x}(t) ; \mathbf{B}) = \varphi(\mathbf{y}(t)) \mathbf{y}^T(t) - \mathbf{I},$$

donde $\varphi(\mathbf{y}(t)) = [\varphi_1(y_1(t)), \varphi_2(y_2(t)), ..., \varphi_N(y_N(t))]^T$, es la *función de estimación* asociada al *estimador de máxima verosimilitud* de la matriz de mezcla.

Demostración. Se deja la prueba para el siguiente Capítulo.

No obstante, la *f.d.p marginal* de las fuentes es, por definición, desconocida. De todas maneras, las funciones de estimación *no sesgadas* que son de la forma

$$\mathbf{F}(\mathbf{x}(t) ; \mathbf{B}) = \varphi(\mathbf{y}(t)) \mathbf{y}^T(t) - \mathbf{I} \tag{2.34}$$

presentan *siempre* unas características excelentes en cuanto a la varianza de sus estimaciones, incluso aunque la elección de la función vectorial φ sea arbitraria. Ello se va a establecer por medio del Teorema 2.3, que se presenta a continuación. Antes, nótese que para (2.34)

$$E[\mathbf{F}(\mathbf{x}(t) ; \mathbf{B})] = \mathbf{0}$$

equivale a

$$E[\, \varphi(\, \mathbf{y}(t)\,)\, \mathbf{y}^T(t)\,] = \mathbf{I}, \tag{2.35}$$

lo que se interpreta como que $\mathbf{y}(t)$ y la variable aleatoria vectorial $\varphi(\, \mathbf{y}(t)\,)$ están incorreladas. En particular, si las componentes de $\mathbf{y}(t)$ son *independientes* y están convenientemente escaladas, se satisface (2.35).

Si $\varphi(\, \mathbf{y}(t)\,) = [y_1(t), y_2(t), ...,y_N(t)]^T$, entonces (2.35) sólo fuerza a que las componentes de $\mathbf{y}(t)$ están incorreladas; pero no que a sean independientes, así que no basta para separar las fuentes: en general, φ debe ser una función no lineal de $\mathbf{y}(t)$.

Sea $\mathfrak{X}$ el espacio de todas las funciones que son de la forma $\varphi_i(y_i)\, y_j$ (para i distinto de j), incluyendo sus combinaciones lineales. Nótese que los elementos no diagonales de la *función de estimación* (2.34) pertenecen a este espacio. Cualquier función escalar $g(\, \mathbf{y}(t)\,)$ puede ser proyectada en el espacio $\mathfrak{X}$, es decir, siempre se puede escribir $g(\, \mathbf{y}(t)\,)$ como una combinación lineal de funciones pertenecientes a $\mathfrak{X}$ más un resto que se dice ortogonal a este espacio. Los coeficientes de este desarrollo se computan a partir del producto escalar de $g(\, \mathbf{y}(t)\,)$ por las funciones base de $\mathfrak{X}$, donde el *producto escalar* entre dos funciones cualesquiera $w_1(t)$ y $w_2(t)$ se define como [Amari97*a*]

$$< w_1(t), w_2(t) > = E[\, w_1(t)\, w_2(t)\,]$$

Pues bien,

Teorema 2.3. (*Eficiencia de las Funciones de Estimación*) En el problema de la Separación de Fuentes, se verifica que

a) La proyección de cualquier *función de estimación* sobre $\mathfrak{X}$ también es una *función de estimación.*
b) Cualquier *función de estimación* tiene siempre *mayor* varianza que su proyección sobre $\mathfrak{X}$.

Demostración. La prueba es compleja y aporta poco a esta exposición. Por ello, remitimos al lector al artículo de Amari y Cardoso [Amari97*a*].

Este teorema es clave. Consideremos las *funciones de estimación* que se construyen a base de términos de la forma: $\varphi_i(y_i)\phi_j(y_j)$, $\varphi_i(y_i)\phi_j(y_j)\lambda_k(y_k)$, etc. Por ejemplo, sea

$$\mathbf{F}(\,\mathbf{x}(t)\,;\,\mathbf{B}\,) = \varphi(\,\mathbf{y}(t)\,)\psi(\,\mathbf{y}(t)\,)^{\mathrm{T}} - \mathbf{I},$$

siendo $\varphi(\,\mathbf{y}(t)\,)$ y $\psi(\,\mathbf{y}(t)\,)$ vectores de $N \times 1$ elementos, es decir, imponiendo que

$$E[\,\mathbf{F}(\,\mathbf{x}(t)\,;\,\mathbf{B}\,)\,] = \mathbf{0}$$

se obtendría que

$$E[\, \varphi(\, \mathbf{y}(t)\,)\psi(\, \mathbf{y}^T(t)\,)\,] = \mathbf{I}$$

lo que parece ser una condición más fuerte que (2.35). Por ello, aparentemente, estas funciones de estimación son más generales que (2.34). Sin embargo, el Teorema 2.3 afirma que con estas "generalizaciones" tan sólo se consigue incrementar la varianza de la estimación.

Por esta razón, a las funciones de estimación que tienen la forma (2.34) las llamaremos de forma genérica "*eficientes*" en lo que sigue. Esta denominación no es del todo ortodoxa, por cuanto que, generalmente, se entiende que "*eficiente*" es aquél estimador que alcanza la cota de Cramér-Rao. A pesar de todo, resulta natural llamar "*eficientes*" a las funciones (2.34) en el contexto de la Separación de Fuentes [Amari98*b*].

De todas formas, hay que decir que la prueba del Teorema 2.3 se basa en aproximaciones locales de la funciones de estimación. Por lo tanto, las funciones de estimación que no son "eficientes" podrían tener mejores propiedades globales de convergencia (por ejemplo, cuando todas sus raíces sean matrices que separen las fuentes, lo cuál no se ha garantizado para (2.34)). También podrían ser más robustas frente al ruido y, en general, cuando hay menos sensores que fuentes. Ésta es una línea de investigación que permanece abierta.

2.4.3 Estimación Supereficiente

El Lema 2.2 afirma que, en general, una vez estimadas las fuentes

$$E[\, y_i(t)\, y_j(t)\,] = O(\frac{1}{T})$$

para $i \neq j$, siendo T el número de muestras que se han utilizado para estimar la matriz de mezcla. Interesa que $E[\, y_i(t)\, y_j(t)\,]$ sea lo más pequeña posible para $i \neq j$; de hecho, esta correlación debería anularse si $y_i(t)$ e $y_j(t)$ fuesen independientes.

Sea, como en (2.34),

$$\mathbf{F}(\mathbf{x}(t)\,;\mathbf{B}) = \varphi(\mathbf{y}(t))\,\mathbf{y}^{T}(t) - \mathbf{I}$$

donde $\varphi(\mathbf{y}(t)) = [\varphi_1(y_1(t)), \varphi_2(y_2(t)), ...,\varphi_N(y_N(t))]^T$ y supongamos que se satisface la hipótesis

$$E[\,\varphi_i(s_i(t))\,] = 0 \text{ para todo } i$$

lo que ocurre en cualquiera de los siguientes casos [Amari98*b*]:

1.- Que $\varphi_i(s_i) = -\frac{\partial}{\partial s_i}\log q_i(s_i)$, siendo $q_i(s_i)$ la *f.d.p marginal* de la i-ésima fuente. De hecho, como se vio en el Teorema 2.2, así se obtiene el *estimador de máxima verosimilitud*. Por desgracia, $q_i(s_i)$ es desconocida y es difícil escoger φ_i para que cumpla este criterio.

2.- La *f.d.p* de las fuentes es una función *par* mientras que φ_i es una función *impar*. Éste es un caso que se puede suponer frecuente en la práctica.

Entonces, la Función de Estimación (2.34) es *supereficiente* [Amari98*b*], esto es,

$$E[\,y_i(t)\,y_j(t)\,] = O(\frac{1}{T^2})$$

para $i \neq j$. Esto explica los excelentes resultados que suelen tener los algoritmos de Separación de Fuentes basados en las funciones de estimación de la forma (2.34). Nótese que, en otros contextos, se llama *supereficientes* a aquéllos estimadores *sesgados* cuya varianza es menor que la del estimador de máxima

verosimilitud (MLE) [Stoica96] (esto no es ninguna contradicción: la cota de Cramér-Rao para la varianza de los estimadores *sesgados* no es la misma (quizás es menor) que la cota para la varianza de los estimadores *no sesgados*, que es alcanzada por el MLE [Stoica96, Johnson93, pág. 279]).

2.4.4 Supereficiencia de los Algoritmos Adaptativos

Los resultados anteriores se aplican sólo a algoritmos de *bloque*, es decir, a aquéllos que resuelven *ecuaciones de estimación* como (2.2). Sea, como antes en (2.34),

$$\mathbf{F}(\,\mathbf{x}(t)\,;\,\mathbf{B}\,) = \varphi(\,\mathbf{y}(t)\,)\,\mathbf{y}^{T}(t) - \mathbf{I}$$

donde $\varphi(\,\mathbf{y}(t)\,) = [\varphi_1(y_1(t)),\ \varphi_2(y_2(t)),\ ...,\varphi_N(y_N(t))]^T$. Al aplicar el algoritmo de adaptación (2.16) de Cardoso y Laheld para buscar las raíces de $\mathbf{F}(\,\mathbf{x}(t)\,;\,\mathbf{B}\,)$ se obtiene

$$\Delta\mathbf{B} = -\,\mu\,\mathbf{F}(\,\mathbf{x}(t)\,;\,\mathbf{B}\,)\,\mathbf{B} = \mu\,(\,\mathbf{I} - \varphi(\,\mathbf{y}(t)\,)\,\mathbf{y}^{T}(t)\,)\,\mathbf{B} \qquad (2.36)$$

donde μ es una constante positiva y muy pequeña. Si se cumple la condición de *supereficiencia*, esto es, $E[\ \varphi_i(s_i(t))\ ] = 0$ para todo i, tras iterar un número suficientemente grande de veces se tiene que

$$E[\ y_i(t)\,y_j(t)\ ] = O(\,\mu^2\,)$$

para $i \neq j$ [Amari98*b*]. El resultado es válido siempre que el algoritmo converja a una matriz separe las fuentes, lo cuál no se puede garantizar *a priori.*

2.5 Conclusiones

A lo largo de este Capítulo, hemos presentado los fundamentos de la estimación de la matriz de mezcla. En realidad, se han aplicado las técnicas clásicas de Inferencia Estadística al problema de la Separación de Fuentes, aunque empleando una terminología y casuística propia. En particular, se ha tratado con detalle el problema de la *eficiencia* de las Funciones de Estimación, entendiéndose que un estimador es *más eficiente* que otro si *su varianza es menor* y se ha mostrado la forma que deben tener los estimadores de varianza mínima (entre los que se encuentra el estimador de máxima verosimilitud). Sin embargo, no hemos podido afirmar que estos estimadores *siempre* obtengan una matriz de separación. De hecho, como se mostrará con detalle en el siguiente Capítulo, es muy probable que las Funciones de Estimación *eficientes* tengan raíces que *no* consiguen la Separación de las Fuentes.

3. El Estimador de Máxima Verosimilitud

3.1 Introducción

Dentro de la teoría clásica de la estimación de parámetros, el *estimador de máxima verosimilitud* (MLE, acrónimo de 'Maximum likelihood estimator') es el estimador *no sesgado* que goza de mejores propiedades, en el sentido de que su varianza satisface asintóticamente la cota de Cramér-Rao (ver Apéndice C). En este Capítulo vamos a plantear las ecuaciones del MLE de la matriz de mezcla y a discutir su utilidad y prestaciones.

En la Sección 3.2 se presenta el MLE y su interpretación en el contexto de la Separación de Fuentes. Después, en §3.3 se relaciona el MLE con los algoritmos basados en el "Principio de Maximización de la Información" (Infomax), que ha ganado gran popularidad en los últimos años. La Sección 3.4 se dedica a la discusión sobre el MLE. Finalmente, en §3.5 se presentan nuestras conclusiones.

Salvo indicación en contra, se admiten las siguientes *hipótesis*:

- La matriz de mezcla **A** es *invertible*.
- Las señales fuente son realizaciones de procesos aleatorios *reales* y *estacionarios*, siendo *estadísticamente independientes* entre sí y de media *cero*.
- La varianza de cada fuente es igual a *uno* en todo instante.

La primera y segunda suposiciones son las habituales. La tercera hipótesis es una simple convención que normaliza las fuentes, sin mayor trascendencia.

Mantenemos la notación del Capítulo anterior: se emplea la cursiva para las cantidades escalares, la negrita minúscula para los vectores y la negrita mayúscula para las matrices, sin perjuicio de que, a su vez, $s_i(t)$ o $x_i(t)$ (respectivamente $\mathbf{s}(t)$ o $\mathbf{x}(t)$) denoten realizaciones de un proceso aleatorio y, para un instante t particular, sean además variables aleatorias escalares (respectivamente vectoriales).

3.2 El estimador de Máxima verosimilitud

En 3.2.1 presentaremos las ecuaciones que tiene el estimador en sí, para después, en 3.2.2, estudiar la variante de que la matriz de separación sea ortogonal. Finalmente, en 3.2.3 se definirá la llamada "distancia de Kullback-Leibler", que servirá para dar una interpretación al MLE.

En esta Sección se utilizan las siguientes *hipótesis* de trabajo, que se añaden a las formuladas previamente

- En todo instante t, cada *fuente* $s_i(t)$ es una variable aleatoria *continua.*
- Además, $s_i(t)$ y $s_i(t')$ están idénticamente distribuidas y son *independientes* si t es distinta de t'.

Se dice que una variable aleatoria Z es *continua* si su *función de distribución* F_z es absolutamente continua (es decir, existe una función p_z: $\Re \rightarrow \Re_+$ la *función de densidad de probabilidad* (*f.d.p*) *de* Z, tal que

$$F_z(z) = \text{Prob}(Z \leq z) = \int_{-\infty}^{z} p_z(x)\,dx$$

[Gnedenko75, pág. 132]. Esta definición se supone conocida por el lector; pero sirve para fijar nuestra notación. En contraposición, se dice que una variable aleatoria es *discreta* si los valores que toma forman un conjunto *numerable* [Gnedenko75, pág. 131]. El estimador de máxima verosimilitud de **A** se obtendrá suponiendo que las fuentes para el instante t son *variables aleatorias continuas*. Sólo al final discutiremos la problemática asociada a las fuentes de naturaleza *discreta*.

3.2.1 Formulación de las ecuaciones del estimador

El modelo de mezcla *lineal* e *instantánea* viene dado por la siguiente ecuación, sobre la que centraremos nuestro análisis:

$$\mathbf{x}(t) = \mathbf{A}\,\mathbf{s}(t), \tag{3.1}$$

siendo **A** la matriz N×N de *mezcla*, $\mathbf{s}(t) = [\, s_1(t), s_2(t), \ldots s_N(t) \,]^T$ el vector que recoge las N *fuentes* y $\mathbf{x}(t) = [\, x_1(t), x_2(t), \ldots x_N(t) \,]^T$ el vector de las N *observaciones*. Deseamos estimar la matriz de mezcla **A** a partir de un conjunto dado de T observaciones vectoriales, $\mathbf{X}_T = \{\mathbf{x}(1), \ldots, \mathbf{x}(T)\}$, como paso previo a la Separación. El modelo se representa en la Figura 3.1.

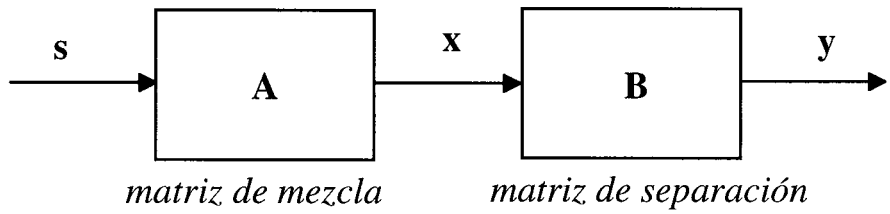

Figura 3.1. Modelo de Mezcla y Sistema de Separación.

Sea $q(\cdot)$ la *f.d.p conjunta* de las fuentes. Se puede demostrar fácilmente que la *f.d.p conjunta* de las observaciones es igual a [Papoulis91, pág.183]

$$p_{\mathbf{x}}(\mathbf{x}(t)\,;\,\mathbf{A}^{-1},\,q) = |\det \mathbf{A}^{-1}|\, q(\mathbf{A}^{-1}\mathbf{x}(t)) \tag{3.2}$$

Como ya hemos apuntado, no utilizamos símbolos distintos para referirnos a la variable aleatoria y al valor instantáneo que toma. Dado que las fuentes $s_i(t)$ son *independientes*, $q(\cdot)$ es igual al producto de sus funciones de densidad de probabilidad marginal, denotadas a su vez por $q_i(s_i(t))$. Es decir,

$$q(\mathbf{s}(t)) = \prod_{i=1}^{N} q_i(s_i(t)) \tag{3.3}$$

Dada una muestra de T observaciones vectoriales *independientes* e *idénticamente distribuidas* $\mathbf{x}(1)$, ..., $\mathbf{x}(T)$, el logaritmo de su función de verosimilitud (ver Apéndice C) toma el valor

$$\begin{aligned} l(\mathbf{B}) &= \frac{1}{T}\sum_{t=1}^{T} \log p_{\mathbf{x}}(\mathbf{x}(t)\,;\,\mathbf{B}\,,\,q) = \\ &= \log|\det \mathbf{B}| + \frac{1}{T}\sum_{t=1}^{T} \log q(\mathbf{B}\,\mathbf{x}(t)) \end{aligned} \tag{3.4}$$

La matriz $\mathbf{B}$ es genérica y no tiene por qué coincidir con $\mathbf{A}^{-1}$; por otra parte, nótese que $l(\mathbf{B})$ es un escalar. La estimación de máxima verosimilitud de $\mathbf{A}^{-1}$ es justamente el valor de $\mathbf{B}$ que maximiza $l(\mathbf{B})$, es decir

$$\mathbf{A}^{-1} = \arg\max_{\mathbf{B}}\; l(\mathbf{B})$$

Por lo tanto, admitiendo que $l(\mathbf{B})$ es diferenciable, el MLE es una solución de la ecuación:

$$\nabla_{\mathbf{B}}\, l(\mathbf{B}) = \mathbf{0} \tag{3.5}$$

siendo $\nabla_{\mathbf{B}}\, l(\mathbf{B})$ la *matriz* cuyo elemento (i, j) vale $\frac{\partial l(\mathbf{B})}{\partial b_{ij}}$, la derivada de $l(\mathbf{B})$ respecto a la componente de $\mathbf{B}$ que está en la posición (i, j). Es posible demostrar que [Bell95]

$$\nabla_{\mathbf{B}} \log |\det \mathbf{B}| = (\mathbf{B}^{\mathrm{T}})^{-1} \tag{3.6}$$

y, por otra parte, de (3.3),

$$\nabla_{\mathbf{B}} \log q(\mathbf{y}(t)) = \sum_{i=1}^{N} \nabla_{\mathbf{B}} \log q_i(y_i(t)) \tag{3.7}$$

donde $\mathbf{y}(t) = \mathbf{B}\,\mathbf{x}(t)$ representa la salida del sistema, como en la Figura 3.1. Por definición, el elemento (k, j) de la matriz $\nabla_{\mathbf{B}} \log q_i(y_i(t))$ es, según (3.3), igual a

$$\frac{\partial}{\partial b_{kj}} \log q_i(y_i) = \frac{1}{q_i(y_i)} \frac{d}{dy_i} q_i(y_i) \frac{\partial}{\partial b_{kj}} y_i \tag{3.8}$$

Dado que $\frac{\partial}{\partial b_{kj}} y_i = \delta_{ki}\, x_j$, donde δ_{ki} denota la delta de Kronecker, resulta

$$\frac{\partial}{\partial b_{kj}} \log q_i(y_i(t)) = -\varphi_i(y_i(t))\, \delta_{ki}\, x_j(t) \tag{3.9}$$

siendo $\varphi_i(y_i) = -\frac{1}{q_i(y_i)} \frac{d}{dy_i} q_i(y_i)$ (*"score function"* en inglés). Combinando todas estas expresiones, se llega al sorprendentemente sencillo resultado:

$$\nabla_{\mathbf{B}} \log q(\mathbf{y}(t)) = -\varphi(\mathbf{y}(t))\, \mathbf{x}^{\mathrm{T}}(t) \tag{3.10}$$

donde $\varphi(\mathbf{y}(t)) = [\, \varphi_1(y_1(t)), \ldots, \varphi_N(y_N(t)) \,]^{\mathrm{T}}$. Utilizando (3.6) y (3.10), obtenemos las ecuaciones del MLE:

$$\nabla_{\mathbf{B}}\, l(\mathbf{B}) = (\mathbf{B}^{\mathrm{T}})^{-1} - \frac{1}{\mathrm{T}} \sum_{t=1}^{\mathrm{T}} \varphi(\mathbf{y}(t))\, \mathbf{x}^{\mathrm{T}}(t) = \mathbf{0} \qquad (3.11a)$$

o bien, considerando que la función de verosimilitud es una *función contraste* y aplicando *la regla del gradiente natural* [Amari98*a*] (ver el Teorema 2.1):

$$\nabla_{\mathbf{B}}^{\mathrm{nat}}\, l(\mathbf{B}) = \nabla_{\mathbf{B}}\, l(\mathbf{B})\, \mathbf{B}^{\mathrm{T}}\mathbf{B} =$$
$$= \{\, \mathbf{I} - \frac{1}{\mathrm{T}} \sum_{t=1}^{\mathrm{T}} \varphi(\mathbf{y}(t))\, \mathbf{y}^{\mathrm{T}}(t) \,\}\, \mathbf{B} = \mathbf{0} \qquad (3.11b)$$

Desde luego, (3.11*a*) y (3.11*b*) tienen exactamente las mismas raíces, por cuanto que $\mathbf{B}^{\mathrm{T}}\mathbf{B}$ es *invertible*. Es más, la función de estimación (2.34), que reproducimos a continuación (el signo '–' no es relevante y aparece sólo por conveniencia de presentación):

$$-\mathbf{F}(\mathbf{x}(t)\,;\mathbf{B}) = \mathbf{I} - \varphi(\mathbf{y}(t))\, \mathbf{y}^{\mathrm{T}}(t)$$

está íntimamente ligada a (3.11*b*) por cuanto que se puede considerar que (3.11*b*) es un estimador de las matrices $\mathbf{B}$ que verifican $\mathrm{E}[\mathbf{F}(\mathbf{x}(t)\,;\mathbf{B})] = \mathbf{0}$. De hecho, ésta es la forma que, según dijimos en el Capítulo anterior, tienen las funciones de estimación *eficientes* (ver el Apartado 2.4.2 y, específicamente, el Teorema 2.2. en la página 57. Ver también el Apartado 2.4.3). También resulta sugerente comparar (3.11*b*) con el algoritmo adaptativo (2.36), que pasamos a recordar:

$$\Delta\mathbf{B} = \mu\, (\mathbf{I} - \varphi(\mathbf{y}(t))\, \mathbf{y}^{\mathrm{T}}(t))\, \mathbf{B}$$

y que también se puede utilizar para determinar el MLE, como se estudiará con detalle en el Capítulo 4. Además, según lo visto en el Apartado 2.4.3 (pág.60), el algoritmo va a ser *supereficiente*.

Ejemplo 3.1. Probemos que el MLE es un estimador *equivariante* [Cardoso95*a*]. La función de verosimilitud de la muestra es igual a

$$\log p_{\mathbf{x}}(\mathbf{X}_T ; \mathbf{B}) = \log \det | \mathbf{B} | + \frac{1}{T} \sum_{t=1}^{T} \log q(\mathbf{B}\, \mathbf{x}(t))$$

La inversa de la matriz de mezcla se estima justamente como el valor de **B** que hace máxima la expresión anterior:

$$\hat{\mathbf{A}}^{-1} = \arg\max_{\mathbf{B}} \log p_{\mathbf{x}}(\mathbf{X}_T ; \mathbf{B})$$

Definamos ahora la función

$$l(\mathbf{S}_T ; \mathbf{C}) = \log \det | \mathbf{C} | + \frac{1}{T} \sum_{t=1}^{T} \log q(\mathbf{C}\, \mathbf{s}(t))$$

donde $\mathbf{S}_T = \{\mathbf{s}(1), ..., \mathbf{s}(T)\}$ y sea $\mathbf{C}_{max}(\mathbf{S}_T) = \arg\max_{\mathbf{C}} l(\mathbf{S}_T ; \mathbf{C})$ la matriz **C** que la maximiza. Podemos comprobar con facilidad que

$$l(\mathbf{S}_T ; \mathbf{B}\, \mathbf{A}) = \log p_{\mathbf{x}}(\mathbf{X}_T ; \mathbf{B}) + \log \det | \mathbf{A} |$$

por lo tanto, $\mathbf{C}_{max}(\mathbf{S}_T) = \hat{\mathbf{A}}^{-1}\, \mathbf{A}$, ya que **A** es un parámetro constante que no tiene efecto en la maximización. Así,

$$\mathbf{s}_{est}(t) = \hat{\mathbf{A}}^{-1}\, \mathbf{x}(t) = \hat{\mathbf{A}}^{-1}\, \mathbf{A}\, \mathbf{s}(t) = \mathbf{C}_{max}(\mathbf{S}_T)\, \mathbf{s}(t)$$

y, por lo tanto, la calidad de la Separación no depende de la matriz de mezcla, sólo de $\mathbf{S}_T$. Ésta es la condición de *equivarianza*, como se quería demostrar.

3.2.2 El MLE Como Contraste Ortogonal

Si suponemos que las observaciones están espacialmente incorreladas y tienen varianza unidad (esto es, $E[x_i(t)\, x_j(t)] = \delta_{ij}$), entonces tanto $\mathbf{A}$ como $\mathbf{B}$ deben ser matrices *ortogonales*, o sea, $\mathbf{A}\,\mathbf{A}^T = \mathbf{B}\,\mathbf{B}^T = \mathbf{I}$. Se deduce que $\det(\mathbf{B}\,\mathbf{B}^T) = \det(\mathbf{I})$ o, lo que es lo mismo, $\det(\mathbf{B})^2 = 1 \Rightarrow \log|\det \mathbf{B}| = 0$. Por ello, la función de verosimilitud se simplifica de forma apreciable, comparada con (3.4)

$$l_o(\mathbf{B}) = \frac{1}{T}\sum_{t=1}^{T} \log q(\mathbf{B}\,\mathbf{x})$$

aunque ahora debe ser optimizada sabiendo que $\mathbf{B}$ es una matriz ortogonal.

Sea $g_{jk}(\mathbf{B}) = \mathbf{b}_j^T\,\mathbf{b}_k - \delta_{jk}$, siendo $\mathbf{b}_i^T$ la i-ésima fila de $\mathbf{B}$. La condición

$$\mathbf{B}\,\mathbf{B}^T = \mathbf{I}$$

equivale al conjunto de restricciones

$$g_{jk}(\mathbf{B}) = 0 \text{ para todo } j,\, k \tag{3.12}$$

El **método de los multiplicadores de Lagrange** ([Thomas87]) permite encontrar los máximos y mínimos locales de $l_o(\mathbf{B})$ sujetos a (3.12), satisfaciendo simultáneamente las ecuaciones:

$$\nabla_{\mathbf{B}}\, l_o(\mathbf{B}) = \sum_{j,k\geq j} \lambda_{jk} \nabla_{\mathbf{B}}\, g_{jk}(\mathbf{B}) \tag{3.13a}$$

$$\mathbf{B}\mathbf{B}^{\mathrm{T}} = \mathbf{I} \tag{3.13b}$$

Ahora bien, resulta que $\sum_{j,k \geq j} \lambda_{jk} \nabla_{\mathbf{B}}\, g_{jk}(\,\mathbf{B}\,) = \Lambda\,\mathbf{B}$, siendo

$$\Lambda = \begin{bmatrix} 2\lambda_{11} & \lambda_{12} & \cdots & \lambda_{1N} \\ \lambda_{12} & 2\lambda_{22} & \cdots & \lambda_{2N} \\ & \ddots & \ddots & \\ \lambda_{1N} & \lambda_{2N} & \cdots & 2\lambda_{NN} \end{bmatrix}$$

Por lo tanto, al multiplicar (3.13*a*) por $\mathbf{B}^{\mathrm{T}}$ desde la derecha se obtiene, teniendo en cuenta (3.13*b*):

$$\nabla_{\mathbf{B}}\, l_o(\,\mathbf{B}\,)\,\mathbf{B}^{\mathrm{T}} = \Lambda \tag{3.14a}$$

y, trasponiendo,

$$\mathbf{B}\,\nabla_{\mathbf{B}}\, l_o(\,\mathbf{B}\,)^{\mathrm{T}} = \Lambda^{\mathrm{T}} = \Lambda \tag{3.14b}$$

Por lo tanto, al restar (3.14*b*) de (3.14*a*) se elimina la dependencia con Λ. Sabiendo que $\nabla_{\mathbf{B}}\, l_o(\,\mathbf{B}\,) = \varphi(\,\mathbf{y}(t)\,)\,\mathbf{x}^{\mathrm{T}}(t)$ (ver (3.10)), se obtiene el par de sistemas de ecuaciones

$$\frac{1}{\mathrm{T}} \sum_{t=1}^{\mathrm{T}} \varphi(\,\mathbf{y}(t)\,)\,\mathbf{y}^{\mathrm{T}}(t) - \mathbf{y}(t)\,\varphi(\,\mathbf{y}^{\mathrm{T}}(t)\,) = \mathbf{0} \tag{3.15a}$$

$$\mathbf{B}\mathbf{B}^{\mathrm{T}} = \mathbf{I} \tag{3.15b}$$

que sustituyen a (3.11). Por otra parte, como ahora $\mathbf{B}\mathbf{B}^{\mathrm{T}} = \mathbf{I}$ resulta que $\nabla_{\mathbf{B}}\, l_o(\mathbf{B})$ es ya de por sí el gradiente *natural* y no tenemos que preocuparnos por este aspecto.

3.2.3 Interpretación del MLE: La Distancia de Kullback-Leibler

Dadas dos *funciones de densidad de probabilidad* p_1 y p_2, se define la distancia de Kullback-Leibler (*dKL*) [Borovkov88, 207] entre las distribuciones como

$$dKL[\, p_1 \| p_2 \,] = \int_{-\infty}^{\infty} p_1(\mathbf{u}) \log \frac{p_1(\mathbf{u})}{p_2(\mathbf{u})} d\mathbf{u} \tag{3.16}$$

Debemos explicar la notación: se entiende que, en general, p_1 y p_2 son *f.d.p* conjuntas y, por lo tanto, funciones escalares de una variable vectorial **u**. El símbolo 'integral' en (3.16) representa una integral múltiple extendida a toda la región en la que el integrando está bien definido.

Resulta que $dKL[\, p_1 \| p_2 \,] \geq 0$, dándose la igualdad si y sólo si $p_1 = p_2$, por lo que a menudo se utiliza como una medida de la *distancia* entre las distribuciones, a pesar de que $dKL[\, p_1 \| p_2 \,] \neq dKL[\, p_2 \| p_1 \,]$.

La distancia de Kullback-Leibler está muy relacionada con el e*stimador de máxima verosimilitud,* como veremos a continuación [Cardoso97*a*]:

Sea **x** una variable aleatoria vectorial con *f.d.p.* conjunta igual a p_*. Supongamos que

$$\Pi = \{\, p_\theta \mid \theta \in \Theta \,\}$$

es un modelo paramétrico de p_*, donde no necesariamente $p_* = p_\theta$ para algún $\theta \in \Theta$. Dadas T muestras de **x**, **x**(1), ..., **x**(T), todas independientes entre sí, el MLE del parámetro se define como el valor de θ que maximiza

$$l(\,\theta\,) = \frac{1}{\mathrm{T}} \sum_{t=1}^{\mathrm{T}} \log p_\theta(\,\mathbf{x}(t)\,) \tag{3.17}$$

Suponiendo que $\log p_\theta(\mathbf{x}(t))$ está bien definido para todo t, entonces $l(\theta)$ converge en probabilidad[1] a $L(\theta) = E[\log p_\theta(\mathbf{x})]$ [Papoulis91, pág. 213] :

$$l(\theta) \underset{T\to\infty}{\to} L(\theta) = \int p_*(\mathbf{x}) \log p_\theta(\mathbf{x})\, d\mathbf{x}$$

(como antes, se entiende que la integral se extiende a todo el espacio muestral). Por ello, para T suficientemente grande, el error que se comete al sustituir $l(\theta)$ por $L(\theta)$ es, a efectos prácticos, despreciable. Podemos reescribir $L(\theta)$ como:

$$L(\theta) = -H[p_*] - dKL[\, p_* \,\|\, p_\theta\,]$$

donde $H[p_*] = -\int p_*(\mathbf{x}) \log p_*(\mathbf{x})\, d\mathbf{x}$ es la llamada entropía diferencial de la variable $\mathbf{x}$ (ver Apéndice A). Entonces, al no depender $H[p_*]$ del parámetro θ, *el MLE es el valor de* θ *que minimiza la distancia de Kullback-Leibler entre* p_* *y* p_θ. Esta relación es conocida de antiguo por los matemáticos [Borovkov88].

Para aplicar este resultado al problema de la Separación Ciega de Fuentes debemos identificar $\theta = \mathbf{B}$ (el estimador de la inversa de la matriz de mezcla) y, de (3.2),

$$p_*(\mathbf{x}(t)) = p_x(\mathbf{x}(t)\,;\, \mathbf{A}^{-1}, q), \qquad p_\theta(\mathbf{x}(t)) = p_x(\mathbf{x}(t)\,;\, \mathbf{B}, r)$$

donde $q(\cdot)$ es la *f.d.p* conjunta de la fuentes, que, por definición, es desconocida y $r(\cdot)$ es una aproximación a $q(\cdot)$. Es decir, $p_*(\mathbf{x}(t))$ es la auténtica distribución de $\mathbf{x}(t)$ mientras que $p_\theta(\mathbf{x}(t))$ es la distribución que se supone que tiene esta variable. Si volvemos a (3.11), resulta que la función vectorial $\varphi(\mathbf{y}(t))$ debe depender de $q(\cdot)$; entonces, basta que la elección de $\varphi(\mathbf{y}(t))$ sea arbitraria para que, de forma implícita, modelemos de forma errónea la distribución de las fuentes.

[1]Para todo $\varepsilon > 0$, cuando T crece la probabilidad del evento $|\, l(\theta) - L(\theta)\,| \le \varepsilon$ tiende a *uno*.

Incidamos en esta idea: no es necesario que $q(\cdot)$ y $r(\cdot)$ coincidan; ni siquiera cuando $\mathbf{B} = \mathbf{A}^{-1}$, para que el MLE pueda ser calculado: a la luz de esta discusión y para ser rigurosos, es mejor pensar que "*las ecuaciones del estimador de máxima verosimilitud encuentran, en realidad, la matriz* **B** *que minimiza la distancia de Kullback-Leibler entre la auténtica distribución de las observaciones y la que, por hipótesis, suponemos que tienen* ".

En consecuencia, podemos muy bien preguntarnos si las fuentes realmente se llegan a separar cuando la discrepancia entre $r(\cdot)$ y $q(\cdot)$ es grande. A pesar de que las matrices de separación **B** correctas *siempre* son solución de las ecuaciones (3.11), puede que el mínimo de la distancia de Kullback-Leibler no se corresponda con una matriz de separación admisible. Este problema será tratado en una Sección posterior. Por ahora, nos hemos limitado a mostrar su existencia.

Ejemplo 3.1. En su momento, nos apoyamos en el Teorema de Darmois-Skitovich para argumentar que no es posible separar fuentes de distribución gaussiana; sin embargo, no llegamos a demostrar este Teorema. La *dKL* servirá para argumentarlo ahora [Cardoso96*a*].

La *dKL* entre dos variables vectoriales gaussianas, **s** e $\mathbf{y} = \mathbf{C}\,\mathbf{s}$, de media cero y matrices de covarianza **I** e $\mathbf{R}_\mathbf{Y} = \mathbf{C}\,\mathbf{C}^T$, respectivamente, es igual a

$$dKL = (\text{Traza}(\mathbf{R}_\mathbf{Y}) - \log \det \mathbf{R}_\mathbf{Y} - \mathrm{N}) / 2$$

[Cover91], siendo N es la dimensión de ambos vectores. Las componentes de **s** son mutuamente independientes por hipótesis, así que la *dKL* será cero sólo cuando las componentes de **y** también sean independientes. Denotemos por $\mu_1, \ldots, \mu_N$ los autovalores de $\mathbf{R}_\mathbf{Y}$. Como $\text{Traza}(\mathbf{R}_\mathbf{Y}) = \sum_{i=1,N} \mu_i$ y $\log \det \mathbf{R}_\mathbf{Y} = \sum_{i=1,N} \log \mu_i$, resulta que $dKL = \Sigma_{i=1,N} (\mu_i - 1 - \log \mu_i) / 2$. Ahora bien, $\mu_i - 1 \geq \log \mu_i$, dándose la

igualdad si y sólo si $\mu_i = 1$. Por lo tanto, $dKL = 0$ (alcanza su mínimo) sólo si $\mu_i = 1$ para todo i, en cuyo caso $\mathbf{R}_Y = \mathbf{I}$, por lo que $\mathbf{C}$ debe ser forzosamente una matriz *ortogonal* pero no necesariamente una matriz de separación. En conclusión, el que las componentes de $\mathbf{y}$ estén *incorreladas* ya implica su independencia estadística. Esta propiedad es exclusiva de las variables aleatorias gaussianas.

3.3 El Principio de Maximización de la Información

En 1994, en un trabajo conjunto, el francés Nadal y el español Nestor Parga [Nadal94] mostraron que, con relaciones señal a ruido altas, *la máxima transferencia de información* entre la entrada y la salida de una red de neuronas se produce cuando la distribución estadística de las señales de salida es factorial, es decir, cuando su *f.d.p* conjunta puede ser factorizada en un producto de *f.d.p* marginales y, por lo tanto, las salidas son estadísticamente independientes. Casi de inmediato, se utilizó este principio para resolver el problema de la Separación de Fuentes [Bell95]. También merece la pena citar el trabajo precursor, aunque limitado, de Ukrainec y Haykin publicado en 1992 [Haykin94*b*, pág. 466].

Sea una red de neuronas que puede ser representada como en la Figura 3.2 por las relaciones:

$$\mathbf{u}(t) = \mathbf{g}(\,\mathbf{y}(t)\,), \qquad \text{donde} \qquad \mathbf{y}(t) = \mathbf{B}\,\mathbf{x}(t) \tag{3.18}$$

con $\mathbf{u}(t) = \mathbf{g}(\,\mathbf{y}(t)\,) = [\,g_1(y_1(t)), \ldots, g_N(y_N(t))\,]^T$, siendo $g_i(\,\cdot\,)$ una función monótona, de modo que su inversa $y_i = g_i^{-1}(\,u_i\,)$ está siempre bien definida.

Dadas dos variables aleatorias vectoriales $\mathbf{u}_1$ y $\mathbf{u}_2$ cuya *f.d.p* conjunta se denota por $p(\,\mathbf{u}_1,\mathbf{u}_2\,)$ y las *f.d.p* marginales son, respectivamente, $p_1(\,\mathbf{u}_1\,)$ y $p_2(\,\mathbf{u}_2\,)$, medimos su grado de independencia estadística con la cantidad [Haykin94*b*, pág. 448]

$$I[\,\mathbf{u}_1, \mathbf{u}_2\,] = dKL[\; p(\mathbf{u}_1,\mathbf{u}_2) \,\|\, p_1(\,\mathbf{u}_1\,)\, p_2(\,\mathbf{u}_2\,)\;] \tag{3.19}$$

que es llamada *información mutua* entre las variables. Se sabe que $I[\,\mathbf{u}_1, \mathbf{u}_2\,] = 0$ si y sólo si $p(\,\mathbf{u}_1, \mathbf{u}_2\,) = p(\,\mathbf{u}_1\,)\, p(\,\mathbf{u}_2\,)$ o, lo que es lo mismo, cuando las variables aleatorias son independientes. Para una exposición detallada de las propiedades de la *información mutua* y de otras magnitudes relacionadas con ellas, ver el Apéndice A.

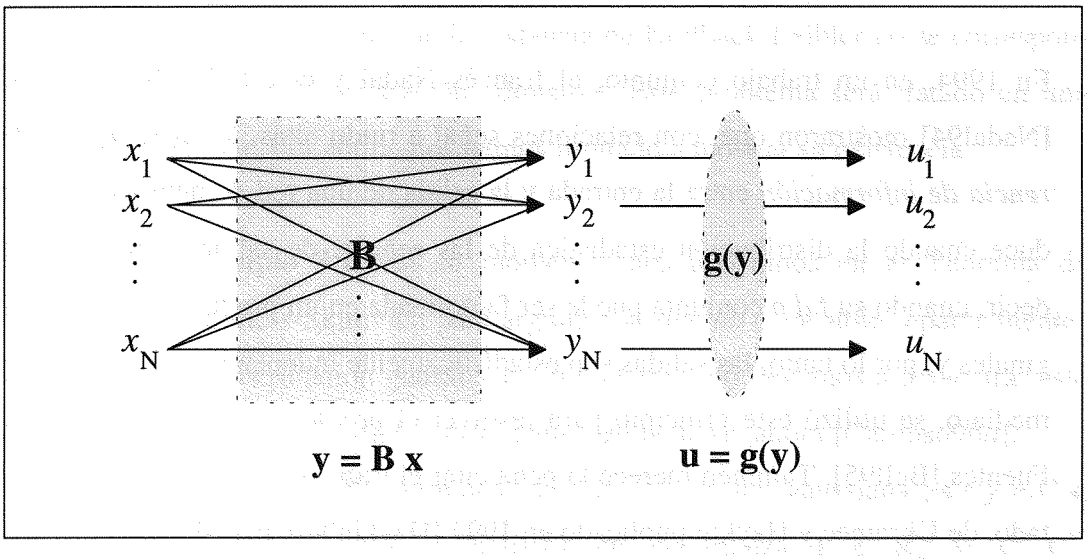

Figura 3.2. Modelo de red neuronal utilizado por Infomax

Es muy sencillo demostrar que (3.19) equivale para las variables aleatorias vectoriales $\mathbf{x}(t)$ y $\mathbf{u}(t)$ definidas en (3.18) a

$$I[\,\mathbf{u}(t), \mathbf{x}(t)\,] = H[\,\mathbf{u}(t)\,] - H[\,\mathbf{u}(t) \mid \mathbf{x}(t)\,] \tag{3.20}$$

donde $H[\,\mathbf{u}(t)\,]$ es la entropía diferencial de la variable aleatoria $\mathbf{u}(t)$ y

$$H[\,\mathbf{u}(t) \mid \mathbf{x}(t)\,] = H[\,\mathbf{u}(t), \mathbf{x}(t)\,] - H[\,\mathbf{x}(t)\,]$$

es la entropía diferencial de $\mathbf{u}(t)$ condicionada al valor de $\mathbf{x}(t)$ [Haykin94*b*, pág.448]. Se dice que $H[\,\mathbf{u}(t) \mid \mathbf{x}(t)\,]$ mide la incertidumbre remanente en $\mathbf{u}(t)$ supuesto que $\mathbf{x}(t)$ es conocida. En el caso de nuestro interés, $\mathbf{u}(t) = \mathbf{g}(\,\mathbf{B}\,\mathbf{x}(t)\,)$ es una función determinista de $\mathbf{x}(t)$ y entonces $H[\,\mathbf{u}(t) \mid \mathbf{x}(t)\,] = 0$.

Por lo tanto, de (3.20),

$$I[\,\mathbf{u}(t), \mathbf{x}(t)\,] = H[\,\mathbf{u}(t)\,] \quad \Rightarrow \quad \nabla_{\mathbf{B}}\, I[\,\mathbf{u}(t), \mathbf{x}(t)\,] = \nabla_{\mathbf{B}}\, H[\,\mathbf{u}(t)\,] \tag{3.21}$$

por lo que *la transferencia de información a través de la red se optimiza haciendo máxima la entropía de las salidas* $\mathbf{u}(t)$ [Bell95]. Para hacer el desarrollo pertinente, el siguiente teorema es necesario:

Teorema 3.1 (*Transformación de la Entropía de variables continuas*). Sea $\mathbf{x}(t)$ una variable aleatoria *continua*. Si $\mathbf{u}(t) = \mathbf{g}(\mathbf{B}\mathbf{x}(t))$, entonces

$$H[\,\mathbf{u}(t)\,] = H[\,\mathbf{x}(t)\,] + \mathrm{E}\log \mid \mathbf{J} \mid$$

siendo $H[\,\mathbf{x}(t)\,]$ y $H[\,\mathbf{u}(t)\,]$ las respectivas entropías diferenciales de las variables y $\mathbf{J}$ es el Jacobiano de la transformación que las liga.

Demostración. En [Papoulis91, pág. 565], se prueba el teorema pero sólo cuando las variables aleatorias son escalares. La demostración es inmediata si $\mathbf{x}$ es un vector: por definición, $H[\,\mathbf{u}\,] = -\,\mathrm{E}[\,\log p_{\mathbf{u}}(\,\mathbf{u}\,)\,]$. Ahora bien, la ley de transformación de variables *continuas* establece que $p_{\mathbf{u}}(\,\mathbf{u}\,) = p_{\mathbf{x}}(\,\mathbf{x}\,) \,/\, \det \mid \mathbf{J} \mid$. Entonces, resolviendo la integral con el adecuado cambio de variables, obtenemos, efectivamente, $H[\,\mathbf{u}\,] = -\mathrm{E}\,[\log p_{\mathbf{x}}(\,\mathbf{x}\,) \,/\, \det \mid \mathbf{J} \mid] = H[\mathbf{x}] + \mathrm{E}\log \mid \mathbf{J} \mid$.

Usando el Teorema 3.1, (3.21) queda como

$$\nabla_{\mathbf{B}} I[\, \mathbf{u}(t), \mathbf{x}(t)\,] = \nabla_{\mathbf{B}} \left(H[\, \mathbf{x}(t)\,] + \mathrm{E} \log |\, \mathbf{J}\, | \right) \tag{3.22}$$

Como $H[\, \mathbf{x}(t)\,]$ no depende de $\mathbf{B}$, resulta que $\nabla_{\mathbf{B}} I[\, \mathbf{u}(t), \mathbf{x}(t)\,] = \nabla_{\mathbf{B}} \mathrm{E} \log |\, \mathbf{J}\, |$. Ahora bien, por definición, $\mathbf{J}$ es la matriz:

$$\mathbf{J} = \begin{bmatrix} \frac{\partial u_1}{\partial x_1} & \cdots & \frac{\partial u_1}{\partial x_N} \\ \vdots & \ddots & \vdots \\ \frac{\partial u_N}{\partial x_1} & \cdots & \frac{\partial u_N}{\partial x_N} \end{bmatrix} \tag{3.23}$$

siendo $\frac{\partial u_i}{\partial x_j} = \frac{\mathrm{d}\, u_i}{\mathrm{d}\, y_i} \frac{\partial y_i}{\partial x_j} = g_i'(\, y_i\,)\, b_{ij}$. Se sabe que al escalar cualquier fila de una matriz, escalamos su determinante por la misma cantidad. Entonces, es sencillo probar que

$$\mathbf{J} = (\det \mathbf{B})\, \Pi_{i=1}^{\mathrm{N}}\, g_i'(\, y_i\,) \tag{3.24}$$

y, por tanto, [Bell95]

$$\mathrm{E} \log |\, \mathbf{J}\, | = \log |\det \mathbf{B}\, | + \mathrm{E}\, \Sigma_{i=1}^{\mathrm{N}}\, \log\, g_i'(\, y_i\,) \tag{3.25}$$

Si los procesos $y_i(\, t\,)$ son *ergódicos*, entonces podemos sustituir las esperanzas matemáticas en (3.25) por medias muestrales. Entonces, (3.25) resulta ser *formalmente idéntica* a (3.4) siempre que $g_i'(\, y_i(t)) = q_i(\, \mathbf{B}\, \mathbf{x}(t)\,)$. Por lo tanto, la matriz $\mathbf{B}$ *que maximiza la transferencia de información en la red es justamente solución de las ecuaciones* (3.11), las mismas que daban el MLE.

Sin embargo, hay una diferencia crucial: para obtener este resultado, hemos supuesto que los procesos son ergódicos; pero no que las *observaciones* sean independientes entre sí, esto es, $\mathbf{x}(t)$ y $\mathbf{x}(t')$ no han de ser, necesariamente,

independientes si $t \neq t'$ (lo que sí se postuló para el MLE). Si, como caso particular, esto sucede, obtenemos el estimador de máxima verosimilitud de $\mathbf{A}^{-1}$, que es asintóticamente el más eficiente.

Otra cosa distinta es que los procesos sean ergódicos pero $\mathbf{x}(t)$ y $\mathbf{x}(t')$ no sean independientes: entonces puede que la sustitución de la esperanza matemática por la media muestral sólo sea apropiada (con un error pequeño) para un número T de muestras muy grande [Papoulis91, pág. 427]; pero esto no invalida el desarrollo y el principio de funcionamiento de la red neuronal.

Relación entre Infomax y el MLE.

Entonces, maximizar la transferencia de información en la red hace que separemos las fuentes. Este resultado es sorprendente y debe ser estudiado con más detalle. Es sencillo probar que [Bell95]

$$H[\,\mathbf{u}(t)\,] = H[\,u_1(t)\,] + \ldots + H[\,u_N(t)\,] - I[\,\mathbf{u}(t)\,] \qquad (3.26)$$

Por lo tanto, optimizar $H[\,\mathbf{u}(t)\,]$ es un compromiso entre buscar el máximo de la suma $H[\,u_1(t)\,] + \ldots + H[\,u_N(t)\,]$ y minimizar $I[\,\mathbf{u}(t)\,]$.

Cuando $\mathbf{B} = \mathbf{A}^{-1}$ (separamos las fuentes), $\mathbf{y}(t) = \mathbf{s}(t)$ y, por consiguiente, las variables $u_i(t) = g_i(\,y_i(t)\,) = g_i(\,s_i(t)\,)$ son estadísticamente independientes entre sí, de manera que $I[\,\mathbf{u}(t)\,] = 0$.

Por otra parte, la comparación entre las ecuaciones del MLE y (3.24) sugiere que la derivada de $g_i(\,\cdot\,)$ coincida con $q_i(\,\cdot\,)$, de manera que $g_i(\,\cdot\,)$ debe ser la *función de distribución* de la i-ésima fuente. Entonces, de $u_i = g_i(\,s_i\,)$ se deduce que u_i está acotada y, además, es sencillo probar que se distribuye uniformemente en el intervalo (0,1). Por lo tanto,

$$H[\,u_i(t)\,] = -\,\mathrm{E}[\,\log p(\,u_i(t)\,)] = 0 \qquad (3.26)$$

que es el máximo valor que puede tomar la entropía diferencial de una variable que está acotada en el intervalo (0, 1) [Te-Won99, pág. 12]. En realidad, las variables aleatorias uniformes son las que tienen la mayor entropía de entre todas las variables que están acotadas en el mismo intervalo. Sin embargo, que la entropía diferencial sea cero no quiere decir que la incertidumbre sobre la variable haya desaparecido, todo lo contrario: aquí es máxima. No debe crearse confusión con el hecho de que, para una variable aleatoria *discreta*, entropía nula implique *determinismo*.

Finalmente, como $I[\ \mathbf{u}(t)\ ]$ es mínima y $H\ [\ u_i(t)\ ]$ es máxima para todo i, resulta que $H[\ \mathbf{u}(t)\ ]$, la entropía de $\mathbf{u}(t)$, es *máxima* cuando separamos las fuentes, como queríamos demostrar (coincide, además, que $H[\ \mathbf{u}(t)\ ] = 0$). Es decir, la transferencia de información es máxima cuando las salidas u_i son *estadísticamente independientes*.

Sin embargo, $q_i(\ \cdot\ )$ es desconocida por definición del problema. Si $g_i(\ \cdot\ )$ no coincide con la función de distribución de la fuente, cabe esperar que, además de la raíz correcta $\mathbf{B} = \mathbf{A}^{-1}$, las ecuaciones tengan soluciones que no sean matrices de separación. Este problema ya se puso de manifiesto en la Sección anterior, con otro enfoque. Ahora podemos ilustrarlo proponiendo el siguiente Ejemplo

> **Ejemplo 3.2.** (adaptado de [Bell95]). Supongamos que el problema nos da dos fuentes, $s_1(t)$ y $s_2(t)$, de distribución uniforme. Por definición, se toma $g_i(\ y_i\ ) = 1\ /\ (1 + \exp(-\ y_i\ ))$. Si $\mathbf{A} = \mathbf{I}$, Bell y Sejnowski han apuntado que $H[\ \mathbf{u}\ ]$ es mayor para $y_1(t) = s_1(t) + s_2(t)$ e $y_1(t) = s_1(t) - s_2(t)$ que para $y_1(t) = s_1(t)$ e $y_2(t) = s_2(t)$. Se puede argumentar que las fuentes son sub-gaussianas, mientras que $g_i(\ \cdot\ )$ está ajustada a una *f.d.p* super-gaussiana y, por ello, el resultado no es satisfactorio.

En cualquier caso, de la discusión anterior podemos extraer una consecuencia muy importante: sea cual sea la función $g_i(\cdot)$, u_i debe estar acotada. De esta forma, se garantiza que $H[u_i]$ tiene una cota superior ($H[u_i] \leq 0$). Como, además, $I[\mathbf{u}]$ es siempre mayor o igual que cero, resulta que $H[\mathbf{u}]$ también está acotada superiormente: $H[\mathbf{u}] \leq 0$. En consecuencia, de maximizar $H[\mathbf{u}]$ mediante un algoritmo de optimización adaptativo, podemos confiar en que éste no va a diverger [Bell95].

Comentarios sobre Infomax.

Finalmente, desearíamos hacer tres observaciones:

- En primer lugar, las conclusiones que se obtienen *no* son válidas si las variables aleatorias son *discretas*. En tal caso, se puede demostrar que la información mutua $I[\mathbf{u}(t), \mathbf{x}(t)] = H[\mathbf{u}(t)] = H[\mathbf{x}(t)]$ (ver Figura 2.2), con independencia del valor que tome $\mathbf{B}$ [Papoulis91, pág. 565]. Es decir, no tiene sentido maximizar $I[\mathbf{u}(t), \mathbf{x}(t)]$ porque ésta siempre toma el mismo valor.

- En segundo lugar, veamos cómo se puede explicar la regla Infomax a partir del concepto de distancia de Kullback-Leibler [Cardoso97*a*]. Sea $\mathbf{v} = [v_1,...,v_N]^T$ una variable aleatoria vectorial cualquiera (es decir, no se la identifica con ninguna de las que han aparecido en el desarrollo) con *f.d.p* conjunta:

$$q(\mathbf{v}) = \prod_{i=1}^{N} q_i(v_i)$$

Sea una función $g_i(\cdot)$ tal que su derivada $g_i'(\cdot) = q_i(\cdot)$. Entonces, $g_i(v_i)$ está distribuida uniformemente en el intervalo (0, 1) y, en consecuencia, $\mathbf{w} = \mathbf{g}(\mathbf{v})$ se distribuye uniformemente en $(0,1)^N$. Como $p(\mathbf{w}) = 1$ para $\mathbf{w} \in (0,1)^N$, es evidente que la entropía de $\mathbf{u} = \mathbf{g}(\mathbf{Bx})$, donde también se cumple que $\mathbf{u} \in (0,1)^N$, se puede escribir como:

$$H[\,\mathbf{u}\,] = -\,dKL[\,\mathbf{u} \,\|\, \mathbf{w}\,]$$

Ahora bien, la *dKL* es invariante ante las transformaciones invertibles del espacio muestral (ver Apéndice A). Por lo tanto, $dKL[\,\mathbf{u} \,\|\, \mathbf{w}\,] = dKL[\mathbf{g}^{-1}(\mathbf{u}) \,\|\, \mathbf{g}^{-1}(\mathbf{w})] = dKL[\,\mathbf{B}\,\mathbf{x} \,\|\, \mathbf{v}\,]$. En consecuencia, maximizar la entropía de **u** es equivalente a minimizar la *dKL entre la distribución de las salidas* $\mathbf{y} = \mathbf{B}\,\mathbf{x}$ *y la distribución supuesta de las fuentes*; pero esto es lo mismo que dijimos para el MLE. Fue Cardoso [Cardoso97*a*] el primero en notar la equivalencia entre Infomax y el MLE, aunque restringió su estudio al caso en el que las observaciones son temporalmente independientes.

◆ Por último, una curiosidad: Infomax modela matemáticamente un principio biológico de funcionamiento de las neuronas. Para que el modelo sea correcto, debemos sumar ruido a la salida de la neurona, como realmente ocurre en cada sinapsis. Este ruido tiene una importante función en las tareas de aprendizaje; no obstante, no altera el desarrollo previo, como veremos: sea el ruido aditivo **n**, cuya *f.d.p* es $p_{\mathbf{n}}(\,\mathbf{n}\,)$. Suponemos que la de la red neuronal responde ahora a la expresión:

$$\mathbf{z} = \mathbf{u} + \mathbf{n}$$

Resulta que

$$I[\,\mathbf{z}, \mathbf{x}\,] = H[\,\mathbf{z}\,] - H[\,\mathbf{n}\,]$$

pues, como se definió en (3.20), $I[\,\mathbf{z}, \mathbf{x}\,] = H[\,\mathbf{z}\,] - H[\,\mathbf{z} \mid \mathbf{x}\,]$. Si la entrada **x** de la red es conocida, toda la incertidumbre sobre el valor de **z** se debe **n**. Por lo tanto, $p(\,\mathbf{z} \mid \mathbf{x}\,) = p_{\mathbf{n}}(\,\mathbf{z} - \mathbf{u}\,) = p_{\mathbf{n}}(\,\mathbf{z} - \mathbf{g}(\,\mathbf{B}\,\mathbf{x}\,)\,)$ y haciendo el cambio $\mathbf{z} = \mathbf{n} + \mathbf{u}$ en la

expresión de $H[\,\mathbf{z} \mid \mathbf{x}\,]$, resulta $H[\,\mathbf{z} \mid \mathbf{x}\,] = H[\,\mathbf{n}\,]$. Ahora bien, dado que $H[\,\mathbf{n}\,]$ no depende de **B**, obtenemos $\nabla_{\mathbf{B}}\, I[\,\mathbf{z}, \mathbf{x}\,] = \nabla_{\mathbf{B}}\, H[\,\mathbf{z}\,]$, igual que en ausencia de ruido [Bell95].

3.4 Discusión sobre el estimador de máxima verosimilitud

Para utilizar con propiedad el estimador de máxima verosimilitud, tenemos forzosamente que conocer la distribución estadística de las fuentes. Ahora bien, así negamos la propia esencia del problema de la Separación, enfatizada por el adjetivo "ciega". Por lo tanto, o bien se da un valor *a priori* a la *f.d.p* de las fuentes (con el riesgo de que una mala elección nos dé peores resultados) o bien estimamos esta *f.d.p* junto con la matriz de mezcla. Por ahora, tan sólo mencionamos esta segunda posibilidad, dejando su análisis para el siguiente Capítulo.

Indeterminaciones en el orden y el signo de las Fuentes

Sea $r(\,\cdot\,)$ la *f.d.p* con la que aproximamos $q(\,\cdot\,)$, la auténtica *f.d.p* de las fuentes (idealmente $r(\,\cdot\,) = q(\,\cdot\,)$). Si $r(\,\cdot\,)$ es una función par tal que

$$p_{\mathbf{x}}(\,\mathbf{x}(t)\,;\,\mathbf{B},\,r\,) = p_{\mathbf{x}}(\,\mathbf{x}(t)\,;\,-\mathbf{B},\,r\,)$$

la función de verosimilitud tendrá, al menos, dos máximos globales y uno de ellos cambia el signo de las fuentes estimadas. Del mismo modo, si $p_{\mathbf{x}}(\,\cdot\,)$ es invariante ante cambios en el orden de sus argumentos, tendremos nuevos máximos globales que se corresponden con permutaciones en el orden de las fuentes. En cualquier caso, todos pueden ser soluciones admisibles del problema.

¿ Siempre separamos las Fuentes ?

Las ecuaciones (3.11), de las que se obtiene el MLE, presuponen que las observaciones $\mathbf{x}(t)$ son *temporalmente* independientes entre sí. Pero ¿Qué ocurre cuando esto no se cumple? Afortunadamente, a partir del principio de Infomax también se deducen las ecuaciones (3.11), las mismas del MLE, siempre que los procesos sean *ergódicos*. Ahora bien, Infomax no es propiamente un principio de Separación de Fuentes: siempre preferirá maximizar la entropía de las salidas antes que extraer componentes independientes. Este inconveniente que presenta Infomax equivale al que tiene el MLE por presuponer el conocimiento exacto de la *f.d.p* de las fuentes.

Entonces, acabamos de exponer dos problemas fundamentales:

- Se necesita un conocimiento *exacto* de la *f.d.p* de las fuentes. Si no es así (lo que será frecuente), no podemos garantizar la Separación.
- Las observaciones deben ser temporalmente independientes. Si no lo son, Infomax garantiza que las ecuaciones (3.11) aún separan las fuentes (suponiendo que su *f.d.p* es conocida); pero la estimación no es de verosimilitud máxima y, por ello, se pierde eficiencia.

Vamos a ilustrar mejor la discusión mediante experimentos. Sea, como siempre, $\mathbf{s}(t) = \mathbf{A}\ \mathbf{x}(t)$ el modelo de la mezcla. Dadas T observaciones *independientes* $\mathbf{x}(1)$, ..., $\mathbf{x}(T)$, el MLE de $\mathbf{A}^{-1}$ es la matriz $\mathbf{B}$ que, de acuerdo con (3.4), maximiza:

$$l(\mathbf{B}) = \log|\det \mathbf{B}| + \frac{1}{T}\sum_{t=1}^{T} \log r(\mathbf{B}\,\mathbf{x}) \qquad (3.27)$$

siendo r la (*supuesta*) *f.d.p* conjunta de las fuentes, que no tiene por qué coincidir con la auténtica. Las siguientes figuras ilustran la dependencia de l con $\mathbf{B}$, sin variar r. Para cada experimento, utilizaremos dos fuentes y un número T = 1000 observaciones. La expresión (3.27) se va a evaluar en $\mathbf{B} = (\mathbf{A}\,\mathbf{M}(u, v))^{-1}$, siendo

$$\mathbf{M}(u, v) = \begin{bmatrix} \cosh u & \text{senh}u \\ \text{senh}u & \cosh u \end{bmatrix} \cdot \begin{bmatrix} \cos v & -\text{sen}v \\ \text{sen}v & \cos v \end{bmatrix}$$

pues, para u y v pequeñas,

$$\mathbf{M}(u, v) \approx \mathbf{I} + u \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} + v \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

de forma que, variando u y v podemos "explorar" fácilmente la vecindad de $\mathbf{A}^{-1}$ [Cardoso98*a*]. A u se la llama coordenada *simétrica* y a v coordenada *antisimétrica*.

- **Experimento 3.1.** Las fuentes se generan elevando al cubo una variable *gaussiana* de media cero, donde la varianza del resultado se normaliza para que valga uno. El histograma de una de las fuentes se muestra a continuación en la Figura 3.3.

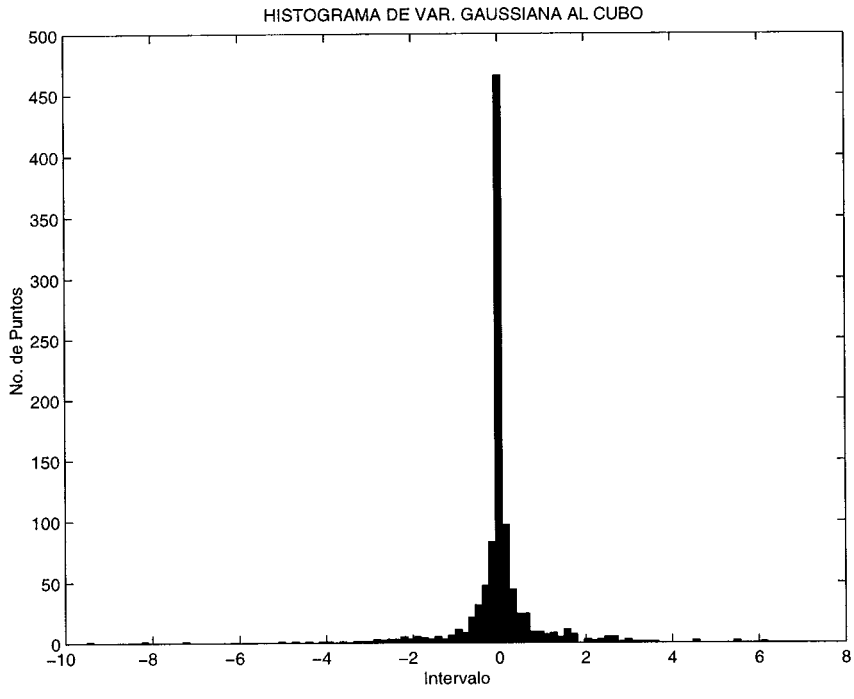

Figura 3.3. Histograma de las fuentes del Experimento 3.1

Para la distribución r de las fuentes se hacen dos presupuestos. En la hipótesis A tomamos $r(\ s_i\ ) \propto \exp(\ -\ |\ s_i\ |\ )$ (*f.d.p* de Laplace). En la hipótesis B, se supone

$$r(\ s_i\ ) \propto N(\ 1'5,\ 1\ ) + N(\ -\ 1'5,\ 1\ )$$

(*f.d.p* bimodal), siendo $N(\ \mu,\ 1\ )$ una *f.d.p* gaussiana de media μ y varianza unidad. Los resultados se muestran en las Figuras 3.4 (hipótesis A) y 3.5 (hipótesis B). Claramente, el óptimo bajo la hipótesis A corresponde a una matriz

de separación. Por el contrario, la matriz **A** es un punto de silla de la función de verosimilitud bajo la hipótesis B.

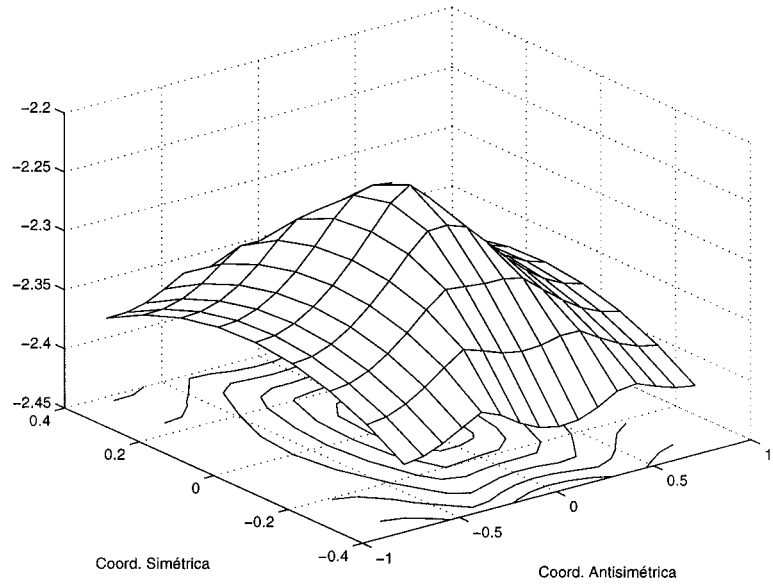

Figura 3.4. *Función de verosimilitud.* Hipótesis A

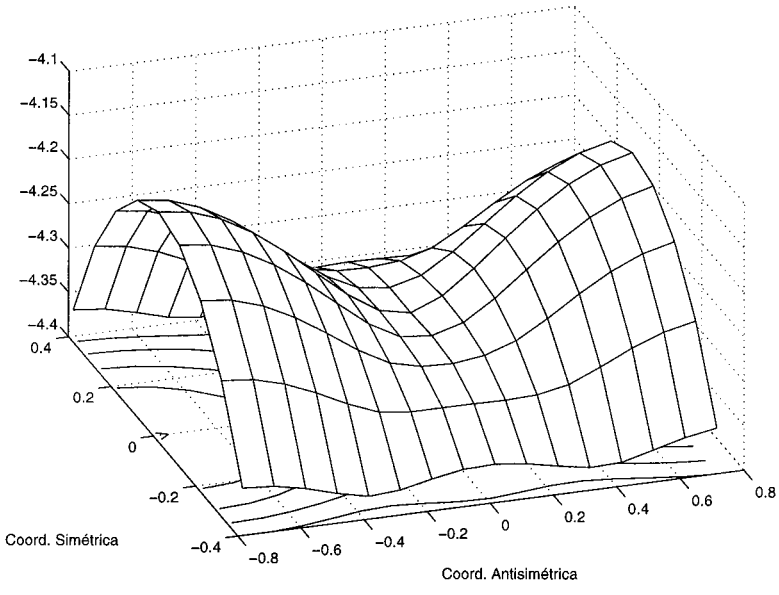

Figura 3.5. *Función de verosimilitud.* Hipótesis B

Interpretación. A la vista del histograma de las fuentes, la *f.d.p* de Laplace parece una elección razonable de r, la supuesta *f.d.p* de las fuentes. De hecho, así lo confirma el experimento. Por el contrario, la elección B es catastrófica. Ya que la *curtosis* es un parámetro que caracteriza justamente la *forma* de las *f.d.p* (ver Apéndice B), se tiende a pensar que basta con que la curtosis de r tenga el mismo signo que la curtosis de las fuentes para que se alcance la Separación [Girolami97]. De hecho, las fuentes de nuestro experimento son *super-gaussianas*, al igual que la *f.d.p* de Laplace mientras que la *f.d.p* bimodal es *sub-gaussiana*. Sin embargo, esto no es rigurosamente cierto: al repetir el experimento con *nuevas* fuentes, ahora de distribución *uniforme* (sub-gaussianas) y media cero, se obtiene bajo la hipótesis B el resultado de la Figura 3.6. Si bien $l(\mathbf{B})$ alcanza su máximo en la matriz de Separación, la superficie es demasiado plana como para poder pensar que los algoritmos de optimización van a tener una convergencia rápida y exenta de problemas. De hecho, al sustituir la *f.d.p* bimodal por $r(s_i) \propto N(2, 1) + N(-2, 1)$, que también es *sub-gaussiana*, el máximo en el origen desparece.

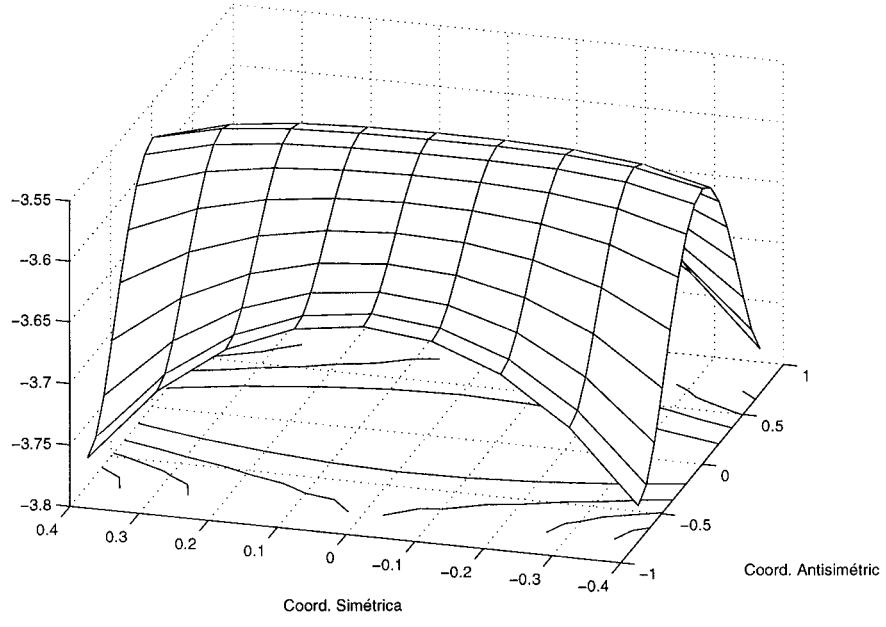

Figura 3.6. Func. de verosim. para fuentes uniformes bajo la hipótesis B.

- **Experimento 3.2.** Vamos a repetir el Experimento 3.1 tomando observaciones que *no* son temporalmente independientes. Para ello, cada fuente generada en dicho experimento se trata con un filtro paso de baja IIR de primer orden que tiene un polo en $z = 0'9$. De esta forma, se introduce una fuerte correlación entre las muestras. La función de verosimilitud bajo la hipótesis A se muestra en la Figura 3.7, que debe ser comparada con la Figura 3.4. En efecto, $\mathbf{A}^{-1}$ aún optimiza la función de verosimilitud; pero encontramos nuevos *puntos de silla* que aparecerán en la solución analítica de (3.11).

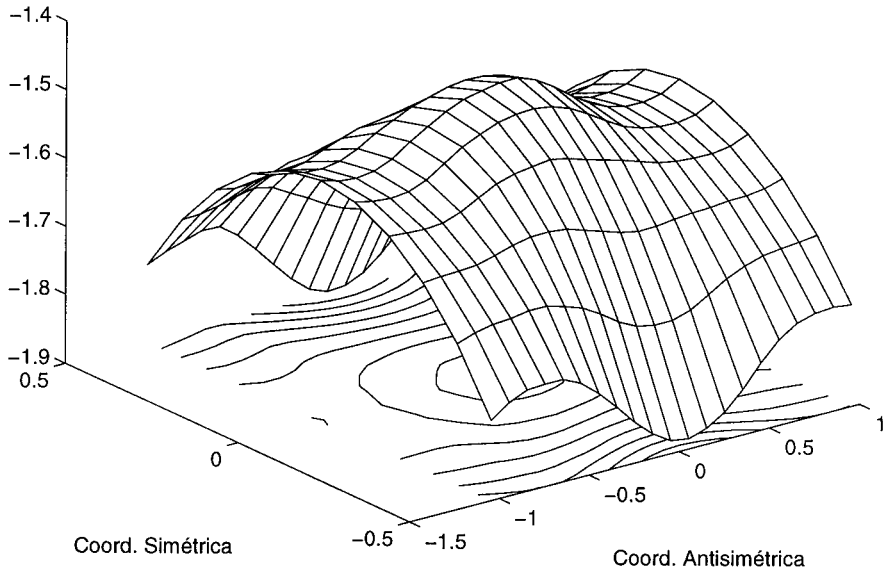

Figura 3.7. *Función de Verosimilitud.* Observaciones no independientes.

Nótese además que las *matrices de separación* han resultado siempre ser *máximos* o *puntos de silla* de $l(\mathbf{B})$ (en general, también pueden *minimizar* $l(\mathbf{B})$).

La razón es sencilla: según (3.11), cualquier matriz de separación soluciona las ecuaciones $\nabla_{\mathbf{B}}\, l(\mathbf{B}) = \mathbf{0}$.

Cabe preguntarse si del análisis de las derivadas segundas, que distinguen entre *óptimo* y *punto de silla*, puede extraerse alguna información útil. En efecto, así es. No obstante, pospondremos este estudio hasta el Capítulo 5 de la Tesis.

Fuentes de Naturaleza Discreta y otros Problemas

Por último, ya se discutió que el principio de Infomax no tiene sentido si las variables aleatorias son discretas (porque, entonces, la entropía de la salida no depende de la matriz de separación **B**). Desde el punto de vista de la estimación de máxima verosimilitud, resulta que

$$p_{\mathbf{x}}(\mathbf{x}\,;\mathbf{A}^{-1}, q) = q(\mathbf{A}^{-1}\mathbf{x}).$$

cuando **s** (y, por tanto, **x**) es una variable aleatoria *discreta* [Papoulis91]. Entonces, la función de verosimilitud para fuentes discretas vale:

$$\begin{aligned} l(\mathbf{B}) &= \frac{1}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}} \log p_{\mathbf{x}}(\mathbf{x}(t); \mathbf{B}, q) = \\ &= \frac{1}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}} \log q(\mathbf{B}\,\mathbf{x}(t)) \end{aligned} \tag{3.28}$$

(compárese con la expresión (3.4) para fuentes *continuas*). Que las fuentes sean discretas significa que **B x** pertenece al mismo espacio muestral que **s** si y sólo si $\mathbf{B} = \mathbf{A}^{-1}$ (o, en general, **B** igual a una matriz de separación admisible). En otro caso, la probabilidad q de que **s** tome el valor **B x** es *cero*.

Ejemplo 3.3. Supongamos dos fuentes binarias s_1 y $s_2 \in \{-1, 1\}$. Es decir, $\mathbf{s}^{\mathrm{T}} \in \mathfrak{S} = \{[-1,-1], [-1, 1], [1,-1], [1, 1]\}$ de donde $q(\mathbf{s}) = 0$ si

$\mathbf{s}^T \notin \mathfrak{S}$. Las observaciones $\mathbf{x}$ vendrán dadas por $\mathbf{x} = \mathbf{A}\,\mathbf{s}$, donde la matriz de mezcla vale

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \quad \Rightarrow \quad \mathbf{A}^{-1} = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

es decir, $\mathbf{x}^T \in \{[-2, 0], [0, -2], [0, 2], [2, 0]\}$. Tomando $\mathbf{y} = \mathbf{B}\,\mathbf{x}$, con $\mathbf{B} = 2\,\mathbf{A}^{-1}$ ¿Cuál es la probabilidad de cada evento $\mathbf{y}$? Pues bien, resulta que $\mathbf{y}^T \in \{[-2, -2], [-2, 2], [2, -2], [2, 2]\}$, $\mathbf{y}^T \notin \mathfrak{S}$, por lo que $q(\mathbf{y}) = 0$.

Entonces, la función de verosimilitud (3.28) no está definida porque $q(\mathbf{B}\,\mathbf{x}) = 0$, a no ser que, en efecto, $\mathbf{B}$ sea una matriz de separación. Es decir, la función de verosimilitud (3.28) no es continua en sus máximos y, como resultado, estos no pueden ser determinados con las técnicas del cálculo diferencial.

De todas formas, no queremos decir que con las ecuaciones (3.11) sea imposible obtener una *matriz de separación*. De hecho, las matrices de separación son siempre solución de (3.11) como se razonó en la Sección anterior. Lo que ocurre es que, ahora forzosamente, r debe corresponder a la *f.d.p* de una variable *continua* para no tener problemas y por lo tanto $r \neq q$, con todos los inconvenientes que ello apareja. En particular, la varianza de las estimaciones ya *no* es mínima porque ya *no* tenemos el MLE.

Esta discusión no puede ser trivial cuando advertimos que todas las señales digitales son variables aleatorias *discretas*. Los métodos geométricos de Separación de Fuentes [Puntonet95*a*, Prieto97] han sido desarrollados específicamente para señales discretas y no necesitan que las fuentes sean estadísticamente independientes (ello no debe sorprendernos, la hipótesis de independencia es muy útil gracias al teorema de Darmois-Skitovich; pero no tiene por qué ser la única oportunidad de diseño de algoritmos).

Belouchrani [Belouchrani94] ha añadido ruido de distribución continua a las estimaciones de las fuentes. De esta manera convierte estas variables de discretas a continuas. Sin embargo, también debe estimar los parámetros del ruido, complicando la solución.

En realidad, las fuentes de naturaleza discreta no son las únicas con una problemática específica. Consideremos una variable aleatoria continua de distribución *uniforme*, es decir, q = *constante*. De resolver (3.11) tomando, como se debe, $\varphi_i(y_i) = -\frac{1}{q_i(y_i)} \frac{d}{dy_i} q_i(y_i)$, resulta $\varphi(\mathbf{y}(t)) = \mathbf{0}$. Si trasladamos este resultado a las ecuaciones (3.11) resulta la identidad carente de sentido

$$(\mathbf{B}^T)^{-1} = \mathbf{0},$$

La raíz del problema se encuentra en que las observaciones $\mathbf{x}(t)$ para $t = 1, \ldots, T$ *no contienen información suficiente sobre la matriz de mezcla*, salvo que sean puntos de la frontera del espacio muestral.

Ejemplo 3.4 Supongamos que las fuentes pueden tomar cualquier valor en el intervalo [−1, 1] con igual probabilidad. Pensemos en una situación con sólo dos fuentes y la matriz de mezcla ortogonal

$$\mathbf{A} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$$

La distribución de las fuentes en el plano se muestra en la Figura 3.8 *a)* mientras que la de las observaciones se muestra en la Figura 3.8 *b)*. Como vemos, se obtiene una a partir de la otra rotando las *figuras* un ángulo de $\pi/4$ *rads*. Mediante un giro, cualquier subconjunto abierto de la región mostrada en la Figura 3.8 *b)* puede quedar comprendido en la región de la Figura 3.8 *a)*. El problema es que, para ello, el ángulo de giro no tiene que ser necesariamente $-\pi/4$ *rads*. Para

entenderlo, es clarificador descomponer dicho giro en dos rotaciones: una primera de $-\pi/4$ *rads* y la otra del ángulo restante. Evidentemente, este segundo ángulo será tanto más pequeño cuanto mayor sea la región que giramos; pero ello no invalida el razonamiento.

Sea $\mathfrak{B}$ el conjunto de todas estas matrices de giro *permitidas* (que no son necesariamente matrices de separación), entre las que se encuentra $\mathbf{A}^{-1}$. Resulta que $\mathbf{y}(t) = \mathbf{B}\,\mathbf{x}(t)$ es, para cada t, un punto de la Figura 3.8 *a)*. Como todos son equiprobables, $q(\mathbf{y}(t))$ es constante y, de nuevo, $\varphi(\mathbf{y}(t)) = \mathbf{0}$. Dicho de otra manera, todas las matrices de $\mathfrak{B}$ son equivalentes y el MLE no tiene preferencia por ninguna.

Así, las únicas observaciones $\mathbf{x}(t)$ que identifican sin ambigüedad a las matrices de separación son aquéllas que pertenecen a la frontera del espacio muestral. De hecho, el análisis de estos puntos recuerda bastante al que hicimos sobre las fuentes discretas, pues los puntos de la frontera sólo pertenecen al espacio muestral de $\mathbf{s}$ cuando la matriz de separación es correcta.

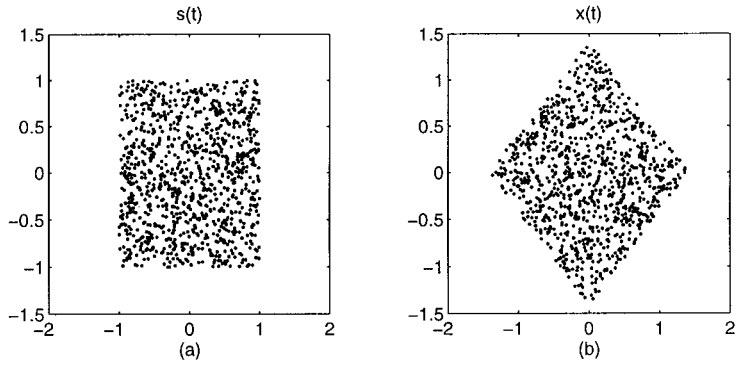

Figura 2.8. *a) Fuentes uniformes b) Mezclas*

Esta discusión tendrá validez, en general, siempre que la *f.d.p* conjunta de las fuentes presente simetrías, teniendo especial interés el caso en el que, además, la

región de soporte de la distribución sea finita, como ocurre con las variables *uniformes*.

3.5 Conclusiones

En este Capítulo hemos formulado las ecuaciones *del estimador de máxima verosimilitud de la matriz de mezcla*. La forma de las ecuaciones muestra que este estimador es *eficiente* en el sentido descrito en el Capítulo anterior (ver la Sección 2.4.2 y *ss*.). De hecho, el estimador de máxima verosimilitud es el estimador no sesgado *más* eficiente por cuanto que su varianza alcanza la cota de Cramér-Rao. Sin embargo, las hipótesis que han llevado a la obtención del MLE son restrictivas y su incumplimiento hace que aumente la varianza de las estimaciones:

- En primer lugar, no es válido para fuentes de naturaleza *discreta* y tiene problemas si la región de soporte de las fuentes es finita, como ocurre con las distribuciones uniformes.

- En segundo lugar, se asume que las observaciones son estadísticamente independientes entre sí. Cuando esto no ocurre, aún podemos dar una justificación teórica a las ecuaciones del MLE y viene dada por el Principio de Infomax, aunque Infomax sea un algoritmo que prefiera maximizar entropías antes que separar fuentes.

- El inconveniente más serio que tiene el MLE es que presupone conocida la *f.d.p* de las fuentes. Estrictamente, la *f.d.p* sería otro parámetro a estimar, al igual que la matriz de mezcla. Sin embargo, mientras que la matriz de mezcla pertenece a un espacio de dimensión N^2, el espacio de las funciones reales tiene *dimensión infinita*. Si la

estimación de la *f.d.p* de las fuentes *no* es apropiada, *no* conseguiremos estimar la matriz de mezcla.

4. Algoritmos de Separación de Fuentes

4.1 Introducción.

Dedicaremos este Capítulo a presentar una selección de los algoritmos de Separación de Fuentes que se basan en las propiedades *estadísticas* de las señales. No consideraremos otros criterios alternativos, como el uso de la caracterización algebraica o geométrica del problema [Puntonet95, Prieto97, Veen98, Zarzoso99]. Por criterios meramente formales, distinguiremos entre algoritmos adaptativos, que se estudiarán en la Sección 4.2 y no adaptativos o algoritmos de bloque, que se tratarán en la Sección 4.3. Finalmente, la Sección 4.4 se reserva para las Conclusiones.

4.2 Los Algoritmos Adaptativos de Separación de Fuentes

En esta Sección tratamos fundamentalmente el algoritmo de "Maximización de la Información" (Infomax) y otros que guardan una relación estrecha con él. En el Apartado 4.2.1 se presenta el algoritmo Infomax propiamente dicho. En el Apartado 4.2.2 se trata el algoritmo EASI, que puede ser considerado una variante de Infomax. En los Apartados 4.2.3 y 4.2.4 se estudia la selección de los parámetros de los algoritmos y su estabilidad asintótica. Finalmente, en los Apartados 4.2.5 y 4.2.6 se presentan los algoritmos Infomax Extendido y MMI, que, como veremos, corrigen las deficiencias de Infomax preservando su potencia. Los algoritmos seleccionados ofrecen una panorámica suficiente del estado actual de la investigación. De hecho, muchos otros métodos no vistos aquí son equivalentes a Infomax [Te-Won98, pág. 67 y ss.] (por ejemplo, el algoritmo de maximización de la

"Negentropía" [Girolami97]) o guardan una relación muy estrecha con él (PCA no lineal [Karhunen94], algoritmos basados en la propiedad de 'bussgang' [Lambert97], etc.).

4.2.1 El algoritmo Infomax

En el Capítulo anterior hemos estudiado con detalle los fundamentos de Infomax, que es el estimador de *máxima verosimilitud* cuando las muestras de cada fuente son *independientes* entre sí. Con referencia a la Figura 4.1, decimos que Infomax es el algoritmo que maximiza la *entropía diferencial* de la variable de salida $\mathbf{u}(t)$ respecto a la matriz $\mathbf{B}$.

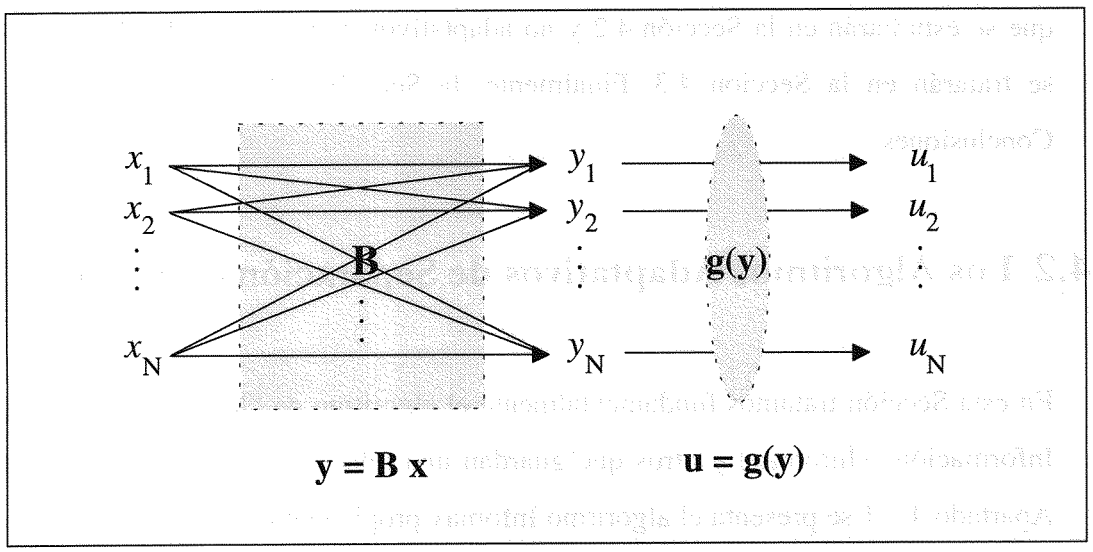

Figura 4.1. Modelo empleado por Infomax

Sea $H[\ \mathbf{u}(t)\ ]$ la entropía diferencial de la variable aleatoria vectorial $\mathbf{u}(t)$. De acuerdo con lo estudiado en el Capítulo anterior,

$$\nabla_{\mathbf{B}} H[\ \mathbf{u}(t)\ ] = (\ \mathbf{B}^{\mathrm{T}}\ )^{-1} - \mathrm{E}[\varphi(\ \mathbf{y}(t)\ )\ \mathbf{x}(t)^{\mathrm{T}}] \qquad (4.1)$$

donde $\nabla_{\mathbf{B}} H[\,\mathbf{u}(t)\,]$ es la matriz $\mathrm{N} \times \mathrm{N}$ cuya componente $(\,i, j\,)$ es $\dfrac{\partial H[\mathbf{u}]}{\partial b_{ij}}$ y hemos definido el vector de dimensiones $\mathrm{N} \times 1$ $\varphi(\,\mathbf{y}(t)\,) = [\,\varphi_1(\,y_1(t)\,), \ldots, \varphi_N(\,y_N(t)\,)\,]^T$ La optimización de $H[\,\mathbf{u}(t)\,]$ se lleva a cabo mediante el algoritmo de *gradiente natural*, es decir,

$$\Delta\mathbf{B} \propto \nabla_{\mathbf{B}} H[\,\mathbf{u}(t)\,]\,\mathbf{B}^T\mathbf{B} = (\,\mathbf{I} - \mathrm{E}[\varphi(\,\mathbf{y}(t)\,)\,\mathbf{y}^T(t)]\,)\,\mathbf{B} \tag{4.2}$$

Finalmente, se suprime el operador "esperanza matemática" para dar lugar a un algoritmo de *gradiente estocástico* [Bell95]:

$$\boxed{\Delta\mathbf{B} \propto (\,\mathbf{I} - \varphi(\,\mathbf{y}(t)\,)\,\mathbf{y}(t)^T\,)\,\mathbf{B}} \tag{4.3}$$

que es, propiamente, la expresión del algoritmo Infomax, donde ' $\propto$ ' denota *proporcionalidad*. Por otra parte, el algoritmo es *idéntico* al (2.36) (pág. 62) y, por ello, puede ser *supereficiente* si se dan las condiciones apropiadas.

Los *puntos estacionarios* del algoritmo son precisamente aquéllos en los que

$$\mathrm{E}[\,\Delta\mathbf{B}\,] = \mathbf{0} \Rightarrow \mathrm{E}[\,\mathbf{I} - \varphi(\,\mathbf{y}(t)\,)\,\mathbf{y}(t)^T\,] = \mathbf{0},$$

entre los que se cuentan las matrices correctas de separación, como ya se ha remarcado varias veces (ver la Sección 2.4.2).

Cuando las muestras vectoriales de las fuentes son estadísticamente independientes entre sí, es decir $\mathbf{s}(t)$ es independiente de $\mathbf{s}(t')$ si t es distinto de t', Infomax es el estimador de máxima verosimilitud de la matriz de mezcla [Cardoso97*a*] (ver la Sección 3.2.1). Bell [Bell95] aconseja que se aleatorice el orden en que se presentan las observaciones al algoritmo, buscando así romper su estructura temporal. Sin embargo, al hacerlo dejamos de tener un algoritmo capaz de trabajar en *tiempo*

real, puesto que tenemos que registrar las observaciones durante un periodo largo de tiempo para poder *barajar* su orden de presentación.

4.2.2 El algoritmo EASI

El algoritmo EASI (acrónimo de 'Equivariant Adaptive Separation via Independence') fue desarrollado por Cardoso y Laheld [Cardoso96*a*]. Este algoritmo optimiza la *función de verosimilitud* de las observaciones bajo la restricción de que la matriz de separación debe ser ortogonal.

Se demostró en el Capítulo 3 (ver la Sección 3.2.2) que la formulación del problema es la siguiente: resolver

$$\frac{1}{T}\sum_{i=1}^{T}\{\ \varphi(\ \mathbf{y}(t)\ )\ \mathbf{y}^{T}(t) - \mathbf{y}(t)\ \varphi(\ \mathbf{y}^{T}(t)\ )\ \} = \mathbf{0}\ \ \textit{sujeto a}\ \mathbf{B}\ \mathbf{B}^{T} = \mathbf{I} \qquad (4.4)$$

donde $\varphi(\ \mathbf{y}(t)\ )$ tiene la misma definición que en (4.1) y T es el número de muestras vectoriales $\mathbf{x}(1)$, ..., $\mathbf{x}(T)$ de que se dispone.

Como se sabe, para que $\mathbf{B}\ \mathbf{B}^{T} = \mathbf{I}$ necesitamos que las observaciones estén incorreladas, es decir

$$E[\ \mathbf{x}(t)\mathbf{x}^{T}(t)\ ] = \mathbf{I} \qquad (4.5)$$

supuesto que la varianza de las fuentes es la *unidad*. Denotemos, momentáneamente, por $\mathbf{z}(t)$ las observaciones tal y como se obtienen de los sensores y que aún no están *incorreladas*. De esta manera, reservamos la denominación $\mathbf{x}(t)$ en exclusiva para las observaciones que verifican (4.5). La Figura 4.2 ilustra las relaciones entre las variables.

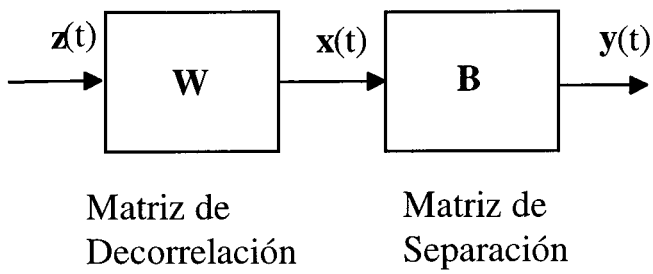

Figura 4.2. Se detalla la decorrelación de las observaciones dentro del esquema general.

El algoritmo de decorrelación que proponen Cardoso y Laheld [Cardoso96*a*] es:

$$\Delta \mathbf{W} \propto -\left(\mathbf{x}(t)\mathbf{x}^{\mathrm{T}}(t) - \mathbf{I} \right) \mathbf{W} \tag{4.6}$$

En efecto, E[$\Delta \mathbf{W}$] = $\mathbf{0}$ cuando E[$\mathbf{x}(t)\mathbf{x}^{\mathrm{T}}(t)$] = $\mathbf{I}$. Merece la pena comentar que multiplicando ($\mathbf{x}(t)\mathbf{x}^{\mathrm{T}}(t) - \mathbf{I}$) por $\mathbf{W}$ se obtiene una *regla de adaptación* en serie como la presentada por Cardoso y Laheld (ver la expresión (2.16) en la pág. 49), con su propiedad de comportamiento uniforme (Corolario 2.2 en la pág. 50). En contraste, la ley de adaptación

$$\Delta \mathbf{W} \propto -\left(\mathbf{x}(t)\mathbf{x}^{\mathrm{T}}(t) - \mathbf{I} \right)$$

no gozaría de tal propiedad. El signo 'menos' que aparece en (4.6) afecta a la estabilidad del algoritmo, que será considerada más adelante.

En un segundo paso, se intenta solucionar la ecuación del *estimador de máxima verosimilitud*. Para ello, se propone el siguiente algoritmo estocástico:

$$\Delta \mathbf{B} \propto -\left(\varphi(\mathbf{y}(t)) \mathbf{y}^{\mathrm{T}}(t) - \mathbf{y}(t) \varphi(\mathbf{y}^{\mathrm{T}}(t)) \right) \mathbf{B} \tag{4.7}$$

Se puede comprobar que esta regla preserva, en primera aproximación, la ortogonalidad de la matriz **B** [Cardoso96*a*].

Finalmente, sea **U** = **B W** la matriz que liga las observaciones originales $\mathbf{z}(t)$ con la salida $\mathbf{y}(t)$ del sistema, esto es, con referencia a la Figura 4.2:

$$\mathbf{y}(t) = \mathbf{U}\,\mathbf{z}(t)$$

Combinando (4.6) y (4.7) es posible demostrar, tras algunas manipulaciones [Cardoso96*a*], que

$$\boxed{\Delta\mathbf{U} \propto -\left(\mathbf{y}(t)\mathbf{y}^{\mathrm{T}}(t) - \mathbf{I} + \varphi(\mathbf{y}(t))\mathbf{y}^{\mathrm{T}}(t) - \mathbf{y}(t)\varphi(\mathbf{y}^{\mathrm{T}}(t)) \right)\mathbf{U}} \qquad (4.8)$$

la cuál también es una regla de *adaptación en serie*. Los puntos estacionarios de este algoritmo son aquéllos en los que $\mathrm{E}[\,\mathbf{y}\mathbf{y}^{\mathrm{T}} - \mathbf{I} + \varphi(\mathbf{y})\,\mathbf{y}^{\mathrm{T}} - \mathbf{y}\,\varphi(\mathbf{y})^{\mathrm{T}}\,] = \mathbf{0}$. Es inmediato comprobar que es el caso de aquéllas matrices **U** que consiguen que las componentes de $\mathbf{y}(t)$ sean *independientes*. Por otra parte, también deducimos que

$$\mathbf{F}(\mathbf{x}(t), \mathbf{B}) = \mathbf{y}(t)\mathbf{y}(t)^{\mathrm{T}} - \mathbf{I} + \varphi(\mathbf{y}(t))\,\mathbf{y}(t)^{\mathrm{T}} - \mathbf{y}(t)\varphi(\mathbf{y}(t))^{\mathrm{T}}$$

es la *función de estimación* asociada a EASI. Es interesante comprobar que esta misma función de estimación se asocia a otros estimadores, como los presentados en los Ejemplos 2.2 y 2.5 (págs. 30 y 40 respectivamente).

4.2.3 Selección de la no linealidad del algoritmo

Hemos visto que las funciones φ_i deben ser escogidas de acuerdo a la *f.d.p* de las fuentes para tener un resultado óptimo; aunque, por definición, esta *f.d.p* es desconocida. En concreto,

$$\varphi_i(y_i) = -\frac{1}{q_i(y_i)} \frac{\partial}{\partial y_i} q_i(y_i)$$

donde q_i es la *f.d.p* marginal de la i-ésima fuente. Para modelar fuentes cuya *f.d.p* es simétrica, resulta habitual utilizar la *distribución de potencia exponencial* o *gaussiana generalizada*, que viene dada por la expresión [Haykin94*a*, pág. 12]:

$$q_i(s_i) = \frac{\gamma}{2} \frac{e^{-|\gamma s_i|^\alpha}}{\Gamma(1+\frac{1}{\alpha})}$$

donde Γ es la función Gamma, $\alpha \geq 0$ y $\gamma > 0$. Se demuestra que esta distribución tiene curtosis positiva para $1 \leq \alpha < 2$, para $\alpha = 2$ la *f.d.p* es gaussiana y para $\alpha > 2$ la curtosis de la distribución es negativa. Así, por ejemplo, para $\alpha = 1$ la distribución es de Laplace mientras que la *f.d.p* uniforme se obtiene haciendo tender α hacia el infinito. La varianza de la distribución se puede ajustar mediante el parámetro γ [Haykin94*a*, pág. 12].

Entonces, resulta que

$$\varphi_i(y_i) = -\frac{q_i'(y_i)}{q_i(y_i)} = \gamma\alpha \, sign(y_i) \, |\gamma y_i|^{\alpha-1} \qquad (4.9)$$

Volviendo a los ejemplos, para la *f.d.p* de Laplace ($\alpha = 1$) se obtiene

$$\varphi_i(y_i) = \gamma\, sign(y_i)$$

En cambio, cuando la *f.d.p* es gaussiana ($\alpha = 2$)

$$\varphi_i(y_i) = 2\gamma^2 y_i$$

que, llevada al algoritmo Infomax da (tras el escalado de las señales)

$$\Delta\mathbf{B} \propto (\mathbf{I} - \mathbf{y}(t)\mathbf{y}(t)^{\mathrm{T}})\,\mathbf{B}$$

que es meramente un algoritmo de decorrelación (compárese con (4.6)). En efecto, variables gaussianas incorreladas *también* son independientes.

Por último, cuando la *f.d.p* es uniforme encontramos que φ_i tiende a cero a medida que α crece (piénsese que $y_i < 1$ y, por lo tanto, $|\gamma y_i|^{\alpha-1}$ es muy pequeño). En este caso, se suele tomar $\alpha = 4$ ó 6, a fin de que $\varphi_i(y_i)$ sea una función impar y, por lo tanto, la estimación sea *supereficiente* (véanse las Secciones 2.4.3 y 2.4.4, págs. 60 y *ss.*).

4.2.4 Estabilidad Asintótica de los Algoritmos

Los puntos de equilibrio de la regla de adaptación Infomax son aquéllos en los que $E[\,\Delta\mathbf{B}\,] = \mathbf{0}$. Concretamente, las *matrices de separación* son (convenientemente escaladas) puntos de equilibrio de la iteración. Sin embargo, aún no hemos estudiado la convergencia de los algoritmos. El estudio se lleva a cabo *linealizando* el algoritmo en torno a alguna de las matrices de separación. Definamos

$$\xi_i = E\left[\varphi_i'(s_i)\right]E[s_i^2] - E[\varphi_i(s_i)s_i] \qquad (4.10)$$

donde $\varphi_i'(s_i)$ denota a la derivada de la función $\varphi_i(s_i)$ respecto a s_i. Se puede demostrar que el algoritmo Infomax es *localmente* estable si se verifican simultáneamente las condiciones [Amari97*b*, Cardoso98*a*]:

$$-1 + \xi_i > 0$$
$$-\xi_i + \xi_j > 0$$

para $1 \leq i < j \leq n$. Curiosamente, las condiciones de estabilidad dependen de las relaciones que se establecen entre *parejas* de fuentes.

El algoritmo EASI es *localmente* estable si se satisface la *segunda condición*; pero no necesita la primera de ellas [Cardoso96*a*].

Ejemplo 4.1. Si $\xi_i < -1$ pero aún se verifica que $\xi_i + \xi_j > 0$, EASI es localmente estable mientras que Infomax es inestable en las matrices de separación. En este sentido, se dice que EASI *es más robusto*.

Para garantizar la *estabilidad* es suficiente que ξ_i sea *positivo* para todo i. Se puede demostrar [Amari97*b*] que, en efecto, ξ_i es mayor o igual que cero si la función φ_i se construye a partir de la auténtica *f.d.p* de las fuentes, dándose la igualdad sólo si la fuente es *gaussiana*. En la práctica, la discrepancia entre las funciones φ_i y la *f.d.p* de las fuentes no debe ser tan grande como para que ξ_i sea *negativa*. En todo caso, [Amari97*b*] propone algoritmos derivados de Infomax que pueden ser estables cuando el algoritmo original no lo es.

Ejemplo 4.2. Consideremos una fuente de distribución *uniforme* que tiene media cero y varianza unidad. La Tabla 4.1 muestra el valor que toma la cantidad ξ_i cuando $\varphi_i(y_i) = \gamma\alpha\, sign(y_i)\,|\gamma y_i|^{\alpha-1}$, como se definió antes en (4.9). Se ha tomado $\gamma = 1$, mientras que α es variable. Recordemos que φ_i corresponde a una distribución de curtosis positiva para $\alpha < 2$ y curtosis negativa para $\alpha > 2$. En concreto, la distribución

uniforme tiene curtosis negativa y la expresión elegida para φ_i se le ajusta tanto mejor cuanto mayor sea α.

α	0	1	2	3	4	5	...	17
ξ_i	0	−3	0	$\frac{9}{2}$	$\frac{48\sqrt{3}}{5}$	45	...	185.895

Tabla 4.1. Estabilidad del algoritmo Infomax para *f.d.p* uniforme

Como muestra la Tabla, Infomax es muy robusto al ser estable para cualquier $\alpha > 2$. Sin embargo, cuando φ_i no está *ajustada con exactitud* a la *f.d.p* de las fuentes, se pierde *eficiencia* (aumenta la varianza de la estimación), como ya hemos discutido.

Ejemplo 4.3. Si $\varphi_i(s_i) = a\, s_i^3$, entonces

$$\xi_i = a(3\,E[s_i^2]\text{–}E[s_i^4]) \quad = -a\,\kappa_{si},$$

siendo κ_{si} la curtosis de la i-ésima fuente. El algoritmo Infomax es estable si el signo de a es el opuesto al de la curtosis. Por lo tanto, cuando se conoce el signo de las curtosis de las fuentes se puede garantizar fácilmente la estabilidad local del algoritmo. Una interpretación interesante se vio con el Ejemplo 2.5 (pág. 40).

4.2.5 Infomax Extendido

Como hemos visto, el comportamiento de Infomax depende de que las funciones φ_i del algoritmo estén bien adaptadas a la *f.d.p* de las fuentes. En general, debemos suponer que esta *f.d.p* es desconocida, por lo que se corre el riesgo de escoger mal las funciones φ_i. En [Pham96] se propone un algoritmo para proyectar φ_i sobre una base previamente seleccionada del espacio de las señales. Otra alternativa la

constituye el algoritmo Infomax Extendido, que aborda de manera sencilla este problema [Girolami97].

Toda densidad de probabilidad simétrica y sub-gaussiana puede ser modelada de acuerdo con la siguiente expresión, propuesta originariamente por Pearson en 1894 [Te-Won98, pág. 43]:

$$p(u) = \frac{1}{2}\left(N(\mu, \sigma^2) + N(-\mu, \sigma^2) \right) \tag{4.11}$$

donde $N(\mu, \sigma^2)$ es una *f.d.p* normal de media $\mu > 0$ y varianza σ^2. La curtosis puede ser ajustada variando μ y σ^2; pero se puede demostrar que siempre es *negativa.*

Resulta que:

$$\varphi(u) = -\frac{1}{p(u)}\frac{\partial p(u)}{\partial u} = \frac{u}{\sigma^2} - \frac{\mu}{\sigma^2}\tanh\left(\frac{\mu}{\sigma^2}u\right) \tag{4.12}$$

donde $\tanh(x)$ representa la tangente hiperbólica de x.

Para fijar ideas, se toma $\mu = 1$ y $\sigma^2 = 1$. Por lo tanto,

$$\varphi(y) = y - \tanh(y) \tag{4.13}$$

Llevando (4.13) a la ley de adaptación (4.3) del algoritmo Infomax se obtiene

$$\Delta\mathbf{B} \propto (\mathbf{I} - \varphi(\mathbf{y})\,\mathbf{y}^T)\,\mathbf{B} = (\mathbf{I} + \tanh(\mathbf{y})\,\mathbf{y}^T - \mathbf{y}\,\mathbf{y}^T)\,\mathbf{B} \tag{4.14}$$

Por las razones expuestas, cabe esperar que esta elección de $\varphi(y)$ sea apropiada para la separación de fuentes de curtosis negativa. De otro lado, gracias a una casualidad muy afortunada, resulta que:

$$\Delta\mathbf{B} \propto (\mathbf{I} - \tanh(\mathbf{y})\,\mathbf{y}^T - \mathbf{y}\,\mathbf{y}^T)\,\mathbf{B} \tag{4.15}$$

está ajustada a la *f.d.p* de curtosis positiva $p(u) \propto p_G(u)\,\mathrm{sech}^2(u)$, donde $p_G(u)$ es una *f.d.p* normal de media cero y varianza unidad. De igual forma, se espera que el algoritmo (4.15) nos permita separar una clase amplia de fuentes de curtosis positiva.

Finalmente, aprovechando que, salvo por el signo de la tangente hiperbólica, las expresiones de (4.14) y (4.15) son idénticas, se propone el siguiente algoritmo (Infomax Extendido) [Girolami97]:

$$\Delta \mathbf{B} \propto (\mathbf{I} + \mathbf{K}\tanh(\mathbf{y})\,\mathbf{y}^T - \mathbf{y}\,\mathbf{y}^T)\,\mathbf{B} \tag{4.16}$$

donde $\mathbf{K}$ es una matriz diagonal tal que $K_{ii} = -1$ si la i-ésima fuente a recuperar tiene curtosis negativa o bien $K_{ii} = +1$ en caso contrario. El signo de cada componente K_{ii} se actualiza iteración a iteración de acuerdo al criterio de estabilidad visto en el Apartado 4.2.4, como sigue: resulta que (4.16) equivale al algoritmo Infomax (4.3) con:

$$\varphi_i(y_i) = y_i + K_{ii}\tanh(y_i) \tag{4.17}$$

Como se vio en el Apartado anterior, garantizamos la estabilidad *local* del algoritmo si

$$\xi_i = \mathrm{E}\left[\varphi_i'(s_i)\right]\mathrm{E}[s_i^2] - \mathrm{E}[\varphi_i(s_i)s_i] > 0 \tag{4.18}$$

para todo i. Al sustituir (4.17) en (4.18) se obtiene una condición *suficiente* para que el algoritmo sea estable [Te-Won98, pág. 47]:

$$K_{ii} = signo\{\ \mathrm{E}[\mathrm{sech}^2(\ s_i\ )]\ \mathrm{E}[\ s_i^2\ ] - \mathrm{E}[\ s_i \tanh(\ s_i\ )\ ]\ \}$$

Puesto que los estadísticos de las fuentes son desconocidos, en la práctica se supone que $s_i(t) \sim y_i(t)$ y, por lo tanto, se toma

$$K_{ii} = signo\{\ \mathrm{E}[\mathrm{sech}^2(\ y_i\ )]\ \mathrm{E}[\ y_i^2\ ] - \mathrm{E}[\ y_i \tanh(\ y_i\ )\ ]\ \}$$

donde la esperanzas se estiman a partir de varias muestras de la señal. Las simulaciones muestran resultados satisfactorios [Te-Won98, pág. 51].

4.2.6 El algoritmo MMI

La información mutua entre las componentes del vector de salida $\mathbf{y}(t) = \mathbf{B}\ \mathbf{x}(t)$ se puede escribir como:

$$I\ [\ \mathbf{y}(t)\ ] = H[\ y_1(t)\ ] + H[\ y_2(t)\ ] + \ldots. + H[\ y_N(t)\ ] - H[\ \mathbf{y}(t)\ ] \quad (4.19)$$

donde $H[\ \mathbf{y}(t)\ ]$ es la entropía diferencial de $\mathbf{y}(t)$ mientras que $H[\ y_i(t)\ ]$ es la entropía marginal de la variable $y_i(t)$. Como sabemos, $I\ [\ \mathbf{y}(t)\ ]$ toma su valor mínimo si y sólo si las componentes del vector son estadísticamente independientes. De aquí que este algoritmo se plantee para minimizar la información mutua $I\ [\ \mathbf{y}(t)\ ]$ (MMI es el acrónimo de 'Minimum Mutual Information') [Amari96].

Dada cualquier variable aleatoria Z, su entropía diferencial $H[\ Z\ ]$ se define

$$H[\ Z\ ] = -\int_{-\infty}^{\infty} p\ (\ z\ ) \log\ p(\ z\ )\ dz \quad (4.20)$$

siendo $p(\ z\ )$ la *f.d.p* de Z. Siempre podemos aproximar $p(\ z\ )$ con un desarrollo de Gram-Charlier o Edgeworth y escribir (ver Apéndice B)

$$p(z) \approx \mathrm{N}(0,1)\,(1 + h(z)) \tag{4.21}$$

donde N(0,1) es la *f.d.p* de una variable gaussiana de media cero y varianza unidad, es decir,

$$\mathrm{N}(0,1) = \frac{1}{\sqrt{2\pi}} \exp(-z^2)$$

Llevando (4.21) a (4.20) se obtiene:

$$H[Z] \approx -\int_{-\infty}^{\infty} \mathrm{N}(0,1)\,(1 + h(z)) \log \{\, \mathrm{N}(0,1)\,(1 + h(z)) \,\}\, dz \tag{4.22}$$

Particularizando $h(z)$ para un desarrollo de Gram-Charlier (ver Apéndice B) se obtiene, tras algunas aproximaciones [Amari96]

$$H[Z] \approx \log \sqrt{2\pi e} - \frac{\kappa_3^2}{54} - \frac{\kappa_4^2}{216} + \frac{5\,\kappa_3^2 \kappa_4}{8} + \frac{\kappa_4^3}{16} \tag{4.23}$$

que es la entropía expresada en función de los cumulantes de Z (en concreto, κ_3 es el coeficiente de asimetría mientras que κ_4 es la curtosis). Evidentemente, (4.23) también nos permite calcular las entropías diferenciales $H[y_i(t)]$ que aparecen en (4.19) sin más que sustituir Z por $y_i(t)$.

Por otra parte, de $\mathbf{y}(t) = \mathbf{B}\,\mathbf{x}(t)$, resulta que aplicando el Teorema 3.1 (pág. 79)

$$H[\mathbf{y}(t)] = H[\mathbf{x}(t)] + \log |\det \mathbf{B}| \tag{4.24}$$

donde $H[\mathbf{x}(t)]$ es la entropía diferencial de $\mathbf{x}(t)$, que no depende de $\mathbf{B}$. Por lo tanto, llevando (4.24) y la expresión aproximada (4.23) de las entropías diferenciales $H[y_i(t)]$ a (4.19), se obtiene la información mutua $I[\mathbf{y}(t)]$ explícitamente en función de los estadísticos de $\mathbf{y}(t)$ y de la matriz $\mathbf{B}$. Entonces, se puede demostrar [Amari96] que

$$\Delta \mathbf{B} \propto (\mathbf{I} - \varphi(\mathbf{y}(t)) \mathbf{y}(t)^{\mathrm{T}}) \mathbf{B}^{-\mathrm{T}} \tag{4.25}$$

es un algoritmo de gradiente estocástico que minimiza $I[\ \mathbf{y}(t)\ ]$, siendo

$$\varphi(\mathbf{y}) = [\varphi(y_1), ..., \varphi(y_N)]^{\mathrm{T}}$$

para

$$\varphi(y) = \frac{3}{4} y^{11} + \frac{25}{4} y^9 - \frac{14}{3} y^7 - \frac{47}{4} y^5 + \frac{29}{4} y^3 \tag{4.26}$$

que es una función impar no monótona. Por último, multiplicando por la derecha (4.25) por $\mathbf{B}^{\mathrm{T}}\mathbf{B}$ conseguimos *un algoritmo de gradiente natural*, llegando a la expresión final:

$$\Delta \mathbf{B} \propto (\mathbf{I} - \varphi(\mathbf{y}(t)) \mathbf{y}^{\mathrm{T}}(t)) \mathbf{B} \tag{4.27}$$

Como vemos, este algoritmo es justamente Infomax. La gran aportación está en proponer una no linealidad (4.26) que es, en principio, apropiada tanto para fuentes sub- como super-gaussianas. La potencia de la función (4.26) está condicionada a la validez de las aproximaciones hechas, especialmente del desarrollo de Gram-Charlier. En el problema de la Separación de Fuentes, el desarrollo de Edgeworth es preferible al de Gram-Charlier [Comon94] y un algoritmo similar que utiliza este último se puede consultar en la referencia [Harroy96]. De todas formas, como (4.26) es una función impar, la estimación siempre es *supereficiente* (ver las Secciones 2.4.3 y 2.4.4, págs. 60 y *ss.*).

La equivalencia del criterio MMI con Infomax no debe sorprendernos. Con relación a (4.19), Infomax maximiza $H[\ \mathbf{y}(t)\ ]$ y, bajo las condiciones expuestas en el Capítulo 3, ello equivale a minimizar $I\ [\ \mathbf{y}(t)\ ]$.

4.3 Algoritmos de Bloque

Esta Sección se dedica a presentar algoritmos de 'bloque' o que no son adaptativos. Todos tienen en común que obtienen la matriz de mezcla a partir de la estimación de un conjunto de estadísticos de las observaciones. En el Apartado 4.3.1 se trata el algoritmo ICA de Comon. Después, en el Apartado 4.3.2 presentamos el algoritmo FastIca. El Apartado 4.3.3 se dedica al algoritmo JADE. Finalmente, en el Apartado 4.3.4 se estudian las relaciones entre algunos de estos algoritmos.

4.3.1 El algoritmo ICA.

Reescribamos la expresión (4.19) de la información mutua entre las componentes del vector de salida $\mathbf{y}(t) = \mathbf{B}\,\mathbf{x}(t)$:

$$I\,[\,\mathbf{y}(t)\,] = H[\,y_1(t)\,] + H[\,y_2(t)\,] + \ldots. + H[\,y_N(t)\,] - H[\,\mathbf{y}(t)\,] \qquad (4.28)$$

siendo $H[\,\mathbf{y}(t)\,]$ la entropía diferencial de $\mathbf{y}(t)$ mientras que $H[\,y_i(t)\,]$ es la entropía marginal de la variable $y_i(t)$. El algoritmo ICA [Comon94] minimiza $I\,[\,\mathbf{y}(t)\,]$ con la restricción de que $\mathbf{B}$ sea una matriz ortogonal, como también hacía EASI. Recordemos que $\mathbf{A}$ también se supone ortogonal por lo que, en suma,

$$\mathbf{R}_x = \mathrm{E}[\,\mathbf{x}(t)\mathbf{x}^T(t)\,] = \mathbf{I} \text{ y, por consiguiente, } \mathbf{R}_y = \mathrm{E}[\,\mathbf{y}(t)\mathbf{y}^T(t)\,] = \mathbf{I} \qquad (4.29)$$

La *negentropía* de una variable aleatoria Z es la distancia de Kullback-Leibler entre la correspondiente densidad de probabilidad $p(\,z\,)$ y la densidad $p_G(\,z\,)$ de la variable gaussiana que tiene los mismos momentos de primer y segundo

orden (ver Apéndice A) y se denota por $nH[\ Z\ ]$. Cuanto más se parezca la distribución de Z a la gaussiana, menor será $nH[\ Z\ ]$.

Como en (4.21), siempre podemos aproximar $p(\ z\ )$ con un desarrollo de Gram-Charlier o Edgeworth y escribir (ver Apéndice B)

$$p(\ z\ ) \approx \mathrm{N}(0,1)\ (\ 1 + h(\ z\ )\ ) \tag{4.30}$$

siendo N(0,1) la *f.d.p* de una variable gaussiana de media cero y varianza unidad. Resulta que, de acuerdo a su definición,

$$nH[\ Z\ ] = \int_{-\infty}^{\infty} p(z)\ \log \frac{p(z)}{\mathrm{N}(0,\ 1)}\, dz \approx \int_{-\infty}^{\infty} p(z)\ \log\ (1 + h(z))\ dz \tag{4.31}$$

Si $h(\ y\ )$ viene dada por la expansión de Edgeworth de tipo A (ver Apéndice B), se puede demostrar [Comon94] que

$$nH[\ Z\ ] \approx \frac{1}{12}\kappa_3^2 + \frac{1}{48}\kappa_4^2 + \frac{7}{48}\kappa_3^4 + \frac{1}{8}\kappa_3^2\kappa_4 \tag{4.32}$$

siendo κ_3 y κ_4, respectivamente, el coeficiente de asimetría y la curtosis de Z. Si la *f.d.p* de Z es simétrica, entonces $\kappa_3 = 0$ y

$$nH[\ Z\ ] \approx \frac{1}{48}\kappa_4^2 \tag{4.33}$$

Se puede demostrar con relativa facilidad que, supuesto (4.29), entonces [Comon94]

$$I\ [\ \mathbf{y}(t)\ ] = nH[\ \mathbf{y}(t)\ ] - \Sigma_{i=1,\mathrm{N}}\ nH[\ y_i(t)\ ] \tag{4.34}$$

siendo $nH[\ \mathbf{y}(t)\ ]$ la negentropía de $\mathbf{y}(t)$ y $nH[\ y_i(t)\ ]$ la negentropía de cada variable $y_i(t)$.

Resulta que $nH[\ \mathbf{y}(t)\ ]$ es una distancia de Kullback-Leibler y, como tal, es invariante ante cualquier transformación invertible que se aplique a la variable vectorial $\mathbf{y}(t)$. Dicho de otra manera, $nH[\ \mathbf{y}(t)\ ]$ no depende de $\mathbf{B}$ [Comon94]. Entonces, minimizar $I\ [\ \mathbf{y}(t)\ ]$ equivale a maximizar la suma de las negentropías de las variables $y_i(t)$. La interpretación es muy interesante, planteada como alternativa al principio de Maximización de la Información (Infomax): de acuerdo al Teorema Central del Límite [Papoulis91, pág. 214] podemos admitir que las observaciones tienen una distribución asintóticamente gaussiana. Entonces, de (4.34) se deduce que hemos separado las fuentes cuando las distribuciones de las señales de salida $y_i(t)$ son tan distintas de la gaussiana como es posible [Girolami97].

Finalmente, de (4.33) y (4.34)

$$\nabla_{\mathbf{B}} I\ [\ \mathbf{y}(t)\ ] \approx -\frac{1}{48} \sum_{i=1,N} \nabla_{\mathbf{B}}\ \kappa_4^2(i) \tag{4.35}$$

siendo $\nabla_{\mathbf{B}}$ el operador que calcula el gradiente con respecto a las componentes de la matriz $\mathbf{B}$ y κ_4 la *curtosis* de $y_i(t)$.

Con un gusto encomiable por las cosas bien terminadas, Comon prueba que, pese a todas las aproximaciones que ha hecho,

$$\psi_{\mathrm{ICA}}(\ \mathbf{B}\ ) = \sum_{i=1,N} \kappa_4^2(i) \tag{4.36}$$

es un *contraste ortogonal discriminante* [Comon94], es decir, *todos* los máximos globales de esta función son realmente *matrices de separación* (ver la Sección 2.2.4 y el Ejemplo 2.3). Finalmente, el contraste se optimiza por medio de un algoritmo que estima las curtosis de las señales de salida en función de la matriz $\mathbf{B}$ y de los estadísticos de las observaciones.

4.3.2 El Algoritmo FastIca

Denotemos, como siempre, por $\mathbf{x}(t) = \mathbf{A}\ \mathbf{s}(t)$ al vector de observaciones. Se supone que las componentes de $\mathbf{x}(t)$ son estacionarias, tienen media cero y la matriz de mezcla $\mathbf{A}$ es ortogonal (y, por lo tanto, la matriz de separación $\mathbf{B}$ también lo es).

La *curtosis* de la observación $x_i(t)$ vale

$$\kappa_4(x_i) = \mathrm{E}[\, x_i^4 \,] - 3\ \mathrm{E}[\, x_i^2 \,]^2 \tag{4.37}$$

donde se omite la dependencia con el tiempo gracias a la hipótesis de estacionariedad de las señales. Sea

$$y_i(t) = \mathbf{b}_i^{\mathrm{T}}\ \mathbf{x}(t) \tag{4.38}$$

la i-ésima señal de salida, donde $\mathbf{b}_i^{\mathrm{T}}$ es la i-ésima fila de la matriz de separación $\mathbf{B}$. En particular,

$$\| \mathbf{b}_i \|_2 = 1$$

La curtosis de $y_i(t)$ será precisamente igual que:

$$\begin{aligned} \mathrm{E}[\, y_i^4 \,] - 3\ \mathrm{E}[\, y_i^2 \,]^2 &= \mathrm{E}[\, (\, \mathbf{b}_i^{\mathrm{T}}\ \mathbf{x}\,)^4 \,] - 3\ \mathrm{E}[\, (\, \mathbf{b}_i^{\mathrm{T}}\ \mathbf{x}\,)^2 \,]^2 = \\ &= \mathrm{E}[\, (\, \mathbf{b}_i^{\mathrm{T}}\ \mathbf{x}\,)^4 \,] - 3\ \| \mathbf{b}_i \|_2^4 \end{aligned} \tag{4.39}$$

Delfosse y Loubaton [Delfosse95] han probado que cuando, por lo menos, una de las fuentes es super-gaussiana y otra es sub-gaussiana, *todos* los puntos extremos de la curtosis de $y_i(t)$ se alcanzan cuando $y_i(t)$ coincide con alguna de las fuentes (ver Ejemplo 2.6 en la pág. 40). Entonces, determinamos $\mathbf{b}_i$ como alguno de los extremos de la función de coste [Hyvärinen97]

$$J(\mathbf{b}_i) = E[(\mathbf{b}_i^T \mathbf{x})^4] - 3 \| \mathbf{b}_i \|_2^4 + F(\| \mathbf{b}_i \|_2) \tag{4.40}$$

siendo $F(\| \mathbf{b}_i \|_2)$ un factor cualquiera de penalización que fuerza $\| \mathbf{b}_i \|_2 = 1$. Derivando $J(\mathbf{b}_i)$ con respecto a $\mathbf{b}_i$ e igualando a cero las derivadas se obtiene la ecuación [Hyvärinen97]

$$E[\mathbf{x} (\mathbf{b}_i^T \mathbf{x})^3] - 3 \| \mathbf{b}_i \|_2^2 \mathbf{b}_i + f(\| \mathbf{b}_i \|_2) \mathbf{b}_i = \mathbf{0} \tag{4.41}$$

donde $f(\| \mathbf{b}_i \|_2)$ es la derivada de $F(\| \mathbf{b}_i \|_2)/2$. Reordenando los términos, se obtiene una nueva ecuación:

$$\mathbf{b}_i = -(E[\mathbf{x} (\mathbf{b}_i^T \mathbf{x})^3] - 3 \| \mathbf{b}_i \|_2^2 \mathbf{b}_i) / f(\| \mathbf{b}_i \|_2) \tag{4.42}$$

que puede ser resuelta mediante una iteración de *punto fijo* [Hyvärinen97]. Es más, la función $f(\| \mathbf{b}_i \|_2)$ es *irrelevante* por cuanto que sólo introduce un factor de escala. Del mismo modo, el signo *menos* que aparece en (4.42) no juega ningún papel importante ya que no podemos verificar el signo de las fuentes. De esta forma, el algoritmo FastIca es aquél que resuelve mediante un algoritmo de punto fijo la ecuación [Hyvärinen97]

$$\boxed{\mathbf{b}_i = E[\mathbf{x} (\mathbf{b}_i^T \mathbf{x})^3] - 3 \| \mathbf{b}_i \|_2^2 \mathbf{b}_i} \tag{4.43}$$

con el cuidado de normalizar a uno el módulo de $\mathbf{b}_i$ en cada iteración.

Sea el vector

$$\mathbf{z}^T = \mathbf{b}_i^T \mathbf{A} \tag{4.44}$$

Es decir,

$$y_i(t) = \mathbf{b}_i^T \mathbf{x}(t) = \mathbf{z}^T \mathbf{s}(t) \tag{4.45}$$

Por supuesto, $\| \mathbf{z} \|_2 = 1$. La separación se consigue siempre que $\mathbf{z}^T$ sea un vector *canónico* (una fila de la matriz identidad). Sea $z_i(n)$ la i-ésima componente del vector $\mathbf{z}^T$ en la n-ésima iteración del algoritmo FastIca. Se puede demostrar que [Hyvärinen97]

$$\frac{|z_i(n)|}{|z_k(n)|} = \sqrt{\gamma} \left(\sqrt{\gamma} \frac{|z_i(0)|}{|z_k(0)|}\right)^{3n} \tag{4.46}$$

siendo γ el módulo del cociente de las curtosis de la k-ésima y la i-ésima fuentes. Si

$$i = \arg\max_j \; |z_j(0)|$$

entonces, teniendo en cuenta que $\| \mathbf{z} \|_2 = 1$, $z_i(n)$ crecerá en cada iteración tendiendo a uno mientras que las restantes componentes del vector $\mathbf{z}$ se harán cero, con lo que $\mathbf{z}$ tenderá a un vector canónico. Es decir, el algoritmo *siempre* converge y separa las fuentes con independencia del signo de las curtosis de las señales (siempre que a lo más una de ellas sea gaussiana). Además, la convergencia es muy rápida, cúbica, como muestra (4.46).

El algoritmo estima *una* de las filas de la matriz de separación. Para determinar las restantes, basta con ejecutar FastIca varias veces, teniendo el cuidado de forzar que las soluciones sean ortogonales a las filas de la matriz $\mathbf{B}$ que han sido obtenidas previamente. Un estudio detallado de las propiedades estadísticas de este estimador, así como algunas generalizaciones, se puede encontrar en [Hyvärinen99].

4.3.3 El algoritmo JADE

Este algoritmo fue propuesto por Cardoso y Souloumiac [Cardoso93] en 1993. Se ha considerado siempre un algoritmo potente y seguro. El análisis se desarrolla como sigue: en primer lugar, se presentan las ideas fundamentales como una generalización del clásico Análisis de Componentes Principales. A continuación, se estudia el algoritmo en sí.

Una extensión del Análisis de Componentes Principales

Dada una matriz **M** cualquiera de dimensiones N × N y el vector N × 1 de observaciones $\mathbf{x}(t)$, que se suponen *incorreladas* ($E[\,\mathbf{x}(t)\mathbf{x}^T(t)\,] = \mathbf{I}$), se define la matriz $\mathbf{N} = \mathbf{Q}(\mathbf{M})$ componente a componente como sigue

$$n_{ij} = \Sigma_{k,l=1,N}\ cum(\, x_i(t), x_j(t), x_k(t), x_l(t)\,)\ m_{lk} \tag{4.47}$$

para todo i,j, siendo n_{ij} el elemento $(\, i, j\,)$ de la matriz **N**, m_{lk} el elemento $(\, l, k\,)$ de la matriz **M** y *cum* es la definición usual de cumulante (ver Apéndice B). Cardoso y Souloumiac [Cardoso93] prueban que la siguiente relación siempre se satisface

$$\mathbf{B}\,\mathbf{N}\,\mathbf{B}^T = \Lambda \tag{4.48}$$

donde **B** es una *matriz ortogonal de separación* y Λ una matriz *diagonal*, cuyas entradas dependen de **M**. Dicho de otra manera, las filas de la matriz de separación son los *autovectores* de cualquier matriz **N** construida como en (4.47). Este resultado es muy potente por cuanto que de él se induce de inmediato un algoritmo de Separación de Fuentes:

"Tómese cualquier matriz **M** y constrúyase a partir de ella **N** = **Q**(**M**) como en (4.47). Los autovectores de **N** forman las filas de la matriz de separación buscada".

Sin embargo, una mala elección de **M** puede hacer que los *autovalores* de **N** se repitan y, por lo tanto, no se pueda obtener un conjunto de N autovectores linealmente independientes. De hecho, el primer algoritmo propuesto por Cardoso [Cardoso89], que recibió el nombre de FOBI (Fourth-Order Blind Identification) tenía este grave inconveniente y no era capaz de separar fuentes que tuviesen la misma curtosis.

Por otra parte, los cumulantes en (4.47) no son conocidos y deben ser estimados. Se debe esperar que los errores en la estimación de los cumulantes se propaguen a la matriz de separación.

Debido a estas razones, Cardoso y Souloumiac deciden incrementar la robustez de su algoritmo utilizando la información de diferentes matrices $\mathbf{N}_i$, todas ellas definidas como en (4.47) para distintas **M**. Los detalles se dejan para el siguiente Apartado.

El Análisis de Componentes Principales [Haykin94*b*, pág. 363] consiste en la extracción de componentes *incorreladas* de una señal y, para ello, calcula los vectores propios de la matriz de *covarianza* de los datos. En este sentido, se puede considerar que Cardoso y Soulomiac definen con (4.47) el equivalente de la matriz de *covarianza* para los estadísticos de cuarto orden y la utilizan para extraer componentes *independientes*.

Diagonalización conjunta de matrices

Cardoso y Souloumiac determinan la matriz de separación como aquélla que minimiza el siguiente contraste [Cardoso93]:

$$\Psi_{\mathrm{JADE}}(\mathbf{B}) = \sum_{i,k,l=1,\mathrm{N}} cum^2(y_i, y_i, y_k, y_l) \tag{4.49}$$

donde *cum* denota la definición usual de cumulante. De hecho, Comon ha probado que la suma del cuadrado de *todos* los cumulantes de orden r ($r = 4$ en este caso) es independiente de $\mathbf{B}$ [Comon94] (ver Ejemplo 2.4 en la página 38). Entonces, maximizar (4.49) equivale a *minimizar* la suma del cuadrado de los cumulantes que tienen al menos *dos* índices distintos. Dicho *mínimo* se alcanza cuando las variables $y_i(t)$ son independientes. Por otra parte, se debe notar el parecido de este contraste con el propuesto por Comon (ver (4.36)):

$$\psi_{\mathrm{ICA}}(\mathbf{B}) = \sum_{i=1,\mathrm{N}} \kappa_4^2(i) = \sum_{i=1,\mathrm{N}} cum^2(y_i, y_i, y_i, y_i)$$

De hecho, como veremos, ambos se pueden considerar equivalentes. La elección del contraste $\psi_{\mathrm{JADE}}(\mathbf{B})$ se justifica a continuación. Resulta [Cardoso93] que existen $\mathrm{P} = \mathrm{N}^2$ matrices $\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_\mathrm{P}$ tales que

$$\mathbf{Q}(\mathbf{P}_i) = \lambda_i \mathbf{P}_i \tag{4.50}$$

donde $\mathbf{Q}(\mathbf{P}_i)$ es como la definida en (4.47). La prueba no es difícil [Cardoso93]: cada par de matrices $\mathbf{P}_i$ y $\mathbf{Q}(\mathbf{P}_i)$ de dimensiones $\mathrm{N} \times \mathrm{N}$ puede ser transformado en dos vectores $\mathbf{p}_i$ y $\mathbf{q}(\mathbf{p}_i)$ de dimensiones $\mathrm{N}^2 \times 1$ concatenando las columnas de las matrices. De acuerdo con (4.47), existe una matriz de cumulantes Θ de dimensiones $\mathrm{N}^2 \times \mathrm{N}^2$ tal que

$$\mathbf{q}(\mathbf{p}_i) = \Theta \mathbf{p}_i \tag{4.51}$$

Como $\mathbf{Q}(\mathbf{P}_i) = \lambda_i \mathbf{P}_i$, entonces $\mathbf{q}(\mathbf{p}_i) = \lambda_i \mathbf{p}_i$, de donde

$$\lambda_i \mathbf{p}_i = \Theta \mathbf{p}_i \tag{4.52}$$

En definitiva, $\mathbf{p}_i$ es un autovector de Θ. La matriz $\mathbf{P}_i$ correspondiente se construye a partir del vector $\mathbf{p}_i$ sin ninguna dificultad añadida. Convengamos en que

$$| \lambda_1 | \geq | \lambda_2 | \geq | \lambda_3 | \geq ... \geq | \lambda_P |$$

Cardoso y Souloumiac [Cardoso97] prueban que λ_i es igual a la *curtosis* de la i-ésima fuente para $i = 1, ..., N$, mientras que los restantes $N (N - 1)$ autovalores λ_j de Θ se anulan. Finalmente, se puede probar que [Cardoso97]

$$\psi_{JADE}(\mathbf{B}) = \sum_{i=1,N} \left| diag(\mathbf{B}^T \mathbf{Q}(\mathbf{P}_i) \mathbf{B}) \right|^2 \qquad (4.53)$$

(compárese con (4.49)). Precisamente, cuando **B** separa las fuentes también está diagonalizando las matrices $\mathbf{Q}(\mathbf{P}_i)$, como se vio en (4.48). De hecho JADE es el acrónimo en inglés de "Joint Approximate Diagonalization of Eigen-Matrices" o "Diagonalización aproximada de Matrices Propias". Las "matrices propias" a que se hace referencia son las matrices $\mathbf{Q}(\mathbf{P}_i) = \lambda_i \mathbf{P}_i$ ya definidas. El término "aproximada" enfatiza el hecho de que, debido a errores en la estimación de los cumulantes, no va a ser posible diagonalizar de forma exacta todas las matrices a la vez.

La diagonalización simultánea [Therrien92, Cardoso93] puede ser llevada a cabo con una generalización del algoritmo de Jacobi [Golub96] para obtener los vectores propios de una única matriz . El coste computacional es, aproximadamente, N veces mayor que el de calcular los autovectores de una única matriz [Cardoso93].

Por último, diremos que es muy sencillo modificar JADE para que trabaje con señales y matrices complejas [Cardoso93].

4.4 Relación entre los Algoritmos

Recordemos el concepto de *función de estimación*: como se vio en (2.1), es toda función $\mathbf{F}(\mathbf{x}(t), \mathbf{B})$ tal que

$$E[\mathbf{F}(\mathbf{x}(t), \mathbf{B})] = \mathbf{0} \tag{4.54}$$

cuando $\mathbf{B}$ es una matriz de separación. Dadas T muestras vectoriales de las observaciones, $\mathbf{x}(1)$, ..., $\mathbf{x}(T)$ y presuponiendo la ergodicidad de los procesos, es posible sustituir (4.54) por la *ecuación de estimación*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbf{F}(\mathbf{x}(t), \mathbf{B}) = \mathbf{0} \tag{4.55}$$

Si determinamos *el gradiente natural* del contraste $\psi_{JADE}(\mathbf{B})$, definido en (4.49) y lo igualamos a cero, estimando los cumulantes a partir de las T muestras de las observaciones, se obtiene [Cardoso97*b*]

$$\frac{1}{T}\sum_{t=1}^{T}\mathbf{F}_{JADE}(\mathbf{x}(t), \mathbf{B}) + \mathbf{O}_{JADE}(T^{-1/2}) = \mathbf{0} \tag{4.56}$$

siendo $\mathbf{O}_{JADE}(T^{-1/2})$ una matriz cuyos elementos son del orden de $T^{-1/2}$ y, por tanto, despreciables si T es suficientemente grande, donde

$$\mathbf{F}_{JADE}(\mathbf{x}, \mathbf{B}) = \mathbf{y}\mathbf{y}^T - \mathbf{I} + \varphi(\mathbf{y})\,\mathbf{y}^T - \mathbf{y}\,\varphi^T(\mathbf{y}) \tag{4.57}$$

siendo $\varphi(\mathbf{y}) = [-\kappa_{s1} y_1^3, -\kappa_{s2} y_2^3, \ldots, -\kappa_{sN} y_{sN}^3]^T$ y κ_{si} la curtosis de la i-ésima fuente. Es decir, (4.57) es el mismo tipo de función de estimación que se asocia a

EASI (ver (4.8)). Además, esta elección de las funciones no lineales $\varphi(\mathbf{y}(t))$ garantizaría la estabilidad local de EASI (ver el Ejemplo 4.3). También es la función de estimación que se asocia al "método de los momentos" (ver el Ejemplo 2.2 en la pág. 30).

Si repetimos el mismo desarrollo para el contraste $\psi_{ICA}(\mathbf{B})$, definido en (4.36) se obtiene una función de estimación tal que

$$\frac{1}{T}\sum_{t=1}^{T}\mathbf{F}_{ICA}(\mathbf{x}(t),\mathbf{B}) + \mathbf{O}_{ICA}(T^{-1/2}) = \mathbf{0} \tag{4.58}$$

donde, asombrosamente, $\mathbf{F}_{ICA}(\mathbf{x}(t),\mathbf{B}) = \mathbf{F}_{JADE}(\mathbf{x}(t),\mathbf{B})$ [Cardoso97*b*]. Esto pone de manifiesto la relación entre ICA y JADE: asintóticamente, para T creciente, obtienen las mismas estimaciones. No obstante, el algoritmo JADE es computacionalmente mucho más eficiente que ICA cuando el número de fuentes es moderado, como prueban las simulaciones. Por otra parte, la relación entre ambos algoritmos y EASI (y, por extensión, con Infomax y el MLE) no debe sorprendernos, pues, a fin de cuentas, el algoritmo ICA de Comon y EASI se basan en el mismo principio: la minimización de la Información Mutua entre las variables de salida.

Por último, el algoritmo FastIca tiene una cierta relación de parentesco con ICA. Hyvärinen [Hyvärinen98] muestra que la función de coste (4.40) de su algoritmo es una aproximación a la negentropía (4.31) de las variables de salida. Por lo tanto, FastIca también minimiza la Información Mutua de las variables de salida.

Por lo tanto, es posible relacionar de forma clara unos algoritmos con otros. No debe sorprendernos que, por lo tanto, en la práctica los resultados que ofrecen sean muy similares.

4.5 Conclusiones

A lo largo de este Capítulo, hemos presentado una selección de los algoritmos de Separación de Fuentes. Implícita o explícitamente, son algoritmos que tratan de

minimizar la información mutua de las variables de salida. La diferencia estriba, por supuesto, en la implementación final de cada uno de ellos.

Suponiendo que las muestras $\mathbf{x}(1)$, ..., $\mathbf{x}(T)$ sean independientes y que la *f.d.p* de las fuentes está correctamente estimada, la minimización de la información mutua es una estimación de máxima verosimilitud de la matriz de mezcla. Por ello, cabe esperar que los algoritmos que estiman los estadísticos de las observaciones tengan un mejor funcionamiento pues, implícitamente, utilizan dichos estadísticos para aproximar la distribución de las fuentes. A cambio, su coste computacional es mayor.

Parte III: Aportaciones Originales

5. Ecuaciones Cuadráticas para la Separación de Fuentes.

5.1 Introducción

La *estimación de máxima verosimilitud* está en la raíz de la mayor parte de los algoritmos de Separación de Fuentes. No obstante, no hay garantía de que la solución obtenida sea siempre correcta, en especial cuando la distribución estadística de las fuentes no se modela con exactitud.

De hecho, vimos ejemplos en el Capítulo 3 en los que las matrices de separación *no* optimizan la *función de verosimilitud* de las muestras. En realidad, en el Capítulo 3 se probó que *las primeras derivadas* de la función de verosimilitud siempre se cancelan cuando las evaluamos en las matrices de separación; pero el carácter de estos puntos críticos (máximos, mínimos y puntos de silla) depende tanto de la distribución estadística de las fuentes como de las funciones no lineales que caracterizan a los algoritmos. Tampoco se ha podido probar que *todos* los máximos globales de las funciones de verosimilitud se correspondan con las matrices que separan las fuentes.

Este Capítulo aborda el problema desde un nuevo enfoque; a través de ecuaciones, plantearemos condiciones que son *necesarias* y *suficientes* para garantizar la Separación de las Fuentes. Nuestra propuesta *aprovecha la información que contienen las derivadas de alto orden* de los cumulantes de las salidas. El Capítulo se desarrolla como sigue: en 5.2 presentamos una primera aplicación, válida para dos fuentes, que servirá para introducir las ideas principales. En 5.3 se generalizan los resultados para cualquier número de fuentes. Finalmente, la Sección 5.4 se dedica a las Conclusiones.

5.2 Separación de dos Fuentes mediante Ecuaciones de Segundo Grado

En el Apartado 5.2.1 vamos a mostrar que la información que contienen las derivadas de los cumulantes de las señales de salida es suficiente para garantizar la Separación. Después, en 5.2.2 se trata la Separación de dos Fuentes mediante la resolución de ecuaciones de *segundo* grado. Finalmente, en 5.2.3 se presenta un algoritmo adaptativo para la Separación de dos fuentes.

5.2.1 Las Derivadas de los Cumulantes de Salida

Sean dos fuentes $s_1(t)$ y $s_2(t)$, de *media cero* y estadísticamente independientes, que se relacionan con las mezclas $x_1(t)$ y $x_2(t)$ como sigue:

$$\mathbf{x}(t) = \mathbf{A}\,\mathbf{s}(t);\quad \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} 1 & a_{12} \\ a_{21} & 1 \end{bmatrix} \begin{bmatrix} s_1(t) \\ s_2(t) \end{bmatrix} \tag{5.1}$$

siendo la matriz de mezcla *invertible* ($1 - a_{12}a_{21} \neq 0$). Para simplificar el desarrollo, hemos supuesto que los elementos de la diagonal de **A** valen uno.

Recuperamos las fuentes con el sistema que se indica en la siguiente ecuación:

$$\mathbf{y}(t) = \mathbf{B}\,\mathbf{x}(t);\quad \begin{bmatrix} y_1(t) \\ y_2(t) \end{bmatrix} = \begin{bmatrix} 1 & b_{12} \\ b_{21} & 1 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} =$$

$$= \begin{bmatrix} 1 & b_{12} \\ b_{21} & 1 \end{bmatrix} \begin{bmatrix} 1 & a_{12} \\ a_{21} & 1 \end{bmatrix} \begin{bmatrix} s_1(n) \\ s_2(n) \end{bmatrix} \tag{5.2}$$

La Figura 5.1 ilustra las relaciones entre las variables, de acuerdo con las ecuaciones (5.1) y (5.2).

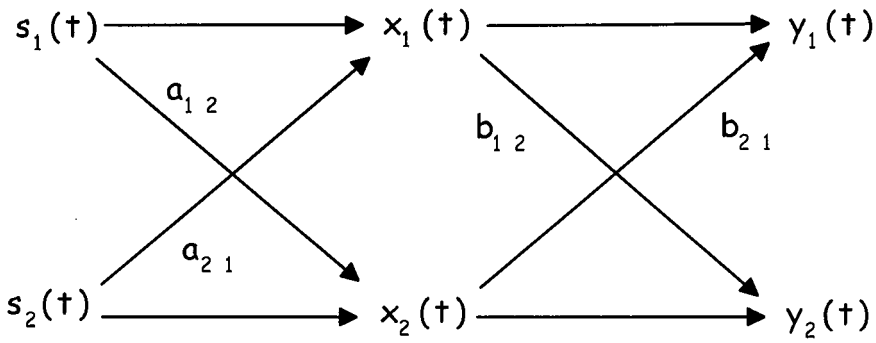

Figura 5.1. Modelos empleados en el desarrollo

Sea **G** = **B A** *la matriz global de transferencia*, que relaciona **s**(t) e **y**(t). Resulta que

$$\mathbf{G} = \begin{bmatrix} 1+b_{12}a_{21} & b_{12}+a_{12} \\ b_{21}+a_{21} & 1+b_{21}a_{12} \end{bmatrix} \tag{5.3}$$

Según (5.3), la *separación* se consigue en cualquiera de los siguientes casos:

$$\begin{cases} b_{12} = -a_{12} \\ b_{21} = -a_{21} \end{cases} \tag{5.5a}$$

o bien

$$\begin{cases} b_{12} = -1/a_{21} \\ b_{21} = -1/a_{12} \end{cases} \tag{5.5b}$$

Nótese que las condiciones (5.5) implican que las fuentes se recuperan *escaladas*. Esto no supone ningún problema, puesto que siempre podemos ajustar su potencia *a posteriori*. Tomemos ahora la función de estimación

$$\mathbf{H}(\mathbf{x}(t);\mathbf{B}) = \varphi(\mathbf{y}(t))\,\mathbf{y}^{T}(t) - \mathbf{D} \tag{5.6}$$

siendo $\varphi(\mathbf{y}(t)) = [\,\varphi_1(y_1(t)), \ldots, \varphi_N(y_N(t))\,]$ y $\mathbf{D}$ una matriz diagonal. Supongamos, como es habitual, que las funciones no lineales φ_i están dominadas por una potencia cúbica, es decir:

$$\varphi(y_i(t)) \approx y_i^3(t) \tag{5.7}$$

Resulta sencillo entonces demostrar que (omitimos la dependencia con t)

$$\mathrm{E}[\,y_1^3 y_2\,] = g_{11}^3 g_{21}\kappa_{s1} + g_{12}^3 g_{22}\kappa_{s2} \tag{5.8a}$$

$$\mathrm{E}[\,y_1 y_2^3\,] = g_{11} g_{21}^3\kappa_{s1} + g_{12} g_{22}^3\kappa_{s2} \tag{5.8b}$$

siendo g_{ij} la componente (i, j) de $\mathbf{G}$ y κ_{si} la curtosis de la i-ésima fuente.

Si suponemos que (5.5*a*) se verifica, resulta que $g_{21} = g_{12} = 0$ mientras que (5.5*b*) lleva a que $g_{11} = g_{22} = 0$. En cualquiera de los dos casos, los estadísticos (5.8) se anulan, como era previsible.

Para fijar nuestras ideas, vamos a centrar el estudio en (5.8*a*). Resulta que

$$\frac{\partial}{\partial b_{12}}\mathrm{E}[\,y_1^3 y_2\,] = 3g_{11}^2 g_{21} a_{21}\,\kappa_{s1} + 3g_{12}^2 g_{22}\kappa_{s2} \tag{5.9a}$$

$$\frac{\partial^2}{\partial b_{12}^2}\mathrm{E}[\,y_1^3 y_2\,] = 6g_{11} g_{21} a_{21}^2\,\kappa_{s1} + 6g_{12} g_{22}\kappa_{s2} \tag{5.9b}$$

$$\frac{\partial^3}{\partial b_{12}^3}\mathrm{E}[\,y_1^3 y_2\,] = 6g_{21} a_{21}^3\,\kappa_{s1} + 6g_{22}\kappa_{s2} \tag{5.9c}$$

Notamos un hecho, cuando menos, *sorprendente*: si $\mathbf{G}$ garantiza la Separación de las Fuentes (es decir, $g_{21} = g_{12} = 0$ ó $g_{11} = g_{22} = 0$) tanto (5.9*a*) como (5.9*b*) *se cancelan.* Por otra parte, obtenemos resultados similares al derivar (5.8*b*) respecto a b_{21}. Todo ello *con independencia de la distribución de las fuentes.*

Entonces, el estadístico (5.8*a*) (respectivamente (5.8*b*)) es poco sensible a las variaciones de b_{12} (respectivamente b_{21}) cerca de las *matrices de separación*. La Figura 5.2 muestra la forma de la superficie (5.8*a*) ((5.8*b*) es similar) en función de las componentes de **B**, así como sus curvas de nivel, supuesto que ambas fuentes son *uniformes* (sus curtosis valen −1'2) y **A** es la matriz identidad, de forma que la separación se consigue cuando $b_{12} = b_{21} = 0$ y $\mathbf{B} = \mathbf{I}$).

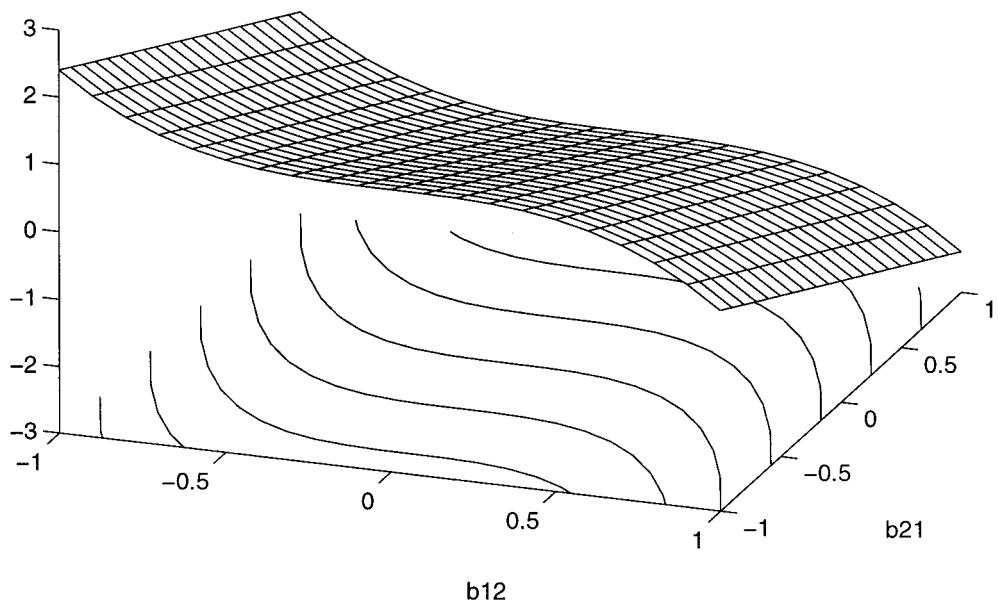

Figura 5.2. Representación de la superficie E[$y_1^3 y_2$].

A la vista de la Figura 5.2, cabe esperar que la convergencia de un algoritmo adaptativo que busque la matriz de separación sea lenta. Sin embargo, no tenemos un mal resultado: (5.8*a*) y (5.8*b*) son polinomios de cuarto grado en los coeficientes de **B** pero (5.9*b*) es *sólo* un polinomio de *segundo grado*. La complejidad de encontrar las *matrices de separación* a partir de (5.9*b*) debe ser, por lo tanto, mucho

menor. A este asunto vamos a dedicar el siguiente Apartado. Antes, consideramos que es interesante destacar que los estadísticos (5.8) son en realidad cumulantes de las señales de salida,

$$\mathrm{E}[\, y_i^3(t)\, y_j(t)] = cum_{31}(y_i(t),\, y_j(t)),\ \text{para}\ i \neq j$$

Las funciones de estimación basadas en la cancelación de cumulantes han sido utilizadas profusamente por Jutten y sus colaboradores [Mansour96], [Nguyen95]. De hecho, Jutten y Mansour [Mansour96] caracterizaron las matrices de separación, *sólo para una mezcla de dos fuentes*, mediante una ecuación polinómica de segundo grado. Por otra parte, Jutten y Nguyen presentan resultados interesantes sobre dichas funciones de estimación en el artículo [Nguyen95].

Si las curtosis de todas las fuentes se anulasen, entonces (5.8*a*) y (5.8*b*) siempre se cancelarían. De todas formas, esto no plantea ninguna dificultad: siempre que la función no lineal $\varphi(\, y_i\,)$ esté dominada por una potencia de y_i de grado mayor o igual que dos se llega a conclusiones similares. Nótese que, en todo caso, no conviene utilizar estadísticos que involucren a los cumulantes de orden impar de las fuentes, puesto que se cancelan siempre que las *f.d.p* sean simétricas respecto a su media.

5.2.2 Obtención de las ecuaciones

Vamos a adoptar la siguiente notación simplificada para los *cumulantes cruzados* de las observaciones $x_1(t)$ y $x_2(t)$:

$$cum_{kl}(\, x_1(t),\, x_2(t)\,) = c_{kl} \tag{5.10}$$

Además, recordemos que la *curtosis* de la fuente $s_k(t)$ es, por definición, el estadístico:

$$\kappa_{si} = cum(s_i^4(n)) = \mathrm{E}[s_i^4(n)] - 3\mathrm{E}^2[s_i^2(n)] \tag{5.11}$$

Utilizando (5.1), (5.2) y (5.11), es sencillo obtener las siguientes relaciones:

$$c_{31} = a_{21}\kappa_{s1} + a_{12}^3\kappa_{s2} \tag{5.12a}$$

$$c_{13} = a_{21}^3\kappa_{s1} + a_{12}\kappa_{s2} \tag{5.12b}$$

$$c_{22} = a_{21}^2\kappa_{s1} + a_{12}^2\kappa_{s2} \tag{5.12c}$$

$$\kappa_{x1} = c_{40} = \kappa_{s1} + a_{12}^4\kappa_{s2} \tag{5.12d}$$

$$\kappa_{x2} = c_{04} = a_{21}^4\kappa_{s1} + \kappa_{s2} \tag{5.12e}$$

Todos los *cumulantes* c_{kl} pueden ser estimados con facilidad a partir del conjunto de *observaciones* disponibles.

Operando de igual forma, se puede demostrar la validez de las siguientes expresiones para los cumulantes cruzados de $y_1(t)$ e $y_2(t)$:

$$cum(y^3{}_1(n),y_2(n)) = c_{31}+b_{21}\kappa_{x1}+3b_{12}[c_{22}+b_{21}c_{31}]+3b^2{}_{12}[c_{13}+b_{21}c_{22}]+b^3{}_{12}[\kappa_{x2}+b_{21}c_{13}] \tag{5.13}$$

$$cum(y_1(n),y^3{}_2(n)) = c_{13}+b_{12}\kappa_{x2}+3b_{21}[c_{22}+b_{12}c_{13}]+3b^2{}_{21}[c_{31}+b_{12}c_{22}]+b^3{}_{21}[\kappa_{x1}+b_{12}c_{31}] \tag{5.14}$$

Al sustituir las relaciones (5.12) en (5.13) y (5.14), se encuentra que, en efecto, (5.13) y (5.14) valen cero cuando $b_{ij} = -a_{ij}$ o $b_{ij} = -1/a_{ji}$, para $i, j = 1, 2$ e $i \neq j$, como ya sabemos.

Resulta mucho más interesante el hecho de que *las siguientes derivadas parciales también se anulan al ser evaluadas en las matrices de separación*

$$\frac{\partial^2 cum(y_1^3(n), y_2(n))}{\partial b_{12}^2} = 6[c_{13} + b_{21}c_{22}] + 6b_{12}[\kappa_{x2} + b_{21}c_{13}] = 0 \tag{5.16}$$

$$\frac{\partial^2 cum(y_1(n), y_2^3(n))}{\partial b_{21}^2} = 6[c_{31} + b_{12}c_{22}] + 6b_{21}[\kappa_{x1} + b_{12}c_{31}] = 0 \qquad (5.17)$$

como puede ser también demostrado trayendo (5.12) a (5.16) y (5.17). Este resultado corrobora la argumentación de los Apartados precedentes.

De las ecuaciones (5.16) y (5.17) se puede derivar una solución inmediata. Primero, despejamos b_{21} de la ecuación (5.17), obteniendo:

$$b_{21} = -\frac{c_{31} + c_{22}b_{12}}{\kappa_{x1} + c_{31}b_{12}} \qquad (5.18)$$

Al sustituir (5.18) en (5.16) obtenemos la siguiente ecuación de segundo grado (se puede encontrar otra ecuación similar para b_{21}):

$$c_{13}\kappa_{x1} - c_{22}c_{31} + [\kappa_{x1}\kappa_{x2} - c_{22}^2]b_{12} + [c_{31}\kappa_{x2} - c_{13}c_{22}]b_{12}^2 = 0 \qquad (5.19)$$

cuyas soluciones,

$$b_{12a} = -a_{12}, \quad b_{12b} = -1/a_{21},$$

son, efectivamente, los coeficientes de las matrices de separación. Como los *cumulantes* c_{ij} pueden ser estimados, se obtiene **B** fácilmente resolviendo (5.19). Por supuesto, la principal virtud de las ecuaciones (5.19) es que *garantizan* que **B** es una matriz de separación correcta.

Por supuesto, la exactitud de la solución depende del grado de acierto en la estimación de los cumulantes. La Figura 5.3 ilustra este aspecto. En ella se muestra *la relación señal a ruido* (SNR), definida como el cociente entre las potencias de la fuente y del error residual tras la separación, frente al número de muestras

utilizadas para estimar los cumulantes. Cada gráfica es el resultado de promediar cincuenta experimentos independientes, en los que la matriz de mezcla y las fuentes se generaron de forma aleatoria. Las fuentes tienen una distribución uniforme. Con línea punteada se muestra la SNR *antes de la separación*, es decir, cuando se supone que las fuentes coinciden con las observaciones.

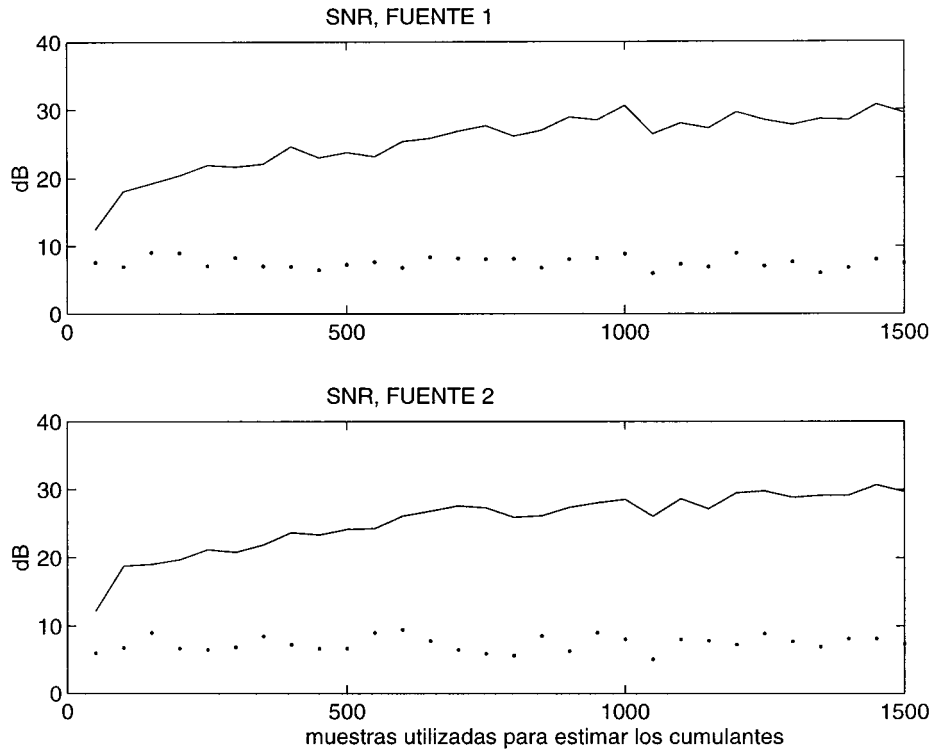

Figura 5.3. Estudio experimental del error cometido al usar (5.19)

Complementariamente, podemos llevar a cabo el siguiente análisis: el error de la raíz σ_i puede ser expresado en términos del error en el coeficiente α_j de la ecuación como:

$$\Delta\sigma_i = \frac{\partial\sigma_i}{\partial\alpha_j}\Delta\alpha_j \tag{5.12}$$

Pues bien, resulta que la derivada parcial en (5.12) es proporcional a [Martín97]:

$$\frac{\sigma_i}{\sigma_i - \sigma_k} \text{ para } k \neq i$$

Por lo tanto, *las raíces son tanto más sensibles cuanto más próximas se encuentren entre sí.* Dado que ambas raíces son $-a_{12}$ y $-1/a_{21}$, diremos que la *precisión disminuye* cuando el producto $a_{12}a_{21}$ *tiende a uno* o, lo que es lo mismo, cuando la matriz de mezcla **A** se hace singular. Además, en cualquier caso, la raíz

$$b_{12} = -1/a_{21}$$

no está bien condicionada si a_{21} tiende a cero. Ambos resultados se mantienen con *independencia de la distribución de las fuentes.*

Mansour y Jutten [Mansour96] han propuesto una ecuación similar que presenta los mismos problemas de sensibilidad. En realidad, que la matriz de separación **B** tenga los coeficientes de su diagonal fijados a uno *no conduce a estimadores equivariantes* [Cardoso95*a*] (ver también el Apartado 2.2.3), por lo que la calidad de la estimación depende de la propia matriz de mezcla.

5.2.3 Resolución adaptativa de las ecuaciones

Las ecuaciones anteriores también pueden ser resueltas mediante un algoritmo adaptativo, como veremos. Vamos a suponer que la varianza de ambas fuentes vale uno y que las observaciones están *incorreladas.* Esto nos lleva a que la mezcla queda descrita por la siguiente relación:

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} \alpha & \beta \\ -\beta & \alpha \end{bmatrix} \begin{bmatrix} s_1(t) \\ s_2(t) \end{bmatrix} \tag{5.13}$$

siendo $\alpha^2 + \beta^2 = 1$. Podemos *decorrelar las observaciones en tiempo real* utilizando cualquiera de los algoritmos adaptativos propuestos para ello (ver, por ejemplo, [Douglas97] y las referencias que contiene), así que la hipótesis no es, de ninguna manera, restrictiva.

Para transformar (5.13) en (5.1) y así aprovechar los resultados anteriores, hacemos los siguientes cambios de variable:

$$s_1(t) \leftarrow \alpha\, s_1(t) \quad \text{y} \quad s_2(t) \leftarrow \alpha s_2(t)$$

Entonces, (5.13) se convierte en

$$\mathbf{x}(t) = \mathbf{A}\ \mathbf{s}(t); \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} 1 & a \\ -a & 1 \end{bmatrix} \begin{bmatrix} s_1(t) \\ s_2(t) \end{bmatrix} \tag{5.14}$$

siendo $a = \beta / \alpha$. Supondremos que $| a | < 1$; lo que significa que la primera fuente está más cerca del primer sensor que la segunda fuente y viceversa (si no fuese así, bastaría con intercambiar los subíndices de las fuentes).

La matriz de mezcla en (5.13) es ortogonal, por lo que el problema de encontrar su inversa está muy bien condicionado. Ello se refleja en que la matriz de mezcla en (5.14) *no* puede ser singular.

Los resultados del Apartado anterior son perfectamente aplicables al modelo (5.14). En particular, las matrices de separación son aquéllas que tienen los coeficientes

$$b_{12} = -a \quad \text{y} \quad b_{21} = a \qquad \text{o bien} \qquad b_{12} = -1/a \quad \text{y} \quad b_{21} = 1/a$$

Vamos a proponer el siguiente algoritmo para buscar las raíces de las ecuaciones (5.8) y (5.9):

$$\mathbf{b}(n+1) = \mathbf{b}(n) + \mu\,[e_1(n)\; e_2(n)]^{\mathrm{T}} \tag{5.15}$$

siendo

$$\mathbf{b}(n) = [b_{12}(n)\; b_{21}(n)]^{\mathrm{T}} \tag{5.16a}$$

$$e_1(n) = y_1(n)y_2(n)x_2^2(n) - b_{21}(n)\sigma_{x1}^2\sigma_{x2}^2 - 3\,b_{12}(n)(\,\sigma_{x2}^2\,)^2 \tag{5.16b}$$

$$e_1(n) = y_1(n)y_2(n)x_1^2(n) - b_{12}(n)\sigma_{x1}^2\sigma_{x2}^2 - 3\,b_{21}(n)(\,\sigma_{x1}^2\,)^2 \tag{5.16c}$$

donde σ_{xi}^2 es la varianza de la i-ésima observación. Resulta que:

$$\mathrm{E}[\,e_1(n)\,] = \frac{1}{6}\frac{\partial^2 cum(y_1^3(n), y_2(n))}{\partial b_{12}^2} \text{ y } \mathrm{E}[\,e_2(n)\,] = \frac{1}{6}\frac{\partial^2 cum(y_1(n), y_2^3(n))}{\partial b_{21}^2}$$

por lo que, en efecto, E[$\mathbf{b}(n+1)$] = E[$\mathbf{b}(n)$] cuando estas derivadas parciales se cancelan: justo *en las matrices de separación*, como ya demostramos. Entonces, las matrices de separación son los *puntos fijos* o *estacionarios* del algoritmo.

El análisis de la convergencia del algoritmo se aborda a continuación.

Lema 5.1. (*Estabilidad de los puntos de equilibrio*) Si ambas fuentes tienen la misma *curtosis*, entonces el algoritmo (5.15) es *globalmente* estable [Martín99*a*].

Demostración.

Se supone que ambas fuentes tienen la misma curtosis: $\kappa_{S1} = \kappa_{S2} = \kappa_S$. Entonces, según (5.12*d*) y (5.12*e*), las curtosis de las observaciones son

$\kappa_{X1} = \kappa_{X2} = \kappa_X$ y, según (5.12*a*) y (5.12*b*), los cumulantes cruzados de las salidas guardan la relación $c_{13} = -\, c_{31}$. Además, por ahora, supondremos que $\kappa_S < 0$ y $\mu > 0$ en (5.15). Vamos a usar la técnica ODE (*Ordinary Diferential Equation*) [Benveniste90, pág. 40 y ss.], es decir, estudiaremos la ecuación diferencial asociada a (5.15):

$$\frac{d}{dt}\,\mathbf{b}(t) = [\mathrm{E}[\, e_1(t)\,]\,\mathrm{E}[\, e_2(t)\,]]^{\mathrm{T}} \qquad \text{(L5.1.1)}$$

Sea V(t) la siguiente función de $b_{12}(t)$ y $b_{21}(t)$:

$$\mathrm{V}(t) = (\, b_{12}(t) + b_{21}(t)\,)^2 \geq 0 \qquad \text{(L5.1.2)}$$

Derivando (L5.1.2) se obtiene (dejamos de indicar la dependencia con t, para simplificar la notación):

$$\frac{d}{dt}\,\mathrm{V} = \frac{\partial \mathrm{V}}{\partial b_{12}}\,\frac{db_{12}}{dt} + \frac{\partial \mathrm{V}}{\partial b_{21}}\,\frac{db_{21}}{dt} =$$

$$= 2\,(\, b_{12} + b_{21}\,)\left(\frac{db_{12}}{dt} + \frac{db_{21}}{dt}\right) \qquad \text{(L5.1.3)}$$

que, operando, lleva a

$$\frac{d}{dt}\,\mathrm{V} = 2\,(\, b_{12} + b_{21}\,)^2\,[\, 1 + 2\,a^2 + a^4\,]\,\kappa_s \qquad \text{(L5.1.4)}$$

Como $\kappa_s < 0$, resulta que $\frac{d}{dt}\,\mathrm{V} < 0$: V($t$) decrece monótonamente. Además, según (L5.1.2), V(t) ≥ 0. Entonces, V(t) ≈ 0 para $t \geq t_0$, siendo t_0 suficientemente grande o, equivalentemente, $b_{21}(\mathrm{t}) = -\, b_{12}(t)$ para $t \geq t_0$. Desgraciadamente, V(t) no es una función de Liapunov ya que no se anula en un único punto sino a lo largo de toda la recta $b_{21}(\mathrm{t}) = -\, b_{12}(t)$,

por lo que aún no hemos demostrado nada. Al sustituir $b_{21}(t)$ por $-b_{12}(t)$ en (L5.1.1) obtenemos la siguiente ecuación diferencial, válida para t mayor o igual que t_0:

$$\frac{d}{dt}b_{12} = b_{12}(\kappa_x - c_{22}) + (1 - b_{12}^2)c_{13} \qquad \text{(L5.1.5)}$$

Sea $v(t) = b_{12}(t) + a$. Si $v(t)$ tiende a cero entonces $b_{12}(t)$ tenderá a $-a$ ($b_{21}(t)$ tenderá a a) y las fuentes quedarán separadas. Se prueba que $v(t)$ satisface la ecuación diferencial [Martín99*a*]:

$$\frac{d}{dt}v = v(A - c_{13}v) \qquad \text{(L5.1.6)}$$

siendo $A = \kappa_x - c_{22} + 2\,a\,c_{13} = (1 - a^4)\,\kappa_s$. Como $|a| < 1$ y $\kappa_s < 0$, resulta que $A < 0$. La ecuación (L5.1.6) es equivalente a:

$$\int \frac{dv}{v(A - c_{13}v)} = \int dt \qquad \text{(L5.1.7)}$$

cuya solución es $v(t) = C\,\dfrac{A\exp(At)}{1 + C\,c_{13}\exp(At)}$, siendo C una constante de integración. Como $A < 0$ resulta que, en efecto, $v(t)$ tiende a cero y *separamos las fuentes*. En conclusión, acabamos de probar que (L5.1.1) es *globalmente estable*. Por lo tanto, el algoritmo estocástico también converge y es *globalmente estable*, siempre que μ cumpla algunas condiciones nada restrictivas [Benveniste90]. La convergencia de este algoritmo es *independiente* de la *f.d.p* de las fuentes, siempre que la curtosis sea negativa. Además, el Lema es destacable por cuanto que la mayoría de los autores sólo prueban la estabilidad *local* de sus algoritmos.

Se ha supuesto $\kappa_s < 0$ y $\mu > 0$. Si, por el contrario, $\kappa_s > 0$ tomando $\mu < 0$, la prueba se mantiene. Si las curtosis de las fuentes son

distintas (incluso en el signo) se puede probar que la estabilidad es local haciendo un desarrollo similar.

5.3 Separación de Fuentes mediante ecuaciones polinómicas de segundo grado

Hemos probado, para dos fuentes, que los coeficientes de la *matriz de separación* satisfacen una ecuación de segundo grado. Ahora vamos a generalizar este resultado para cualquier número de fuentes. En el Apartado 5.3.1 se determina la segunda derivada de los cumulantes de cuarto orden. En el Apartado 5.3.2 discutimos la conveniencia de suponer que las matrices de mezcla y de separación son *ortogonales*. Por último, en el Apartado 5.3.3 se prueba el resultado principal.

5.3.1 Derivadas de los *cumulantes*

Recordemos las ecuaciones del modelo, ahora para N fuentes:

$$\mathbf{x}(t) = \mathbf{A}\,\mathbf{s}(t) \tag{5.17}$$

donde $\mathbf{s}(t)$ y $\mathbf{x}(t)$ son, respectivamente, los vectores $N \times 1$ de fuentes y observaciones y $\mathbf{A}$ es la matriz $N \times N$ de mezcla. El vector de salida $\mathbf{y}(t)$ será obtenido combinando las observaciones como sigue:

$$\mathbf{y}(t) = \mathbf{B}\,\mathbf{x}(t) = \mathbf{G}\,\mathbf{s}(t) \tag{5.18}$$

donde $\mathbf{G} = \mathbf{B}\,\mathbf{A}$ es la matriz global de transferencia del sistema. Vamos a determinar los *cumulantes cruzados* de las componentes del vector $\mathbf{y}(t)$. Utilizando las propiedades de los cumulantes, se prueba que

$$cum_{31}(y_i, y_j) = \sum_{p=1}^{N} g_{ip}^3 \, g_{jp} \, \kappa_{sp} \tag{5.19}$$

donde κ_{sp} es la curtosis de la fuente s_p y g_{ip} es la componente (i,p) de $\mathbf{G}$, es decir:

$$g_{ip} = \sum_{l=1}^{N} b_{il} a_{lp} \tag{5.20}$$

Al igualar (5.19) a cero para todo i, j se obtiene un sistema de ecuaciones, que es satisfecho por toda matriz $\mathbf{G}$ que implique que las fuentes están separadas ($\mathbf{G}$ diagonal, por ejemplo). Desgraciadamente, los cumulantes (5.19) pueden cancelarse aunque las señales $y_i(t)$ e $y_j(t)$ no sean independientes ([Nguyen95]).

Vamos a calcular las derivadas de (5.19). Resulta que, de (5.19) y (5.20),

$$\frac{1}{6}\frac{\partial^2 c_{31}}{\partial b_{ij}^2} = \sum_{p=1}^{N} g_{ip} \, g_{jp} a_{jp}^2 \, \kappa_{sp} \qquad (i \neq j) \tag{5.21}$$

Como $cum(x_l, x_q, x_j, x_j) = \Sigma_p \, a_{lp} a_{qp} a_{jp}^2 \kappa_{sp}$, lo que, de nuevo, es una consecuencia directa de las propiedades de los cumulantes, podemos reescribir (5.21) como sigue:

$$\chi_{ij} \overset{def}{=} \sum_{p=1}^{N} g_{ip} \, g_{jp} a_{jp}^2 \, \kappa_p = \sum_{l=1}^{N} \sum_{m=1}^{N} b_{il} b_{jm} \, cum(x_l, x_m, x_j, x_j) \tag{5.22}$$

En realidad, (5.21) es válida para $i \neq j$; no obstante, vamos a convenir que su consecuencia (5.22) se mantiene incluso para $i = j$, definiendo así implícitamente la cantidad χ_{ii}.

Efectivamente, es sencillo comprobar que (5.22) *se cancela para $i \neq j$ cuando* $\mathbf{G}$ *es una matriz asociada a la separación de fuentes* (volvemos a pensar en una matriz $\mathbf{G}$ diagonal, por ejemplo).

5.3.2 Separación mediante matrices ortogonales

Si la *matriz de separación* fuese completamente arbitraria, el problema tendría N^2 incógnitas a resolver, tantas como elementos hay en la matriz.

De suponer que todos los coeficientes *diagonales* de la matriz de separación son iguales a la *unidad* [Jutten91*a*, Martín97] como en (5.1) la estimación podría *no ser equivariante* [Cardoso95*a*].

Alternativamente, podemos admitir que *la matriz de mezcla es ortogonal.* De esta forma se reduce también el número de grados de libertad del problema, ya que los coeficientes de una matriz ortogonal no son completamente independientes entre sí. Además, una matriz ortogonal *no puede ser singular* y, sobre todo, *no está próxima a ninguna matriz singular*, en el sentido de que el número de condición de las matrices ortogonales es siempre pequeño. Así evitamos el tipo de mezcla, *a priori*, más delicado. Por último, las propiedades de las matrices ortogonales (por ejemplo, que su inversa y su transpuesta coinciden) facilitan mucho el trabajo.

Recordemos que las hipótesis que se formulan son:

- Que la potencia de todas las fuentes vale uno, es decir, $E[s_i^2(t)] = 1$, sin perder por ello generalidad.
- Que las observaciones están incorreladas y su varianza es uno, es decir, $E[\mathbf{x}(t)\mathbf{x}^T(t)] = \mathbf{I}$. En realidad, de $\mathbf{x}(t) = \mathbf{A}\,\mathbf{s}(t)$ resulta que

$$E[\,\mathbf{x}(t)\,\mathbf{x}^T(t)\,] = \mathbf{A}\,E[\,\mathbf{s}(t)\,\mathbf{s}^T(t)\,]\,\mathbf{A}^T = \mathbf{A}\,\mathbf{A}^T = \mathbf{I}$$

Si la matriz de mezcla es ortogonal, entonces *la matriz de separación* también ha de serlo,

$$\mathbf{B}\,\mathbf{B}^T = \mathbf{I}$$

En todo caso, remitimos al lector al Capítulo 2 (ver el Apartado 2.2.2, pág. 32) para una discusión más detallada.

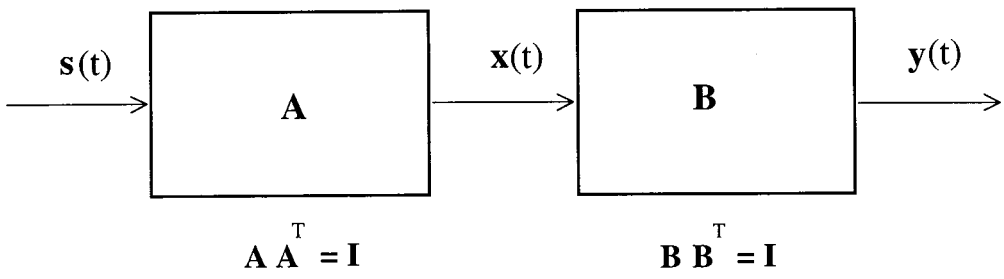

Figura 5.4. Todas las matrices del problema son ortogonales

5.3.3 Ecuaciones polinómicas para la Separación de Fuentes

Entonces, usando que E[$\mathbf{x}(t)\mathbf{x}^T(t)$] = **I** y **B** es ortogonal, (5.22) se puede escribir como:

$$\chi_{ij} = \{ \sum_{l=1}^{N} b_{il}\, b_{jl} \{ \mathrm{E}[x_l^2 x_j^2] - k_l \} + \sum_{m=1, m\neq l}^{N} b_{il}\, b_{jm}\, \mathrm{E}[x_l x_m x_j^2] \} \qquad (5.23)$$

siendo $k_l = 1$ salvo cuando $j = l$ porque, en tal caso, $k_l = 3$. Para estudiar (5.23) el siguiente *lema* nos será de mucha utilidad:

Lema 5.2. Sea $\mathbf{a}_i^T$ la i-ésima fila de la matriz de mezcla **A**. En particular, a_{ij} denota la j-ésima componente de $\mathbf{a}_i^T$. Igualmente, sea Λ_i la matriz diagonal cuya entrada (j, j) es $a_{ij}^2\, \kappa_{sj}$, siendo κ_{sj} la curtosis de la j-ésima fuente. Entonces, se verifica:

$$\mathrm{E}[\,\mathbf{s}(t)\mathbf{s}(t)^{\mathrm{T}} x_i^2(t)] = \Lambda_i + 2\,\mathbf{a}_i\,\mathbf{a}_i^{\mathrm{T}} + \mathbf{I}$$

Demostración.

Por definición, $x_i(t) = \Sigma_j\, a_{ij}\, s_j(t)$, entonces (omitimos la dependencia con t)

$$\mathrm{E}[\,\mathbf{s}\,\mathbf{s}^{\mathrm{T}}\, x_i^2\,] = \mathrm{E}[\,\mathbf{s}\,\mathbf{s}^{\mathrm{T}}\,(\,\Sigma_j\, a_{ij}\, s_j\,)^2\,] =$$

$$= \mathrm{E}[\,\mathbf{s}\,\mathbf{s}^{\mathrm{T}}\,(\,\Sigma_j\, a_{ij}^{\,2} s_j^2 + \Sigma_{j\neq k}\, a_{ij} a_{ik} s_j s_k\,)\,] \qquad \text{(L5.2.1)}$$

Las fuentes son independientes entre sí, tienen media cero y varianza unidad. Por lo tanto, se verifica

1.- $\mathrm{E}[\, s_i s_j s_k s_l\,] = 0$, a no ser que $i = j$, $k = l$, en cuyo caso:

2.- $\mathrm{E}[\, s_i s_i s_k s_k\,] = \mathrm{E}[\, s_i^2\,]\,\mathrm{E}[\, s_k^2\,] = 1$

3.- $\mathrm{E}[\, s_i s_i s_i s_i\,] = \mathrm{E}[\, s_i^4\,] = \kappa_{sl} + 3.$

Teniendo en cuenta estas tres propiedades, se comprueba que

$$\mathrm{E}[\,\mathbf{s}\,\mathbf{s}^{\mathrm{T}}(\,\Sigma_j\, a_{ij}^{\,2} s_j^2\,)\,]$$

es una matriz *diagonal*, cuya componente en la posición $(\,p, p\,)$ vale

$$a_{ip}^{\,2}\,(\kappa_{sp} + 3) + \Sigma_{j\neq p}\, a_{ij}^{\,2} = a_{ip}^{\,2}\,\kappa_{sp} + 2\,a_{ip}^{\,2} + \Sigma_j\, a_{ij}^{\,2} = a_{ip}^{\,2}\,\kappa_{sp} + 2\,a_{ip}^{\,2} + 1$$

donde hemos utilizado que $\|\,\mathbf{a}_i\,\|_2 = 1$. Por otra parte, resulta que el elemento $(\,j, k\,)$ de la matriz

$$\mathrm{E}[\,\mathbf{s}\,\mathbf{s}^{\mathrm{T}}\,\Sigma_{j\neq k}\, a_{ij} a_{ik} s_j s_k\,]$$

es igual que 2 $a_{ij}\, a_{ik}$ siempre que $j \neq k$. Por el contrario, todos los elementos diagonales de esta matriz valen cero.

Uniendo todos estos resultados parciales, la demostración del *Lema* es inmediata.

Ya estamos en condiciones de presentar el resultado principal de este Capítulo. Probaremos que se puede construir una *función de estimación* a partir de las magnitudes χ_{ij} definidas en (5.23). Es más, no sólo las *matrices de separación* solucionan las correspondientes *ecuaciones de estimación* sino que son las *únicas* raíces de las mismas.

Teorema 5.1. (*Ecuaciones cuadráticas para la Separación de Fuentes*). Si, a lo más, una fuente tiene curtosis *nula*, entonces el conjunto de ecuaciones

$$\chi_{ij} = 0$$

para todo i, j ($i \neq j$), estando χ_{ij} definida en (5.23), nos provee de condiciones *necesarias y suficientes* para determinar una matriz de separación [Martín99*b*]

Demostración.

Resulta que χ_{ij} es igual a

$$\chi_{ij} = \mathbf{b}_i^{\mathrm{T}} \left[\mathbf{C}_i - \mathbf{C}_j \right] \mathbf{b}_j \qquad \text{(T5.1.1)}$$

siendo $\mathbf{b}^T_k$ la k-ésima fila de $\mathbf{B}$, y las matrices (se omite la dependencia con t)

$$\mathbf{C}_i = \mathrm{E}[\, \mathbf{x}\,\mathbf{x}^T x_i^2 \,] = \mathbf{A}\,\mathrm{E}[\, \mathbf{s}\,\mathbf{s}^T x_i^2 \,]\,\mathbf{A}^T \qquad \text{(T5.1.2}a\text{)}$$

$$\mathbf{C}_j = 2\,\mathbf{e}_j\mathbf{e}_j^T + \mathbf{I} \qquad \text{(T5.1.2}b\text{)}$$

donde $\mathbf{I}$ es la matriz identidad y $\mathbf{e}_j$ el j-ésimo vector canónico, esto es, la j-ésima columna de $\mathbf{I}$. La matriz $\mathbf{C}_i$ contiene todas las esperanzas matemáticas que aparecen en la expresión de χ_{ij}, mientras que $\mathbf{C}_j$ contiene los términos k_l de (5.23). El Lema 5.2 establece que

$$\mathrm{E}[\, \mathbf{s}\,\mathbf{s}^T x_i^2 \,] = \Lambda_i + 2\,\mathbf{a}_i\,\mathbf{a}_i^T + \mathbf{I} \qquad \text{(T5.1.3)}$$

Entonces, llevando (T5.1.3) a (T5.1.2a) se obtiene:

$$\mathbf{C}_i = \mathbf{A}\,\Lambda_i\,\mathbf{A}^T + \mathbf{C}_j \qquad \text{(T5.1.4)}$$

y, entonces, $$\chi_{ij} = \mathbf{b}_i^T\,[\,\mathbf{C}_i - \mathbf{C}_j\,]\,\mathbf{b}_j = \mathbf{b}_i^T\,\mathbf{A}\,\Lambda_i\,\mathbf{A}^T\,\mathbf{b}_j \qquad \text{(T5.1.5)}$$

La matriz Λ_i sólo se va a cancelar cuando todas las fuentes tengan su curtosis igual a cero o bien cuando $a_{ij}= 0$ para todo j, lo que significaría que la matriz de mezcla es singular. Sin embargo, ninguna de estas hipótesis es válida.

Recordemos que Λ_i es una matriz diagonal cuya entrada en la posición (j, j) es igual a $a_{ij}^2\,\kappa_{sj}$, siendo κ_{sj} la curtosis de la j-ésima fuente. Ya que, como mucho, sólo una de las fuentes tiene su curtosis igual a cero, podemos suponer *sin pérdida de generalidad* que *todos los elementos de la diagonal de* Λ_i *son diferentes*. De no ser así, basta con

ajustar la matriz de mezcla **A** multiplicando las observaciones por alguna matriz ortogonal **U**; así se obtiene una nueva matriz de mezcla, igual al producto **U A**, para la que esta hipótesis sobre la diagonal de Λ_i es admisible. Cardoso encontró un inconveniente similar en el algoritmo FOBI [Cardoso89], predecesor de JADE [Cardoso93]: este algoritmo *no funciona a no ser que las curtosis de todas las fuentes sean diferentes.* Sin embargo, este problema no se puede solucionar, como en nuestro caso, cambiando la matriz de mezcla.

Definamos un nuevo vector $\mathbf{r}_k = \mathbf{A}^T \mathbf{b}_k$. Como tanto **A** como **B** son matrices ortogonales, se sigue que $\mathbf{r}_i^T \mathbf{r}_j = \delta_{ij}$, donde el símbolo δ_{ij} representa a la delta de Kronecker. Por otra parte, es inmediato comprobar que $\chi_{ij} = \mathbf{r}_i^T \Lambda_i \mathbf{r}_j$. La demostración se concluye a continuación:

(*Prueba de Necesidad*) Si **B** es una matriz de separación, entonces cada vector $\mathbf{r}_k$ es un vector canónico diferente, por lo que, en efecto, $\chi_{ij} = 0$ para todo i, j ($i \neq j$).

(*Prueba de Suficiencia*) El vector $\Lambda_i \mathbf{r}_i$ puede ser escrito como una combinación lineal de los vectores $\mathbf{r}_k$, ya que el conjunto $\{\mathbf{r}_1,\mathbf{r}_2,...,\mathbf{r}_N\}$constituye una base ortonormal del espacio. Los coeficientes del desarrollo serán justamente el producto escalar de los vectores $\mathbf{r}_k$ por $\Lambda_i \mathbf{r}_i$; pero esto es, por definición, el valor de χ_{ij}: $\chi_{ij} = \mathbf{r}_i^T (\Lambda_i \mathbf{r}_j)$. Entonces,

$$\Lambda_i \mathbf{r}_i = \chi_{ii} \mathbf{r}_i + \Sigma_{i \neq j} \chi_{ij} \mathbf{r}_j \qquad \text{(T5.1.6)}$$

Supongamos que **B** es una matriz tal que, en efecto, $\chi_{ij} = 0$ para todo i, j ($i \neq j$). Entonces, de (T5.1.6) resulta que $\Lambda_i \mathbf{r}_i = \chi_{ii} \mathbf{r}_i$, lo que implica que $\mathbf{r}_i$

es un autovector de una matriz diagonal. Por lo tanto, $\mathbf{r}_i$ es un vector canónico y, en consecuencia, **B** *es una matriz de separación.*

5.4 Conclusiones.

Al igual que Mansour y Jutten [Mansour96], hemos propuesto una fórmula que da la matriz de separación para dos fuentes. Después, hemos ampliado este resultado mostrando que, sea cual sea el número de las fuentes, las matrices de separación siempre satisfacen ecuaciones polinómicas de segundo grado.

Estas ecuaciones no tienen la forma de las funciones de estimación *eficientes*, es decir, las que son de la forma (ver la Sección 2.4.2)

$$\mathbf{F}(\mathbf{x}(t), \mathbf{B}) = \varphi(\mathbf{y}(t))\,\mathbf{y}^T(t) - \mathbf{I};$$

pero, a cambio, garantizan la Separación con independencia de la *distribución estadística* de las fuentes. En cualquier caso, se mostrará que la precisión de las soluciones es más que suficiente y comparable a la de otros algoritmos.

Dada esta caracterización de las matrices de separación, tenemos que encontrar una forma sencilla de solucionar las ecuaciones. Se propuso un algoritmo adaptativo para dos fuentes; sin embargo, salvo que las curtosis de las fuentes fuesen idénticas, sólo se pudo garantizar la *estabilidad local.* Este hecho hace que optemos por desarrollar otro tipo de métodos, a los que se dedicará el siguiente Capítulo.

6. Ecuaciones Lineales para la Separación de Fuentes

6.1 Introducción

En el Capítulo anterior, hemos presentado el Teorema 5.1, que es fundamental en lo que sigue. Lo enunciamos de nuevo ahora para comodidad del lector: sea, como en (5.23),

$$\chi_{ij} = \{ \sum_{l=1}^{N} b_{il}\, b_{jl} \{ \mathrm{E}[x_l^2 x_j^2] - k_l \} + \sum_{m=1, m\neq l}^{N} b_{il}\, b_{jm}\, \mathrm{E}[x_l\, x_m\, x_j^2] \}$$

$$= \mathbf{b}_i^{\mathrm{T}}\, \mathbf{A}\, \Lambda_i\, \mathbf{A}^{\mathrm{T}}\, \mathbf{b}_j \qquad (6.1)$$

donde $\mathbf{b}_i^{\mathrm{T}}$ es la i-ésima fila de $\mathbf{B}$ y $\Lambda_i = diag(\, a_{i1}^{\,2}\, \kappa_{s1},\, ...,\, a_{iN}^{\,2}\, \kappa_{sN}\,)$, siendo κ_{si} la curtosis de la i-ésima fuente. La segunda igualdad de (6.1) coincide con (T5.1.5), que, a su vez, aparece en la demostración del Teorema 5.1. Entonces,

> **Teorema 5.1** (*Ecuaciones cuadráticas para la Separación de Fuentes*). Si, a lo más, una fuente tiene curtosis, entonces el conjunto de ecuaciones
>
> $$\chi_{ij} = 0$$
>
> para todo i, j ($i \neq j$), estando χ_{ij} definida en (5.23), nos provee de condiciones *necesarias y suficientes* para determinar una matriz de separación.

Para que este resultado sea válido, las observaciones deben estar incorreladas y tener varianza unidad, esto es $E[\ \mathbf{x}(t)\mathbf{x}^T(t)\ ] = \mathbf{I}$. En consecuencia, tanto la matriz de mezcla como la de separación son *ortogonales* (su transpuesta coincide con su inversa).

6.2 Conjunto de ecuaciones lineales

Vamos a definir $\mathbf{J}(\ \mathbf{x}(t),\ \mathbf{B}\ )$ como la matriz cuyo elemento en la posición (i, j) es igual a la cantidad χ_{ij}. Utilizando (6.1) se puede demostrar que

$$\mathbf{J}(\ \mathbf{x},\ \mathbf{B}\ ) = E[\ diag(\ x_1^2, \ldots, x_N^2\ )\ \mathbf{B}\ \mathbf{x}\ \mathbf{x}^T\ ]\ \mathbf{B}^T - \mathbf{B}\mathbf{B}^T - 2\ diag(\ b_{11}, \ldots, b_{NN}\ )\ \mathbf{B}^T \tag{6.2}$$

siendo N el número de fuentes (nótese que, a lo largo del Capítulo, a menudo omitiremos la dependencia explícita de las señales con el tiempo). Cuando se satisfacen las hipótesis del Teorema, es *necesario y suficiente* que $\mathbf{J}(\ \mathbf{x},\ \mathbf{B}\ )$ sea una matriz *diagonal*, esto es,

$$\mathbf{J}(\ \mathbf{x},\ \mathbf{B}\ ) = diag(\ \chi_{11},\ \chi_{22},\ \ldots,\ \chi_{NN}\ ) \tag{6.3}$$

para que $\mathbf{B}$ sea una matriz de separación. No hemos determinado aún las cantidades χ_{ii} para $i = 1, \ldots, N$; así que, por el momento, supondremos que su valor es conocido. Entonces, (6.3) es un sistema de ecuaciones polinómicas de segundo grado del que se puede obtener $\mathbf{B}$.

Como $\mathbf{B}$ es una matriz ortogonal, (6.3) da lugar a la expresión

$$E[\ diag(\ x_1^2, \ldots, x_N^2)\ \mathbf{B}\ \mathbf{x}\ \mathbf{x}^T\ ] - (\ \mathbf{I} + diag(\ \chi_{11},\ \chi_{22},\ \ldots,\ \chi_{NN}\ )\ )\ \mathbf{B} = \\ = 2\ diag(\ b_{11}, \ldots, b_{NN}\ )$$

que, reescrita *fila* a *fila*, resulta en

$$\mathbf{P}_i \mathbf{b}_i = 2\, b_{ii}\, \mathbf{e}_i \tag{6.4}$$

para $i = 1, 2, ..., N$, donde $\mathbf{b}_i^T$ es la i-ésima fila de $\mathbf{B}$, b_{ii} es la i-ésima componente de $\mathbf{b}_i$, $\mathbf{e}_i$ es el i-ésimo vector canónico y hemos definido

$$\mathbf{P}_i = E[\, \mathbf{x}\, \mathbf{x}^T\, x_i^2\,] - \alpha_i\, \mathbf{I}, \text{ donde} \tag{6.5a}$$

$$\alpha_i = 1 + \chi_{ii} \tag{6.5b}$$

Si $\mathbf{P}_i$ es *invertible*, como probaremos, entonces $b_{ii} \neq 0$ ya que, en otro caso, (6.4) implicaría que $\mathbf{b}_i = \mathbf{0}$. Definamos ahora un nuevo vector

$$\mathbf{v}_i = \mathbf{b}_i / 2b_{ii} \tag{6.7}$$

Llevando (6.7) a (6.6) obtenemos el conjunto de *ecuaciones lineales:*

$$\boxed{\mathbf{P}_i \mathbf{v}_i = \mathbf{e}_i} \tag{6.8}$$

para $i = 1, 2, ..., N$. Estas ecuaciones no son idénticas a (6.4) o (6.3); sin embargo, tienen, esencialmente, las mismas soluciones: nótese que $\mathbf{b}_i$ y $\mathbf{v}_i$ son, por definición, vectores *colineales* o, si se prefiere, perpendiculares a las mismas (todas salvo la i-ésima) $N - 1$ filas de la matriz $\mathbf{P}_i$, como indican las ecuaciones (6.4) y (6.8). Por lo tanto, la fila $\mathbf{b}_i^T$ de la matriz de separación guarda la siguiente relación con $\mathbf{v}_i$:

$$\mathbf{b}_i = \pm\, \mathbf{v}_i / \| \mathbf{v}_i \|_2$$

no siendo el signo relevante. En resumen, (6.8) no contiene información sobre el módulo y sentido del vector $\mathbf{b}_i$; pero conserva su dirección. A cambio, (6.8) es un *sistema lineal* de ecuaciones mientras que (6.3) es *cuadrático*.

De hecho, para determinar $\mathbf{v}_i$ sólo hacen falta N − 1 ecuaciones. Por ejemplo, tomando $i = 1$ se obtiene de (6.8):

$$\mathbf{P}_1 \mathbf{v}_1 = \mathbf{e}_1 ; \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} \begin{bmatrix} v_{11} \\ v_{12} \\ v_{13} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

Tan sólo las dos últimas ecuaciones -que dicen que $\mathbf{v}_1$ es perpendicular a la segunda y tercera filas de $\mathbf{P}_1$- son relevantes, ya que determinan la dirección del vector. Como la magnitud de $\mathbf{v}_1$ es indiferente, podemos dar cualquier valor no nulo a v_{11}, por ejemplo $v_{11} = 1/2$, y resolver sólo para v_{12} y v_{13} :

$$\begin{bmatrix} e & f \\ h & i \end{bmatrix} \begin{bmatrix} v_{12} \\ v_{13} \end{bmatrix} = -\frac{1}{2} \begin{bmatrix} d \\ g \end{bmatrix}$$

Esta observación es importante porque gracias a ella reducimos el tamaño de los sistemas de ecuaciones, disminuyendo con ello el tiempo de cómputo.

Es más, los N vectores $\mathbf{v}_i$ son ortogonales entre sí por lo que al conocer N−1 de ellos es posible determinar la dirección del restante, con un gasto computacional menor del que resultaría de resolver las ecuaciones para este vector. En consecuencia, sólo es necesario utilizar (6.8) para $i = 1, 2, ..., N-1$.

Por todo ello, determinamos las N filas de la matriz de separación resolviendo N−1 sistemas de ecuaciones con N−1 incógnitas cada uno. Esto supone un número de operaciones del orden de $(N-1)^4$, siendo N el número de fuentes. Nótese que no está incluido el coste de estimar las matrices $\mathbf{P}_i$,

6.3 Análisis algebraico de las ecuaciones.

Vamos a estudiar el sistema de ecuaciones (6.8). En particular, nos interesa caracterizar mejor sus soluciones y probar que las matrices $\mathbf{P}_i$ no son singulares.

Para cualquier número i entero comprendido entre 1 y N, si, como mucho, *una de las fuentes tiene curtosis nula*, podemos suponer sin pérdida de generalidad que las cantidades $a_{ij}^2 \, \kappa_{sj}$ son diferentes para $j = 1, ..., N$, siendo κ_{sj} la curtosis de la j-ésima fuente y a_{ij} la componente (i, j) de $\mathbf{A}$. La validez de esta hipótesis se discutió ya en la demostración del Teorema 5.1, en el Capítulo anterior: al multiplicar las observaciones $\mathbf{x}(t)$ por cualquier matriz ortogonal, implícitamente cambia el valor de la matriz de mezcla. Entonces, $\mathbf{A}$ *siempre* puede ser ajustada por este procedimiento hasta que $a_{ij}^2 \, \kappa_{sj} \neq a_{ik}^2 \, \kappa_{sk}$ si $j \neq k$. De la misma forma, supondremos que $a_{ij} \neq 0$ para todo j, aunque esta última hipótesis podría ser relajada con facilidad. Ambas suposiciones se mantendrán a lo largo de toda la Sección.

De acuerdo con (6.5*a*) y (6.5*b*), la matriz $\mathbf{P}_i$ que contiene los coeficientes de las ecuaciones se define como

$$\mathbf{P}_i = E[\, \mathbf{x} \, \mathbf{x}^T \, x_i^2 \,] - \alpha_i \, \mathbf{I}$$

siendo $\alpha_i = 1 + \chi_{ii}$. Según (6.1), $\chi_{ii} = \mathbf{b}_i^T \, \mathbf{A} \, \Lambda_i \, \mathbf{A}^T \, \mathbf{b}_i$, donde $\mathbf{b}_i^T$ es la i-ésima fila de $\mathbf{B}$ y Λ_i es la matriz *diagonal* cuya entrada en la posición (i,i) vale $a_{ii}^2 \, \kappa_{si}$. Cuando $\mathbf{B} = \mathbf{A}^T$, la inversa de la matriz de mezcla, es inmediato comprobar que

$$\chi_{ii} = a_{ii}^2 \, \kappa_{si} \tag{6.9}$$

de donde

$$\mathbf{P}_i = \mathrm{E}[\, \mathbf{x}\,\mathbf{x}^{\mathrm{T}} x_i^2 \,] - (1 + a_{ii}^2 \kappa_{si})\, \mathbf{I} \tag{6.10}$$

o bien, teniendo en cuenta que $\mathbf{x} = \mathbf{A}\,\mathbf{s}$ y $\mathbf{A}\,\mathbf{A}^{\mathrm{T}} = \mathbf{I}$,

$$\mathbf{P}_i = \mathbf{A} \left(\mathrm{E}[\, \mathbf{s}\,\mathbf{s}^{\mathrm{T}} x_i^2 \,] - (1 + a_{ii}^2 \kappa_{si})\, \mathbf{I} \right) \mathbf{A}^{\mathrm{T}} \tag{6.11}$$

El Lema 5.2 del Capítulo anterior afirma que

$$\mathrm{E}[\, \mathbf{s}\,\mathbf{s}^{\mathrm{T}} x_i^2 \,] = \mathbf{I} + \Lambda_i + 2\, \mathbf{a}_i\, \mathbf{a}_i^{\mathrm{T}} \tag{6.12}$$

siendo $\Lambda_i = diag(a_{i1}^2 \kappa_{s1}, \ldots, a_{iN}^2 \kappa_{sN})$, donde $\mathbf{a}_i^{\mathrm{T}}$ es la i-ésima fila de $\mathbf{A}$. Entonces, llevando (6.12) a (6.11) obtenemos

$$\mathbf{P}_i = \mathbf{A} \left(\Lambda_i - (a_{ii}^2 \kappa_{si})\, \mathbf{I} + 2\, \mathbf{a}_i\, \mathbf{a}_i^{\mathrm{T}} \right) \mathbf{A}^{\mathrm{T}} \tag{6.13}$$

Por construcción, $\Lambda_i - (a_{ii}^2 \kappa_{si})\, \mathbf{I}$ es una matriz cuya i-ésima columna sólo contiene ceros. Esto nos permite afirmar que

$$(\Lambda_i - (a_{ii}^2 \kappa_{si})\, \mathbf{I})\, \mathbf{e}_i = \mathbf{0}$$

y, por lo tanto,

$$(\Lambda_i - (a_{ii}^2 \kappa_{si})\, \mathbf{I} + 2\, \mathbf{a}_i\, \mathbf{a}_i^{\mathrm{T}})\, \mathbf{e}_i = 2\, a_{ii}\, \mathbf{a}_i \tag{6.14}$$

Retomemos ahora el argumento principal. Planteemos de nuevo el *sistema de ecuaciones lineal* (6.8),

$$\mathbf{P}_i\, \mathbf{v}_i = \mathbf{e}_i \tag{6.15}$$

que, trayendo (6.13) a (6.15), se convierte en:

$$\mathbf{A}\,(\,\Lambda_i - (\,a_{ii}^{\,2}\,\kappa_{si}\,)\,\mathbf{I} + 2\,\mathbf{a}_i\,\mathbf{a}_i^{\mathrm{T}}\,)\,\mathbf{A}^{\mathrm{T}}\,\mathbf{v}_i = \mathbf{e}_i \tag{6.16}$$

de donde, tras multiplicar por $\mathbf{A}^{\mathrm{T}}$ ambos lados de la igualdad

$$(\,\Lambda_i - (\,a_{ii}^{\,2}\,\kappa_{si}\,)\,\mathbf{I} + 2\,\mathbf{a}_i\,\mathbf{a}_i^{\mathrm{T}}\,)\,\mathbf{A}^{\mathrm{T}}\,\mathbf{v}_i = \mathbf{a}_i \tag{6.17}$$

siendo $\mathbf{a}_i$ la i-ésima fila de $\mathbf{A}$. De la comparación de (6.17) y (6.14), se deduce que la solución de (6.17) viene dada precisamente por

$$\mathbf{A}^{\mathrm{T}}\,\mathbf{v}_i = \mathbf{e}_i / 2\,a_{ii} \;\Rightarrow\; \mathbf{v}_i = \mathbf{A}\,\mathbf{e}_i / 2\,a_{ii}, \tag{6.18}$$

es decir, $\mathbf{v}_i$ es proporcional a la i-ésima columna de $\mathbf{A}$. Este resultado era, en realidad, previsible atendiendo a la discusión realizada en la Sección 6.2. Resulta más interesante la siguiente generalización:

Retomemos la matriz de coeficientes $\mathbf{P}_i$,

$$\mathbf{P}_i = \mathrm{E}[\,\mathbf{x}\,\mathbf{x}^{\mathrm{T}}\,x_i^{\,2}\,] - \alpha_i\,\mathbf{I}$$

Siendo $\alpha_i = 1 + \chi_{ii}$. Según (6.1)

$$\chi_{ii} = \mathbf{b}_i^{\mathrm{T}}\,\mathbf{A}\,\Lambda_i\,\mathbf{A}^{\mathrm{T}}\,\mathbf{b}_i$$

donde $\Lambda_i = diag(\,a_{i1}^{\,2}\,\kappa_{s1},\,...,\,a_{iN}^{\,2}\,\kappa_{sN}\,)$. Al suponer que la matriz de separación $\mathbf{B}$ es igual que $\mathbf{A}^{\mathrm{T}}$, hemos obtenido que $\chi_{ii} = a_{ii}^{\,2}\,\kappa_{si}$; pero, en realidad, éste no es el único valor admisible de $\mathbf{B}$. En general, el producto $\mathbf{A}^{\mathrm{T}}\,\mathbf{b}_i$ puede ser un vector canónico cualquiera, así que χ_{ii} puede tomar los siguientes valores:

$$a_{ij}^2 \kappa_{sj} \tag{6.19}$$

para $j = 1, ..., N$. Pues bien, repitiendo los desarrollos previos resulta inmediato demostrar el siguiente Lema:

Lema 6.1. Sea $\alpha_i = 1 + a_{ij}^2 \kappa_{sj}$, donde j es cualquier número entero entre 1 y N y definamos $\mathbf{P}_i = E[\mathbf{x}\mathbf{x}^T x_i^2] - \alpha_i \mathbf{I}$. Entonces, la solución del sistema de ecuaciones

$$\mathbf{P}_i \mathbf{v}_i = \mathbf{e}_i$$

donde i es cualquier número entero entre 1 y N es proporcional a la j-ésima columna de la matriz de mezcla $\mathbf{A}$,

$$\mathbf{v}_i = \mathbf{A}\,\mathbf{e}_j / 2\,a_{ij}$$

Es decir, si estimásemos la matriz $\Lambda_i = diag(a_{i1}^2 \kappa_{s1}, ..., a_{iN}^2 \kappa_{sN})$, el Lema 6.1 nos garantiza que podremos separar con seguridad las fuentes. Abordaremos este problema en la próxima Sección. Antes, veamos un último resultado.

Lema 6.2. Las matrices $\mathbf{P}_i$ definidas en el Lema 6.1 son *invertibles* para todo j.

Demostración. Del Lema 6.1 y (6.12) resulta que

$$\mathbf{P}_i = \mathbf{A} \left(\Lambda_i - \left(a_{ij}^2 \kappa_{sj} \right) \mathbf{I} + 2 \mathbf{a}_i \mathbf{a}_i^T \right) \mathbf{A}^T$$

Sea $\mathbf{Q}_i = \mathbf{A}^T \mathbf{P}_i \mathbf{A} = \left(\Lambda_i - \left(a_{ij}^2 \kappa_{sj} \right) \mathbf{I} + 2 \mathbf{a}_i \mathbf{a}_i^T \right)$. Basta demostrar que $\mathbf{Q}_i$ es invertible, es decir, $\mathbf{Q}_i \mathbf{r} = \mathbf{0}$ si y sólo si $\mathbf{r} = \mathbf{0}$. Que

$$\mathbf{Q}_i \mathbf{r} = \mathbf{0} \Rightarrow \{ \Lambda_i - \left(a_{ij}^2 \kappa_{sj} \right) \mathbf{I} \} \mathbf{r} = -2 \mathbf{a}_i \{ \mathbf{a}_i^T \mathbf{r} \} \qquad \text{(L6.2.1)}$$

A la izquierda, multiplicamos $\mathbf{r}$ por una matriz cuya j-ésima columna sólo contiene ceros. En consecuencia, el j-ésimo elemento del vector resultante es cero. Por lo tanto, la misma componente de $2 \mathbf{a}_i \{ \mathbf{a}_i^T \mathbf{r} \}$ se anula, $2 a_{ii} \{ \mathbf{a}_i^T \mathbf{r} \} = 0$ y, como $a_{ij} \neq 0$ por hipótesis, $\Rightarrow \mathbf{a}_i^T \mathbf{r} = 0$. Llevando este resultado a (L6.2.1) obtenemos

$$\{ \Lambda_i - \left(a_{ii}^2 \kappa_{si} \right) \mathbf{I} \} \mathbf{r} = \mathbf{0} \qquad \text{(L6.2.2)}$$

La columna j-ésima de la matriz $\{ \Lambda_i - \left(a_{ii}^2 \kappa_{si} \right) \mathbf{I} \}$ se anula. Es más, sólo esta columna vale cero debido a que las cantidades $a_{ij}^2 \kappa_{sj}$ son distintas por hipótesis. Como esta matriz es *diagonal,* (L6.2.2) sólo se va a satisfacer cuando $\mathbf{r} = \lambda \mathbf{e}_j$, siendo λ un factor de escala. Ahora bien, como $\mathbf{a}_i^T \mathbf{r} = 0$ y $a_{ij} \neq 0$ se concluye que $\lambda = 0$. Por lo tanto, necesariamente, $\mathbf{r} = \mathbf{0}$. En conclusión, $\mathbf{P}_i$ es, efectivamente *invertible.*

6.4 Estimación de los parámetros de las ecuaciones

Hemos supuesto que los parámetros α_i de las ecuaciones son conocidos. Sin embargo, en la práctica los parámetros α_i deben ser estimados a partir de las observaciones $\mathbf{x}(t)$. Para ello, hay que relacionarlos con alguna cantidad que podamos medir. Resulta que

$$E[\, \mathbf{x}\,\mathbf{x}^T x_i^2 \,] = \mathbf{A}\, E[\, \mathbf{s}\,\mathbf{s}^T x_i^2 \,]\, \mathbf{A}^T \tag{6.20}$$

Como $\mathbf{A}$ es una matriz ortogonal, las matrices $E[\, \mathbf{x}\,\mathbf{x}^T x_i^2 \,]$ y $E[\, \mathbf{s}\,\mathbf{s}^T x_i^2 \,]$ tienen los mismos autovalores. Además, el Lema 5.2 del anterior Capítulo afirma que

$$E[\, \mathbf{s}\,\mathbf{s}^T x_i^2 \,] = \mathbf{I} + \Lambda_i + 2\, \mathbf{a}_i\, \mathbf{a}_i^T$$

siendo $\Lambda_i = diag(\, a_{i1}^2\, \kappa_{s1},\ ...,\ a_{iN}^2\, \kappa_{sN}\,)$. Sabemos que α_i puede tomar N valores distintos ($\alpha_i = 1 + a_{ij}^2\, \kappa_{sj}$, $j = 1,\ ...\ ,\ N$): fijémonos ahora que son justamente los elementos de la diagonal de $\mathbf{I} + \Lambda_i$. Como en la Sección anterior, suponemos que las cantidades $a_{i1}^2\, \kappa_{s1},\ ...,\ a_{iN}^2\, \kappa_{sN}$ son todas distintas.

El siguiente Teorema es clave:

Teorema 6.2. (*Estimación de los parámetros* α_i) Sea λ_i el i-ésimo autovalor de $E[\, \mathbf{x}\,\mathbf{x}^T x_i^2 \,]$, donde $\lambda_1 \geq ... \geq \lambda_N$. Igualmente, supondremos que $a_{i1}^2\, \kappa_{s1} > ... > a_{iN}^2\, \kappa_{sN}$. Entonces,

a) $\quad 0 \leq \lambda_k - (\, 1 + a_{ik}^2\, \kappa_{sk}\,) \leq 2$, para $k = 1,\ ...,\ N$

b) $\quad \lambda_k \in [\, 1 + a_{ik-1}^2\, \kappa_{sk-1},\ 1 + a_{ik}^2\, \kappa_{sk}\,]$, para $k = 2,\ ...,\ N$

Demostración. Utilizando (6.20) y el libro de Golub y Van Loan (Teoremas 8.1.5 y 8.1.8)

Por lo tanto, las cantidades α_i pueden ser aproximadas (quizás de forma burda) por los autovalores de $E[\mathbf{x}\,\mathbf{x}^T x_i^2]$. De todas formas, veremos que esta caracterización es suficiente para desarrollar algoritmos *muy robustos*.

6.5 Análisis de Sensibilidad

El Teorema 6.2 no permite determinar con total exactitud el valor de los parámetros α_i. Teniendo esto en cuenta, dedicaremos esta Sección a estudiar la precisión de las soluciones que se obtienen al resolver las ecuaciones (6.8).

Recordemos que α_i debe tomar uno cualquiera de los siguientes valores, como prueba el Lema 6.1:

$$1 + a_{ij}^2 \kappa_{sj} \text{ para } j = 1, ..., N$$

Utilizando el Teorema 6.1 sólo se puede determinar $a_{ij}^2 \kappa_{sj}$ de forma aproximada, así que, en general, sólo podemos afirmar que $\alpha_i \approx 1 + a_{ij}^2 \kappa_{sj}$ para algún (algunos) valor (valores) de j. Sea

$$\varepsilon_{ij} = 1 + a_{ij}^2 \kappa_{sj} - \alpha_i \tag{6.21}$$

el error de estimación. Se supone que $\varepsilon_{ij} \neq 0$ para todo j.

A partir de (6.11) es sencillo introducir los errores ε_{ij} en la expresión de la matriz de coeficientes $\mathbf{P}_i$

$$\mathbf{P}_i = \mathbf{A} \left(E[\mathbf{s}\,\mathbf{s}^T x_i^2] - \alpha_i \mathbf{I} \right) \mathbf{A}^T =$$

$$= \mathbf{A} \left(\mathbf{I} + \Lambda_i - \alpha_i \mathbf{I} + 2\, \mathbf{a}_i\, \mathbf{a}_i^T \right) \mathbf{A}^T =$$
$$= \mathbf{A} \left(\mathbf{E} + 2\, \mathbf{a}_i\, \mathbf{a}_i^T \right) \mathbf{A}^T \qquad (6.22)$$

siendo $\mathbf{E} = diag(\, \varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{iN} \,)$. El *Lema de Inversión de Matrices* [Golub96], nos permite calcular la *inversa* de $\mathbf{P}_i$:

$$\mathbf{P}_i^{-1} = \mathbf{A} \left(\mathbf{E}^{-1} - 2 \frac{\mathbf{E}^{-1} \mathbf{a}_i \mathbf{a}_i^T \mathbf{E}^{-1}}{1 + 2\, \mathbf{a}_i^T \mathbf{E}^{-1} \mathbf{a}_i} \right) \mathbf{A}^T \qquad (6.23)$$

Llevando esta expresión al sistema de ecuaciones (6.8) se obtiene:

$$\mathbf{P}_i\, \mathbf{v}_i = \mathbf{e}_i \Rightarrow \mathbf{v}_i = \mathbf{P}_i^{-1}\, \mathbf{e}_i$$

de donde, tras algunas operaciones,

$$\mathbf{v}_i = \mathbf{A}\, \mathbf{E}^{-1}\, \mathbf{a}_i / k =$$
$$= \mathbf{A} \left[\, a_{i1}/\varepsilon_{i1}, \ldots, a_{iN}/\varepsilon_{iN} \, \right]^T / k \qquad (6.24)$$

siendo k una constante escalar cuyo valor es irrelevante. Llegados a este punto, un razonamiento cualitativo es suficiente para nuestros fines:

Si ε_{i1}, por ejemplo, fuese mucho más pequeño, en módulo, que $\varepsilon_{i2}, \ldots, \varepsilon_{iN}$, la siguiente aproximación sería válida

$$\left[\, a_{i1}/\varepsilon_{i1}, \ldots, a_{iN}/\varepsilon_{iN} \, \right]^T \approx a_{i1}/\varepsilon_{i1} \left[\, 1, 0, \ldots, 0 \, \right]^T$$

de tal forma que $\mathbf{v}_i$ vendrá a ser una buena aproximación a la primera columna de $\mathbf{A}$, salvo por el factor de escala $a_{i1} / (k\, \varepsilon_{i1})$. En cambio, cuando $| \varepsilon_{i1} | \approx | \varepsilon_{i2} | << | \varepsilon_{ij} |$ para $j \neq 1, 2$ obtenemos que $\mathbf{v}_i$ es una combinación de la primera y segunda columnas de $\mathbf{A}$. Que $| \varepsilon_{i1} | \approx | \varepsilon_{i2} |$ ocurre cuando $a_{i1}^2\, \kappa_{s1} \approx a_{i2}^2\, \kappa_{s2}$, lo que hace que la matriz $\mathbf{P}_i$ *esté mal condicionada.*

Así pues, el grado de acierto de las soluciones disminuye cuando las cantidades $a_{ij}^2 \kappa_{sj}$ no están bien diferenciadas.

Ejemplo 6.1. Consideremos una mezcla de dos fuentes uniformes, de media cero y varianza unidad (sus curtosis son iguales, $\kappa_{s1} = \kappa_{s2} = \kappa = -1'2$). La matriz de mezcla ortogonal **A** se supone igual que

$$\mathbf{A} = \begin{bmatrix} a & b \\ -b & a \end{bmatrix}, \qquad \sqrt{a^2 + b^2} = 1$$

Con esta notación, resulta que $a_{11}^2 \kappa_{s1} = a^2 \kappa$ y $a_{12}^2 \kappa_{s2} = b^2 \kappa$. Si la mezcla no es fuerte, esto es, por ejemplo, $b \sim 0$, entonces $1 + a^2 \kappa \sim 0$ y, por ello, parece que tomar $\alpha_i = 0$ es una buena aproximación al valor de $1 + a^2 \kappa$. Cabe esperar, por tanto, que las ecuaciones den un buen resultado. Por el contrario, cuando a^2 y b^2 son parecidos, esto es, cuando $a^2 \approx 1/2$ ($a \approx 0'7$), se debe pensar que la calidad de la Separación será mala.

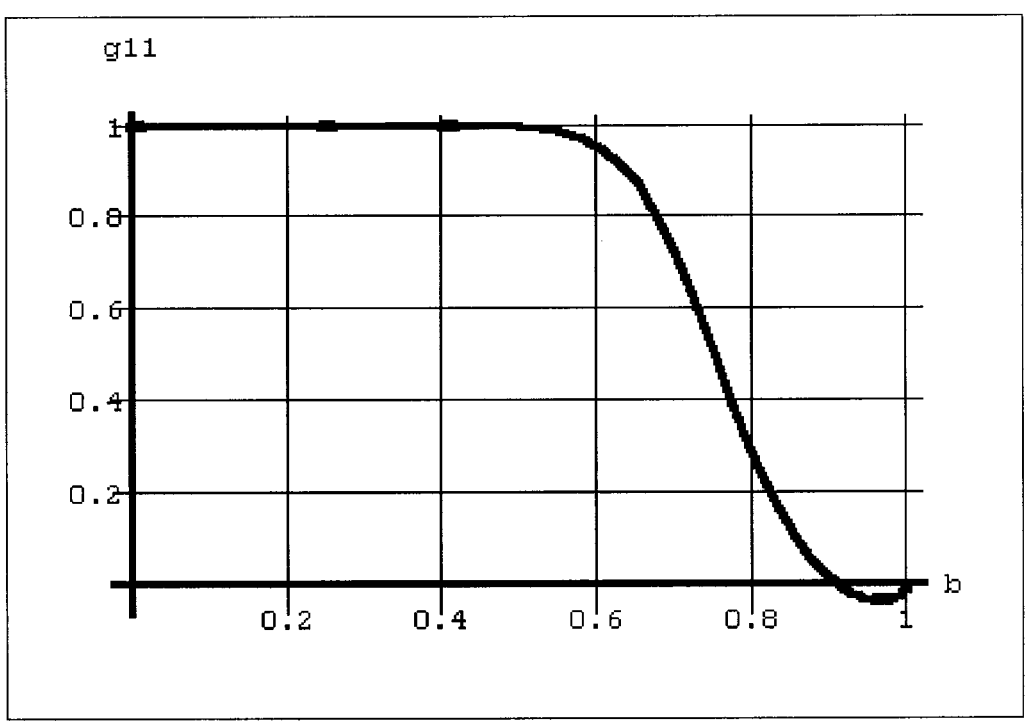

Figura 6.1. Coeficiente de la matriz **G** del Ejemplo 6.1

La matriz *global* del sistema **G** = **B A** se denotará como

$$\mathbf{G} = \begin{bmatrix} g_{11} & g_{12} \\ -g_{12} & g_{11} \end{bmatrix}, \qquad \sqrt{g_{11}^2 + g_{12}^2} = 1$$

La Figura 6.1 muestra la evolución del coeficiente g_{11} de la matriz **G** en función del parámetro b de la matriz de mezcla. La Figura confirma que la calidad de la Separación es buena excepto cuando a y b son parecidos.

En todo caso, las soluciones del sistema (6.8) son útiles aún cuando no sean muy precisas. Según (6.24), el producto escalar entre el vector $\mathbf{v}_i$ y la observación $\mathbf{x}(t)$ vale

$$\mathbf{v}_i^{\mathrm{T}}\, \mathbf{x}(t) \propto [\, a_{i1}/\varepsilon_{i1}, \ldots, a_{iN}/\varepsilon_{iN} \,]^{\mathrm{T}}\, \mathbf{s}(t) = \sum_{j=1}^{N} \frac{a_{ij}\, s_j(t)}{\varepsilon_{ij}} \tag{6.25}$$

donde $\propto$ denota "proporcionalidad". Si, como antes, resultara que $|\, \varepsilon_{i1} \,| \approx |\, \varepsilon_{i2} \,| \ll |\varepsilon_{ij}\,|$ para $j \neq 1, 2$, entonces $\mathbf{v}_i^{\mathrm{T}}\, \mathbf{x}(t)$ sería una mezcla casi en exclusiva de la *primera* y *segunda* fuentes. Es decir, al proyectar $\mathbf{x}(t)$ sobre $\mathbf{v}_i$ *eliminaríamos la contribución de varias de las señales fuente.*

Este razonamiento sugiere un *procedimiento iterativo*: definamos un nuevo conjunto de observaciones $\mathbf{x'}(t) = [\, x'_1(t), \ldots, x'_N(t) \,]^{\mathrm{T}}$, siendo $x'_i(t) = \mathbf{v}_i^{\mathrm{T}}\, \mathbf{x}(t)$ para $i = 1, \ldots, N$. Cabe esperar que $\mathbf{x'}(t)$ sea más parecido al vector de fuentes que $\mathbf{x}(t)$. Si ahora utilizamos $\mathbf{x'}(t)$ para estimar las matrices $\mathbf{P}_i$ y repetimos el mismo procedimiento una y otra vez, es razonable confiar en que separaremos las fuentes.

6.6 El Algoritmo SEVILLA

Vamos a proponer un algoritmo de bloques que es la implementación obvia de las ideas desarrolladas hasta ahora. Después, analizaremos sus prestaciones y expondremos las líneas generales que hacen eficiente su implementación [Martín99*b*].

SEVILLA

(**SE**PARACIÓN POR LA **VÍ**A DE UNA FORMU**LA** **LINEAL**)

Paso 1. Sea $\mathbf{x}(t)$ el vector de observaciones, ya *decorreladas*.

Paso 2. Desde $i = 1$ hasta $i = N$

Estimar la matriz $\mathbf{P}_i$

Resolver $\mathbf{P}_i \mathbf{v}_i = \mathbf{e}_i$

Tomar $\mathbf{b}_i = \mathbf{v}_i / \| \mathbf{v}_i \|_2$, siendo $\mathbf{b}_i^T$ la i-ésima fila $\mathbf{B}$.

Fin *desde i*

Paso 3. Convertir las filas de $\mathbf{B}$ en un conjunto ortonormal.

Paso 4. Sea $\mathbf{x}(t) = \mathbf{B}\,\mathbf{x}(t)$ el *nuevo* vector de observaciones.

Paso 5. Si $\mathbf{B}$ está "suficientemente" próxima a una matriz de permutación, tomar $\mathbf{y}(t) = \mathbf{x}(t)$ y terminar. En otro caso, volver al **Paso 2.**

Debemos hacer algunos comentarios:

Para estimar $\mathbf{P}_i$ en el Paso 2, tenemos que elegir primero un valor de α_i apropiado. Una elección adecuada se basa en el criterio de estabilidad que presentaremos en la próxima Sección. Según se vio en la Sección 6.2, es suficiente resolver las ecuaciones $i = 1, \ldots, N - 1$. De tal manera, donde dice 'Desde $i = 1$ hasta $i = N$' podría decir 'Desde $i = 1$ hasta $i = N - 1$'.

Con el Paso 3 evitamos que varias filas de **B** converjan a la misma columna de la matriz de mezcla, lo que podría ocurrir si la estimación de α_i es defectuosa. También asegura que **B** es una matriz ortogonal, como se requiere. Por otra parte, la estimación de los estadísticos de $\mathbf{x}(t)$ en una iteración puede hacerse a partir de los estadísticos y del valor de **B** calculados en la iteración previa, con el consiguiente ahorro computacional (de hecho, el Paso 4 se suprime en una implementación eficiente). En este caso, se puede calcular con facilidad, aunque omitiremos los detalles, que el algoritmo efectúa del orden de $2N^3T$ operaciones de punto flotante, donde T es el número de observaciones disponibles [Martín99*b*]. Para hacer este cálculo se ha supuesto, de acuerdo con las simulaciones, que el algoritmo itera aproximadamente N veces antes de separar las fuentes. Es importante destacar que la mayor parte del esfuerzo computacional se emplea en la estimación de los estadísticos de las señales y no en la resolución de las ecuaciones.

6.6.1 Análisis de Convergencia: ley de adaptación

Reproduzcamos la ecuación (6.24), por conveniencia:

$$\mathbf{v}_i = \mathbf{A}\,\mathbf{E}^{-1}\,\mathbf{a}_i / k$$

donde $\mathbf{E}^{-1}$ es una matriz *diagonal* y k una irrelevante constante escalar. Por definición, $\mathbf{b}_i^T$, la i-ésima fila de **B**, se toma

$$\mathbf{b}_i = + \mathbf{v}_i / \| \mathbf{v}_i \|_2 = sign(k)\, \mathbf{A}\,\mathbf{E}^{-1}\,\mathbf{a}_i / \| \mathbf{E}^{-1}\,\mathbf{a}_i \|_2 \qquad (6.26)$$

donde $\| \mathbf{E}^{-1}\,\mathbf{a}_i \|_2 = \| \mathbf{A}\,\mathbf{E}^{-1}\,\mathbf{a}_i \|_2$ debido a que **A** es una matriz ortogonal. La proyección del vector de observaciones $\mathbf{x}(t)$ sobre $\mathbf{b}_i$ da

$$x'_i(t) = \mathbf{b}_i^T\,\mathbf{x}(t) = \{\, \mathbf{b}_i^T\,\mathbf{A}\,\}\,\mathbf{s}(t) =$$

$$= sign(\, k \,)\, (\, \mathbf{E}^{-1}\, \mathbf{a}_i \,/\, \| \mathbf{E}^{-1}\, \mathbf{a}_i \|_2 \,)^{\mathrm{T}}\, \mathbf{s}(t) \tag{6.27}$$

que nos sirve para definir $\mathbf{x}'(t) = [x'_1(t), \ldots, x'_N(t)]^{\mathrm{T}}$ que es, de hecho, el nuevo vector de observaciones tal y como se calcula en el Paso 4 del algoritmo. En lo referente a la matriz de mezcla, sea $\mathbf{a}_i^n$ la i-ésima fila de la matriz de mezcla en la n-ésima iteración del algoritmo. Nótese que (6.27), en realidad, establece una ley de recursión:

$$\mathbf{a}_i^{n+1} = signo(\, k \,)\, \mathbf{E}^{-1}\, \mathbf{a}_i^n \,/\, \| \mathbf{E}^{-1}\, \mathbf{a}_i^n \|_2 \tag{6.28}$$

donde tanto $signo(\, k \,)$ como $\mathbf{E}$ son funciones de $\mathbf{a}_i^n$.

Ejemplo 6.2. En el caso de tener dos fuentes, tenemos para $i = 1$:

$$a_{11}^{n+1} = \frac{1}{D}\, \frac{a_{11}^n}{1 + \kappa_{s1}(a_{11}^n)^2 - \alpha_1^n}$$

y

$$a_{12}^{n+1} = \frac{1}{D}\, \frac{a_{12}^n}{1 + \kappa_{s2}(a_{12}^n)^2 - \alpha_1^n}$$

donde

$$D = \frac{1}{signo(k)} \sqrt{\left(\frac{a_{11}^n}{1 + \kappa_{s1}(a_{11}^n)^2 - \alpha_1^n}\right)^2 + \left(\frac{a_{12}^n}{1 + \kappa_{s2}(a_{12}^n)^2 - \alpha_1^n}\right)^2}$$

6.6.2 Convergencia y estabilidad

A la vista de (6.28), queda claro que los *vectores canónicos* son puntos de equilibrio de la iteración. Es más, si las filas de $\mathbf{A}$ llegaran a ser vectores canónicos distintos, las fuentes quedarían naturalmente separadas. En la presente Sección

estableceremos las condiciones que garantizan que el algoritmo converja hacia estos puntos de equilibrio.

Supongamos que existe un índice M (de '*M*ayor') tal que

$$(a_{iM}^n)^2\kappa_{s_M} > (a_{ij}^n)^2\kappa_{s_j} \tag{6.29}$$

para todo $M \neq j$, donde $\kappa_{s_M} > 0$ (es decir, al menos una fuente es super-gaussiana).

Al dividir a_{iM}^n entre a_{ij}^n se obtiene

$$\frac{\left|a_{iM}^{n+1}\right|}{\left|a_{ij}^{n+1}\right|} = \left|L_{Mj}^n\right| \frac{\left|a_{iM}^n\right|}{\left|a_{ij}^n\right|} \tag{6.30}$$

donde, según (6.28)

$$L_{Mj}^n = \frac{\alpha_i^n - 1 - (a_{ij}^n)^2\kappa_{s_j}}{\alpha_i^n - 1 - (a_{iM}^n)^2\kappa_{s_M}} \tag{6.31}$$

Elijamos un número χ tal que

$$\chi > (a_{iM}^n)^2\kappa_{s_M} \tag{6.32}$$

y tomemos $\alpha_i^n = 1 + \chi$. Esta elección garantiza que $L_{Mj}^n > 1$ para todo $j \neq M$ ya que $0 < \chi - (a_{iM}^n)^2\kappa_{s_M} < \chi - (a_{ij}^n)^2\kappa_{s_j}$. En consecuencia, si $j \neq M$,

$$\frac{\left|a_{iM}^{n+1}\right|}{\left|a_{ij}^{n+1}\right|} > \frac{\left|a_{iM}^n\right|}{\left|a_{ij}^n\right|} \tag{6.33}$$

Por otra parte, el que tanto $\| \mathbf{a}_i^{n+1} \|_2$ como $\| \mathbf{a}_i^n \|_2$ valgan uno implica que

$$(a_M^{n+1})^2\left(1 + \sum_{j\neq M}\left(\frac{a_{ij}^{n+1}}{a_{iM}^{n+1}}\right)^2\right) = (a_{iM}^n)^2\left(1 + \sum_{j\neq M}\left(\frac{a_{ij}^n}{a_{iM}^n}\right)^2\right) \tag{6.34}$$

Por lo tanto, se deduce que $\left|a_{iM}^{n+1}\right| > |a_{iM}^{n}|$ y $(a_{iM}^{n+1})^2\kappa_{s_M} > (a_{ij}^{n+1})^2\kappa_{s_j}$, por lo que la misma discusión puede hacerse ahora en la iteración n+1. En conclusión,

$$\left|a_{iM}^{n+1}\right| > |a_{iM}^{n}| \text{ para todo } n \tag{6.35}$$

Ya que la norma de $\mathbf{a}_i$ es siempre igual a uno, esta desigualdad implica que $|a_{iM}|$ crece mientras que los otros coeficientes tienden a cero. Es decir, $\mathbf{a}_i$ *tiende hacia un vector canónico. Por lo tanto, se alcanza la separación.*

Por el contrario, supongamos que todas las fuentes son sub-gaussianas (es decir, $\kappa_{s_j} < 0$ para todo j). Entonces, debe existir un índice m (de '*m*enor') tal que $(a_{im}^{n})^2\kappa_{s_m} < (a_{ij}^{n})^2\kappa_{s_j}$ para todo $j \neq m$. Razonando como antes, la elección $\alpha_i^n = 1 + \chi$ con

$$\chi < (a_{im}^{n})^2\kappa_{s_m} \tag{6.36}$$

garantiza igualmente la convergencia de $\mathbf{a}_i$ hacia un vector canónico.

En conclusión, hemos probado que la convergencia del algoritmo a una solución que separe las fuentes es *global*, siempre que se respeten las condiciones (6.32) ó (6.36). Por otra parte, estas condiciones son *suficientes*; pero no se ha probado que, además, sean necesarias.

Ejemplo 6.3. Supongamos una mezcla de fuentes de distribución uniforme, de curtosis $\kappa = -1'2$. La aplicación estricta del criterio (6.36) se asegura tomando

$$\chi < \kappa \leq (a_{im}^{n})^2\kappa_{s_m}, \text{ es decir, } \chi < -1'2$$

lo que implica $\alpha_i^n = 1 + \chi < -0'2$ para todo n. Esto garantiza la convergencia global del algoritmo y la Separación de las Fuentes. No obstante, numerosas simulaciones corroboran que la elección $\alpha_i^n = 0$ también consigue la Separación en el caso de tener dos o más fuentes, a pesar de que $\alpha_i^n = 0$ no verifica siempre (6.36). Como se ha dicho, que $\alpha_i^n < -0'2$ es suficiente pero no necesario y una cierta relajación de la desigualdad no parece afectar al

comportamiento del algoritmo. En cambio, se comprueba que $\alpha_i^n > 0$ hace que el algoritmo diverja.

Ejemplo 6.4. El estudio clásico de estabilidad se basa en la linealización del algoritmo alrededor de alguno de sus puntos de equilibrio. Por supuesto, este análisis sólo es local, es decir no aporta información sobre la convergencia del algoritmo cuando las condiciones iniciales distan de los puntos de equilibrio.

Así, suponiendo que $\mathbf{a}_i^n$ está en las proximidades del equilibrio $\mathbf{e}_i$, la regla de adaptación (6.28) se puede aproximar por la serie de Taylor

$$\mathbf{a}_i^{n+1} = \mathbf{e}_i + \mathbf{J}\,(\,\mathbf{a}_i^n - \mathbf{e}_i\,) \Rightarrow (\,\mathbf{a}_i^{n+1} - \mathbf{e}_i\,) = \mathbf{J}\,(\,\mathbf{a}_i^n - \mathbf{e}_i\,)$$

donde $\mathbf{J}$ es la matriz Jacobiana de la iteración, es decir,

$$[\,\mathbf{J}\,]_{kl} = \frac{\partial\, a_{ik}^{n+1}}{\partial\, a_{il}^n}$$

Resulta que

$$(\,\mathbf{a}_i^{n+1} - \mathbf{e}_i\,) = \mathbf{J}^n\,(\,\mathbf{a}_i^1 - \mathbf{e}_i\,)$$

La norma de $\mathbf{J}^n$ tenderá a cero y, por lo tanto, se alcanzará el equilibrio, cuando todos los autovalores de $\mathbf{J}$ sean, en módulo, menores que la unidad. Es posible demostrar que ello equivale a

$$|\alpha_i - 1| > |\alpha_i - 1 - \kappa_{s_i}|,$$

lo que, según (6.31), hace que L_{ij}^n sea mayor que uno. Esto mismo garantiza la *estabilidad global* según nuestro análisis.

En conclusión, el algoritmo es muy robusto. Converge aunque la estimación de α_i sea grosera: basta que la magnitud de α_i sea suficientemente grande para que se satisfagan las condiciones (6.32) ó (6.36), según corresponda. Ahora bien, para escoger α_i debemos tener en cuenta que

- El signo de las curtosis de las fuentes ha ser conocido o estimado, a fin de que se pueda escoger entre la condición (6.32) y la (6.36). Resulta interesante que conocido el signo de las curtosis también se pueda garantizar la estabilidad local del algoritmo EASI (ver el Ejemplo 4.3 en la pág. 108). Ante este problema, la solución habitual que se encuentra en la literatura es *el método de prueba y error* hasta conseguir la convergencia del algoritmo [Girolami97].
- Si la magnitud de α_i es excesivamente grande, la velocidad de convergencia del algoritmo será pequeña, ya que $L_{ij}^n \approx 1$ como muestra su definición (6.31).

Afortunadamente, el Teorema 6.2 es una herramienta poderosa: sea λ_M el mayor valor propio de la matriz $\mathrm{E}[\mathbf{x}\,\mathbf{x}^{\mathrm{T}}\,x_i^2]$. Entonces, este Teorema garantiza que

$$0 \le \lambda_M - (1 + a_{iM}^2 \kappa_{s_M}) \le 2 \Rightarrow \begin{cases} a_{iM}^2 \kappa_{s_M} \le \lambda_M - 1 \\ \lambda_M - 3 \le a_{iM}^2 \kappa_{s_M} \end{cases} \tag{6.37}$$

Entonces, proponemos un procedimiento heurístico pero muy simple y efectivo: sea

$$\chi = \mu\,(\lambda_M - 3) \tag{6.38}$$

y tomemos $\alpha_i^n = 1 + \chi$. De (6.37) se deduce que $\lambda_M - 3 \le (a_{iM}^2)\kappa_{s_M}$. Por ello,

- Si $\lambda_M - 3$es positivo, entonces κ_{s_M}es positivo y se aplica la *condición de estabilidad* (6.32). Debe existir un número $\mu_0 \geq 1$ para el que χ satisface (6.32) siempre que μ sea mayor o igual que μ_0.
- Si todas las fuentes son sub-gaussianas (es decir, todas las curtosis son negativas), entonces $\lambda_M - 3$ es negativo y debe existir un número $\mu'_0 \geq 1$ para el que χ satisface (6.36) cuando μ es mayor o igual que μ'_0.
- Si $\lambda_M - 3$ es negativo pero κ_{s_M}es positivo (lo que se detecta porque, según (6.37), $a_{iM}^2 \kappa_{s_M} \leq \lambda_M - 1$ es positivo, entonces existe un número $\mu''_0 \leq -1$ para el que χ satisface (6.32) cuando μ es menor o igual que μ''_0.

Por otra parte, no tiene sentido determinar con exactitud el autovalor λ_M ya que no conocemos μ_0, μ'_0 o μ''_0. Afortunadamente, el Teorema de Gersghorin [Noble89], que enunciamos a continuación, permite obtener una aproximación suficiente a este autovalor.

Teorema 6.2 (*Teorema de los Círculos de Gerschgorin*) Todo autovalor λ de una matriz $\mathbf{M}$ de dimensiones $N \times N$ satisface cuando menos una de las siguientes desigualdades

$$| \lambda - m_{ii} | \leq r_i, \text{ siendo } r_i = \sum_{j=1, j \neq i}^{N} | m_{ij} |$$

donde m_{ij} es la componente (i, j) de $\mathbf{M}$.

En la práctica, $E[\, \mathbf{x}\, \mathbf{x}^T x_i^2 \,]$ es una matriz dominada por su diagonal. Por lo tanto, el Teorema de Gersghorin nos permite aproximar los autovalores de esta matriz por los elementos que están en su diagonal. En particular, λ_M puede ser sustituido en buena aproximación por el mayor de los elementos diagonales de la

matriz E[$\mathbf{x}\,\mathbf{x}^T x_i^2$], que suele ser [$\mathbf{P}_i$]$_{ii}$ = E[x_i^4]. En conclusión, de acuerdo con (6.38) proponemos:

$$\chi = \mu\, \kappa_{xi} \tag{6.39}$$

siendo κ_{xi} = E[x_i^4] − 3 es decir, la curtosis de la observación *i*-ésima y μ se escoge de acuerdo con la discusión hecha en la página anterior. En la práctica, la elección $|\mu| \in [1, 1'5]$ ofrece un buen resultado, como mostraremos a través de simulaciones.

Ejemplo 6.5. Consideremos una mezcla de tres fuentes super-gaussianas que son generadas elevando al cubo una variable aleatoria normal. La matriz de mezcla (no ortogonal, de forma que, en primer lugar hay que decorrelar las observaciones) se escoge de forma aleatoria como

$$\mathbf{M} = \begin{bmatrix} 1.61 & -1.76 & 0.09 \\ 1.05 & 1.68 & 0.65 \\ 0.43 & -0.42 & -0.67 \end{bmatrix}$$

Los estadísticos de las observaciones se computan a partir de 1000 muestras. De acuerdo con (6.39), se toma $\alpha_i^n = 1 + \kappa_{xi}$. Después de *sólo una iteración*, el algoritmo SEVILLA consigue que la matriz *global* de separación **G** (matriz de separación por matriz de mezcla) sea igual que

$$\mathbf{G} = \begin{bmatrix} -0.99 & 0.11 & 0.016 \\ 0.10 & 0.99 & -0.29 \\ 0.03 & -0.04 & -1.00 \end{bmatrix}$$

Como se aprecia, el resultado es bastante bueno y las fuentes están prácticamente separadas. Éste es sólo un ejemplo favorable; pero ilustra la potencia del método.

6.7 Conclusiones

Presentamos un algoritmo de bloques para la Separación de Fuentes. El algoritmo converge y separa las fuentes *siempre*, con independencia de su distribución estadística. Además, como en todos los algoritmos de bloque, el ajuste de sus parámetros es muy simple.

7. Experimentos

7.1 Introducción

Vamos a dedicar este Capítulo a caracterizar experimentalmente el comportamiento del algoritmo SEVILLA. Básicamente, pretendemos estudiar aspectos del algoritmo que serían muy difíciles de tratar analíticamente.

En la Sección 7.2 presentamos los índices con los que vamos a medir la calidad de la Separación de Fuentes. Después, los experimentos se plantean en el orden que sigue:

- La Sección 7.3 presenta una simulación simple pero muy ilustrativa de la capacidad del algoritmo SEVILLA.
- En la Sección 7.4 exploramos las prestaciones del algoritmo en función del número de muestras que se utilicen para estimar los estadísticos.
- En la Sección 7.5 se plantea un experimento en el que se trabaja con señales *no estacionarias.*
- En la Sección 7.6 se compara SEVILLA con otros algoritmos.
- En la Sección 7.7 se muestra la capacidad de manejar con SEVILLA un gran número de fuentes.

Finalmente, la Sección 7.8 se dedica a las Conclusiones.

7.2 Medidas de Trabajo

Sea **G** = **B A** la matriz *global* del sistema, esto es, que relaciona las fuentes con las salidas. Para medir la calidad de la Separación de las Fuentes, Amari y sus colaboradores han propuesto el siguiente índice

$$I_{\mathrm{AM}}(\mathbf{G}) = \sum_{i=1}^{\mathrm{N}} \left(\sum_{j=1}^{\mathrm{N}} \frac{|g_{ij}|}{\max_k |g_{ik}|} - 1 \right) + \sum_{j=1}^{\mathrm{N}} \left(\sum_{i=1}^{\mathrm{N}} \frac{|g_{ij}|}{\max_k |g_{kj}|} - 1 \right) \qquad (7.1)$$

Es sencillo verificar que $I_{\mathrm{AM}} \geq 0$, dándose la igualdad si y sólo si **G** = **P D**, donde **P** es una matriz de permutación y **D** es una matriz diagonal, o sea, $I_{\mathrm{AM}} = 0$ cuando hemos separado las fuentes. Es muy sencillo utilizar este índice, por cuanto no requiere que las filas / columnas de la matriz **G** sean reordenadas y / o escaladas.

Un índice muy parecido ha sido propuesto por E. Moreau y O. Macchi:

$$I_{\mathrm{MM}}(\mathbf{G}) = \frac{1}{2}\left[\sum_{i=1}^{\mathrm{N}} \left(\sum_{j=1}^{\mathrm{N}} \frac{|g_{ij}|^2}{\max_k |g_{ik}|^2} - 1 \right) + \sum_{j=1}^{\mathrm{N}} \left(\sum_{i=1}^{\mathrm{N}} \frac{|g_{ij}|^2}{\max_k |g_{kj}|^2} - 1 \right) \right] \qquad (7.2)$$

En la literatura reciente es muy normal el uso de uno u otro índice para estudiar el comportamiento de los algoritmos, por lo que nosotros usaremos ambos para que sea más sencilla la comparación.

El índice de Amari es menos "vistoso" aunque, a nuestro parecer, más fiable. Veámoslo con un ejemplo: sea **G** la matriz 10×10 tal que los elementos de su diagonal valen $[\,\mathbf{G}\,]_{ii} = 1$ para todo i, mientras que $[\,\mathbf{G}\,]_{ij} = 0'1$ para i distinto de j. Para esta matriz

$$I_{\mathrm{AM}}(\mathbf{G}) = 18$$

mientras que

$$I_{MM}(\mathbf{G}) = 0'9$$

Con esta matriz **G** cabe esperar que las fuentes estén bien separadas "casi siempre" y así lo indica el índice de Moreau y Macchi. Sin embargo, dado que

$$\mathbf{y}(t) = \mathbf{G}\,\mathbf{s}(t)$$

si en un determinado instante las diez fuentes tienen magnitudes parecidas pero el signo de $s_1(t)$ es el opuesto al de $s_2(t)$, $s_3(t)$, ..., $s_{10}(t)$, la primera salida $y_1(t)$ se cancelará y la separación es "mala". De esto nos previene el índice de Amari.

Así pues, utilizaremos el índice de Moreau y Macchi como una medida de la calidad que, en condiciones normales, tiene la Separación. Complementariamente, el índice de Amari evaluará el comportamiento que tendría el resultado en una situación desfavorable.

7.3 Separación de cuatro fuentes con distinta distribución

En primer lugar, preparamos un experimento "vistoso" que permita visualizar los resultados y comprobar *in situ* la calidad de la separación. Generamos cuatro fuentes con distribuciones estadísticas distintas: uniforme, binaria (ambas sub-gaussianas), exponencial y el cubo de una variable normal (ambas super-gaussianas). Las muestras de cada fuente son independientes entre sí. La matriz de mezcla se escoge como

$$\mathbf{M} = \begin{bmatrix} -1.475 & 1.432 & -1.050 & -0.246 \\ 0.602 & -0.190 & 0.162 & -0.290 \\ 0.418 & 1.111 & -0.170 & -2.100 \\ 1.315 & 0.190 & -1.990 & 1.349 \end{bmatrix}$$

Esta matriz tiene un número de condición muy alto $c(\mathbf{M}) = 623'8$ por lo que se puede considerar casi singular. Las cuatro fuentes se muestran en la Figura 7.1 y las correspondientes observaciones en 7.2 (sólo se representan las primeras cien muestras). Obsérvese que las señales no tienen media cero ni varianza unidad y las observaciones no están *incorreladas* entre sí. Además, las mezclas están fuertemente dominadas por el cubo de la variable gaussiana.

Lógicamente, el primer paso consiste en eliminar el nivel de continua de las observaciones, normalizar su potencia a la unidad y decorrelarlas. Después se aplica el algoritmo SEVILLA.

Se utilizan 1000 muestras de las observaciones para estimar todos los estadísticos. Después de la Separación, la *matriz global* **G** toma el valor

$$\mathbf{G} = \begin{bmatrix} -0.0706 & 0.0079 & -0.2565 & 0.1006 \\ -0.0046 & 0.0008 & -0.0196 & -2.0013 \\ -0.1176 & 0.972 & 0.0064 & 0.0277 \\ 3.4534 & 0.0498 & -0.0007 & -0.1629 \end{bmatrix}$$

La estimación de las fuentes se muestra en la Figura 7.3. Nótese que el algoritmo no puede distinguir entre $s_i(t)$ y $-s_i(t)$. Además, el orden de presentación de las fuentes está cambiado.

Como variante, sustituimos la cuarta fuente (binaria) por otra de distribución gaussiana (curtosis nula) y repetimos los cálculos. La *matriz global* **G** es ahora

$$\mathbf{G} = \begin{bmatrix} 0.0611 & -0.0073 & 0.2572 & 0.0095 \\ 0.1125 & -0.0875 & -0.0019 & -1.0549 \\ -0.1210 & 0.9720 & 0.0059 & -0.0985 \\ 3.4407 & 0.0544 & 0.0007 & 0.0491 \end{bmatrix}$$

Como vemos, las fuentes se separan sin ningún problema. Por último, se sustituyen la cuarta (binaria) y tercera (cubo de gaussiana) fuentes por otras dos de distribución gaussiana. El resultado es

$$\mathbf{G} = \begin{bmatrix} -0.0996 & 0.9566 & -0.1494 & 0.0111 \\ -0.3985 & 0.1034 & 0.4364 & -0.9134 \\ -3.4189 & -0.0495 & 0.0090 & 0.1085 \\ 0.1224 & 0.1714 & 0.8569 & 0.5309 \end{bmatrix}$$

Hemos extraído las fuentes que *no* tienen una distribución gaussiana, que se corresponden con la primera y tercera señales de salida. En cambio, la segunda y cuarta salidas son una combinación de las fuentes gaussianas. Sin embargo, estas dos señales están incorreladas lo que, para una distribución gaussiana, quiere decir que son *independientes*. En conclusión, hemos extraído cuatro señales estadísticamente independientes aunque sólo dos de ellas se corresponden con las fuentes originales.

7.4 Experimentos en los que se varía el número de muestras

Las prestaciones del algoritmo SEVILLA dependen por completo de la precisión con la que se estimen los estadísticos de las observaciones. Ahora preparamos una serie de experimentos en los que se evalúa la calidad de la Separación en función del número de muestras de que se disponga.

En 7.4.1 se considera que las muestras de cada fuente son independientes entre sí, esto es, $s_i(t)$ es independiente de $s_i(t')$ si t es distinto de t'. Se experimenta con mezclas de dos, cinco y diez fuentes y SEVILLA itera, respectivamente, cuatro, diez y veinte veces. Las Tablas que se presentan son el resultado de promediar veinte experimentos independientes, en los que las matrices de mezcla fueron generadas de forma aleatoria.

En 7.4.2 se introduce correlación entre las muestras de cada fuente. En este caso, la estimación de los estadísticos tiene mayor varianza [Papoulis91, pág. 427 y *ss.*] y así se refleja después en los resultados.

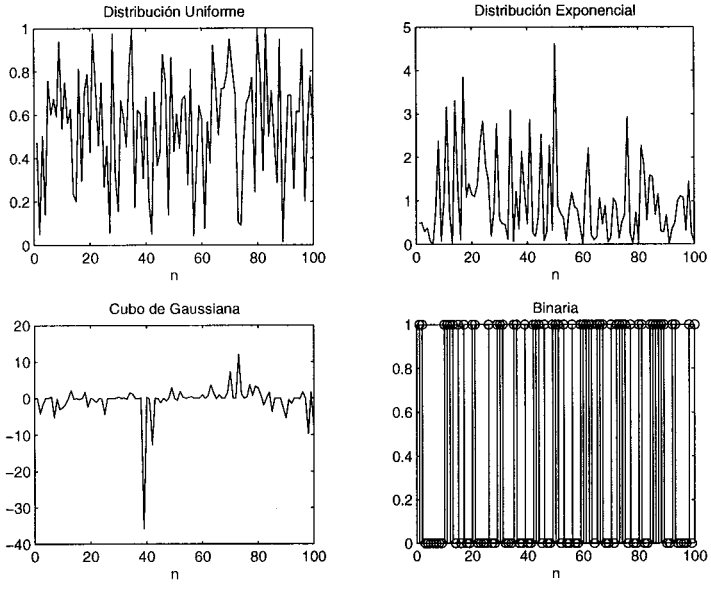

Figura 7.1. Señales Fuente del Experimento de la Sección 4.3

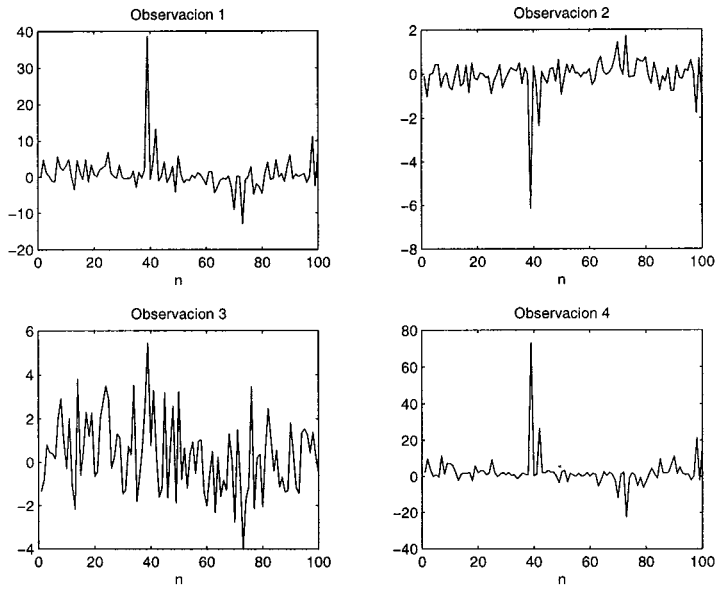

Figura 7.2. Observaciones del Experimento de la Sección 4.3

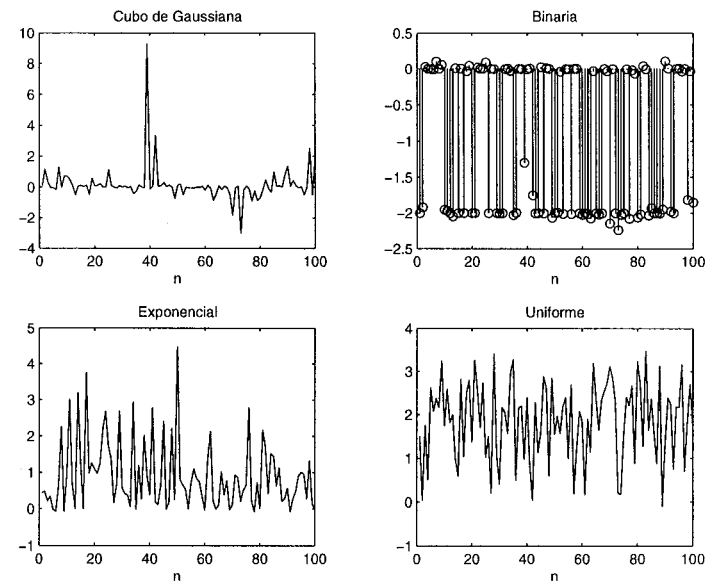

Figura 7.3. Señales de Salida del Experimento de la Sección 4.3

7.4.1 Las muestras de las Fuentes son independientes

La varianza en la estimación de los estadísticos de las observaciones depende de la propia distribución de las fuentes. Por esta razón, vamos a experimentar con fuentes:

-De distribución uniforme

-Obtenidas elevando al cubo una variable gaussiana.

Las razones que nos llevan a elegir estas distribuciones se aclaran mediante el siguiente ejemplo: por cinco veces, generamos *cien* muestras independientes de una variable uniforme, de media cero y varianza uno, con el comando **`rand`** de Matlab. La curtosis de la variable, estimada a partir de las cien muestras y para cada uno de los cinco experimentos, es

−1'1025	−1'3559	−1'2882	−1'0702	−1'2649

La media es −1'2163 y la desviación típica 0'1238. La *auténtica* curtosis de la variable vale −1'2.

Después, repetimos el mismo procedimiento pero con la variable que es el cubo de una gaussiana (obtenida, a su vez, con el comando **`randn`** de Matlab). Los resultados son ahora

32'1179	11'5278	41'8815	5'5311	22'2470

cuya media es 22'6611 y su desviación típica 14'8022. Así pues, resulta que los estadísticos de la primera variable se pueden estimar con mucha mayor precisión que los de la segunda. Esto justifica la elección de ambas distribuciones: se consideran representativas de los casos más y menos favorable, respectivamente.

Las Tablas recogen los valores medios de los índices de Amari (I_{AM}) y de Moreau-Macchi (I_{MM}) así como su desviación típica (entre paréntesis) en función del número de muestras que se han utilizado para estimar los estadísticos de las observaciones.

MEZCLA DE *DOS* VARIABLES UNIFORMES

Número de muestras	50	100	250	500	1000	2000
I_{AM}	0'38 (0'25)	0'38 (0'34)	0'24 (0'19)	0'15 (0'08)	0'093 (0'051)	0'067 (0'041)
I_{MM}	0'03 (0'03)	0'03 (0'06)	0'01 (0'02)	0'0045 (0'0047)	0'0018 (0'0016)	0'00091 (0'0013)

MEZCLA DE *CINCO* VARIABLES UNIFORMES

Número de muestras	50	100	250	500	1000	2000
I_{AM}	1'68 (3'69)	1'07 (2'39)	0'62 (1'70)	0'32 (0'52)	0'42 (1'84)	0'19 (0'25)
I_{MM}	0'33 (0'94)	0'15 (0'51)	0'062 (0'43)	0'0089 (0'016)	0'062 (0'58)	0'0030 (0'0049)

MEZCLA DE *DIEZ* VARIABLES UNIFORMES

Número de muestras	50	100	250	500	1000	2000	5000	10000
I_{AM}	23'43 (8'20)	15'34 (4'80)	10'10 (2'35)	7'95 (2'16)	5'09 (1'03)	4'06 (0'83)	2'91 (0'65)	2'25 (0'49)
I_{MM}	3'87 (2'36)	2'00 (1'27)	0'94 (0'60)	0'59 (0'42)	0'20 (0'16)	0'11 (0'077)	0'075 (0'064)	0'036 (0'022)

MEZCLA DEL CUBO DE *DOS* VARIABLES GAUSSIANAS

Número de muestras	50	100	250	500	1000	2000
I_{AM}	0'57 (0'94)	0'55 (0'94)	0'16 (0'097)	0'11 (0'096)	0'11 (0'15)	0'060 (0'050)
I_{MM}	0'15 (0'43)	0'15 (0'39)	0'0061 (0'0064)	0'0035 (0'0057)	0'0044 (0'015)	0'0010 (0'0016)

MEZCLA DEL CUBO DE *CINCO* VARIABLES GAUSSIANAS

Número de muestras	50	100	250	500	1000	2000
I_{AM}	4'18 (2'18)	3'34 (2'28)	2'14 (1'21)	1'71 (0'95)	1'19 (0'50)	0'85 (0'29)
I_{MM}	0'60 (0'62)	0'49 (0'79)	0'22 (0'34)	0'17 (0'26)	0'055 (0'072)	0'0020 (0'0174)

MEZCLA DEL CUBO DE *DIEZ* VARIABLES GAUSSIANAS

Número de muestras	50	100	250	500	1000	2000	5000	10000
I_{AM}	23'16 (6'02)	15'28 (4'00)	10'05 (2'69)	7'45 (2'59)	5'26 (1'66)	4'06 (1'23)	2'84 (0'41)	2'24 (0'34)
I_{MM}	3'76 (1'63)	1'93 (1'18)	0'87 (0'74)	0'48 (0'42)	0'24 (0'29)	0'18 (0'34)	0'050 (0'020)	0'038 (0'042)

Como se podía esperar, los resultados mejoran a medida que crece el número de muestras a partir de las que se estiman los estadísticos. Las desviaciones típicas son altas, lo que se debe a que algunos experimentos *dan resultados bastante peores que la media*. Curiosamente, no se puede afirmar que la distribución de las fuentes sea claramente determinante en los resultados.

7.4.2 Las muestras de las Fuentes están correladas

Como en el caso anterior, para obtener cada fuente se generan muestras independientes distribuidas uniformemente, con media cero y varianza unidad. Como paso previo a la mezcla, cada fuente se filtra con el sistema que tiene la respuesta

$$H(z) = \frac{0.5}{1 - 0.5\,z^{-1}}$$

cuyo único objeto es el de introducir correlación entre las muestras. Después, se repiten los experimentos del apartado anterior, aunque sólo con la variable uniforme. A continuación se recogen los resultados.

MEZCLA DE *DOS* VARIABLES UNIFORMES

Número de muestras	50	100	250	500	1000	2000
I_{AM}	1'31 (0'85)	1'00 (0'85)	0'60 (0'46)	0'53 (0'64)	0'50 (0'92)	0'20 (0'12)
I_{MM}	0'32 (0'40)	0'21 (0'34)	0'07 (0'10)	0'0085 (0'19)	0'13 (0'43)	0'0072 (0'0075)

El algoritmo ha iterado cuatro veces. Comparados con los del experimento anterior, los resultados evidentemente empeoran. En realidad, las propiedades de SEVILLA no dependen de que las fuentes estén correladas o dejen de estarlo; pero la estimación que se hace de los estadísticos de las observaciones sí [Papoulis91, pág. 427 y ss.]. Cualquier otro algoritmo de bloque va a tener el mismo problema. De igual manera, los algoritmos adaptativos del tipo de INFOMAX no van a ser *estimadores de máxima verosimilitud* de la matriz de mezcla en estas condiciones. En cambio, los algoritmos que sólo emplean estadísticos de 2° orden pueden operar en esta situación [Belouchrani97, Tong91].

7.5 Señales no estacionarias

Una de las hipótesis que se ha mantenido a lo largo de toda la Tesis es que las fuentes son realizaciones de procesos estacionarios. Ahora vamos a mezclar cuatro señales de audio que no verifican esta suposición. Para ello, grabamos la voz de tres personas distintas (un hombre y dos mujeres) mientras leen, por separado, un texto. La cuarta fuente es un fragmento de una pieza musical. Las fuentes y su correspondiente desviación típica (σ), que sirve para tener una idea acerca de su potencia, son:

- *Fuente 1* (hombre): " Desde muy niño destacó ", $\sigma = 13'76$
- *Fuente 2* (mujer): " Bélgica, Bermudas, Brasil ", $\sigma = 7'98$
- *Fuente 3* (mujer): " Grupo poblacional ", $\sigma = 10'80$
- *Fuente 4*: Fragmento de " Las Cuatro Estaciones" de Vivaldi, $\sigma = 3'15$

La frecuencia de muestreo de las señales es de 11.000 Hz. Las cuatro se muestran en la Figura 7.4. La matriz de mezcla se toma

$$\mathbf{A} = \begin{bmatrix} 1.16 & -0.70 & 0.26 & 1.25 \\ 0.63 & 1.70 & 0.87 & -0.64 \\ 0.07 & 0.06 & -1.45 & 0.58 \\ 0.35 & 1.80 & -0.70 & -0.36 \end{bmatrix}$$

que tiene un número de condición bastante alto, $c(\mathbf{A}) = 275'06$. Las observaciones están en la Figura 7.5. En la primera, segunda y cuarta observación, la mezcla es fuerte y casi ininteligible.

Los estadísticos se estiman a partir de 12.500 muestras (poco más de un segundo) de las observaciones. Tras la Separación, se obtiene la *matriz global de transferencia*

$$\mathbf{G} = \begin{bmatrix} 0.075 & 0.003 & -0.001 & -0.003 \\ -0.002 & 0.125 & 0.003 & -0.003 \\ 0.001 & 0.003 & -0.092 & 0.002 \\ -0.001 & 0.003 & -0.003 & -0.318 \end{bmatrix}$$

Las fuentes se han atenuado mucho; pero esto no es relevante. La calidad subjetiva de la Separación, tras la audición de las señales, es muy buena. De hecho, los índices de esta matriz **G** son I_{AM} = 0'0461 e I_{MM} = 0'006. Las señales de salida se presentan en la Figura 7.6.

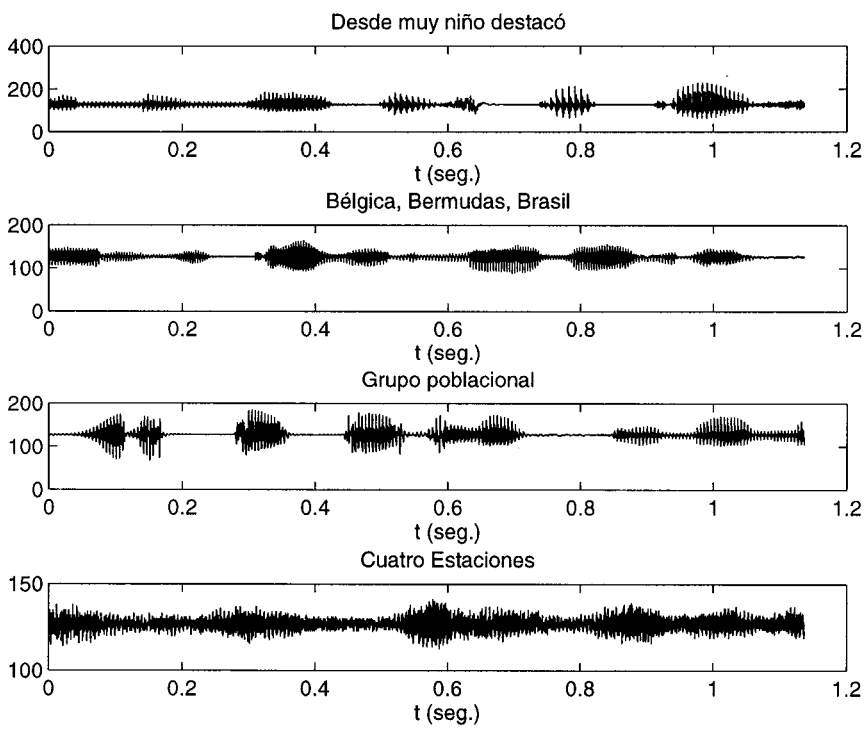

Figura 7.4. Fuentes del Experimento 4.5

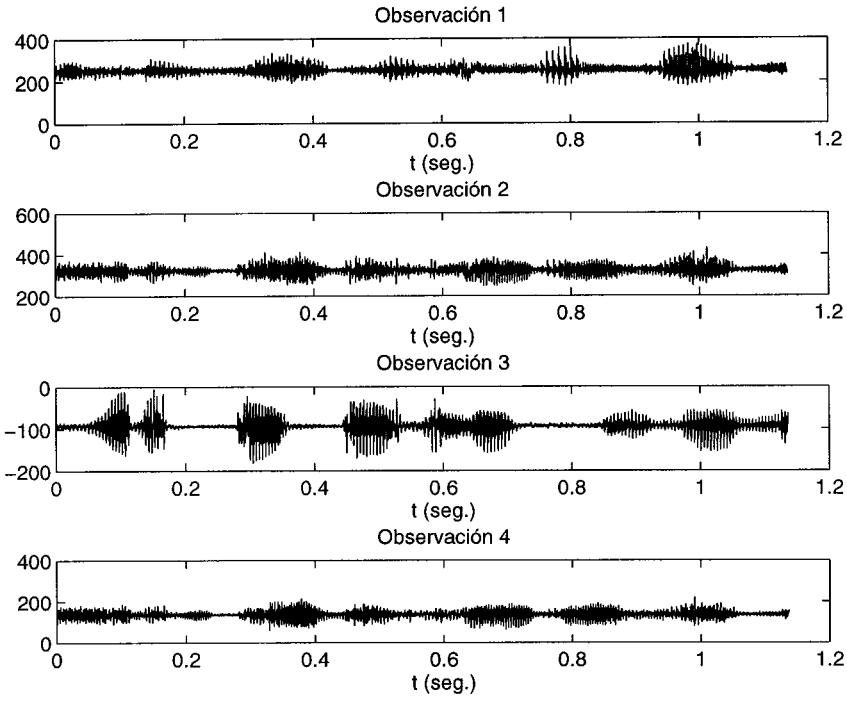

Figura 7.5. Observaciones del Experimento 4.5

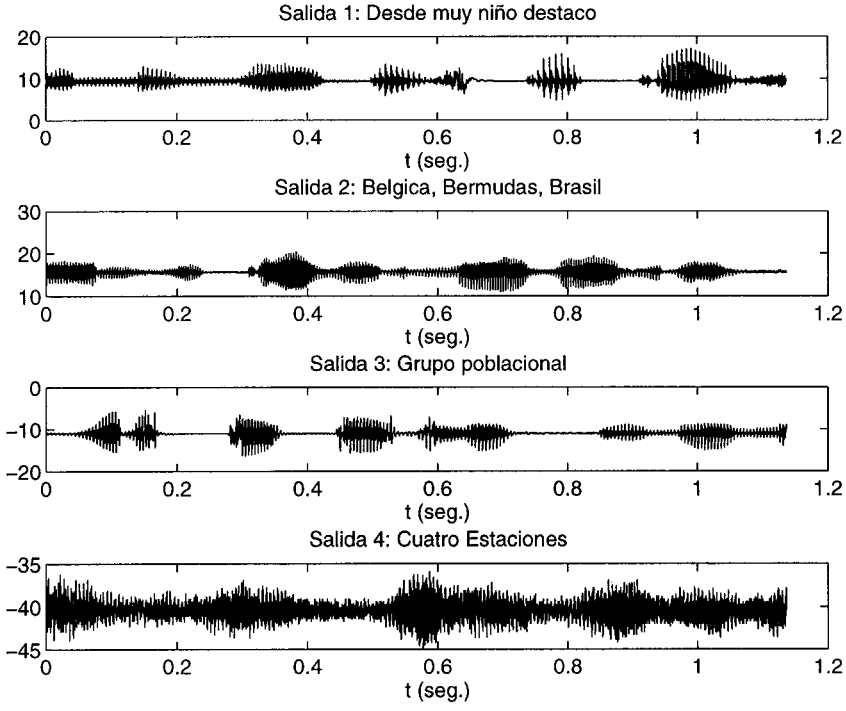

Figura 7.6. Observaciones del Experimento 4.5

7.6 Comparación entre Algoritmos

A fin de comparar con otros algoritmos, hemos creado una mezcla de *diez* fuentes, cinco de distribución uniforme (sub-gaussiana) y cinco generadas como el cubo de una variable gaussiana (super-gaussiana), cada una con 5000 muestras. La *matriz de mezcla* se escogió aleatoriamente para ser estimada mediante cuatro algoritmos de *bloque*:

- SEVILLA, presentado en esta Tesis.
- JADE [Cardoso93]
- FastIca [Hyvärinen97]
- ICA [Comon94]

Para este fin, se han obtenido a través de Internet las implementaciones que los propios autores han hecho de los algoritmos. Hemos utilizado los programas *tal y como se encuentran en la red*, sin ninguna modificación salvo en el caso del algoritmo ICA de Comon, al que se han añadido las oportunas líneas de código para que pueda manejar señales cuya media no es cero.

Se miden los siguientes parámetros:

- *La complejidad computacional*, como el número de operaciones que realiza el algoritmo, medidos con el comando 'flops' de Matlab. Por convención, la complejidad de SEVILLA se toma igual a *uno*.
- *La relación señal a ruido media* (SNR) después de conseguir la Separación
- Los índices de Separación definidos en (7.1) y (7.2).

Los resultados de la siguiente Tabla son el promedio de 20 experimentos independientes. Cada entrada en la Tabla consta de dos números: el primero es la

media del resultado mientras que el segundo, entre paréntesis, es su desviación típica.

	Operaciones	SNR (dB)	I_{AM}	I_{MM}
SEVILLA	1 (0)	14.40 (1.26)	1.88 (0.13)	0.016 (0.003)
JADE	5.28 (0.19)	15.78 (1.11)	1.41 (0.13)	0.008 (0.001)
FastIca	1.92 (0.08)	14.42 (1.35)	1.88 (0.13)	0.016 (0.003)
ICA	2.57 (0.04)	15.70 (0.41)	1.45 (0.14)	0.010 (0.002)

La Tabla muestra resultados similares para todos los algoritmos en cuanto a la calidad de la Separación. Sin embargo, SEVILLA es, computacionalmente, el más simple. También se muestra algo llamativo: se suele afirmar [Cardoso97] que JADE es mucho más eficiente que ICA; sin embargo, nuestros experimentos muestran que sólo es así cuando el número de fuentes es pequeño.

Tiene ya menos sentido comparar SEVILLA con algoritmos típicamente adaptativos. Los algoritmos INFOMAX, INFOMAX Extendido y EASI se encuentran en la red Internet, programados respectivamente por los propios Bell, Lee y Cardoso. Sin embargo, difícilmente se puede decir que mejoren los resultados obtenidos por los algoritmos de bloque, ni siquiera en lo referente al número de operaciones: los algoritmos adaptativos dependen por completo del *paso de adaptación que se escoja.* Por ejemplo, cuando el paso de adaptación es pequeño, el MLE consigue resultados poco mejores (del mismo orden de magnitud) que SEVILLA; pero necesita muchas más muestras de las observaciones para ello. No entramos a valorar qué ocurre en entornos no estacionarios.

7.7 Mezcla de un número grande de fuentes

En este experimento, mezclamos 30 fuentes de distribución uniforme. La matriz de mezcla se escoge de forma aleatoria. La Figura 7.7 muestra la evolución del índice de Moreau y Macchi en función del número de iteraciones del algoritmo. La curva es el resultado de promediar cinco experimentos independientes en los que los estadísticos de las observaciones se estiman a partir de 5000 muestras.

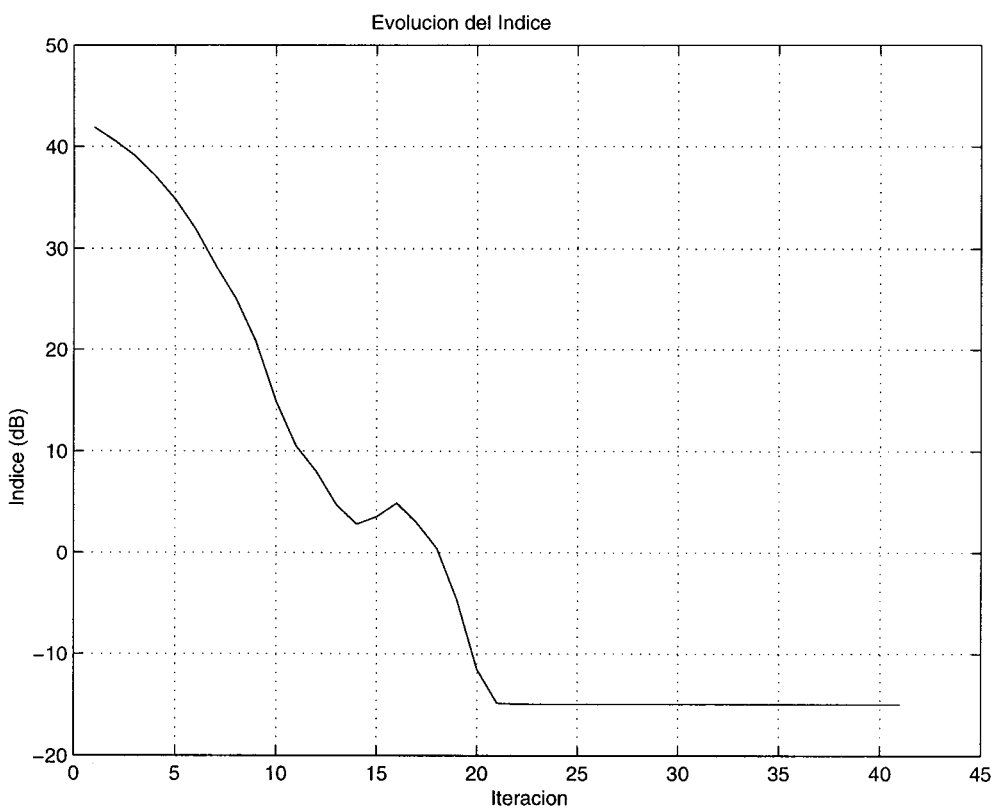

Figura 7.7. Evolución del Índice en el Experimento de la Sección 4.7.

7.8 Conclusiones

El algoritmo SEVILLA ha demostrado ser capaz de Separar las Fuentes de forma satisfactoria en los experimentos que hemos llevado a cabo. En particular, merece la pena que destaquemos su capacidad de tratar mezclas con un gran número de fuentes, lo que no todos los algoritmos pueden hacer en la práctica. Por otra parte, ha funcionado bien con señales no estacionarias, aunque este aspecto no se ha estudiado analíticamente. Finalmente, la comparación con otros algoritmos muestra que SEVILLA es muy competitivo.

Uno de los puntos fuertes de SEVILLA, que comparte con el resto de los algoritmos de bloque, es que no tiene parámetros que requieran un ajuste cuidadoso. Sin embargo, como los otros algoritmos de bloque, sus prestaciones se degradan cuando las fuentes están *temporalmente* correladas.

8. Conclusiones y Líneas Futuras de Investigación

8.1 Conclusiones

Hemos probado que la Separación Ciega de Fuentes es perfectamente posible, supuesto que el modelo de la mezcla es *lineal*, que las fuentes son *estadísticamente independientes* y que, a lo más, *una de ellas tiene distribución gaussiana*, siempre que el número de sensores sea *igual* que el número de fuentes presentes en la mezcla.

Por supuesto, como también se ha apuntado, éstas no son las únicas hipótesis a partir de las cuáles es posible separar las fuentes (en particular, se puede sustituir la de independencia estadística por otras relativas a la forma de onda de las señales o a su carácter discreto, por ejemplo); pero sí son las únicas con las que se ha trabajado en esta Tesis.

El estimador de máxima verosimilitud de la matriz de mezcla es *eficiente* en el sentido de que su varianza tiende asintóticamente a la cota de Cramér-Rao. Sin embargo, sus ecuaciones se formulan a partir de las funciones de densidad de probabilidad de las fuentes, que, por hipótesis, son desconocidas. Aunque estas distribuciones de probabilidad podrían ser estimadas [Pham96, Amari96, Comon94], la mayor parte de los algoritmos simplemente las aproximan por medio de funciones fijadas de antemano. Este procedimiento simple se justifica porque, bajo condiciones muy generales, la estimación es robusta (*supereficiente*) [Amari98*b*]. Sin embargo, también es sabido que una mala elección impide la convergencia de los algoritmos y la Separación de las Fuentes [Bell95].

Alternativamente, en esta Tesis Doctoral hemos propuesto una nueva caracterización de las *matrices de separación* que se basa en la cancelación de las

derivadas de segundo orden de ciertos cumulantes cruzados. Se ha demostrado que esta caracterización no sólo es *necesaria*, sino también *suficiente*. A partir de ella se consigue un sistema de ecuaciones *lineales* cuya solución es la *matriz de separación* deseada.

Finalmente, se ha propuesto un algoritmo iterativo (SEVILLA) que hace uso de las ideas expuestas. Se ha probado que el algoritmo *siempre* converge a una matriz de separación, independientemente de sus condiciones iniciales (esto es destacable, por cuanto que la mayor parte de los autores sólo prueban la convergencia *local* de sus métodos). No es la única manera de utilizar la información que contienen las derivadas de los cumulantes (ver por ejemplo en [Martín00*b*] un procedimiento alternativo, que separa las fuentes por parejas), aunque SEVILLA parece ser la mejor opción.

La práctica muestra que el algoritmo resultante es capaz de trabajar con un gran número de fuentes. Además, es fuertemente competitivo por cuanto que la precisión de sus soluciones es similar a la obtenida por otros algoritmos, teniendo un menor coste computacional.

8.2 Líneas Futuras de Investigación

En primer lugar, suponer que el número de sensores es mayor o igual que el número de fuentes es una hipótesis demasiado restrictiva. Cuando no se cumple (por ejemplo, en el caso de añadir ruido a las observaciones), comprobamos mediante simulaciones que los algoritmos y procedimientos presentados en esta Tesis son, con frecuencia, capaces de extraer alguna(s) fuente(s) con una relación señal a ruido aceptable. No obstante, la forma en la que operan los algoritmos en estas condiciones debe ser investigada en profundidad. En particular, no se han presentado aún condiciones que aseguren la convergencia de los algoritmos y la separación, por tanto, de alguna de las señales. El uso de información adicional

sobre las fuentes, por ejemplo, su distribución estadística en aquéllos casos en los que se conozca [Te-Won99], parece dar buenos resultados y debería ser considerada en el futuro.

En lo que se refiere al algoritmo SEVILLA, es la estimación de los estadísticos de las señales y *no la resolución de las ecuaciones* quien consume la mayor parte del esfuerzo computacional. En [Martín99*b*] se proponen métodos para el cálculo eficiente de estos estadísticos. Por otra parte, hemos presentado una serie de condiciones necesarias y suficientes que garantizan la Separación. Ahora bien, comprobamos que se puede obtener un buen resultado utilizando un subconjunto reducido de condiciones; es decir, forzando sólo algunas de ellas (al menos, la mitad del total), los algoritmos tienden a satisfacer de forma natural las restantes. Este es un hecho experimental que debe ser confirmado por el estudio analítico.

Por último, hay que investigar en profundidad las consecuencias de que, en la práctica, no se satisfagan todas las hipótesis de partida; en particular, que la mezcla no sea lineal o invariante en el tiempo o que las señales no sean estacionarias o completamente independientes entre sí.

Apéndices

APÉNDICE A.

Algunos conceptos propios de la Teoría de la Información.

La Distancia de Kullback-Leibler

Sean p_1 y p_2 dos funciones de densidad de probabilidad. Se llama distancia de Kullback-Leibler (*dKL*) entre las correspondientes distribuciones a la magnitud [Cover91, Borovkov88]

$$dKL[\ p_1 \parallel p_2\ ] = \int_{-\infty}^{\infty} p_1(\mathbf{u}) \log \frac{p_1(\mathbf{u})}{p_2(\mathbf{u})} d\mathbf{u} = \\ = -\int_{-\infty}^{\infty} p_1(\mathbf{u})\ \log \frac{p_2(\mathbf{u})}{p_1(\mathbf{u})} d\mathbf{u} \qquad (A.1)$$

(donde los argumentos de *dKL* se escriben entre corchetes y no entre paréntesis para enfatizar el hecho de que *dKL* depende de las distribuciones estadísticas, no de las variables en sí). En realidad, la $dKL[\ p_1 \parallel p_2\ ]$ no es realmente una distancia o una métrica en sentido estricto, ya que ni es una función simétrica de p_1 y p_2 ni cumple la desigualdad del triángulo. No obstante, veremos que $dKL[\ p_1 \parallel p_2\ ]$ caracteriza la desviación de p_2 respecto a p_1.

De la desigualdad $\log(1 + x) \leq x$ para $x \geq -1$, donde el signo de igualdad se da sólo para $x = 0$, se sigue que

$$\log \frac{p_2}{p_1} = \log(1 + (\frac{p_2}{p_1} - 1)) \leq \frac{p_2}{p_1} - 1 \qquad (A.2)$$

y el signo de igualdad sólo es posible si $p_1 = p_2$. Entonces,

$$-dKL[\ p_1 \,||\, p_2\ ] \leq \int_{\mathbf{u}} p_1(\mathbf{u})\left(\frac{p_2(\mathbf{u})}{p_1(\mathbf{u})} - 1\right)d\mathbf{u} = \int p_2 d\mathbf{u} - \int p_1 d\mathbf{u} = 0 \quad \text{(A.3)}$$

De aquí se deduce la propiedad más importante de la distancia Kullback-Leibler: "La $dKL[\ p_1 \,||\, p_2\ ] \geq 0$, dándose la igualdad si y sólo si $p_1(\mathbf{u}) = p_2(\mathbf{u})$ (para casi toda $\mathbf{u}$)". Otras propiedades interesantes de la *dKL* se enuncian a continuación sin ser demostradas [Cover91]

- La *dKL* es invariante ante cualquier transformación invertible g del espacio muestral:

$$dKL[\ p_1(\mathbf{u}) \,||\, p_2(\mathbf{u})\ ] = dKL[\ p_1(\ g(\boldsymbol{u})\ ) \,||\, p_2(\ g(\boldsymbol{u})\ )\ ]$$

- $dKL[\ p_1 \,||\, p_2\ ] = dKL[\ p_1 \,||\, q_1\ ] + dKL[\ q_1 \,||\, p_2\ ]$, donde $q_1(\ \mathbf{u}\ ) = \prod_i p_{1i}(u_i)$ y $p_{1i}(u_i)$ es la *f.d.p* marginal de la componente u_i de $\mathbf{u}$, esto es,

$$p_{1i}(\ u_i\ ) = \int_{\mathbf{u}} p_1(\ \mathbf{u}\ )\, du_1 \ldots du_{i-1}\, du_{i+1} \ldots du_N$$

Aprovechamos para definir la información mutua entre las componentes de $\mathbf{u}$, que se denota por $I[\ \mathbf{u}\ ]$, como la cantidad

$$I[\ \mathbf{u}\ ] = dKL[\ p_1 \,||\, q_1\ ] \quad \text{(A.4)}$$

Se sigue de inmediato que $I[\ \mathbf{u}\ ] \geq 0$, dándose la igualdad si y sólo si las componentes de $\mathbf{u}$ son mutuamente independientes.

Como se dijo, la *dKL* no es rigurosamente una distancia, a diferencia de, por ejemplo, la raíz cuadrada de $\int(\sqrt{p_1} - \sqrt{p_2})^2$ (distancia de Hellinger) [Borovkov88].

Definición. Se llama *negentropía* [Comon94] de una distribución a la distancia de Kullback-Leibler entre la correspondiente densidad de probabilidad p y la densidad

p_G de la variable gaussiana que tiene los mismos estadísticos de primer y segundo orden. Denotaremos la negentropía por $nH[\,p\,]$. Así:

$$nH[\,p\,] = dKL[\;p \,\|\, p_G\,] \tag{A.5}$$

De las propiedades de la dKL se deduce que $nH[\,p\,] \geq 0$ y, precisamente, de aquí viene su nombre, como justificaremos ahora: por definición

$$nH[\,p\,] = \int p(\mathbf{u}) \log \frac{p(\mathbf{u})}{p_G(\mathbf{u})} d\mathbf{u} = -H[\,p\,] - \int p(\mathbf{u}) \log p_G(\mathbf{u}) d\mathbf{u}$$

ahora bien, como $\log p_G(\,\mathbf{u}\,)$ es una forma cuadrática y los momentos de primer y segundo orden de p y p_G coinciden, resulta que

$$\int p(\mathbf{u}) \log p_G(\mathbf{u}) d\mathbf{u} = \int p_G(\mathbf{u}) \log p_G(\mathbf{u}) d\mathbf{u} = -H[\,p_G\,]$$

por lo tanto, $nH[\,p\,] = H[\,p_G\,] - H[\,p\,]$,

así que la negentropía no es más que la entropía desplazada. Como $nH[\,p\,] \geq 0$ mientras que $H[\,p\,] \leq 0$, $nH[\,p\,]$ se interpreta como la negación de la entropía (*neg*-entropía). Nótese que hemos probado igualmente que $H[\,p_G\,] \geq H[\,p\,]$, que es la conocida propiedad que dice que las variables gaussianas tienen la mayor entropía de entre todas las variables aleatorias con igual media y varianza [Cover91].

APÉNDICE B.

Recordatorio de Estadística

B.1) MOMENTOS Y CUMULANTES

Se llama función característica de la variable aleatoria x a la función α definida por la fórmula [Nikias93]:

$$\alpha(t) = E[e^{jtx}], \qquad \text{con } j = \sqrt{-1} \tag{B.1}$$

Si existe el momento k-ésimo de la distribución, la función característica puede desarrollarse en serie de McLaurin:

$$\alpha(t) = 1 + \sum_{q=1}^{k} \frac{m_q}{q!} (jt)^q + O(t^k) \tag{B.2}$$

donde $m_q = E[x^q]$. Para la función $\log(1 + v)$ tenemos el desarrollo correspondiente

$$\log(1 + v) = v - v^2/2 + \ldots \pm v^k/k + O(v^k) \tag{B.3}$$

Si sustituimos en el desarrollo anterior $1 + v$ por $\alpha(t)$, obtenemos, después de ordenar de nuevo los términos, un desarrollo de la forma

$$\log \alpha(t) = \sum_{q=1}^{k} \frac{\kappa_q}{q!} (jt)^q + O(t^k) \tag{B.4}$$

Los coeficientes κ_q fueron introducidos por T. N. Thiele en 1903 [Cramér70] y reciben el nombre de semiinvariantes o *cumulantes* de la distribución. Para deducir la relación entre los momentos y los cumulantes, emplearemos la identidad

$$\log \alpha(t) = \log\left(1 + \sum_{q=1}^{\infty} \frac{m_q}{q!} (jt)^q\right) = \sum_{q=1}^{\infty} \frac{\kappa_q}{q!} (jt)^q \qquad \text{(B.5)}$$

que, sin prestar atención a las cuestiones de existencia de los momentos o convergencia de las series, muestra que los *cumulantes* son funciones polinómicas de los *momentos* y viceversa. En particular:

$\kappa_1 = m_1 = \mu$ (esperanza de x)

$\kappa_2 = m_2 - m_1^2 = \sigma^2$ (varianza de x)

$\kappa_3 = m_3 - 3\, m_1 m_2 + 2\, m_1^3$ (coeficiente de asimetría o de deformación)

$\kappa_4 = m_4 - 3\, m_2^2 - 4\, m_1 m_3 + 12\, m_1^2 m_2 - 6\, m_1^4$ (coeficiente de exceso o *curtosis*)

La variable aleatoria x se llama "*platicúrtica*", "*mesocúrtica*" o "*leptocúrtica*" según que $\kappa_4 > 0$, $\kappa_4 = 0$ ó $\kappa_4 < 0$, respectivamente. Las variables aleatorias gaussianas son *mesocúrticas*, es decir, su curtosis vale cero. Si bien "*platicúrtica*" y "*leptocúrtica*" son los términos tradicionales, también son empleadas (aunque con ciertas reservas [Mansour99]) las voces "*super-gaussiana*" (variable de curtosis positiva) y "*sub-gaussiana*" (variable de curtosis negativa), tomadas directamente del inglés. En esta Tesis preferimos utilizar estas últimas. La curtosis es una medida del grado de "*aplastamiento*" de la función de densidad alrededor de su moda (máximo). Se supone que cuando la curtosis es positiva (super-gaussiana) la curva de densidad es más alta y esbelta que la de la densidad normal en las proximidades de la moda, ocurriendo lo contrario para las sub-gaussianas. De todas formas, estas afirmaciones deben acogerse con reservas [Cramér70].

De igual forma que hemos definido los cumulantes de una variable, podemos introducir el concepto de *cumulante cruzado* de un conjunto de variables aleatorias $\{x_1, ..., x_N\}$: la *función característica* de las variables es ahora:

$$\alpha(a_1, ..., a_N) = E[\exp j\,(a_1 x_1 + ... + a_N x_N)] \tag{B.6}$$

Los coeficientes del desarrollo en serie de $\log[\alpha(a_1, ..., a_N)]$ reciben el nombre de *cumulantes cruzados* entre las variables. En [Nikias93] se introducen como:

$$cum(x_1, ..., x_N) = (-j)^N \left.\frac{\partial^N \log \alpha(a_1, ..., a_N)}{\partial a_1 ... \partial a_N}\right|_{a_1 ... a_N = 0} \tag{B.7}$$

En particular, si todas las variables tienen *media cero*, resultan de especial interés los *cumulantes cruzados* que vamos a definir ahora:

1.- $cum_{31}(x_1, x_2) = E[x_1^3 x_2] - 3\,E[x_1^2]\,E[x_1 x_2]$
2.- $cum_{13}(x_1, x_2) = E[x_1 x_2^3] - 3\,E[x_2^2]\,E[x_1 x_2]$
3.- $cum_{22}(x_1, x_2) = E[x_1^2 x_2^2] - E[x_1^2]\,E[x_2^2] - 2\,E^2[x_1 x_2]$

B. 2) PROPIEDADES DE LOS CUMULANTES

Entre otras, las más interesantes para nuestros fines son [Nikias93]:

1.- Todos los cumulantes de orden superior a dos de una variable aleatoria gaussiana se *anulan*. En particular, la *curtosis* de una variable gaussiana siempre vale *cero*.

2.- Los *cumulantes cruzados* de variables aleatorias *estadísticamente independientes* siempre *se anulan*.

3.- El *cumulante* de una suma de N variables *independientes* es igual a la suma de los cumulantes de orden N de cada una de ellas.

4.- Si la función de densidad de probabilidad de las variable aleatoria es *par*, entonces se anulan todos sus cumulantes de orden *impar*.

B.3) DESARROLLOS DERIVADOS DE LA DISTRIBUCIÓN NORMAL

La manera usual de determinar de forma aproximada una *f.d.p* desconocida es desarrollar la misma a partir de una *f.d.p* gaussiana de igual media y varianza. Dada la variable aleatoria y, de media cero y varianza unidad, cuya *f.d.p* es $p(\ y\ )$, se escribe [Cramér70, pág. 253 y ss.; Stuart93, pág. 228 y ss.]:

$$p(\ y\ ) \approx \mathrm{N}(0,1)(\ 1 + h(\ y\ )\ ) \qquad \text{(B.8)}$$

donde $\mathrm{N}(0,1) = \frac{1}{\sqrt{2\pi}} \exp(\ -\ y^2/2\ )$ y la función $h(\ y\ )$ depende de los estadísticos de la variable y del tipo de desarrollo que se decida emplear.

El desarrollo ortogonal de Gram-Charlier

Escribamos $p(\ y\ )$ desarrollándola en lo que se conoce como serie A de Gram-Charlier:

$$p(\ y\ ) = c_0\, \mathrm{N}(0,1) + \frac{c_1}{1!}\, \mathrm{N'}(0,1) + \frac{c_2}{2!}\, \mathrm{N''}(0,1) + \ldots$$

donde los c_i son los coeficientes del desarrollo y el apóstrofe denota *derivación*. Resulta que

$$\mathrm{N}^{(k)}(0,1) = (\ -1\ )^k\, CH_k(\ y\ )\, \mathrm{N}(\ 0,\ 1\ )$$

donde $CH_k(\ y\ )$ es el polinomio de Chebyshev-Hermite de orden k, que también se define mediante la recursión [Comon94]:

$$CH_0(y) = 1 \qquad CH_1(y) = y \qquad CH_{k+1}(y) = y\,CH_k(y) - \frac{\partial}{\partial y} CH_k(y)$$

Los coeficientes c_i del desarrollo se obtienen multiplicando $p(\ y\ )$ por el polinomio $CH_k(\ y\ )$ e integrando a continuación, de manera similar a como se evalúan los coeficientes de una serie de Fourier. Este procedimiento está justificado por la ortogonalidad de las funciones base del desarrollo de Gram-Charlier.

Así, con referencia a la fórmula anterior, se obtiene para los primeros términos del desarrollo:

$$h(\ y\ ) = \frac{\kappa_3}{3!} CH_3(\ y\ ) + \frac{\kappa_4}{4!} CH_4(\ y\ )$$

Puede demostrarse que la serie es *convergente*; aunque necesita sumar muchos términos para que la aproximación a $p(\ y\ )$ sea razonable, al menos cuando y es la suma de unas variables aleatorias independientes, como es el caso en el problema de Separación de Fuentes. En esta situación, resulta más preciso el desarrollo de Edgeworth, que introducimos a continuación.

El desarrollo asintótico de Edgeworth

El desarrollo de Edgeworth de tipo A y orden cuatro es aquél que define:

$$\begin{aligned} h(\ y\ ) = {} & \frac{\kappa_3}{3!} CH_3(y) + \frac{\kappa_4}{4!} CH_4(y) + \frac{10\kappa_3^2}{3!} CH_6(y) + \\ & + \frac{\kappa_5}{5!} CH_5(y) + \frac{35\kappa_3\kappa_4}{7!} CH_7(y) + \frac{280\kappa_3^3}{9!} CH_9(y) + \\ & + \frac{\kappa_6}{6!} CH_6(y) + \frac{56\kappa_3\kappa_5}{8!} CH_8(y) + \frac{35\kappa_4^2}{8!} CH_8(y) + \end{aligned}$$

$$+ \frac{2100\kappa_3^2\kappa_4}{10!}CH_{10}(y) + \frac{15400\kappa_3^4}{12!}CH_{12}(y)$$

Como vemos, los primeros términos coinciden con los de la expansión de Gram-Charlier. No interesa añadir más términos al desarrollo de Edgeworth, pues para órdenes superiores al cuarto puede haber excesivas fluctuaciones en la serie, pudiendo ocurrir incluso que la *f.d.p* tome valores negativos. Si y es la suma de N variables aleatorias independientes con cumulantes finitos, como va a ser el caso, se puede demostrar que κ_i es del orden de $N^{(2-i)/2}$ [Comon94]. Entonces, los coeficientes de la expansión de Edgeworth decrecen uniformemente.

APÉNDICE C.

El Estimador de Máxima Verosimilitud

Sea una variable aleatoria X cuya distribución está parametrizada por θ, donde este parámetro *no* es una cantidad *aleatoria*. Dadas T muestras $x(1)$, ..., $x(T)$ de X se define la función de *verosimilitud* de la muestra como [Johnson93, Casella90]

$$e(\theta) = p(x(1), ..., x(T); \theta) \qquad (C.1)$$

siendo p la *función de densidad de probabilidad* conjunta de las T muestras.

Definición. El *estimador de máxima verosimilitud* (MLE) de θ es aquél valor del parámetro que maximiza $e(\theta)$.

Nota. Ya que las funciones de densidad de probabilidad suelen depender exponencialmente de sus variables, en la práctica no se opera con la función de *verosimilitud* sino con su logaritmo. Ello no altera las estimaciones pues $e(\theta)$ y $\log\{e(\theta)\}$ tienen los mismos máximos debido a que el logaritmo es una *función monótona creciente.*

Es importante destacar que, estrictamente, *no* se puede hacer una estimación de máxima verosimilitud si *no* se conoce p, la *función de densidad de probabilidad* conjunta de las muestras.

El MLE es el estimador *más utilizado* debido a que:

1.- A*sintóticamente* (para un número de muestras creciente), no tiene sesgo. En muchos casos, el estimador es insesgado para cualquier número de muestras.

2.- Es *consistente*, es decir, su varianza tiende a cero si el número de muestras tiende a infinito.

3.- Es asintóticamente *eficiente*, es decir, su varianza tiende hacia la cota de Cramér-Rao a medida que crece el número de muestras (en muchos casos, la varianza del estimador coincide con esta cota independientemente del número de muestras utilizado). Recordemos que la cota de Cramér-Rao [Cramér70, pág. 549; Johnson93, pág. 281]es la cota inferior de la varianza de cualquier estimador, para un conjunto de muestras dado; es decir, en particular la cota de Cramér-Rao depende del número de muestras disponibles, por lo que no hay contradicción con la propiedad de *consistencia*: si el número de muestras crece, la cota de Cramér-Rao tiende a cero.

4.- El MLE *asintóticamente* se distribuye como una variable aleatoria gaussiana.

Estas definiciones y propiedades se extienden rápidamente al caso en el que el parámetro θ es un vector, del que hay que estimar sus componentes.

Ejemplo. Sea el proceso AR de primer orden [Söderström89, pág. 199]:

$$y(t) + a\, y(t-1) = e(t),$$

siendo $| a | < 1$ y $e(t)$ un ruido blanco gaussiano de media cero y varianza σ^2. En este caso, $\theta = [a \; \sigma]^T$ y la función de verosimilitud puede ser evaluada con facilidad empleando la *regla de Bayes*:

$$p(y(1), ..., y(T)) = p(y(2), ..., y(T) \mid y(1)) \; p(y(1)) =$$

$$= p(y(3), ..., y(T) \mid y(1) \; y(2)) \; p(y(2) \mid y(1)) \; p(y(1)) =$$

$$= \{ \Pi_{k=2}^{T} \; p(y(k) \mid y(k-1) \; ... \; y(1)) \} \; p(y(1))$$

donde, para $k > 1$,

$$p(y(k) \mid y(k-1) \; ... \; y(1)) = p(e(k)) =$$

$$= \frac{1}{\sqrt{2\pi\sigma}} \exp(-(y(k) + a \; y^2(k-1))/2\sigma^2)$$

siendo

$$p(y(1)) = \frac{1}{\sqrt{2\pi\tau}} \exp(- y^2(1) / 2\tau^2), \text{ con } \tau^2 = \sigma^2 /(1 - a^2).$$

Esta última expresión se debe a que $y(t)$ es, para todo t, una variable aleatoria gaussiana, de media cero y varianza τ^2, como es sencillo probar.

Se ha escogido el ejemplo anterior por una razón: a menudo se utilizan los estimadores de *máxima verosimilitud* bajo la presunción de que las distintas muestras de la variable aleatoria son *estadísticamente independientes* entre sí. Sin embargo, como vemos, esto no es necesario.

Referencias

REFERENCIAS

[Amari96] S. Amari, A. Cichocki y H. Yang, A new Learning Algorithm for Blind Signal Separation, en Advances in Neural Information Processing Systems, editores D. Touretzky, M. Mozer y M.Hasselmo, págs. 757-763, MIT Press, 1996.

[Amari97*a*] S. Amari y J.F. Cardoso, Blind Source Separation-Semiparametric Statistical Approach, IEEE. Trans. on SP., 45(11): 2692-2700, 1997

[Amari97*b*] S. Amari, T. Chen y A. Cichocki, Stability Analysis of Adaptive Blind Source Separation, *Neural Networks*, 10 (8): 1345-1351, 1997.

[Amari98*a*] S. Amari, Natural Gradient Works Efficiently in Learning, *Neural Computation*, 10:251-276, 1998.

[Amari98*b*] S. Amari, Superefficiency in Blind Source Separation, *IEEE. Trans on Signal Processing*, 47 (8): 936-944.

[Barlett97] M. Barlett y T. Sejnowski, Viewpoint invariant face recognition using independent component analysis and attractor networks, *Advances in Neural Information Processing Systems*, 9:817-823, MIT Press, 1997.

[Bell95] A. Bell, T. Sejnowski, An Information-Maximization Approach to Blind Separation and Blind Deconvolution, *Neural Computation*, 7:1129-1159.

[Bell97] A. Bell, T. Sejnowski, The 'independent components' of natural scenes are edge filters, *Vision Research*, 37(23):3327-3338.

[Belouchrani94] A. Belouchrani y J.F. Cardoso, Maximum Likelihood Source Separation for Discrete Sources, Actas del Congreso EUSIPCO'94, Edimburgo, 1994.

[Belouchrani97] A. Belouchrani, K. Abed-Meraim, J.F. Cardoso y E. Moulines, A Blind Source Separation Technique Using Second-Order Statistics, *IEEE Trans. on Signal Proc.*, 45(2): 434-444, 1997.

[Benveniste90] A. Benveniste, M. Métivier, P. Priouret, "Adaptive Algorithms and Stochastic Approximations", Editorial Springer-Verlag, 1990.

[Borovkov88] A.A. Borovkov, "Estadística Matemática", Editorial MIR, Moscú, 1988.

[Cao96] X.R. Cao, R. Liu, General Approach to Blind Source Separation, *IEEE Trans. on Signal Proc.*, 44(3): 562-571, 1996.

[Cardoso89] J.F. Cardoso, Source Separation Using Higher Order Moments, Actas del Congreso ICASSP'89, 1989.

[Cardoso93] J.F. Cardoso y A. Souloumiac, Blind Beamforming for Non-Gaussian Signals, *IEE Proceedings-F*, 140(46):362-370, 1993.

[Cardoso95] J.F. Cardoso, The Invariant Approach to Source Separation, Actas del Congreso NOLTA95.

[Cardoso96*a*] J.F. Cardoso y B. Laheld, Equivariant Adaptive Source Separation, IEEE. Trans on S.P., 45(2): 434-444.

[Cardoso96*b*] J.F. Cardoso, S. Bose y B. Friedlander, On Optimal Source Separation Based on Second and Fourth Order Cumulants, Actas del Congreso IEEE SSAP, Corfú, 1996.

[Cardoso97*a*] J.F. Cardoso, Infomax and Maximum Likelihood for Blind Source Separation, IEEE Signal Proc. Letters, 4(4):112-114.

[Cardoso97*b*] J.F. Cardoso, Estimating Equations for Source Separation, Actas del Congreso ICASSP'97, Munich, 1997.

[Cardoso98*a*] J.F. Cardoso, Blind Signal Separation: Statistical Principles, *Proceedings of the IEEE*, 86(10): 2009-2025, 1998.

[Cardoso98*b*] J.F. Cardoso, Multidimensional independent component analysis, Actas del Congreso ICASSP'98, 1998.

[Casella90], G. Casella, R. Berger, "Statistical Inference", 1ª Ed., Ed. Wadsorth & Brooks/Cole, Pacific Grove, EEUU, 1990.

[Chevalier99] P. Chevalier, V. Capdevielle, P. Comon, Performance of HO blind source separation methods: experimental results on ionospheric HF links, Actas del Congreso ICA'99, Aussois, Francia, 1999.

[Comon91] P. Comon, C. Jutten y J. Herault, Blind Separation of Sources, part II: Problems Statement, *Signal Processing,* 24:11-20, 1991.

[Comon94] P. Comon, Independent Component Analysis - A New Concept?, *Signal Processing,* 36(3):287-314, 1994.

[Comon96] P. Comon, Contrasts for multichannel blind deconvolution, *Signal Processing Letters*, 3(7):209-211, 1996.

[Cover91] T. Cover y J. Thomas (Editores), "Elements of Information Theory", Vol. 1, John Wiley and Sons, Nueva York, 1991.

[Cramér70] H. Cramér, "Métodos Matemáticos de Estadística", 1ª Ed., Ed. Aguilar, 1970.

[Delfosse95] N. Delfosse y P. Loubaton, Adaptive Blind Separation of Independent Sources: A Deflation Approach, *Signal Processing*, 45(1): 59-83, 1995.

[Deville96] Y. Deville, A Unified Stability Analysis of the Herault-Jutten source separation neural network, *Signal Processing*, 51: 229-233, 1996.

[Douglas97] S. Douglas y A. Cichocki, Neural Networks for Blind Decorrelation of Signals, *IEEE Trans. on Signal Proc.*, 45(11):2829-2842, 1997.

[Girolami97] M. Girolami y C. Fyfe, Extraction of independent signal sources using a deflationary exploratorary projection pursuit network with lateral inhibition, *I.E.E Proceedings on Vision, Image and Signal Processing Journal*, 14(5):299-306.

[Gnedenko75] B. Gnedenko, "The Theory of Probability", Ed. MIR, Moscú, 1975.

[Golub96] G. Golub y C. Van Loan, "Matrix Computations", Ed. Johns Hopkins, 1996

[Gray98] M. Gray, T,. Sejnowski, J. Movellan, A comparison of visual representations for speechreading, *IEEE Pattern Analysis and Machine Intelligence*, (remitido para su publicación), 1998.

[Haykin94*a*] S. Haykin (Editor), "Blind Deconvolution", 1ª Ed., Prentice-Hall, New Jersey, 1994.

[Haykin94*b*] S. Haykin, "Neural Networks: A comprehensive foundation", Prentice-Hall, New Jersey, 1994.

[Herault86] J. Herault y C. Jutten, Space or time adaptive signal processing by neural network models. Capítulo del libro de J. Denker (editor) "Neural Networks for Computing: AIP Conference Proceedings 151", American Institute for Physics, Nueva York, 1986.

[Harroy96] F. Harroy y J. Lacoume, Maximum Likelihood Estimators and Cramer-Rao Bounds in source separation, *Signal Processing*, 55: 166-177, 1996.

[Johnson93] D. H. Johnson, D. E. Dudgeon, "Array Signal Processing", 1ª Ed., Ed. Prentice-Hall, Englewood Cliffs, 1993.

[Hyvärinen97] A. Hyvärinen y E. Oja, A fast fixed-point algorithm for Independent Component Analysis, *Neural Computation*, 9:1483-1492, 1997.

[Hyvärinen98] A. Hyvärinen, New approximations of differential entropy for independent component analysis and projection pursuit, *Advances in Neural Information Processing Systems*, 10: 273-279, MIT Press, 1998.

[Hyvärinen99] A. Hyvärinen, Fast and Robust Fixed-Point Algorithms for Independent Component Analysis, *aceptado para su publicación en IEEE Trans. on Neural Networks*, 1999.

[Jung98] T. Jung, C. Humphries, T. Lee, S. Makeig, M. McKeown, V. Iragui, T. Sejnowski, Extended ICA removes artifacts from EEG data, *Advances in Neural Information Processing Systems*, 10:894-900.

[Jutten91] C. Jutten y J. Herault, Blind Separation of Sources, part I:An Adaptive Algorithm Based on Neuromimetic Structure, *Signal Processing,* 24:1-10, 1991.

[Karhunen94] J. Karhunen y J. Joutsensalo, Representation and Separation of Signals using Nonlinear PCA type learning, *Neural Networks*, 7:113-127.

[Lambert97] R. Lambert y A. Bell, Blind Separation of multiple speakers in a multipath enviroment, *Actas del Congreso ICASSP*, Munich, 1997.

[Macchi97] O. Macchi, E. Moreau, Self-Adaptive Source Separation, Part I: Convergence Analysis of a Direct Linear Network Controled by the Herault-Jutten Algorithm, *IEEE Trans. on Signal Proc.*, 45 (4): 918-926, 1997.

[Makeig96] S. Makeig, A. Bell, T. Jung y T. Sejnowski, Independent Component Analysis of Electroencephalographic Data, *Advances in Neural Information Processing Systems*, 8: 145-151.

[Makeig97] S. Makeig, T. Jung, A. Bell, D. Gharemani y T. Sejnowski, Blind Separation of event-related brain responses into independent components, *Proc. Natl. Acad. Sci. USA*, 94: 10979-10984

[McKeown98] M. McKeown, S. Makeig, G. Brown, T. Jung, S. Kindermann, A. Bell, T. Sejnowski, Analysis of fmri by blind separation into independent spatial components, *Human Brain Mapping*, 6:1-31, 1998.

[Mansour96] A. Mansour y C. Jutten, A direct Solution for Blind Separation of Sources, *IEEE Trans. on Signal Proc.*, 44: 746-748, 1996.

[Mansour99] A. Mansour y C. Jutten, What should we say about the kurtosis?, *IEEE Signal Proc. Letters*, 6: 321-322, 1999.

[Martín97*a*] R. Martín Clemente y J. I. Acha, Blind Separation of Sources using a New Polynomial Equation, *Electronics Letters*, 33(3):176-177, 1997.

[Martín97*b*] R. Martín Clemente y Laura Roa, Técnica no invasiva para la obtención del ECG fetal mediante una transformada ortogonal, Actas del Congreso CASEIB'97, Valencia, España, 1997.

[Martín99*a*] R. Martín Clemente y J. I. Acha, A new Algorithm for the Adaptive Separation of Sources, Actas del Congreso ICA'99, Aussois, Francia, 1999.

[Martín99*b*] R. Martín Clemente y J. I. Acha, Linear equations for the Blind Separation of Sources, *remitido para su publicación.*

[Martín00*a*] R. Martín Clemente y J. I. Acha, Blind Separation of Sources by differentiating the output cumulants, *remitido para su publicación.*

[Martín00*b*] R. Martín Clemente y J. I. Acha, Fast algorithm for Blind Separation of Sources based on the derivatives of the output cumulants, *remitido para su publicación.*

[Mishchenko88] A. Mishchenko y A. Fomenko, "A Course of Differential Geometry and Topology", 1 ª Ed., Ed. MIR, Moscú, 1988.

[Moreau96] E. Moreau, O. Macchi, High-Order Contrasts for Self-Adaptive Source Separation, *International Journal of Adaptive Control and Signal Proc.*, 10: 19-46, 1996.

[Nadal94] J.P. Nadal y N. Parga, Non Linear Neurons in the Low Noise Limit: A Factorial Code Maximizes Information Transfer, *Network*, 5:565-581.

[Nikias93] C. Nikias y A. Petropulu, "Higher-order spectra analysis", Ed. Prentice-Hall, 1993.

[Noble89] B. Noble y J. Daniel, "Álgebra Lineal Aplicada", Ed. Prentice-Hall Hispanoamericana, 3ª Ed., 1989.

[Nguyen95] H-L. Nguyen Thi y C. Jutten, Blind Source Separation for Convolutive Mixtures, *Signal Processing*, 45:209-229, 1995

[Papoulis91] A. Papoulis, "Probability, Random Variables and Stochastic Processes", Ed. Mc.Graw-Hill, Nueva York, 1991.

[Pham96] D.T. Pham, Blind separation of instantaneous mixture of sources via an independent component analysis, *IEEE Trans. Signal Proc.*, 44:2768-2779, 1996.

[Prieto97] A. Prieto, C. Puntonet y B. Prieto, Separation of Sources based on Geometrical Properties, *Signal Processing*, 64(3), 1997.

[Puntonet95] C. Puntonet, A. Prieto, C. Jutten y J.O.M Rodríguez-Álvarez, Separation of Sources: A Geometry-Based Procedure For Reconstruction of N-Valued Signals, *Signal Processing*, 46(3):267-284, 1995

[Shalvi90] O. Shalvi y E. Weinstein, New criteria for blind deconvolution of non minimum phase systems (channels), *IEEE Trans. Inf. Theory*, 36(2): 312-321, 1990

[Söderström89] T. Söderström, P. Stoica, "System Identification", 1ª Ed., Ed. Prentice-Hall, 1989.

[Sorouchyari91] E. Sorouchyari, Blind Separation of Sources, part III: Stability Analysis, *Signal Processing,* 24:21-29, 1991.

[Stoica96] P. Stoica, B. Ottersten, "The evil of superefficiency", *Signal Processing*, 55:133-136, 1996.

[Stuart93] A. Stuart y K. Ord, "Kendall's advanced theory of statistics", Vol. I, 6ª Ed., Ed. Edward Arnold, 1993.

[Taleb99*a*] A. Taleb, C. Jutten, Batch algorithm for Source Separation in Postnonlinear mixtures, Actas del Congreso ICA'99, Aussois, Francia, 1999.

[Taleb99*b*] A. Taleb, C. Jutten, Source Separation in Post-Nonlinear Mixtures, *IEEE Trans. on Signal Proc.*, 47:2807-2820, 1999.

[Te-Won98] Te-Won Lee, "Independent Component Analysis", 1° Ed., Kluwer Academic Publishers, Boston, 1998.

[Te-Won99] Te-Won Lee, M Lewicki, M. Girolami y T. Sejnowski, Blind Source Separation of more sources than mixtures using overcomplete representations, *IEEE Signal Proc. Letters*, 6:87-90, 1999.

[Therrien92] C. Therrien, "Discrete Random Signals and Statistical Signal Processing", Ed. Prentice-Hall, 1992.

[Thomas87] G. Thomas y R. Finney, "Cálculo con Geometría Analítica", 6ª Ed., Addison-Wesley Iberoamericana, México D.F., 1987.

[Tong91] L. Tong, R. Liu, V. Soon y Y. Huang, Indeterminacy and Identifiability of Blind Identification, *IEEE Trans. on Circuits and Systems*, 38(5): 499-508, 1991.

[VanGerven95] S. Van Gerven, D. Van Compernolle, Signal Separation by symmetric adaptive decorrelation: stability, convergence and uniqueness, *IEEE Trans. on Signal Proc.*, 43: 1602-1612, 95

[Veen98] A. Van der Veen, Algebraic methods for deterministic blind beamforming, *Proceedings of the IEEE*, 86 (10): 1987-2008, 1998.

[Vigario96] R. Vigario, A. Hyvaerinen, E. Oja, Ica fixed-point algorithm in extraction of artifacts from EEG, Actas del Congreso *IEEE Nordic Signal Processing Symposium*, Espoo, Finlandia, 1996.

[Zarzoso98] V. Zarzoso, A. Nandi, Generalization of a Maximum-Likelihood Aproach to Blind Source Separation, Actas del Congreso *EUSIPCO*, Rodas, Grecia, 1998.

[Zarzoso99] V. Zarzoso, A. Nandi, Blind Separation of Sources for Virtually any Source Probability Density Function, *IEEE Trans. on Signal Proc*, 47: 2419-2432, 1999.

[Zhang99] L. Zhang, S. Amari, A. Cichocki, Natural gradient approach to blind separation of over- and undercomplete mixtures, Actas del Congreso ICA'99, Aussois, Francia, 1999.