Universidad de Sevilla
Escuela Superior de Ingenieros

Tesis Doctoral

Sobre el Análisis en Componentes Independientes de Imágenes Naturales

Susana Hornillo Mellado

Sevilla 2005

Universidad de Sevilla
Escuela Superior de Ingenieros
Dpto. de Teoría de la Señal y Comunicaciones

Tesis Doctoral

Sobre el Análisis en Componentes Independientes de Imágenes Naturales

Autora:

Susana Hornillo Mellado

Ingeniera de Telecomunicación

Director:

Rubén Martín Clemente

Profesor Titular del Dpto. de Teoría de la Señal y Comunicaciones

Sevilla 2005

Índice general

Índice de figuras

Índice de cuadros

Resumen de la Tesis

En esta Tesis se estudian matemática y experimentalmente, los resultados obtenidos al realizar el análisis en componentes independientes (abreviadamente, **ICA**, del inglés *Independent Component Analysis*) de imágenes naturales. El trabajo publicado en 1995 por Bell y Sejnowski [BellSej95], estableciendo una conexión entre los resultados obtenidos al aplicar ICA a imágenes naturales y el comportamiento de ciertas neuronas de la corteza visual primaria, suscitó un gran interés y motivó la aparición de numerosos artículos en los que, mediante diversos experimentos, se ofrecían distintos matices de esta conexión (por citar algunos ejemplos, [CaywWT04, HyvHH03, vanHat98a]). En esta Tesis se aporta, por primera vez, una prueba matemática que explica por qué se observa este interesante comportamiento cuando ICA es aplicado a imágenes naturales.

Gracias a las investigaciones de David H. Hubel y Torsten N. Wiesel [HubW62, HubW68] sobre el modo en el que la corteza visual analiza la información captada por la retina, y por las cuales recibieron el Premio Nobel de Fisiología o Medicina en 1981, sabemos que la mayoría de las neuronas corticales responden con mayor intensidad en presencia de estímulos visuales consistentes en contornos orientados. Por otro lado, tras analizar los histogramas de las respuestas de estas neuronas, David J. Field concluyó que éstos están caracterizados por una elevada curtosis, lo que podía asociarse con una distribución «dispersa» [Field87, Field94]. Cuando aplicamos ICA a imágenes naturales obtenemos unos resultados que recuerdan, sorprendentemente, a este comportamiento de las neuronas de la corteza visual:

- La mayoría de las «bases ICA» obtenidas contienen «bordes» con distintas orientaciones y localizaciones.

- Las componentes independientes tienen una distribución «dispersa».

Esta similitud entre ICA y el sistema visual, junto con la teoría de Barlow [Barlow61, Barlow89, Barlow01] sobre el proceso de reducción de redundancia que llevan a cabo los distintos sistemas sensoriales del cuerpo humano, ha suscitado mucho interés en cuanto a que sugiere que el sistema visual podría realizar una especie de «análisis en componentes independientes» de la información captada por la retina.

Esta Tesis analiza esta situación desde otra óptica y, dejando a un lado la semejanza existente entre los resultados de ICA y el sistema visual humano, plantea la siguiente cuestión: ¿**por qué** al aplicar ICA a imágenes naturales obtenemos unas «bases ICA» que contienen «bordes» y unas componentes independientes caracterizadas por una distribución «dispersa»?

Para responder a esta pregunta, hay que abandonar necesariamente el modo en el que normalmente se plantea el análisis en componentes independientes. En primer lugar, el objetivo en este caso no es analizar la independencia estadística de las componentes, sino su estructura y las ecuaciones que las determinan. En segundo lugar, en esta Tesis se demuestra que aplicar ICA a una imagen natural es equivalente a realizar un filtrado en dos dimensiones de dicha imagen con los denominados «filtros ICA», rotados convenientemente, y muestreando posteriormente el resultado final en los puntos apropiados. Como se podrá comprobar, este nuevo modo de expresar el análisis en componentes independientes de imágenes naturales es necesario para la correcta interpretación de los resultados obtenidos.

De entre todos los posibles algoritmos ICA, esta Tesis se centra en aquellos basados en maximizar estadísticos de orden superior. Como caso particular, se analizan matemáticamente las componentes independientes que determina el popular algoritmo FastICA [HyvOja97, FastICA], cuando los estadísticos que se maximizan son la curtosis y el coeficiente de asimetría.

A continuación se expone la estructura de este documento, indicando brevemente el contenido de cada capítulo.

El Capítulo 1 está dedicado al análisis en componentes independientes. Se cen-

tra en los métodos que encuentran las componentes independientes maximizando la «no gaussianeidad» de los datos y, más concretamente, en aquellos que miden esa «no gaussianeidad» mediante estadísticos de orden superior.

El Capítulo 2 trata de las semejanzas entre el sistema visual humano y los resultados obtenidos al aplicar ICA a imágenes naturales. En este Capítulo se revisarán las motivaciones que llevaron a relacionar ICA con el proceso de extracción de patrones que tiene lugar en la corteza visual, así como los resultados que muestran el parecido entre ICA y el comportamiento de ciertas neuronas de la corteza visual primaria.

El Capítulo 3 se estudian en detalle los algoritmos ICA que usan como criterio el maximizar estadísticos de orden superior, particularizando para los casos en los que estos estadísticos son coeficiente de asimetría y la curtosis. En primer lugar se tratará la extracción de una única componente independiente para después hacer la extensión a múltiples componentes. Como caso particular, se analizan matemáticamente las componentes independientes que determina el popular algoritmo FastICA[HyvKO01, FastICA]. Finalmente, se discuten los resultados obtenidos al aplicar ICA a una imagen.

En el Capítulo 4 se muestran distintos resultados experimentales que corroboran las conclusiones del Capítulo anterior.

Para finalizar, se expondrán las conclusiones de esta Tesis, así como las líneas futuras de investigación.

Capítulo 1

El Análisis en Componentes Independientes (ICA)

1.1. Definición del Análisis en Componentes Independientes

La técnica de Análisis en Componentes Independientes (abreviadamente conocida como ICA, del inglés *Independent Component Analysis*) ha sido descrita convenientemente en numerosos libros (ver, por ejemplo, [HyvKO01]). Por esta razón, vamos a describir ahora sólo los aspectos que consideramos más relevantes. Básicamente, ICA es una técnica de análisis de datos multivariantes cuya finalidad es «descubrir» componentes *estadísticamente independientes* presentes en dichos datos [CaoLiu96, CichAm02, HyvKO01]. Debido a su gran potencial, esta técnica ha recibido una considerable atención desde que P. Comon la acuñase en [Comon94] pudiéndose decir que, a día de hoy, se aplica en campos tales como el procesado de señal, el tratamiento de imágenes o la ingeniería biomédica, entre otros.

En análisis multivariante, se denomina «variable latente» a todo «concepto supuesto y no observado que sólo puede ser aproximado mediante variables medibles u observables»[1]. Normalmente, las «variables latentes» se construyen como

[1]Por ejemplo, la variable latente «satisfacción del usuario de un modelo de coche» puede ser

combinaciones lineales de las variables observables. Pues bien, ICA asocia a cada conjunto de variables latentes una función llamada «función contraste». Esta función alcanza su valor máximo cuando las variables latentes son *estadísticamente independientes* entre sí. El objetivo de ICA no es otro que el de determinar las variables latentes que maximizan dicha «función contraste».

1.2. Formulación matemática de ICA

Sea $\mathbf{X}$ la matriz $N \times T$ de datos (normalmente, la matriz que contiene T observaciones de un vector aleatorio de dimensiones $N \times 1$). Inspirándonos en `Matlab`, denotaremos la k-ésima columna de $\mathbf{X}$ como $\mathbf{x}_{:k} = [x_{1k}, \ldots, x_{Nk}]^\dagger$ donde el superíndice † denota transposición. De igual manera, la i-ésima fila de la matriz será $\mathbf{x}_{i:} = [x_{i1}, \ldots, x_{iT}]$. Sin ninguna pérdida de generalidad se acepta que el valor medio de cada fila de $\mathbf{X}$ es *cero*:

$$\boxed{\frac{1}{T}\sum_{k=1}^{T} x_{ik} = 0 \qquad \forall i} \tag{1.1}$$

Sea $\mathbf{Y}$ la matriz $N \times T$ de variables latentes. Por definición

$$\mathbf{Y} = \mathbf{B}\,\mathbf{X} \tag{1.2}$$

donde $\mathbf{B}$ es una matriz invertible $N \times N$. Se dice que la i-ésima fila de $\mathbf{Y}$, es decir, $\mathbf{y}_{i:} = [y_{i1}, \ldots, y_{iT}]$, es un vector que contiene T realizaciones de la variable latente $\mathcal{Y}_i$. Obsérvese que, a consecuencia de (1.1), la media muestral de las variables $\mathcal{Y}_i$ es siempre cero. *El objetivo de ICA es en principio el siguiente:* obtener unas variables $\mathcal{Y}_1, \ldots, \mathcal{Y}_N$ «tan estadísticamente independientes entre sí como sea posible». Para ello se define una «función contraste» Φ que mide la dependencia estadística entre las variables latentes (el contraste alcanza su valor máximo cuando las variables son estadísticamente independientes). El contraste se puede expresar de las siguientes formas

$$\Phi(\mathbf{Y}) = \Phi(\mathbf{B}\,\mathbf{X}) = \Phi(\mathbf{B})$$

aproximada mediante una función de las variables medibles «tamaño», «tiempo entre averías», «precio», «potencia», etc.

En la práctica ICA se lleva a cabo «encontrando la matriz $\mathbf{B}$ que maximiza $\Phi(\mathbf{B})$».

Si $\mathbf{A} = \mathbf{B}^{-1}$ se puede escribir

$$\mathbf{X} = \mathbf{A}\,\mathbf{Y} \tag{1.3}$$

Esta ecuación indica que cada columna de $\mathbf{X}$ pertenece al espacio vectorial generado por las columnas de $\mathbf{A}$. Por ello es corriente llamar «bases» a estas últimas. Además, en ICA se suele utilizar la siguiente nomenclatura (heredada del problema de la «Separación Ciega de Fuentes», que describiremos más adelante):

- A la matriz $\mathbf{B}$ se la llama «matriz de separación».
- A la matriz $\mathbf{A}$ se la llama «matriz de mezcla».
- A $\mathcal{Y}_1, \ldots, \mathcal{Y}_N$ también se las llama «fuentes» o «componentes independientes».

1.3. Ambigüedades de ICA

Si analizamos detenidamente el modelo ICA mostrado en (1.2) nos daremos cuenta de que existen ambigüedades:

1. No precisa las **varianzas** de las componentes independientes.
2. No precisa el **orden** de las componentes independientes.

Por convención es suponer que cada fuente tiene varianza unidad:

$$\boxed{\frac{1}{T}\sum_{k=1}^{T} y_{ik}^2 = 1 \qquad \forall i}$$

y se ajusta la matriz $\mathbf{B}$ de tal forma que tenga en cuenta esta restricción.

En cuanto al orden de las fuentes, cualquier matriz de permutación $\mathbf{P}$ y su inversa $\mathbf{P}^{-1}$ pueden ser insertadas en el modelo ICA de la siguiente forma:

$$\mathbf{X} = \mathbf{A}\,\mathbf{P}^{-1}\,\mathbf{P}\,\mathbf{Y} \tag{1.4}$$

La matriz $\mathbf{P}\,\mathbf{Y}$ contiene las componentes independientes originales $\mathcal{Y}_i$ cambiadas de orden (en distintas filas). La indeterminación existe porque es imposible distinguir en la práctica el modelo generativo (1.4) del modelo (1.3).

1.4. Aplicación: el problema de la «Separación Ciega de Fuentes»

La aplicación más popular de ICA se da en el problema llamado de «Separación Ciega de Fuentes», que pasamos a describir: consideremos una situación en la que existen una serie de señales emitidas por algún tipo de generador. Supongamos que disponemos de unos sensores que recogen estas señales y que están colocados en diferentes posiciones, de tal forma que cada uno de estos sensores recibe distintas contribuciones de estas señales. Podría darse el caso, por ejemplo, cuando se registran con un grupo de micrófonos las voces de un grupo de personas hablando simultáneamente, o cuando se recogen las señales electromiográficas emitidas por los músculos mediante varios electrodos colocados en la piel.

Para facilitar la explicación, supongamos que son tres las fuentes y tres las señales observadas (es decir, disponemos de tres sensores), denotadas las primeras por $s_1(t)$, $s_2(t)$ y $s_3(t)$, y las segundas por $x_1(t)$, $x_2(t)$ y $x_3(t)$ para $t = 1, \ldots, T$. Supongamos también que las observaciones $x_i(t)$ pueden obtenerse mediante una suma ponderada de las fuentes $s_i(t)$, donde los coeficientes de ponderación dependen de la distancia entre cada uno de los sensores y las fuentes:

$$\begin{aligned} x_1(t) &= a_{11}\, s_1(t) + a_{12}\, s_2(t) + a_{13}\, s_3(t) \\ x_2(t) &= a_{21}\, s_1(t) + a_{22}\, s_2(t) + a_{23}\, s_3(t) \\ x_3(t) &= a_{31}\, s_1(t) + a_{32}\, s_2(t) + a_{33}\, s_3(t) \end{aligned} \tag{1.5}$$

donde los coeficientes a_{ij} son desconocidos, al igual que las fuentes $s_i(t)$. No es muy arriesgado suponer que los coeficientes a_{ij} son lo suficientemente independientes entre sí como para que la matriz que forman, $\mathbf{A} = [a_{ij}]$, sea invertible. Así, existirá una matriz $\mathbf{B} = \mathbf{A}^{-1}$, de coeficientes b_{ij} que pueden separar las fuentes, tal y como se muestra a continuación:

$$\begin{aligned} s_1(t) &= b_{11}\, x_1(t) + b_{12}\, x_2(t) + b_{13}\, x_3(t) \\ s_2(t) &= b_{21}\, x_1(t) + b_{22}\, x_2(t) + b_{23}\, x_3(t) \\ s_3(t) &= b_{31}\, x_1(t) + b_{32}\, x_2(t) + b_{33}\, x_3(t) \end{aligned} \tag{1.6}$$

El problema que se pretende resolver en «separación de fuentes» es la determinación de la matriz **B** y, a partir de ella, de las señales «fuente». Siguiendo el ejemplo, trataríamos de separar las voces de los distintos locutores que han hablado simultáneamente. Pues bien, Comon [Comon94] ha demostrado que bajo hipótesis muy generales *ICA lleva a cabo la resolución de este problema*. Es decir, se verifica que la matriz **Y** de componentes independientes obtenida al aplicar un método ICA a la matriz **X** de observaciones coincide con la matriz de fuentes **S** salvo, quizás, factores de escala o permutaciones en el orden de las filas. Hoy en día, ésta es la principal aplicación de ICA.

1.5. ICA basado en maximizar la «no gaussianeidad»

Una forma simple e intuitiva de estimar el modelo ICA se basa en maximizar la «no gaussianeidad» de las observaciones. Según el teorema central del límite, la distribución de una suma de variables aleatorias independientes tiende hacia una distribución gaussiana. En otras palabras, la suma de un cierto número de variables aleatorias independientes tiene una distribución que «se parece» más a una distribución gaussiana que la distribución de cualquiera de las variables por separado. Por ejemplo, consideremos dos variables aleatorias independientes con distribución uniforme, de media cero y varianza unidad. En las Figuras 1.2 y 1.3 mostramos la distribución conjunta de dichas variables independientes (obtenida como una muestra de las variables representadas en el plano bidimensional), y la distribución de una de estas variables uniformes (estimada a partir de su histogra-

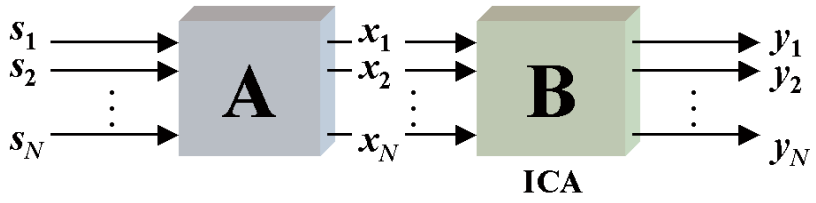

Figura 1.1: Problema del análisis en componentes independientes

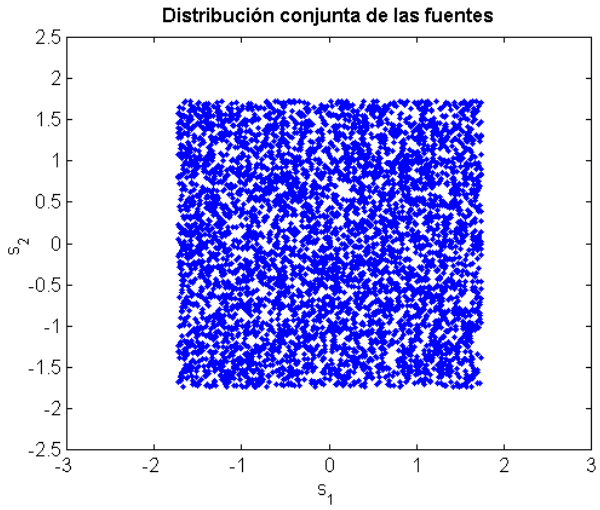

Figura 1.2: Distribución conjunta estimada de dos variables independientes uniformes.

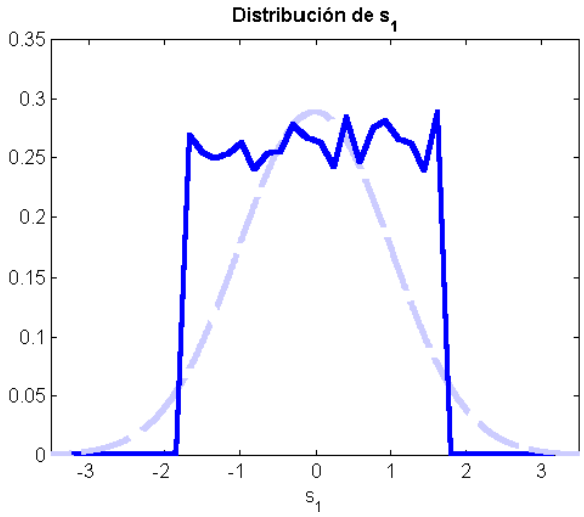

Figura 1.3: Distribución estimada de una de las variables uniformes comparada con una distribución gaussiana (línea discontinua).

ma) comparada con una distribución gaussiana.

La distribución conjunta de una combinación lineal de ambas variables se muestra en la Figura 1.4 (donde la matriz que contienen los coeficientes de ponderación es ortogonal). Si nos fijamos en la Figura 1.5, podemos ver que la nueva distribución tras la suma se acerca a la de una distribución gaussiana, verificando lo expuesto anteriormente.

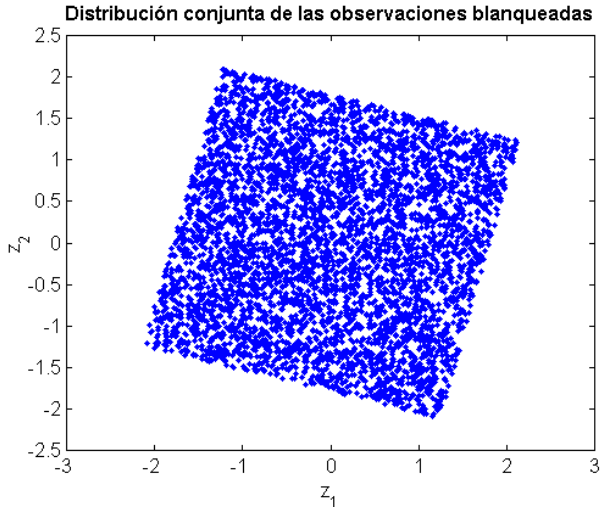

Figura 1.4: Distribución conjunta estimada de la mezcla blanqueada de dos variables aleatorias uniformes independientes

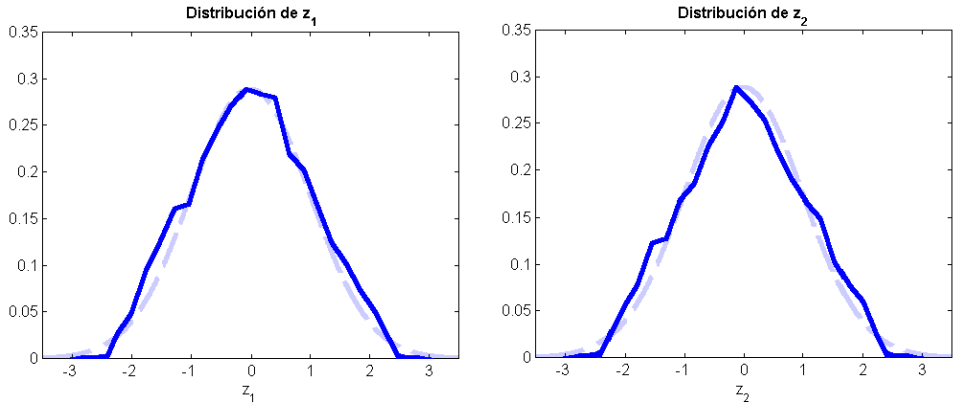

Figura 1.5: Distribuciones marginales estimadas de las variables obtenidas tras la mezcla comparada con una distribución gaussiana (línea discontinua).

¿Cómo se aplica esta idea a ICA? Sean $\mathcal{Y}_1, \ldots, \mathcal{Y}_N$ variables aleatorias estadísticamente independientes *no gaussianas*. Sea

$$\mathbf{X} = \mathbf{A}\,\mathbf{Y} \tag{1.7}$$

donde la fila i-ésima de $\mathbf{Y}$ contiene las realizaciones de la variable $\mathcal{Y}_i$ y $\mathbf{A}$ es una matriz $N \times N$. Gracias al teorema central del límite se puede esperar que la distribución de las filas de $\mathbf{X}$ sea «bastante gaussiana». Por la misma razón, la

distribución de las filas de

$$\mathbf{B}\,\mathbf{X} = \mathbf{B}\,\mathbf{A}\,\mathbf{Y} = \mathbf{G}\,\mathbf{Y}, \tag{1.8}$$

donde $\mathbf{B}$ es una matriz $N \times N$, también debe ser «normal». La excepción a esta regla no es otra que la situación en la que $\mathbf{G}$ es *la matriz identidad* (o una permutación en las filas de la matriz identidad), en cuyo caso, simplemente

$$\mathbf{B}\,\mathbf{X} \equiv \mathbf{Y} \tag{1.9}$$

De todo esto nace la siguiente idea:

> La búsqueda de la matriz $\mathbf{B}$ que produce componentes independientes se puede sustituir por la búsqueda de la matriz $\mathbf{B}$ que produce componentes tan «poco gaussianas» como sea posible.

Este principio reduce ICA a la técnica conocida como «Projection Pursuit» [Fried87], de la que existe abundante bibliografía. La suposición de que las variables $\mathcal{Y}_1, \ldots, \mathcal{Y}_N$ no son *de Gauss*, por otra parte, distingue ICA del llamado «Análisis Factorial» de los datos [HyvKO01].

1.6. Medida de la «no gaussianeidad» mediante estadísticos de alto orden

La búsqueda de componentes «no gaussianas» se suele llevar a cabo maximizando funciones contraste que se definen a partir de estadísticos de las variables latentes, como el *momento central de tercer orden* (o *coeficiente de asimetría*) y la *curtosis* [HyvKO01]. Por ejemplo, la *curtosis*, o cumulante de cuarto orden de una variable aleatoria nos da una medida de la distancia que separa la distribución de dicha variable aleatoria de una distribución gaussiana, que posee kurtosis cero (de hecho, todos los cumulantes de orden mayor que dos de una variable aleatoria gaussiana son nulos [Nandi99]). En [HyvKO01] se demuestra que maximizando la kurtosis podemos obtener las componentes independientes subyacentes en las observaciones $\mathbf{X}$ del modelo ICA.

En esta Tesis consideraremos únicamente las funciones contraste *coeficiente de asimetría* y *curtosis*. Las razones son dos:

- Son las funciones contraste más sencillas de estudiar matemáticamente.
- Prácticamente, cualquier otra función contraste conocida proporciona resultados equivalentes a los obtenidos al maximizar el *coeficiente de asimetría* y la *curtosis* [HyvKO01, Lee98].

Por su especial interés, repasaremos en lo que queda de Sección la definición de la *curtosis*. Para una variable aleatoria $\mathcal{Y}$, cuya media es cero, la curtosis viene dada por:

$$\text{curt} = E\{\mathcal{Y}^4\} - 3E\{\mathcal{Y}^2\}^2 \tag{1.10}$$

Si además imponemos que la variable aleatoria y tenga varianza unidad, curt $= E\{\mathcal{Y}^4\} - 3$. Las variables aleatorias con curtosis negativa reciben el nombre de *subgaussianas* o *platicúrticas*, mientras que las que tienen curtosis positiva reciben el nombre de *supergaussianas* o *leptocúrticas*. Por otro lado, mientras que las variables supergaussianas pueden llegar a tener curtosis infinitamente grandes (en teoría), las subgaussianas tienen acotado el valor mínimo de su curtosis a -2 (en el caso de varianza unidad) [HyvKO01].

Las variables aleatorias supergaussianas suelen presentar una función de densidad de probabilidad picuda, tomando valores grandes en torno a la media y pequeños lejos de ella. Dicho de otra forma, estas variables aleatorias toman valores cercanos a su media con mucha probabilidad, tomando valores alejados de su media en raras ocasiones, por lo que presentan una distribución «dispersa». De hecho, cuanto mayor sea la curtosis de una variable aleatoria, más «dispersa» será su distribución [HyvKO01]. Un ejemplo típico lo tenemos en la distribución de Laplace, cuya función de densidad de probabilidad viene dada por (considerando media cero y varianza unidad):

$$p(y) = \frac{1}{\sqrt{2}} \exp\left(-\sqrt{2}\,|y|\right) \tag{1.11}$$

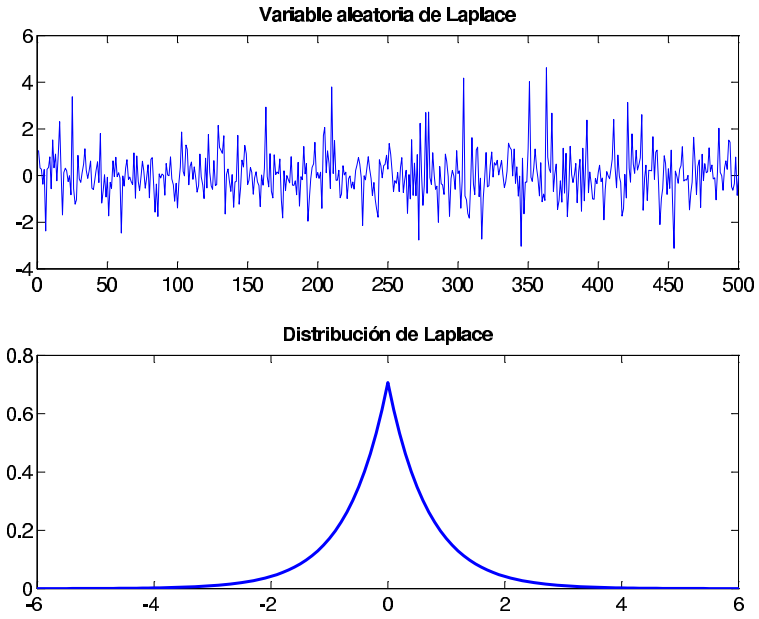

Figura 1.6: Ejemplo de una distribución supergaussiana (curtosis positiva): muestra de una variable aleatoria de Laplace de media cero y varianza unidad (arriba) y su correspondiente función de densidad de probabilidad (abajo).

En la Figura 1.6 mostramos un ejemplo de una variable aleatoria laplaciana, de media cero y varianza unidad y su correspondiente función de densidad de probabilidad. Se puede observar el aspecto «disperso» de dicha variable aleatoria.

En cuanto a las variables aleatorias subgaussianas suelen presentar una función de densidad de probabilidad achatada, prácticamente constante en torno a la media y casi nula lejos de ella. Un ejemplo típico es la variable uniforme, cuya función de densidad de probabilidad viene dada por (considerando media cero y varianza unidad):

$$p(y) = \begin{cases} \frac{1}{\sqrt{3}}\,, & \text{si } |y| \leq \sqrt{3} \\ 0\,, & \text{e.o.c.} \end{cases} \tag{1.12}$$

En la Figura 1.7 mostramos el ejemplo de una variable aleatoria uniforme de media cero y varianza unidad, así como su correspondiente función de densidad de probabilidad.

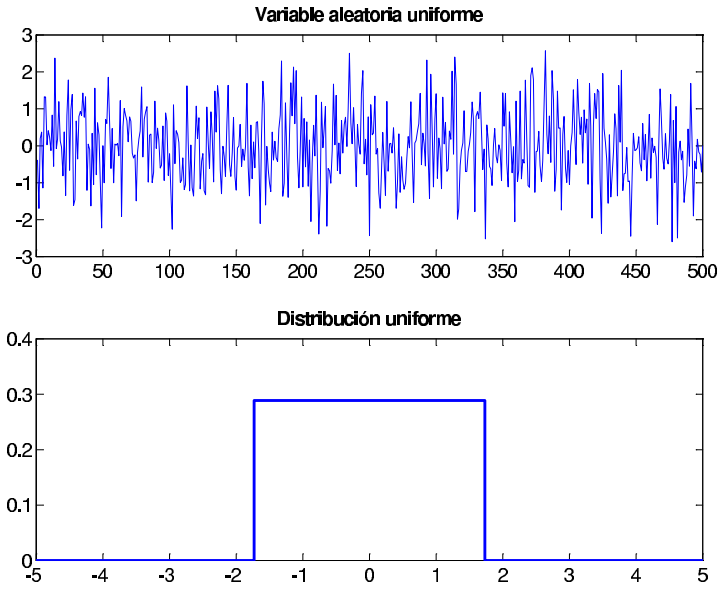

Figura 1.7: Ejemplo de una distribución subgaussiana (curtosis negativa): muestra de una variable aleatoria uniforme de media cero y varianza unidad (arriba) y su correspondiente función de densidad de probabilidad (abajo).

1.7. Análisis en componentes principales (PCA)

El análisis en componentes principales (abreviadamente, PCA, del inglés *Principal Component Analysis*) es una técnica clásica usada en el análisis estadístico de datos, extracción de características y compresión de datos. Dado un conjunto de variables, el propósito es encontrar otro conjunto más pequeño de variables, con menos redundancia, que nos permita representar al primero lo mejor que sea posible. Este objetivo está muy ligado al de ICA, pero en PCA la redundancia se mide en función de la correlación entre las variables, mientras que en ICA se tienen en cuenta dependencias estadísticas de mayor orden. Por su conexión con ICA y porque haremos mención a esta técnica matemática en numerosas ocasiones, dedicamos esta sección a ofrecer unas nociones básicas sobre PCA (para más información, consúltese [GonzWo92, HyvKO01]).

Dado un conjunto de variables de media cero, $\mathcal{X}_1, \mathcal{X}_2, ..., \mathcal{X}_N$, mediante PCA obtenemos otro conjunto de variables no correlacionadas entre sí, $\mathcal{X}_1^{\perp}, \mathcal{X}_2^{\perp}, ..., \mathcal{X}_N^{\perp}$. Esta transformación se consigue mediante la rotación del sistema de coordenadas

usado para la representación de las variables, de tal forma que el nuevo sistema de coordenadas vendrá determinado por los *autovectores* de la matriz de correlación de $\mathcal{X}_1, \mathcal{X}_2, ..., \mathcal{X}_N$, que además identifican las direcciones de máxima varianza [GonzWo92, HyvKO01]. En la práctica, la matriz $\mathbf{X}^{\perp}$ de dimensiones $N \times T$ que contiene las nuevas variables viene dada por:

$$\mathbf{X}^{\perp} = \mathbf{W}\,\mathbf{X} = \mathbf{D}^{-1/2}\,\mathbf{V}^{\dagger}\,\mathbf{X} \tag{1.13}$$

donde $\mathbf{V} = [\mathbf{v}_1|...|\mathbf{v}_N]$ es la matriz ortogonal[2] que contiene, por columnas, los autovectores de la matriz de correlación de los datos $\mathbf{R}_x = \frac{1}{T}\mathbf{X}\,\mathbf{X}^{\dagger}$, y $\mathbf{D} = \text{diag}(\lambda_1, ..., \lambda_N)$ es la matriz diagonal de sus autovalores. Normalmente, los autovalores se ordenan de mayor a menor en la diagonal de la matriz $\mathbf{D}$, al igual que los autovectores de la matriz $\mathbf{V}$ (esto es, la matriz $\mathbf{V}$ estaría ordenada de tal forma que su primera columna sería el autovector asociado al mayor autovalor y la última columna sería el autovector asociado al menor autovalor). Mientras que la matriz de autovectores es la responsable de la rotación del sistema de coordenadas, la matriz de autovalores, concretamente $\mathbf{D}^{-1/2} = \text{diag}(\lambda_1^{-1/2}, ..., \lambda_n^{-1/2})$, hace que las varianzas de las variables resultantes estén normalizadas a la unidad. Este proceso también es conocido como «blanqueado» o «whitening» debido a que las variables $\mathcal{X}_1^{\perp}, \mathcal{X}_2^{\perp}, ..., \mathcal{X}_N^{\perp}$ son «espacialmente blancas», es decir, su matriz de correlación es la matriz identidad. Por ello, a la matriz $\mathbf{W} \stackrel{def}{=} \mathbf{D}^{-1/2}\,\mathbf{V}^{\dagger}$ se la llama «matriz de blanqueado».

[2] Es decir, se verifica que $\mathbf{V}\,\mathbf{V}^{\dagger} = \mathbf{V}^{\dagger}\,\mathbf{V} = \mathbf{I}$, siendo $\mathbf{I}$ la matriz unidad.

Capítulo 2

Conexión entre ICA y el sistema visual humano

En los últimos años ha suscitado un gran interés el hecho de que gran parte de las neuronas de la corteza visual parecen responder específicamente a «líneas» o «bordes» con distintas orientaciones. Se han propuesto diversas técnicas matemáticas que imitan, con mayor o menor éxito, esta especie de proceso de extracción de patrones que realiza nuestro sistema visual. Entre estas técnicas se encuentra el análisis en componentes independientes (ICA).

En este Capítulo analizaremos la conexión entre ICA y el sistema visual humano. Para ello, previamente debemos ofrecer algunas nociones de cómo el sistema visual procesa las señales luminosas captadas por la retina. Nos basaremos en los estudios que sobre el sistema visual realizaron David H. Hubel y Torsten N. Wiesel, galardonados con el Premio Nobel de Fisiología o Medicina en 1981[1].

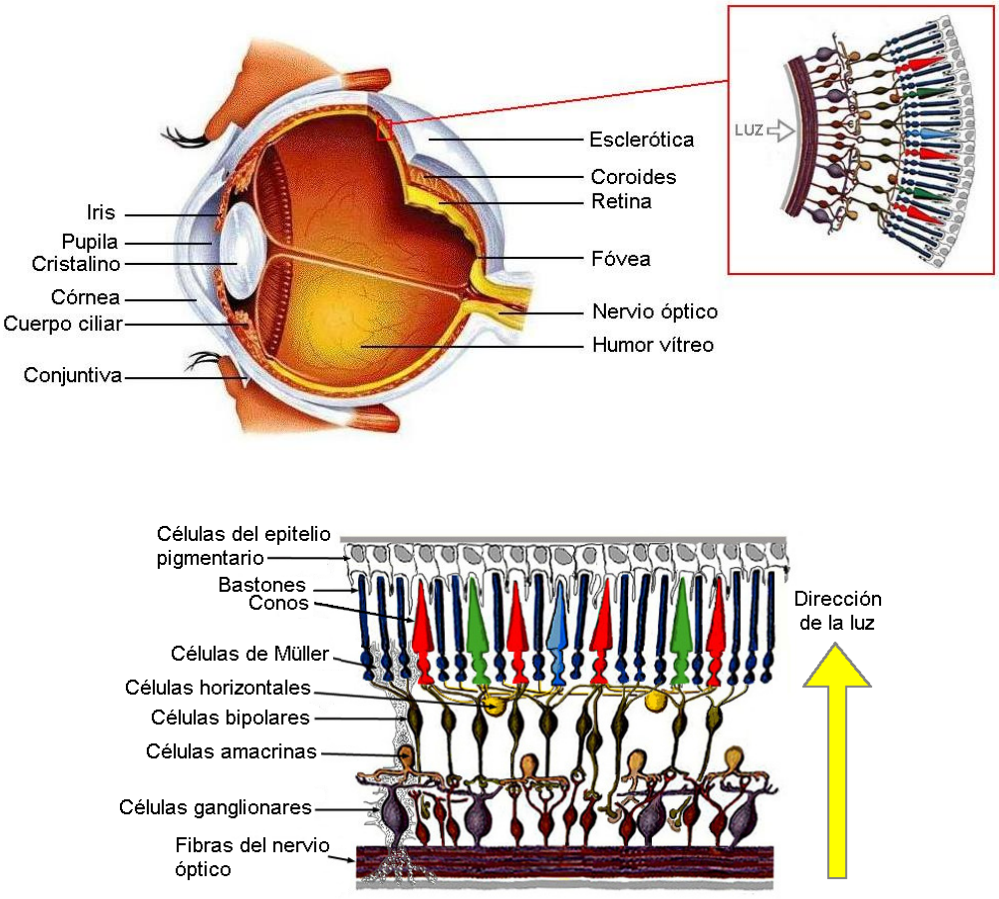

Figura 2.1: Anatomía del ojo humano (vista lateral, sección vertical) y estructura de la retina.

2.1. Procesado de las señales en el sistema visual central

El ojo es un órgano especializado en la recepción de la luz. Está formado por tres capas: *retina*, *esclerótica* y *úvea*, la última de las cuales está compuesta a su vez por la *coroides*, el *iris* y el *cuerpo ciliar* (Figura 2.1). Las células especializadas de la *retina* son las que hacen posible que la energía luminosa se transforme en potencial nervioso. El resto de estructuras que forman el ojo (en las que no profundizaremos) sostienen la retina o sirven para enfocar las imágenes que le llegan del exterior (más

[1] Recientemente se han publicado nuevos trabajos en los que se matizan las conclusiones de Hubel y Wiesel.

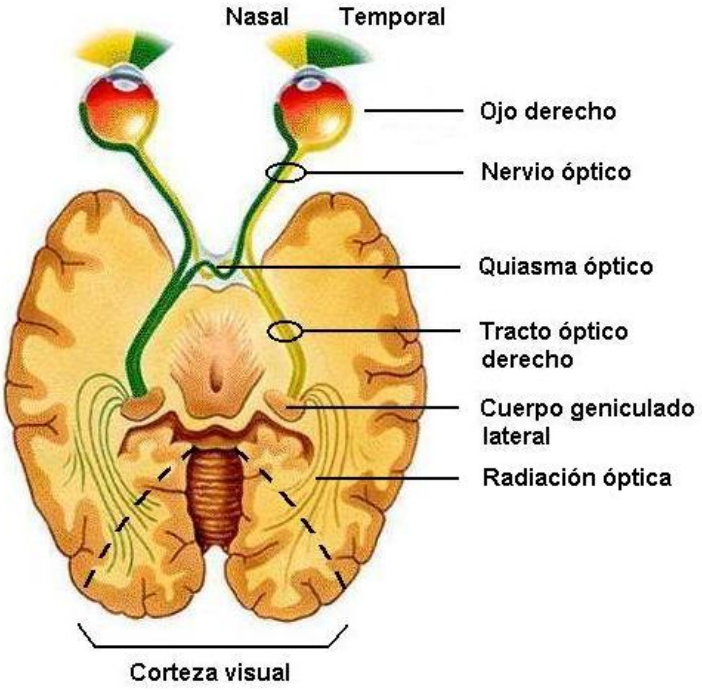

Figura 2.2: Esquema de la vía óptica del cerebro humano.

información en [GuyHal00, SchmTh93] o en cualquier otro texto sobre fisiología humana).

La retina es un área fotosensible, es decir, que experimenta una reacción específica a las radiaciones luminosas. Es la capa más interna del ojo y está organizada en estratos caracterizados por distintos tipos de células neuronales (Figura 2.1). En la capa más externa se encuentran unas células con capacidad fotorreceptora, que son de dos tipos: *conos* (responsables de la visión en color) y *bastones* (responsables de la visión en la oscuridad). La información luminosa captada por conos y bastones se propaga por las sucesivas capas de células neuronales hasta alcanzar el *nervio óptico* que transmite la información a la *corteza visual*, en la parte posterior del cerebro (Figura 2.2).

Los nervios ópticos de ambos ojos se unen en la base del cráneo en el *quiasma óptico*, en el que las fibras nerviosas procedentes de la mitad *nasal* de la retina cruzan al lado derecho y las procedentes de la mitad *temporal* cruzan al lado izquierdo (Figura 2.2), componiendo, respectivamente, los *tractos ópticos derecho*

e izquierdo. Cada uno de estos tractos ópticos desembocan en su respectivo *cuerpo geniculado lateral*[2]. La mayoría de los axones[3] de las neuronas del geniculado se dirigen a través de la *radiación óptica* a las neuronas de la *corteza visual primaria* o V1, también llamada *corteza estriada* debido a su aspecto «rayado», y que es la primera y más importante zona de toda una jerarquía de áreas de la corteza visual (el lector puede encontrar más información sobre cómo se relacionan las neuronas del cuerpo geniculado lateral y las de la corteza visual primaria en). Las zonas de la corteza visual externas a la corteza visual primaria reciben el nombre de *corteza visual extraestriada* [GuyHal00, SchmTh93].

2.1.1. Campos receptivos de las neuronas del sistema visual

Consideremos una célula neuronal de la retina, por ejemplo, una célula ganglionar. A través de las células horizontales, bipolares y amacrinas, esta célula ganglionar recibe información de un conjunto de fotorreceptores localizados en una determinada región de la retina (Figura 2.1). Esta región sería el *campo receptivo* de la célula ganglionar. Para cualquier otra neurona del sistema visual, ya sea de la retina, del cuerpo geniculado lateral o de la corteza visual, existe un conjunto de fotorreceptores (o lo que es igual, una determinada área de capa más externa de la retina) cuya estimulación provoca un cambio en la respuesta de dicha neurona, y que sería, por tanto, su campo receptivo.

La forma de caracterizar un campo receptivo es mediante la colocación de electrodos en las neuronas correspondientes y la medición de las respuestas eléctricas ante estímulos luminosos con diferentes formas, tamaños, orientaciones, etc. Un determinado estímulo puede incrementar la respuesta de la neurona (excitación) mientras que otro puede hacer que decrezca (inhibición). De esta forma se puede configurar una especie de «mapa» del campo receptivo. Para profundizar más en este tema, consúltense las referencias [HubW62, HubW68].

Así, por ejemplo, muchas de las células ganglionares de la retina y las neuronas

[2] Llamado así por su forma acodada.

[3] El axón es la fibra nerviosa que permite la salida del impulso nervioso fuera del núcleo de la neurona.

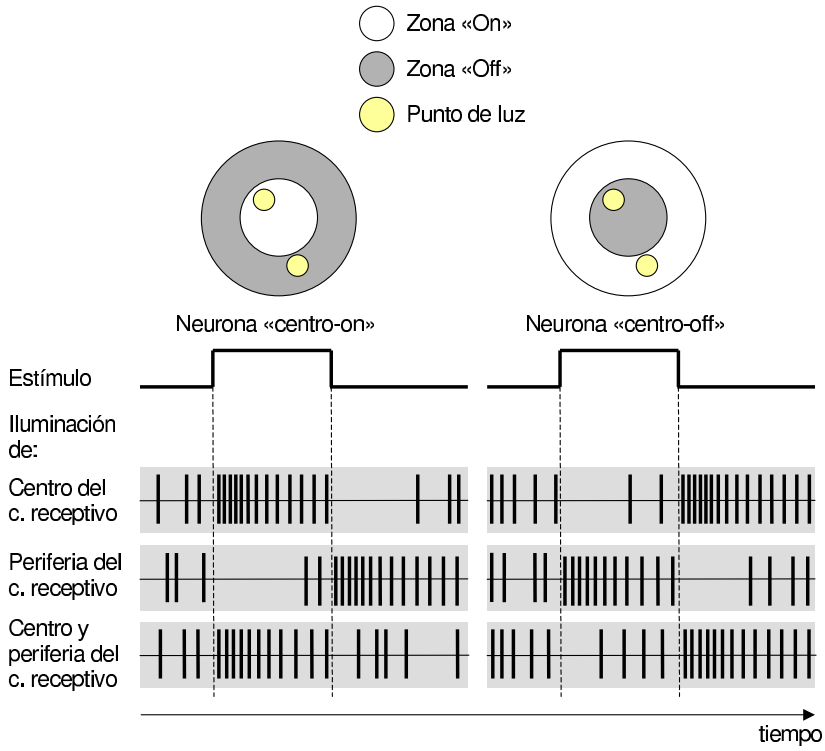

Figura 2.3: Respuesta de una neurona con campo receptivo de estructura concéntrica y organización antagonista: la neurona se excita o inhibe en función de la zona del campo receptivo que se ilumine. Duración del estímulo = 1 seg.

del cuerpo geniculado lateral tienen campos receptivos concéntricos y organizados de forma *antagonista*: para una determinada neurona, el mismo estímulo visual produce en distintas zonas de su campo receptivo reacciones contrarias (este tipo de comportamiento es compartido por muchas de las neuronas del sistema visual). En la Figura 2.3 mostramos la organización funcional de los campos receptivos de dos tipos de células ganglionares de la retina. Para su análisis se proyectan puntos de luz en el centro o la periferia del campo receptivo (CR). Las células ganglionares con «centro-on» responden a la iluminación del centro del CR. La iluminación de la periferia del CR produce una inhibición pasajera de la actividad neuronal. Si se iluminan simultáneamente el centro y la periferia del CR, domina la respuesta del centro, aunque su activación es más débil que cuando se ilumina solo. Las células ganglionares con «centro-off» tienen un comportamiento contrario a las anteriores.

En la corteza visual primaria encontramos básicamente dos tipos de células

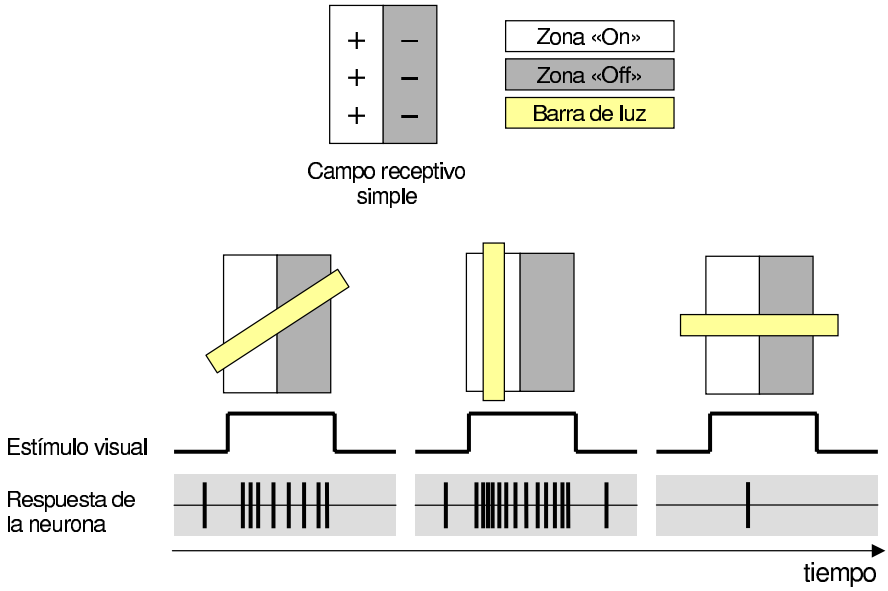

Figura 2.4: Respuesta de una neurona con campo receptivo simple orientado verticalmente. Duración del estímulo = 1 seg.

neuronales, caracterizadas por la estructura de sus campos receptivos [HubW62, HubW68]:

- **Células simples:** Son las más abundantes y están caracterizadas por campos receptivos denominados *simples*, que principalmente responden a contornos paralelos claro-oscuros con una determinada orientación[4].

- **Células complejas:** Caracterizadas por campos receptivos *complejos*, que responden a contornos claro-oscuros con una determinada orientación y extensión espacial, interrupciones en contornos y esquinas. Estas neuronas se activan mucho más fuertemente por patrones de estímulos móviles o por un cambio de patrón de estímulo.

En la Figura 2.4 ilustramos el caso de una célula simple cuyo campo receptivo consiste en dos líneas paralelas verticales, una de tipo «on» o excitatoria y otra de tipo «off» o inhibitoria. En este caso también observamos un comportamiento

[4] En la corteza visual primaria también existen neuronas simples cuyos campos receptivos están organizados concéntricamente, como en el cuerpo geniculado lateral, aunque son mucho menos numerosas [SchmTh93].

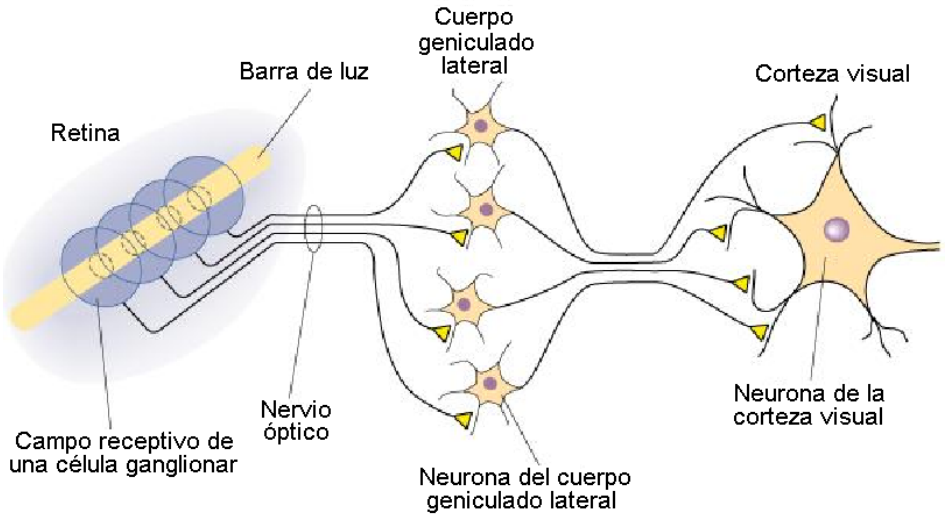

Figura 2.5: Organización retinotópica del sistema visual: la información procedente de células neuronales adyacentes de la retina es procesada por células también adyacentes del cuerpo geniculado lateral y de la corteza visual.

antagonista. Si se proyecta una barra de luz con la orientación y posición «adecuadas» al campo receptivo, este estímulo provoca una fuerte activación neuronal. Si la barra está orientada perpendicularmente con respecto a la «dirección óptima», por regla general la neurona no responde.

Las regiones la corteza visual extraestriada, externas a la corteza visual primaria, están especializadas en el procesamiento de otro tipo de cualidades de la visión. Por ejemplo, el área V3 tiene una marcada sensibilidad al movimiento, mientras que el área V4 tiene campos receptivos encargados de la percepción del color [SchmTh93].

2.1.2. Arquitectura de la corteza visual

Gran parte del sistema visual central está caracterizado por una «organización retinotópica», de tal forma que la información proveniente de puntos adyacentes de la imagen de la retina es procesada por neuronas también adyacentes (Figura 2.5). Sin embargo, esta proyección retinotópica de la retina sobre el sistema visual central *no es lineal*: la región de la *fóvea central* (pequeña depresión en el centro de la retina, Figura 2.1) se proyecta sobre una región mucho mayor de la corteza

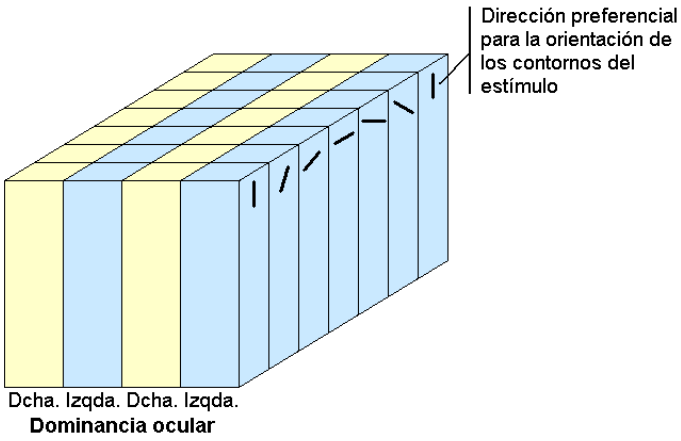

Figura 2.6: Esquema de la organización vertical de la corteza visual.

visual que, por ejemplo, un área de igual superficie de la periferia de la retina. Esto es debido a que la densidad de fotorreceptores va decreciendo desde la fóvea central a la periferia de la retina [GuyHal00, SchmTh93].

La proyección retinotópica de la retina sobre la corteza visual está caracterizada por una orientación horizontal, es decir, paralela a la superficie del cerebro. Adicionalmente, según los estudios de Hubel y Wiesel [HubW62, HubW68], existe un segundo principio organizativo vertical, orientado perpendicularmente a la superficie del cerebro. Según este principio, las neuronas de la corteza visual se agrupan en columnas denominadas *columnas corticales*, las cuales son consideradas como la unidad mínima de procesamiento de la corteza y están caracterizadas por las siguientes propiedades:

- Los campos receptivos de las neuronas dentro de una columna cortical están en la misma región de la retina.

- Las neuronas de una misma columna son sensibles a los estímulos luminosos recibidos de un mismo ojo, alternándose de tal forma que si una columna se activa por el ojo derecho, la adyacente se activa por el ojo izquierdo.

- Los campos receptivos de las neuronas de cada columna responden a orientaciones muy definidas de los contornos del estímulo.

La Figura 2.6 muestra un esquema que ayuda a comprender mejor esta organización en columnas de la corteza visual.

2.2. El sistema visual humano y el reconocimiento de patrones

Según Barlow [Barlow61, Barlow89] la percepción que tenemos de nuestro entorno viene condicionada por la *redundancia* existente en la información percibida. Esta redundancia consistiría en aquellos patrones repetitivos o regularidades que distinguen un determinado estímulo de algo aleatorio, y su reconocimiento es lo que permite al cerebro elaborar una especie de modelo o «mapa» del mundo en que vivimos. Por otro lado, Barlow propone que el cerebro llevaría a cabo un proceso de *reducción de redundancia* con el objetivo de disminuir al máximo la cantidad de información a procesar. Para Barlow, la redundancia vendría caracterizada por una dependencia estadística entre estímulos, y su reducción consistiría, primero en minimizar esta dependencia estadística y, segundo, en la combinación de estímulos fuertemente dependientes (lo que puede indicar que transportan la misma información o muy parecida) en uno sólo [Barlow89].

Estas ideas podrían aplicarse al proceso de extracción de patrones llevado a cabo por el sistema visual, descrito en la Sección anterior. Si el objetivo de todo sistema sensorial es detectar la redundancia (dependencia estadística) que caracteriza un determinado estímulo y luego «aprovecharla» para reducir la cantidad de información a procesar, entonces las características de los campos receptivos del sistema visual (el hecho de que la mayoría respondan a contornos con una determinada orientación, por ejemplo) deben ser tales que permitan detectar esa redundancia. La cuestión que surge entonces es la siguiente: si pudiésemos caracterizar la redundancia existente en las imágenes que vemos, ¿entenderíamos mejor este proceso de extracción de patrones?. El primer paso es plantear un modelo que permita describir cualquier imagen. Lo más evidente es considerar que las imágenes observadas están compuestas por una combinación de ciertos patrones o *funciones*

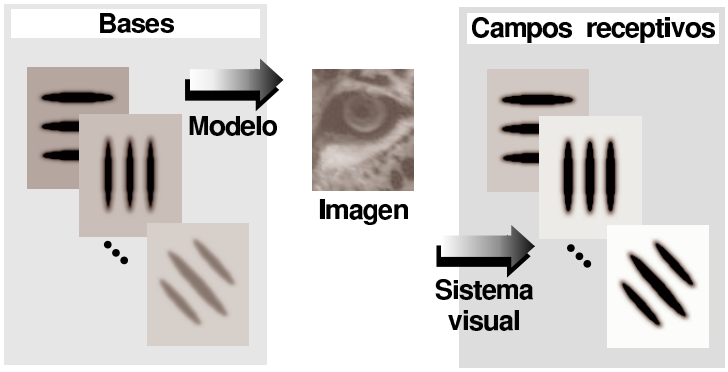

Figura 2.7: Modelo generativo de una imagen: suponemos que cada imagen puede ser obtenida a partir de ciertas funciones base que guardan relación con los campos receptivos de las neuronas del sistema visual.

base que deben parecerse a los patrones extraídos por el sistema visual (ver Figura 2.7). Además, como el análisis que se lleva a cabo de los estímulos visuales es local[5] (el estado de cada neurona depende de la respuesta de un conjunto limitado de fotorreceptores, o lo que es lo mismo, de una pequeña porción del campo visual), nuestro modelo también deberá referirse a pequeñas porciones de la imagen. Lo más simple es suponer que la combinación a la que nos hemos referido es *lineal*, de tal forma que, dado un bloque $I(x, y)$ de una imagen, lo expresaremos de la siguiente forma:

$$I(x, y) = \sum_i s_i \, a_i(x, y) \tag{2.1}$$

donde (x, y) hacen referencia a las coordenadas espaciales, $a_i(x, y)$ son las funciones base y $s_i \in \mathbb{R}$ son unos coeficientes que indican en qué medida está presente cada base en el trozo de la imagen que estamos considerando. El objetivo es encontrar funciones base que, verificando este modelo para todas las posibles «$I(x, y)$», se asemejen lo más posible a los campos receptivos de las neuronas del sistema visual. En la práctica para caracterizar el modelo (2.1) se emplea un conjunto de T porciones o *bloques*, todos de igual tamaño, tomados de un gran número de *imágenes*

[5] Fueron el médico e histólogo español Santiago Ramón y Cajal (1852-1934), galardonado con el Premio Nobel de Fisiología o Medicina en 1906, y, más tarde, su discípulo Rafael Lorente de Nó (1902-), quienes pusieron de manifiesto que las operaciones que realiza la corteza sobre la información que recibe son locales.

naturales[6], digitalizadas y representadas en *escala de grises*[7]. Usamos imágenes naturales para nuestro propósito porque el objetivo que perseguimos es modelar ciertos aspectos del sistema visual humano, y parece lógico suponer que éste ha evolucionado en función de escenas naturales y no generadas de forma artificial.

Estos T bloques, que consideraremos de tamaño $\sqrt{N} \times \sqrt{N}$ píxeles, son reorganizados en vectores columna de tamaño $N \times 1$, y se usan para componer una matriz $\mathbf{X}$, dando lugar a la versión matricial del modelo (2.1):

$$\mathbf{X} = \mathbf{A}\,\mathbf{S} \tag{2.2}$$

con $\mathbf{X} = [\mathbf{x}_1|\mathbf{x}_2|...|\mathbf{x}_T]^\dagger$ († indica «transpuesta»), donde $\mathbf{x}_k$ hace referencia al k-ésimo bloque de los T usados. De la misma forma, la matriz $\mathbf{A} = [\mathbf{a}_1|\mathbf{a}_2|...|\mathbf{a}_N]^\dagger$ contiene por columnas las bases que previamente, igual que los bloques, han sido organizadas en vectores columna. Por su parte, cada elemento s_{ij} de la matriz $\mathbf{S}$ nos da una medida de la contribución de la base $\mathbf{a}_i$ en la generación del bloque $\mathbf{x}_j$, con $i = 1, 2, ..., N$ y $j = 1, 2, ..., T$.

2.3. Las componentes principales de las imágenes naturales

En su trabajo [Barlow61], Barlow ya apuntó que había indicios para suponer que el proceso de reducción de redundancia llevado a cabo por los distintos sistemas sensoriales podría estar basado en la *correlación* entre los estímulos de entrada. Sugirió que este proceso podría asemejarse al *análisis en componentes principales*, una técnica matemática muy usada para el análisis de datos y señales, y que hemos

[6] Una imagen natural posee características como reflejos, transparencias, sombras, transiciones suaves, brillo y color no uniformes, etc., que la diferencia de una imagen artificial, caracterizada por transiciones bruscas, superficies homogéneas, etc. Algunos autores restringen aún más el conjunto de imágenes naturales a aquellas en las que no aparece ninguna estructura construida por el hombre, como edificios, carreteras, etc. (véanse, por ejemplo, [BellSej95, Field94, HyvHH03]).

[7] Como hemos comentado en la Sección anterior, la percepción del color se lleva a cabo en una de las últimas fases del proceso de análisis de la información visual, por lo que, en una primera aproximación, se suele trabajar con imágenes representadas en escala de grises.

revisado en el Capítulo 1. Es lógico, por tanto, que el primer intento relevante de caracterizar el modelo (2.2) estuviese basado en esta técnica [HancBS92].

Mediante PCA obtenemos un conjunto de bases, algunas de las cuales recuerdan a los campos receptivos de las neuronas simples de la corteza visual primaria. A continuación mostramos un ejemplo. Tomamos un total de 18 240 bloques de 12×12 píxeles, extraídos aleatoriamente de un conjunto de imágenes naturales representadas en escala de grises (tomadas de la colección de imágenes naturales [NatCollec]), y componemos la matriz $\mathbf{X}$ tal y como se explicó en la Sección anterior. Consideramos las columnas de $\mathbf{X}$ como muestras de distintas realizaciones, filas de $\mathbf{X}$, de un mismo proceso, y estimamos la matriz de covarianzas, $\mathbf{R_x}$. Las *bases PCA* en este caso serían las columnas de la matriz $\mathbf{W} = \mathbf{D}^{-1/2}\,\mathbf{V}^{\dagger}$, donde $\mathbf{D}$ y $\mathbf{V}$ son, respectivamente, las matrices de los *autovalores* y los *autovectores* de $\mathbf{R_x}$. Reorganizando en matrices y representando como imágenes estas bases, tendrían el aspecto mostrado en la Figura 2.8. Como podemos observar, algunas de las primeras bases, correspondientes a los autovectores asociados con los autovalores más grandes, recuerdan a los patrones asociados a los campos receptivos de las neuronas simples de la corteza visual primaria, debido a su estructura orientada y organizada en zonas claro-oscuras. Sin embargo, la mayoría de estas bases no muestran un patrón característico, asemejándose a un ruido aleatorio. Por otro lado, esa especie de disminución en la escala que presentan las bases (la magnitud o el tamaño de las «formas» que aparecen va decreciendo progresivamente) no concuerda con ninguna de las características de los campos receptivos de las neuronas del sistema visual.

2.3.1. Las componentes independientes de imágenes naturales

Barlow también planteó la posibilidad de que los distintos sistemas sensoriales, entre ellos, el sistema visual, analizaran en cierta forma las dependencias estadísticas de alto orden (mayor que dos, que sería el caso de la correlación) para llevar a cabo el proceso de reducción de redundancia [Barlow61]. Esta idea fue la que

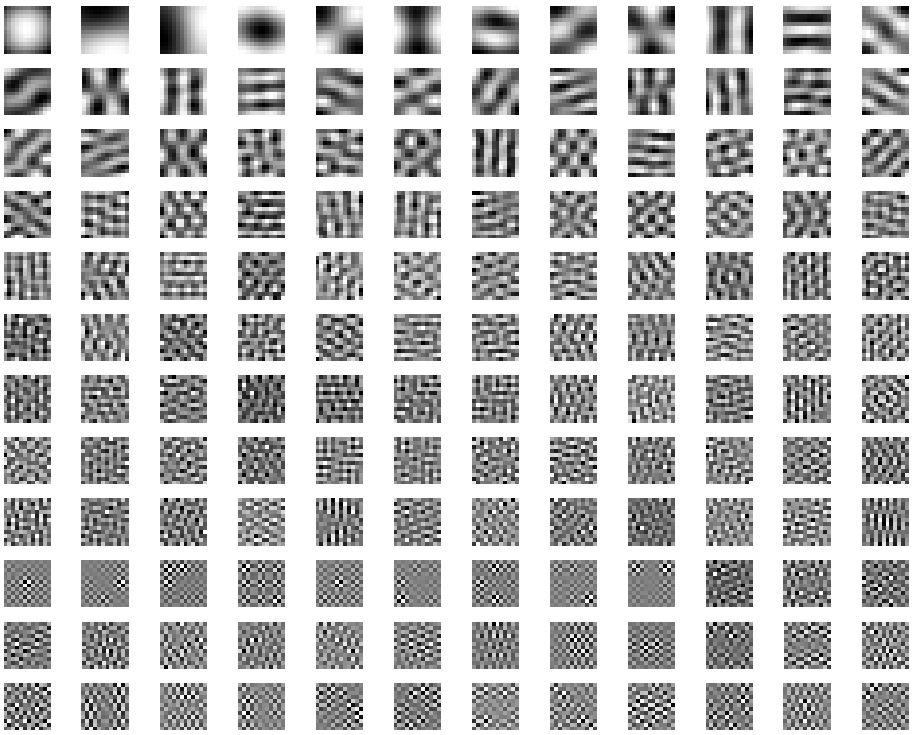

Figura 2.8: Bases PCA obtenidas a partir de 18 240 bloques de tamaño 12×12 extraídos aleatoriamente de un conjunto de imágenes naturales representadas en escala de grises (tomadas de la colección de imágenes naturales [NatCollec]).

motivó otro tipo de soluciones para la caracterización del modelo (2.2), basado en el *análisis en componentes independientes* (ICA), la primera de las cuales fue propuesta por Bell y Sejnowski en [BellSej95].

Mediante su conocido algoritmo *infomax* [BelSej95b] (disponible en [Infomax][8]), obtuvieron un conjunto de bases (columnas de la matriz $\mathbf{A}$), a las que denominaron *bases ICA*, cuyas características recordaban a las de los campos receptivos de las neuronas simples del sistema visual humano.

En la Figura 2.9 mostramos las bases ICA obtenidas en un experimento realizado con 18 240 bloques de 12×12 píxeles cada uno, tomados de la colección de imágenes naturales representadas en escala de grises disponible en [NatCollec]. Los parámetros de entrada del algoritmo *infomax* fueron: *número de pasadas* (*sweep*)

[8] Posteriormente, se publicó una nueva versión de este algoritmo [LeeGS99], denominado *infomax extendido*, que puede encontrarse en [InfomaxE].

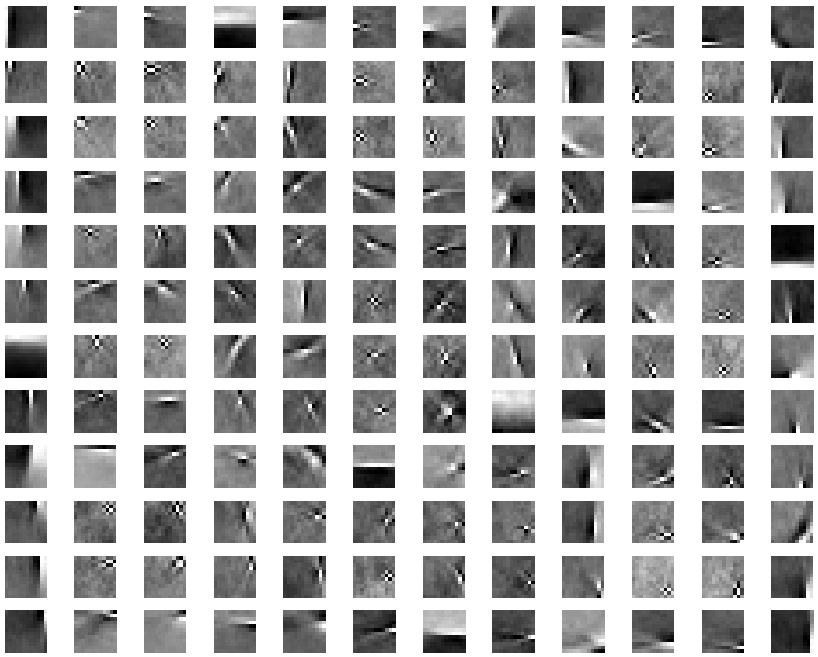

Figura 2.9: Bases (columnas de la matriz **A**) obtenidas aplicando el algoritmo *infomax* a 18 240 bloques de tamaño 12×12 extraídos de un conjunto de imágenes naturales representadas en escala de grises (tomadas de la colección de imágenes naturales [NatCollec]). La estructura localizada y orientada de la mayoría de estas bases recuerda a los campos receptivos simples de las neuronas de la corteza visual primaria.

$= 50$, *tamaño de bloque* $(B) = 30$ y *tasa de aprendizaje* $(L) = 0.0001$ (consúltense [BelSej95b, Infomax] para más información sobre estos parámetros).

Por otra parte, la distribución de los coeficientes de la matriz **S** resultó ser «dispersa», caracterizada por una elevada curtosis. En la Figura 2.11 y mostramos las componentes independientes y en la Figura 2.10 comparamos las curtosis de las observaciones con las de las componentes independientes obtenidas en el experimento anterior. Esta propiedad encuentra conexión, una vez más, con el sistema visual humano. En efecto, según los estudios de Field [Field87, Field94], existen evidencias para afirmar que la distribución de las respuestas de las neuronas de la corteza visual es «dispersa». En este sentido, resulta interesante la propuesta de Olshausen y Field [OlshF96a, OlshF96b]: maximizando la dispersión de la distribución de los coeficientes s_i, modelo (2.1), obtuvieron unos resultados prácticamente idénticos a los de Bell y Sejnowski (algoritmo disponible en [Sparsenet]).

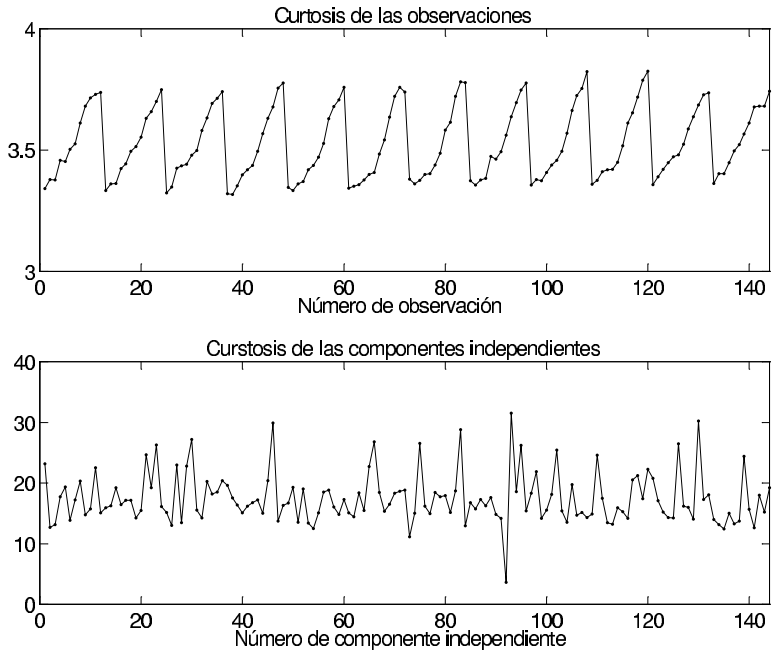

Figura 2.10: Comparativa entre las curtosis de las observaciones (de media cero) y las de las componentes independientes obtenidas al aplicar el algoritmo *infomax* a 18 240 bloques de tamaño 12×12 extraídos de un conjunto de imágenes naturales representadas en escala de grises (tomadas de la colección de imágenes naturales [NatCollec]).

Resultados similares a los de Bell y Sejnowski son obtenidos con otro conocido algoritmo ICA, el FastICA [FastICA, HyvOja97], cuyo criterio para encontrar las componentes independientes se basa en maximizar estadísticos de orden superior como la curtosis [HyvHH03, HyvKO01, vanHat98a, vanHat98b]. En [HyvHH03] se propone además un modelo de organización espacial de las bases ICA que trata de imitar la organización retinotópica de la corteza visual (sección 2.1.2, página 19). Recientemente se han publicado otros trabajos interesantes, como [CaywWT04], en el que se analizan las similitudes entre el sistema visual humano e ICA en relación al procesado del color, o [HyvGH05], en el que se establece una conexión entre ICA y ciertos aspectos de la corteza visual secundaria (V2).

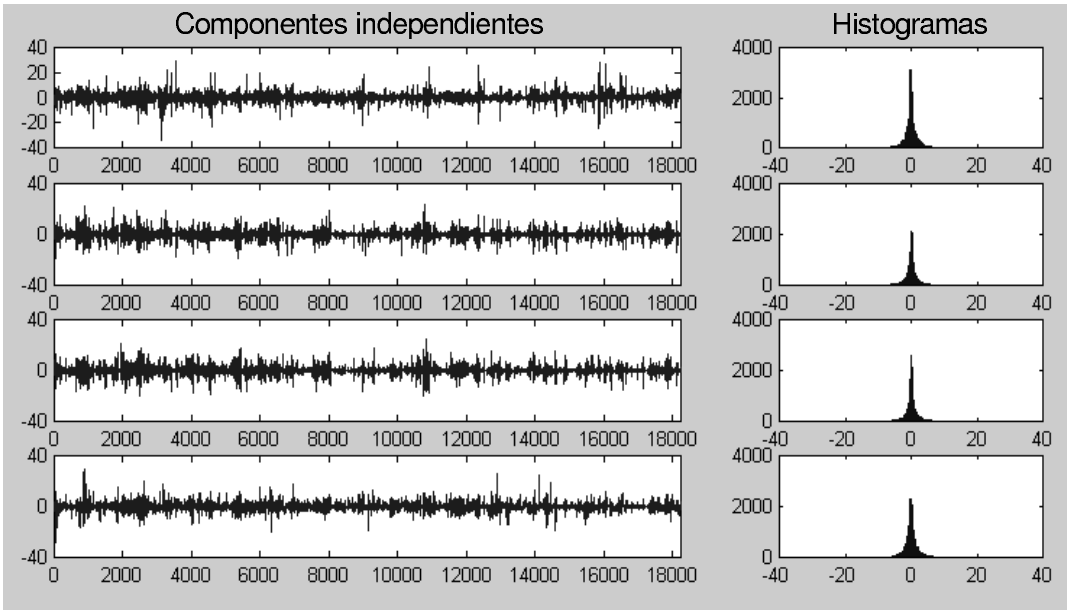

Figura 2.11: Algunas de las componentes independientes obtenidas al aplicar el algoritmo *infomax* a 18 240 bloques de tamaño 12×12 extraídos de un conjunto de imágenes naturales representadas en escala de grises (tomadas de la colección de imágenes naturales [NatCollec]). A través de sus histogramas podemos observar que su distribución es dispersa.

2.3.2. ¿Por qué las bases ICA de imágenes naturales tienen el aspecto de «bordes» y sus componentes independientes presentan una distribución «dispersa»?

Realmente es muy llamativo el hecho de que las bases ICA de imágenes naturales muestren el aspecto de bordes, recordando a los campos receptivos de las neuronas simples de la corteza visual primaria. No lo es menos el que, además, la distribución de las componentes independientes sea «dispersa», la misma distribución observada para las respuestas de las neuronas de la corteza visual. Como ya hemos comentado, han sido muchos los artículos que se han publicado añadiendo nuevos matices a esta conexión entre ICA y el sistema visual humano. Sin embargo, en todos ellos las conclusiones aportadas han estado basadas únicamente en observaciones experimentales, sin ofrecer nunca ninguna prueba matemática que explique los resultados obtenidos. Aquí surge la motivación de esta Tesis: demostrar **matemáticamente**

1. por qué las bases ICA de imágenes naturales tienen el aspecto de bordes, y
2. por qué la distribución de las componentes independientes de imágenes naturales es «dispersa».

En el siguiente Capítulo daremos respuesta a estas cuestiones, que se verán confirmadas en el Capítulo 4 por los resultados experimentales.

Capítulo 3

Interpretación de ICA. ICA aplicado a imágenes

3.1. Presentación

Hay muchos métodos para llevar a cabo ICA. Puesto que la mayoría son equivalentes [HyvKO01], nos ocuparemos sólo de los que son matemáticamente más tratables: aquéllos que se basan en maximizar o minimizar estadísticos de alto orden. En concreto, caracterizaremos las componentes independientes que se obtienen al maximizar el coeficiente de asimetría y la curtosis. *De esta forma iremos presentando las principales aportaciones de esta Tesis:* demostraremos por primera vez que las componentes independientes sólo pueden ser de dos clases, esto es, según la notación que emplearemos, de «Clase 1» o «Clase 2», siendo la «Clase 1» un caso muy particular que raramente se presenta en la práctica. Después, como aplicación, desarrollaremos una original teoría para explicar los resultados que se obtienen al llevar a cabo el análisis en componentes independientes de imágenes. La estructura que tiene este Capítulo es la siguiente:

1. Se estudian en detalle los criterios ICA basados en el coeficiente de asimetría y la curtosis. Comenzaremos con los métodos para la extracción de una única componente independiente para después tratar los métodos de extracción de varias componentes .

2. Se analizan matemáticamente las componentes independientes que determina el popular algoritmo FastICA [HyvKO01, FastICA] para demostrar que pertenecen a la categoría que hemos llamado «Clase 2».

3. En último lugar se discuten los resultados obtenidos al aplicar ICA a una imagen.

3.1.1. Notación

Utilizaremos los mismos convenios que en Capítulos anteriores (se reproducen aquí sólo para la comodidad del lector): $\mathbf{X}$ será la matriz $N \times T$ de las observaciones; $\mathbf{X}^{\perp}$ será la matriz de las observaciones «blanqueadas» o «no correlacionadas», es decir,

$$\mathbf{X}^{\perp} = \mathbf{W}\,\mathbf{X} \tag{3.1}$$

donde $\mathbf{W}$ es la matriz $N \times N$ de *whitening* o «blanqueado» (ver Capítulo **??**); la k-ésima columna de $\mathbf{X}$ (respectivamente $\mathbf{X}^{\perp}$) va a ser denotada como $\mathbf{x}_{:k} = [x_{1k}, \ldots, x_{Nk}]^{\dagger}$ (respectivamente $\mathbf{x}^{\perp}_{:k} = [x^{\perp}_{1k}, \ldots, x^{\perp}_{Nk}]^{\dagger}$) donde † indica «transposición» — obviamente

$$\mathbf{x}^{\perp}_{:k} = \mathbf{W}\,\mathbf{x}; \tag{3.2}$$

la matriz de separación será $\mathbf{B}$; la i-ésima fila de $\mathbf{B}$ va a ser representada con el vector $\mathbf{b}_{i:} = [b_{i1}, \ldots, b_{iN}]$, cuyas dimensiones son $1 \times N$; *la matriz ortogonal de separación* será $\mathbf{B}^{\perp}$, cuya i-ésima fila se representará como $\mathbf{b}^{\perp}_{i:}$, verificándose que

$$\mathbf{B} = \mathbf{B}^{\perp}\,\mathbf{W}. \tag{3.3}$$

Finalmente, sea $\mathbf{Y}$ la matriz $N \times T$ que contiene las componentes independientes. Por construcción:

$$\mathbf{Y} = \mathbf{B}\,\mathbf{X} = \mathbf{B}^{\perp}\,\mathbf{W}\,\mathbf{X} = \mathbf{B}^{\perp}\,\mathbf{X}^{\perp}. \tag{3.4}$$

Cualquier otro símbolo que se utilice en el texto será definido en el momento de su aparición.

3.2. Criterios para la extracción de una componente independiente basados en estadísticos de alto orden

3.2.1. El Coeficiente de Asimetría como criterio

Sea $\mathbf{y}_{1:} = [y_{11}, \ldots, y_{1T}]$ el vector fila obtenido como

$$\mathbf{y}_{1:} = \mathbf{b}_{1:}\mathbf{X}. \tag{3.5}$$

Puesto que $\mathbf{b}_{1:} = [b_{11}, \ldots, b_{1N}]$ representa a la primera fila de la matriz de separación $\mathbf{B}$, resulta claro que $\mathbf{y}_{1:}$ es la realización de la primera componente independiente $\mathcal{Y}_1$.

Los coeficientes $b_{11}, \ldots, b_{1N}$ deben resolver el problema **[Referencia]**

$$\begin{aligned} &\max_{\mathbf{b}_{1:}} J_3(\mathcal{Y}_1) = E\{\mathcal{Y}_1^3\} \\ &\text{sujeto a la restricción } E^2\{\mathcal{Y}_1^2\} = 1 \end{aligned} \tag{3.6}$$

donde $J_3(\mathcal{Y}_1)$ es el momento central de tercer orden o *coeficiente de asimetría* de $\mathcal{Y}_1$. La restricción $E^2\{\mathcal{Y}_1^2\} = 1$ evita que la componente independiente crezca sin control. Se prefiere la restricción $E^2\{\mathcal{Y}_1^2\} = 1$ a la equivalente $E\{\mathcal{Y}_1^2\} = 1$ por razones de conveniencia de cálculo. En la práctica, supuesto que $\mathcal{Y}_1$ es un proceso ergódico, se puede hacer la aproximación $E\{\mathcal{Y}_1^p\} \approx \frac{1}{T}\sum_{k=1}^{T} y_{1k}^p$ de donde (3.6) acaba siendo sustituido por:

$$\begin{aligned} &\max_{\mathbf{b}_{1:}} \hat{J}_3(\mathcal{Y}_1) \stackrel{def}{=} \frac{1}{T}\sum_{k=1}^{T} y_{1k}^3 \\ &\text{sujeto a la restricción } \left(\frac{1}{T}\sum_{k=1}^{T} y_{1k}^2\right)^2 = 1 \end{aligned} \tag{3.7}$$

El lagrangiano de (3.7) resulta ser[1]:

$$L_3(\mathcal{Y}_1, \lambda_3) = \frac{1}{T}\sum_{k=1}^{T} y_{1k}^3 - \lambda_3 \left\{ \left(\frac{1}{T}\sum_{k=1}^{T} y_{1k}^2 \right)^2 - 1 \right\} \tag{3.8}$$

donde λ_3 es el multiplicador de Lagrange. Finalmente, el problema (3.7) acaba siendo equivalente a:

$$\max_{\mathbf{b}_{1:}, \lambda_3} L_3(\mathcal{Y}_1, \lambda_3) \tag{3.9}$$

3.2.2. La Curtosis como criterio

En este caso, los coeficientes $b_{11}, \ldots, b_{1N}$ se obtienen al resolver el problema **[Referencia]**

$$\begin{aligned} &\max_{\mathbf{b}_{1:}} J_4(\mathcal{Y}_1) = E\{\mathcal{Y}_1^4\} - 3\,E^2\{\mathcal{Y}_1^2\} \\ &\text{sujeto a la restricción } E^2\{\mathcal{Y}_1^2\} = 1 \end{aligned} \tag{3.10}$$

donde se conserva la notación empleada en el apartado anterior. Nótese que $J_4(\mathcal{Y}_1)$ es la curtosis de $\mathcal{Y}_1$. Procediendo como antes, en la práctica resolveremos (3.10) a partir del problema equivalente

$$\max_{\mathbf{b}_{1:}, \lambda_4'} L_4'(\mathcal{Y}_1, \lambda_4') \tag{3.11}$$

donde el lagrangiano tiene la expresión

$$\begin{aligned} L_4'(\mathcal{Y}_1, \lambda_4') &= \frac{1}{T}\sum_{k=1}^{T} y_{1k}^4 - 3\left(\frac{1}{T}\sum_{k=1}^{T} y_{1k}^2 \right)^2 - \lambda_4' \left\{ \left(\frac{1}{T}\sum_{k=1}^{T} y_{1k}^2 \right)^2 - 1 \right\} \\ &= \frac{1}{T}\sum_{k=1}^{T} y_{1k}^4 - (3 + \lambda_4') \left\{ \left(\frac{1}{T}\sum_{k=1}^{T} y_{1k}^2 \right)^2 - 1 \right\} - 3 \end{aligned} \tag{3.12}$$

Obviamente, podemos sumar la constante «3» a (3.12) sin alterar la posición de sus máximos y mínimos. Haciéndolo, obtenemos la formulación definitiva del problema de optimización:

$$\max_{\mathbf{b}_{1:}, \lambda_4} L_4(\mathcal{Y}_1, \lambda_4) \tag{3.13}$$

[1] Estrictamente hablando, (3.7) puede ser considerado un abuso de notación porque $\hat{J}_3$ no es una función del proceso $\mathcal{Y}_1$, como se da a entender, sino de la realización concreta $y_{11}, \ldots, y_{1T}$. No obstante, se mantiene por ser una notación más compacta.

siendo

$$\begin{aligned} L_4(\mathcal{Y}_1, \lambda_4) &\stackrel{def}{=} L'_4(\mathcal{Y}_1, \lambda'_4) + 3 \\ &= \frac{1}{T}\sum_{k=1}^{T} y_{1k}^4 - \lambda_4 \left\{ \left(\frac{1}{T}\sum_{k=1}^{T} y_{1k}^2 \right)^2 - 1 \right\} \end{aligned} \tag{3.14}$$

donde hemos definido

$$\lambda_4 = 3 + \lambda'_4 \tag{3.15}$$

Obsérvese la gran similitud que existe entre (3.8) y (3.14): sólo difieren en que (3.8) utiliza el momento de tercer orden de $\mathcal{Y}_1$ donde (3.14) usa el momento de cuarto orden y viceversa. De hecho, esto es lo que buscábamos con las manipulaciones y cambios de variable llevados a cabo.

Como detalle adicional, nótese que hubiera sido posible una formulación alternativa en la que, teniendo en cuenta la restricción $E^2\{\mathcal{Y}_1^2\} = 1$, hubiésemos definido en (3.10) la función objetivo $J_4(\mathcal{Y}_1)$ simplemente como $J_4(\mathcal{Y}_1) = E\{\mathcal{Y}_1^4\} - 3$. De haber procedido así, es fácil comprobar que también hubiésemos obtenido un lagrangiano idéntico a (3.14).

3.3. Las derivadas del lagrangiano

Sea

$$L_p(\mathcal{Y}_1, \lambda_p) \stackrel{def}{=} \frac{1}{T}\sum_{k=1}^{T} y_{1k}^p - \lambda_p \left\{ \left(\frac{1}{T}\sum_{k=1}^{T} y_{1k}^2 \right)^2 - 1 \right\} \tag{3.16}$$

Según sea $p = 3$ ó $p = 4$, se obtienen los lagrangianos (3.8) o (3.14). De esta manera, nótese que el siguiente desarrollo es aplicable a ambos. Los puntos estacionarios (o puntos críticos) del lagrangiano son las soluciones del siguiente sistema de ecuaciones:

$$\begin{aligned} \frac{\partial L_p}{\partial \lambda_p} &= \left(\frac{1}{T}\sum_{k=1}^{T} y_{1k}^2 \right)^2 - 1 = 0 \\ \frac{\partial L_p}{\partial b_{1n}} &= \frac{p}{T}\left(\sum_{k=1}^{T} y_{1k}^{p-1}\, x_{nk} \right) - \frac{4\lambda_p}{T}\left(\sum_{k=1}^{T} y_{1k}^2 \right)\left(\sum_{k=1}^{T} y_{1k}\, x_{nk} \right) = 0 \end{aligned} \tag{3.17}$$

para $n = 1, \ldots, N$. Para reescribir (3.17) en forma matricial, definimos el vector fila $1 \times N$

$$\frac{\partial L_p}{\partial \mathbf{b}_{1:}} \stackrel{def}{=} \left[\frac{\partial L_p}{\partial b_{11}}, \ldots, \frac{\partial L_p}{\partial b_{1N}}\right] \tag{3.18}$$

y el vector fila $1 \times T$

$$\mathbf{y}_{1:}^{p} \stackrel{def}{=} [y_{11}^{p}, \ldots, y_{1T}^{p}]. \tag{3.19}$$

Por ejemplo $\mathbf{y}_{1:}^{1} \equiv \mathbf{y}_{1:} = [y_{11}, \ldots, y_{1T}]$, que es la primera fila de la matriz $\mathbf{Y}$ de componentes independientes. Con estos nuevos símbolos, (3.17) se puede escribir de manera compacta como:

$$\begin{aligned} \frac{\partial L_p}{\partial \lambda_p} &= \left(\frac{1}{T}\,\|\mathbf{y}_{1:}\|^2\right)^2 - 1 = 0 \\ \frac{\partial L_p}{\partial \mathbf{b}_{1:}} &= \frac{1}{T}\left\{p\,\mathbf{y}_{1:}^{p-1} - 4\,\lambda_p\,\|\mathbf{y}_{1:}\|^2\,\mathbf{y}_{1:}\right\}\,\mathbf{X}^{\dagger} = \mathbf{0} \end{aligned} \tag{3.20}$$

donde $\mathbf{X} = (x_{ij})$ es la matriz que contiene a las observaciones e $\|\mathbf{y}_{1:}\|^2 = \sum_{k=1}^{T} y_{1k}^2$ es la norma-2 del vector $\mathbf{y}_{1:}$.

3.4. Caracterización de las componentes independientes

En esta Sección se demuestra que, utilizando (3.20), podemos clasificar las componentes independientes en dos grandes grupos («Clase 1» y «Clase 2»).

Primera clase: soluciones «ralas»

Nos referimos aquí a los vectores $\mathbf{y}_{1:}$ de componentes independientes que resuelven (3.20) porque verifican

$$\begin{aligned} \left(\frac{1}{T}\,\|\mathbf{y}_{1:}\|^2\right)^2 - 1 = 0 \\ p\,\mathbf{y}_{1:}^{p-1} - 4\,\lambda_p\,\|\mathbf{y}_{1:}\|^2\,\mathbf{y}_{1:} = \mathbf{0} \end{aligned} \tag{3.21}$$

o, desglosando la segunda ecuación elemento a elemento

$$\begin{aligned} \left(\frac{1}{T}\,\|\mathbf{y}_{1:}\|^2\right)^2 - 1 = 0 \\ p\,y_{1k}^{p-1} - 4\,\lambda_p\,\|\mathbf{y}_{1:}\|^2\,y_{1k} = 0 \Longrightarrow p\,y_{1k}^{p-1} = 4\,\lambda_p\,\|\mathbf{y}_{1:}\|^2\,y_{1k} \end{aligned} \tag{3.22}$$

para $k = 1, \ldots, T$. Se obtiene inmediatamente que, para cada valor de k,

$$y_{1k} = 0 \qquad \text{ó} \qquad y_{1k} = \left(\frac{4\,\lambda_p\,T}{p} \right)^{\frac{1}{p-2}} \tag{3.23}$$

Teniendo en cuenta la restricción $\frac{1}{T}\,\|\mathbf{y}_{1:}\|^2 = \frac{1}{T}\sum_{k=1}^{T} y_{1k}^2 = 1$, se deduce que:

- Cualquier vector $1 \times T$ con $1 \leq L \leq T$ elementos no nulos e iguales a $\sqrt{T/L}$ es solución de (3.21) [y por lo tanto también de (3.20)].
- Lo mismo puede decirse de cualquier vector $1 \times T$ con L elementos no nulos e iguales a $-\sqrt{T/L}$.
- Asimismo, si $p = 4$ entonces (3.21) también admite como solución a cualquier vector $1 \times T$ con L elementos no nulos e iguales, *en módulo*, a $\sqrt{T/L}$. Se permite que no todos los elementos del vector tengan el mismo signo.
- Si $p = 3$, $\lambda_3 = \frac{3}{4}\frac{1}{\sqrt{LT}}$ cuando los elementos no nulos de $\mathbf{y}_{1:}$ son positivos o $\lambda_3 = -\frac{3}{4}\frac{1}{\sqrt{LT}}$ cuando dichos elementos son negativos. Si $p = 4$, $\lambda_4 = \frac{1}{L}$ siempre.

Estas características quedan recogidas en el Cuadro 3.1.

Son vectores con L $(1 \leq L \leq T)$ elementos no nulos y $T - L$ iguales a cero	
Si $p = 3$	
$\lambda_3 = \frac{3}{4}\frac{1}{\sqrt{LT}}$	$\lambda_3 = -\frac{3}{4}\frac{1}{\sqrt{LT}}$
$y_{1k} = 0$ ó $y_{1k} = \sqrt{\frac{T}{L}}$	$y_{1k} = 0$ ó $y_{1k} = -\sqrt{\frac{T}{L}}$
Si $p = 4$	
$\lambda_4 = \frac{1}{L}$	
$y_{1k} = 0$ ó $y_{1k} = \pm\sqrt{\frac{T}{L}}$	

Cuadro 3.1: Puntos estacionarios del lagrangiano de primera clase («soluciones ralas»).

En cuanto al carácter de máximo o mínimo de las soluciones, éste se explica a continuación:

Teorema 3.1. *Sea* $\mathbf{y}_{1:}$ *un vector* $1 \times T$ *tal que* $\frac{1}{T}\|\mathbf{y}_{1:}\|^2 = 1$. *Entonces,*

1. *Si* $\mathbf{y}_{1:}$ *sólo tiene un elemento no nulo y éste es positivo, entonces* $\mathbf{y}_{1:}$ *es un máximo global de (3.6) (coeficiente de asimetría) y (3.10) (curtosis).*

2. *Si* $\mathbf{y}_{1:}$ *sólo tiene un elemento no nulo pero, a diferencia de antes, éste es negativo, entonces* $\mathbf{y}_{1:}$ *es un máximo global de (3.10) (curtosis) pero un mínimo global de (3.6) (coeficiente de asimetría).*

3. *Si todos los elementos de* $\mathbf{y}_{1:}$ *tienen módulo unidad, entonces* $\mathbf{y}_{1:}$ *es un mínimo global de (3.10) (curtosis).*

4. *Cualquier otro valor de* $\mathbf{y}_{1:}$ *es un punto de silla de (3.6) y (3.10).*

Demostración. Vamos a demostrar sólo la proposición 1. La prueba de las restantes es similar. Supongamos que $\mathbf{y}_{1:}$ tiene un único elemento no nulo. Por ejemplo, sin pérdida de generalidad se toma:

$$y_{11} = \sqrt{T}, \qquad y_{1k} = 0 \quad \text{para } 2 \leq k \leq T \tag{3.24}$$

Es bien sabido que la determinación de si es máximo o mínimo se lleva a cabo averiguando el signo de la segunda diferencial de la función de Lagrange

$$d^2\, L_p(\mathcal{Y}_1, \lambda_p) = \sum_{i,j=1}^{T} \frac{\partial^2 L_p}{\partial y_{1i} \partial y_{1j}}\, dy_{1i}\, dy_{1j} \tag{3.25}$$

sujeta a la restricción

$$d\left\{\left(\frac{1}{T}\sum_{k=1}^{T} y_{1k}^2\right)^2 - 1\right\} = 0 \longrightarrow \sum_{k=1}^{T} y_{1k}\, dy_{1k} = 0 \tag{3.26}$$

donde

$$dy_{1k} = \sum_{k=1}^{N} x_{ik} db_{1i}$$

Derivando se obtiene

$$d^2\, L_p(\mathcal{Y}_1, \lambda_p) = \sum_{i=1}^{T} \left[\frac{p^2 - p}{T}\, y_{1i}^{p-2} - \frac{4\,\lambda_p}{T} \right] (dy_{1i})^2 - \frac{8\lambda_p}{T^2} \sum_{i,j=1}^{T} y_{1i}\, y_{1j}\, (dy_{1i})\, (dy_{1j})$$

que, teniendo en cuenta (3.26), se puede simplificar a

$$d^2\, L_p(\mathcal{Y}_1, \lambda_p) = \frac{p^2 - p}{T} \sum_{i=1}^{T} y_{1i}^{p-2}\, (dy_{1i})^2 - \frac{4\,\lambda_p}{T} \sum_{i=1}^{T} (dy_{1i})^2 \tag{3.27}$$

Sustituyendo (3.24) en (3.26) y (3.27) se llega en primer lugar a que $dy_{11} = 0$ y en segundo lugar a que

$$d^2 L_p = -\frac{4\,\lambda_p}{T} \sum_{i=2}^{T} (dy_{1i})^2$$

Leemos en el Cuadro 3.1 que el multiplicador de Lagrange asociado a (3.24) es

$$\lambda_p = \begin{cases} 3/(4\sqrt{T}) & \text{si } p = 3 \\ 1 & \text{si } p = 4 \end{cases}$$

Dado que λ_p es siempre positivo, $d^2 L_p < 0$. Por lo tanto, (3.24) es un *máximo*.

En realidad, (3.24) no es un simple máximo: también es un *máximo global*. Para demostrar esto, estudiemos un caso más general de solución de clase 1: sea $\mathbf{y}_{1:}$ un vector con sólo L muestras no nulas e iguales a $\sqrt{T/L}$ (en conformidad con el Cuadro 3.1). En este caso, la función de coste (3.10) (maximizar la *curtosis*) queda simplemente reducida a

$$\sum_{i=1}^{T} y_{1k}^4 - 3 \left(\sum_{i=1}^{T} y_{1k}^2 \right)^2 = \frac{T^2}{L} - 3$$

la cual alcanza su valor máximo cuando $L = 1$. Ello pone de manifiesto, como defendíamos, el carácter de máximo global. Nótese que no es preciso contemplar los casos en los que $\mathbf{y}_{1:}$ toma valores negativos porque el signo desaparece al elevar a potencias pares. Finalmente, un razonamiento similar vale para demostrar que (3.24) también es un máximo global de (3.6) (coeficiente de asimetría). □

Como ilustración del Teorema, la Figura 3.1 muestra la superficie y una gráfica de nivel del coeficiente de asimetría para el caso $T = 3$. Los valores de y_{11}, y_{12}, y_{13} se parametrizan en coordenadas esféricas para cumplir la restricción que afecta a su varianza, es decir: $y_{11} = \sqrt{3}\cos(\theta)\cos(\phi)$, $y_{12} = \sqrt{3}\cos(\theta)\sin(\phi)$, $y_{13} = \sqrt{3}\sin(\theta)$. La figura muestra, por ejemplo, los máximos y mínimos en $\theta = 0$, $\phi = 0, \pi/2, \pi, 3\pi/2$. Todos ellos son extremos globales (aún cuando la perspectiva no permita apreciarlo perfectamente).

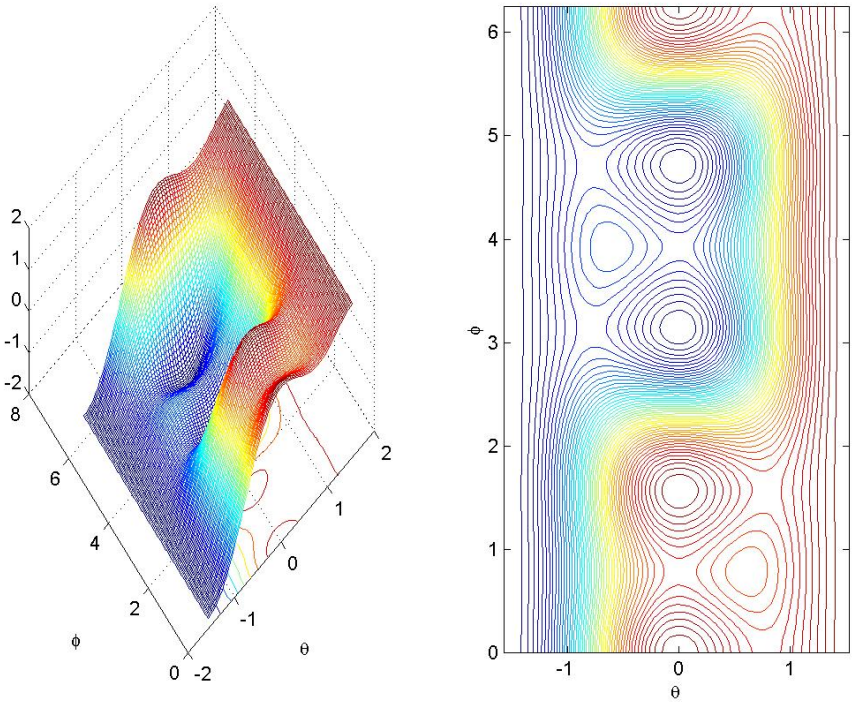

Figura 3.1: El coeficiente de asimetría ($T = 3$).

Segunda clase

El segundo grupo de soluciones de (3.20) está formado por aquellos vectores $\mathbf{y}_{1:}$ que verifican

$$\left\{ p\,\mathbf{y}_{1:}^{p-1} - 4\,\lambda_p \left\|\mathbf{y}_{1:}\right\|^2 \mathbf{y}_{1:} \right\} \mathbf{X}^{\dagger} = \mathbf{0}, \tag{3.28}$$

porque $p\,\mathbf{y}_{1:}^{p-1} - 4\,\lambda_p\,\|\mathbf{y}_{1:}\|^2\,\mathbf{y}_{1:}$ es **perpendicular** a todas las **filas** de la matriz de observaciones $\mathbf{X}$. Teniendo en cuenta la restricción $\frac{1}{T}\|\mathbf{y}_{1:}\|^2 = 1$, el sistema de ecuaciones (3.28) se puede reescribir elemento a elemento como

$$\sum_{k=1}^{T} x_{ik}\{p\,y_{1k}^{p-1} - 4\lambda_p\,T\,y_{1k}\} = 0 \tag{3.29}$$

para $i = 1, \ldots, N$. De aquí se despeja

$$\frac{\sum_{k=1}^{T} x_{ik}\,y_{1k}^{p-1}}{\sum_{k=1}^{T} x_{ik}\,y_{1k}} = \frac{4}{p}\,\lambda_p\,T \tag{3.30}$$

para $i = 1, \ldots, N$. Como el término de la derecha en la última igualdad no depende del índice i, se deduce finalmente que

$$\boxed{\frac{\sum_{k=1}^{T} x_{1k}\,y_{1k}^{p-1}}{\sum_{k=1}^{T} x_{1k}\,y_{1k}} = \ldots = \frac{\sum_{k=1}^{T} x_{Nk}\,y_{1k}^{p-1}}{\sum_{k=1}^{T} x_{Nk}\,y_{1k}}} \tag{3.31}$$

siendo ésta la relación que define las componentes independientes de «Clase 2».

3.5. El algoritmo «FastICA»

El algoritmo «FastICA» [HyvOja97, FastICA] es el más popular de los algoritmos que llevan a cabo ICA maximizando estadísticos de alto orden. Básicamente, «FastICA» es un algoritmo de punto fijo que se usa para resolver la ecuación (3.20) (página 36) o ecuación que verifican las componentes independientes. Reproducimos aquí íntegra esa ecuación (3.20) para comodidad del lector:

$$\begin{aligned} \frac{\partial L_p}{\partial \lambda_p} &= \left(\frac{1}{T}\,\|\mathbf{y}_{1:}\|^2\right)^2 - 1 = 0 \\ \frac{\partial L_p}{\partial \mathbf{b}_{1:}} &= \frac{1}{T}\left\{p\,\mathbf{y}_{1:}^{p-1} - 4\,\lambda_p\,\|\mathbf{y}_{1:}\|^2\,\mathbf{y}_{1:}\right\}\,\mathbf{X}^{\dagger} = \mathbf{0} \end{aligned} \tag{3.32}$$

donde $p = 3, 4$, $\mathbf{b}_{1:} = [b_{11}, \ldots, b_{1N}]$ es el vector fila de separación, $\mathbf{y}_{1:} = [y_{11}, \ldots, y_{1T}]$ es la componente independiente, $\mathbf{y}_{1:}^{p-1} = [y_{11}^{p-1}, \ldots, y_{1T}^{p-1}]$, $\mathbf{X}$ es la matriz $N \times T$ de las observaciones y λ_p es un multiplicador de Lagrange. Aunque el algoritmo «FastICA» es bien conocido [HyvOja97, HyvKO01, FastICA] vamos no obstante a

enfocar su formulación desde el punto de vista de las ecuaciones que hemos desarrollado en las secciones anteriores pues será útil para derivaciones posteriores. A partir de (3.32) se deduce fácilmente que $\mathbf{y}_{1:}$ verifica

$$\frac{4}{T}\lambda_p \left\|\mathbf{y}_{1:}\right\|^2 \mathbf{y}_{1:}\,\mathbf{X}^\dagger = \frac{p}{T}\,\mathbf{y}_{1:}^{p-1}\,\mathbf{X}^\dagger$$

como, por definición, $\mathbf{y}_{1:} = \mathbf{b}_{1:}\,\mathbf{X}$, se tiene que

$$\frac{4}{T}\lambda_p \left\|\mathbf{y}_{1:}\right\|^2 \mathbf{b}_{1:}\,\mathbf{X}\,\mathbf{X}^\dagger = \frac{p}{T}\,\mathbf{y}_{1:}^{p-1}\,\mathbf{X}^\dagger$$

Ahora bien, $\frac{1}{T}\,\mathbf{X}\,\mathbf{X}^\dagger \overset{def}{=} \mathbf{R}_x$ es la matriz de correlación de los datos. Se sabe que

$$\mathbf{W}\,\mathbf{R}_x\,\mathbf{W}^\dagger = \mathbf{I} \Longrightarrow \mathbf{R}_x \;= \mathbf{W}^{-1}\,\mathbf{W}^{-\dagger}$$

donde $\mathbf{W}$ es la matriz de «whitening» o «blanqueado». Sustituyendo y operando se tiene que

$$4\lambda_p \left\|\mathbf{y}_{1:}\right\|^2 \mathbf{b}_{1:}\,\mathbf{W}^{-1} = \frac{p}{T}\,\mathbf{y}_{1:}^{p-1}\,\mathbf{X}^\dagger\,\mathbf{W}^\dagger$$

pero $\mathbf{b}_{1:}\,\mathbf{W}^{-1} = \mathbf{b}_{1:}^{\perp}$ es la primera fila de la *matriz ortogonal* de separación $\mathbf{B}^{\perp}$, mientras que $\mathbf{W}\,\mathbf{X} = \mathbf{X}^{\perp}$ es la matriz de observaciones no correlacionadas. De aquí se llega a que

$$4\lambda_p \left\|\mathbf{y}_{1:}\right\|^2 \mathbf{b}_{1:}^{\perp} = \frac{p}{T}\,\mathbf{y}_{1:}^{p-1}\;\mathrm{trans}\left(\mathbf{X}^{\perp}\right) \tag{3.33}$$

donde trans$(\,\cdot\,)$ denota «transpuesta» (empleamos aquí esta notación para evitar el uso excesivo de superíndices). Se comprende que (3.33) es el equivalente a (3.32) cuando expresamos las ecuaciones en función de $\mathbf{b}_{1:}^{\perp}$ y $\mathbf{X}^{\perp}$ en lugar de $\mathbf{b}_{1:}$ y $\mathbf{X}$. De igual manera, la restricción

$$\left(\frac{1}{T}\left\|\mathbf{y}_{1:}\right\|^2\right)^2 - 1 = 0$$

se transforma con facilidad usando que $\mathbf{y}_{1:} = \mathbf{b}_{1:}^{\perp}\,\mathbf{X}^{\perp}$ en

$$\left(\left\|\mathbf{b}_{1:}^{\perp}\right\|^2\right)^2 - 1 = 0 \Rightarrow \left\|\mathbf{b}_{1:}^{\perp}\right\| = 1 \tag{3.34}$$

3.5.1. FastICA para el coeficiente de asimetría

Haciendo $p = 3$, (3.33) se reescribe

$$\mathbf{b}_{1:}^{\perp} = \frac{3}{4\,T\,\lambda_3\,\|\mathbf{y}_{1:}\|^2}\,\mathbf{y}_{1:}^2\ \mathrm{trans}\left(\mathbf{X}^{\perp}\right)$$

que indica que, en los puntos extremos del criterio, $\mathbf{b}_{1:}^{\perp}$ ha de ser *proporcional* al vector $\mathbf{y}_{1:}^2\ \mathrm{trans}\left(\mathbf{X}^{\perp}\right)$. Es decir,

$$\mathbf{b}_{1:}^{\perp} \propto \mathbf{y}_{1:}^2\ \mathrm{trans}\left(\mathbf{X}^{\perp}\right) \tag{3.35}$$

La búsqueda de las soluciones de (3.35) puede llevarse a cabo utilizando un método de punto fijo: en cada iteración $\mathbf{b}_{1:}^{\perp}$ se sobreescribe con el valor de $\mathbf{y}_{1:}^2\ \mathrm{trans}\left(\mathbf{X}^{\perp}\right)$. Después, para satisfacer (3.34), normalizamos $\mathbf{y}_{1:}^2$. De este modo se asegura que la longitud de $\mathbf{y}_{1:}^2$ es siempre *uno* y, por eso, no es preciso controlarla de forma explícita en (3.35). El algoritmo completo se transcribe en el Cuadro 3.2.

---

1. Sea $\mathbf{b}_{1:}^{\perp}$ un vector $1 \times N$ cualquiera de módulo unidad
2. Calcular $\mathbf{y}_{1:} = \mathbf{b}_{1:}^{\perp}\,\mathbf{X}^{\perp}$
3. Sustituir $\mathbf{b}_{1:}^{\perp}$ por $\mathbf{y}_{1:}^2\ \mathrm{trans}\left(\mathbf{X}^{\perp}\right)$
4. Normalizar $\mathbf{b}_{1:}^{\perp}$ dividiéndolo por su módulo
5. Volver al paso 2 y repetir hasta que se alcance la convergencia.

Cuadro 3.2: Algoritmo FastICA que optimiza el coeficiente de asimetría de las componentes independientes

3.5.2. FastICA para maximizar la curtosis

Tomando $p = 4$, (3.33) se reescribe

$$\lambda_4\,\|\mathbf{y}_{1:}\|^2\,\mathbf{b}_{1:}^{\perp} = \frac{1}{T}\,\mathbf{y}_{1:}^3\ \mathrm{trans}\left(\mathbf{X}^{\perp}\right) \tag{3.36}$$

que, como vimos previamente, indica que en la solución

$$\mathbf{b}_{1:}^{\perp} \propto \mathbf{y}_{1:}^{3} \text{ trans}\left(\mathbf{X}^{\perp}\right) \tag{3.37}$$

Ahora podríamos proceder como en el apartado anterior. No obstante, resulta sencillo demostrar que el algoritmo converge mejor haciendo lo que sigue: definimos (algo arbitrariamente) una cantidad β como $\lambda_4 \left\|\mathbf{y}_{1:}\right\|^2 = \beta + 3$. Entonces, (3.36) se puede reescribir como $\beta\, \mathbf{b}_{1:}^{\perp} = \frac{1}{T}\, \mathbf{y}_{1:}^{3} \text{ trans}\left(\mathbf{X}^{\perp}\right) - 3\, \mathbf{b}_{1:}^{\perp}$ que indica que, en la solución, también es cierto que

$$\mathbf{b}_{1:}^{\perp} \propto \frac{1}{T}\, \mathbf{y}_{1:}^{3} \text{ trans}\left(\mathbf{X}^{\perp}\right) - 3\, \mathbf{b}_{1:}^{\perp} \tag{3.38}$$

Debe entenderse que (3.37) y (3.38) son relaciones completamente equivalentes: si $\mathbf{b}_{1:}^{\perp}$ es paralelo a $\mathbf{y}_{1:}^{3} \text{ trans}\left(\mathbf{X}^{\perp}\right)$, se deduce inmediatamente que todos los vectores de la forma $\gamma\, \mathbf{y}_{1:}^{3} \text{ trans}\left(\mathbf{X}^{\perp}\right) - \delta\, \mathbf{b}_{1:}^{\perp}$, ($\gamma, \delta$ = constantes) son también paralelos a $\mathbf{b}_{1:}^{\perp}$. Puede comprobarse fácilmente que la afirmación recíproca también es cierta. Finalmente, a partir de (3.38) se formula el algoritmo FastICA. Los pasos de este algoritmo se enumeran en el Cuadro 3.3.

---

1. Sea $\mathbf{b}_{1:}^{\perp}$ un vector $1 \times N$ cualquiera de módulo unidad
2. Calcular $\mathbf{y}_{1:} = \mathbf{b}_{1:}^{\perp}\, \mathbf{X}^{\perp}$
3. Sustituir $\mathbf{b}_{1:}^{\perp}$ por $\frac{1}{T}\, \mathbf{y}_{1:}^{3} \text{ trans}\left(\mathbf{X}^{\perp}\right) - 3\, \mathbf{b}_{1:}^{\perp}$
4. Normalizar $\mathbf{b}_{1:}^{\perp}$ dividiéndolo por su módulo
5. Volver al paso 2 y repetir hasta que se alcance la convergencia.

Cuadro 3.3: Algoritmo FastICA que optimiza la curtosis de las componentes independientes

3.6. Los puntos estacionarios de «FastICA»

En esta Sección se va a presentar un Lema que determina cómo son las componentes independientes que determina «FastICA». Este Lema es otro de los resultados relevantes obtenidos en esta Tesis.

Como se ha visto anteriormente, el algoritmo del Cuadro 3.2 se detiene cuando el vector $\mathbf{b}_{1:}^{\perp}$ es tal que

$$\mathbf{b}_{1:}^{\perp} \propto \mathbf{y}_{1:}^{2} \text{ trans}\left(\mathbf{X}^{\perp}\right)$$

Es decir, $\mathbf{b}_{1:}^{\perp}$ sería un vector propio de la transformación no lineal

$$\mathbf{b}_{1:}^{\perp} \longrightarrow \mathbf{y}_{1:}^{2} \text{ trans}\left(\mathbf{X}^{\perp}\right)$$

por lo que la iteración del Cuadro 3.2 recuerda poderosamente al «método de las potencias» utilizado para determinar autovectores en Álgebra Lineal. Del mismo modo, el algoritmo del Cuadro 3.3 se detiene cuando

$$\mathbf{b}_{1:}^{\perp} \propto \mathbf{y}_{1:}^{3} \text{ trans}\left(\mathbf{X}^{\perp}\right) - 3\,\mathbf{b}_{1:}^{\perp}$$

o, lo que es lo mismo, cuando

$$\mathbf{b}_{1:}^{\perp} \propto \mathbf{y}_{1:}^{3} \text{ trans}\left(\mathbf{X}^{\perp}\right)$$

En general, se podrá decir entonces que los algoritmos se van a detener en el momento en que

$$\boxed{\mathbf{b}_{1:}^{\perp} \propto \mathbf{y}_{1:}^{p-1} \text{ trans}\left(\mathbf{X}^{\perp}\right)} \tag{3.39}$$

donde $p = 3, 4$ según sea el caso. Con este resultado, es posible demostrar el siguiente lema:

Lema 3.1. *Si* $\mathbf{X}$ *es cuadrada e invertible, los puntos estacionarios de los algoritmos FastICA son soluciones de «Clase 1» o de «Clase 2». En otro caso, dichos puntos estacionarios son siempre soluciones de «Clase 2».*

El lema es importante por cuanto que asegura que, en la práctica, los algoritmos FastICA no van a generar componente independiente $\mathcal{Y}_1$ alguna[2] que no verifique

[2] Recuérdese que, por ahora, sólo estamos considerando la generación de la componente independiente $\mathcal{Y}_1$.

las condiciones de la «Clase 2». Vamos a dedicar el resto de la presente Sección a demostrar este lema.

Prueba del Lema. Utilizando que

$$\mathbf{X}^{\perp} = \mathbf{W}\,\mathbf{X} \Rightarrow \mathrm{trans}\left(\mathbf{X}^{\perp}\right) = \mathbf{X}^{\dagger}\,\mathbf{W}^{\dagger}$$

donde $\mathbf{W}$ es la matriz de «blanqueado», se llega a que

$$\mathbf{b}_{1:}^{\perp} \propto \mathbf{y}_{1:}^{p-1}\,\mathbf{X}^{\dagger}\,\mathbf{W}^{\dagger} \tag{3.40}$$

(nótese que, como $\mathbf{y}_{1:}^{p-1}\,\mathbf{X}^{\dagger}$ es un vector fila $1\times N$, se puede deducir inmediatamente que $\mathbf{b}_{1:}^{\perp}$ es una combinación lineal de las columnas de $\mathbf{W}$). Por definición

$$\mathbf{b}_{1:} = \mathbf{b}_{1:}^{\perp}\,\mathbf{W} \Rightarrow \mathbf{b}_{1:}^{\perp} = \mathbf{b}_{1:}\,\mathbf{W}^{-1}$$

luego podemos reescribir (3.40) como

$$\mathbf{b}_{1:}\,\mathbf{W}^{-1} \propto \mathbf{y}_{1:}^{p-1}\,\mathbf{X}^{\dagger}\,\mathbf{W}^{\dagger} \quad \overset{(a)}{\Longrightarrow} \quad \mathbf{b}_{1:}\,\mathbf{W}^{-1}\,\mathbf{W}^{-\dagger} \propto \mathbf{y}_{1:}^{p-1}\,\mathbf{X}^{\dagger} \tag{3.41}$$

donde el paso (a) en (3.41) puede llevarse a cabo pues el producto por $\mathbf{W}^{-\dagger}$ no altera la proporcionalidad entre los vectores ya que $\mathbf{W}$ es una matriz diagonal multiplicada por una matriz de rotación. Se sabe que $\mathbf{R}_x = \mathbf{W}^{-1}\,\mathbf{W}^{-\dagger}$ por lo que

$$\mathbf{b}_{1:}\,\mathbf{R}_x \propto \mathbf{y}_{1:}^{p-1}\,\mathbf{X}^{\dagger} \tag{3.42}$$

a su vez, $\mathbf{R}_x = \frac{1}{T}\mathbf{X}\,\mathbf{X}^{\dagger}$. Entonces se obtiene

$$\mathbf{b}_{1:}\,\mathbf{X}\,\mathbf{X}^{\dagger} \propto \mathbf{y}_{1:}^{p-1}\,\mathbf{X}^{\dagger}$$

y como $\mathbf{y}_{1:} = \mathbf{b}_{1:}\,\mathbf{X}$, llegamos finalmente al resultado

$$\mathbf{y}_{1:}\,\mathbf{X}^{\dagger} \propto \mathbf{y}_{1:}^{p-1}\,\mathbf{X}^{\dagger} \tag{3.43}$$

Si $\mathbf{X}^{\dagger}$ es invertible (y, por lo tanto, cuadrada), entonces se deduce que $\mathbf{y}_{1:} \propto \mathbf{y}_{1:}^{p-1}$, siendo ésta la condición para ser solución de «Clase 1» o «rala» (ver página 36). En cualquier caso, escribiendo (3.43) elemento a elemento se obtiene

$$\sum_{k=1}^{T} y_{1k}\,x_{ik} \propto \sum_{k=1}^{T} y_{1k}^{p-1}\,x_{ik}$$

para todo $i = 1, \ldots, N$. Como la constante de proporcionalidad es la misma para todo i, se llega rápidamente a que ésta es la condición que verifican las soluciones de «Clase 2» (ver ecuación (3.31), página 41). □

3.7. Más sobre los puntos estacionarios

En esta Sección se profundiza sobre la naturaleza de los puntos estacionarios de FastICA. En particular, se prepara el camino para interpretar los resultados del «análisis en componentes independientes» de una imagen. Partimos de (3.42), que establece que

$$\mathbf{b}_{1:}\,\mathbf{R}_x \propto \mathbf{y}_{1:}^{p-1}\,\mathbf{X}^\dagger$$

de donde

$$\mathbf{b}_{1:} \propto \mathbf{y}_{1:}^{p-1}\,\mathbf{X}^\dagger\,\mathbf{R}_x^{-1}$$

Sea $\mathbf{R}_x = \mathbf{V}\,\mathbf{D}\,\mathbf{V}^\dagger$, donde $\mathbf{V}$ es la matriz que contiene a los autovectores de $\mathbf{R}_x$ y $\mathbf{D} = \mathrm{diag}(\lambda_1, \ldots, \lambda_N)$ es la matriz diagonal con sus autovalores. En lo que sigue, además, supondremos que $\lambda_1 > \ldots > \lambda_N$. Entonces:

$$\mathbf{b}_{1:} \propto \mathbf{y}_{1:}^{p-1}\,\mathbf{X}^\dagger\,\mathbf{V}\,\mathbf{D}^{-1/2}\,\mathbf{D}^{-1/2}\,\mathbf{V}^\dagger$$

pero $\mathbf{X}^\perp = \mathbf{D}^{-1/2}\,\mathbf{V}^\dagger\,\mathbf{X}$ luego

$$\mathbf{b}_{1:} \propto \mathbf{y}_{1:}^{p-1}\ \mathrm{trans}\left(\mathbf{X}^\perp\right)\ \mathbf{D}^{-1/2}\,\mathbf{V}^\dagger$$

Transponiendo se llega a la relación fundamental:

$$\boxed{\mathbf{b}_{1:}^\dagger \propto \mathbf{V}\,\mathbf{D}^{-1/2}\,\mathbf{X}^\perp\,\mathrm{trans}\left(\mathbf{y}_{1:}^{p-1}\right)} \tag{3.44}$$

También podemos desarrollar (3.44) como sigue:

$$\boxed{\mathbf{b}_{1:}^\dagger \propto \textstyle\sum_{i=1}^{N} \gamma_i\,\mathbf{v}_i} \tag{3.45}$$

donde $\mathbf{v}_i$ es la i-ésima columna de $\mathbf{V}$ (es decir, el i-ésimo autovector de $\mathbf{R}_x$) y los coeficientes γ_i se definen como

$$\gamma_i \overset{def}{=} \frac{\sum_{k=1}^{T} x_{ik}^\perp\,y_{1k}^{p-1}}{\sigma_{x_i}} \overset{(a)}{\propto} \frac{\sum_{k=1}^{T} x_{ik}^\perp\,y_{1k}}{\sigma_{x_i}} = \frac{b_{1i}^\perp}{\sigma_{x_i}} \tag{3.46}$$

para $p = 3, 4$, donde (a) se sigue del hecho de que las soluciones son de «Clase 2». En otras palabras, $\mathbf{b}_{1:}$, la primera fila de la matriz de separación, es una combinación lineal de los vectores $\mathbf{v}_i$. Esto en principio no tiene nada de particular:

a fin de cuentas los vectores $\mathbf{v}_i$ componen una base ortonormal del espacio. Además, (3.45) es una relación *implícita*, no *explícita*, por cuanto que los coeficientes γ_i son una función de $\mathbf{b}_{1:}$ a través de la relación $\mathbf{y}_{1:} = \mathbf{b}_{1:}\mathbf{X}$. Por eso, (3.45) no es útil para despejar el valor de $\mathbf{b}_{1:}$. Por otra parte nótese que el numerador de (3.46) está acotado: puesto que $\sum_{i=1}^{n}(b_{1i}^{\perp})^2 = 1$ no cabe esperar en la práctica que el numerador de (3.46) tome valores «excesivamente grandes» o «excesivamente pequeños». Consideremos ahora el caso particular en el que los coeficientes (3.46) son «grandes» porque su denominador es «pequeño»: suponiendo que σ_{x_i} es «suficientemente pequeño» (por ejemplo, $\sigma_{x_i} < \epsilon$ para $N_\epsilon \leq i \leq N$ donde ϵ es una cantidad «suficientemente pequeña») (3.45) se puede simplificar a

$$\mathbf{b}_{1:}^{\dagger} \propto \sum_{i=N_\epsilon}^{N} \gamma_i \, \mathbf{v}_i \tag{3.47}$$

En este caso $\mathbf{b}_{1:}$ va a resultar ser una combinación lineal de los autovectores asociados a los autovalores más pequeños de $\mathbf{R}_x$.

3.8. Extensión a más de una componente independiente

En la práctica se ejecuta «FastICA» tantas veces como componentes independientes se desee obtener [HyvOja97]. Para evitar que la misma componente se calcule más de una vez, en cada repetición del algoritmo se parte de condiciones distintas (incluso se reinicia el algoritmo una y otra vez hasta que tener la seguridad de que éste no va a converger a una solución calculada ya, utilizando técnicas más o menos sofisticadas [HyvOja97]). Si bien cada ejecución *condiciona* a las siguientes, se acepta que las propiedades del algoritmo no se ven afectadas significativamente [HyvOja97] (como se comprueba experimentalmente). En este caso, las soluciones a que llega cada ejecución del algoritmo siguen estando descritas con suficiente aproximación por la teoría anterior y, en particular, por (3.45). Nótese que los coeficientes $b_{ij}^{\perp}$ que aparece en el numerador de (3.45) se corresponden con los elementos de una matriz *ortogonal*, lo que añade una condición adicional sobre

los elementos de la matriz $\mathbf{B}$ cuyos efectos prácticos dependen del problema en cuestión que se esté abordando.

3.9. Aplicación: ICA e imágenes

Consideremos que $\mathbf{X}$ proviene de una imagen como en el Capítulo anterior ¿Qué caracteriza a una imagen natural? Hay un hecho incuestionable: los píxeles de una imagen natural guardan una alta correlación con sus vecinos. De hecho, como casi siempre se observa en la práctica, píxeles contiguos son «casi idénticos» (exceptuando aquellos situados en diferentes lados de las fronteras entre texturas). Esta gran redundancia que existe en la imagen es la base para que funcionen los algoritmos de compresión de imágenes naturales. Ahora bien, toda «compresión» no es sino una proyección de los datos que se desea comprimir sobre un espacio de menos dimensiones. Ilustraremos esto repasando brevemente el algoritmo óptimo de compresión (en el sentido de que minimiza el error cuadrático medio). Con ello además introduciremos en la discusión la matriz de correlación de los datos: sea $\mathbf{x}_{:i}$ la i-ésima columna de $\mathbf{X}$, sean $\mathbf{v}_1, \ldots, \mathbf{v}_N$ los autovectores asociados con los N autovalores $\lambda_1, \ldots, \lambda_N$ de la matriz de correlación $\mathbf{R}_x = \frac{1}{T}\,\mathbf{X}\,\mathbf{X}^\dagger$ (es decir, $\mathbf{R}_x\,\mathbf{v}_i = \lambda_i\,\mathbf{v}_i$), donde se supone que $\lambda_1 > \ldots > \lambda_N$. Como $\mathbf{R}_x$ es una matriz simétrica, se sabe que sus autovectores forman una base ortonormal del espacio. Por lo tanto, siempre es posible escribir

$$\mathbf{x}_{:i} = \sum_{k=1}^{N} \left(\mathbf{v}_k^T\,\mathbf{x}_{:i}\right)\,\mathbf{v}_k \tag{3.48}$$

que es, precisamente, la proyección de $\mathbf{x}_{:i}$ en dicha base (nótese que tanto $\mathbf{x}_{:i}$ como los $\mathbf{v}_k$ son vectores columna $N \times 1$). Truncando la serie en el término $k = r < N$ se obtiene la siguiente *aproximación* al vector $\mathbf{x}_{:i}$

$$\hat{\mathbf{x}}_{:i} = \sum_{k=1}^{r} \left(\mathbf{v}_k^T\,\mathbf{x}_{:i}\right)\,\mathbf{v}_k \tag{3.49}$$

El error cuadrático medio cometido con esta aproximación es

$$\begin{aligned}\epsilon &= \frac{1}{T}\sum_{i=1}^{T} \|\mathbf{x}_{:i} - \hat{\mathbf{x}}_{:i}\|^2 \\ &= \frac{1}{T}\sum_{i=1}^{T} \left\| \sum_{k=r+1}^{N} \left(\mathbf{v}_k^\dagger \mathbf{x}_{:i}\right) \mathbf{v}_k \right\|^2 \\ &= \frac{1}{T}\sum_{i=1}^{T} \sum_{k=r+1}^{N} \left(\mathbf{v}_k^\dagger \mathbf{x}_{:i}\right)^2 \\ &= \sum_{k=r+1}^{N} \lambda_k\end{aligned} \tag{3.50}$$

donde se ha utilizado que $\frac{1}{T}\sum_{i=1}^{T}\left(\mathbf{v}_k^\dagger \mathbf{x}_{:i}\right)^2 = \lambda_k$. De aquí se deduce que la aproximación (3.49) es *buena* siempre que los $\lambda_{r+1}, \ldots, \lambda_N$ sean suficientemente pequeños[3]. De hecho, no sólo es *buena* también es *óptima* en el sentido de que *es la aproximación que hace mínimo el error cuadrático medio*; no existe, por lo tanto, mejor forma de aproximar las columnas $\mathbf{X}$ en un espacio de $r < N$ dimensiones (nótese que $\hat{\mathbf{x}}_{:i}$ pertenece a un espacio de r dimensiones).

En la práctica *se acepta sin discusión* que, debido a su gran redundancia[4], se puede reconstruir cualquier imagen natural con (3.49) para r pequeño ($r \ll N$). Ello justifica la siguiente afirmación:

> *La matriz de correlación de una imagen natural posee pocos autovalores significativos.*

Esta conjetura caracteriza las imágenes naturales suficientemente bien. Como mera ilustración, la Figura 3.2 muestra los sesenta y cuatro autovalores de la matriz de correlación correspondiente a la imagen «Lena», donde las columnas de $\mathbf{X}$ representan a bloques de tamaño 8×8 de la imagen. Se observa con claridad que sólo unos pocos ($r = 10$ u 11) autovalores son significativos. De hecho, la

[3] Para reducir el espacio necesario para almacenar la información no se almacenan explícitamente los T vectores $\hat{\mathbf{x}}_{:i}$ sino los r vectores $\mathbf{v}_i$ y las $r \times T$ coordenadas $\mathbf{v}_i^T \mathbf{x}_{:i}$. Ésta es la base de los algoritmos de compresión.

[4] Como se dijo, píxeles contiguos de una imagen natural son casi idénticos. En la práctica esto hace que los bloques en que dividimos las imágenes pertenezcan a los mismos hiperplanos, facilitando así la proyección de los mismos sobre espacios de menos dimensiones.

Figura 3.3 muestra el resultado de recomponer «Lena» a partir de la fórmula (3.49) para distintos valores de r: la elección $r = 10$ (proyección de «Lena» sobre un espacio de sólo diez dimensiones en vez de la sesenta y cuatro originales) parece ser suficiente para lograr una reconstrucción satisfactoria.

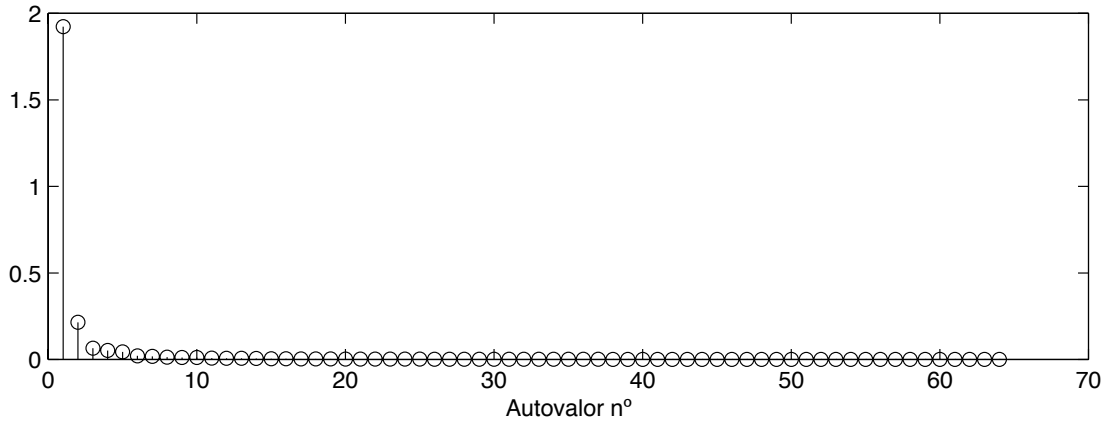

Figura 3.2: Autovalores de la matriz de correlación de «Lena» ordenados de mayor a menor (tamaño de bloque: 8×8 píxeles).

3.9.1. Los autovectores de la matriz de correlación

Sea $\mathbf{v}_1$ el autovector asociado al mayor autovalor de la matriz de correlación de la imagen. Se deduce inmediatamente de la discusión precedente que este vector es quien mejor aproxima a la imagen en el sentido de que el error cuadrático medio cometido al proyectar las columnas de $\mathbf{X}$ en su dirección[5] es *menor* que el error que se comete al proyectar sobre *cualquier otra* dirección. De hecho, la Figura 3.3 (b) muestra el resultado de recomponer bloque a bloque «Lena» como su proyección sobre $\mathbf{v}_1$: la fotografía es perfectamente reconocible aun cuando se han perdido los «detalles» de la imagen en la operación.

Como las columnas de $\mathbf{X}$ representan a bloques de la imagen, si interpretamos $\mathbf{v}_1$ como un «bloque» más (el que se obtiene reordenando los N elementos de $\mathbf{v}_1$ en una matriz $\sqrt{N} \times \sqrt{N}$), que llamaremos $\mathbf{V}_1$, podremos parafrasear lo anterior

[5]Es decir, al usar la fórmula (3.49) con $r = 1$.

Figura 3.3: Reconstrucción de «Lena» a partir de la fórmula (3.49) para distintos valores de r: (a) Imagen original, (b) $r = 1$, (c) $r = 5$, (d) $r = 10$.

diciendo que $\mathbf{V}_1$ representa a los bloques de la imagen en el sentido de que minimiza el error cuadrático medio. El salto al dominio de la frecuencia es ahora inmediato sin más que utilizar el «Teorema de Parseval»: dado que los bloques de las imágenes naturales tienen características «paso de baja» (cada bloque es el resultado de «enventanar» la imagen y la imagen es típicamente «paso de baja»), se deduce fácilmente aplicando este Teorema que $\mathbf{V}_1$ es un bloque «paso de baja» asimismo.

Sean $\mathbf{V}_2, \ldots, \mathbf{V}_N$ los bloques de imagen asociados a los autovectores $\mathbf{v}_2, \ldots, \mathbf{v}_N$. Cada uno de ellos aporta mayor nivel de «detalle» a la reconstrucción de la imagen.

Como los autovectores son ortogonales se tiene que

$$\sum_{n_1=1}^{\sqrt{N}} \sum_{n_2=1}^{\sqrt{N}} V_i(n_1, n_2)\, V_j(n_1, n_2) = 0$$

si $i \neq j$, donde $V_i(n_1, n_2)$ es el elemento (n_1, n_2) de la matriz $\mathbf{V}_i$. En el dominio de Fourier, la relación anterior se escribe usando el «Teorema de Parseval» como

$$\frac{1}{N_1\, N_2} \sum_{k_1=1}^{\sqrt{N}} \sum_{k_2=1}^{\sqrt{N}} \mathbb{V}_i(k_1, k_2)\, \mathbb{V}_j(k_1, k_2) = 0$$

si $i \neq j$, donde hemos llamado $\mathbb{V}_i(n_1, n_2)$ a los elementos de la Transformada Discreta de Fourier en dos dimensiones de la matriz $\mathbf{V}_i$. Esta ecuación también se interpreta como una relación de «ortogonalidad» entre las transformadas. Teniendo en cuenta además que:

1. los autovectores reconstruyen *toda* la imagen (las componentes de baja, *media* y *alta* frecuencia) por constituir una base ortonormal del espacio

2. los autovectores reconstruyen los «detalles» (alta frecuencia) de la imagen a medida que el índice r crece en (3.49) o, dicho de otra forma, las componentes de baja frecuencia se concentran en los autovectores,

se reconoce que las transformadas $\mathbb{V}_i(n_1, n_2)$ *van desplazando su soporte desde la baja hasta la alta frecuencia a medida que* i *crece desde* 1 *hasta* N (lo que recuerda los bancos de filtros de la Transformada «Wavelet»). Por ejemplo, la Figura 3.4 muestra el módulo de la Transformada de Fourier bidimensional de los bloques $\mathbf{V}_1, \ldots, \mathbf{V}_4$ correspondientes a la imagen «Lena». Se aprecia que $\mathbf{V}_1$ –Figura 3.4 (a)– es claramente «paso de baja» mientras que $\mathbf{V}_2, \mathbf{V}_3, \mathbf{V}_4$ cubren la zona de baja frecuencia del espectro. Para completar, la Figura 3.5 muestra la transformada de los bloques $\mathbf{V}_1, \mathbf{V}_{22}, \mathbf{V}_{43}, \mathbf{V}_{64}$. Se observa cómo éstos van ocupando progresivamente la alta frecuencia a medida que el índice i crece. Se obtienen resultados comparables para todas las imágenes naturales con las que hemos experimentado.

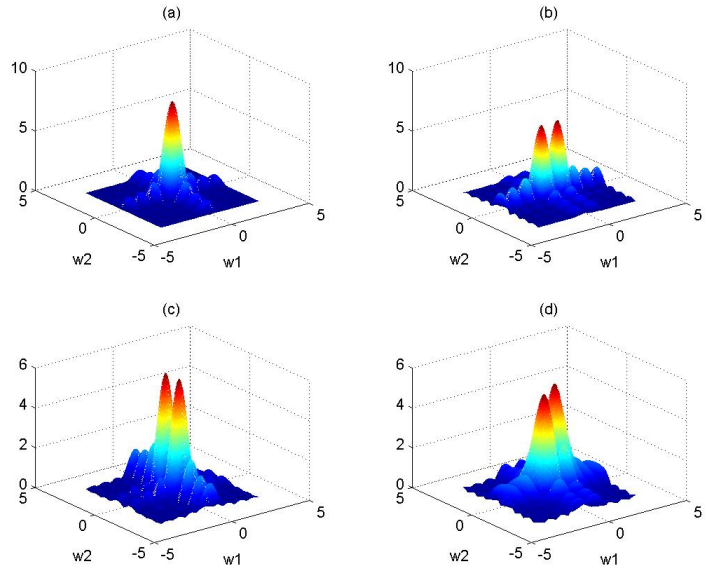

Figura 3.4: De izquierda a derecha y de arriba abajo, módulo de la Transformada de Fourier bidimensional de $\mathbf{V}_1, \mathbf{V}_2, \mathbf{V}_3, \mathbf{V}_4$.

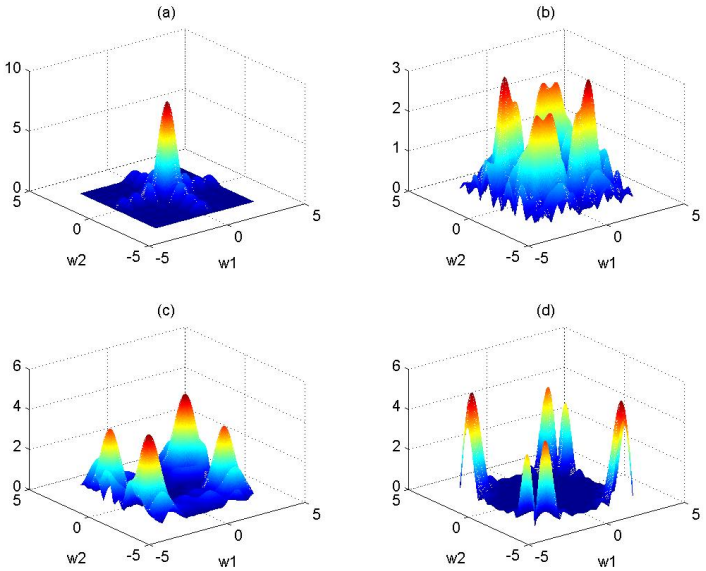

Figura 3.5: De izquierda a derecha y de arriba abajo, módulo de la Transformada de Fourier bidimensional de $\mathbf{V}_1, \mathbf{V}_{22}, \mathbf{V}_{43}, \mathbf{V}_{64}$.

3.9.2. Obtención de las componentes independientes de una imagen natural mediante un proceso de «filtrado – muestreo»

El hecho de que la mayoría de los autovalores de la matriz de correlación de una imagen natural sean muy pequeños hace que en la práctica (3.45) pueda ser aproximada por (3.47). En este caso $\mathbf{b}_{1:}$ va a resultar ser una combinación lineal de los autovectores asociados a los autovalores más pequeños de $\mathbf{R}_x$: es decir, contemplada como bloque, $\mathbf{b}_{1:}$ tiene características «paso de alta». Este hecho ha sido observado en todos nuestros experimentos con imágenes naturales (ver siguiente Capítulo). El tratar ahora las filas de la matriz de separación como filtros en dos dimensiones, abre una nueva interpretación de ICA cuando es aplicado a imágenes naturales. En particular, en esta Sección demostraremos que podemos obtener cada una de las componentes independientes de una imagen natural filtrándola con el correspondiente filtro ICA, convenientemente rotado, y realizando un muestreo apropiado del resultado.

Consideremos la imagen natural I, digitalizada y representada en escala de grises, a partir de la cual obtenemos la matriz $\mathbf{X}$ de observaciones. Sea $\mathbf{I}_k$ el k-ésimo bloque de la imagen I:

$$\mathbf{I}_k = \begin{bmatrix} i_{k,11} & i_{k,12} & \cdots & i_{k,1\sqrt{N}} \\ i_{k,21} & i_{k,22} & \cdots & i_{k,2\sqrt{N}} \\ \vdots & \vdots & \ddots & \vdots \\ i_{k,\sqrt{N}1} & i_{k,\sqrt{N}2} & \cdots & i_{k,\sqrt{N}\sqrt{N}} \end{bmatrix}_{(\sqrt{N}\times\sqrt{N})} \tag{3.51}$$

Por conveniencia, trabajaremos con la secuencia en dos dimensiones dada por:

$$i_k(n_1, n_2) = i_{k,(n_1+1)(n_2+1)} \tag{3.52}$$

con $n_1, n_2 = 0, 1, ..., \sqrt{N}-1$. Obtendríamos la k-ésima columna de $\mathbf{X}$ de la siguiente forma:

$$\mathbf{x}_{:k} = [i_{k,11}, i_{k,21}, \ldots, i_{k,\sqrt{N}1}, i_{k,12}, i_{k,22}, \ldots, i_{k,\sqrt{N}2}, \ldots, i_{k,1\sqrt{N}}, i_{k,1\sqrt{N}}, \ldots, i_{k,\sqrt{N}\sqrt{N}}]^\dagger \tag{3.53}$$

esto es, el elemento (k_1, k_2) de $\mathbf{I}_k$ pasaría a ser el j-ésimo elemento de $\mathbf{x}_{:k}$, con $j = (k_2 - 1)\sqrt{N} + k_1$.

Sea $\mathbf{b}_{1:} = [b_{11}, \ldots, b_{1N}]$ la primera fila de la matriz de separación $\mathbf{B}$. Definimos la matriz $\mathbf{H} = (h_{n_1 n_2})$, de tamaño $\sqrt{N} \times \sqrt{N}$, como aquella compuesta por los elementos de $\mathbf{b}_{1:}$ reorganizados siguiendo el proceso inverso al utilizado para obtener $\mathbf{x}_{:k}$ a partir de $\mathbf{I}_k$, esto es:

$$\mathbf{H} = \begin{bmatrix} b_{11} & b_{1\,(\sqrt{N}+1)} & \cdots & b_{1\,(N-\sqrt{N}+1)} \\ b_{12} & b_{1\,(\sqrt{N}+2)} & \cdots & b_{1\,(N-\sqrt{N}+2)} \\ \vdots & \vdots & \ddots & \vdots \\ b_{1\,\sqrt{N}} & b_{1\,(2\sqrt{N})} & \cdots & b_{1N} \end{bmatrix}_{(\sqrt{N}\times\sqrt{N})} \tag{3.54}$$

es decir, el elemento $h_{k_1 k_2}$ de $\mathbf{H}$ es el elemento j-ésimo del vector $\mathbf{b}_{1:}$, con $j = (k_2 - 1)\sqrt{N} + k_1$. También en este caso, en lugar de trabajar con esta matriz, lo haremos con la secuencia en dos dimensiones definida por:

$$h(n_1, n_2) = b_{1(n_2\sqrt{N}+n_1+1)} \tag{3.55}$$

con $n_1, n_2 = 0, 1, ..., \sqrt{N} - 1$. Hemos demostrado que esta secuencia representa un filtro paso de alta, y recibe el nombre de *filtro ICA*.

Sea $\mathbf{y}_{1:} = [y_{11}, \ldots, y_{1T}]$ el vector fila obtenido como

$$\mathbf{y}_{1:} = \mathbf{b}_{1:}\,\mathbf{X} \tag{3.56}$$

Cada elemento y_{1k}, $k = 1, 2, ..., T$, del vector $\mathbf{y}_{1:}$ puede obtenerse como

$$y_{1k} = \sum_{i=1}^{N} b_{1i}\, x_{ik} \tag{3.57}$$

o bien

$$y_{1k} = \sum_{n_1=0}^{\sqrt{N}-1} \sum_{n_2=0}^{\sqrt{N}-1} h(n_1, n_2)\, i_k(n_1, n_2) \tag{3.58}$$

es decir, el resultado de multiplicar cada elemento del k-ésimo bloque de la imagen $\mathbf{I}_k$ por el correspondiente elemento de la matriz $\mathbf{H}$, sumando posteriormente todos los productos.

Sea $\mathbf{H}^{\curvearrowleft}$ la matriz $\mathbf{H}$ rotada 180° en el sentido contrario al de las agujas del reloj, esto es:

$$\mathbf{H}^{\curvearrowleft} = \begin{bmatrix} b_{1N} & b_{1\,(N-\sqrt{N})} & \cdots & b_{1\,\sqrt{N}} \\ b_{1(N-1)} & b_{1\,(N-\sqrt{N}-1)} & \cdots & b_{1\,\sqrt{N}-1} \\ \vdots & \vdots & \ddots & \vdots \\ b_{1\,(N-\sqrt{N}+1)} & b_{1\,(N-2\sqrt{N}+1)} & \cdots & b_{11} \end{bmatrix}_{(\sqrt{N}\times\sqrt{N})} \tag{3.59}$$

o, simplemente:

$$h^{\curvearrowleft}(n_1, n_2) = h(\sqrt{N} - 1 - n_1, \sqrt{N} - 1 - n_2) \tag{3.60}$$

donde $h^{\curvearrowleft}(n_1, n_2)$ es el elemento $(n_1 + 1, n_2 + 1)$ de la matriz $\mathbf{H}^{\curvearrowleft}$, con $n_1, n_2 = 0, 1, ..., \sqrt{N} - 1$. Teniendo en cuenta (3.55), $h^{\curvearrowleft}(n_1, n_2) = b_{1\{(\sqrt{N}-n_2)\sqrt{N}-n_1\}}$. Por otro lado, el filtro representado por la secuencia $h^{\curvearrowleft}(n_1, n_2)$ también es paso de alta, como $h(n_1, n_2)$, ya que la rotación sólo afecta a la fase.

Sea $z(n_1, n_2)$ el resultado de filtrar en dos dimensiones la secuencia $i_k(n_1, n_2)$, con el filtro dado por $h^{\curvearrowleft}(n_1, n_2)$:

$$\begin{aligned} z(n_1, n_2) &= i_k(n_1, n_2) \star h^{\curvearrowleft}(n_1, n_2) \\ &= \sum_{k_1=-\infty}^{\infty} \sum_{k_2=-\infty}^{\infty} i_k(k_1, k_2)\, h^{\curvearrowleft}(n_1 - k_1, n_2 - k_2) \\ &= \sum_{k_1=0}^{\sqrt{N}-1} \sum_{k_2=0}^{\sqrt{N}-1} i_k(k_1, k_2)\, h^{\curvearrowleft}(n_1 - k_1, n_2 - k_2) \end{aligned} \tag{3.61}$$

donde el símbolo «$\star$» representa la convolución en dos dimensiones. Teniendo en cuenta (3.58) y la relación (3.60), el elemento $z(\sqrt{N} - 1, \sqrt{N} - 1)$ vendrá dado por:

$$\begin{aligned} z(\sqrt{N} - 1, \sqrt{N} - 1) &= \sum_{k_1=0}^{\sqrt{N}-1} \sum_{k_2=0}^{\sqrt{N}-1} i_k(k_1, k_2)\, h^{\curvearrowleft}(\sqrt{N} - 1 - k_1, \sqrt{N} - 1 - k_2) \\ &= \sum_{k_1=0}^{\sqrt{N}-1} \sum_{k_2=0}^{\sqrt{N}-1} i_k(k_1, k_2)\, h(k_1, k_2) \\ &= y_{1k} \end{aligned} \tag{3.62}$$

Extendiendo este desarrollo a todos los elementos de $\mathbf{y}_{1:}$, concluimos que podemos obtener la primera componente independiente mediante el filtrado paso de alta de la imagen con un filtro obtenido a partir de la primera fila de la matriz $\mathbf{B}$ de separación. En la Figura 3.6 esquematizamos el proceso descrito.

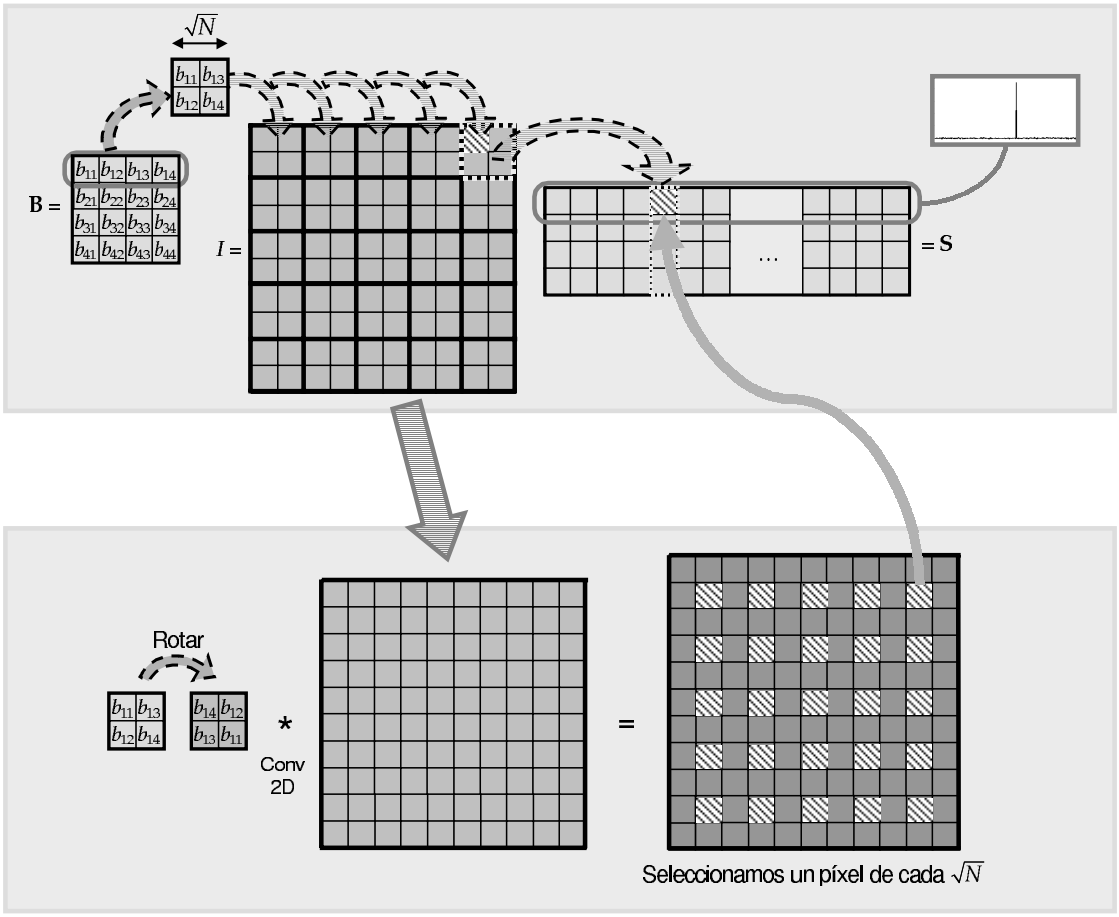

Figura 3.6: Obtención de las componentes independientes mediante el producto $\mathbf{B}\,\mathbf{X}$ (arriba) y mediante filtrado y muestreo (abajo).

Capítulo 4

Resultados experimentales

Hemos demostrado que debido a la distribución «dispersa» o «rala» de las componentes independientes de las imágenes naturales, condicionada por las características de los filtros ICA, las bases ICA correspondientes deben ser similares a fragmentos de la propia imagen. El objetivo de este Capítulo es mostrar que los experimentos confirman todo lo expuesto en el Capítulo 3.

En este Capítulo presentamos dos experimentos. En el primero mostramos los resultados obtenidos al aplicar ICA a dos imágenes naturales, representadas en escala de grises, eligiendo para ello un tamaño de bloque de 16×16 píxeles. A pesar de que estas imágenes son muy distintas entre sí, comprobamos que los resultados obtenidos son similares, verificando en todo caso las conclusiones obtenidas en el Capítulo 3.

En el segundo experimento analizamos los resultados al aplicar ICA a una imagen natural representada en escala de grises, pero ahora eligiendo un tamaño de bloque de 2×2 píxeles. Compararemos los resultados obtenidos con los del experimento anterior y mostraremos posibles aplicaciones de los mismos.

4.1. Presentación de los experimentos

Sea una imagen I genérica, natural y representada en escala de grises, de tamaño $n_1 \times n_2$ píxeles. La dividimos en T bloques de $\sqrt{N} \times \sqrt{N}$ y componemos la

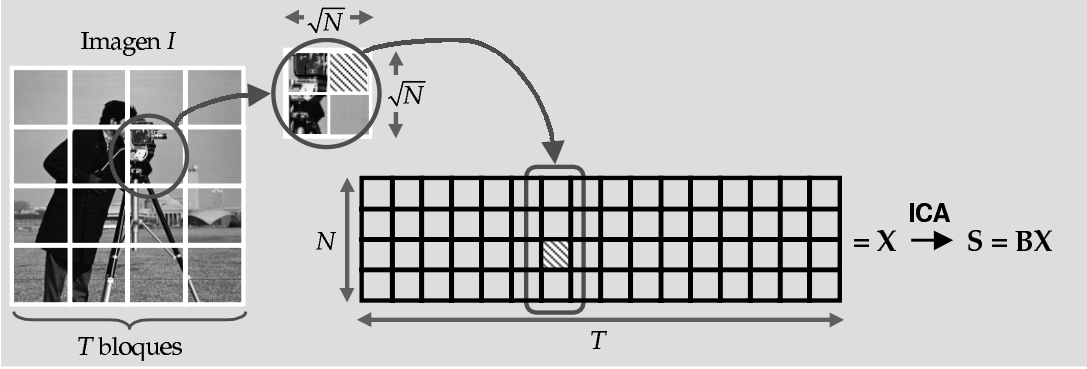

Figura 4.1: Tras componer la matriz **X** de observaciones, aplicamos ICA para obtener la matriz **S** de componentes independientes, la matriz **A** de bases ICA y la matriz **B** de filtros ICA.

matriz **X** ($N \times T$) de observaciones según se explicó en la Sección **??** del Capítulo 3, página **??**. Para comodidad del lector, reproducimos en la Figura 4.1 el esquema empleado en dicha Sección. Al aplicar ICA a esta imagen, obtendremos:

- **Matriz de componentes independientes, S**. Mostraremos en distintos experimentos que la mayoría de las componentes independientes de una imagen presenta una distribución «dispersa», tal y como se demostró en el Capítulo 3. Estas componentes independientes serán tanto más «dispersas» cuanto mayores sean los bloques en los que dividimos la imagen.

- **Matriz de filtros ICA, B**. De la misma forma, demostraremos que la mayoría de los filtros ICA (filas de **B**) son paso de alta, que es lo que condiciona la distribución de las componentes independientes.

- **Matriz de bases ICA, A**. Debido a la distribución dispersa de las componentes independientes, comprobaremos que las bases ICA resultan ser prácticamente idénticas a fragmentos de la propia imagen.

En todos los casos, los experimentos presentados han sido realizados con el algoritmo FastICA [HyvOja97, FastICA], configurando los parámetros opcionales de la siguiente forma:

- **Aproximación por deflación**. De esta forma forzamos que las componentes independientes sean extraídas una a una. El algoritmo impone que cada

componente obtenida sea ortogonal a la anterior, por lo que sólo la primera componente independiente cumplirá los criterios descritos en el Capítulo 3.

- **Coeficiente de asimetría como no linealidad**. Tomando la curtosis se obtienen resultados similares, pero supone mayor carga computacional.

4.2. Experimento 1

Consideremos las imágenes naturales, representadas en escala de grises, mostradas en las Figuras 4.2 y 4.3. Hemos elegido estas imágenes por poseer características muy distintas entre sí, tanto en las formas que se encuentran presentes como en la iluminación. Las dividimos en bloques de 16×16 y aplicamos el algoritmo FastICA. A continuación analizamos los resultados obtenidos.

Figura 4.2: Exp. 1: Imagen «Reloj» (464×336).

Figura 4.3: Exp. 1: Imagen «Saltamontes» (256×512).

4.2.1. Exp. 1: Filtros ICA

Comprobamos que, tal y como demostramos, los filtros ICA (filas de **B** organizadas en matrices) resultan ser **paso de alta**. Sin embargo, como ya hemos comentado, debido a que el algoritmo FastICA impone que cada componente independiente extraída debe ser ortogonal a la anterior, los filtros ICA son «cada vez menos paso de alta». En concreto, la componente de continua de dichos filtros va creciendo, permitiendo el paso de componentes de baja frecuencia. En las Figuras 4.4 y 4.5 representamos las magnitudes de las transformadas de Fourier de los primeros y últimos nueve filtros ICA obtenidos para la imagen «Reloj». Los filtros correspondientes a la imagen «Saltamontes» se muestran en las Figuras 4.6 y 4.7.

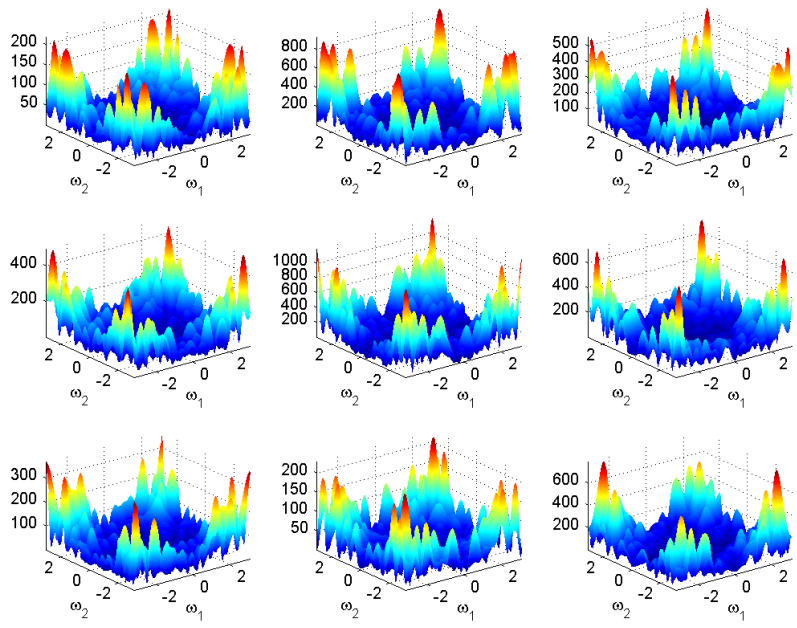

Figura 4.4: Exp. 1: Filtros ICA 16×16 correspondientes a las primeras componentes independientes extraídas para la imagen «Reloj». Son filtros claramente paso de alta.

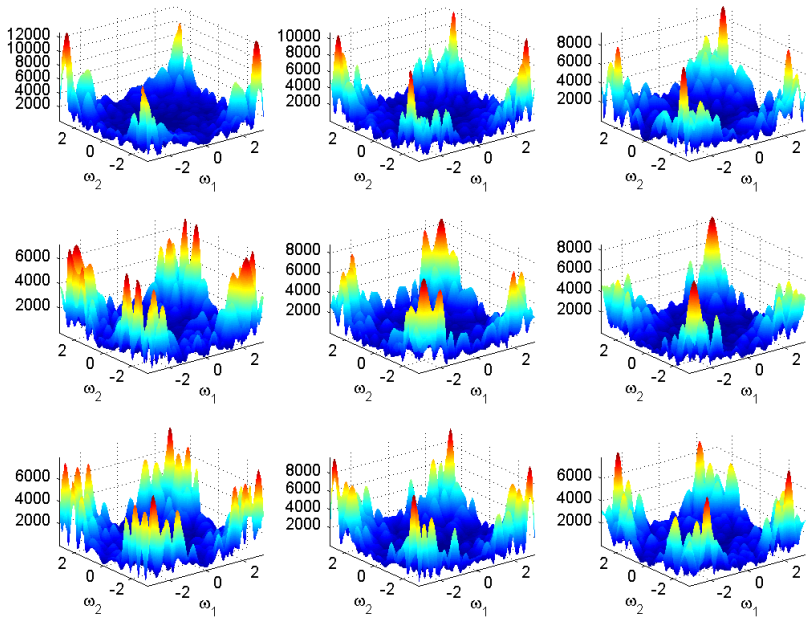

Figura 4.5: Exp. 1: Filtros ICA 16×16 correspondientes a las últimas componentes independientes extraídas para la imagen «Reloj». En este caso, la elevada componente de continua permite el paso de las bajas frecuencias.

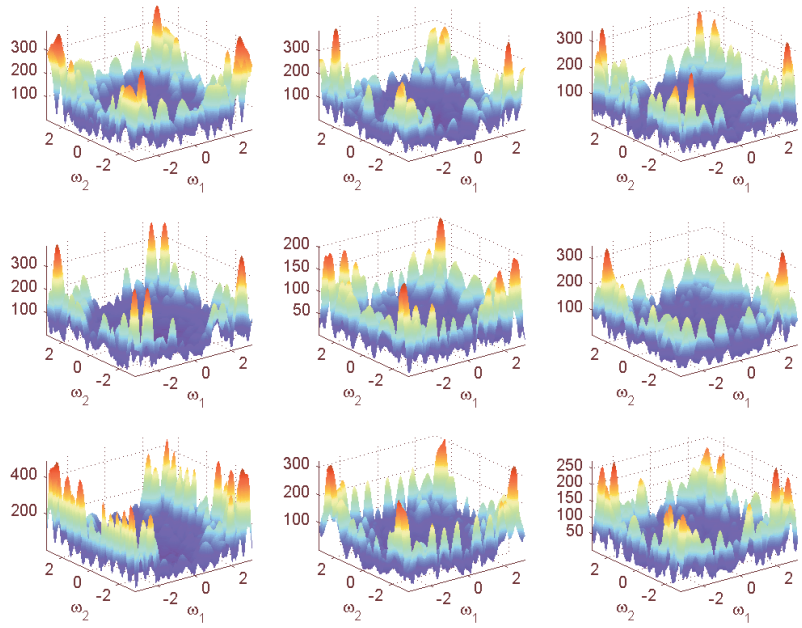

Figura 4.6: Exp. 1: Filtros ICA 16×16 correspondientes a las primeras componentes independientes extraídas para la imagen «Saltamontes», todos claramente paso de alta.

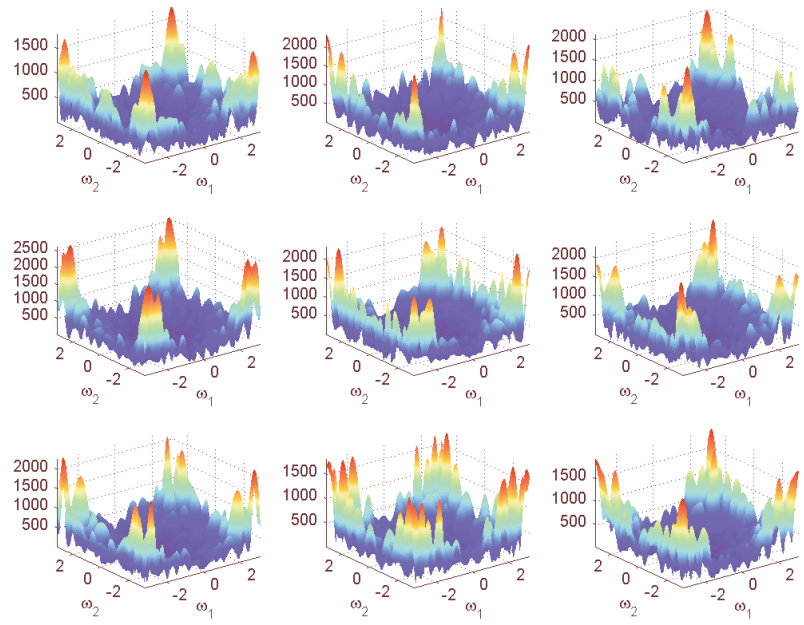

Figura 4.7: Exp. 1: Filtros ICA 16×16 correspondientes a las últimas componentes independientes extraídas para la imagen «Saltamontes». La elevada componente de continua permite el paso de las bajas frecuencias.

4.2.2. Exp.1: Componentes independientes

Hemos demostrado que, debido a las características de los filtros ICA, las componentes independientes de imágenes naturales presentan una distribución «dispersa». Sin embargo, igual que en el caso de los filtros ICA, a medida que se van obteniendo componentes independientes, esta distribución resulta ser cada vez menos «dispersa». En efecto, si analizamos las primeras componentes independientes obtenidas para la imagen «Reloj», Figura 4.8, y la imagen «Saltamontes», Figura 4.9, en ambos casos observamos una marcada distribución «dispersa». En concreto, tienden hacia las soluciones ideales, es decir, aquellas en las que sólo un elemento es distinto de cero e igual a $\sqrt[p]{T}$, con $p = 3$, al estar trabajando con el coeficiente de asimetría como no linealidad, y T el número de muestras de cada componente independiente e igual al número de bloques en los que se divide la imagen.

Las últimas componentes independientes extraídas, sin embargo, presentan una distribución mucho menos dispersa, tal y como se muestra en las Figuras 4.10 y 4.11.

Por otro lado, comprobamos que la primera componente independiente extraída

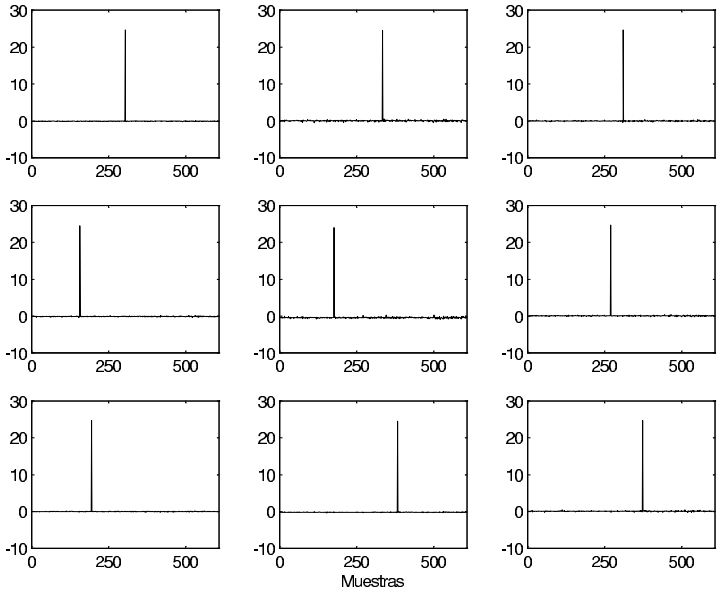

Figura 4.8: Exp. 1: Componentes independientes extraídas en primer lugar para la imagen «Reloj» cuando ésta ha sido dividida en bloques de 16×16 píxeles para su análisis en componentes independientes. Tienden a la solución ideal, esto es, todas las muestras igual a cero salvo una.

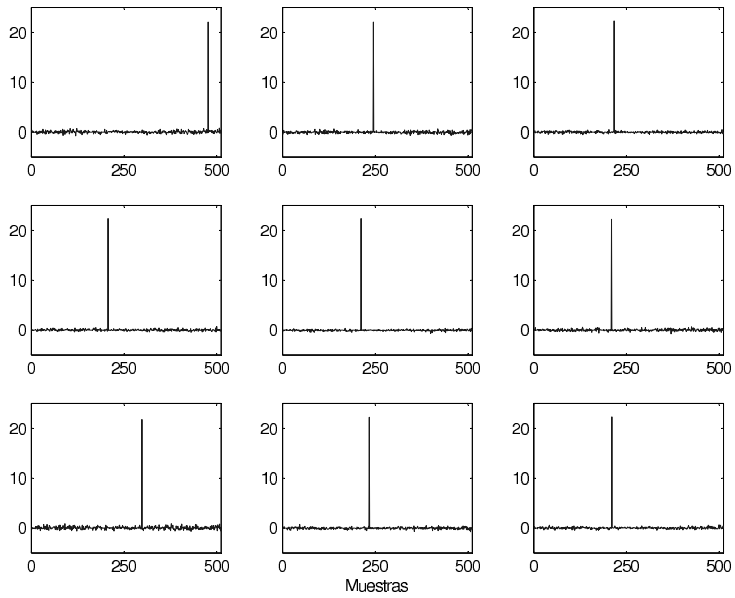

Figura 4.9: Exp. 1: Componentes independientes extraídas en primer lugar para la imagen «Saltamontes» cuando ésta ha sido dividida en bloques de 16×16 píxeles para su análisis en componentes independientes. Tienden a la solución ideal, esto es, todas las muestras igual a cero salvo una.

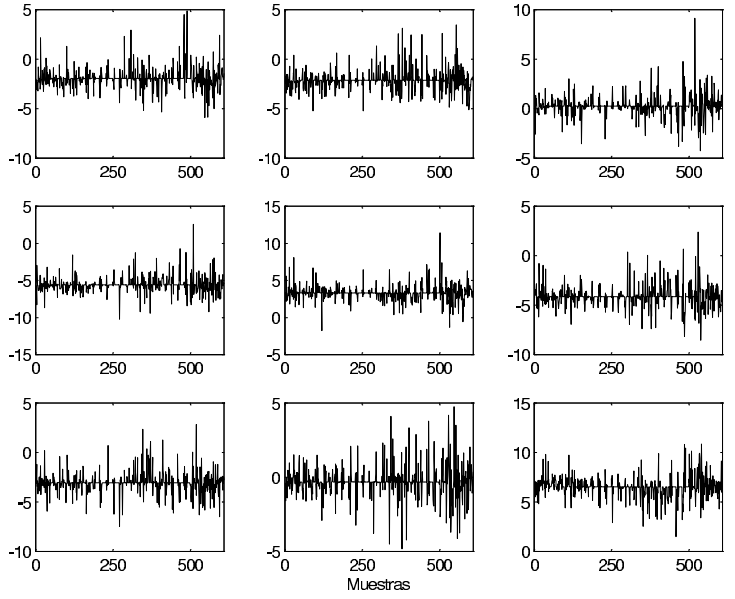

Figura 4.10: Exp. 1: Componentes independientes extraídas en último lugar para la imagen «Reloj», cuando ésta ha sido divida en bloques de 16×16 píxeles para su análisis en componentes independientes.

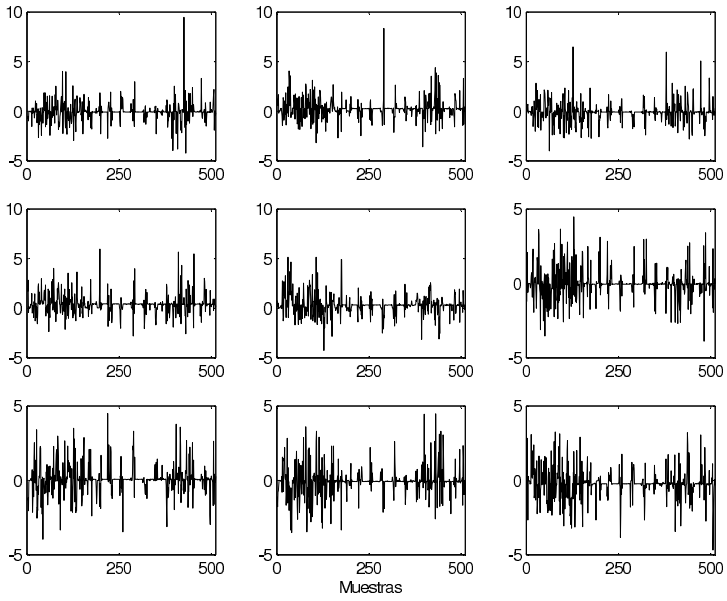

Figura 4.11: Exp. 1: Componentes independientes extraídas en último lugar para la imagen «Saltamontes», cuando ésta ha sido divida en bloques de 16×16 píxeles para su análisis en componentes independientes.

en ambos ejemplos verifican la solución de segunda clase descrita en el Capítulo 3, página 41, fórmula (3.31), y que, por conveniencia, repetimos a continuación:

$$\frac{\sum_{k=1}^{T} x_{1k}\, y_{1k}^{p-1}}{\sum_{k=1}^{T} x_{1k}\, y_{1k}} = \ldots = \frac{\sum_{k=1}^{T} x_{Nk}\, y_{1k}^{p-1}}{\sum_{k=1}^{T} x_{Nk}\, y_{1k}} \tag{4.1}$$

donde T es el número de bloques en los que dividimos la imagen, x_{jk} e y_{jk}, $j = 1, 2, ..., N$, $k = 1, 2, ..., T$, son, respectivamente, el elemento k-ésimo de la fila j-ésima de la matriz $\mathbf{X}$ de observaciones y de la matriz $\mathbf{S}$ de componentes independientes, y $p = 3$ en el caso que nos ocupa (hemos elegido el coeficiente de asimetría para la obtención de las componentes independientes). Como ya hemos comentado, debido a la imposición de ortogonalidad entre componentes independientes que lleva a cabo el algoritmo FastICA, es la primera de ellas la que estrictamente verifica el criterio anterior. En el Cuadro 4.1 se muestran la media y varianza de todos los cocientes de (4.1) para la primera de las fuentes extraídas y cada una de las imágenes consideradas.

En el Capítulo 3 mostramos que cada componente independiente puede obtenerse tras un proceso de «filtrado - muestreo» de la imagen original, utilizando

Imagen	«Reloj»	«Saltamontes»
Media	24.55917	20.9974
Varianza	$4.7351\dot{1}0^{-12}$	$1.4741\dot{1}0^{-10}$

Cuadro 4.1: Exp. 1: Media y varianza de los cocientes de (4.1) para la primera componente independiente obtenida y cada una de las imágenes consideradas.

el correspondiente filtro ICA convenientemente rotado. En las Figuras 4.12 y 4.13 mostramos los resultados obtenidos tras filtrar las imágenes «Reloj» y «Saltamontes» con el primer y último filtro ICA en cada caso. En las mismas figuras representamos las posiciones de los puntos a muestrear para obtener una de las componentes independientes.

4.2.3. Exp. 1: Bases ICA

Las bases ICA (columnas de la matriz **A**) obtenidas para cada imagen se muestran en las Figuras 4.14 y 4.15. Como podemos observar, en ambos casos las bases son muy parecidas a fragmentos de la imagen original. Este parecido es más notorio en en el caso de la imagen «Reloj», entre cuyas bases podemos observar claramente los números y divisiones de la esfera. Este comportamiento, como ya demostramos, viene determinado por la distribución «dispersa» de las componentes independientes. Sin embargo, acabamos de ver que las últimas componentes independientes obtenidas por el algoritmo FastICA presentan una distribución menos «dispersa» que las primeras, lo que se traduce, como puede apreciarse en las Figuras 4.14 y 4.15, en que las últimas bases ICA ya no guardan tanta similitud con fragmentos de la imagen original.

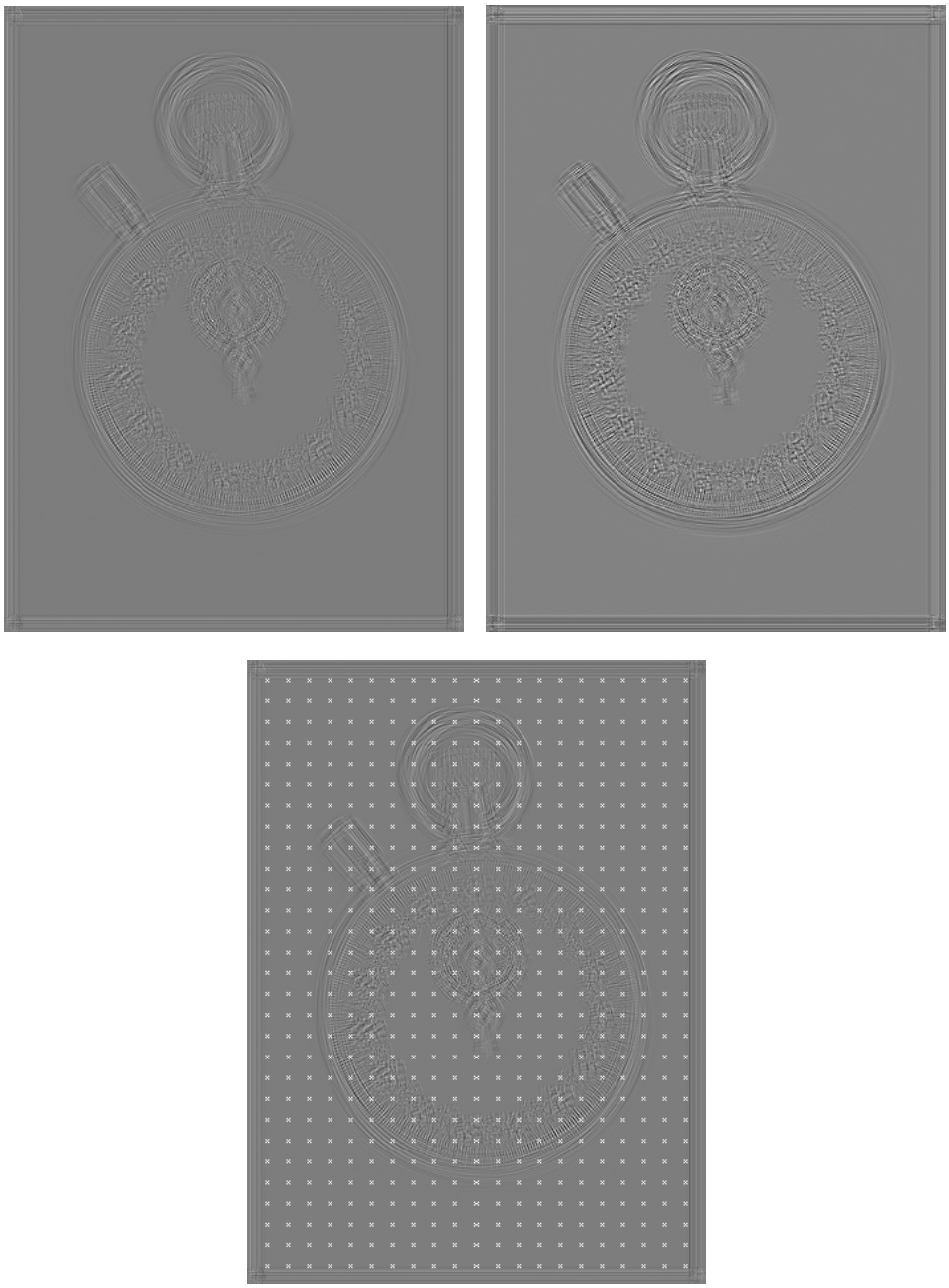

Figura 4.12: Exp. 1: Arriba: Imagen «Reloj» filtrada con el primer (izquierda) y último (derecha) filtro ICA obtenido. Debajo: Puntos (marcados con «×») recogidos por la primera componente independiente.

4.3. Experimento 2

Sea la imagen natural representada en escala de grises de la Figura 4.16. En este experimento analizaremos los resultados obtenidos al aplicar el algoritmo FastICA

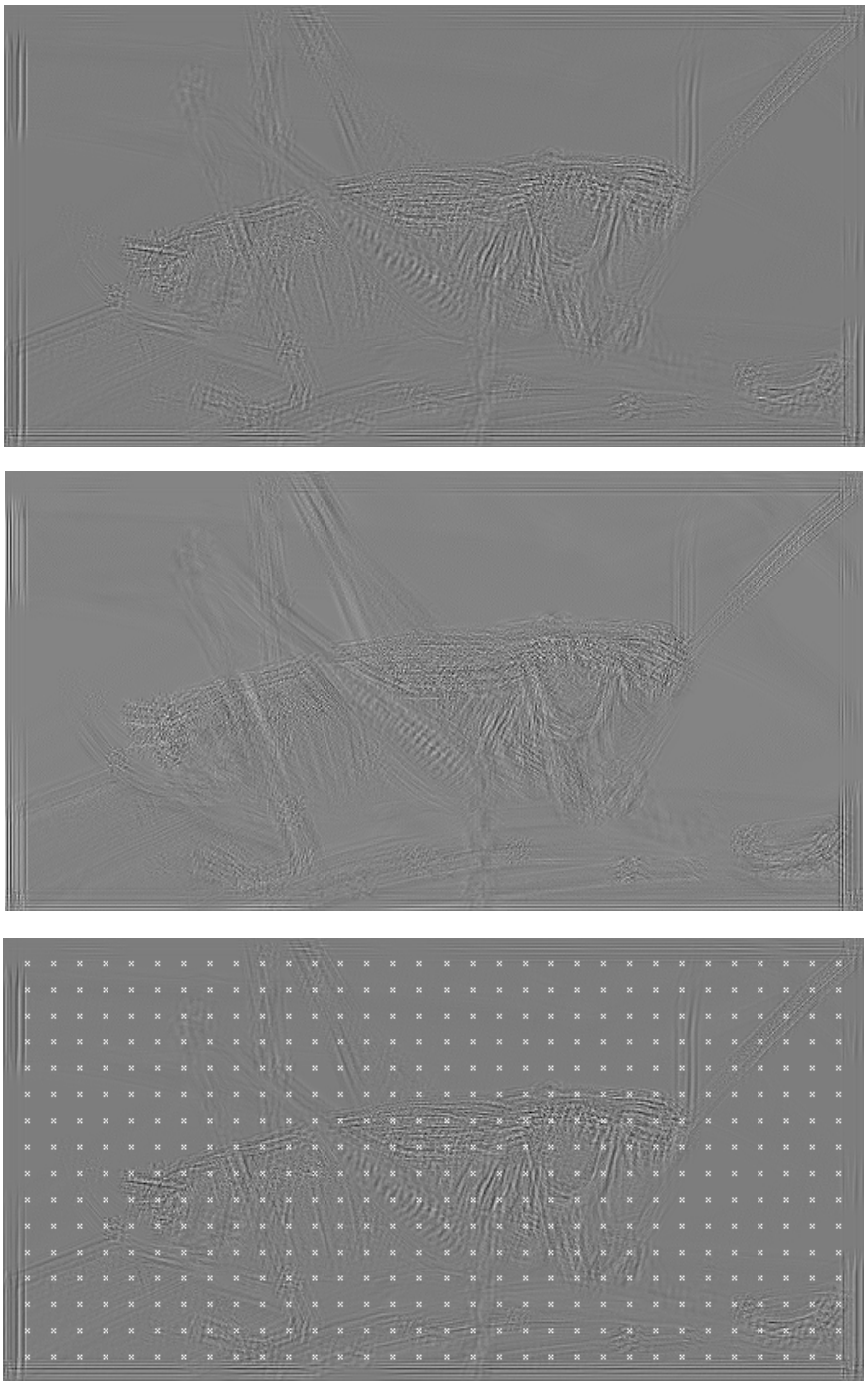

Figura 4.13: Exp. 1: De arriba a abajo, imagen «Saltamontes» filtrada con el primer y último filtro ICA obtenido, y puntos (marcados con «×») recogidos por la primera componente independiente.

a esta imagen cuando la dividimos en bloques con el mínimo tamaño posible, es decir, 2×2 píxeles.

4.3.1. Exp. 2: Filtros ICA

En este caso, de los cuatro filtros ICA, el único que no tiene un marcado carácter paso de alta es el cuarto, de nuevo debido a su componente de continua (Figura 4.17).

4.3.2. Exp. 2: Componentes independientes

En la Figura 4.18 mostramos las componentes independientes obtenidas en este experimento. Aquellas componentes correspondientes a los filtros ICA paso de alta (las tres primeras en este experimento) tienen una distribución «dispersa»,

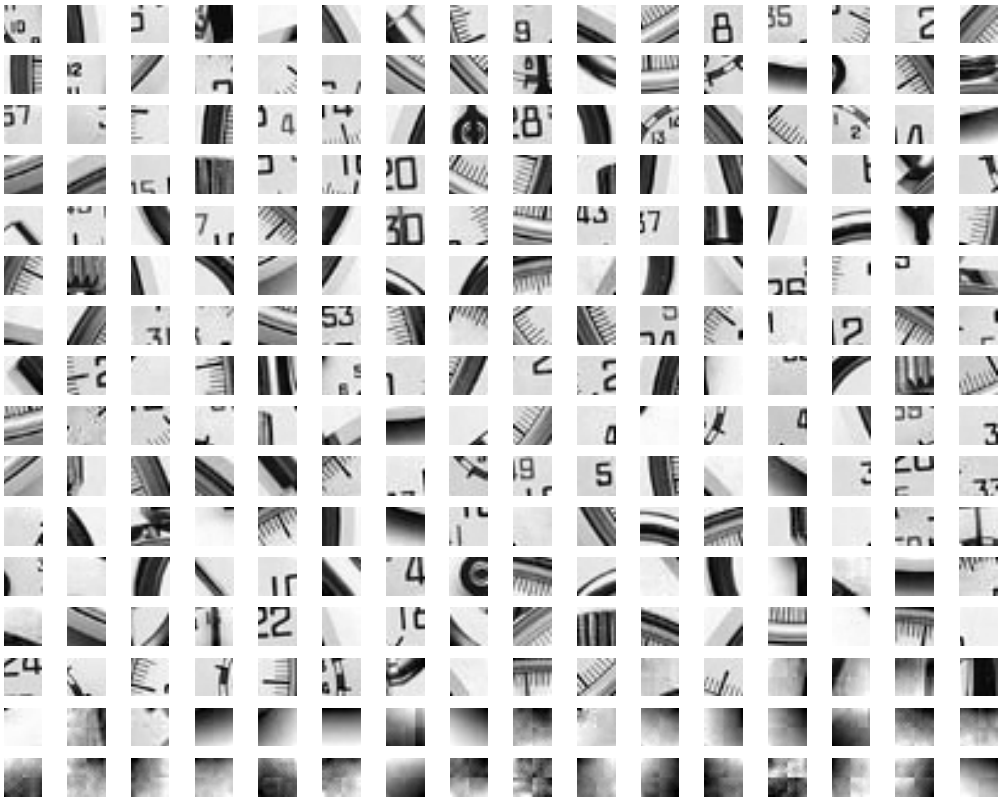

Figura 4.14: Exp. 1: Bases ICA 16×16 para la imagen «Reloj».

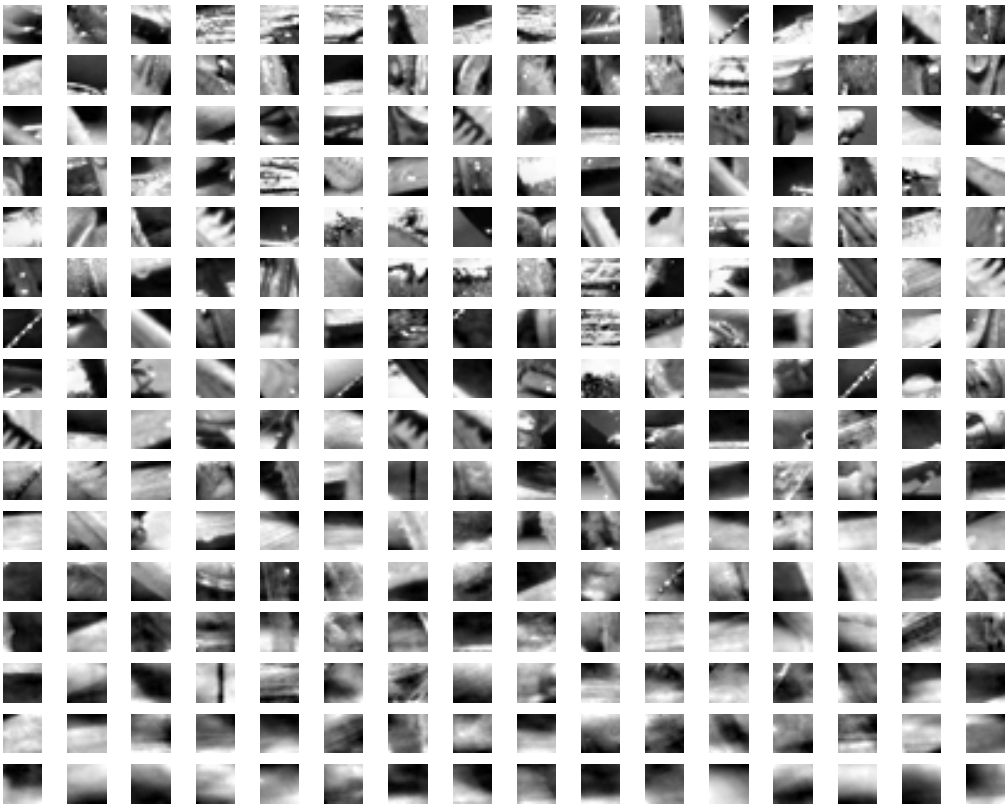

Figura 4.15: Exp. 1: Bases ICA 16×16 para la imagen «Saltamontes».

Figura 4.16: Exp. 2: Imagen «Lena», 256×256.

caracterizadas por un histograma de aspecto «picudo», mientras que la cuarta posee una distribución muy similar a la de la imagen original, aunque cambiada de signo. En la Figura 4.19 representamos los histogramas de esta última componente independiente y de la imagen con más detalle.

En este caso, volvemos a comprobar que la solución obtenida es de segunda clase, es decir, que se cumple (4.1) para la primera de las componentes extraídas (el proceso de ortogonalización de FastICA impide que se verifique para todas las componentes). En concreto, los cocientes de (4.1) toman, en media, el valor 1.2213, con una varianza de $1.9775 \cdot 10^{-7}$.

En la Figura 4.20 mostramos las imágenes obtenidas tras filtrar la imagen «Lena» con cada uno de los cuatro filtros ICA. En los tres primeros casos, al ser filtros paso de alta, obtenemos claramente los contornos de la imagen. Al usar con el cuarto de los filtros, obtenemos una versión suavizada y negativa de la imagen original.

Por último, indicar que estos resultados pueden ser empleados en aplicaciones como la detección de contornos [Hornillo04] o el marcado digital de imágenes[1]

[1] El lector puede encontrar amplia información sobre esta técnica en [HartKut99, LangSL00,

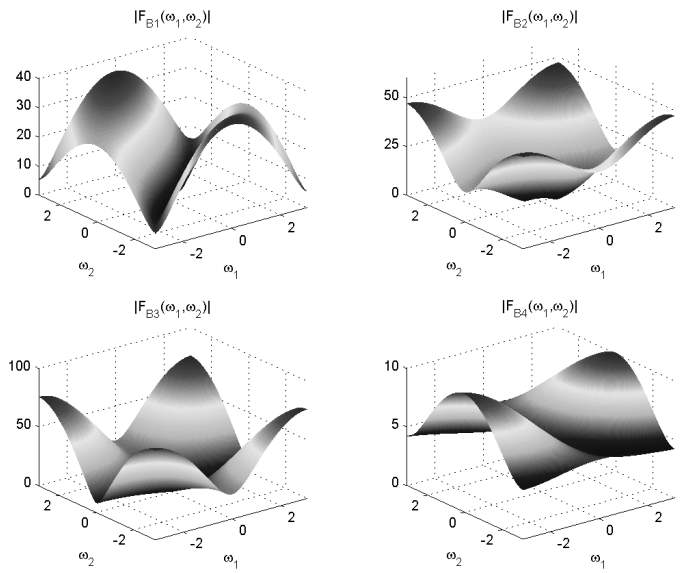

Figura 4.17: Exp. 2: Filtros ICA 2×2 para la imagen «Lena».

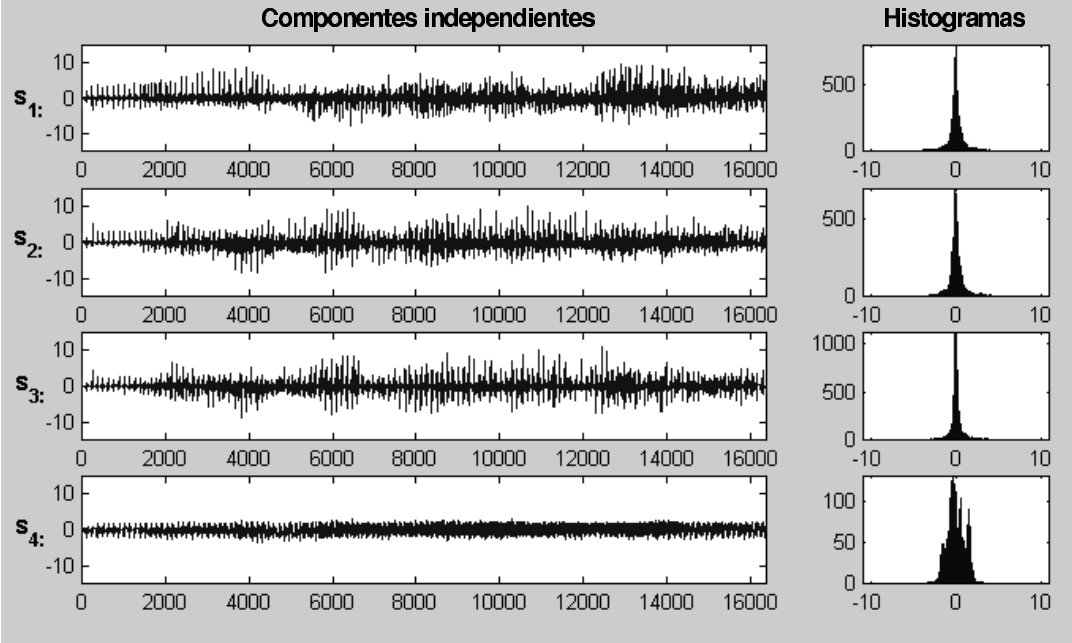

Figura 4.18: Exp. 2: Componentes independientes de la imagen «Lena», cuando ésta ha sido dividida en bloques de 2×2 píxeles para su análisis en componentes independientes. Todas las componentes tienen una distribución claramente «dispersa», excepto la última, cuyo histograma es similar al de la imagen original (en este caso, cambiado de signo).

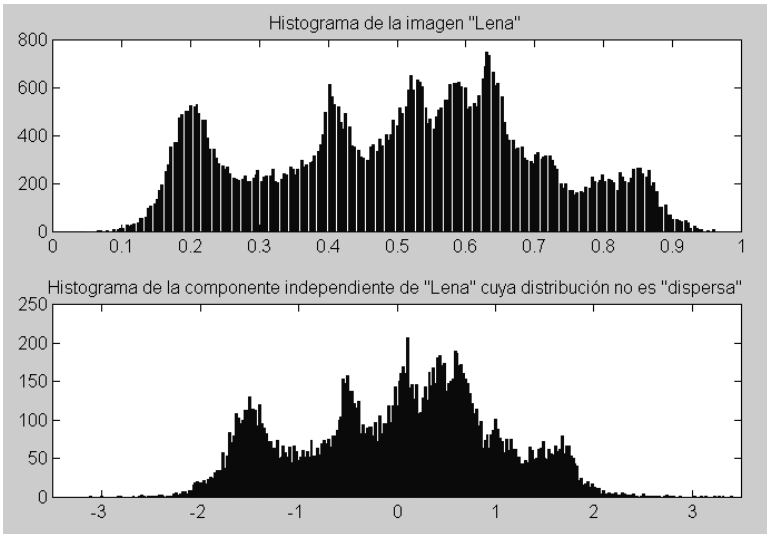

Figura 4.19: Exp. 2: Comparación del histograma de la imagen original y el de la componente independiente (cambiada de signo) cuya distribución no es «dispersa».

[Hornillo03].

Petit00, PodilD01] y en la web [WMorg]

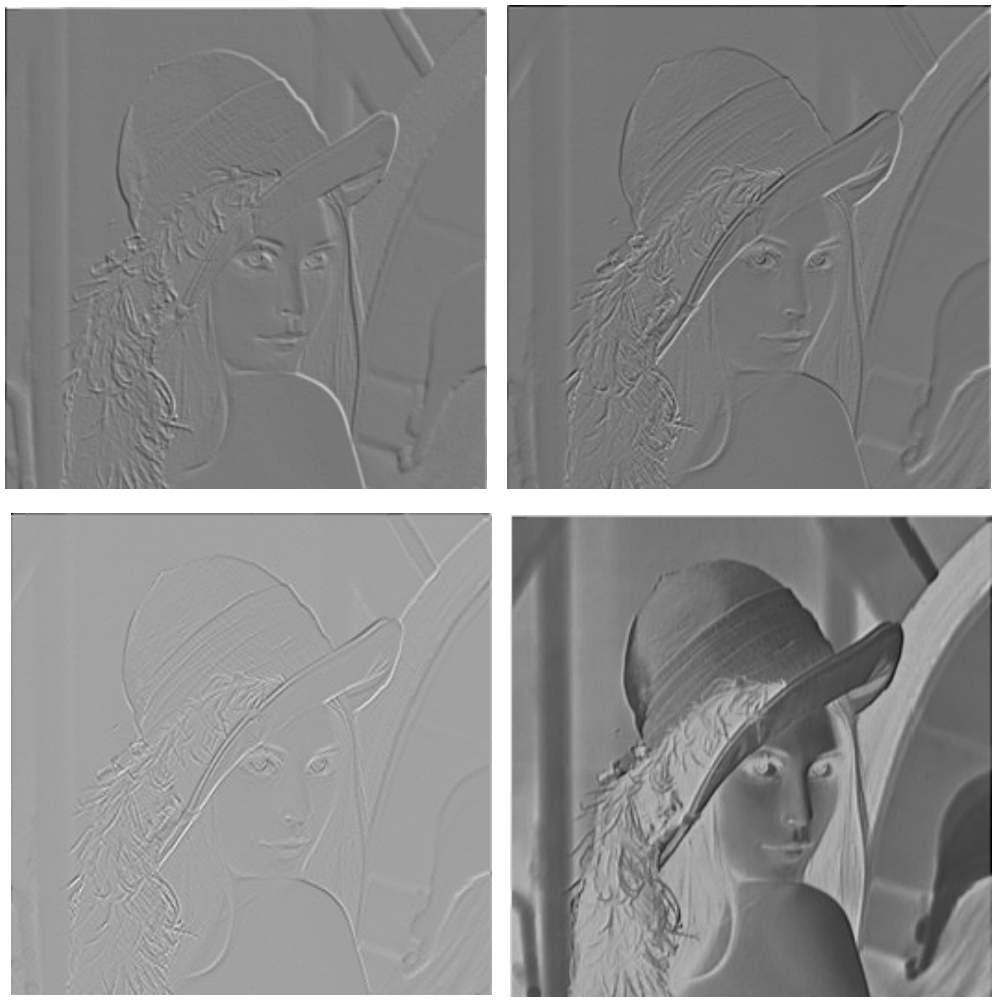

Figura 4.20: Exp. 2: Resultados obtenidas al filtrar la imagen «Lena» con los cuatro filtros ICA obtenidos en este experimento.

4.3.3. Exp. 2: Bases ICA

Las bases ICA obtenidas en este experimento se muestran en la Figura 4.21. En este caso ya no podemos decir que haya un claro parecido entre estas bases y bloques de la imagen original, debido al tamaño de bloque usado.

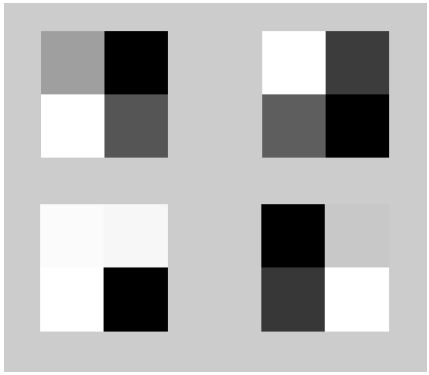

Figura 4.21: Exp. 2: Bases ICA 2×2 para la imagen «Lena».

Conclusiones y líneas futuras de investigación

El análisis en componentes independientes (conocido abreviadamente como ICA) de imágenes naturales despertó un gran interés cuando se mostró su conexión con el comportamiento de ciertas neuronas de la corteza visual primaria. En concreto, las «bases ICA» de imágenes naturales muestran el aspecto de «bordes», lo que se ha relacionado con el hecho de que las neuronas simples de la corteza visual primaria alcanzan su máxima respuesta en presencia de estímulos visuales consistentes en contornos con una determinada orientación. Por otro lado, las componentes independientes de imágenes naturales presentan una distribución «dispersa», coincidiendo con el comportamiento observado para las respuestas de las citadas neuronas simples.

En esta Tesis se muestra, por primera vez, una teoría matemática que explica por qué los resultados obtenidos al aplicar ICA a imágenes naturales presentan unas características tan especiales. A continuación se enumeran las conclusiones derivadas del estudio realizado:

1. Se han analizado las componentes independientes obtenidas por los algoritmos ICA basados en maximizar estadísticos de orden superior. La novedad es que este análisis no se ha realizado desde el punto de vista de la independencia estadística de las componentes, como es lo habitual, sino que se ha centrado en su estructura y la de las ecuaciones a partir de las cuales son calculadas. Como resultado, se concluye que estas componentes independientes sólo pueden ser de dos clases, las cuales se han denominado de «Clase 1» y

de «Clase 2».

2. En la práctica, cuando se realiza el análisis en componentes independientes de imágenes naturales, sólo se van a obtener componentes de «Clase 2» debido a que las soluciones de «Clase 1» exigen que la matriz **B** de separación sea perpendicular a todas las columnas de la matriz **X** de observaciones, esto es, a todos los bloques en los que ha sido dividida la imagen original. De ser así, todas las columnas de **X** (bloques de la imagen) estarían en el mismo plano, algo poco probable para imágenes naturales. Además, las soluciones de «Clase 1» no cumplen la propiedad necesaria de que las componentes independientes deben tener media cero.

3. Las soluciones obtenidas por el algoritmo FastICA cuando es aplicado a imágenes naturales siempre son de «Clase 2». La condición para una solución de «Clase 1» es que la matriz **X** de observaciones sea invertible y, por tanto, cuadrada, algo muy poco probable en la práctica.

4. Los «filtros ICA» (filas de la matriz **B** de separación, reorganizados en matrices) se obtienen a partir de la combinación lineal de los autovectores asociados a los autovalores más pequeños de la matriz de correlación de las observaciones. En el caso de imágenes naturales se deduce que estos filtros son **paso de alta**.

5. El hecho de que los «filtros ICA» de imágenes naturales sean paso de alta implica que sus componentes independientes van a tener una distribución «dispersa» y que las «bases ICA» se correspondan con los bordes de la imagen.

Apéndice A

Blanqueado

Previamente a la estimación de la matriz de mezclas del modelo ICA:

$$\mathbf{x} = \mathbf{A}\,\mathbf{s} \tag{A.1}$$

las observaciones dadas por el vector de media cero $\mathbf{x}$ suelen someterse a un proceso de *blanqueado*. Este proceso consiste en aplicar una transformación lineal $\mathbf{W}$ a las observaciones de tal forma que obtengamos un nuevo vector de variables

$$\mathbf{z} = \mathbf{W}\,\mathbf{x} \tag{A.2}$$

cuyas componentes no estén correlacionadas entre sí y que además tengan varianza unidad. En otras palabras, estamos haciendo que $\mathrm{E}\{\mathbf{z}\mathbf{z}^T\} = \mathbf{I}$, donde $\mathbf{I}$ es la matriz identidad. Una forma de obtener esta matriz de blanqueado $\mathbf{W}$ es a partir de la descomposición en autovalores y autovectores de la matriz de correlación de $\mathbf{x}$:

$$\mathbf{R}_\mathbf{x} = \mathrm{E}\{\mathbf{x}\mathbf{x}^T\} = \mathbf{V}\mathbf{D}\mathbf{V}^T \tag{A.3}$$

donde $\mathbf{V} = [\mathbf{v}_1|...|\mathbf{v}_n]$ es la matriz ortogonal (lo que quiere decir que $\mathbf{V}\,\mathbf{V}^T = \mathbf{V}^T\,\mathbf{V} = \mathbf{I}$, donde $\mathbf{I}$ es la matriz unidad) que contiene, por columnas, los autovectores de $\mathbf{R}_\mathbf{x}$ y $\mathbf{D} = \mathrm{diag}(\lambda_1, ..., \lambda_n)$ es la matriz diagonal de sus autovalores. Así, podemos conseguir una matriz de blanqueado mediante el producto:

$$\mathbf{W} = \mathbf{D}^{-1/2}\mathbf{V}^T \tag{A.4}$$

donde $\mathbf{D}^{-1/2} = \text{diag}(\lambda_1^{-1/2}, ..., \lambda_n^{-1/2})$. Para ver si la matriz $\mathbf{W}$ es una matriz de blanqueado, basta con comprobar que la matriz de correlación de $\mathbf{z}$ es igual a la matriz unidad:

$$\mathbf{R_z} = \mathbf{W}\text{E}\{\mathbf{x}\mathbf{x}^T\}\mathbf{W}^T = \mathbf{D}^{-1/2}\mathbf{V}^T\mathbf{V}\mathbf{D}\mathbf{V}^T\mathbf{V}\mathbf{D}^{-1/2} = \mathbf{I} \tag{A.5}$$

Para llegar a este resultado hemos tenido en cuenta que la matriz $\mathbf{V}$ es ortogonal, es decir, que $\mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I}$.

La matriz de blanqueado dada por $\mathbf{W}$ no es la única posible. Para cualquier matriz ortogonal, $\mathbf{U}$, es fácil comprobar que la matriz dada por $\mathbf{UW}$ también es una matriz de blanqueado. Esto es debido a que, para $\mathbf{z} = \mathbf{UWx}$:

$$\mathbf{R_z} = \mathbf{UW}\text{E}\{\mathbf{x}\mathbf{x}^T\}\mathbf{W}^T\mathbf{U}^T = \mathbf{U}\mathbf{I}\mathbf{U}^T = \mathbf{I} \tag{A.6}$$

Un ejemplo interesante es la matriz $\mathbf{V}\mathbf{D}^{-1/2}\mathbf{V}^T$, llamada «raíz cuadrada inversa» de $\mathbf{R_x}$ y denotada por $\mathbf{R_x}^{-1/2}$.

Realizando este proceso de blanqueado antes de aplicar análisis en componentes independientes, obtenemos una nueva matriz de mezclas $\tilde{\mathbf{A}}$:

$$\mathbf{z} = \mathbf{WAs} = \tilde{\mathbf{A}}\mathbf{s} \tag{A.7}$$

que resulta ser ortogonal:

$$\text{E}\{\mathbf{z}\mathbf{z}^T\} = \tilde{\mathbf{A}}\text{E}\{\mathbf{s}\mathbf{s}^T\}\tilde{\mathbf{A}}^T = \tilde{\mathbf{A}}\tilde{\mathbf{A}}^T = \mathbf{I} \tag{A.8}$$

De esta forma, al restringir nuestra búsqueda de la matriz de mezclas al conjunto de las matrices ortogonales, en lugar de estimar los n^2 elementos de la matriz $\mathbf{A}$, necesitamos estimar sólo $n(n-1)/2$, disminuyendo la carga computacional [HyvKO01].

Apéndice B

Función de densidad de probabilidad de una transformación

Supongamos que $\mathbf{x}$ e $\mathbf{y}$ son vectores aleatorios de dimensión n y que están relacionados mediante la transformación:

$$\mathbf{y} = \mathbf{g}(\mathbf{x}) \tag{B.1}$$

para la cual la transformación inversa:

$$\mathbf{x} = \mathbf{g}^{-1}(\mathbf{y}) \tag{B.2}$$

existe y es única. Puede demostrarse [Papoul91] que la función de densidad de probabilidad $p_{\mathbf{y}}(\mathbf{y})$ de $\mathbf{y}$ se obtiene a partir de la densidad de probabilidad $p_{\mathbf{x}}(\mathbf{x})$ de $\mathbf{x}$ de la siguiente forma:

$$p_{\mathbf{y}}(\mathbf{y}) = \frac{1}{\mid \det J_{\mathbf{g}}\left(\mathbf{g}^{-1}(\mathbf{y})\right) \mid} p_{\mathbf{x}}(\mathbf{g}^{-1}(\mathbf{y})) \tag{B.3}$$

donde $J_{\mathbf{g}}$ es la matriz jacobiana:

$$\begin{pmatrix} \frac{\partial g_1(\mathbf{x})}{\partial x_1} & \frac{\partial g_2(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial g_n(\mathbf{x})}{\partial x_1} \\ \frac{\partial g_1(\mathbf{x})}{\partial x_2} & \frac{\partial g_2(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial g_n(\mathbf{x})}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_1(\mathbf{x})}{\partial x_n} & \frac{\partial g_2(\mathbf{x})}{\partial x_n} & \cdots & \frac{\partial g_n(\mathbf{x})}{\partial x_n} \end{pmatrix} \tag{B.4}$$

y $\mathbf{g}_j(\mathbf{x})$ es la componente j-ésima de la función vectorial $\mathbf{g}(\mathbf{x})$.

En el caso particular de que la transformación (B.1) sea lineal y no singular, de tal forma que $\mathbf{y} = \mathbf{A}\mathbf{x}$ y $\mathbf{x} = \mathbf{A}^{-1}\mathbf{y}$, la fórmula (B.3) se simplifica de la siguiente forma:

$$p_{\mathbf{y}}(\mathbf{y}) = \frac{1}{|\det \mathbf{A}|}\, p_{\mathbf{x}}(\mathbf{A}^{-1}\mathbf{y}) \tag{B.5}$$

Apéndice C

El teorema central del límite

Sea

$$x_k = \sum_i = 1^k z_i \tag{C.1}$$

la suma parcial de una secuencia $\{z_i\}$ de variables aleatorias independientes e idénticamente distribuidas, z_i. Debido a que la media y la varianza de x_k puede crecer sin límite a medida que k tiende a infinito, consideremos las variables normalizadas:

$$y_k = \frac{x_k - m_{x_k}}{\sigma_{x_k}} \tag{C.2}$$

donde m_{x_k} y σ_{x_k} son la media y la desviación típica de x_k, respectivamente.

Puede demostrarse [Papoul91] que la distribución de y_k converge a una distribución gaussiana de media cero y varianza unidad cuando $k \to \infty$. Este resultado es conocido como el *teorema central del límite* y es la razón primordial por la que muchos fenómenos aleatorios se modelan como variables gaussianas (por ejemplo, un ruido aditivo, que suele ser considerado como la suma de un gran número de señales de distintos orígenes).

El teorema central del límite puede generalizarse a variables aleatorias vectoriales $\mathbf{z}_i$ independientes e idénticamente distribuidas con una media común $\mathbf{m}_z$ y una matriz de covarianzas igual a $\mathbf{C}_z$. La distribución conjunta de la variable aleatoria vectorial:

$$\mathbf{y}_k = \frac{1}{\sqrt{k}} \sum_{i=1}^{k} (\mathbf{z}_i - \mathbf{m}_\mathbf{z}) \tag{C.3}$$

es gaussiana con media cero y matriz de covarianzas $\mathbf{C}_z$.

Apéndice D

Cumulantes y momentos

Sea x una variable aleatoria escalar, continua, real y de media cero, y cuya función de densidad de probabilidad es $p_x(x)$.

La *primera función característica* $\varphi(\omega)$ de x se define como la transformada de Fourier de $p_x(x)$:

$$\varphi(\omega) = E\{\exp(j\omega x)\} = \int_{-\infty}^{+\infty} p_x(x) \exp(j\omega x) dx \tag{D.1}$$

donde $j = \sqrt{-1}$ y ω es la variable transformada correspondiente a x. Cada distribución de probabilidad está especificada de forma única por su función característica y viceversa [Papoul91]. Expandiendo la función característica $\varphi(\omega)$ en su serie de Taylor, obtenemos [Nandi99, Papoul91]:

$$\varphi(\omega) = \int_{-\infty}^{+\infty} \left(\sum_{k=0}^{\infty} \frac{x^k (j\omega)^k}{k!} \right) p_x(x) dx = \sum_{k=0}^{\infty} E\{x^k\} \frac{(j\omega)^k}{k!} \tag{D.2}$$

Los coeficientes de esta expansión, $E\{x^k\}$, son los momentos de x (suponiendo que existen). Por esta razón, la función característica $\varphi(\omega)$ también recibe el nombre de *función generadora de momentos*.

A menudo es útil el uso de la segunda función característica $\phi(\omega)$ de x, o *función generadora de cumulantes*, y que viene dada por el logaritmo neperiano de la primera función característica:

$$\phi(\omega) = \ln(\varphi(\omega)) = \ln(E\{\exp(j\omega x)\}) \tag{D.3}$$

Por su parte, los cumulantes κ_k de x se definen a partir del desarrollo en serie de Taylor de la segunda función característica:

$$\phi(\omega) = \sum_{k=0}^{\infty} \kappa_k \frac{(j\omega)^k}{k!} \tag{D.4}$$

donde el k-ésimo cumulante se obtiene a partir de la siguiente derivada:

$$\kappa_k = (-j)^k \left. \frac{d^k \phi(\omega)}{d\omega^k} \right|_{\omega=0} \tag{D.5}$$

Para una variable aleatoria x de media cero, los primeros cuatro cumulantes son [Nandi99]:

$$\begin{aligned} \kappa_1 &= 0 \\ \kappa_2 &= E\{x^2\} \\ \kappa_3 &= E\{x^3\} \\ \kappa_4 &= E\{x^4\} - 3\left[E\{x^2\}\right]^2 \end{aligned} \tag{D.6}$$

Como vemos, los primeros tres cumulantes son idénticos a los tres primeros momentos (para una variable aleatoria x de media cero), y el cuarto cumulante κ_4 se conoce con el nombre de *kurtosis*.

Si la variable aleatoria x no tuviese media cero [HyvKO01]:

$$\begin{aligned} \kappa_1 &= E\{x\} \\ \kappa_2 &= E\{x^2\} - \left[E\{x\}\right]^2 \\ \kappa_3 &= E\{x^3\} - 3E\{x^2\}E\{x\} + 2\left[E\{x\}\right]^3 \\ \kappa_4 &= E\{x^4\} - 3\left[E\{x^2\}\right]^2 - 4E\{x^3\}E\{x\} + 12E\{x^2\}\left[E\{x\}\right]^2 - 6\left[E\{x\}\right]^4 \end{aligned} \tag{D.7}$$

Consideremos ahora brevemente el caso multivariable. Sea $\mathbf{x}$ una variable aleatoria vectorial y $p_{\mathbf{x}}(\mathbf{x})$ su función de densidad de probabilidad. La función característica de $\mathbf{x}$ vuelve a ser la transformada de Fourier de su función de densidad de probabilidad:

$$\varphi(\omega) = E\{\exp(j\omega\mathbf{x})\} = \int_{-\infty}^{+\infty} p_{\mathbf{x}}(\mathbf{x}) \exp(j\omega x) d\mathbf{x} \tag{D.8}$$

donde ω es ahora un vector de igual dimensión que $\mathbf{x}$, y la integral se extiende a todas sus componentes. Los momentos y cumulantes de $\mathbf{x}$ se obtienen como los

coeficientes del desarrollo en serie de Taylor de la primera y la segunda función característica, respectivamente, de forma similar al caso escalar. En el caso multivariable, los cumulantes suelen llamarse *cumulantes cruzados*, por analogía con las covarianzas cruzadas.

Puede demostrarse que los cumulantes de segundo, tercer y cuarto orden, para un vector aleatorio $\mathbf{x}$ de media cero, son [Nandi99]:

$$\begin{aligned} \text{cum}\,(x_i, x_j) &= E\{x_i\, x_j\} \\ \text{cum}\,(x_i, x_j, x_k) &= E\{x_i\, x_j\, x_k\} \\ \text{cum}\,(x_i, x_j, x_k, x_l) &= E\{x_i\, x_j\, x_k\, x_l\} - E\{x_i\, x_j\}E\{x_k\, x_l\} - \\ &\quad E\{x_i\, x_k\}E\{x_j\, x_l\} - E\{x_i\, x_l\}E\{x_j\, x_k\} \end{aligned} \tag{D.9}$$

La importancia de los momentos y cumulantes reside en que nos permiten caracterizar completamente un proceso aleatorio sin necesidad de conocer su función de densidad de probabilidad, salvo en algunas excepciones como es el caso de la distribución log-normal [Nandi99]. Aunque momentos y cumulantes contienen básicamente la misma información estadística sobre una variable aleatoria, se suele preferir trabajar con los cumulantes porque verifican las siguientes propiedades, no cumplidas por los momentos [Nandi99]:

P1 Los cumulantes de orden superior (mayor que dos) de un proceso aleatorio gaussiano valen exactamente cero. Por ello, a menudo se toma el cumulante como la distancia entre el proceso aleatorio para el que se ha calculado y un proceso estocástico de una distribución gaussiana.

P2 Si dos o más conjuntos de variables aleatorias $\{x_1, x_2, ..., x_K\}$ y $\{v_1, v_2, ..., v_K\}$ son estadísticamente independientes, entonces el cumulante de orden n-ésimo de la variable aleatoria $y_i = x_i + v_i$, $i = 1, 2, ..., K$, es igual a la suma de los cumulantes de orden n-ésimo de ambos conjuntos por separado.

Apéndice E

Información y entropía

Sea X un suceso que puede presentarse con probabilidad $P(X)$. Cuando X tiene lugar, decimos que hemos recibido

$$I(X) = \log\left(\frac{1}{P(X)}\right) \tag{E.1}$$

unidades de información. La elección de la base del logaritmo que interviene en la anterior definición equivale a elegir una determinada unidad. Si el logaritmo usado es de base 2, la unidad correspondiente es el bit [Abram86].

E.1. Fuente de información de memoria nula

Sea una fuente de información discreta que emite una secuencia de símbolos que pertenecen a un alfabeto finito y fijo, $S = \{s_1, s_2, ..., s_q\}$. Los símbolos emitidos se eligen de acuerdo con una ley fija de probabilidad.

La fuente más sencilla es aquella que emite los símbolos estadísticamente independientes entre sí, y diremos que es una *fuente de memoria nula*. La cantidad media de información por símbolo de la fuente recibe el nombre de *entropía* de la fuente, S. Para el caso de una fuente de memoria nula [Ash65, Reza94]:

$$H(S) = -\sum_{S} P(s_i)\log(P(s_i)) \tag{E.2}$$

E.2. Fuente de información de Markov

La fuente de memoria nula resulta demasiado limitada en algunas aplicaciones. Un tipo de fuente de información más general es aquella en que la presencia de un determinado símbolo si depende de un número finito m de símbolos precedentes. Este tipo de fuente recibe el nombre de *fuente de Markov* de orden m, y viene definida por su alfabeto, S, y el conjunto de probabilidades condicionales [Abram86]:

$$P(s_i|s_{j_1}, s_{j_2}, ..., s_{j_m}) \tag{E.3}$$

para $i = 1, 2, ...$ y $j_p = 1, 2, ...$

En una fuente de Markov de orden m, la probabilidad de un símbolo cualquiera viene determinada por los m símbolos que lo preceden. En cualquier momento, por lo tanto, definiremos el *estado* de la fuente de Markov de orden m por los m símbolos precedentes. Puesto que existen q símbolos distintos, una fuente de Markov de orden m admitirá q^m estados posibles. Al emitir la fuente nuevos símbolos, el estado cambia.

En el estudio de las fuentes de información de Markov de orden m nos limitaremos a considerar las denominadas *fuentes ergódicas*. Una fuente ergódica es aquella que, observada durante un tiempo suficientemente largo, emite (con probabilidad 1) una secuencia «típica» de símbolos [Abram86].

El problema de calcular las probabilidades de estado de una fuente ergódica de Markov a partir de las probabilidades condicionales de la fuente es una tarea complicada. Sin embargo, estas probabilidades de estado también pueden calcularse a partir de las probabilidades condicionales de los símbolos [Abram86], como veremos a continuación.

La información media suministrada por una fuente de Markov de orden m (en adelante omitiremos la palabra ergódico al hablar de tales fuentes) puede calcularse de la forma siguiente: si nos encontramos en el estado definido por $(s_{j_1}, s_{j_2}, ..., s_{j_m})$ (es decir, los m símbolos emitidos anteriormente fueron $s_{j_1}, s_{j_2}, ..., s_{j_m}$), la probabilidad condicional de recibir el símbolo s_i es $P(s_i|s_{j_1}, s_{j_2}, ..., s_{j_m})$. La información obtenida si s_i se presenta cuando estamos en el estado $(s_{j_1}, s_{j_2}, ..., s_{j_m})$, según (E.1),

es:

$$I(s_i|s_{j_1}, s_{j_2}, ..., s_{j_m}) = \log \frac{1}{P(s_i|s_{j_1}, s_{j_2}, ..., s_{j_m})} \tag{E.4}$$

Por lo tanto, la cantidad media de información por símbolo cuando nos encontramos en el estado $(s_{j_1}, s_{j_2}, ..., s_{j_m})$ viene dada por la ecuación:

$$H(s_i|s_{j_1}, s_{j_2}, ..., s_{j_m}) = \sum_S P(s_i|s_{j_1}, s_{j_2}, ..., s_{j_m})\ I(s_i|s_{j_1}, s_{j_2}, ..., s_{j_m}) \tag{E.5}$$

La cantidad media de información o *entropía* de la fuente de Markov de orden m, se tendrá calculando el valor medio de esta cantidad, extendida a los q^m estados posibles:

$$H(S) = \sum_{S^m} P(s_{j_1}, s_{j_2}, ..., s_{j_m})\ H(s_i|s_{j_1}, s_{j_2}, ..., s_{j_m}) \tag{E.6}$$

donde S^m es el espacio de los símbolos $(s_{j_1}, s_{j_2}, ..., s_{j_m})$.

Debido a que $H(S)$, dada por (E.2), nos da la cantidad de información media por símbolo de una fuente, suponiendo que no conocemos el símbolo de salida, la diferencia entre $H(S)$ y $H(s_i|s_{j_1}, s_{j_2}, ..., s_{j_m})$ es la cantidad de información media recibida a partir de la observación de un único símbolo de salida. Esta diferencia, denotada por $I(s_1, s_2, ..., s_m)$ y llamada *información mutua* de $s_1, s_2, ..., s_m$, es:

$$I(s_1, s_2, ..., s_m) = H(S) - H(s_i|s_{j_1}, s_{j_2}, ..., s_{j_m}) \tag{E.7}$$

Bibliografía

[Abram86] Abramson, N.: *Teoría de la información y codificación*. Paraninfo (1986)

[Ash65] Ash, R. B.: *Information theory*. Dover Publications (1965)

[Barlow61] Barlow, H. B.: *Sensory communication. Possible principles underlying the transformation of sensory messages*. MIT Press (1961) 217–234

[Barlow89] Barlow, H. B.: *Unsupervised learning*. Neural Computation, vol. 1 (1989) 295–311

[Barlow01] Barlow, H. B.: *Redundancy reduction revisited*. Network: Computation in Neural Systems, vol. 12 (2001) 241–253

[BellSej95] Bell, A. J., Sejnowski, T. J.: *The independent component of natural images are edge filters*. Vision Research, vol. 37, no. 23 (1995) 3327–3338

[BelSej95b] Bell, A. J., Sejnowski, T. J.: *An information maximisation approach to blind separation and blind deconvolution*. Neural Computation, vol. 7, no. 6 (1995) 1129–1159

[CaywWT04] Caywood, M., Willmore, B., Tolhurst, D.: *Independent Components of Color Natural Scenes Resemble V1 Neurons in Their Spatial and Color Tuning*. Journal of Neurophysiology, vol. 91 (2004) 2859–2873

[CaoLiu96] Cao, X., Liu, R.: *General approach to blind source separation*. IEEE Transactions on Signal Processing, vol. 44, no. 3 (1996) 562–571

[CichAm02] Cichocki, A., Amari, S. I.: *Adaptive blind signal and image processing*. John Wiley & Sons (2002)

[Comon94] Comon, P.: *Independent component analysis, a new concept?*. Signal Processing, vol. 36, no. 3 (1994) 287–314

[DerWeb04] Derrington, A. M., Webb, B. S.: *Visual System: How Is the Retina Wired up to the Cortex?*. Current Biology, vol. 14 (2004) R14–R15

[Field87] Field, D. J.: *Relations between the statistics of natural images and the response properties of cortical cells*. Journal of the Optical Society of America A, vol. 4 (1987) 2379–2394

[Field94] Field, D. J.: *What is the goal of sensory coding?*. Neural Computation, vol. 6 (1994) 559–601

[Fried87] Friedman, J.H.: *Exploratory projection pursuit*. Journal of the American Statistical Association, 82 (1987) 249–266

[GolLoan96] Golub, G., Van Loan, C.: *Matrix computations*. The Johns Hopkins University Press (1996)

[GonzWo92] González, R. C., Woods, R. E.: *Digital image processing*. Addison-Wesley (1992)

[GuyHal00] Guyton, A. C., Hall, J. E.: *Textbook of medical physiology*. Saunders (2000)

[HancBS92] Hancock P. J. B., Baddeley R. J., Smith L.S.: *The principle components of natural images*. Network, vol. 3 (1992) 61–72

[HartKut99] Hartung, F., Kutter, M.: *Multimedia watermarking technique*. Proceedings of IEEE, vol. 87, no. 7 (1999) 1079–1107

[HerJA85] Hérault, J., Jutten, C., Ans, B.: *Détection de grandeurs primitives dans un message composite par une architecture de calcul neuromimétique en apprentissage non supervisé*. 10éme Colloque GRETSI, Nice, Francia (1985) 1017–1022

[Hornillo04] Hornillo-Mellado, S., Martín-Clemente, R., Puntonet, C. G., Acha, J. I.: *Application of independent component analysis to edge detection*. Proc. World Automation Congress (WAC 2004). Sevilla, Spain (2004)

[Hornillo03] Hornillo-Mellado, S., Martín-Clemente, R., Acha, J. I., Puntonet, C. G.: *Application of independent component analysis to edge detection and watermarking*. Lecture Notes in Computer Science, vol. 2 (2003) 273–280

[HyvGH05] A. Hyvärinen, M. Gutmann and P.O. Hoyer. *Statistical model of natural stimuli predicts edge-like pooling of spatial frequency channels in V2*. BMC Neuroscience, (2005) 6–12

[HyvHoy00] Hyvärinen, A., Hoyer, P. O.: *Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces*. Neural Computation 2000, vol. 12, no. 7 (2000) 1705–1720

[HyvHH03] Hyvärinen, A., Hoyer, P. O., Hurri, J.: *Extensions of ICA as models of natural images and visual processing*. Proceedings of 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003), Nara, Japón (abril 2003) 963–974

[HyvKO01] Hyvärinen, A., Karhunen, J., Oja, E.: *Independent component analysis*. John Wiley & Sons (2001)

[HyvOja97] Hyvärinen, A., Oja, E.: *A fast fixed-point algorithm for independent component analysis*. Neural Computation, vol. 6 (1997) 1484–1492

[HubW62] Hubel, D. H., Wiesel T. N.: *Receptive fields, binocular interaction and functional architecture in cat's visual cortex*. Journal of Physiology, vol. 160 (1962) 106–154

[HubW68] Hubel D. H., and Wiesel T. N.: *Receptive fields and functional architecture of monkey striate cortex*. Journal of Physiology, vol. 195 (1968) 215–243

[LangSL00] Langelaar, G. C., Setyawan, I., Lagendijk, R. L.: *Watermarking digital image and video data. A state-of-the-art overview*. IEEE Signal Processing Magazine (2000) 20–46

[Lee98] Lee, T-W.: *Independent Component Analysis. Theory and Applications*. Kluwer Academic Publishers (1998)

[LeeGS99] Lee, T-W., Girolami, M., Sejnowski, T. J.: *Independent Component Analysis using an Extended Infomax Algorithm for Mixed Sub-Gaussian and Super-Gaussian Sources*. Neural Computation, vol.11, no. 2 (1999) 417–441

[LindHyv04] Lindgren, J. T., Hyvärynen, A.: *Learning high-level independent component of images through a spectral representation*. Proceedings of International Conference on Pattern Recognition (ICPR2004), Cambridge, UK (2004)

[Nandi99] Nandi, A. K.: *Blind estimation using higher-order statistics*. Ed. Kluwer (1999)

[Olsh03] Olshausen, B. A.: *Principles of image representation in visual cortex*. The Visual Neurosciences, L.M. Chalupa, J.S. Werner, Eds. MIT Press (2003) 1603–1615

[OlshF96a] Olshausen, B. A., Field, D. J.: *Natural image statistics and efficient coding*. Network, vol. 7 (1996) 333–339

[OlshF96b] Olshausen, B. A., Field, D. J.: *Emergence of simple-Cell receptive field properties by learning a sparse code for natural images*. Nature, vol. 381 (1996) 607–609

[OlshF97] Olshausen, B. A., Field, D. J.: *Sparse coding with an overcomplete basis set: a strategy employed by V1?* Vision Research, vol. 37 (23) (1997) 3311–3325

[Papoul91] Papoulis, A.: *Probability, random variables and stochastic processes*. McGraw-Hill, 3rd Ed. (1991)

[Petit00] Petitcolas, F. A. P.: *Watermarking schemes evaluation*. IEEE Signal Processing Magazine (2000) 58–64

[PodilD01] Podilchuk, C. I., Delp, E. J.: *Digital watermarking: algorithms and applications*. IEEE Signal Processing Magazine (2001) 33–46

[Reza94] Reza, F. M. Reza: *An introduction to information theory*. Dover Publications (1994)

[SchmTh93] Schmidt, R. B., Thews, G.: *Fisiología humana*. 24 Ed., Interamericana, McGraw-Hill (1993)

[vanHat98a] van Hateren, J. H., van der Schaaf, A.: *Independent component filters of natural images compared with simple cells in primary visual cortex*. Proc. Roy. Soc. Lond. B 265 (1998) 359–366

[vanHat98b] van Hateren, J. H., Ruderman, D. L.: *Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex*. Proc. R. Soc. Lond. B 265 (1998) 2315–2320

[WMorg] `http://www.watermarkingworld.org`

[NatCollec] *Natural stimuli collection*. J. H. van Hateren. `http://hlab.phys.rug.nl/imlib/index.html`

[NatColICA] *Natural image collection for ICA experiments*. ICA Group. Laboratory of Computer and Information Science. Helsinki University of Technology. `http://www.cis.hut.fi/projects/ica/data/images`

[FastICA] Algoritmo *FastICA*. Disponible en: `http://www.cis.hut.fi/projects/ica/fastica/`

[Infomax] Algoritmo *Infomax*, disponible en:
`ftp://ftp.cnl.salk.edu/pub/tony/sep96.public`

[InfomaxE] Algoritmo *Infomax Extendido*, disponible en:
`http://inc2.ucsd.edu/~tewon/ica_cnl.html`

[Sparsenet] Algoritmo *Sparsenet*, disponible en:
`http://redwood.ucdavis.edu/bruno/sparsenet.html`