MÁSTER EN MATEMÁTICAS AVANZADAS

Especialidad de Estadística e Investigación Operativa

TRABAJO DE FIN DE MÁSTER

COMPONENTES SUBYACENTES COMUNES EN SERIES TEMPORALES

Juan Vicente Bógalo Román

Universidad Nacional de Educación a Distancia

Directora del TFM

Maria do Rosário Ramos

Universidade Aberta de Portugal

Octubre 2012

Comprender las cosas que nos rodean es la mejor preparación para comprender las cosas que hay más allá.

Hipatia de Alejandría

Índice General

Anexos

Introducción

El presente documento constituye la memoria del Trabajo de Fin de Máster correspondiente al Máster de Matemáticas Avanzadas de la UNED titulado "*Componentes subyacentes comunes en series temporales*".

El primer objetivo de este trabajo es describir con cierto detalle las técnicas estadísticas para obtener unas componentes no observables comunes a un conjunto de datos que pueden ser reconstruidos por dichas componentes. De este modo, se dispone de las técnicas del Análisis de Componentes Principales, PCA, y del Análisis de Componentes Independientes, ICA.

En este sentido, el segundo objetivo es determinar cómo y bajo qué condiciones se pueden utilizar estos métodos cuando los datos observados son series temporales y poner de relieve sus diferencias conceptuales y su significado. Consecuentemente, el fin de los métodos, PCA e ICA, sería estimar componentes oscilatorias no observables o subyacentes que son comunes a un conjunto de series temporales.

El último objetivo consiste en aplicar las técnicas estudiadas a un caso real que permita comparar las diferentes técnicas y poder obtener conclusiones prácticas.

El presente trabajo se organiza en tres capítulos. El primero se dedica al estudio del PCA. Se parte de la caracterización de las componentes principales de la población para centrarse en la derivación, estimación y posibles inferencias de las componentes principales muestrales. La siguiente sección se dedica al, siempre controvertido e inacabado, tema de determinar el número de componentes principales a estimar. El capítulo finaliza con las implicaciones que para el PCA tienen las series temporales y se dedica especial atención a un método específico como es el Análisis Espectral Singular, SSA, con el objetivo de filtrar las series temporales.

El segundo capítulo se aplica a la exposición del ICA. Se inicia el estudio con la caracterización de las componentes independientes del modelo sin ruido para proseguir con las diferentes funciones que permiten su estimación con cierto énfasis en las medidas de no-gaussianidad. El capítulo continúa con el estudio de cómo se deben preparar los datos para una mejor realización del ICA. Para finalizar, se derivan y

exponen dos algoritmos para la estimación de las componentes independientes: el algoritmo FastICA, basado en la maximización de la no-gaussianidad, y el algoritmo AMUSE, más adecuado para realizar un ICA con series temporales.

En el tercer capítulo se realiza una aplicación práctica sobre las series de las tasas de empleo de las CC.AA. Para ello, se utilizan las diferentes técnicas estudiadas en los capítulos anteriores. Mediante el SSA, aplicado a las series estacionarias en varianza aunque no necesariamente en media tal y como se probará, se obtienen unas series limpias de ruido con las que se realiza un PCA y un ICA. El propósito es computar unas componentes oscilatorias subyacentes comunes al conjunto de series de las tasas de empleo. La determinación del número de componentes principales se realiza de forma empírica según algunas reglas expuestas en el primer capítulo. Sin embargo, para determinar el número de componentes independientes se motiva y propone una regla basada en la aproximación de la matriz de covarianzas. Para concluir, se realiza una comparación de los métodos estudiados mediante el análisis de las diferentes componentes estimadas. Esta comparación no se puede efectuar en términos de distancias, como se demostrará, y se desarrolla con técnicas geométricas y espectrales.

En los anexos que figuran a continuación, se han adjuntado, tanto de forma gráfica como en tablas, los resultados detallados de las diferentes técnicas utilizadas. Se ha preferido hacerlo de este modo, y no incluir todo este material dentro del texto, para elaborar un trabajo de fácil y cómoda lectura pero sin prescindir de los pormenores.

El software utilizado para la realización de todos los cálculos y estimación de las diferentes componentes ha sido MATLAB, bien empleando las funciones disponibles de sus *toolboxes*, bien programando nuevas funciones cuyo código se incluye en el Anexo F. También, MATLAB ha servido para la realización de algunos gráficos mientras que Excel se ha empleado como sistema de almacenamiento de resultados y para la confección de la mayoría de las figuras.

Las referencias bibliográficas mencionadas al final de este documento, no son una mera relación de las citadas a lo largo del texto sino que, han servido para la elaboración del presente trabajo y han constituido una base exhaustiva de estudio.

Capítulo 1

Análisis de Componentes Principales

1.1 Introducción

Un problema importante en el análisis de datos multivariantes es la reducción de la dimensión, es decir, describir con la mayor precisión posible el comportamiento de p variables mediante un pequeño subconjunto $r < p$ de ellas a costa de la menor pérdida posible de información.

El análisis de componentes principales tiene este objetivo: dado un conjunto de datos con un gran número de variables interrelacionadas, analizar si es posible representar adecuadamente esta información con un número menor de variables (construidas como combinaciones lineales de las originales) que no están correlacionadas y que conservan, tanto como sea posible, la mayor parte de la variabilidad presente en el conjunto de datos de las variables originales. De este modo, el análisis de componentes principales, PCA, está interesado en explicar la estructura de varianzas-covarianzas de un conjunto de p variables a través de unas pocas $r < p$ combinaciones lineales de dichas variables que se denominan componentes principales.

A pesar de que se requieren p componentes para reproducir la variabilidad total del sistema, con frecuencia gran parte de esta variabilidad puede ser explicada por un pequeño número $r < p$ de componentes. Si esto es así, existe tanta información en las r componentes como en las p variables, de esta forma, las r componentes pueden remplazar a las p variables iniciales.

La técnica de PCA se debe a los trabajos pioneros de Pearson (1901), desde un punto de vista geométrico, y al desarrollo de Hotelling (1933), desde un punto de vista algebraico. La utilidad de esta técnica es doble:

- Permite representar de forma óptima observaciones de un espacio p-dimensional en un espacio de reducida dimensión. Ello, es un primer paso para identificar variables no observadas o subyacentes que generan los datos.
- Facilita la interpretación de los datos al transformar las variables originales correlacionadas en otras nuevas incorrelacionadas.

En este capítulo se va a estudiar el análisis de componentes principales comenzando por su derivación y el examen de sus propiedades en la población para posteriormente desarrollar y extender los resultados a una muestra con el fin de poder realizar inferencias. Además, se discutirá, el siempre controvertido tema, de la elección del número de componentes, y, finalmente, se abordará como ampliar el tratamiento del PCA a realizaciones de procesos estocásticos o series temporales.

1.2 Componentes principales de la población

Algebraicamente, las componentes principales son combinaciones lineales particulares de p variables aleatorias $x_1, x_2, \cdots, x_p$. Geométricamente, estas combinaciones lineales representan la selección de un nuevo sistema de coordenadas que se obtiene por rotación del sistema original que tiene a $x_1, x_2, \cdots, x_p$ como los ejes de coordenadas. Los nuevos ejes representan las direcciones de máxima variabilidad y proporcionan una descripción más simple de la estructura de covarianzas.

Según se verá, las componentes principales dependen únicamente de la matriz de covarianzas $\mathbf{\Sigma}$ (o de la matriz de correlaciones $\boldsymbol{\rho}$) de $x_1, x_2, \cdots, x_p$. Su desarrollo no requiere suponer una distribución gaussiana multivariante, sin embargo, las componentes principales de poblaciones gaussianas multivariantes tienen una interpretación útil en términos de elipsoides de densidad constante y, además, se pueden realizar inferencias de las componentes de la muestra cuando la población es gaussiana multivariante.

1.2.1 Derivación y propiedades de las componentes principales

Sea un vector aleatorio $\mathbf{x}=\left(x_1,x_2,\cdots,x_p\right)^T$ que tiene vector de medias $\boldsymbol{\mu}$ y matriz de covarianzas $\boldsymbol{\Sigma}$ con autovalores $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_p\geq 0$. Se consideran las siguientes combinaciones lineales

$$\begin{aligned} y_1 &= \mathbf{a}_1^T\mathbf{x}=a_{11}x_1+a_{12}x_2+\cdots+a_{1\mathrm{p}}x_{\mathrm{p}} \\ y_2 &= \mathbf{a}_2^T\mathbf{x}=a_{21}x_1+a_{22}x_2+\cdots+a_{2\mathrm{p}}x_{\mathrm{p}} \\ &\vdots \\ y_{\mathrm{p}} &= \mathbf{a}_{\mathrm{p}}^T\mathbf{x}=a_{\mathrm{p}1}x_1+a_{\mathrm{p}2}x_2+\cdots+a_{\mathrm{pp}}x_{\mathrm{p}} \end{aligned} \tag{1.1}$$

para las cuales se obtiene

$$\mathrm{Var}\left(y_k\right)=\mathrm{Var}\left(\mathbf{a}_k^T\mathbf{x}\right)=E\left[\mathbf{a}_k^T\left(\mathbf{x}-\boldsymbol{\mu}\right)\left(\mathbf{x}-\boldsymbol{\mu}\right)^T\mathbf{a}_k\right]=\mathbf{a}_k^T\boldsymbol{\Sigma}\mathbf{a}_k \tag{1.2}$$

$$\mathrm{Cov}\left(y_i,y_k\right)=\mathrm{Cov}\left(\mathbf{a}_i^T\mathbf{x},\mathbf{a}_k^T\mathbf{x}\right)=E\left[\mathbf{a}_i^T\left(\mathbf{x}-\boldsymbol{\mu}\right)\left(\mathbf{x}-\boldsymbol{\mu}\right)^T\mathbf{a}_k\right]=\mathbf{a}_i^T\boldsymbol{\Sigma}\mathbf{a}_k \tag{1.3}$$

donde los $\mathbf{a}_k$ con $k=1,2,\cdots,p$ son vectores de coeficientes. Las componentes principales son estas combinaciones lineales incorrelacionadas, $y_k \quad k=1,2,\cdots,p$, cuyas varianzas dadas en (1.2) sean tan grandes como sea posible.

La primera componente principal es la combinación lineal $\mathbf{a}_1^T\mathbf{x}$ con máxima varianza, es decir, que maximiza $\mathrm{Var}\left(y_1\right)=\mathbf{a}_1^T\boldsymbol{\Sigma}\mathbf{a}_1$. Está claro que $\mathbf{a}_1^T\boldsymbol{\Sigma}\mathbf{a}_1$ se puede incrementar multiplicando $\mathbf{a}_1$ por una constante así que, para evitar esta indeterminación, se impone la restricción de normalización $\mathbf{a}_1^T\mathbf{a}_1=1$.

La segunda componente principal es la combinación $\mathbf{a}_2^T\mathbf{x}$ incorrelacionada con $\mathbf{a}_1^T\mathbf{x}$, $\mathrm{Cov}(\mathbf{a}_1^T\mathbf{x},\mathbf{a}_2^T\mathbf{x})=0$, con máxima varianza, es decir, que maximiza $\mathrm{Var}\left(y_2\right)=\mathbf{a}_2^T\boldsymbol{\Sigma}\mathbf{a}_2$ y a la que también se impone la restricción de normalización $\mathbf{a}_2^T\mathbf{a}_2=1$.

En general, la k-ésima componente principal es la combinación lineal $\mathbf{a}_k^T\mathbf{x}$ incorrelacionada con las anteriores, $\mathrm{Cov}(\mathbf{a}_i^T\mathbf{x},\mathbf{a}_k^T\mathbf{x})=0 \quad \forall i<k$, que maximiza $\mathrm{Var}\left(y_k\right)=\mathbf{a}_k^T\boldsymbol{\Sigma}\mathbf{a}_k$ sujeta, además, a la restricción de normalización $\mathbf{a}_k^T\mathbf{a}_k=1$.

Proposición 1.1 *Sea* $\mathbf{x}=\left(x_1,x_2,\cdots,x_p\right)^T$ *un vector aleatorio cuya matriz de covarianzas* $\boldsymbol{\Sigma}$ *tiene como pares de autovalores-autovectores a* $\left(\lambda_1,\boldsymbol{\upsilon}_1\right),\left(\lambda_2,\boldsymbol{\upsilon}_2\right),\cdots,\left(\lambda_p,\boldsymbol{\upsilon}_p\right)$ *donde* $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_p\geq 0$. *Entonces, la k-ésima componente principal esta dada por*

$$y_k=\boldsymbol{\upsilon}_k^T\mathbf{x}=\upsilon_{k1}x_1+\upsilon_{k2}x_2+\cdots+\upsilon_{kp}x_p \qquad k=1,2,\cdots,p \tag{1.4}$$

y, por tanto, se tiene que

$$\begin{aligned} \mathrm{Var}\left(y_k\right) &= \boldsymbol{\upsilon}_k^T \boldsymbol{\Sigma} \boldsymbol{\upsilon}_k = \lambda_k \qquad k = 1, 2, \cdots, p \\ \mathrm{Cov}\left(y_j, y_k\right) &= \boldsymbol{\upsilon}_j^T \boldsymbol{\Sigma} \boldsymbol{\upsilon}_k = 0 \qquad j \neq k \end{aligned} \tag{1.5}$$

Demostración. Para demostrar esta proposición se utilizan las propiedades de las matrices normales, ver Meyer (2000), de forma similar a la demostración de Johnson y Wichern (2007). Este enfoque es alternativo al habitual que emplea la técnica de los multiplicadores de Lagrange tal y como lo efectúa Jolliffe (2002).

La matriz de covarianzas $\boldsymbol{\Sigma}$ es simétrica y definida positiva con lo cual admite una diagonalización ortogonal $\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$, o equivalentemente $\mathbf{U}^T\boldsymbol{\Sigma}\mathbf{U} = \boldsymbol{\Lambda}$, donde $\boldsymbol{\Lambda} = \mathrm{diag}\left(\lambda_1, \lambda_2, \cdots, \lambda_p\right)$ siendo $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ los autovalores de $\boldsymbol{\Sigma}$ y $\mathbf{U} = \left[\boldsymbol{\upsilon}_1, \boldsymbol{\upsilon}_2, \cdots, \boldsymbol{\upsilon}_p\right]$ una matriz ortogonal, $\mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{U}^T = \mathbf{I}$, es decir, los autovectores $\boldsymbol{\upsilon}_1, \boldsymbol{\upsilon}_2, \cdots, \boldsymbol{\upsilon}_p$ son ortonormales. Puesto que $\mathbf{a}^T\mathbf{a} = 1 \Leftrightarrow \mathbf{b}^T\mathbf{b} = 1$ para $\mathbf{b} = \mathbf{U}^T\mathbf{a}$, se llega a

$$\max_{\mathbf{a}^T\mathbf{a}=1} \mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a} = \max_{\mathbf{b}^T\mathbf{b}=1} \mathbf{b}^T\boldsymbol{\Lambda}\mathbf{b} = \max_{\mathbf{b}^T\mathbf{b}=1} \sum_{j=1}^{p} \lambda_j b_j^2 \leq \max_{\mathbf{b}^T\mathbf{b}=1} \lambda_1 \sum_{j=1}^{p} b_j^2 = \lambda_1 \tag{1.6}$$

Estableciendo $\mathbf{a} = \boldsymbol{\upsilon}_1$ se tiene que $\mathbf{b} = \mathbf{U}^T\boldsymbol{\upsilon}_1 = \left(1, 0, \cdots, 0\right)^T$ con lo cual en (1.6) se alcanza la igualdad

$$\max_{\mathbf{a}^T\mathbf{a}=1} \mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a} = \lambda_1 = \boldsymbol{\upsilon}_1^T\boldsymbol{\Sigma}\boldsymbol{\upsilon}_1 = \mathrm{Var}\left(\boldsymbol{\upsilon}_1^T\mathbf{x}\right) = \mathrm{Var}\left(y_1\right) \tag{1.7}$$

Ahora, $\mathbf{a}^T\mathbf{a} = 1 \Leftrightarrow \mathbf{b}^T\mathbf{b} = 1$ para $\mathbf{a} = \mathbf{U}\mathbf{b} = b_1\boldsymbol{\upsilon}_1 + b_2\boldsymbol{\upsilon}_2 + \cdots + b_p\boldsymbol{\upsilon}_p$, de este modo $\mathbf{a} \perp \boldsymbol{\upsilon}_1, \boldsymbol{\upsilon}_2, \cdots, \boldsymbol{\upsilon}_{k-1}$ implica que

$$0 = \boldsymbol{\upsilon}_j^T\mathbf{a} = b_1\boldsymbol{\upsilon}_j^T\boldsymbol{\upsilon}_1 + b_2\boldsymbol{\upsilon}_j^T\boldsymbol{\upsilon}_2 + \cdots + b_j\boldsymbol{\upsilon}_j^T\boldsymbol{\upsilon}_j + \cdots + b_p\boldsymbol{\upsilon}_j^T\boldsymbol{\upsilon}_p = b_j \quad \forall\, j < k, k = 2, 3, \cdots, p$$

con lo cual

$$\max_{\substack{\mathbf{a}^T\mathbf{a}=1 \\ \mathbf{a}\perp\boldsymbol{\upsilon}_1,\boldsymbol{\upsilon}_2,\cdots,\boldsymbol{\upsilon}_{k-1}}} \mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a} = \max_{\substack{\mathbf{b}^T\mathbf{b}=1 \\ \mathbf{a}\perp\boldsymbol{\upsilon}_1,\boldsymbol{\upsilon}_2,\cdots,\boldsymbol{\upsilon}_{k-1}}} \mathbf{b}^T\boldsymbol{\Lambda}\mathbf{b} = \max_{\mathbf{b}^T\mathbf{b}=1} \sum_{j=k}^{p} \lambda_j b_j^2 \leq \max_{\mathbf{b}^T\mathbf{b}=1} \lambda_k \sum_{j=k}^{p} b_j^2 = \lambda_k \tag{1.8}$$

Tomando $b_k = 1, b_{k+1} = 0, \cdots, b_p = 0$ se tiene que $\mathbf{a} = \mathbf{U}\mathbf{b} = \boldsymbol{\upsilon}_k$ de forma que en (1.8) se alcanza la igualdad

$$\max_{\substack{\mathbf{a}^T\mathbf{a}=1 \\ \mathbf{a}\perp\boldsymbol{\upsilon}_1,\boldsymbol{\upsilon}_2,\cdots,\boldsymbol{\upsilon}_{k-1}}} \mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a} = \lambda_k = \boldsymbol{\upsilon}_k^T\boldsymbol{\Sigma}\boldsymbol{\upsilon}_k = \mathrm{Var}\left(\boldsymbol{\upsilon}_k^T\mathbf{x}\right) = \mathrm{Var}\left(y_k\right) \quad k = 2, 3, \cdots, p \tag{1.9}$$

Finalmente, resta demostrar que si $\boldsymbol{\upsilon}_i \perp \boldsymbol{\upsilon}_k$, es decir $\boldsymbol{\upsilon}_i^T\boldsymbol{\upsilon}_k = 0$ para $i \neq k$, entonces $\mathrm{Cov}\left(y_i, y_k\right) = 0$. Los autovectores de $\boldsymbol{\Sigma}$ son ortogonales por ser una matriz simétrica aunque no todos los autovalores sean distintos. Por ello, para cualquier par de autovectores $\boldsymbol{\upsilon}_i$ y $\boldsymbol{\upsilon}_k$, $\boldsymbol{\upsilon}_i^T\boldsymbol{\upsilon}_k = 0$ para $i \neq k$. Puesto que $\boldsymbol{\Sigma}\boldsymbol{\upsilon}_k = \lambda_k\boldsymbol{\upsilon}_k$, se llega a

$$\mathrm{Cov}\left(y_i, y_k\right) = \boldsymbol{\upsilon}_i^T \boldsymbol{\Sigma} \boldsymbol{\upsilon}_k = \boldsymbol{\upsilon}_i^T \lambda_k \boldsymbol{\upsilon}_k = \lambda_k \boldsymbol{\upsilon}_i^T \boldsymbol{\upsilon}_k = 0 \quad \forall i \neq k \tag{1.10}$$

■

Según esta proposición, las componentes principales son una transformación lineal ortogonal de $\mathbf{x}$, $\mathbf{y} = \mathbf{U}^T\mathbf{x}$ ($\mathbf{y}$ es el vector aleatorio con dichas componentes), están incorrelacionadas y sus varianzas son iguales a los autovalores de la matriz $\boldsymbol{\Sigma}$.

Proposición 1.2 *Sea* $\mathbf{x} = \left(x_1, x_2, \cdots, x_p\right)^T$ *un vector aleatorio cuya matriz de covarianzas* $\boldsymbol{\Sigma}$ *tiene como pares de autovalores-autovectores a* $\left(\lambda_1, \boldsymbol{\upsilon}_1\right), \left(\lambda_2, \boldsymbol{\upsilon}_2\right), \cdots, \left(\lambda_p, \boldsymbol{\upsilon}_p\right)$ *donde* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ *y sean* $y_1 = \boldsymbol{\upsilon}_1^T\mathbf{x}, y_2 = \boldsymbol{\upsilon}_2^T\mathbf{x}, \cdots, y_p = \boldsymbol{\upsilon}_p^T\mathbf{x}$ *las componentes principales. Entonces*

$$\sum_{k=1}^{p} \mathrm{Var}\left(x_k\right) = \sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sum_{k=1}^{p} \mathrm{Var}\left(y_k\right)$$

Demostración. Por definición, $\mathrm{trace}\left(\boldsymbol{\Sigma}\right) = \sum_{k=1}^{p} \sigma_{kk}$. La matriz $\boldsymbol{\Sigma}$ es simétrica con lo cual admite una diagonalización ortogonal $\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$ donde $\boldsymbol{\Lambda} = \mathrm{diag}\left(\lambda_1, \lambda_2, \cdots, \lambda_p\right)$ y $\mathbf{U} = \left[\boldsymbol{\upsilon}_1, \boldsymbol{\upsilon}_2, \cdots, \boldsymbol{\upsilon}_p\right]$ es una matriz ortogonal, $\mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{U}^T = \mathbf{I}$, y se tiene

$$\mathrm{trace}\left(\boldsymbol{\Sigma}\right) = \mathrm{trace}\left(\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T\right) = \mathrm{trace}\left(\boldsymbol{\Lambda}\mathbf{U}^T\mathbf{U}\right) = \mathrm{trace}\left(\boldsymbol{\Lambda}\right) = \sum_{k=1}^{p} \lambda_k$$

De este modo,

$$\sum_{k=1}^{p} \mathrm{Var}\left(x_k\right) = \mathrm{trace}\left(\boldsymbol{\Sigma}\right) = \mathrm{trace}\left(\boldsymbol{\Lambda}\right) = \sum_{k=1}^{p} \mathrm{Var}\left(y_k\right)$$

Y, así, queda demostrada la proposición. ■

Esta proposición muestra que

$$\begin{aligned} \text{Varianza total de la población} &= \sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} \\ &= \lambda_1 + \lambda_2 + \cdots + \lambda_p \end{aligned} \tag{1.11}$$

y, consecuentemente, la ratio

$$\frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} \qquad k = 1, 2, \cdots, p \tag{1.12}$$

es la proporción de la varianza total de la población que es debida a la k-ésima componente principal.

Si la mayoría (por ejemplo más del 80%) de la varianza total de la población, para un p grande, se puede atribuir a las uno, dos o tres primeras componentes principales, entonces estas componentes pueden remplazar a las p variables originales sin gran pérdida de información.

Proposición 1.3 *Si* $y_1 = \mathbf{v}_1^T\mathbf{x}, y_2 = \mathbf{v}_2^T\mathbf{x}, \cdots, y_p = \mathbf{v}_p^T\mathbf{x}$ *son las componentes principales obtenidas de la matriz de covarianzas* $\mathbf{\Sigma}$ *que tiene como pares de autovalores-autovectores a* $(\lambda_1, \mathbf{v}_1), (\lambda_2, \mathbf{v}_2), \cdots, (\lambda_p, \mathbf{v}_p)$ *donde* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, *entonces*

$$\rho_{y_k, x_i} = \frac{\upsilon_{ki}\sqrt{\lambda_k}}{\sqrt{\sigma_{ii}}} \qquad i, k = 1, 2, \cdots, p \tag{1.13}$$

son los coeficientes de correlación entre las componentes y_k *y las variables* x_i.

Demostración. Si $\mathbf{e}_i = (0, \cdots, 0, 1, 0, \cdots, 0)^T$ es el vector con todos sus elementos nulos excepto el de la posición i que es 1, entonces se tiene que $x_i = \mathbf{e}_i^T\mathbf{x}$ y $\mathrm{Cov}(x_i, y_k) = \mathrm{Cov}(\mathbf{e}_i^T\mathbf{x}, \mathbf{v}_k^T\mathbf{x}) = \mathbf{e}_i^T\mathbf{\Sigma}\mathbf{v}_k$ .Como $\mathbf{\Sigma}\mathbf{v}_k = \lambda_k\mathbf{v}_k$, $\mathrm{Cov}(x_i, y_k) = \mathbf{e}_i^T\lambda_k\mathbf{v}_k = \lambda_k\upsilon_{ki}$. Entonces $\mathrm{Var}(y_k) = \lambda_k$ y $\mathrm{Var}(x_i) = \sigma_{ii}$ producen

$$\rho_{y_k, x_i} = \frac{\mathrm{Cov}(x_i, y_k)}{\sqrt{\mathrm{Var}(y_k)}\sqrt{\mathrm{Var}(x_i)}} = \frac{\lambda_k\upsilon_{ki}}{\sqrt{\lambda_k}\sqrt{\sigma_{ii}}} = \frac{\upsilon_{ki}\sqrt{\lambda_k}}{\sqrt{\sigma_{ii}}} \qquad i, k = 1, 2, \cdots, p$$

Con lo cual, queda demostrada la proposición. ■

Proposición 1.4 *La descomposición espectral de la matriz de covarianzas* $\mathbf{\Sigma}$ *del vector aleatorio* $\mathbf{x} = (x_1, x_2, \cdots, x_p)^T$ *está dada por*

$$\mathbf{\Sigma} = \lambda_1\mathbf{v}_1\mathbf{v}_1^T + \lambda_2\mathbf{v}_2\mathbf{v}_2^T + \cdots + \lambda_p\mathbf{v}_p\mathbf{v}_p^T$$

donde $(\lambda_1, \mathbf{v}_1), (\lambda_2, \mathbf{v}_2), \cdots, (\lambda_p, \mathbf{v}_p)$ *son los pares de autovalores-autovectores de* $\mathbf{\Sigma}$.

Demostración. La matriz $\mathbf{\Sigma}$ es simétrica con lo cual admite una diagonalización ortogonal $\mathbf{\Sigma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ donde $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_p)$ y $\mathbf{U} = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_p]$ es una matriz ortogonal, $\mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{U}^T = \mathbf{I}$. Si se expande el lado derecho del producto de la diagonalización se tiene la descomposición buscada

$$\mathbf{\Sigma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T = \lambda_1\mathbf{v}_1\mathbf{v}_1^T + \lambda_2\mathbf{v}_2\mathbf{v}_2^T + \cdots + \lambda_p\mathbf{v}_p\mathbf{v}_p^T$$ ■

Según esta proposición, los elementos σ_{kk} de la diagonal principal de $\mathbf{\Sigma}$, es decir, las varianzas de las variables originales x_k se pueden escribir como combinación lineal de

las varianzas de las componentes principales

$$\sigma_{kk} = \mathrm{Var}\left(x_k\right) = \sum_{i=1}^{p} \lambda_i \upsilon_{ik}^2 \tag{1.14}$$

1.2.2 Interpretación de las componentes principales poblacionales

Las componentes principales tienen una sencilla interpretación geométrica con implicaciones estadísticas según se desprende de la siguiente proposición.

Proposición 1.5 *Las componentes principales definen los ejes principales de la familia de elipsoides p-dimensionales*

$$\left(\mathbf{x}-\boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right) = c^2 \tag{1.15}$$

donde $\boldsymbol{\mu}$ *es el vector de medias de* $\mathbf{x}$, $\boldsymbol{\mu} = E\left[\mathbf{x}\right]$.

Demostración. Sin pérdida de generalidad, se considera que $\boldsymbol{\mu} = \mathbf{0}$ puesto que la transformación $\mathbf{w} = \mathbf{x} - \boldsymbol{\mu}$ verifica que $E\left[\mathbf{w}\right] = \mathbf{0}$ y, sin embargo, $\mathrm{Cov}\left(\mathbf{w}\right) = \mathrm{Cov}\left(\mathbf{x}\right)$. De este modo, la ecuación (1.15) se puede reformular como

$$\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x} = c^2 \tag{1.16}$$

Por la Proposición 1.1 se tiene que $\mathbf{y} = \mathbf{U}^T\mathbf{x}$ con lo cual $\mathbf{x} = \mathbf{U}\mathbf{y}$. Sustituyendo $\mathbf{x}$ por esta última expresión en (1.16) y teniendo en cuenta que $\mathbf{U}^T\boldsymbol{\Sigma}^{-1}\mathbf{U} = \boldsymbol{\Lambda}^{-1}$ se llega a

$$\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} = \mathbf{y}^T\mathbf{U}^T\boldsymbol{\Sigma}^{-1}\mathbf{U}\mathbf{y} = \mathbf{y}^T\boldsymbol{\Lambda}^{-1}\mathbf{y} = c^2 \tag{1.17}$$

donde $\boldsymbol{\Lambda} = \mathrm{diag}\left(\lambda_1, \lambda_2, \cdots, \lambda_p\right)$ con $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$ son los autovalores de la matriz $\boldsymbol{\Sigma}$. Finalmente, la ecuación (1.17) se puede rescribir como

$$\frac{y_1^2}{\lambda_1} + \frac{y_2^2}{\lambda_2} + \cdots + \frac{y_p^2}{\lambda_p} = c^2 \tag{1.18}$$

que representa la ecuación de un elipsoide referido a sus ejes principales $y_1, y_2, \cdots, y_p$ que se sitúan en las direcciones de $\boldsymbol{\upsilon}_1, \boldsymbol{\upsilon}_2, \cdots, \boldsymbol{\upsilon}_p$ respectivamente. Esta ecuación implica que las longitudes de los semiejes principales del elipsoide son proporcionales a $\lambda_1^{1/2}, \lambda_2^{1/2}, \cdots, \lambda_p^{1/2}$. ■

El resultado anterior cobra más importancia cuando el vector aleatorio $\mathbf{x}$ tiene una distribución gaussiana p-dimensional. En esta situación los elipsoides dados por (1.16), si están centrados en el origen, o por (1.15), si están centrados en $\boldsymbol{\mu}$, definen contornos de densidad constante para la distribución de $\mathbf{x}$.

El primer o más largo de los ejes principales de los elipsoides definirá la dirección en la que la variación estadística es mayor, es decir, el más largo de los ejes principales permanece en la dirección de $\boldsymbol{\upsilon}_1$. El segundo eje principal maximiza la variación estadística sujeto a que debe ser ortogonal al primero, es decir, permanece en la dirección de $\boldsymbol{\upsilon}_2$. Este resultado se ilustra en la Figura 1.1 para un vector aleatorio gaussiano bidimensional. En ella, se puede apreciar que las componentes principales se obtienen rotando los ejes de coordenadas hasta que coinciden con los ejes de la elipse de densidad constante.

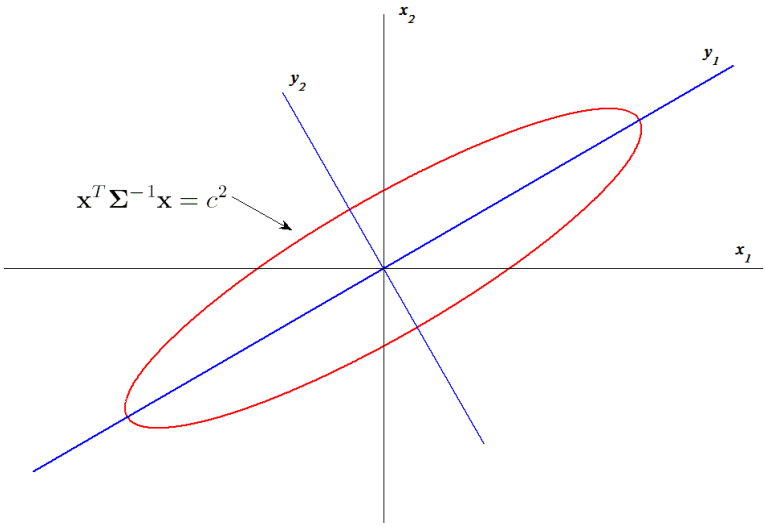

Figura 1.1 Elipse de densidad constante y componentes principales.

En resumen, las componentes principales $y_1 = \boldsymbol{\upsilon}_1^T\mathbf{x}, y_2 = \boldsymbol{\upsilon}_2^T\mathbf{x}, \cdots, y_p = \boldsymbol{\upsilon}_p^T\mathbf{x}$ permanecen en las direcciones de los ejes de un elipsoide de densidad constante. Por ello, cualquier punto sobre el k-ésimo eje del elipsoide tiene coordenadas proporcionales a $\boldsymbol{\upsilon}_k = \left(\upsilon_{k1}, \upsilon_{k2}, \cdots, \upsilon_{kp}\right)^T$ y las coordenadas respecto del sistema de componentes principales son de la forma $\left(0, \cdots, 0, y_k, 0, \cdots, 0\right)$.

Cuando $\boldsymbol{\mu} \neq \mathbf{0}$, significa que la componente principal está centrada en media $y_k = \boldsymbol{\upsilon}_k^T\left(\mathbf{x} - \boldsymbol{\mu}\right)$, así, tiene media cero y se encuentra en la dirección de $\boldsymbol{\upsilon}_k$.

1.2.3 Componentes principales poblacionales estandarizadas

Las componentes principales también se pueden obtener a partir de la matriz de correlaciones $\boldsymbol{\rho}$, es decir, se pueden derivar para las variables estandarizadas

$$z_k = \frac{x_k - \mu_k}{\sqrt{\sigma_{kk}}} \qquad k = 1, 2, \cdots, p \tag{1.19}$$

Si $\mathbf{z} = \left(z_1, z_2, \cdots, z_p\right)^T$ es el vector aleatorio con las variables estandarizadas, en notación matricial se escribe

$$\mathbf{z} = \mathbf{V}^{-1/2}\left(\mathbf{x} - \boldsymbol{\mu}\right) \tag{1.20}$$

donde $\mathbf{V}^{1/2} = \text{diag}\left(\sqrt{\sigma_{11}}, \sqrt{\sigma_{22}}, \cdots, \sqrt{\sigma_{pp}}\right)$. Obviamente, su vector de medias y su matriz de covarianzas son $E\left[\mathbf{z}\right] = \mathbf{0}$ y $\text{Cov}\left(\mathbf{z}\right) = \mathbf{V}^{-1/2}\boldsymbol{\Sigma}\mathbf{V}^{-1/2} = \boldsymbol{\rho}$, respectivamente.

Las componentes principales de $\mathbf{z}$ se obtienen de la matriz de correlación de $\mathbf{x}$ que es la matriz de covarianzas de $\mathbf{z}$. Todos los resultados obtenidos de las proposiciones previas se pueden aplicar con alguna simplificación puesto que la varianza de cada variable z_k es la unidad.

Proposición 1.6 *La k-ésima componente principal de las variables estandarizadas* $\mathbf{z} = \left(z_1, z_2, \cdots, z_p\right)^T$ *con* $\text{Cov}\left(\mathbf{z}\right) = \boldsymbol{\rho}$*, está dada por*

$$y_k = \boldsymbol{\upsilon}_k^T\mathbf{z} = \boldsymbol{\upsilon}_k^T\mathbf{V}^{-1/2}\left(\mathbf{x} - \boldsymbol{\mu}\right) \qquad k = 1, 2, \cdots, p \tag{1.21}$$

Además,

$$\sum_{k=1}^{p}\text{Var}\left(y_k\right) = \sum_{k=1}^{p}\text{Var}\left(z_k\right) = p \tag{1.22}$$

y

$$\rho_{y_k, x_i} = \upsilon_{ki}\sqrt{\lambda_k} \qquad i, k = 1, 2, \cdots, p \tag{1.23}$$

donde $\left(\lambda_1, \boldsymbol{\upsilon}_1\right), \left(\lambda_2, \boldsymbol{\upsilon}_2\right), \cdots, \left(\lambda_p, \boldsymbol{\upsilon}_p\right)$ *son los pares de autovalores-autovectores de la matriz* $\boldsymbol{\rho}$ *con* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.

Demostración. Esta proposición es el resultado de las Proposiciones 1.1, 1.2 y 1.3 sustituyendo $\mathbf{x} = \left(x_1, x_2, \cdots, x_p\right)^T$ por $\mathbf{z} = \left(z_1, z_2, \cdots, z_p\right)^T$ y $\boldsymbol{\Sigma}$ por $\boldsymbol{\rho}$. ■

En la ecuación (1.22) se observa que el total de la varianza de la población estandarizada es simplemente p, la suma de los elementos de la diagonal de la matriz $\boldsymbol{\rho}$. Por tanto, usando (1.12) con $\mathbf{z}$ en lugar de $\mathbf{x}$, la proporción de la varianza total de la población explicada por la k-ésima componente principal está dada por

$$\frac{\lambda_k}{p} \qquad k = 1, 2, \cdots, p \tag{1.24}$$

donde los λ_k son los autovalores de la matriz $\boldsymbol{\rho}$.

Las componentes principales derivadas de $\boldsymbol{\Sigma}$ son diferentes a las obtenidas de $\boldsymbol{\rho}$. Puede parecer que las componentes para la matriz $\boldsymbol{\rho}$ podrían derivarse bastante fácilmente de las correspondientes a $\boldsymbol{\Sigma}$ puesto que $\mathbf{z}$ se relaciona con $\mathbf{x}$ por una transformación sencilla. Sin embargo, esto no es así, no existe una relación simple entre ambos conjuntos de componentes. Si las componentes principales obtenidas de $\boldsymbol{\rho}$ se expresan en términos de $\mathbf{x}$ (por la transformación inversa de $\mathbf{z}$ a $\mathbf{x}$), entonces estas componentes no coinciden con las obtenidas de $\boldsymbol{\Sigma}$. La razón de esta circunstancia es que las componentes principales son invariantes bajo transformaciones ortogonales de $\mathbf{x}$ pero no bajo otras transformaciones, ver Zwiers (1999), y la transformación de $\mathbf{x}$ a $\mathbf{z}$ no es ortogonal. Por ello, las componentes principales para las matrices $\boldsymbol{\Sigma}$ y $\boldsymbol{\rho}$ no facilitan información equivalente ni se pueden derivar unas a partir de otras.

El mejor argumento para usar la matriz $\boldsymbol{\rho}$ en lugar de $\boldsymbol{\Sigma}$, para derivar las componentes, tal vez sea que proporciona resultados, para diferentes conjuntos de variables aleatorias, que son más directamente comparables. El gran problema del PCA basado en la matriz de covarianzas $\boldsymbol{\Sigma}$ es la sensibilidad de las componentes a las unidades de medida usadas para cada uno de las variables de $\mathbf{x}$. Si entre las varianzas de las variables de $\mathbf{x}$ existen grandes diferencias, entonces las variables cuyas varianzas son las más elevadas tienden a dominar las pocas primeras componentes principales. Por ello, con el fin de evitar que esto suceda, es conveniente utilizar las matrices de correlaciones para derivar las componentes principales.

1.2.4 Componentes principales en situaciones especiales

Existen situaciones en las que las variables x_k producen matrices de covarianzas o correlaciones con estructuras especiales cuyas componentes principales se pueden expresar de formas simples.

- **Las variables x_k están incorrelacionadas**

En esta situación, la matriz de covarianzas $\boldsymbol{\Sigma}$ es diagonal con la forma

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & 0 & \cdots & 0 \\ 0 & \sigma_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{pp} \end{pmatrix}$$

Si $\mathbf{e}_k$ es el vector con todos sus elementos iguales a cero excepto el de la posición k que es uno, se observa que $\boldsymbol{\Sigma}\mathbf{e}_k = \sigma_{kk}\mathbf{e}_k$ con lo cual $(\sigma_{kk}, \mathbf{e}_k)$ es el k-ésimo par autovalor-autovector. La componente principal k-ésima es $y_k = \mathbf{e}_k^T\mathbf{x} = x_k$, es decir, el conjunto de componentes principales coincide con las variables incorrelacionadas.

En estas condiciones no existe alguna ganancia al obtener las componentes principales. Además, si $\mathbf{x}$ tiene una distribución gaussiana p-dimensional $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, los contornos de densidad constante son elipses cuyos ejes ya se encuentran en las direcciones de máxima variación y, por tanto, no hay necesidad de rotar el sistema de coordenadas.

Si las variables incorrelacionadas se estandarizan, la matriz de correlaciones es la matriz identidad, $\boldsymbol{\rho} = \mathbf{I}$. Obviamente, $\boldsymbol{\rho}\mathbf{e}_k = 1\mathbf{e}_k$ con lo cual el autovalor 1 tiene multiplicidad p y los autovectores son los vectores $\mathbf{e}_k$ ya definidos. Consecuentemente, las componentes principales derivadas de $\boldsymbol{\rho}$ coinciden con las variables estandarizadas.

- **Las variables x_k están igualmente correlacionadas**

En esta ocasión la matriz de correlaciones de las variables x_k, la matriz de covarianzas de las variables estandarizadas z_k, tiene la forma

$$\boldsymbol{\rho} = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix} \tag{1.25}$$

Los p autovalores, según describen Johnson y Wichern (2007), se dividen en dos grupos. Cuando ρ es positivo, el autovalor mayor es

$$\lambda_1 = 1 + (p-1)\rho \tag{1.26}$$

asociado al autovector

$$\boldsymbol{\upsilon}_1 = \left(1/\sqrt{p}, 1/\sqrt{p}, \cdots, 1/\sqrt{p}\right) \tag{1.27}$$

y el resto de $p-1$ autovalores son

$$\lambda_2 = \lambda_3 = \ldots = \lambda_p = 1 - \rho$$

La primera componente principal es proporcional a la suma de las p variables estandarizadas

$$y_1 = \boldsymbol{\upsilon}_1^T\mathbf{z} = \frac{1}{\sqrt{p}}\sum_{k=1}^{p} z_k$$

y explica una proporción de la varianza total de la población igual a

$$\frac{\lambda_1}{p} = \rho + \frac{1-\rho}{p} \tag{1.28}$$

Se puede observar que λ_1 / p es similar a 1 cuando ρ está próximo a 1 y, por tanto, las restantes de $p-1$ componentes contribuyen muy poco de forma colectiva al total de la varianza. En esta situación, es suficiente con retener sólo la primera componente para explicar prácticamente la totalidad de la varianza de la población.

Si las variables estandarizadas z_k tienen una distribución gaussiana p-dimensional con matriz de covarianzas dada por (1.25), entonces los elipsoides de densidad constante tienen forma de cigarro con el eje mayor proporcional a la primera componente principal. Esta primera componente principal es la proyección de $\mathbf{z}$ sobre el subespacio generado por $\mathbf{1} = (1,1,\cdots,1)^T$. Los restantes ejes se encuentran en direcciones esféricamente perpendiculares al eje mayor.

1.3 Componentes principales de la muestra

Antes de desarrollar las componentes principales de la muestra y analizar sus propiedades, es necesario efectuar algunas observaciones. Se supone que se realizan n observaciones independientes de un vector aleatorio p-dimensional $\mathbf{x}$ con vector de medias $\boldsymbol{\mu}$ y matriz de covarianzas $\boldsymbol{\Sigma}$. Las n observaciones p-dimensionales se indican por $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ de forma que $\mathbf{x}_i = \left(x_{i1}, x_{i2}, \cdots, x_{ip}\right)^T$, $i = 1, 2, \cdots, n$. El vector de medias muestral es $\overline{\mathbf{x}} = \left(\overline{x}_1, \overline{x}_2, \cdots, \overline{x}_p\right)^T$ donde

$$\overline{x}_k = \frac{1}{n}\sum_{i=1}^{n} x_{ik} \qquad k = 1, 2, \cdots, p$$

También, las observaciones de la muestra producen la matriz de covarianzas muestral $\mathbf{S} = \left[s_{jk}\right]$ cuyo elemento (j,k) está dado por

$$s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij} - \overline{x}_j\right)\left(x_{ik} - \overline{x}_k\right)$$

y la matriz de correlaciones muestral $\mathbf{R} = \left[r_{jk}\right]$ cuyo elemento (j,k) se obtiene como

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}}\sqrt{s_{kk}}}$$

Además, se define la matriz $\mathbf{X}=\left[\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_n\right]^T=\left[x_{ik}\right]$ de dimensión $(n\times p)$ cuyo elemento (i,k) es x_{ik} y representa el valor de la variables k-ésima para la observación i-ésima de la muestra. Así mismo, se define la matriz $\tilde{\mathbf{X}}=\left[\tilde{x}_{ik}\right]$ de dimensión $(n\times p)$ cuyo elemento (i,k) es $\tilde{x}_{ik}=x_{ik}-\overline{x}_k$ y representa la desviación respecto a la media de la variable k-ésima para la observación i-ésima de la muestra. En estas condiciones, la matriz de covarianzas muestral $\mathbf{S}$ se puede obtener como

$$\mathbf{S}=\frac{1}{n-1}\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}$$

1.3.1 Derivación y propiedades de las componentes principales

El objetivo es construir combinaciones lineales no correlacionadas que den cuenta de gran parte de la variación en la muestra. Los n valores de cualquier combinación lineal

$$\mathbf{a}_k^T\mathbf{x}_i=a_{k1}x_{i1}+a_{k2}x_{i2}+\cdots+a_{kp}x_{ip}\quad i=1,2,\cdots,n\quad k=1,2,\cdots,p$$

tienen media muestral $\mathbf{a}_k^T\overline{\mathbf{x}}$ y varianza muestral $\mathbf{a}_k^T\mathbf{S}\mathbf{a}_k$. También, los pares de valores $\left(\mathbf{a}_j^T\mathbf{x}_i,\mathbf{a}_k^T\mathbf{x}_i\right)$, para dos combinaciones lineales, tienen covarianza muestral $\mathbf{a}_j^T\mathbf{S}\mathbf{a}_k$.

Las componentes principales muestrales se definen como esas combinaciones que tienen máxima varianza muestral.

La primera componente principal muestral es la combinación lineal $\mathbf{a}_1^T\mathbf{x}_i$ que maximiza su varianza muestral $\mathbf{a}_1^T\mathbf{S}\mathbf{a}_1$ sujeta a $\mathbf{a}_1^T\mathbf{a}_1=1$ para evitar la indeterminación de que $\mathbf{a}_1^T\mathbf{S}\mathbf{a}_1$ aumenta al multiplicar $\mathbf{a}_1$ por una constante.

La segunda componente muestral es la combinación lineal $\mathbf{a}_2^T\mathbf{x}_i$ que maximiza su varianza muestral $\mathbf{a}_2^T\mathbf{S}\mathbf{a}_2$ sujeta a $\mathbf{a}_2^T\mathbf{a}_2=1$ y a que la covarianza muestral del par $\left(\mathbf{a}_1^T\mathbf{x}_i,\mathbf{a}_2^T\mathbf{x}_i\right)$ es cero, $\mathbf{a}_1^T\mathbf{S}\mathbf{a}_2=0$.

En general, la k-ésima componente principal de la muestra es la combinación $\mathbf{a}_k^T\mathbf{x}_i$ que maximiza su varianza muestral $\mathbf{a}_k^T\mathbf{S}\mathbf{a}_k$ sujeta a $\mathbf{a}_k^T\mathbf{a}_k=1$ y a que la covarianza de los pares $\left(\mathbf{a}_j^T\mathbf{x}_i,\mathbf{a}_k^T\mathbf{x}_i\right)$ es cero, $\mathbf{a}_j^T\mathbf{S}\mathbf{a}_k=0\quad\forall\, j<k$.

De manera equivalente a las Proposiciones 1.1, 1.2, 1.3 y 1.4 para las componentes de la población, la siguiente proposición describe la obtención y las diferentes propiedades de las componentes principales muestrales.

Proposición 1.7 *Si* $\mathbf{S}=\left[s_{jk}\right]$ *es la matriz de covarianzas muestral de dimensión* $(p\times p)$ *cuyos pares de autovalores-autovectores son* $(l_1,\mathbf{u}_1),(l_2,\mathbf{u}_2),\cdots,(l_p,\mathbf{u}_p)$ *con* $l_1\geq l_2\geq\cdots\geq l_p\geq 0$*, entonces la k-ésima componente principal muestral está dada por*

$$\hat{y}_k=\mathbf{u}_k^T\mathbf{x}=u_{k1}x_1+u_{k2}x_2+\cdots+u_{kp}x_p \qquad k=1,2,\cdots,p \tag{1.29}$$

donde $\mathbf{x}$ *es cualquier observación sobre las variables* $x_1,x_2,\cdots,x_p$*.*

En estas condiciones

$$\begin{aligned}\operatorname{var}(\hat{y}_k)&=l_k \qquad k=1,2,\cdots,p\\ \operatorname{cov}(\hat{y}_j,\hat{y}_k)&=0 \qquad j\neq k\end{aligned} \tag{1.30}$$

Además,

$$\textit{Varianza total muestral}=\operatorname{trace}(\mathbf{S})=\sum_{k=1}^{p}s_{kk}=\sum_{k=1}^{p}l_k \tag{1.31}$$

el coeficiente de correlación muestral entre $\hat{y}_j$ *y* x_k *es*

$$r_{\hat{y}_j,x_k}=\frac{u_{jk}\sqrt{l_j}}{\sqrt{s_{kk}}} \qquad j,k=1,2,\cdots,p \tag{1.32}$$

y la descomposición espectral de la matriz $\mathbf{S}$ *está dada por*

$$\mathbf{S}=l_1\mathbf{u}_1\mathbf{u}_1^T+l_2\mathbf{u}_2\mathbf{u}_2^T+\cdots+l_p\mathbf{u}_p\mathbf{u}_p^T \tag{1.33}$$

Demostración. Seguir las demostraciones de las Proposiciones 1.1, 1.2, 1.3 y 1.4. ■

La matriz de covarianzas muestral $\mathbf{S}$ es simétrica y definida positiva con lo cual es diagonalizable ortogonalmente, $\mathbf{S}=\mathbf{PLP}^T$ donde $\mathbf{L}=\operatorname{diag}(l_1,l_2,\cdots,l_p)$ siendo $l_1\geq l_2\geq\cdots\geq l_p\geq 0$ los autovalores de $\mathbf{S}$ y $\mathbf{P}=\left[\mathbf{u}_1,\mathbf{u}_2,\cdots,\mathbf{u}_p\right]$ es una matriz ortogonal, $\mathbf{PP}^T=\mathbf{P}^T\mathbf{P}=\mathbf{I}$, así, los autovectores $\mathbf{u}_1,\mathbf{u}_2,\cdots,\mathbf{u}_p$ son ortonormales.

Sea la matriz $\hat{\mathbf{Y}}=\left[\hat{y}_{ik}\right]$ de dimensión $(n\times p)$ cuyo elemento (i,k) es y_{ik} y representa el valor de la k-ésima componente principal para la observación i-ésima de la muestra o, igualmente, $\hat{\mathbf{Y}}=\left[\hat{\mathbf{y}}_1,\hat{\mathbf{y}}_2,\cdots,\hat{\mathbf{y}}_n\right]^T$ donde $\hat{\mathbf{y}}_i=\left(\hat{y}_{i1},\hat{y}_{i2},\cdots,\hat{y}_{ip}\right)^T$ es el vector que contiene los valores de las componentes principales muestrales para la observación i-ésima de la muestra. Entonces, según esta última proposición, los valores de las componentes principales muestrales son una transformación lineal ortogonal de los valores observados de las variables que se puede expresar matricialmente como

$$\hat{\mathbf{Y}}=\mathbf{XP} \tag{1.34}$$

Con frecuencia los vectores de observaciones $\mathbf{x}_i$ se centran respecto a la media muestral $\bar{\mathbf{x}}$. Esto no afecta a la matriz de covarianzas $\mathbf{S}$ y la k-ésima componente principal muestral se obtiene como

$$\hat{y}_k = \mathbf{u}_k^T \left(\mathbf{x} - \bar{\mathbf{x}} \right) \tag{1.35}$$

para cualquier vector de observaciones $\mathbf{x}$. Si el arbitrario vector $\mathbf{x}$ en (1.35) se sustituye por cada uno de los vectores de observaciones $\mathbf{x}_i$, se obtienen los valores muestrales de la k-ésima componente

$$\hat{y}_{ik} = \mathbf{u}_k^T \left(\mathbf{x}_i - \bar{\mathbf{x}} \right) \qquad i = 1, 2, \cdots, n$$

y la media muestral para cada componente principal es cero

$$\bar{\hat{y}}_k = \frac{1}{n} \sum_{i=1}^{n} y_{ik} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{u}_k^T \left(\mathbf{x}_i - \bar{\mathbf{x}} \right) = \frac{1}{n} \mathbf{u}_k^T \sum_{i=1}^{n} \left(\mathbf{x}_i - \bar{\mathbf{x}} \right) = \frac{1}{n} \mathbf{u}_k^T \mathbf{0} = 0$$

En cualquier caso, a partir de (1.31), la ratio dada por

$$\frac{l_k}{\sum_{j=1}^{p} l_j}$$

indica la varianza total de la muestra debida a la k-ésima componente principal.

1.3.2 Interpretación de las componentes principales muestrales

En primer lugar, aunque no es necesario asumir gaussianidad para los resultados de la Proposición 1.7, se supone que la distribución subyacente de $\mathbf{x}$ es $N_p \left(\boldsymbol{\mu}, \boldsymbol{\Sigma} \right)$. Entonces las componentes principales muestrales $\hat{y}_k = \mathbf{u}_k^T \left(\mathbf{x} - \bar{\mathbf{x}} \right)$ son realizaciones de las componentes principales poblacionales $y_k = \boldsymbol{\upsilon}_k^T \left(\mathbf{x} - \boldsymbol{\mu} \right)$ que tienen una distribución $N_p \left(\mathbf{0}, \boldsymbol{\Lambda} \right)$ donde $\boldsymbol{\Lambda} = \operatorname{diag} \left(\lambda_1, \lambda_2, \cdots, \lambda_p \right)$ y $\left(\lambda_k, \boldsymbol{\upsilon}_k \right)$ $k = 1, 2, \cdots, p$ son los pares de autovalores-autovectores de la matriz $\boldsymbol{\Sigma}$.

Los valores muestrales $\mathbf{x}_i$ aproximan $\boldsymbol{\mu}$ con $\bar{\mathbf{x}}$ y $\boldsymbol{\Sigma}$ con $\mathbf{S}$. Puesto que $\mathbf{S}$ es una matriz simétrica, la superficie formada por todos los vectores $\mathbf{x}$ que verifican

$$\left(\mathbf{x} - \bar{\mathbf{x}} \right)^T \mathbf{S}^{-1} \left(\mathbf{x} - \bar{\mathbf{x}} \right) = c^2 \tag{1.36}$$

estima el contorno de densidad constante $\left(\mathbf{x} - \boldsymbol{\mu} \right)^T \boldsymbol{\Sigma}^{-1} \left(\mathbf{x} - \boldsymbol{\mu} \right) = c^2$ de la distribución gaussiana subyacente.

Geométricamente, las observaciones x_{ik} pueden verse como n puntos en un espacio p-dimensional. Entonces, los datos se pueden expresar en un nuevo sistema de

coordenadas que coinciden con los ejes del elipsoide p-dimensional definido en (1.36) que son los autovectores de $\mathbf{S}$. De esta forma, las longitudes de los ejes de este elipsoide son proporcionales a $l_1^{1/2}, l_2^{1/2}, \cdots, l_p^{1/2}$ donde $l_1 \geq l_2 \geq \cdots \geq l_p \geq 0$ son los autovalores de la matriz $\mathbf{S}$.

Además, las componentes principales muestrales $\hat{y}_k = \mathbf{u}_k^T (\mathbf{x} - \bar{\mathbf{x}})$ se encuentran sobre los ejes del elipsoide y sus valores absolutos son las longitudes de las proyecciones de $\mathbf{x} - \bar{\mathbf{x}}$ sobre los ejes en las direcciones de $\mathbf{u}_k$ porque los vectores $\mathbf{u}_k$ son unitarios.

Por todo ello, las componentes principales pueden verse como el resultado de trasladar el origen del sistema de coordenadas original a $\bar{\mathbf{x}}$ y a continuación rotar los ejes de coordenadas hasta que se sitúen en las direcciones perpendiculares de máxima varianza según las observaciones muestrales.

Dibujar los contornos aproximados de densidad constante sobre un gráfico de dispersión puede resultar de gran ayuda para determinar las componentes principales de la muestra. Así, para el caso $p = 2$, en las Figuras 1.2 y 1.3 se ilustra la interpretación geométrica anterior. La Figura 1.2 muestra una elipse de densidad constante, centrada en $\bar{\mathbf{x}}$, con $l_1 > l_2$. Las componentes principales muestrales se sitúan sobre los ejes de la elipse en las direcciones perpendiculares de máxima varianza. En este caso, las componentes están bien determinadas.

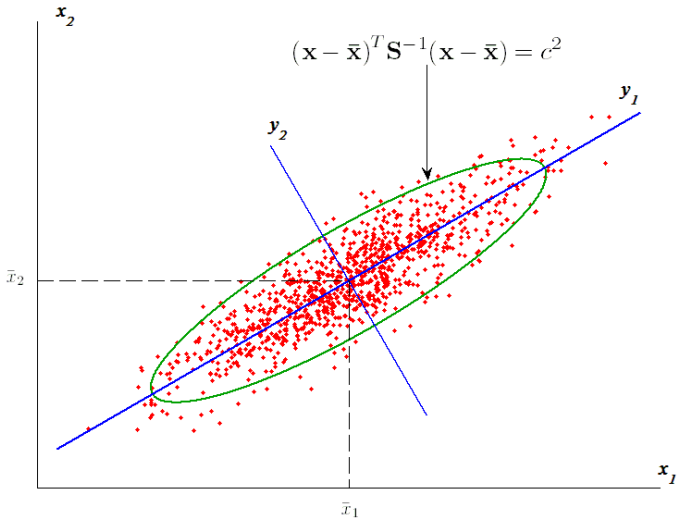

Figura 1.2 Componentes principales muestrales con $l_1 > l_2$.

El caso inverso se ilustra en la Figura 1.3 en la que se muestra una elipse, centrada en $\bar{\mathbf{x}}$, con $l_1 \approx l_2$. Cuando $l_1 = l_2$, la elipse degenera en un círculo de densidad constante y

los ejes no están determinados de forma única ya que se pueden situar en cualquier par de direcciones perpendiculares incluidas las direcciones de los ejes de coordenadas originales. Por ello, cuando los autovalores de $\mathbf{S}$ son aproximadamente iguales, las componentes principales muestrales están mal determinadas y no es posible representar bien los datos con menos de p dimensiones.

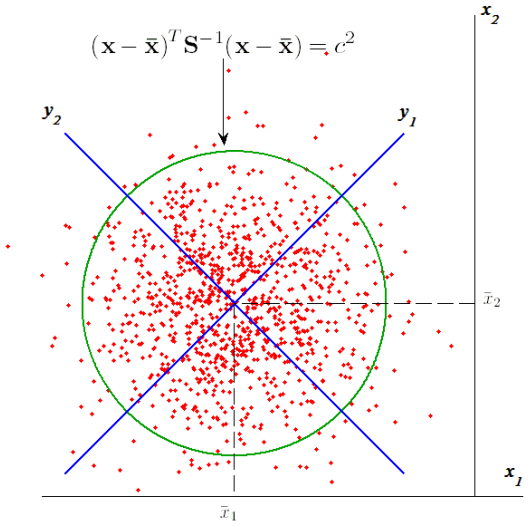

Figura 1.3 Componentes principales muestrales con $l_1 = l_2$.

1.3.3 Componentes principales muestrales estandarizadas

Según se comentó en el tratamiento de las componentes principales poblacionales, las componentes principales no son invariantes ante cambios de escala. Por ello, cuando las variables están medidas en diferentes escalas se estandarizan con frecuencia. Para la muestra, la estandarización se realiza con las siguientes transformaciones para las variables de la i-ésima observación

$$z_{ik} = \frac{x_{ik} - \bar{x}_k}{\sqrt{s_{kk}}} \qquad k = 1, 2, \cdots, p$$

dando lugar al vector i-ésimo de observaciones estandarizadas $\mathbf{z}_i = \left(z_{i1}, z_{i2}, \cdots, z_{ip}\right)^T = \mathbf{D}^{-1/2}\left(\mathbf{x}_i - \bar{\mathbf{x}}\right) \quad i = 1, 2, \cdots, n$ con $\mathbf{D}^{1/2} = \text{diag}\left(\sqrt{s_{11}}, \sqrt{s_{22}}, \cdots, \sqrt{s_{pp}}\right)$ y a la matriz $\mathbf{Z} = \left[z_{ik}\right]$ de dimensión $\left(n \times p\right)$ con las observaciones estandarizadas.

La matriz $\mathbf{Z}$ produce el vector de medias muestral

$$\bar{\mathbf{z}} = \frac{1}{n}\mathbf{Z}^T\mathbf{1} = \mathbf{0} \tag{1.37}$$

y la matriz de covarianzas muestral

$$\mathbf{S}_z = \frac{1}{n-1}\mathbf{Z}^T\mathbf{Z} = \mathbf{R} \tag{1.38}$$

donde el elemento (i, j) de la matriz $\mathbf{R}$ está dado por

$$r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}}\sqrt{s_{jj}}}$$

Las componentes principales muestrales de las observaciones estandarizadas $\mathbf{Z}$, se obtienen de la matriz de correlación $\mathbf{R}$ que es la matriz de covarianzas de los datos originales $\mathbf{X}$. Los resultados obtenidos en la Proposición 1.7 se pueden aplicar con alguna simplificación porque la varianza de cada variable estandarizada es la unidad.

Proposición 1.8 *Si* $\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_p$ *son los vectores de observaciones estandarizadas cuya matriz de covarianzas muestral* $\mathbf{R}$ *de dimensión* $(p \times p)$ *tiene como pares de autovalores-autovectores a* $(l_1, \mathbf{u}_1), (l_2, \mathbf{u}_2), \cdots, (l_p, \mathbf{u}_p)$ *con* $l_1 \geq l_2 \geq \cdots \geq l_p \geq 0$*, entonces la k-ésima componente principal muestral está dada por*

$$\hat{y}_k = \mathbf{u}_k^T\mathbf{z} = \mathbf{u}_k^T\mathbf{D}^{-1/2}(\mathbf{x} - \bar{\mathbf{x}}) = u_{k1}z_1 + u_{k2}z_2 + \cdots + u_{kp}z_p \qquad k = 1, 2, \cdots, p \tag{1.39}$$

donde $\mathbf{z}$ *es cualquier observación sobre las variables estandarizadas* $z_1, z_2, \cdots, z_p$ *.*

En estas condiciones

$$\begin{aligned} \operatorname{var}(\hat{y}_k) &= l_k \qquad k = 1, 2, \cdots, p \\ \operatorname{cov}(\hat{y}_j, \hat{y}_k) &= 0 \qquad j \neq k \end{aligned} \tag{1.40}$$

Además,

$$\textit{Varianza total muestral} = \operatorname{trace}(\mathbf{R}) = p = \sum_{k=1}^{p} l_k \tag{1.41}$$

el coeficiente de correlación muestral entre $\hat{y}_j$ *y* z_k *es*

$$r_{\hat{y}_j, z_k} = u_{jk}\sqrt{l_j} \qquad j, k = 1, 2, \cdots, p \tag{1.42}$$

Demostración. En la Proposición 1.7 basta con sustituir $\mathbf{x}$ por $\mathbf{z}$ y $\mathbf{S}$ por $\mathbf{R}$. ■

Haciendo uso de (1.41), la proporción del total de la varianza muestral explicada por la k-ésima componente está dada por

$$\frac{l_k}{p} \qquad k = 1, 2, \cdots, p \tag{1.43}$$

1.3.4 Distribuciones de probabilidad para las componentes principales muestrales

Se ha realizado un gran esfuerzo en derivar las distribuciones de probabilidad, la mayoría asintóticas, para los coeficientes de las componentes principales muestrales y para sus varianzas muestrales o, equivalentemente, para los autovectores y autovalores de la matriz de covarianzas muestral. En este apartado se exponen sucintamente los resultados teóricos obtenidos por Anderson (1963) y Girshick (1939).

Se asume que el vector aleatorio $\mathbf{x}$ tiene una distribución gaussiana p-dimensional $N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$. Aunque $\boldsymbol{\mu}$ no necesita ser dada, se asume que $\boldsymbol{\Sigma}$ es conocida.

Sean $(l_k,\mathbf{u}_k)\quad k=1,2,\cdots,p$ los pares de autovalores-autovectores de la matriz $\mathbf{S}$ y sean $(\lambda_k,\boldsymbol{\upsilon}_k)\quad k=1,2,\cdots,p$ los pares de autovalores-autovectores de la matriz $\boldsymbol{\Sigma}$. También, sean $\mathbf{l}=(l_1,l_2,\cdots,l_p)^T$ y $\boldsymbol{\lambda}=(\lambda_1,\lambda_2,\cdots,\lambda_p)^T$ los vectores con los autovalores de $\mathbf{S}$ y $\boldsymbol{\Sigma}$ respectivamente y sean u_{kj} y υ_{kj} los elementos j-ésimos de los autovectores $\mathbf{u}_k$ y $\boldsymbol{\upsilon}_k$ respectivamente. Así mismo, se asume que los autovalores de la población son positivos y distintos, es decir $\lambda_1>\lambda_2>\cdots>\lambda_p>0$. Entonces, los siguientes resultados se alcanzan de forma asintótica:

1) Todos los autovalores l_k son independientes de todos los autovectores $\mathbf{u}_k$.

2) Tanto $\mathbf{l}$ como $\mathbf{u}_k$ tienen una distribución gaussiana multivariante.

3) Las esperanzas de los vectores $\mathbf{l}$ y $\mathbf{u}_k$ son

$$\begin{aligned} E(\mathbf{l}) &= \boldsymbol{\lambda} \\ E(\mathbf{u}_k) &= \boldsymbol{\upsilon}_k \quad k=1,2,\cdots,p \end{aligned} \tag{1.44}$$

4) Las covarianzas de los vectores $\mathbf{l}$ y $\mathbf{u}_k$ están dadas por

$$\mathrm{Cov}(l_k,l_{k'})=\begin{cases} \dfrac{2\lambda_k^2}{n-1} & k=k' \\ 0 & k\neq k' \end{cases} \tag{1.45}$$

$$\mathrm{Cov}(u_{kj},u_{k'j'})=\begin{cases} \dfrac{\lambda_k}{n-1}\displaystyle\sum_{\substack{i=1\\ i\neq k}}^{p}\frac{\lambda_i\upsilon_{ij}\upsilon_{ij'}}{(\lambda_i-\lambda_k)^2} & k=k' \\ \dfrac{-\lambda_k\lambda_{k'}\upsilon_{kj}\upsilon_{k'j'}}{(n-1)(\lambda_k-\lambda_{k'})^2} & k\neq k' \end{cases} \tag{1.46}$$

En Anderson (1963) se encuentra una extensión al caso expuesto donde algunos de los autovalores λ_k pueden ser iguales pero siempre positivos.

1.3.5 Inferencias sobre el modelo de componentes principales

Las distribuciones descritas en el apartado anterior se pueden utilizar para realizar inferencias sobre las componentes principales de la población siempre y cuando los supuestos necesarios sean válidos. El principal supuesto, el vector aleatorio $\mathbf{x}$ tiene una distribución gaussiana multivariante, no se cumple con frecuencia y limita por ello el valor de los resultados. Se podría argumentar que el PCA debería realizarse con datos que son, al menos aproximadamente, gaussianos multivariantes porque, solo entonces, se pueden realizar inferencias correctas respecto a las componentes principales subyacentes de la población. No obstante, esto ofrece una visión más estrecha de lo que puede hacer el PCA ya que es una herramienta de una amplia utilidad cuyo uso principal es más descriptivo que inferencial. Puede proporcionar información valiosa para una gran variedad de datos siempre que las variables sean continuas aunque no tengan una distribución gaussiana.

- **Estimación puntual**

El estimador de máxima verosimilitud (MLE) para la matriz de covarianzas $\boldsymbol{\Sigma}$ de una distribución gaussiana multivariante es $\frac{n-1}{n}\mathbf{S}$, ver, por ejemplo, Johnson y Wichern (2007) para su derivación. Si $\boldsymbol{\lambda}$, $\mathbf{l}$, $\boldsymbol{\upsilon}_k$ y $\mathbf{u}_k$ se definen como en el apartado anterior, entonces los MLEs de $\boldsymbol{\lambda}$ y $\boldsymbol{\upsilon}_k$, $k=1,2,\cdots,p$, pueden obtenerse del MLE para $\boldsymbol{\Sigma}$ y son $\hat{\boldsymbol{\lambda}}=\frac{n-1}{n}\mathbf{l}$ y $\hat{\boldsymbol{\upsilon}}_k=\mathbf{u}_k$, $k=1,2,\cdots,p$, suponiendo que los elementos de $\boldsymbol{\lambda}$ son distintos y positivos. El MLE para λ_k, al igual que el MLE para $\boldsymbol{\Sigma}$, es sesgado pero asintóticamente insesgado.

En el caso que algunos λ_k sean iguales, el MLE para el valor común es la media de los correspondientes l_k multiplicada por $\frac{n-1}{n}$. Los MLEs de los $\boldsymbol{\upsilon}_k$ correspondientes a los λ_k iguales no son únicos. Esto se debe a que la matriz de dimensión $(p\times q)$ cuyas columnas son los MLEs de los $\boldsymbol{\upsilon}_k$ correspondientes a los λ_k iguales se puede multiplicar por otra matriz ortogonal de dimensión $(q\times q)$, donde q es la multiplicidad de los autovalores iguales, para obtener otro conjunto de MLEs.

La mayoría de las veces, las estimaciones puntuales de $\boldsymbol{\lambda}$ y $\boldsymbol{\upsilon}_k$ son simplemente dadas por $\mathbf{l}$ y $\mathbf{u}_k$ respectivamente y rara vez se acompañan de sus errores estándar.

Si no se puede asumir gaussianidad multivariante y no existe una distribución alternativa, entonces puede ser deseable obtener estimaciones robustas para las componentes principales.

- **Estimación por intervalos**

Las distribuciones marginales asintóticas dadas en el apartado anterior se pueden utilizar para construir intervalos de confianza.

Para l_k, a partir de (1.44) y (1.45), la distribución marginal es aproximadamente

$$l_k \sim N\left(\lambda_k, \frac{2\lambda_k^2}{n-1}\right)$$

que permite construir un intervalo de confianza para λ_k con coeficiente de confianza $(1-\alpha)$ de la forma

$$\frac{l_k}{\left(1+z_{\alpha/2}\sqrt{\frac{2}{n-1}}\right)^{1/2}} < \lambda_k < \frac{l_k}{\left(1-z_{\alpha/2}\sqrt{\frac{2}{n-1}}\right)^{1/2}} \tag{1.47}$$

donde $z_{\alpha/2}$ es el $100\alpha/2$ percentil superior de la distribución gaussiana $N(0,1)$.

Los l_k son asintóticamente independientes y, por ello, una región de confianza conjunta para un número m de λ_k con coeficiente de confianza $(1-\alpha)$ se obtiene sustituyendo $z_{\alpha/2}$ por $z_{\alpha/2m}$ en la expresión (1.47).

Para $\mathbf{u}_k$, a partir de (1.44) y (1.46), la distribución marginal es aproximadamente

$$\mathbf{u}_k \sim N(\boldsymbol{\upsilon}_k, \mathbf{T}_k)$$

donde

$$\mathbf{T}_k = \frac{\lambda_k}{n-1}\sum_{\substack{i=1 \\ i\neq k}}^{p} \frac{\lambda_i}{(\lambda_i-\lambda_k)^2}\boldsymbol{\upsilon}_i\boldsymbol{\upsilon}_i^T$$

La matriz $\mathbf{T}_k$ tiene rango $p-1$ con lo cual posee un autovalor nulo que corresponde al autovector $\boldsymbol{\upsilon}_k$. Aunque esto causa más complicaciones, Mardia et al (1979) demuestran que, aproximadamente

$$(n-1)(\mathbf{u}_k - \boldsymbol{\upsilon}_k)^T\left(l_k\mathbf{S}^{-1} + l_k^{-1}\mathbf{S} - 2\mathbf{I}_p\right)(\mathbf{u}_k - \boldsymbol{\upsilon}_k) \sim \chi^2_{p-1} \tag{1.48}$$

Puesto que $(l_k, \mathbf{u}_k)$ es un par autovalor-autovector de la matriz $\mathbf{S}$ se sigue que $l_k^{-1}\mathbf{S}\mathbf{u}_k = l_k^{-1} l_k \mathbf{u}_k = \mathbf{u}_k$, $l_k \mathbf{S}^{-1}\mathbf{u}_k = l_k l_k^{-1}\mathbf{u}_k = \mathbf{u}_k$ y $\left(l_k \mathbf{S}^{-1} + l_k^{-1}\mathbf{S} - 2\mathbf{I}_p\right)\mathbf{u}_k = \mathbf{u}_k + \mathbf{u}_k - 2\mathbf{u}_k = \mathbf{0}$ con lo cual el resultado (1.48) se reduce a

$$(n-1)\boldsymbol{\upsilon}_k{}^T \left(l_k \mathbf{S}^{-1} + l_k^{-1}\mathbf{S} - 2\mathbf{I}_p\right)\boldsymbol{\upsilon}_k \sim \chi^2_{p-1} \tag{1.49}$$

Así, una región de confianza aproximada para $\boldsymbol{\upsilon}_k$, con coeficiente de confianza $(1-\alpha)$, a partir de (1.49) tiene la forma

$$(n-1)\boldsymbol{\upsilon}_k{}^T \left(l_k \mathbf{S}^{-1} + l_k^{-1}\mathbf{S} - 2\mathbf{I}_p\right)\boldsymbol{\upsilon}_k \le \chi^2_{p-1,\alpha} \tag{1.50}$$

donde $\chi^2_{p-1,\alpha}$ es el 100α percentil superior de la distribución χ^2 de $p-1$ grados de libertad.

Para los coeficientes individuales υ_{kj} se pueden construir intervalos de confianza a partir de las distribuciones marginales de los coeficientes u_{kj}, que se derivan de (1.44) y (1.46), de manera similar a los intervalos de confianza para los autovalores λ_k.

- **Contraste de hipótesis**

Los resultados obtenidos de (1.44) a (1.46) sirven también para realizar contrastes de hipótesis.

Así, para los autovalores λ_k, si se desea contrastar las hipótesis

$$H_0: \quad \lambda_k = \lambda_{k0}$$
$$H_1: \quad \lambda_k \neq \lambda_{k0}$$

el estadístico que se utiliza es

$$\frac{l_k - \lambda_{k0}}{\lambda_{k0}\sqrt{\dfrac{2}{n-1}}}$$

que tiene, aproximadamente, una distribución gaussiana estándar $N(0,1)$ bajo la hipótesis H_0. Por ello, la hipótesis H_0 será rechazada al nivel de significación α si

$$\left|\frac{l_k - \lambda_{k0}}{\lambda_{k0}\sqrt{\dfrac{2}{n-1}}}\right| > z_{\alpha/2}$$

Similarmente, el resultado (1.49) permite el contraste de las hipótesis

$$H_0: \quad \boldsymbol{\upsilon}_k = \boldsymbol{\upsilon}_{k0}$$
$$H_1: \quad \boldsymbol{\upsilon}_k \neq \boldsymbol{\upsilon}_{k0}$$

de tal manera que la hipótesis H_0 será rechazada al nivel de significación α si

$$(n-1)\mathbf{v}_{k0}{}^{T}\left(l_k\mathbf{S}^{-1}+l_k^{-1}\mathbf{S}-2\mathbf{I}_p\right)\mathbf{v}_{k0}>\chi^2_{p-1,\alpha}$$

donde $\chi^2_{p-1,\alpha}$ es el 100α percentil superior de la distribución χ^2 de $p-1$ grados de libertad.

Existen contrastes sobre los tipos de estructuras en la matriz de covarianzas $\mathbf{\Sigma}$. El más conocido es el contraste de la hipótesis nula donde los últimos $(p-q)$ autovalores son iguales frente a la alternativa donde al menos dos de los últimos $(p-q)$ son distintos, es decir,

$$H_{0,q}: \quad \lambda_{q+1}=\lambda_{q+2}=\cdots=\lambda_p$$
$$H_{1,q}: \quad \exists a,b \ con \ 1\le a\ne b\le p-q \,/\, \lambda_{q+a}\ne\lambda_{q+b}$$

Este contraste, desde Hotelling (1933), ha sido considerado por varios autores como Bartlett (1950) y se justifica porque puede que cada una de las primeras q componentes principales mida un elemento esencial de la variación en el vector aleatorio $\mathbf{x}$ y las últimas $(p-q)$ componentes principales tienen igual variación y esencialmente miden ruido. Un estadístico para contrastar $H_{0,q}$ frente a $H_{1,q}$ se puede construir a partir de la ratio de verosimilitud (LR), asumiendo una distribución gaussiana multivariante, y tiene la forma

$$Q=\left[\frac{\prod_{k=q+1}^{p} l_k}{\left(\frac{1}{p-q}\sum_{k=q+1}^{p} l_k\right)^{p-q}}\right]^{\frac{n}{2}}$$

La distribución exacta de Q es complicada, pero se utiliza el conocido resultado sobre los contrastes LR, es decir que $-2\ln(Q)$ tiene aproximadamente una distribución χ^2 con ν grados de libertad que Mardia et al (1979) han calculado y que resultan ser $\nu=\frac{1}{2}(p-q+2)(p-q-1)$. Así, bajo la hipótesis $H_{0,q}$, se obtiene que

$$n\left[(p-q)\ln(\bar{l})-\sum_{k=q+1}^{p}\ln(l_k)\right]\sim\chi^2_{\nu} \tag{1.51}$$

donde

$$\bar{l}=\frac{1}{p-q}\sum_{k=q+1}^{p} l_k\,.$$

No obstante, la aproximación dada en (1.51) puede ser mejorada si se sustituye n por $n' = n - (2p+11)/6$. Así, la hipótesis $H_{0,q}$ será rechazada al nivel de significación α si

$$n'\left[(p-q)\ln(\bar{l}) - \sum_{k=q+1}^{p}\ln(l_k)\right] > \chi^2_{\nu,\alpha} \tag{1.52}$$

donde $\chi^2_{\nu,\alpha}$ es el 100α percentil superior de la distribución χ^2 de ν grados de libertad.

El caso especial donde todas las variables están igualmente correlacionadas, $\mathrm{Corr}(x_j, x_k) = \rho \quad \forall\, j \neq k$, equivale a la hipótesis $H_{0,q}$ cuando $q = 0$, en cuyo caso todos los autovalores de $\boldsymbol{\Sigma}$ son iguales y los resultados previos sobre inferencia no pueden aplicarse. Para contrastar esta estructura se establecen las hipótesis

$$\begin{aligned} H_0 &: \quad \boldsymbol{\rho} = \boldsymbol{\rho}_0 = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix} \\ H_1 &: \quad \boldsymbol{\rho} \neq \boldsymbol{\rho}_0 \end{aligned} \tag{1.53}$$

El contraste de H_0 frente a H_1 se puede basar en el estadístico Q, pero Lawley (1963) demostró que se puede construir un test equivalente con los elementos no diagonales de la matriz $\mathbf{R}$. El estadístico de Lawley requiere el cálculo previo de las cantidades

$$\bar{r}_k = \frac{1}{p-1}\sum_{\substack{i=1 \\ i\neq k}}^{p} r_{ik} \quad k = 1, 2, \cdots, p; \quad \bar{r} = \frac{2}{p(p-1)}\sum\sum_{i<k} r_{ik}\,; \quad \hat{\gamma} = \frac{(p-1)^2\left[1-(1-\bar{r})^2\right]}{p-(p-2)(1-\bar{r})^2}$$

Evidentemente, $\bar{r}_k$ es la media de los elementos no diagonales de la k-ésima fila o columna de la matriz $\mathbf{R}$ y $\bar{r}$ es la media global de los elementos no diagonales. La aproximación asintótica que rechaza la hipótesis H_0 al nivel de significación α es

$$\frac{n-1}{(1-\hat{\gamma})^2}\left[\sum\sum_{i<k}(r_{ik}-\bar{r})^2 - \hat{\gamma}\sum_{k=1}^{p}(\bar{r}_k-\bar{r})^2\right] > \chi^2_{(p+1)(p-2)/2,\alpha} \tag{1.54}$$

donde $\chi^2_{(p+1)(p-2)/2,\alpha}$ es el 100α percentil superior de la distribución χ^2 de $(p+1)(p-2)/2$ grados de libertad.

El caso extremo donde todas las variables están incorrelacionadas corresponde al contraste cuyas hipótesis están dadas por (1.53) estableciendo $\boldsymbol{\rho}_0 = \mathbf{I}_p$ donde $\mathbf{I}_p$ es la matriz identidad de dimensión $(p \times p)$.

1.4 Elección del número de componentes

En esta sección se presentan, siguiendo a Jolliffe (2002), varias reglas para decidir cuántas componentes principales deberían retenerse con el fin de explicar la mayor variación posible del conjunto $\mathbf{x}$ de p variables. La idea es intentar reducir la dimensión remplazando las p variables por las primeras m componentes principales $(m < p)$ ignorando las restantes.

1.4.1 Porcentaje acumulado de la variación total

El criterio más obvio para elegir m es seleccionar un porcentaje acumulado de la variación total que se desea que expliquen las componentes principales seleccionadas. El número necesario de componentes es el menor valor de m para el cual se sobrepasa el porcentaje elegido. Puesto que las componentes se seleccionan sucesivamente de forma que tengan la mayor varianza posible l_k y que $\sum_{k=1}^{p} l_k = \sum_{j=1}^{p} s_{jj}$, el porcentaje de la variación que acumulan las primeras m componentes es

$$t_m = 100 \sum_{k=1}^{m} l_k \Big/ \sum_{j=1}^{p} s_{jj} = 100 \sum_{k=1}^{m} l_k \Big/ \sum_{k=1}^{p} l_k$$

que se reduce a $t_m = \frac{100}{p} \sum_{k=1}^{m} l_k$ en el caso de utilizar la matriz de correlaciones.

La regla consiste en seleccionar un punto de corte t^* entre el 70% y 90% y conservar las primeras m componentes principales, donde m es el menor entero para el cual $t_m > t^*$, de forma que explican la mayor parte de la información en $\mathbf{x}$.

El uso de esta regla es equivalente a examinar la descomposición del valor singular (SVD) de la matriz $\tilde{\mathbf{X}}$ de dimensión $(n \times p)$ con las desviaciones respecto a las medias (ver Sección 1.3). Sea $\tilde{\mathbf{X}} = \mathbf{AHP}^T$ la SVD de la matriz $\tilde{\mathbf{X}}$, donde $\mathbf{H} = \text{diag}(h_1, h_2, \cdots, h_r)$ contiene los valores singulares siendo $r \leq p$ el rango de $\tilde{\mathbf{X}}$ y donde $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_r]$ de dimensión $(n \times r)$ y $\mathbf{P} = [\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_r]$ de dimensión $(p \times r)$ son matrices con las columnas ortogonales siendo $\mathbf{u}_j$ el autovector j-ésimo de la matriz $\mathbf{S}$. La matriz $\tilde{\mathbf{X}}$ se puede escribir como $\tilde{\mathbf{X}} = \sum_{j=1}^{r} h_j \mathbf{a}_j \mathbf{u}_j^T$ y su elemento $\tilde{x}_{ik}$ es

$$\tilde{x}_{ik} = \sum_{j=1}^{r} h_j a_{ij} u_{kj} = \sum_{j=1}^{r} \tilde{l}_j^{1/2} a_{ij} u_{kj} \tag{1.55}$$

donde $\tilde{l}_j$ es el j-ésimo autovalor de $\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}$ con $\tilde{l}_j = (n-1)l_j$, siendo l_j autovalor de $\mathbf{S}$. Por el teorema de Eckart-Young, la matriz de rango inferior $m<r$ más próxima a $\tilde{\mathbf{X}}$ según la norma de Frobenius está dada por $\tilde{\mathbf{X}}^{(m)} = \sum_{j=1}^{m} h_j \mathbf{a}_j \mathbf{u}_j^T$ (versión truncada de la SVD con m términos) cuyo elemento $\tilde{x}_{ik}^{(m)}$ se puede escribir como

$$\tilde{x}_{ik}^{(m)} = \sum_{j=1}^{m} h_j a_{ij} u_{kj} = \sum_{j=1}^{m} \tilde{l}_j^{1/2} a_{ij} u_{kj} \tag{1.56}$$

Así, aplicando dicho teorema, la distancia de $\tilde{\mathbf{X}}$ a la matriz $\tilde{\mathbf{X}}^{(m)}$ viene dada por

$$\left\| \tilde{\mathbf{X}} - \tilde{\mathbf{X}}^{(m)} \right\|_F^2 = \sum_{i=1}^{n}\sum_{k=1}^{p} \left(\tilde{x}_{ik} - \tilde{x}_{ik}^{(m)} \right)^2 = \sum_{k=m+1}^{p} h_k^2 = \sum_{k=m+1}^{p} \tilde{l}_k = (n-1) \sum_{k=m+1}^{p} l_k$$

De este modo, $\sum_{k=m+1}^{p} l_k$ es una medida apropiada de la falta de ajuste de los primeros m términos para $\tilde{\mathbf{X}}$, con lo cual se demuestra que la SVD de la matriz $\tilde{\mathbf{X}}$ está estrechamente relacionada con el examen de t_m.

1.4.2 Tamaño de las varianzas de las componentes principales

La regla descrita en este apartado, a diferencia de la anterior, está diseñada para usarla con matrices de correlación aunque podría adaptarse a algunos tipos de matrices de covarianzas. La regla se basa en la idea de que si todas las variables del vector $\mathbf{x}$ son independientes entonces las componentes principales son las mismas variables originales y todas tienen varianza unidad en el caso de una matriz de correlación. De este modo, cualquier componente principal con varianza inferior a uno contiene menos información que una variable original y no merece ser retenida. La regla consiste en retener solo las componentes principales cuya varianza l_k exceda de 1, Kaiser (1960).

En ciertas ocasiones, el punto de corte $l^* = 1$ conserva demasiadas pocas componentes. Para razonarlo, se considera una variable que en la población es prácticamente independiente del resto. En una muestra, tal variable tendrá pequeños coeficientes en $(p-1)$ de las componentes principales pero será dominante en la restante con una varianza l_k próxima a 1, usando la matriz de correlación. Puesto que la variable proporciona información independiente de las otras, sería imprudente rechazarla. Sin embargo, debido a la variación muestral puede ser $l_k < 1$ y, debido a la regla, será rechazada. Por ello, es aconsejable elegir un punto de corte menor que 1 para dejar margen a la variación muestral. Jolliffe (1972) propuso que $l^* = 0,7$ es aproximadamente un valor correcto basándose en simulaciones.

La regla descrita se puede adoptar para las matrices de covarianzas tomando como punto de corte $l^* = \overline{l}$, donde $\overline{l}$ es la media aritmética de los autovalores de la matriz $\mathbf{S}$, o, incluso mejor, $l^* = 0.7\overline{l}$.

1.4.3 Gráfico de sedimentación

El gráfico de sedimentación propuesto por Cattell (1966) es simplemente una representación gráfica de los autovalores l_k frente a k. La regla consiste en decidir el valor de k que define de forma clara un "codo" muy marcado, o un cambio de pendiente, en el gráfico y entonces tomar ese valor de k como el número de componentes m que son retenidas. Una alternativa es graficar $\log(l_k)$ frente a k que se conoce como el diagrama del log-autovalor.

Las dos reglas previas tienen un grado de subjetividad pero la presente es incluso más subjetiva. La regla del Apartado 1.4.1 se basa en $t_m = \sum_{k=1}^{m} l_k$, la regla en el Apartado 1.4.2 examina los valores individuales de l_k y la actual usa $l_{k-1} - l_k$ como criterio pero sin un punto de corte formal para la diferencia entre los autovalores. Así, en la Figura 1.4 se muestra un gráfico de sedimentación en el que, claramente, según la regla actual, a partir de $k = 4$ se forma un "codo" con lo cual el número de componentes a retener es $m = 4$, no obstante, la decisión no será siempre tan evidente.

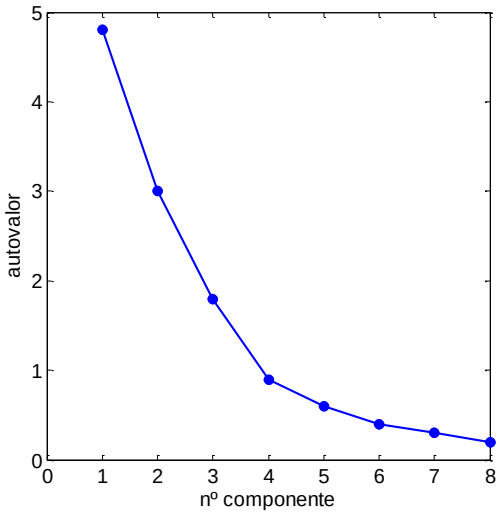

Figura 1.4 Ejemplo de gráfico de sedimentación

1.4.4 El número de componentes con autovalores distintos

En el Apartado 1.3.5 se describió un contraste cuya hipótesis nula establece que los $(p-q)$ últimos autovalores de la matriz $\mathbf{\Sigma}$ son iguales frente a la alternativa de que al menos dos de los últimos $(p-q)$ autovalores son distintos, es decir

$$H_{0,q}: \quad \lambda_{q+1} = \lambda_{q+2} = \cdots = \lambda_p$$
$$H_{1,q}: \quad \exists a,b \ con \ 1 \le a \ne b \le p-q \ / \ \lambda_{q+a} \ne \lambda_{q+b}$$

Si este contraste se utiliza para diferentes valores de q, se puede descubrir cuántas componentes principales contribuyen de manera sustancial a la variación y cuántas son simple ruido. Si el número de componentes principales a retener m se define como el número de componentes que no son ruido, entonces el contraste se utiliza de forma secuencial para obtener el valor de m.

Primero se contrasta $H_{0,p-2}$, es decir $\lambda_{p-1} = \lambda_p$, y si $H_{0,p-2}$ no se rechaza, entonces se contrasta $H_{0,p-3}$. Si $H_{0,p-3}$ no se rechaza, a continuación se contrasta $H_{0,p-4}$ y así sucesivamente hasta la primera $H_{0,q}$ que se rechaza con $q = q^*$. Entonces, se toma $m = q^* + 1$. No obstante, este procedimiento tiene algunos inconvenientes. El primero es que la ecuación (1.51) está basada en el supuesto de que el vector $\mathbf{x}$ tiene una distribución gaussiana multivariante. El segundo problema se refiere al hecho de que, a menos que $H_{0,p-2}$ sea rechazada, se realizan varios contrastes con lo cual el nivel de significación global es diferente al nivel de significación de cada contraste individual. Además, es difícil obtener una idea del nivel de significación global porque le número de contrastes que se realiza no está fijado previamente y los contrastes no son independientes unos de otros.

Sin embargo, el mayor inconveniente, y más práctico, es que en casi todos los ejemplos reales esta posible regla tiende a retener más componentes principales de las necesarias. Así, para matrices de correlaciones, Jolliffe (1970) encontró que esta posible regla se corresponde aproximadamente con un punto de corte l^* entre 0.1 y 0.2, para la regla del Apartado 1.4.2, que es un valor mucho más pequeño que el recomendado.

Aunque más formal, el método de este apartado es similar al gráfico de sedimentación. Buscar la primera pendiente plana en el gráfico corresponde a buscar los dos primeros autovalores consecutivos que son casi iguales. No obstante, el gráfico de sedimentación comienza desde el mayor autovalor comparando autovalores consecutivos dos a dos, mientras que el contraste empieza con los autovalores más pequeños comparando bloques de dos, tres, cuatro y así sucesivamente. Además, el punto que forma el codo es retenido en el gráfico de sedimentación mientras que se excluye en este

procedimiento. El gráfico de sedimentación es más subjetivo pero la objetividad del contraste es una especie de ilusión.

1.4.5 Métodos de validación cruzada

La idea que hay detrás de los dos métodos que se van a describir en este apartado, Wold (1978) y Eastment y Krzanowski (1982), es la aproximación de la matriz $\tilde{\mathbf{X}}$ mediante una matriz de rango inferior basada en la SVD excepto que cada elemento $\tilde{x}_{ik}$ de $\tilde{\mathbf{X}}$ se aproxima con una ecuación como la SVD pero obtenida con una submatriz de $\tilde{\mathbf{X}}$ que no incluye el elemento $\tilde{x}_{ik}$.

En ambos métodos, el número de términos en la estimación de $\tilde{\mathbf{X}}$, es decir el número de componentes principales, se toma igual a 1, 2, ..., y así sucesivamente, hasta que la estimación de $\tilde{x}_{ik}$ no mejora de forma significativa añadiendo componentes adicionales. Entonces, el número de componentes principales que se retiene, m se toma como el mínimo necesario para la estimación aceptable.

Usando la SVD de la matriz $\tilde{\mathbf{X}}$, $\tilde{x}_{ik}$ se puede escribir como en la ecuación (1.55). Una estimación de $\tilde{x}_{ik}$ basada en las primeras m componentes principales usando todos los datos viene dada por

$$\tilde{x}_{ik}^{(m)} = \sum_{j=1}^{m} h_j a_{ij} u_{kj}$$

Sin embargo, se requiere una estimación basada en un subconjunto de los datos que excluya a la observación $\tilde{x}_{ik}$. Esta estimación puede escribirse como

$$\hat{x}_{ik}^{(m)} = \sum_{j=1}^{m} \hat{h}_j \hat{a}_{ij} \hat{u}_{kj} \tag{1.57}$$

donde $\hat{h}_j$, $\hat{a}_{ij}$ y $\hat{u}_{kj}$ se calculan a partir de subconjuntos adecuados de los datos. Entonces, la suma de los cuadrados de las diferencia entre las estimaciones $\hat{x}_{ik}^{(m)}$ y las observaciones $\tilde{x}_{ik}$ es

$$PRESS(m) = \sum_{i=1}^{n} \sum_{k=1}^{p} \left(\tilde{x}_{ik} - \hat{x}_{ik}^{(m)} \right)^2 \tag{1.58}$$

que es equivalente al error cuadrático total en una regresión.

Hasta ahora, todo el razonamiento expuesto es común tanto a Wold (1978) como a Eastment y Krzanowski (1982) pero difieren en cómo elegir el subconjunto de datos para estimar $\tilde{x}_{ik}$ según (1.57) y en cómo utilizar (1.58) para decidir el valor de m.

Eastment y Krzanowski (1982) obtienen $\hat{a}_{ij}$ para (1.57) a partir de la SVD de $\tilde{\mathbf{X}}$ excluyendo la k-ésima variable, también determinan $\hat{u}_{kj}$ a partir de la SVD de $\tilde{\mathbf{X}}$, pero ahora, omitiendo la i-ésima observación y, finalmente, calculan $\hat{h}_j$ como la media geométrica de los valores singulares obtenidos en las dos anteriores SVD para computar $\hat{a}_{ij}$ y $\hat{u}_{kj}$ respectivamente. Wold (1978), por otro lado, divide los datos en g bloques, donde g está entre cuatro y siete pero sin ser divisor de p, sin que ninguno contenga la mayoría de los elementos de una fila o columna de $\tilde{\mathbf{X}}$. Para cada bloque, se calculan las cantidades equivalentes a $\hat{l}_j$, $\hat{a}_{ij}$ y $\hat{u}_{kj}$ y a partir de ellas se estiman las observaciones en el h-ésimo bloque, $h=1,2,\cdots,g$.

Respecto a la elección de m, Wold (1978), para decidir si se incluye la m-ésima componente principal, examina la ratio

$$R=\frac{PRESS(m)}{\sum_{i=1}^{n}\sum_{k=1}^{p}\left(\tilde{x}_{ik}-\tilde{x}_{ik}^{(m-1)}\right)^2}=\frac{PRESS(m)}{(n-1)\sum_{k=m}^{p}l_k} \tag{1.59}$$

Es decir, compara la suma de los cuadrados de los errores de predicción después de ajustar m componentes con la suma de los cuadrados de las diferencias entre los datos observados y los estimados usando $(m-1)$ componentes del conjunto completo de datos. Si $R<1$, implica que se alcanza una mejor predicción usando m en lugar de $(m-1)$ componentes principales, de este modo la componente m-ésima debe incluirse.

El criterio de Eastment y Krzanowski (1982) es similar al usado en el análisis de la varianza. La reducción en la suma de los cuadrados de los errores de predicción al añadir la componente m-ésima al modelo, dividido por sus grados de libertad, se compara con la suma de los cuadrados de los errores de predicción con m componentes principales, después de dividir por sus grados de libertad. De este modo, el criterio es

$$W=\frac{\left[PRESS(m-1)-PRESS(m)\right]/\nu_{m,1}}{PRESS(m)/\nu_{m,2}} \tag{1.60}$$

donde $\nu_{m,1}=n+p-2m$ y $\nu_{m,2}=p(n-1)+m(m+1-n-p)$ son los grados de libertad del numerador y el denominador respectivamente. Si $W>1$, entonces merece la pena incluir la m-ésima componente principal, aunque este punto de corte en la unidad debe considerarse con cierta flexibilidad. Ciertamente, no es apropiado detenerse al añadir componentes principales tan pronto como, en la primera ocasión, (1.61) se sitúe por debajo de la unidad, porque W como función de m no es monotónica decreciente.

Los métodos de validación cruzada para determinar el número de componentes principales son muy expansivos computacionalmente para grandes conjuntos de datos. Besse y Ferré (1993) demostraron que para grandes tamaños muestrales n, $PRESS(m)$ y W son casi equivalentes a las cantidades mucho más simples $(n-1)\sum_{k=m+1}^{p} l_k$ y $l_m \Big/ \sum_{k=m+1}^{p} l_k$, respectivamente.

1.4.6 Correlación parcial

Cuando el PCA se basa en la matriz de correlaciones, Velicer (1976) sugirió que las correlaciones parciales entre las p variables, dadas las primeras m componentes principales, se pueden utilizar para determinar el número de componentes que se retienen. El criterio propuesto es la media aritmética de los cuadrados de las correlaciones parciales

$$V = \sum_{i=1}^{p} \sum_{\substack{j=1 \\ j \neq i}}^{p} \frac{\left(r_{ij}^{*}\right)^2}{p(p-1)}$$

donde r_{ij}^{*} es la correlación parcial entre las variables i-ésima y j-ésima dadas las primeras m componentes principales. El estadístico r_{ij}^{*} se define como la correlación entre los residuos de la regresión lineal de la i-ésima variable sobre las primeras m componentes y los residuos de la correspondiente regresión de la j-ésima variable sobre estas m componentes. Por lo tanto, mide la fuerza de la relación lineal entre las variables i-ésima y j-ésima después de eliminar el efecto común de las primeras m componentes principales.

El criterio V, según m aumenta, primero decrece y después crece. Por ello, Velicer (1976) sugirió que el valor óptimo de m corresponde al que minimiza el valor del criterio V. Sin embargo, este criterio es más convincente para un análisis factorial que para el PCA porque no retendrá las componentes principales dominadas por una sola variable cuya correlación con las demás esté próxima a cero.

1.5 PCA para series temporales

Esta sección trata sobre las implicaciones que para el PCA tiene la no independencia entre los vectores de observaciones $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$, siendo las series temporales el tipo más común de datos no independientes. Los resultados del Apartado 1.3.5, que permiten realizar inferencias formales sobre las componentes principales, se basan en la independencia de $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ así como en la gaussianidad multivariante. Por ello, no pueden ser usados si existe cierta dependencia entre $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$. No obstante, cuando el objetivo del PCA es descriptivo, y no inferencial, complicaciones como la no independencia no afectan seriamente a dicho objetivo.

En las series temporales, la dependencia entre los vectores $\mathbf{x}$ es inducida por su relativa proximidad en el tiempo. De este modo, $\mathbf{x}_i$ y $\mathbf{x}_j$ son a menudo muy dependientes si $|i-j|$ es pequeño, con dependencia decreciente conforme $|i-j|$ aumenta. Este patrón básico puede, además, ser perturbado por una dependencia estacional en datos mensuales o trimestrales.

A continuación, se introducen algunas ideas y definiciones básicas sobre vectores de series temporales, necesarias para el desarrollo de la sección. Se supone que $\mathbf{x}$ es un vector de p variables que son series temporales estacionarias. Estas series pueden ser descritas por sus momentos de primer y segundo orden dados por

$$\boldsymbol{\mu} = E\left[\mathbf{x}_t\right]$$

$$\boldsymbol{\Gamma}_k = E\left[\left(\mathbf{x}_t - \boldsymbol{\mu}\right)\left(\mathbf{x}_{t+k} - \boldsymbol{\mu}\right)^T\right] \tag{1.61}$$

donde $\boldsymbol{\mu}$ es el vector de medias de dimensión $(p \times 1)$ siendo el mismo para todo $\mathbf{x}_t$ cuando las series son estacionarias y donde $\boldsymbol{\Gamma}_k$ es la matriz de dimensión $(p \times p)$ de autocovarianzas entre $\mathbf{x}_t$ y $\mathbf{x}_{t+k}$ que solo depende del retardo k en series estacionarias. La información contenida en la matriz de autocovarianzas se puede expresar equivalentemente en términos del espectro de potencia de las series

$$\mathbf{F}(\omega) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \boldsymbol{\Gamma}_k e^{-ik\omega} \tag{1.62}$$

siendo $\mathbf{F}(\omega)$ una matriz de dimensión $(p \times p)$ donde ω es la frecuencia angular.

El PCA trabaja sobre una matriz de covarianzas o de correlaciones, pero con series temporales se pueden calcular no solo las covarianzas entre las variables medidas en el mismo instante, $\boldsymbol{\Gamma}_0$, sino también las covarianzas entre las variables en diferentes instantes, medidas por $\boldsymbol{\Gamma}_k$, $k \neq 0$. Además de la elección sobre cuál $\boldsymbol{\Gamma}_k$ examinar, el

hecho de que las covarianzas tengan una representación alternativa en el dominio de la frecuencia significa que el PCA puede aplicarse a las series temporales tanto desde el dominio del tiempo usando (1.61) como desde el dominio de la frecuencia usando (1.62).

Antes de examinar las técnicas específicas, resta por definir los conceptos de 'ruido blanco' y 'ruido rojo'. Una serie de ruido blanco es aquella que todos sus términos son independientes e idénticamente distribuidos y su espectro es plano. Una serie de ruido rojo es aquella que sigue un modelo AR(1), $x_t = \phi x_{t-1} + \varepsilon_t$ donde ϕ es una constante tal que $0 < \phi < 1$ y ε_t es una serie de ruido blanco, y su espectro decrece cuando la frecuencia se incrementa.

Entre todas las técnicas del PCA relacionadas con las series temporales, en este trabajo se presentan las que están más difundidas y son más sencillas de implementar dentro de las que tienen un interés teórico y práctico demostrado, para lo cual se sigue principalmente a Jolliffe (2002).

1.5.1 Análisis espectral singular (SSA)

- **SSA univariante**

Elsner y Tsonis (1996) dan referencias sobre el SSA de las dos décadas anteriores en diferentes campos de investigación. El SSA se refiere a una sola serie temporal x_t y su idea básica consiste en realizar un PCA sobre un conjunto de retardos de la serie temporal. Exactamente las p variables retardadas son $\mathbf{x}_t = \left(x_t, x_{t+1}, \cdots, x_{t+p-1}\right)^T$ y, asumiendo que la serie x_t es estacionaria, el elemento σ_{ij} de su matriz de covarianzas $\boldsymbol{\Sigma}$ es la autocovarianza $\gamma_{|i-j|}$ que solo depende de $|i-j|$. De este modo, $\sigma_{ij} = \sigma_{i+1,j+1}$ con lo cual la matriz de covarianzas es una matriz Toeplitz simétrica. La estructura sencilla de las matrices Toeplitz permite deducir el comportamiento de las primeras componentes principales y de sus autovalores y autovectores[1] asociados, los cuales son funciones trigonométricas, para diferentes estructuras de series temporales. Las componentes principales son medias móviles de la serie temporal cuyas ponderaciones las proporcionan los autovectores de forma que cada componente recoge un tipo de oscilación subyacente en la serie, desde tendencia a ruido.

[1] Cuando en un vector de p series temporales todas ellas representan medidas contemporáneas sobre la misma variable pero en p localizaciones espaciales diferentes, p. e. la temperatura en distintos puntos geográficos, los autovectores reciben el nombre de funciones ortogonales empíricas (EOFs).

Para series temporales con un comportamiento oscilatorio, el SSA tiene asociado un par de autovectores con autovalores idénticos. Los coeficientes de ambos autovectores tienen el mismo patrón oscilatorio pero con un desfase de $\pi/2$ radianes. Según Allen y Smith (1996), la mejor aplicación del SSA es descubrir las periodicidades dominantes en una serie. Una ventaja del SSA sobre el análisis espectral tradicional es que las frecuencias de las oscilaciones detectadas pueden tomar cualquier valor en un intervalo dado en lugar de estar restringidas a un conjunto fijo de frecuencias. Sin embargo, una desventaja del SSA es que tiende a encontrar periodicidades aparentes que no existen.

En un contexto muestral, para realizar el SSA se tiene una muestra que se compone de una serie temporal $x_t,\ x_1, x_2, \cdots, x_n,$ la cual se reorganiza para obtener una matriz $\mathbf{X}$ de dimensión $\left(n' \times p\right)$ cuya t-ésima fila es

$$\mathbf{x}_t = \left(x_t, x_{t+1}, \cdots, x_{t+p-1}\right)^T$$

para $t = 1, 2, \cdots, n'$ donde $n' = n - p + 1$.

Una consideración práctica es la elección de un valor elevado de p que permita obtener oscilaciones de periodos largos pero que no deje pocas observaciones n' con las que estimar la matriz de covarianzas de las p variables. Elsner y Tsonis (1996) observan que la elección de $p = n/4$ es una práctica común.

Otra cuestión es que, si las n' observaciones sobre las p variables se tratan como una matriz ordinaria de datos, la correspondiente matriz muestral de covarianzas $\mathbf{S}$ no verificará que $s_{ij} = s_{i+1, j+1}$. Sin embargo, si la serie es estacionaria, la covarianza entre las variables i-ésima y j-ésima debería depender solo de $\left|i - j\right|$. Elsner y Tsonis (1996) presentan técnicas de estimación que se pueden usar para construir matrices de covarianzas con la restricción de que tengan una estructura de Toeplitz.

Si mediante (1.29) se proyecta la serie temporal sobre cada autovector $\mathbf{u}_k$ de la matriz $\mathbf{S}$ se obtiene cada componente principal $\hat{y}_{tk} = \mathbf{u}_k^T \mathbf{x}_t$ con $n' < n$ observaciones. No obstante, se puede reconstruir cada componente oscilatoria (desde tendencia a ruido) de la serie temporal asociada con un autovector mediante el procedimiento definido por Ghil y Vautard (1991)

$$R_{tk} = \frac{1}{M_t} \sum_{j=L_t^{\inf}}^{L_t^{\sup}} \hat{y}_{t-j+1,k} u_{jk} \tag{1.63}$$

donde $M_t = L_t^{\sup} - L_t^{\inf} + 1$ y $\left(L_t^{\inf}, L_t^{\sup}\right) = \begin{cases} (1, t) & 1 \le t \le p-1 \\ (1, p) & p \le t \le n' \\ (t - n + p, p) & n'+1 \le t \le n \end{cases}$.

Las componentes reconstruidas R_{tk} tienen la propiedad de estar en fase con la serie temporal con lo cual pueden superponerse en el tiempo. Además, no existe pérdida de información en la reconstrucción debido a que $x_t = \sum_{k=1}^{p} R_{tk}$. Todo ello permite realizar reconstrucciones parciales con una combinación de los autovectores asociados a un conjunto de determinadas oscilaciones $K = \{k_1, k_2, \cdots, k_s\}$ mediante

$$R_{K,t} = \sum_{k \in K} R_{tk} \tag{1.64}$$

y, así por ejemplo, eliminar el ruido en la serie temporal.

- **SSA multivariante (MSSA)**

La situación más habitual es disponer de p series temporales[2] observadas en n momentos de tiempo aunque la matriz de covarianzas no se calcula a partir de la habitual matriz de dimensión $(n \times p)$ con las observaciones. Los datos se reorganizan en una matriz de dimensión $(n' \times p')$ donde $n' = n - q + 1$, $p' = qp$ y una fila típica de esta matriz es

$$\mathbf{x}_t = \left(x_{t,1}, x_{t+1,1}, \cdots, x_{t+q-1,1}, x_{t,2}, \cdots, x_{t+q-1,2}, \cdots, x_{t,p}, \cdots, x_{t+q-1,p} \right)^T$$

para $t = 1, 2, \cdots, n'$ donde x_{tk} es el valor de la k-ésima variable en el momento t y q juega el mismo papel que p en SSA. La matriz de covarianzas para la matriz de datos reorganizada tiene la forma

$$\begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} & \cdots & \mathbf{S}_{1p} \\ \mathbf{S}_{21} & \mathbf{S}_{22} & \cdots & \mathbf{S}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_{p1} & \mathbf{S}_{p2} & \cdots & \mathbf{S}_{pp} \end{pmatrix}$$

donde $\mathbf{S}_{kk}$ es una matriz de covarianzas de dimensión $(q \times q)$ para varios retardos de la variable k-ésima con una estructura Toeplitz. Las matrices no diagonales $\mathbf{S}_{kl}$, $k \neq l$, tienen como elemento (i, j) la covarianza entre las variables k-ésima y l-ésima con un retardo de $|i - j|$. Plaut y Vautard (1994) reivindican que la propiedad fundamental del MSSA es su habilidad para detectar comportamientos oscilatorios al igual que el SSA pero con patrones comunes al conjunto de variables.

---

[2] Esta situación es equivalente a disponer de una sola variable observada en n momentos de tiempo pero simultáneamente en p localizaciones geográficas diferentes.

El MSSA amplía el SSA de una a varias series temporales pero, si el número de series p es grande, puede llegar a ser inmanejable. Una solución, usada por Benzi et al. (1997), es llevar a cabo el PCA sobre la matriz de datos inicial de dimensión $(n \times p)$ y, entonces, implementar el SSA de forma separada para cada una de las primeras componentes. De forma alternativa, otra solución consiste en realizar el MSSA sobre las primeras componentes en lugar de las propias variables como Plaut y Vautard (1994).

1.5.2 Análisis de patrones de oscilaciones principales (POP)

Tanto el SSA como el MSSA son casos especiales del PCA ya que, una vez definidas las variables de forma adecuada, se realiza un análisis de autovalores-autovectores de la matriz de covarianzas. Sin embargo, el análisis POP, aunque sirve para los mismos propósitos y realiza un análisis de autovalores-autovectores, es algo distinto porque dicho análisis no lo efectúa sobre una matriz de covarianzas. El análisis POP fue introducido por Hasselman (1988).

Los datos están agrupados en una matriz de dimensión $(n \times p)$ de medidas sobre p variables en n instantes de tiempo y se supone que el vector con las p series temporales sigue un proceso VAR(1). Si $\mathbf{x}_t^T$ es la t-ésima fila de dicha matriz, se tiene

$$(\mathbf{x}_{t+1} - \boldsymbol{\mu}) = \boldsymbol{\Phi}(\mathbf{x}_t - \boldsymbol{\mu}) + \boldsymbol{\varepsilon}_t \tag{1.65}$$

donde $\boldsymbol{\Phi}$ es una matriz constante, $\boldsymbol{\mu}$ es el vector de medias de las p variables y $\boldsymbol{\varepsilon}_t$ es un ruido blanco multivariante. El estimador de mínimos cuadrados de $\boldsymbol{\Phi}$ es $\hat{\boldsymbol{\Phi}} = \mathbf{S}_1 \mathbf{S}_0^{-1}$, donde $\mathbf{S}_0$ es la habitual matriz de covarianzas muestrales y $\mathbf{S}_1$ tiene como elemento (i, j) la covarianza muestral entre las variables i-ésima y j-ésima con un retardo de un periodo. El análisis POP calcula los autovalores y autovectores de la matriz $\hat{\boldsymbol{\Phi}}$. Los autovectores se conocen como patrones de oscilaciones principales (POPs) y se denotan por $\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_p$ formando la matriz $\mathbf{P} = \left[\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_p\right]$. Las cantidades $\mathbf{z}_t = \left(z_{t1}, z_{t2}, \cdots, z_{tp}\right)^T$ se usan para reconstruir $\mathbf{x}_t$ como

$$\mathbf{x}_t = \mathbf{P}\mathbf{z}_t \Leftrightarrow \mathbf{X} = \mathbf{Z}\mathbf{P}^T \tag{1.66}$$

donde $\mathbf{Z}$ es la matriz de dimensión $(n \times p)$ cuya t-ésima fila es $\mathbf{z}_t$. Las cantidades $\mathbf{z}_t$ son los llamados coeficientes POP y tienen el mismo papel que las componentes principales en PCA.

Debido a que la matriz $\hat{\boldsymbol{\Phi}}$ no es simétrica, tiene una mezcla de autovectores reales y complejos. Estos últimos aparecen en pares, con cada par compartiendo el mismo

autovalor y teniendo autovectores que son pares complejos conjugados. Los autovectores reales no describen oscilaciones pero los autovectores complejos representan oscilaciones suavizadas y pueden incluir ondas estacionarias y/u ondas de propagación espacial entre las variables, dependiendo de las magnitudes relativas de las partes real e imaginaria de cada POP complejo según detallan von Storch et al. (1988).

Como otras técnicas, los datos pueden ser preprocesados usando PCA para remplazar $\mathbf{x}$ en la ecuación (1.65) por sus componentes principales.

1.5.3 Análisis de autovectores (EOFs) de Hilbert (HEOF)

Esta técnica parece que tuvo su origen con Rasmusson et al. (1981). Se supone que $\mathbf{x}_t,\ t=1,2,\cdots,n,$ es un vector con p series temporales y se establece

$$\mathbf{y}_t = \mathbf{x}_t + i\,\mathbf{x}_t^H \tag{1.67}$$

donde $i=\sqrt{-1}$ y $\mathbf{x}_t^H$ es la transformación de Hilbert de $\mathbf{x}_t$ definida como

$$\mathbf{x}_t^H = \sum_{s=0}^{\infty} \frac{2}{(2s+1)\pi}\left(\mathbf{x}_{t+2s+1} - \mathbf{x}_{t-2s-1}\right)$$

La definición asume que $\mathbf{x}_t$ se observa un número infinito de veces. La estimación de $\mathbf{x}_t^H$ para muestras finitas se puede realizar según Barnett (1983) mediante una convolución dada por

$$\mathbf{x}_t^H = \sum_{k=-L}^{L} \alpha_k \mathbf{x}_{t-k}$$

donde $\alpha_k = \frac{2}{\pi k}\sin^2\left(\frac{\pi k}{2}\right)$ para $k\neq 0$, $\alpha_0 = 0$ y $7 \le L \le 25$ proporciona valores adecuados para L.

Si una serie está compuesta por términos oscilatorios, su transformación de Hilbert adelanta a cada término oscilatorio en $\pi/2$ radianes. Cuando $\mathbf{x}_t$ consta de una sola oscilación periódica, $\mathbf{x}_t^H$ es idéntica a $\mathbf{x}_t$ excepto que está desplazada $\pi/2$ radianes. En el caso más habitual, donde $\mathbf{x}_t$ consta de varias oscilaciones de diferentes frecuencias, el efecto de la transformación a $\mathbf{x}_t^H$ es más complejo porque el desfase de $\pi/2$ radianes se implementa de forma separada para cada frecuencia.

Un análisis HEOF es simplemente un PCA sobre la matriz de covarianzas de $\mathbf{y}_t$ definida en (1.67). Al igual que SSA, MSSA y análisis POP, HEOF encontrará los patrones oscilatorios dominantes, los cuales pueden o no propagarse espacialmente entre las variables.

Similarmente al análisis POP, los autovalores y autovectores del análisis HEOF son complejos. Sin embargo, a diferencia del análisis POP, se obtienen de una matriz de covarianzas correspondiente a variables con valores complejos. El análisis HEOF maximiza varianzas, tiene componentes ortogonales y su base es empírica, propiedades todas ellas compartidas con el PCA. En contraste, el análisis POP no maximiza varianzas, los coeficientes POP no son ortogonales y está basado en un modelo.

1.5.4 PCA y análisis POP para series ciclo-estacionarias

El supuesto de estacionariedad temporal está implícito en las técnicas descritas hasta ahora en esta sección. Existen bastantes tipos de series temporales en las que a menudo existe un ciclo de periodo fijo, generalmente el ciclo anual o estacionalidad pero a veces también existe un ciclo diario. Por ello, las series pueden ser estacionarias a lo largo del tiempo en el mismo punto del ciclo pero no entre diferentes puntos del ciclo. Por ejemplo, para series mensuales la distribución de probabilidad puede ser la misma cada mes de febrero pero diferente en febrero que en julio. Similarmente, la distribución conjunta para febrero y mayo (separados por tres meses) puede ser la misma en diferentes años pero distinta a la distribución conjunta para julio y octubre (también separados por tres meses). Este comportamiento se conoce como ciclo-estacionariedad y obliga a modificar el PCA y el análisis POP.

La modificación es más fácil de aplicar en el análisis POP. Se supone que τ es la longitud del ciclo, así $\tau = 12$ para series mensuales con ciclo anual. Entonces, para las series temporales de longitud $n = n'\tau$ la ecuación (1.65) se remplaza por

$$\left(\mathbf{x}_{s\tau+t+1} - \boldsymbol{\mu}_{t+1}\right) = \boldsymbol{\Phi}_t \left(\mathbf{x}_{s\tau+t} - \boldsymbol{\mu}_t\right) + \boldsymbol{\varepsilon}_{s\tau+t} \tag{1.68}$$

$$t = 0,1,2,\cdots,\tau-1; \quad s = 0,1,2,\cdots,n'-1$$

con $t+1$ sustituido por 0 y s por $s+1$ en el lado izquierdo de (1.68) cuando $t = \tau-1$ y $s = 0,1,2,\cdots,n'-2$. De este modo, para una serie mensual se establecen 12 procesos VAR(1), uno para cada mes de los valores de las series. Aquí, el vector $\boldsymbol{\mu}_t$ y la matriz $\boldsymbol{\Phi}_t$ pueden variar dentro de los ciclos pero no entre los ciclos. El análisis POP ciclo-estacionario estima $\boldsymbol{\Phi}_0, \boldsymbol{\Phi}_1, \cdots, \boldsymbol{\Phi}_{\tau-1}$ y a continuación realiza un análisis de autovalores-autovectores del producto de estas estimaciones $\hat{\boldsymbol{\Phi}} = \sum_{t=0}^{\tau-1} \hat{\boldsymbol{\Phi}}_t$.

La variedad ciclo-estacionaria del PCA la sintetizan Kim y Wu (1999) pero su justificación es menos transparente que la dada para el análisis POP ciclo-estacionario.

En primer lugar, se calculan los vectores $\mathbf{a}_{0,t}, \mathbf{a}_{1,t}, \cdots, \mathbf{a}_{\tau-1,t}$ tal que $\mathbf{x}_t = \sum_{j=0}^{\tau-1} \mathbf{a}_{j,t} e^{\frac{2\pi j t}{\tau} i}$ y, a continuación, se construye un nuevo vector de variables concatenando $\mathbf{a}_{0,t}, \mathbf{a}_{1,t}, \cdots, \mathbf{a}_{\tau-1,t}$. Los autovectores ciclo-estacionarios se obtienen de la matriz de covarianzas calculada para este nuevo vector de variables.

1.5.5 PCA en el dominio de la frecuencia

El PCA en el dominio de la frecuencia no tiene contrapartida en los conjuntos de datos con observaciones independientes. Para ver cómo se derivan las componentes principales en el dominio de la frecuencia, se observa que las componentes principales para un vector aleatorio $\mathbf{x}$ p-dimensional, con media cero, se pueden obtener encontrando unas matrices $\mathbf{B}$, $\mathbf{C}$ de dimensión $(p \times q)$ tales que

$$E\left[\left(\mathbf{x}-\mathbf{C}\mathbf{z}\right)^T\left(\mathbf{x}-\mathbf{C}\mathbf{z}\right)\right]$$

se minimice, donde $\mathbf{z} = \mathbf{B}^T\mathbf{x}$. Se demuestra, p. e. en Jolliffe (2002), que $\mathbf{B} = \mathbf{C}$ y que las columnas de $\mathbf{B}$ son los primeros q autovectores de la matriz de covarianzas $\mathbf{\Sigma}$ del vector $\mathbf{x}$, de modo que los elementos de $\mathbf{z}$ son las primeras q componentes principales para $\mathbf{x}$. Este argumento puede extenderse a un vector de p variables que sean series temporales. Para ello, se supone que la serie de vectores es $\cdots, \mathbf{x}_{-1}, \mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \cdots$ y que $E\left[\mathbf{x}_t\right] = \mathbf{0}$ para todo t. Así, definiendo

$$\mathbf{z}_t = \sum_{s=-\infty}^{\infty} \mathbf{B}_{t-s}^T \mathbf{x}_s$$

se estima $\mathbf{x}_t$ con $\sum_{s=-\infty}^{\infty} \mathbf{C}_{t-s}\mathbf{z}_s$, donde $\cdots, \mathbf{B}_{t-1}, \mathbf{B}_t, \mathbf{B}_{t+1}, \mathbf{B}_{t+2}, \cdots, \mathbf{C}_{t-1}, \mathbf{C}_t, \mathbf{C}_{t+1}, \mathbf{C}_{t+2}, \cdots$ son matrices de dimensión $(p \times q)$ que minimizan

$$E\left[\left(\mathbf{x}_t - \sum_{s=-\infty}^{\infty} \mathbf{C}_{t-s}\mathbf{z}_s\right)^T \left(\mathbf{x}_t - \sum_{s=-\infty}^{\infty} \mathbf{C}_{t-s}\mathbf{z}_s\right)\right]$$

donde $\mathbf{A}^*$ es la transpuesta conjugada de $\mathbf{A}$. La diferencia entre esta última formulación y la anterior es que las relaciones entre $\mathbf{z}$ y $\mathbf{x}$ están en términos de todos los valores de $\mathbf{x}_t$ y $\mathbf{z}_t$, en diferentes instantes, más que de unas solas $\mathbf{x}$ y $\mathbf{z}$. Además, la derivación se realiza de forma general para series complejas mejor que dejarla restringida a series reales. Según Brillinger (1981), ello produce que

$$\mathbf{B}_s^T = \frac{1}{2\pi}\int_0^{2\pi} \tilde{\mathbf{B}}(\omega) e^{is\omega} d\omega \qquad \text{y} \qquad \mathbf{C}_s = \frac{1}{2\pi}\int_0^{2\pi} \tilde{\mathbf{C}}(\omega) e^{is\omega} d\omega$$

donde $\tilde{\mathbf{C}}(\omega)$ es una matriz de dimensión $(p \times q)$ cuyas columnas son los primeros autovectores de la matriz $\mathbf{F}(\omega)$ dada en (1.62) y $\tilde{\mathbf{B}}(\omega)$ es la transpuesta conjugada de $\tilde{\mathbf{C}}(\omega)$. Las q series que forman los elementos de $\mathbf{z}_t$ son las series de las primeras q componentes principales de $\mathbf{x}_t$.

Brillinger (1981) describe la conexión que hay entre el PCA en el dominio del tiempo y el PCA en el dominio de la frecuencia. La conexión implica la transformada de Hilbert con lo cual el PCA en el dominio de la frecuencia guarda relación con el análisis HEOF. Se define el vector de variables $\mathbf{y}_t^H(\omega) = \left[\left(\mathbf{x}_t(\omega)\right)^T, \left(\mathbf{x}_t^H(\omega)\right)^T \right]^T$, donde $\mathbf{x}_t(\omega)$ es la contribución a $\mathbf{x}_t$ en la frecuencia ω y $\mathbf{x}_t^H(\omega)$ su transformada de Hilbert. Entonces, la matriz de covarianzas de $\mathbf{y}_t^H(\omega)$ es proporcional a

$$\begin{bmatrix} \mathrm{Re}(\mathbf{F}(\omega)) & \mathrm{Im}(\mathbf{F}(\omega)) \\ -\mathrm{Im}(\mathbf{F}(\omega)) & \mathrm{Re}(\mathbf{F}(\omega)) \end{bmatrix}$$

Los autovalores que obtiene un PCA de $\mathbf{y}_t^H(\omega)$ son los autovalores de $\mathbf{F}(\omega)$ con el correspondiente par de autovectores

$$\begin{bmatrix} \mathrm{Re}(\tilde{\mathbf{c}}_j(\omega)) \\ \mathrm{Im}(\tilde{\mathbf{c}}_j(\omega)) \end{bmatrix} \quad \text{y} \quad \begin{bmatrix} -\mathrm{Im}(\tilde{\mathbf{c}}_j(\omega)) \\ \mathrm{Re}(\tilde{\mathbf{c}}_j(\omega)) \end{bmatrix}$$

donde $\tilde{\mathbf{c}}_j(\omega)$ es la j-ésima columna de $\tilde{\mathbf{C}}(\omega)$.

Horel (1984) interpreta el análisis HEOF como un PCA en el dominio de la frecuencia promediado sobre toda la banda de frecuencias. Cuando una frecuencia de oscilación domina la variación en un grupo de series temporales, ambas técnicas vienen a ser lo mismo.

Stoffer (1999) describe un tipo diferente de PCA en el dominio de la frecuencia que llama *envolvente espectral*. Este tipo de PCA se realiza sobre la matriz espectral $\mathbf{F}(\omega)$ en relación con la matriz de covarianzas en el dominio del tiempo $\mathbf{\Gamma}_0$. Esto es una forma de generalizar el PCA para $\mathbf{F}(\omega)$ con $\mathbf{\Gamma}_0$ como una métrica y conduce a resolver la ecuación $\left[\mathbf{F}(\omega) - l(\omega)\mathbf{\Gamma}_0\right]\mathbf{u}(\omega) = \mathbf{0}$ variando la frecuencia angular ω. Stoffer (1999) recomienda el método como una forma de descubrir si las p series $x_1(t), x_2(t), \cdots, x_p(t)$ comparten señales comunes.

Capítulo 2

Análisis de Componentes Independientes

2.1 Introducción

Un problema central en la investigación de redes neuronales, así como en estadística y en procesamiento de señales, es encontrar una representación adecuada de los datos mediante una transformación apropiada para facilitar el análisis posterior de los mismos. Un ejemplo sencillo es el ya clásico que se muestra en Hyvärinen y Oja (2000) o en Hyvärinen *et al.* (2001). En él, se considera a tres personas hablando simultáneamente en una sala en la que se dispone de tres micrófonos que registran tres señales temporales de voz denotadas por $x_1(t)$, $x_2(t)$ y $x_3(t)$, con x_1, x_2 y x_3 las amplitudes y t el índice de tiempo. Cada una de las señales registradas es la suma ponderada de las señales de voz emitidas por los oradores que se denotan por $s_1(t)$, $s_2(t)$ y $s_3(t)$. Esto, se podría expresar como un sistema de ecuaciones lineales

$$\begin{aligned} x_1(t) &= a_{11}s_1(t)+a_{12}s_2(t)+a_{13}s_3(t) \\ x_2(t) &= a_{21}s_1(t)+a_{22}s_2(t)+a_{23}s_3(t) \\ x_3(t) &= a_{31}s_1(t)+a_{32}s_2(t)+a_{33}s_3(t) \end{aligned} \tag{2.1}$$

donde los a_{ij} con $i, j = 1,2,3$ son parámetros que dependen de la distancia de los oradores a los micrófonos. Sería de gran utilidad si se pudieran estimar las señales originales de voz, $s_1(t)$, $s_2(t)$ y $s_3(t)$, usando únicamente las señales registradas, $x_1(t)$, $x_2(t)$ y $x_3(t)$. Esta situación se conoce como el problema del cóctel. En

realidad, si se conocieran los parámetros a_{ij}, se podría resolver el sistema de ecuaciones lineales anterior por los métodos algebraicos clásicos. Sin embargo, puesto que no se conocen los a_{ij}, el problema se vuelve considerablemente más difícil.

De forma más general, el problema se centra en la representación de variables multidimensionales continuas. Si se denota por $\mathbf{x}=\left(x_1, x_2, \cdots, x_p\right)^T$ un vector aleatorio p-dimensional, el problema es encontrar una función vectorial $\mathbf{f}$ de modo que la transformada n-dimensional $\mathbf{s}=\left(s_1, s_2, \cdots, s_n\right)^T$ definida por

$$\mathbf{s}=\mathbf{f}\left(\mathbf{x}\right) \tag{2.2}$$

tenga algunas propiedades convenientes. En la mayoría de los casos, la representación que se busca es una transformación lineal de las variables observadas, es decir

$$\mathbf{s}=\mathbf{W}\mathbf{x} \tag{2.3}$$

donde $\mathbf{W}$ es la matriz de separación que debe ser determinada. El uso de transformaciones lineales hace el problema, conceptual y computacionalmente, más simple y facilita la interpretación de los resultados.

Entre los varios métodos desarrollados para encontrar la transformación lineal apropiada se encuentra el Análisis de Componentes Principales, PCA. Sin embargo, si se supone que las componentes o señales originales s_i son estadísticamente independientes unas de otras tanto como sea posible, se puede utilizar el particular método llamado Análisis de Componentes Independientes (ICA) para estimar la transformación lineal buscada. En consecuencia, el ICA se puede aplicar para extraer o estimar las señales fuente o componentes originales s_i, independientes unas de otras tanto como sea posible, a partir de sus mezclas x_i que son los valores observados correspondientes a una realización de una señal temporal discreta p-dimensional $\mathbf{x}(t)$, con $t=1, 2, \cdots, T$.

En este capítulo se caracteriza el análisis de componentes independientes, estableciendo las condiciones necesarias para su puesta en práctica. A continuación, se examinan diferentes funciones de contraste, necesarias para estimar un modelo ICA, resaltando las medidas de no-gaussianidad. Antes de diseñar un algoritmo, necesario para optimizar una función de contraste, se indica cómo preparar los datos con especial atención a las series temporales. Posteriormente, se deriva y detalla el algoritmo FastICA y, finalmente, se describe el algoritmo AMUSE que se adapta mejor a las series temporales.

2.2 Caracterización del ICA

2.2.1 Independencia estadística

Para realizar el ICA se supone que las componentes s_i (señales fuente) son independientes. Por ello, es necesario definir y aclarar qué se entiende por independencia estadística para lo cual se sigue a Peña (1991).

Definición 2.1 *(Independencia estadística) Sean* $y_1, y_2, \cdots, y_n$ *un conjunto de variables aleatorias con función de densidad conjunta* $f(y_1, y_2, \cdots, y_n)$. *Las variables* y_k, *con* $k = 1, 2, \cdots, n$, *son estadísticamente independientes mutuamente si su función de densidad conjunta se puede factorizar como*

$$f(y_1, y_2, \cdots, y_n) = f_1(y_1)\, f_2(y_2) \cdots f_n(y_n) \tag{2.4}$$

donde $f_k(y_k)$ *denota la función de densidad marginal de* y_k.

De esta forma, cualquier subconjunto de variables aleatorias $y_1, y_2, \cdots, y_h$ con $h \leq n$ también será estadísticamente independiente.

Propiedad 2.1 *Si las variables aleatorias* y_k *con* $k = 1, 2, \cdots, n$ *son estadísticamente independientes mutuamente, entonces para cualesquiera funciones* g_1 *y* g_2 *se verifica*

$$\mathrm{E}\left[g_1(y_i)\, g_2(y_j) \right] = \mathrm{E}\left[g_1(y_i) \right] \mathrm{E}\left[g_2(y_j) \right] \quad \forall i \neq j \tag{2.5}$$

Demostración. Para la demostración, ver Peña (1991). ■

Definición 2.2 *(Incorrelación) Se dice que el conjunto de variables aleatorias* y_k, *con* $k = 1, 2, \cdots, n$, *están incorrelacionadas mutuamente si su covarianza es cero*

$$\mathrm{E}\left[y_i y_j \right] - \mathrm{E}\left[y_i \right] \mathrm{E}\left[y_j \right] = 0 \tag{2.6}$$

Propiedad 2.2 *Si las variables aleatorias* y_k, *con* $k = 1, 2, \cdots, n$, *son estadísticamente independientes mutuamente, entonces están incorrelacionadas mutuamente.*

Demostración. En la Propiedad 2.1 basta tomar $g_1(y_i) = y_i$ y $g_2(y_j) = y_j$. ■

Sin embargo, la incorrelación no implica independencia excepto cuando las variables aleatorias tienen una distribución conjunta gaussiana. Debido a ello, el ICA, tal y como se verá en apartados posteriores, no está interesado en variables gaussianas.

2.2.2 Definición de ICA

A la hora de definir el problema de ICA, en este trabajo, se considera el caso lineal aunque también existen formas no lineales de ICA. A continuación, se muestran las tres definiciones más básicas de ICA que, según Comon (1994) e Hyvärinen (1999b), se encuentran en la literatura.

Definición 2.3 *(Definición general de ICA) El ICA de un vector aleatorio* $\mathbf{x}$ *consiste en encontrar una transformación lineal* $\mathbf{s}=\mathbf{W}\mathbf{x}$ *tal que las componentes* s_i *sean independientes tanto como sea posible, en el sentido de maximizar alguna función* $F(s_1, s_2, \cdots, s_n)$ *que mida la independencia.*

Esta definición es la más general porque no realiza supuestos sobre los datos pero bastante imprecisa ya que se debe definir una medida de independencia para las componentes s_i. La definición de independencia no se puede usar porque no es posible encontrar una transformación lineal que produzca componentes estrictamente independientes. Una aproximación diferente se produce con la siguiente definición que tiene una orientación más teórica.

Definición 2.4 *(Modelo ICA con ruido) El ICA de un vector aleatorio* $\mathbf{x}$ *consiste en la estimación del siguiente modelo lineal generador de los datos observados*

$$\mathbf{x}=\mathbf{A}\mathbf{s}+\mathbf{u} \tag{2.7}$$

donde las variables latentes (componentes) s_i *en el vector* $\mathbf{s}=(s_1, s_2, \cdots, s_n)^T$ *se suponen independientes. La matriz* $\mathbf{A}$ *es una matriz de mezcla, constante, de dimensión* $(p\times n)$ *y* $\mathbf{u}$ *es un vector aleatorio de ruido p-dimensional.*

Esta definición presenta el ICA como un problema de estimación de un modelo de variables no observables. No obstante, este problema no es sencillo de resolver y, por ello, la gran mayoría de la investigación sobre el ICA se ha concentrado en la siguiente definición simplificada.

El análisis factorial, cuyo modelo es $\mathbf{x}=\mathbf{A}\mathbf{f}+\mathbf{e}$, guarda una estrecha relación con esta definición de ICA. Los factores comunes $\mathbf{f}$ también son independientes y no observables pero siempre en menor número que las variables observadas $\mathbf{x}$, $\mathbf{A}$ es una matriz constante de dimensión $(p\times n)$ con $n<p$ y los factores específicos (ruido) $\mathbf{e}$ son asimismo independientes con la misma dimensión que $\mathbf{x}$. En cambio, en el modelo factorial se asume que las variables en $\mathbf{f}$ y $\mathbf{e}$ tienen distribución gaussiana.

Definición 2.5 *(Modelo ICA sin ruido) El ICA de un vector aleatorio* $\mathbf{x}$ *consiste en la estimación del siguiente modelo lineal generador de los datos observados*

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{2.8}$$

donde $\mathbf{A}$ *y* $\mathbf{s}$ *son iguales que en la Definición 2.4.*

En esta definición el vector de ruido ha sido omitido. Según Comon (1994), se prueba que, si los datos siguen el modelo de la ecuación (2.8), las Definiciones 2.3 y 2.5 son asintóticamente equivalentes cuando determinadas medidas de independencia se usan en la Definición 2.3 y la relación algebraica $\mathbf{W} = \mathbf{A}^{-1}$ se utiliza con $p = n$.

Este trabajo se centra en la definición del modelo ICA libre de ruido, al que se llamará modelo básico. Ello se debe, en parte, a que la mayoría de las investigaciones sobre el ICA se concentran en esta sencilla definición y, en parte, a que el modelo libre de ruido es una aproximación más manejable que el modelo real con ruido funcionando bastante bien para gran variedad de datos reales.

Las definiciones de ICA dadas anteriormente, al contrario que el PCA, no ordenan las componentes independientes. Sin embargo, es posible introducir alguna forma de ordenar las componentes independientes. Una primera opción, sería usar la norma de las columnas de la matriz de mezcla. Esto determina las contribuciones de las componentes independientes s_i a las varianzas de las variables observadas x_i. Ordenar las s_i según la norma decreciente de las correspondientes columnas de $\mathbf{A}$ recuerda el orden que establece el PCA. Una segunda opción, consiste en usar una función que obtuviera una medida de la no-gaussianidad de las s_i y ordenarlas en sentido decreciente de la no-gaussianidad.

2.2.3 Existencia de un modelo ICA

Para asegurar la existencia y, por tanto, la estimación del modelo básico de ICA definido en (2.8), se realizan las siguientes restricciones y suposiciones.

1) Las componentes no observadas s_i *son estadísticamente independientes.*

Esta es la hipótesis fundamental por lo ya explicado en los apartados anteriores.

2) Los vectores aleatorios $\mathbf{x}$ *y* $\mathbf{s}$ *tienen media cero,* $\mathrm{E}[x_i] = 0$, *con* $i = 1, 2, \cdots, p$ *y* $\mathrm{E}[s_j] = 0$, *con* $j = 1, 2, \cdots, n$.

En realidad, este supuesto no es ninguna restricción porque se cumple al centrar los

datos observados según se ilustrará en una sección posterior. No obstante, ayuda a la computación y simplifica la estimación del modelo ICA.

3) La varianza de las componentes independientes s_i *es la unidad,* $\mathrm{E}\left[s_i^2\right]=1$, *con* $i=1,2,\cdots,n$.

Esta restricción, junto con la primera, significa que la matriz de covarianzas del vector de componentes independientes $\mathbf{s}$ es la matriz unidad, $\mathrm{E}\left[\mathbf{s}\mathbf{s}^T\right]=\mathbf{I}$. Además, garantiza la unicidad de cada una de las componentes independientes salvo un signo multiplicativo que puede ser distinto para cada componente. Esta restricción se impone porque una indeterminación básica en el modelo es que las componentes independientes y las columnas de la matriz de mezcla $\mathbf{A}$ pueden ser calculadas de forma única salvo una constante multiplicativa. Sin embargo, esta indeterminación es bastante insignificante porque cualquier constante que multiplique a una componente independiente en la ecuación (2.8) se puede cancelar dividiendo la columna correspondiente de la matriz de mezcla por esa misma constante.

4) Se asume que la matriz de mezcla desconocida $\mathbf{A}$ *es cuadrada,* $p=n$.

Aunque este supuesto es bastante simplista, queda justificado por el hecho de que si $p>n$ (como es lo más habitual), la dimensión del vector de observaciones o mezclas $\mathbf{x}$ siempre puede reducirse utilizando alguna técnica de reducción de la dimensión, por ejemplo el PCA, de forma que se obtenga $p=n$.

5) La matriz de mezcla $\mathbf{A}$ *debe ser de rango completo por columnas.*

Este supuesto garantiza la invertibilidad de la matriz de mezcla.

6) Las componentes independientes s_i *deben tener distribuciones no-gaussianas.*

Los cumulantes de orden superior son nulos para distribuciones gaussianas, pero estos cumulantes son necesarios para la estimación del modelo ICA como se verá más adelante. Por otro lado, si la matriz de mezcla $\mathbf{A}$ es ortogonal (como ocurre cuando necesariamente se blanquean los datos) y las componentes independientes s_i son gaussianas, entonces las variables mezclas x_i son gaussianas, incorrelacionadas y de varianza unidad. Así, la densidad conjunta de las variables x_i es completamente simétrica y no contiene ninguna información sobre las direcciones que representan las columnas de la matriz de mezcla $\mathbf{A}$. Para probarlo, se supone que la distribución conjunta de dos componentes independientes s_1 y s_2 es gaussiana. Esto significa que su función de densidad conjunta está dada por

$$f(s_1, s_2) = \frac{1}{2\pi} \exp\left(-\frac{s_1^2 + s_2^2}{2} \right) = \frac{1}{2\pi} \exp\left(-\frac{\|\mathbf{s}\|_2^2}{2} \right) \tag{2.9}$$

Puesto que la matriz de mezcla $\mathbf{A}$ es ortogonal, $\mathbf{A}^{-1} = \mathbf{A}^T$, la función de densidad conjunta de las mezclas x_1 y x_2 viene dada por

$$f(x_1, x_2) = \frac{1}{2\pi} \exp\left(-\frac{\|\mathbf{Ax}\|_2^2}{2} \right) \left| \det\left(\mathbf{A}^T \right) \right| \tag{2.10}$$

Y, debido a la ortogonalidad de $\mathbf{A}$, se tiene que $\|\mathbf{Ax}\|_2^2 = \|\mathbf{x}\|_2^2$ y $\left| \det\left(\mathbf{A}^T \right) \right| = 1$. De este modo se llega a que la función de densidad conjunta de las dos mezclas es

$$f(x_1, x_2) = \frac{1}{2\pi} \exp\left(-\frac{\|\mathbf{x}\|_2^2}{2} \right) \tag{2.11}$$

Por ello, ambas distribuciones, original y mezcla, son idénticas y no se puede inferir la matriz de mezcla a partir de las señales mezclas observadas.

Si los elementos de vectores aleatorios $\mathbf{x}$ y $\mathbf{s}$ se interpretan como procesos estocásticos, en lugar de simples variables aleatorias, se necesitan restricciones adicionales. Como mínimo, se debe asumir que los procesos estocásticos son estacionarios en sentido estricto y también es necesario asumir ergodicidad respecto a las estimaciones. Si los procesos estocásticos son independientes idénticamente distribuidos, i.i.d., en el tiempo, entonces estos supuestos se verifican. Después de realizar estos supuestos, se pueden considerar los procesos estocásticos como variables aleatorias.

Respecto a la existencia del modelo ICA con ruido, las restricciones aquí expuestas garantizan parcialmente su existencia si se asume que el ruido es independiente de las componentes s_i. De hecho, el modelo ICA con ruido es un caso especial del modelo ICA sin ruido con $p < n$, porque las variables de ruido podrían considerarse como componentes independientes adicionales. En esta situación, según Cardoso (1991), la matriz de mezcla $\mathbf{A}$ parece identificable mientras que las realizaciones de las componentes s_i no son identificables debido a la no invertibilidad de la matriz $\mathbf{A}$.

2.3 Funciones objetivo o contraste para ICA

La estimación del modelo ICA requiere la formulación de una función objetivo, también llamada función de contraste, y su posterior optimización, maximizar o minimizar la función según corresponda. En este trabajo, como funciones de contraste se presentan la verosimilitud, la información mutua y las medidas de no-gaussianidad, por ser las más representativas en la investigación de ICA. Siguiendo a Hyvärinen (1999b), las funciones de contraste se pueden clasificar según resuelvan el problema de la estimación de las componentes independientes. Si la estimación de todas las componentes se realiza al unísono se denominan funciones de contraste multiunidad, entre las que se encuentran la verosimilitud y la información mutua. Si por el contrario, la estimación se realiza componente a componente se denominan funciones de contraste de una unidad y entre ellas se encuentran las medidas de no-gaussianidad.

2.3.1 Verosimilitud y entropía de red

La función de verosimilitud del modelo ICA sin ruido, ecuación (2.8), la formularon Pham *et al.* (1992) con el fin de estimar dicho modelo por máxima verosimilitud. Si se denota por $\mathbf{W}=\left[\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_n\right]^T$ a la matriz $\mathbf{A}^{-1}$, la log-verosimilitud toma la forma

$$L=\sum_{t=1}^{T}\sum_{i=1}^{n}\log\left[\, f_i\left(\mathbf{w}_i^T\mathbf{x}(t)\right)\right]+T\log\left(\left|\det\left(\mathbf{W}\right)\right|\right) \tag{2.12}$$

donde las funciones f_i son las densidades, supuestamente desconocidas, de las componentes s_i y $\mathbf{x}(t)$, con $t=1, 2, \cdots, T$, son las realizaciones de $\mathbf{x}$. El término $\log\left(\left|\det\left(\mathbf{W}\right)\right|\right)$ aparece por la clásica regla de la función de densidad para transformaciones lineales de variables aleatorias, según describe Ríos (1985). Así, para la estimación del modelo ICA dado en (2.8), si f_s es la función de densidad conjunta de las componentes s_i, la función de densidad de $\mathbf{x}=\mathbf{As}$ viene dada por $f_x(\mathbf{x})=f_s\left(\mathbf{A}^{-1}\mathbf{x}\right)\left|\det\left(\mathbf{A}^{-1}\right)\right|=f_s\left(\mathbf{Wx}\right)\left|\det\left(\mathbf{W}\right)\right|$.

Para describir la función de contraste relacionada con la verosimilitud que se deriva desde un punto de vista de red neuronal se necesita definir primero el concepto de entropía diferencial.

Definición 2.6 *(Entropía diferencial) La entropía diferencial* H *de un vector aleatorio* $\mathbf{y}$ *de dimensión* n *cuya función de densidad conjunta es* f*, se define como*

$$H(\mathbf{y})=-\int f(\mathbf{y})\log\left(f(\mathbf{y})\right)d\mathbf{y}=-\mathrm{E}\left[\log\left(f(\mathbf{y})\right)\right] \tag{2.13}$$

Esta nueva función de contraste, llamada entropía de red, fue desarrollada por Nadal y Parga (1994) y por Bell y Sejnowski (1995). Está basada en la maximización de la entropía producida por una red neuronal con salidas no lineales. Se asume que $\mathbf{x}$ es la entrada de la red neuronal cuyas salidas son de la forma $g_i\left(\mathbf{w}_i^T\mathbf{x}\right)$, donde las g_i son funciones escalares no lineales y los $\mathbf{w}_i$ son los vectores de pesos de las neuronas. Entonces, se quiere maximizar la entropía de las salidas o de red que viene dada por

$$L_2 = H\left(g_1\left(\mathbf{w}_1^T\mathbf{x}\right), g_2\left(\mathbf{w}_2^T\mathbf{x}\right), \cdots, g_n\left(\mathbf{w}_n^T\mathbf{x}\right)\right) \tag{2.14}$$

Si las funciones g_i, utilizadas en la red neuronal, son elegidas como las funciones de distribución correspondientes a las funciones de densidad marginal f_i, es decir $g_i'(\cdot) = f_i(\cdot)$, tanto Cardoso (1997) como Pearlmutter y Parra (1997) han demostrado que maximizar la entropía de red es equivalente a la estimación por máxima verosimilitud.

La aproximación máximo-verosímil posee la ventaja de ser asintóticamente eficiente bajo determinadas condiciones según establece la teoría de la inferencia estadística como en Ríos (1985). Sin embargo, presenta dos inconvenientes. El primero, es que esta aproximación requiere el conocimiento de las funciones de densidad de las componentes independientes s_i y, el segundo, es que puede ser muy sensible a valores atípicos.

2.3.2 Información mutua

Definición 2.7 *(Información mutua) La información mutua I entre n variables aleatorias y_i, con $i = 1, 2, \cdots, n$, se define como*

$$I\left(y_1, y_2, \cdots, y_n\right) = \sum_{i=1}^{n} H\left(y_i\right) - H\left(\mathbf{y}\right) \tag{2.15}$$

donde H denota la entropía diferencial e $\mathbf{y} = \left(y_1, y_2, \cdots, y_n\right)^T$ es un vector aleatorio.

La información mutua es una medida natural de la dependencia de variables aleatorias y, teóricamente, la más satisfactoria de las funciones de contraste multiunidad.

Propiedad 2.3 *La información mutua es no negativa, $I\left(y_1, y_2, \cdots, y_n\right) \geq 0$.*

Demostración. Para la demostración, ver Hyvärinen (2001). ■

Propiedad 2.4 *La información mutua es nula si y solo si las variables aleatorias y_i, con $i = 1, 2, \cdots, n$, son estadísticamente independientes.*

Demostración. Para la demostración, ver Hyvärinen (2001). ■

Propiedad 2.5 *La información mutua de una transformación lineal invertible* $\mathbf{y} = \mathbf{W}\mathbf{x}$, *con* $\mathbf{W}$ *no singular y* $\mathbf{x}$ *un vector aleatorio de dimensión* n, *viene dada por*

$$I(y_1, y_2, \cdots, y_n) = \sum_{i=1}^{n} H(y_i) - H(\mathbf{x}) - \log\left(\left|\det(\mathbf{W})\right|\right) \tag{2.16}$$

Demostración. Para una transformación lineal invertible, $\mathbf{y} = \mathbf{W}\mathbf{x}$, la función de densidad conjunta está dada por

$$f_x(\mathbf{x}) = f_y(\mathbf{W}\mathbf{x})\left|\det(\mathbf{W})\right| \Rightarrow f_y(\mathbf{y}) = f_x(\mathbf{x})\left|\det(\mathbf{W})\right|^{-1}$$

Así, si se expresa la entropía diferencial como una esperanza, $H(\mathbf{y}) = -\mathrm{E}\left[\log\left(f_y(\mathbf{y})\right)\right]$, se obtiene que

$$\mathrm{E}\left[\log\left(f_y(\mathbf{y})\right)\right] = \mathrm{E}\left\{\log\left[f_x(\mathbf{x})\left|\det(\mathbf{W})\right|^{-1}\right]\right\} = \mathrm{E}\left[\log\left(f_x(\mathbf{x})\right)\right] - \log\left(\left|\det(\mathbf{W})\right|\right)$$

Con lo cual, se llega a que la entropía de la transformación lineal es

$$H(\mathbf{y}) = H(\mathbf{x}) + \log\left(\left|\det(\mathbf{W})\right|\right) \tag{2.17}$$

Finalmente, sustituyendo $H(\mathbf{y})$ en (2.15) por la expresión dada en (2.17), queda demostrada la propiedad. ■

Consecuentemente, la información mutua del vector de componentes independientes $\mathbf{s}$ del modelo básico, $\mathbf{x} = \mathbf{A}\mathbf{s} \Leftrightarrow \mathbf{s} = \mathbf{W}\mathbf{x}$ donde $\mathbf{W} = \left[\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_n\right]^T = \mathbf{A}^{-1}$, se puede expresar como sigue

$$\begin{aligned} I(s_1, s_2, \cdots, s_n) &= \sum_{i=1}^{n} H(s_i) - H(\mathbf{s}) = \sum_{i=1}^{n} H(s_i) - H(\mathbf{x}) - \log\left(\left|\det(\mathbf{W})\right|\right) \\ &= \sum_{i=1}^{n} H\left(\mathbf{w}_i^T\mathbf{x}\right) - H(\mathbf{x}) - \log\left(\left|\det(\mathbf{W})\right|\right) \end{aligned} \tag{2.18}$$

Por todo ello, puesto que la información mutua tiene en cuenta la estructura completa de dependencia de las componentes independientes s_i, encontrar una transformación lineal, la matriz de mezcla $\mathbf{A}$, que minimice la información mutua entre dichas componentes es una manera muy lógica de estimar el modelo ICA. No obstante, la información mutua es difícil de estimar porque la definición de entropía necesita una estimación de la función de densidad.

Existe una conexión entre la verosimilitud y la información mutua. Si se considera la esperanza de la log-verosimilitud, se obtiene que

$$\begin{aligned}\frac{1}{T}\mathrm{E}[L] &= \sum_{i=1}^{n}\mathrm{E}\left\{\log\left[\, f_i\left(\mathbf{w}_i^T\mathbf{x}\right)\right]\right\} + \log\left(\left|\det\left(\mathbf{W}\right)\right|\right) \\ &= -\sum_{i=1}^{n} H\left(\mathbf{w}_i^T\mathbf{x}\right) + \log\left(\left|\det\left(\mathbf{W}\right)\right|\right)\end{aligned} \tag{2.19}$$

donde las f_i son las funciones de densidad de las componentes s_i. De este modo, la verosimilitud sería igual, salvo una constante aditiva, a la información mutua negativa dada en la ecuación (2.18).

2.3.3 Medidas de no-gaussianidad

Para usar la no-gaussianidad en la estimación de un modelo ICA, se debe tener una medida cuantitativa de la no-gaussianidad de una variable aleatoria. Entre estas medidas se encuentran la curtosis y la sintropía.

- **Curtosis**

Definición 2.8 *(Curtosis) La curtosis de una variable aleatoria* y*, con media* μ *y varianza* σ^2*, se define como*

$$\mathrm{kurt}\left(y\right) = \mathrm{E}\left[\left(y-\mu\right)^4\right] - 3\sigma^4 \tag{2.20}$$

Si la variable aleatoria y tiene media cero, la curtosis es

$$\mathrm{kurt}\left(y\right) = \mathrm{E}\left[y^4\right] - 3\left(\mathrm{E}\left[y^2\right]\right)^2 \tag{2.21}$$

y si, además, su varianza es la unidad, la curtosis queda como

$$\mathrm{kurt}\left(y\right) = \mathrm{E}\left(y^4\right) - 3 \tag{2.22}$$

Esto último, pone de manifiesto que la curtosis es simplemente una versión normalizada del momento de cuarto orden. Para una variable aleatoria gaussiana se verifica que la curtosis es cero, al contrario que para la mayoría de las variables aleatorias. La curtosis puede ser positiva o negativa. Las variables aleatorias que tienen curtosis negativa se llaman subgaussianas o platicúrticas y si su curtosis es positiva se denominan supergaussianas o leptocúrticas.

Propiedad 2.6 *Si* x_1 *y* x_2 *son dos variables aleatorias estadísticamente independientes, entonces se verifica*

$$\operatorname{kurt}\left(\alpha x_1 + \beta x_2\right) = \alpha^4 \operatorname{kurt}\left(x_1\right) + \beta^4 \operatorname{kurt}\left(x_2\right) \tag{2.23}$$

donde α *y* β *son escalares reales.*

Demostración. La demostración es inmediata aplicando en la definición de curtosis las propiedades de la esperanza para variables independientes. ■

La curtosis es la más sencilla función de contraste de una unidad. Dado el vector de datos observados $\mathbf{x}$, que sigue el modelo básico dado en (2.8), se trata de buscar una combinación lineal $\mathbf{w}^T\mathbf{x}$ tal que su curtosis se maximice o minimice siendo $\mathrm{E}\left[\left(\mathbf{w}^T\mathbf{x}\right)^2\right] = 1$ por la restricción tercera establecida para la existencia del modelo ICA. Se define $\mathbf{z} = \mathbf{A}^T\mathbf{w}$ donde $\mathbf{A}$ es la matriz de mezcla desconocida y, usando la ecuación (2.8), se obtiene que $\mathrm{E}\left[\left(\mathbf{w}^T\mathbf{x}\right)^2\right] = \mathbf{w}^T\mathbf{A}\,\mathrm{E}\left[\mathbf{s}\mathbf{s}^T\right]\mathbf{A}^T\mathbf{w} = \mathbf{w}^T\mathbf{A}\mathbf{A}^T\mathbf{w} = \left\|\mathbf{z}\right\|_2^2 = 1$, puesto que $\mathrm{E}\left[\mathbf{s}\mathbf{s}^T\right] = \mathbf{I}$. Además, por la Propiedad 2.6 se deriva que

$$\operatorname{kurt}\left(\mathbf{w}^T\mathbf{x}\right) = \operatorname{kurt}\left(\mathbf{w}^T\mathbf{A}\mathbf{s}\right) = \operatorname{kurt}\left(\mathbf{z}^T\mathbf{s}\right) = \sum_{i=1}^{m} z_i^4 \operatorname{kurt}\left(s_i\right) \tag{2.24}$$

Bajo la restricción $\left\|\mathbf{z}\right\|_2^2 = 1$, la función definida en (2.24) tiene un máximo y un mínimo local. Según demuestran Delfosse y Loubaton (1995), los puntos extremos de (2.24) son los vectores de la base canónica $\mathbf{z} = \pm\mathbf{e}_j$, es decir, los vectores cuyos elementos son todos cero excepto uno, el de la posición *j*, que es ± 1. Los correspondientes vectores de pesos son $\mathbf{w} = \pm\left(\mathbf{A}^{-1}\right)^T\mathbf{e}_j$, es decir, las filas de la inversa de la matriz de mezcla $\mathbf{A}$ salvo un signo multiplicativo. Así, al maximizar o minimizar la curtosis en la ecuación (2.24) bajo la restricción dada, se obtienen las componentes independientes como $\mathbf{w}^T\mathbf{x} = \pm s_j$. No obstante, esta aparente duplicidad no llega a producirse porque las componentes independientes corresponden siempre al máximo del valor absoluto de la curtosis.

La curtosis ha sido ampliamente usada para estimar componente a componente de un modelo ICA. Ello se debe a la simplicidad de su formulación matemática y, especialmente, a la posibilidad de probar los resultados de convergencia global como lo hacen Delfosse y Loubaton (1995). Sin embargo, en la práctica la curtosis presenta algunos inconvenientes cuando se estima a partir de una muestra. El principal problema es que la curtosis puede ser muy sensible a valores atípicos como demuestra

Huber (1985). Por ello, la curtosis no es una medida robusta de la no-gaussianidad.

- **Sintropía**

Una medida de la no-gaussianidad muy importante es la sintropía que se deriva de la teoría de la información mediante la aplicación del siguiente teorema.

Teorema 2.1 *Una variable gaussiana tiene la mayor entropía de entre todas las variables aleatorias de igual varianza.*

Demostración. Para la demostración del teorema, ver Cover y Thomas (1991). ■

Definición 2.9 *(Sintropía) La sintropía* J *de un vector aleatorio* $\mathbf{y}$ *se define como*

$$J(\mathbf{y}) = H(\mathbf{y}_{gauss}) - H(\mathbf{y}) \tag{2.25}$$

donde H *es la entropía diferencial e* $\mathbf{y}_{gauss}$ *es un vector aleatorio gaussiano con la misma matriz de covarianzas* $\boldsymbol{\Sigma}$ *que el vector aleatorio* $\mathbf{y}$*.*

A partir del Teorema 2.1 se demuestran las siguientes propiedades de la sintropía.

Propiedad 2.7 *La sintropía es no negativa.*

Propiedad 2.8 *La sintropía es cero si y solo si el vector aleatorio* $\mathbf{y}$ *tiene una distribución gaussiana.*

La entropía para un vector aleatorio gaussiano $\mathbf{y}_{gauss}$ con matriz de covarianzas $\boldsymbol{\Sigma}$ puede ser calculada como

$$H(\mathbf{y}_{gauss}) = \frac{1}{2}\log(\det(\boldsymbol{\Sigma})) + \frac{n}{2}(1+\log(2\pi)) \tag{2.26}$$

donde n es la dimensión de $\mathbf{y}_{gauss}$. Este resultado permite la demostración de la propiedad de la invarianza de la sintropía.

Propiedad 2.9 *La sintropía es invariante para transformaciones lineales invertibles.*

Demostración. Sea $\mathbf{y}$ un vector aleatorio de dimensión n y con matriz de covarianzas $\boldsymbol{\Sigma}$. Para una transformación lineal $\mathbf{z} = \mathbf{M}\mathbf{y}$, con $\mathbf{M}$ no singular, se tiene que $\mathrm{E}\left[\mathbf{z}\mathbf{z}^T\right] = \mathbf{M}\boldsymbol{\Sigma}\mathbf{M}^T$. Utilizando el resultado de (2.26) y la demostración de la Propiedad 2.5, la sintropía de $\mathbf{z}$ se calcula como

$$
\begin{aligned}
J(\mathbf{My}) &= H(\mathbf{z}_{gauss}) - H(\mathbf{z}) \\
&= \frac{1}{2}\log\left(\left|\det\left(\mathbf{M\Sigma M}^T\right)\right|\right) + \frac{n}{2}\left(1+\log(2\pi)\right) - \left[H(\mathbf{y}) + \log\left(\left|\det(\mathbf{M})\right|\right)\right] \\
&= \frac{1}{2}\log\left(\left|\det(\mathbf{\Sigma})\right|\right) + 2\frac{1}{2}\log\left(\left|\det(\mathbf{M})\right|\right) + \frac{n}{2}\left(1+\log(2\pi)\right) - H(\mathbf{y}) - \log\left(\left|\det(\mathbf{M})\right|\right) \\
&= \frac{1}{2}\log\left(\left|\det(\mathbf{\Sigma})\right|\right) + \frac{n}{2}\left(1+\log(2\pi)\right) - H(\mathbf{y}) \\
&= H(\mathbf{y}_{gauss}) - H(\mathbf{y}) = J(\mathbf{y})
\end{aligned}
$$

■

La información mutua de un vector aleatorio $\mathbf{y}$ de dimensión n se puede expresar usando la sintropía como

$$I(y_1, y_2, \cdots, y_n) = J(\mathbf{y}) - \sum_{i=1}^{n} J(y_i) + \frac{1}{2}\log\left(\frac{\prod_{i=1}^{n}\sigma_{ii}}{\det(\mathbf{\Sigma})}\right) \tag{2.27}$$

donde $\mathbf{\Sigma}$ es la matriz de covarianzas del vector aleatorio $\mathbf{y}$ y los σ_{ii} son los elementos de su diagonal principal. Si las variables y_i están incorrelacionadas, el tercer término se anula y de este modo se obtiene

$$I(y_1, y_2, \cdots, y_n) = J(\mathbf{y}) - \sum_{i=1}^{n} J(y_i) \tag{2.28}$$

Por esto, y debido a que la sintropía es invariante ante transformaciones lineales, maximizar la sintropía es equivalente a minimizar la información mutua.

Estimar las componentes independientes s_i mediante la maximización de la sintropía es una ventaja puesto que es una medida de la no-gaussianidad que está bien justificada por la teoría estadística. De hecho, la sintropía es de algún modo el estimador óptimo de la no-gaussianidad, tanto como sus propiedades estadísticas lo permiten. Sin embargo, el cálculo de la sintropía es difícil porque requiere una estimación de la función de densidad. Por consiguiente, para aprovechar sus buenas propiedades estadísticas, se realizan aproximaciones como se verá a continuación.

- **Aproximaciones de la sintropía**

Como se ha mencionado anteriormente, estimar la sintropía es difícil y, por ello, tiene principalmente un uso teórico. En la práctica, se tiene que hacer alguna aproximación.

La aproximación clásica utiliza momentos de orden superior tal y como realizan Jones y Sibson (1987) para obtener

$$J(y) \approx \frac{1}{12}\mathrm{E}\left[y^3\right]^2 + \frac{1}{48}\mathrm{kurt}(y)^2 \tag{2.29}$$

donde se asume que la variable aleatoria y tiene media cero y varianza unidad. No obstante, estas aproximaciones, al igual que la curtosis, no son robustas.

Para solucionar los problemas encontrados con la anterior aproximación, Hyvärinen (1998a) desarrolló nuevas aproximaciones basadas en el principio de entropía máxima donde el caso más general tiene la siguiente forma

$$J(y) \approx \sum_{i=1}^{q} k_i \left\{\mathrm{E}\left[G_i(y)\right] - \mathrm{E}\left[G_i(z)\right]\right\}^2 \tag{2.30}$$

siendo los coeficientes k_i constantes positivas, z es una variable aleatoria gaussiana de media cero y varianza unidad, las funciones G_i son funciones no cuadráticas y, como en la aproximación clásica, se asume que la variable aleatoria y tiene media cero y varianza unidad. La aproximación dada en (2.30) es no negativa y es igual a cero si y solo si la variable aleatoria y tiene una distribución gaussiana.

En el caso más sencillo, cuando se usa una sola función no cuadrática G, la aproximación se convierte en

$$J(y) \approx c\left\{\mathrm{E}\left[G(y)\right] - \mathrm{E}\left[G(z)\right]\right\}^2 \propto \left\{\mathrm{E}\left[G(y)\right] - \mathrm{E}\left[G(z)\right]\right\}^2 \tag{2.31}$$

donde, obviamente, c es una constante irrelevante. Esta construcción, si la distribución de y es simétrica, es evidentemente una generalización de la aproximación basada en los momentos dada en (2.29). Así, para el caso en que $G(y) = y^4$ se obtiene una aproximación basada en la curtosis. Y, ciertamente, G no debe ser cuadrática porque la aproximación de la sintropía sería cero para todas las distribuciones.

Ahora, la cuestión es la elección de la función G. Una sabia elección es tomar G de forma que no crezca demasiado rápido porque así se obtienen estimadores más robustos. Las siguientes alternativas de G, según Hyvärinen (1999a), han demostrado ser bastante útiles

$$\begin{aligned} G_1(y) &= \frac{1}{\alpha}\log\left(\cosh(\alpha y)\right) \\ G_2(y) &= -\exp\left(\frac{-y^2}{2}\right) \end{aligned} \tag{2.32}$$

donde $1 \leq \alpha \leq 2$ es una constante adecuada que con frecuencia toma el valor uno.

De este modo, la familia de aproximaciones dada en (2.31) resulta un buen compromiso entre las propiedades de las dos clásicas medidas de la no-gaussianidad, curtosis y sintropía. Estas aproximaciones son conceptualmente sencillas, rápidas de calcular y tienen propiedades estadísticas atractivas como la robustez.

Debido a estas excelentes propiedades, en este trabajo se utilizará esta familia de aproximaciones de la sintropía para desarrollar un método que estime el modelo ICA.

2.4 Preparación de los datos para ICA

Una vez que se ha elegido una función de contraste para estimar el modelo ICA, es necesario utilizar un algoritmo para su puesta en práctica. Para que dicho algoritmo converja de forma rápida y la estimación del modelo ICA sea más sencilla y esté mejor condicionada, es conveniente preparar los datos observados.

2.4.1 Centrado

En las restricciones para la existencia de un modelo ICA se asume que las componentes independientes tienen media cero y, por tanto, también las variables mezcla, según se deduce de la ecuación (2.8) al tomar esperanzas. Si el supuesto de que las variables tienen media cero no es cierto, se centran los datos observados $\mathbf{x}$ restándoles su vector de medias $\mathbf{m} = \mathrm{E}[\mathbf{x}]$

$$\mathbf{x}_0 = \mathbf{x} - \mathbf{m} \tag{2.33}$$

De este modo, se asegura que las componentes independientes también tienen media cero puesto que $\mathrm{E}[\mathbf{s}_0] = \mathbf{A}^{-1}\mathrm{E}[\mathbf{x}_0]$. Después de estimar la matriz de mezcla $\mathbf{A}$ y las componentes independientes con media cero, se puede reconstruir la media añadiendo $\mathbf{A}^{-1}\mathbf{m}$ a las componentes independientes centradas según se muestra

$$\mathbf{s}_0 = \mathbf{A}^{-1}\mathbf{x}_0 = \mathbf{A}^{-1}(\mathbf{x} - \mathbf{m}) = \mathbf{A}^{-1}\mathbf{x} - \mathbf{A}^{-1}\mathbf{m} = \mathbf{s} - \mathbf{A}^{-1}\mathbf{m} \Rightarrow \mathbf{s} = \mathbf{s}_0 + \mathbf{A}^{-1}\mathbf{m} \tag{2.34}$$

2.4.2 Blanqueo

Una propiedad algo más fuerte que la incorrelación es el blanqueo.

Definición 2.10 *Un vector aleatorio de media cero,* $\mathbf{z}$ *, se dice que es blanco cuando sus componentes están incorrelacionadas y sus varianzas son iguales a la unidad. En otras palabras, su matriz de covarianzas es la matriz identidad,* $\mathrm{E}\left[\mathbf{z}\mathbf{z}^T\right]=\mathbf{I}$ *.*

Según la definición anterior, blanquear significa transformar linealmente el vector de datos observados $\mathbf{x}$ (previamente centrados) multiplicándolo por una matriz $\mathbf{Q}$

$$\mathbf{z}=\mathbf{Q}\mathbf{x} \tag{2.35}$$

de forma que el nuevo vector $\mathbf{z}$ sea blanco.

Un conocido método para blanquear se obtiene a partir de la diagonalización ortogonal de la matriz de covarianzas de las variables mezcla observadas

$$\mathrm{E}\left[\mathbf{x}\mathbf{x}^T\right]=\mathbf{C}=\mathbf{U}\mathbf{D}\mathbf{U}^T \tag{2.36}$$

donde $\mathbf{C}=\mathbf{A}\mathbf{A}^T$ se obtiene calculando las covarianzas del modelo dado en (2.8), $\mathbf{U}$ es una matriz ortogonal de autovectores de $\mathbf{C}$ y $\mathbf{D}=\mathrm{diag}\left(\lambda_1,\cdots,\lambda_p\right)$ es la matriz diagonal con los autovalores de $\mathbf{C}$. Así, la matriz blanqueadora para el vector de datos observados $\mathbf{x}$ será la matriz dada por

$$\mathbf{Q}=\mathbf{D}^{-1/2}\mathbf{U}^T \tag{2.37}$$

donde $\mathbf{D}^{-1/2}=\mathrm{diag}\left(\lambda_1^{-1/2},\cdots,\lambda_p^{-1/2}\right)$.

Por otro lado, el blanqueo es de gran utilidad para preparar los datos. De las ecuaciones (2.8) y (2.35) se obtiene una nueva matriz de mezcla dada por

$$\mathbf{z}=\mathbf{Q}\mathbf{x}=\mathbf{Q}\mathbf{A}\mathbf{s}=\mathbf{B}^T\mathbf{s} \tag{2.38}$$

Por tanto, la utilidad del blanqueo reside en el hecho de que la nueva matriz de mezcla $\mathbf{B}^T=\mathbf{Q}\mathbf{A}$ es ortogonal según se puede ver

$$\mathbf{I}=\mathrm{E}\left[\mathbf{z}\mathbf{z}^T\right]=\mathbf{B}^T\,\mathrm{E}\left[\mathbf{s}\mathbf{s}^T\right]\mathbf{B}=\mathbf{B}^T\mathbf{B}\Rightarrow\mathbf{B}^T\mathbf{B}=\mathbf{I} \tag{2.39}$$

De este modo, se puede restringir la búsqueda de las matrices de mezcla dentro del conjunto de matrices ortogonales con la ventaja de que para una matriz ortogonal se deben estimar solo $p\left(p-1\right)/2$ parámetros en lugar de p^2, suponiendo que $p=n$.

No obstante, el blanqueo, aunque esté relacionado con la independencia, no resuelve el problema de la estimación del modelo ICA. Así, si se considera una transformación

ortogonal $\mathbf{V}$ de $\mathbf{z}$, $\mathbf{y} = \mathbf{Vz}$, debido a la ortogonalidad de $\mathbf{V}$, $\mathbf{V}^T = \mathbf{V}^{-1}$, se llega a

$$\mathrm{E}\left[\mathbf{y}\mathbf{y}^T\right] = \mathbf{V}\,\mathrm{E}\left[\mathbf{z}\mathbf{z}^T\right]\mathbf{V}^T = \mathbf{V}\mathbf{I}\mathbf{V}^T = \mathbf{I} \tag{2.40}$$

Es decir, el vector aleatorio $\mathbf{y}$ también es un vector blanco y, por ello, la matriz $\mathbf{U}\mathbf{D}^{-1/2}\mathbf{U}^T$ también es una matriz blanqueadora ya que resulta de multiplicar por la izquierda la matriz $\mathbf{Q}$ de la ecuación (2.37) por la matriz ortogonal $\mathbf{U}$. Debido a que $\mathbf{y}$ podría ser cualquier transformación ortogonal de $\mathbf{z}$, el blanqueo proporciona las componentes independientes salvo una transformación ortogonal o, como dicen Hyvärinen *et al.* (2001), el blanqueo es solo la mitad de la estimación del modelo ICA.

2.4.3 Un ejemplo ilustrativo

Para ilustrar el significado de un modelo ICA y la preparación de los datos (blanqueo de los datos centrados), se consideran dos componentes independientes s_1 y s_2, que siguen una idéntica distribución uniforme con función de densidad

$$f_i(s_i) = \begin{cases} \dfrac{1}{2\sqrt{3}} & \text{si } |s_i| \le \sqrt{3} \\ 0 & \text{en otro caso} \end{cases}, \quad i = 1, 2 \tag{2.41}$$

Esta distribución uniforme tiene media cero y varianza unidad según exigen las hipótesis para la existencia del modelo ICA.

La función de densidad conjunta de s_1 y s_2 es uniforme sobre un cuadrado porque es el producto de las densidades marginales. La densidad conjunta se ilustra en la Figura 2.1 donde se muestran los puntos seleccionados aleatoriamente de dicha distribución junto con las distribuciones de probabilidad marginales de la muestra.

A continuación, se mezclan las dos componentes independientes con la siguiente matriz de mezcla

$$\mathbf{A} = \begin{pmatrix} 6 & 4 \\ 2 & 6 \end{pmatrix}$$

De esta forma, se obtienen dos variables mezcla, x_1 y x_2. La distribución conjunta de los datos mezclados es una distribución uniforme sobre un paralelogramo como se aprecia en la Figura 2.2. En ella, también se observa que las distribuciones de probabilidad marginales tienden a una distribución gaussiana según establece el teorema central del límite para la distribución de la suma de variables aleatorias independientes. Queda patente que ninguna de las variables x_1 y x_2 son independientes porque es posible predecir el valor de una de ellas a partir del valor de la otra.

Claramente, si x_1 alcanza su valor máximo o mínimo, entonces queda determinado completamente el valor de x_2. Esto, sin embargo, no ocurre para las variables s_1 y s_2. Por otro lado, se observa que los lados del paralelogramo, en la Figura 2.2, están en las direcciones de los vectores columna de la matriz de mezcla $\mathbf{A}$. Este hecho podría hacer pensar que el problema de la estimación de la matriz de mezcla tiene una solución sencilla, bastaría estimar la densidad conjunta de las variables observadas y determinar sus lados. En la práctica, esto no es posible porque para la mayoría de las distribuciones, incluidas las gaussianas, estos lados no se pueden encontrar.

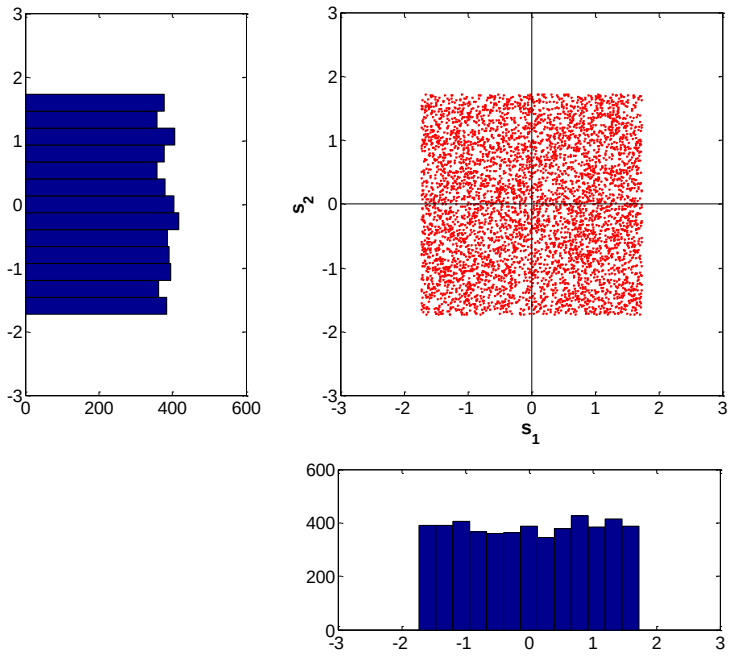

Figura 2.1 La distribución conjunta de las componentes independientes s_1 y s_2 con distribuciones uniformes y sus distribuciones marginales.

Finalmente, los efectos del blanqueo de las variables x_1 y x_2 se ilustran de manera gráfica en la Figura 2.3. El cuadrado que define la distribución conjunta de las variables blanqueadas, z_1 y z_2, es claramente una versión rotada o transformación ortogonal de la versión original de la Figura 2.1. Además, las distribuciones de probabilidad marginales son menos gaussianas que las de la Figura 2.2 como era de esperar. Ahora, todo lo que resta, es estimar un simple ángulo que determine la rotación.

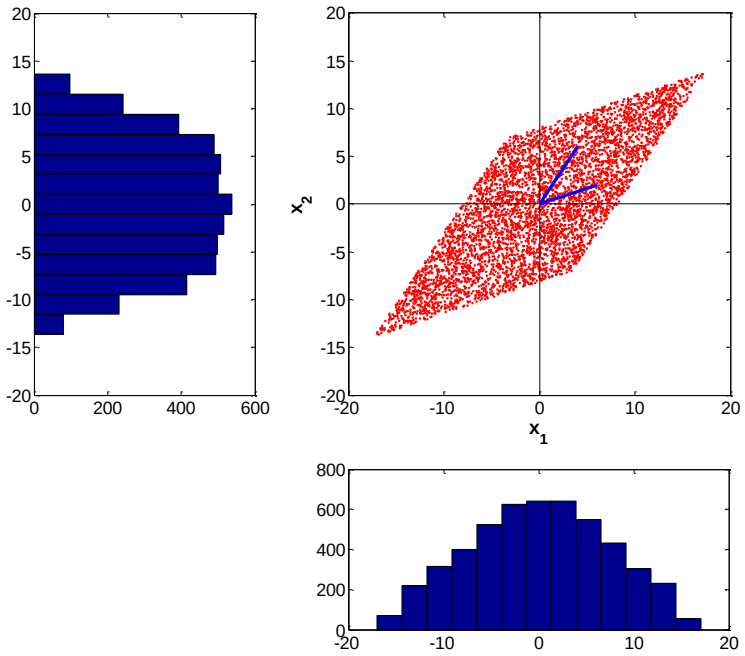

Figura 2.2 La distribución conjunta de las mezclas observadas x_1 y x_2, sus distribuciones marginales y los vectores columna de la matriz $\mathbf{A}$.

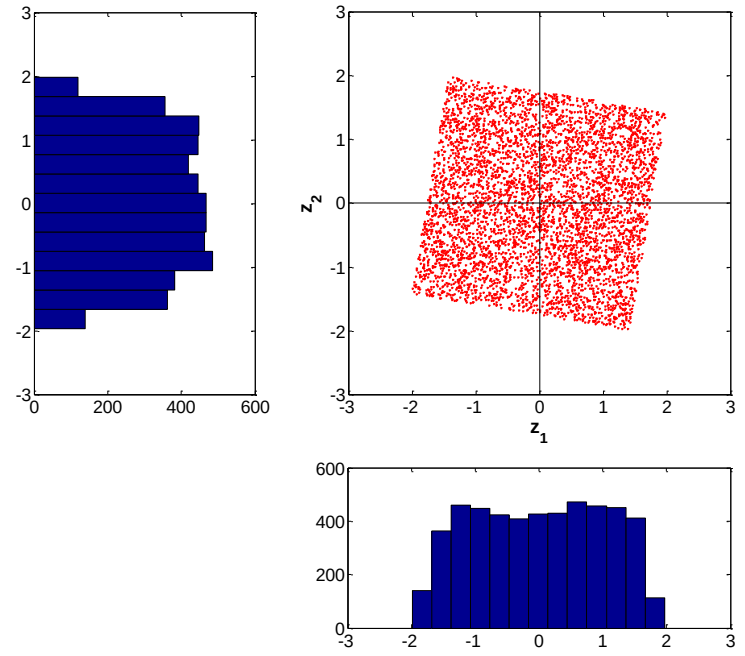

Figura 2.3 La distribución conjunta de las mezclas blanqueadas z_1 y z_2 y sus distribuciones marginales.

2.4.4 Filtros lineales e innovaciones

En múltiples ocasiones, las variables aleatorias observadas son series temporales. De este modo, el índice t en $x_i(t)$ representa un índice temporal. En tales casos, puede ser muy útil filtrar las señales observadas para, por ejemplo, conseguir estacionariedad en sentido estricto o una mayor independencia de los procesos observados según se exige para la existencia de un modelo ICA con series temporales o procesos estocásticos.

Para series temporales, cualquier filtro lineal es permitido porque no cambia el modelo ICA tal y como demuestran Hyvärinen y Oja (2000) en el siguiente teorema.

Teorema 2.2 *Si a las señales observadas* $\mathbf{x}(t)$ *se les aplica un filtro lineal para obtener nuevas señales* $\mathbf{x}^*(t)$*, entonces el modelo ICA es válido para* $\mathbf{x}^*(t)$ *con la misma matriz de mezcla.*

Demostración. Sea $\mathbf{X}$ la matriz de dimensión $(p \times T)$ cuyas columnas son los vectores de observaciones $\mathbf{x}(1), \mathbf{x}(2), \cdots, \mathbf{x}(T)$, es decir $\mathbf{X} = [\mathbf{x}(1) \mid \mathbf{x}(2) \mid \cdots \mid \mathbf{x}(T)]$, y sea $\mathbf{S}$ la matriz de dimensión $(p \times T)$ cuyas columnas son los vectores de componentes independientes $\mathbf{s}(1), \mathbf{s}(2), \cdots, \mathbf{s}(T)$, es decir $\mathbf{S} = [\mathbf{s}(1) \mid \mathbf{s}(2) \mid \cdots \mid \mathbf{s}(T)]$. Entonces, el modelo ICA dado por (2.8) se puede expresar como

$$\mathbf{X} = \mathbf{AS} \tag{2.42}$$

Ahora, sea $\mathbf{M}$ la matriz de dimensión $(T \times T)$ correspondiente a un filtro lineal.

El filtrado de $\mathbf{X}$ se realiza multiplicando dicha matriz por la derecha por la matriz $\mathbf{M}$ para obtener el siguiente resultado

$$\mathbf{X}^* = \mathbf{XM} = \mathbf{ASM} = \mathbf{AS}^* \tag{2.43}$$

Esto demuestra que el modelo ICA continua siendo válido. Las componentes independientes son filtradas con el mismo filtro que se aplica a las mezclas. ■

Debido a que la matriz de mezcla se mantiene sin cambios, se pueden utilizar las observaciones filtradas solo para la estimación de la matriz de mezcla $\mathbf{A}$ del modelo ICA y, posteriormente, aplicar dicha matriz de mezcla estimada sobre los datos observados originales con el fin de obtener las componentes independientes.

A continuación se analizan las propiedades de los diferentes tipos de filtros.

- **Filtros de paso bajo**

Los filtros de paso bajo sustituyen cada valor muestral observado por una media ponderada del valor presente y de los valores anteriores y posteriores con el objetivo de reducir el ruido en los datos para que los métodos de estimación del modelo ICA trabajen mejor. Sin embargo, pueden reducir la información que contienen los datos porque los ciclos asociados a las altas frecuencias se pierden y, a menudo, esto puede suponer también una menor independencia.

- **Filtros de paso alto e innovaciones**

Los filtros de paso alto eliminan la tendencia y los ciclos muy largos asociados a las bajas frecuencias. Un forma habitual de conseguirlo es diferenciando las series, es decir, se sustituye cada valor muestral por la diferencia entre él y su valor precedente. La diferenciación es un caso particular de los procesos de innovaciones, según describen Hyvärinen *et al.* (2001), los cuales se estudian a continuación.

Definición 2.11 (*Procesos de innovaciones) Dado un proceso estocástico vectorial* $\mathbf{s}(t)$, *se define su proceso de innovaciones,* $\tilde{\mathbf{s}}(t)$, *como el error de la mejor predicción (en términos de esperanza condicionada) de* $\mathbf{s}(t)$ *dado su pasado*

$$\tilde{\mathbf{s}}(t) = \mathbf{s}(t) - E\left[\mathbf{s}(t) \,|\, \mathbf{s}(t-1), \mathbf{s}(t-2), \cdots, \mathbf{s}(1)\right] \tag{2.44}$$

Con la expresión "innovación" se quiere indicar que $\tilde{\mathbf{s}}(t)$ contiene toda la nueva información al observar $\mathbf{s}(t)$ en el momento t. El caso más sencillo es el proceso diferencia que se obtiene al diferenciar una serie y viene dado por

$$\tilde{\mathbf{s}}(t) = \Delta\mathbf{s}(t) = \mathbf{s}(t) - \mathbf{s}(t-1) \tag{2.45}$$

El concepto de innovaciones puede utilizarse en la estimación del modelo ICA gracias al siguiente teorema enunciado por Hyvärinen (1998b).

Teorema 2.3 *Si* $\mathbf{x}(t)$ *y* $\mathbf{s}(t)$ *siguen el modelo ICA dado por* $\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t)$, *entonces los procesos de innovaciones* $\tilde{\mathbf{x}}(t)$ *y* $\tilde{\mathbf{s}}(t)$ *siguen dicho modelo ICA,* $\tilde{\mathbf{x}}(t) = \mathbf{A}\tilde{\mathbf{s}}(t)$. *En particular, las componentes* $\tilde{s}_i(t)$ *son independientes.*

Demostración. Para la demostración, ver Hyvärinen (1998b). ■

Por otro lado, la independencia de las innovaciones $\tilde{s}_i(t)$ no implica la independencia de las componentes $s_i(t)$. De este modo, las innovaciones son a menudo

más independientes que los procesos originales. Además, las innovaciones son habitualmente menos gaussianas que los procesos originales porque estos son una media móvil de las innovaciones y, por el teorema central del límite, toda suma de variables aleatorias tiende a ser más gaussiana que las variables originales. Por todo ello, las innovaciones son más susceptibles de cumplir los supuestos y las restricciones para la existencia del modelo ICA.

Debido a que el cálculo de los procesos de innovaciones es equivalente al filtrado de paso alto, los filtros de paso alto gozan de las mismas propiedades que los procesos de innovaciones. Sin embargo, un problema de los filtros de paso alto es que pueden incrementar el ruido en las observaciones filtradas por las mismas razones que los filtros de paso bajo lo reducen.

- **Filtros de paso en banda**

Teniendo en cuenta los resultados anteriores, lo ideal sería obtener un filtro que combine la reducción del ruido con el aumento de la independencia, es decir, que alcance un compromiso entre las ventajas de los filtros de paso bajo y de paso alto. Esto, con muy buenos resultados, lo llevan a cabo los filtros de paso en banda según se describe en Bógalo y Quilis (2003). Los filtros de paso en banda ignoran las frecuencias altas y bajas permitiendo solo una banda de frecuencias entre ellas. Será el investigador, en cada situación, el que determine esta banda de paso según los datos puesto que es imposible ofrecer una respuesta general.

2.5 El algoritmo "FastICA"

Una vez que se ha elegido una función de contraste para maximizar la no-gaussianidad y los datos observados se preparan, se centran y se blanquean, se necesita diseñar un algoritmo para implantar la optimización de la función de contraste.

Desde el trabajo pionero de Jutten y Herault (1991), donde su algoritmo se basaba en la cancelación de correlaciones cruzadas no lineales, se han desarrollado diversos algoritmos para la optimización de las diferentes funciones de contraste. En este trabajo se elige, para su implementación, el algoritmo de punto fijo llamado FastICA, Hyvärinen y Oja (1997). Esta elección se debe a las excelentes propiedades que posee dicho algoritmo: rapidez de convergencia, estabilidad, requiere pocos cálculos y sirve para estimar tanto una sola componente independiente como varias al mismo tiempo.

2.5.1 FastICA para una sola componente

FastICA es un algoritmo de punto fijo diseñado para encontrar una dirección, es decir, un vector $\mathbf{b}$ tal que la proyección $\mathbf{b}^T\mathbf{z}$ del vector blanqueado $\mathbf{z}$ maximice la no-gaussianidad. FastICA fue desarrollado por Hyvärinen y Oja (1997) utilizando la curtosis como medida de la no-gaussianidad y Hyvärinen (1999a) lo adapta para utilizar como medida de la no-gaussianidad la sintropía. En este trabajo, para medir la no-gaussianidad de $\mathbf{b}^T\mathbf{z}$ se utiliza la aproximación de su sintropía, $J\left(\mathbf{b}^T\mathbf{z}\right)$, dada por la ecuación (2.31). La restricción, para la existencia de un modelo ICA, de que la varianza de $\mathbf{b}^T\mathbf{z}$ debe ser la unidad es equivalente a que la norma euclídea del vector $\mathbf{b}$ sea la unidad cuando los datos están blanqueados como $\mathbf{z}$.

El algoritmo FastICA es un método de iteración de punto fijo, en concreto, utiliza la aproximación de Newton para encontrar el máximo de la no-gaussianidad de $\mathbf{b}^T\mathbf{z}$.

- **Derivación del algoritmo**

La derivación del método de Newton tiene en cuenta que maximizar la aproximación de la sintropía de $\mathbf{b}^T\mathbf{z}$ es equivalente a optimizar $\mathrm{E}\left[G\left(\mathbf{b}^T\mathbf{z}\right)\right]$. Esto, junto con la restricción, ya mencionada, de que $\mathrm{E}\left[\left(\mathbf{b}^T\mathbf{z}\right)^2\right]=\|\mathbf{b}\|_2^2=\mathbf{b}^T\mathbf{b}=1$, permite construir el lagrangiano cuya expresión llega a ser

$$L\left(\mathbf{b},\beta\right)=\mathrm{E}\left[G\left(\mathbf{b}^T\mathbf{z}\right)\right]-\beta\left(\mathbf{b}^T\mathbf{b}-1\right) \tag{2.46}$$

donde la función G es alguna de las mostradas en (2.32). Por tanto, los óptimos se obtienen de los puntos para los que el gradiente del lagrangiano se hace cero

$$\frac{\partial L}{\partial \mathbf{b}}=0\Leftrightarrow \mathrm{E}\left[\mathbf{z}\,g\left(\mathbf{b}^T\mathbf{z}\right)\right]-\beta\mathbf{b}=0 \tag{2.47}$$

donde la función g es la primera derivada de la función G, es decir, las alternativas para g están dadas por

$$\begin{aligned} g_1(y)&=\tanh\left(\alpha y\right)\\ g_2(y)&=y\exp\left(\frac{-y^2}{2}\right)\end{aligned} \tag{2.48}$$

Resolver la ecuación (2.47) por el método de Newton es equivalente a encontrar el óptimo del lagrangiano por el método de Newton. Si se define la función

$$F\left(\mathbf{b}\right)=\mathrm{E}\left[\mathbf{z}\,g\left(\mathbf{b}^T\mathbf{z}\right)\right]-\beta\mathbf{b} \tag{2.49}$$

la iteración del método de Newton consiste en hacer

$$\mathbf{b} \leftarrow \mathbf{b} - \left[\frac{\partial F(\mathbf{b})}{\partial \mathbf{b}} \right]^{-1} F(\mathbf{b}) \tag{2.50}$$

La matriz jacobiana de la función F, la matriz hessiana del lagrangiano, es

$$\frac{\partial F(\mathbf{b})}{\partial \mathbf{b}} = \mathrm{E}\left[\mathbf{z}\mathbf{z}^T g'\left(\mathbf{b}^T\mathbf{z}\right) \right] - \beta \mathbf{I} \tag{2.51}$$

donde g' es la derivada de la función g. Con el fin de facilitar los cálculos, se puede simplificar la inversa de la matriz obtenida en (2.51) realizando una aproximación del primer término. Puesto que $\mathbf{z}$ es un vector aleatorio blanqueado, una aproximación razonable parece ser

$$\mathrm{E}\left[\mathbf{z}\mathbf{z}^T g'\left(\mathbf{b}^T\mathbf{z}\right) \right] \approx \mathrm{E}\left[\mathbf{z}\mathbf{z}^T \right] \mathrm{E}\left[g'\left(\mathbf{b}^T\mathbf{z}\right) \right] = \mathrm{E}\left[g'\left(\mathbf{b}^T\mathbf{z}\right) \right] \mathbf{I} \tag{2.52}$$

Así, la matriz jacobiana de la función F se convierte en una matriz diagonal que se puede invertir fácilmente y, de este modo, se obtiene la siguiente aproximación del método de Newton

$$\mathbf{b} \leftarrow \mathbf{b} - \left\{ \mathrm{E}\left[\mathbf{z}\, g\left(\mathbf{b}^T\mathbf{z}\right) \right] - \beta \mathbf{b} \right\} \Big/ \left\{ \mathrm{E}\left[g'\left(\mathbf{b}^T\mathbf{z}\right) \right] - \beta \right\} \tag{2.53}$$

Este algoritmo puede simplificarse bastante multiplicando ambos lados de (2.53) por $\beta - \mathrm{E}\left[g'\left(\mathbf{b}^T\mathbf{z}\right) \right]$ y, teniendo en cuenta que $\beta = \mathrm{E}\left[\mathbf{b}^T\mathbf{z}\, g\left(\mathbf{b}^T\mathbf{z}\right) \right]$, después de sencillas operaciones algebraicas se obtiene

$$\mathbf{b} \leftarrow \mathrm{E}\left[\mathbf{z}\, g\left(\mathbf{b}^T\mathbf{z}\right) \right] - \mathrm{E}\left[g'\left(\mathbf{b}^T\mathbf{z}\right) \right] \mathbf{b} \tag{2.54}$$

El resultado (2.54) constituye la iteración básica del algoritmo FastICA de punto fijo.

- **Descripción del algoritmo**

  La derivación anterior permite detallar el algoritmo FastICA en los siguientes pasos:

  1. Centrar los datos observados $\mathbf{x}$ para que su media sea cero.
  2. Blanquear los datos centrados para obtener el vector aleatorio $\mathbf{z}$.
  3. Elegir, de forma aleatoria, un vector inicial $\mathbf{b}$ de norma unidad.
  4. Hacer

     $$\mathbf{b} \leftarrow \mathrm{E}\left[\mathbf{z}\, g\left(\mathbf{b}^T\mathbf{z}\right) \right] - \mathrm{E}\left[g'\left(\mathbf{b}^T\mathbf{z}\right) \right] \mathbf{b}$$

     donde g es alguna función de las definidas en (2.48).

5. Normalizar

$$\mathbf{b} \leftarrow \mathbf{b} / \|\mathbf{b}\|_2$$

6. Si no converge, volver al paso 4.

La convergencia significa que el valor antiguo y el valor actual del vector $\mathbf{b}$ deben indicar la misma dirección, es decir, el valor absoluto de su producto escalar debe ser (casi) igual a 1. No es necesario que el vector converja a un único punto puesto que $\mathbf{b}$ y $-\mathbf{b}$ definen la misma dirección.

Se debe observar, que, en la práctica, las esperanzas son sustituidas por sus estimaciones realizadas con las medias muestrales.

2.5.2 FastICA para varias componentes

Hasta ahora, se ha estimado una sola componente independiente porque se ha utilizado una función de contraste de una unidad, la sintropía como medida de la no-gaussianidad. En principio, se podrían encontrar más componentes independientes ejecutando el algoritmo en diferentes ocasiones usando distintos vectores iniciales, sin embargo, esto no es un método eficaz porque podrían converger al mismo óptimo.

La matriz de mezcla para los datos blanqueados $\mathbf{B}^T$ es ortogonal con lo cual los vectores $\mathbf{b}_i$, correspondientes a las diferentes componentes independientes, son ortonormales. Además, puesto que los $\mathbf{b}_i$ son las filas de la inversa de la matriz de mezcla, resulta que los $\mathbf{b}_i$ son iguales a las columnas de la matriz de mezcla puesto que $\mathbf{B}^{-1} = \mathbf{B}^T$ debido a la ortogonalidad. Esta propiedad permite extender el método de maximizar la no-gaussianidad para estimar varias componentes independientes. Para ello, se necesita ejecutar el algoritmo para una sola componente varias veces teniendo presente que los vectores $\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_m$ se deben ortogonalizar después de cada iteración para evitar que converjan al mismo máximo. Siguiendo a Hyvärinen *et al.* (2001), en este trabajo se presentan dos métodos para conseguir ortogonalidad.

- **Ortogonalización por reducción: estimación en serie**

Una manera sencilla de ortogonalizar es la ortogonalización por reducción usando el método de Gram-Schmidt. Esto requiere estimar las componentes una a una, es decir, los vectores $\mathbf{b}_i$ se estiman en serie. Cuando se han estimado k componentes independientes, o k vectores $\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_k$, se ejecuta el algoritmo para una sola componente con $\mathbf{b}_{k+1}$ y después de cada iteración se ortogonaliza por reducción respecto

a los k vectores previos estimados. De una forma más detallada, el algoritmo de ortogonalización por reducción consta de los siguientes pasos:

1. Centrar los datos observados $\mathbf{x}$ para que su media sea cero.
2. Blanquear los datos centrados para obtener el vector aleatorio $\mathbf{z}$.
3. Elegir el número de componentes independientes m. Iniciar el contador, $k \leftarrow 1$.
4. Elegir, de forma aleatoria, un vector inicial $\mathbf{b}_k$ de norma unidad.
5. Hacer

$$\mathbf{b}_k \leftarrow \mathrm{E}\left[\mathbf{z}\, g\left(\mathbf{b}_k^T \mathbf{z}\right)\right] - \mathrm{E}\left[g'\left(\mathbf{b}_k^T \mathbf{z}\right)\right]\mathbf{b}_k$$

   donde g es alguna función de las definidas en (2.48).

6. Ortogonalizar por reducción

$$\mathbf{b}_k \leftarrow \mathbf{b}_k - \sum_{j=1}^{k-1}\left(\mathbf{b}_j^T \mathbf{b}_k\right)\mathbf{b}_j \tag{2.55}$$

7. Normalizar

$$\mathbf{b}_k \leftarrow \mathbf{b}_k / \left\|\mathbf{b}_k\right\|_2$$

8. Si $\mathbf{b}_k$ no converge, volver al paso 5.
9. Incrementar el contador, $k \leftarrow k+1$. Si $k \le m$, volver al paso 4.

- **Ortogonalización simétrica: estimación en paralelo**

En determinadas ocasiones puede ser deseable utilizar una ortogonalización simétrica de forma que ningún vector sea privilegiado sobre el resto como exponen Karhunen *et al.* (1997). Esto significa que los vectores $\mathbf{b}_i$ se estimarán en paralelo y no uno a uno. Un motivo para optar por esta estrategia es que la ortogonalización por reducción tiene el inconveniente de que los errores en la estimación de los vectores se van acumulando. Otro motivo es, precisamente, que los métodos de ortogonalización simétrica permiten el cálculo en paralelo, al mismo tiempo, de las componentes independientes.

La ortogonalización simétrica se realiza efectuando primero el paso iterativo del algoritmo para una sola componente sobre todos los vectores $\mathbf{b}_i$ en paralelo y, posteriormente, se ortogonalizan todos los $\mathbf{b}_i$ por métodos simétricos.

El método clásico de ortogonalización simétrica implica raíces cuadradas de matrices puesto que la iteración básica es

$$\mathbf{B} \leftarrow \left(\mathbf{B}\mathbf{B}^T\right)^{-1/2}\mathbf{B} \tag{2.56}$$

donde $\mathbf{B} = \left[\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_m\right]^T$ es la matriz de los vectores y la raíz cuadrada $\left(\mathbf{B}\mathbf{B}^T\right)^{-1/2}$ se obtiene por la diagonalización ortogonal de $\mathbf{B}\mathbf{B}^T = \mathbf{U}\mathbf{D}\mathbf{U}^T$ como

$$\left(\mathbf{B}\mathbf{B}^T\right)^{-1/2} = \mathbf{U}\mathbf{D}^{-1/2}\mathbf{U}^T = \mathbf{U}\,\mathrm{diag}\left(\lambda_1^{-1/2}, \cdots, \lambda_m^{-1/2}\right)\mathbf{U}^T \tag{2.57}$$

siendo $\mathbf{U}$ una matriz ortogonal de autovectores de $\mathbf{B}\mathbf{B}^T$ y $\mathbf{D} = \mathrm{diag}\left(\lambda_1, \cdots, \lambda_m\right)$ la matriz diagonal con los autovalores de $\mathbf{B}\mathbf{B}^T$.

Una alternativa más sencilla es el algoritmo iterativo propuesto por Hyvärinen (1999a) cuyos pasos son:

1. $\mathbf{B} \leftarrow \mathbf{B} / \|\mathbf{B}\|$
2. $\mathbf{B} \leftarrow \frac{3}{2}\mathbf{B} - \frac{1}{2}\mathbf{B}\mathbf{B}^T\mathbf{B}$ (2.58)
3. Si $\mathbf{B}\mathbf{B}^T$ no converge a la matriz identidad, volver al paso 2.

La norma en el paso 1 puede ser cualquier norma matricial excepto la norma de Frobenius.

De una forma más detallada, el algoritmo de ortogonalización simétrica para estimar varias componentes independientes consta de los siguientes pasos:

1. Centrar los datos observados $\mathbf{x}$ para que su media sea cero.
2. Blanquear los datos centrados para obtener el vector aleatorio $\mathbf{z}$.
3. Elegir el número de componentes independientes m.
4. Elegir, de forma aleatoria, valores iniciales para los $\mathbf{b}_i$, con $i = 1, 2, \cdots, m$, cada uno de norma unidad. Ortogonalizar la matriz $\mathbf{B} = \left[\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_m\right]^T$ según el posterior paso 6.
5. Para cada $i = 1, 2, \cdots, m$, hacer

   $$\mathbf{b}_i \leftarrow \mathrm{E}\left[\mathbf{z}\, g\left(\mathbf{b}_i^T\mathbf{z}\right)\right] - \mathrm{E}\left[g'\left(\mathbf{b}_i^T\mathbf{z}\right)\right]\mathbf{b}_i$$

   donde g es alguna función de las definidas en (2.48).
6. Realizar una ortogonalización simétrica de la matriz $\mathbf{B}$, bien por la iteración del método clásico según (2.56), bien por el algoritmo iterativo definido en (2.58).
7. Si la matriz $\mathbf{B}$ no converge, volver al paso 5.

Según se comentó en el algoritmo para una sola componente, la convergencia significa que el valor antiguo y el valor actual de los vectores $\mathbf{b}_i$, las filas de la matriz $\mathbf{B}$, deben indicar las mismas direcciones, es decir, el valor absoluto de su producto escalar debe ser (casi) igual a 1. No es necesario que los vectores converjan a un único punto puesto que $\mathbf{b}_i$ y $-\mathbf{b}_i$ definen la misma dirección.

2.6 ICA para series temporales. EL algoritmo "AMUSE"

El modelo ICA considerado hasta ahora consistía en una mezcla lineal de variables aleatorias independientes donde el orden de las muestras de $\mathbf{x}$ no tiene importancia y puede ser alterado. En muchas aplicaciones, sin embargo, no se mezclan variables aleatorias sino series temporales en las que si tiene importancia el orden.

En esta sección, se considera la estimación del modelo ICA cuando las componentes independientes son series temporales $s_i(t)$ con $t=1,2,\cdots,T$. El modelo se puede expresar matricialmente de la forma

$$\mathbf{x}(t)=\mathbf{A}\mathbf{s}(t) \tag{2.59}$$

donde se asume que $\mathbf{A}$ es cuadrada como de costumbre y las componentes no observadas son independientes. En cambio, las componentes no observadas no necesitan ser no-gaussianas.

Las componentes independientes contienen más información que las variables aleatorias para la resolución del problema de separación porque son series temporales. Así, se pueden calcular las autocovarianzas lo que permite mejorar la estimación del modelo, sobre todo en los casos de dependencia temporal, y evita el uso de estadísticos de orden superior.

Para permitir la estimación del modelo haciendo uso de esa información adicional, es necesario realizar el supuesto adicional de que las componentes independientes deben tener distintas autocovarianzas y estas deben ser no nulas.

2.6.1 Las autocovarianzas como alternativa a la gaussianidad

La forma más sencilla de estructura temporal viene dada por las autocovarianzas, es decir por las covarianzas entre las señales en diferentes instantes de tiempo, $\mathrm{cov}\left(x_i(t), x_j(t-\tau)\right)$ donde τ es algún retardo constante, $\tau=1,2,\cdots$. De este modo, se obtiene la matriz de autocovarianzas o de covarianzas retardadas en el tiempo

$$\mathbf{C}_\tau^{\mathbf{x}} = E\left[\mathbf{x}(t)\mathbf{x}(t-\tau)^T\right] \tag{2.60}$$

donde el elemento (i, j) viene dado por $\text{cov}\left(x_i(t), x_j(t-\tau)\right)$ para un retardo τ.

Según se vio en el apartado 2.4.2, el problema en ICA es que la matriz de covarianzas contemporáneas $\mathbf{C}_0^{\mathbf{x}} = \mathbf{A}\mathbf{A}^T$ no contiene información suficiente que permita estimar $\mathbf{A}$. Quiere decir que para estimar las componentes independientes no es suficiente con encontrar una matriz $\mathbf{Q}$ de forma que sea blanco el vector

$$\mathbf{z}(t) = \mathbf{Q}\mathbf{x}(t) \tag{2.61}$$

Esto es debido a que existen infinitas matrices blanqueadoras $\mathbf{Q}$ que obtienen componentes incorrelacionadas puesto que en el modelo ICA básico se tiene que usar la estructura no-gaussiana. Si las señales fueran gaussianas, puesto que la incorrelación implica independencia, al blanquear las series observadas se convierten en incorrelacionadas y por tanto en independientes de forma que la matriz blanqueadora sería $\mathbf{QA}$, no la matriz $\mathbf{Q}$, no pudiéndose estimar entonces la matriz $\mathbf{A}$.

La cuestión clave es que la información que facilita la matriz de covarianzas con determinado retardo $\mathbf{C}_\tau^{\mathbf{x}}$ se puede utilizar, según Tong el al. (1991), en lugar de estadísticos de orden superior. Se desea encontrar una matriz $\mathbf{V}$ que haga nulas tanto las covarianzas contemporáneas de $\mathbf{y}(t) = \mathbf{V}\mathbf{x}(t)$ como las covarianzas retardadas

$$E\left[y_i(t)\, y_j(t-\tau)\right] = 0 \qquad \forall i \neq j, \tau \tag{2.62}$$

De esta forma se verificará que

$$\mathbf{C}_\tau^{\mathbf{y}} = E\left[\mathbf{y}(t)\mathbf{y}(t-\tau)^T\right] = \mathbf{V}\mathbf{C}_\tau^{\mathbf{x}}\mathbf{V}^T = (\mathbf{V}\mathbf{A})\mathbf{C}_\tau^{\mathbf{s}}(\mathbf{V}\mathbf{A})^T$$

donde $\mathbf{C}_\tau^{\mathbf{s}}$ es una matriz diagonal porque las covarianzas retardadas de $\mathbf{s}(t)$ son todas nulas debido a la independencia de las componentes no observadas $s_i(t)$.

2.6.2 El algoritmo AMUSE

El algoritmo AMUSE (*Algorithm for Multiple Unknown Signals Extraction*), Tong et al. (1990), utiliza una estructura temporal, aplica estadísticos de segundo orden, covarianzas, con el fin de estimar una matriz de separación y obtener las componentes independientes. Este algoritmo se basa en la justificación dada en el apartado anterior de forma que se superan las dificultades que presentan los algoritmos basados en medidas de no-gaussianidad cuando las componentes no observadas son gaussianas.

- **Derivación del algoritmo**

El algoritmo utiliza un solo retardo temporal τ para el cálculo de las matrices de autocovarianzas. Habitualmente se toma $\tau = 1$ por simplicidad. Se desea encontrar una matriz que anule tanto las covarianzas contemporáneas como las correspondientes al retardo τ.

Sean $\mathbf{z}(t) = \mathbf{Q}\mathbf{x}(t)$ los datos blanqueados. Entonces, para la matriz ortogonal de separación $\mathbf{B} = (\mathbf{QA})^T$ se verifica

$$\mathbf{B}\mathbf{z}(t) = \mathbf{s}(t) \tag{2.63}$$

$$\mathbf{B}\mathbf{z}(t-\tau) = \mathbf{s}(t-\tau) \tag{2.64}$$

Se considera una versión ligeramente modificada de la matriz de covarianzas retardadas definida en (2.60) dada por

$$\bar{\mathbf{C}}_\tau^{\mathbf{z}} = \frac{1}{2}\left[\mathbf{C}_\tau^{\mathbf{z}} + \left(\mathbf{C}_\tau^{\mathbf{z}}\right)^T\right] \tag{2.65}$$

Debido a la linealidad de (2.63) y a la ortogonalidad de $\mathbf{B}$ se puede escribir

$$\begin{aligned} \bar{\mathbf{C}}_\tau^{\mathbf{z}} &= \frac{1}{2}\left\{E\left[\mathbf{z}(t)\mathbf{z}(t-\tau)^T\right] + E\left[\mathbf{z}(t-\tau)\mathbf{z}(t)^T\right]\right\} \\ &= \frac{1}{2}\mathbf{B}^T\left\{E\left[\mathbf{s}(t)\mathbf{s}(t-\tau)^T\right] + E\left[\mathbf{s}(t-\tau)\mathbf{s}(t)^T\right]\right\}\mathbf{B} \\ &= \frac{1}{2}\mathbf{B}^T\left[\mathbf{C}_\tau^{\mathbf{s}} + \left(\mathbf{C}_\tau^{\mathbf{s}}\right)^T\right]\mathbf{B} = \mathbf{B}^T\bar{\mathbf{C}}_\tau^{\mathbf{s}}\mathbf{B} \end{aligned} \tag{2.66}$$

A causa de la independencia de las componentes $s_i(t)$, la matriz de covarianzas retardadas $\mathbf{C}_\tau^{\mathbf{s}} = E\left[\mathbf{s}(t)\mathbf{s}(t-\tau)^T\right]$ es diagonal y, si se denota por $\mathbf{S}$, se observa que $\bar{\mathbf{C}}_\tau^{\mathbf{s}} = \frac{1}{2}\left[\mathbf{C}_\tau^{\mathbf{s}} + \left(\mathbf{C}_\tau^{\mathbf{s}}\right)^T\right] = \mathbf{S}$. De este modo se llega a

$$\bar{\mathbf{C}}_\tau^{\mathbf{z}} = \mathbf{B}^T\mathbf{S}\mathbf{B} \tag{2.67}$$

Esta ecuación muestra que las filas de la matriz ortogonal de separación $\mathbf{B}$ son los autovectores correspondientes a la diagonalización ortogonal de la matriz simétrica $\bar{\mathbf{C}}_\tau^{\mathbf{z}}$.

- **Descripción del algoritmo**

En consecuencia, según la derivación anterior, se dispone de un algoritmo sencillo y fácil de computar, llamado AMUSE, para estimar la matriz ortogonal de separación $\mathbf{B}$ para datos blanqueados que consta de los siguientes pasos:

1. Centrar los datos observados $\mathbf{x}$ para que su media sea cero.
2. Blanquear los datos centrados para obtener el vector aleatorio $\mathbf{z}$.
3. Elegir el número de componentes independientes m.
4. Calcular

$$\bar{\mathbf{C}}_{\tau}^{\mathbf{z}} = \frac{1}{2}\left[\mathbf{C}_{\tau}^{\mathbf{z}} + \left(\mathbf{C}_{\tau}^{\mathbf{z}}\right)^{T}\right]$$

   donde $\mathbf{C}_{\tau}^{\mathbf{z}} = E\left[\mathbf{z}(t)\mathbf{z}(t-\tau)^{T}\right]$ es la matriz de covarianzas retardadas para algún retardo τ.
5. Diagonalizar ortogonalmente la matriz $\bar{\mathbf{C}}_{\tau}^{\mathbf{z}}$, $\bar{\mathbf{C}}_{\tau}^{\mathbf{z}} = \mathbf{B}^{T}\mathbf{S}\mathbf{B}$.
6. Las filas de la matriz ortogonal de separación $\mathbf{B}$ son los autovectores de la diagonalización anterior.

El problema es, sin embargo, que el algoritmo solo funciona bien si todos los autovalores de la matriz $\bar{\mathbf{C}}_{\tau}^{\mathbf{z}}$ son distintos. Si algunos de los autovalores son iguales, entonces los correspondientes autovectores no están definidos de forma única, y las correspondientes componentes independientes no pueden ser estimadas. Estos autovalores están dados por $\mathrm{cov}\left(s_i(t), s_i(t-\tau)\right)$ y, de este modo, los autovalores son distintos si y solo si las covarianzas retardadas son diferentes para todas las componentes independientes.

El problema de que puedan existir autovalores iguales restringe considerablemente la aplicabilidad del algoritmo. Una solución puede ser buscar un determinado retardo τ de modo que los autovalores sean distintos aunque esto no siempre es posible. Si las señales $s_i(t)$ tienen idénticas autocovarianzas, entonces ningún valor de τ hace posible la estimación de la matriz $\mathbf{B}$.

Capítulo 3

Aplicación Práctica: Tasas de Empleo de las CC.AA.

3.1 Introducción

En los capítulos anteriores se han presentado y examinado con cierto detalle las técnicas de análisis de componentes principales, PCA, y de análisis de componentes independientes, ICA. Ahora bien, para fijar los contenidos de ambas técnicas así como conocer los detalles en el desarrollo de sus aplicaciones, es conveniente la realización de una aplicación práctica.

En este capítulo, sobre las tasas de empleo de las diferentes CC.AA., se desarrolla un ejemplo real con el fin de aplicar diferentes procedimientos estudiados, tanto de PCA como de ICA. Así, en primer lugar, se utilizará el análisis espectral singular, SSA, con el objetivo de filtrar las series observadas para eliminar las componentes oscilatorias de periodo igual o inferior al año, es decir, para sustraerles el ruido. A continuación se realizará una estimación de un PCA y de un ICA, empleando para este último tanto el algoritmo FastICA como AMUSE, sobre las series filtradas y estacionarias.

Una vez obtenidos los diferentes resultados, el objetivo es realizar comparaciones en términos geométricos y espectrales entre las componentes oscilatorias obtenidas con los tres procedimientos enunciados y, de esta forma, poder obtener conclusiones prácticas y derivar posibles desarrollos futuros.

3.2 Caracterización de los datos

Los datos objeto de estudio son 143 observaciones de las series trimestrales (desde el tercer trimestre de 1976 hasta el primero de 2012) correspondientes a las tasas de empleo para cada una de las diecisiete comunidades autónomas de España a excepción de las series correspondientes a las ciudades autónomas de Ceuta y Melilla porque para ellas no se dispone de datos desde el tercer trimestre de 1976.

Los datos se han obtenido de la Encuesta de Población Activa, EPA, que elabora y publica de forma trimestral el Instituto Nacional de Estadística, INE. En la EPA, se define la tasa de empleo como la ratio (en porcentaje) entre el número total de ocupados y la población total en edad de trabajar (mayores de 16 años según el INE) en un determinado territorio.

Los gráficos de las series de las tasas de empleo para cada una de las CC.AA., junto con el Total Nacional para poder realizar comparaciones, aparecen en el Anexo A desde la Figura A.1 a la Figura A.6. En estos gráficos se observa que en todas las series subyacen oscilaciones de diferentes periodos y, en mayor o menor medida, todas poseen un cierto nivel de ruido. En cuanto a la estacionalidad, a simple vista, esta sólo se aprecia claramente en la serie de Baleares que, además, modifica su perfil entre los años 2001 y 2002 haciéndolo más acusado.

En el mismo Anexo A, se presentan las funciones de autocorrelación de las series observadas, Figuras A.7 y A.8, para cada una de las CC.AA. y el Total Nacional. Estas evidencian claramente que las observaciones están autocorrelacionadas, no obstante, la suposición de independencia entre las observaciones no es necesaria para aplicar las técnicas de PCA e ICA desde un punto de vista descriptivo y no inferencial. Además, como era de esperar por la estacionalidad observada en la serie de Baleares, esta presenta autocorrelación estacional en sus datos.

Así mismo, en el Anexo A también se incluyen las funciones de autocorrelación parcial en las Figuras A.9 y A.10 para cada una de las CC.AA. y el Total Nacional. Todas ellas indican, aunque quizás con alguna duda para la serie de Baleares, que sus correspondientes series de las tasas de empleo tienen una posible raíz unitaria, es decir, no son estacionarias en media.

3.3 Metodología del análisis

Con el fin de obtener una estimación de las componentes oscilatorias o señales cíclicas comunes al conjunto de series temporales de las tasas de empleo de las diecisiete CC.AA. se aplicarán las técnicas de PCA e ICA para dos algoritmos (FastICA y AMUSE) sobre las series estacionarias.

A continuación se detalla la metodología empleada tanto para el PCA como para el ICA. Además, se dedica un apartado particular al SSA como método particular del PCA que se aplica a una única serie temporal para determinar sus diferentes componentes oscilatorias.

3.3.1 Análisis de Componentes Principales, PCA

El objetivo del PCA es reducir la dimensión de un conjunto de datos observados con la menor pérdida posible de información inicial. Las observaciones de un vector de series temporales están autocorrelacionadas de modo que no son independientes. Las inferencias sobre las componentes principales, PCs, se basan en la independencia y en la gaussianidad multivariante de las observaciones muestrales. No obstante, cuando el objetivo es descriptivo y no inferencial, la no independencia no es ninguna complicación que afecte de forma seria a dicho objetivo. Por el contrario, el supuesto de que las series temporales son estacionarias no debe ser omitido.

- **Estimación puntual de las PCs**

Sean $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T$ T observaciones centradas de un vector aleatorio p-dimensional, $\mathbf{x} = \left(x_1, x_2, \cdots, x_p\right)^T$, cuyos elementos son series temporales no necesariamente gaussianas pero estacionarias. A partir de un conjunto de datos agrupados en la matriz $\mathbf{X} = \left[\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\right]^T$ de dimensión $\left(T \times p\right)$, se desea encontrar $m < p$ combinaciones lineales de esas variables, las componentes principales, de forma que no están correlacionadas entre si, tienen máxima varianza muestral y se disponen en orden creciente de la varianza que explican.

Para este fin, a partir de la Proposición 1.7, se utiliza la descomposición espectral de la matriz de covarianzas muestral $\mathbf{C}$ dada por $\mathbf{C} = \frac{1}{T}\mathbf{X}^T\mathbf{X}$. La matriz $\mathbf{C}$ es simétrica y definida positiva con lo cual es diagonalizable ortogonalmente, $\mathbf{C} = \mathbf{U}\mathbf{D}\mathbf{U}^T$ donde $\mathbf{D} = \text{diag}\left(\lambda_1, \lambda_2, \cdots, \lambda_p\right)$ siendo $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ los autovalores de $\mathbf{C}$ y

$\mathbf{U}=\left[\mathbf{u}_1,\mathbf{u}_2,\cdots,\mathbf{u}_p\right]$ es una matriz ortogonal, $\mathbf{UU}^T=\mathbf{U}^T\mathbf{U}=\mathbf{I}$, siendo los autovectores $\mathbf{u}_1,\mathbf{u}_2,\cdots,\mathbf{u}_p$ ortonormales.

Las proyecciones de los datos centrados en el espacio generado por las columnas de la matriz $\mathbf{U}$ son las PCs, cuyas coordenadas se corresponden con las columnas de la matriz $\hat{\mathbf{Y}}$ dada por

$$\hat{\mathbf{Y}}=\mathbf{X}\mathbf{U} \tag{3.1}$$

Así, la estimación de la k-ésima PC se calcula como $\hat{y}_k=\mathbf{u}_k^T\mathbf{x}$ donde $\mathbf{x}$ es cualquier observación sobre las variables $x_1,x_2,\cdots,x_p$. En este contexto, por la Proposición 1.7, se tiene que

$$\begin{aligned}\operatorname{var}\left(\hat{y}_k\right)&=\lambda_k \qquad k=1,2,\cdots,p\\ \operatorname{cov}\left(\hat{y}_j,\hat{y}_k\right)&=0 \qquad j\neq k\end{aligned} \tag{3.2}$$

$$\textit{Varianza total muestral}=\operatorname{trace}\left(\mathbf{C}\right)=\sum_{k=1}^{p}\lambda_k$$

- **Elección del número de PCs a estimar**

Después de calcular las PCs por (3.1), se seleccionan las $m<p$ primeras PCs de forma que la reconstrucción de la matriz de observaciones $\mathbf{X}$ sea lo más fidedigna posible. La elección del número de PCs se realiza de forma empírica debido a que el análisis que se realiza es exploratorio y nunca confirmatorio. Esta elección empírica se realiza teniendo presente de forma conjunta los siguientes tres criterios.

El primero de estos criterios consiste en seleccionar un número suficiente de PCs de forma que la variación total acumulada que explican se encuentre entre el 70 y el 90%.

El segundo criterio se basa en el tamaño de las varianzas de las PCs. Así, se escogerían aquellas PCs cuya varianza sea superior al 70% de la varianza media, es decir, las PCs asociadas a los autovalores que sean mayores que el 70% del valor medio de los autovalores de la matriz de covarianzas muestral $\mathbf{C}$.

Finalmente, el tercer criterio consistirá en determinar el autovalor que forma el codo en el gráfico de sedimentación, λ_k frente a k, o, equivalentemente en el diagrama del log-autovalor, $\log\left(\lambda_k\right)$ frente a k, puesto que este último refleja la diferencia relativa entre los autovalores e indica a partir del autovalor donde la variabilidad acumulada no va a crecer de forma significativa.

3.3.2 Análisis Espectral Singular, SSA

La idea básica del SSA es realizar un PCA sobre un conjunto de retardos de una serie temporal x_t, es decir, el vector aleatorio p-dimensional está dado por $\mathbf{x}(t)=\left(x_t, x_{t+1}, \cdots, x_{t+p-1}\right)^T$ y, asumiendo que la serie x_t es estacionaria, el elemento (i, j) de su matriz de covarianzas solo depende de $|i-j|$. De este modo, los elementos (i, j) y $(i+1, j+1)$ son iguales con lo cual dicha matriz de covarianzas es una matriz Toeplitz simétrica.

La exigencia de estacionariedad se puede limitar a la estacionariedad en varianza ya que por el siguiente teorema, como extensión del Teorema 2.2 al PCA, la estacionariedad en media se puede relajar u omitir.

Teorema 3.1 *Si a las series observadas* $\mathbf{x}(t)$ *se les aplica un filtro lineal para obtener nuevas señales* $\mathbf{x}^*(t)$*, entonces el modelo PCA es válido para* $\mathbf{x}^*(t)$ *con la misma matriz de autovectores.*

Demostración. Sea $\mathbf{X}$ la matriz de dimensión $(T \times p)$ cuyas filas son los vectores de observaciones $\mathbf{x}(1), \mathbf{x}(2), \cdots, \mathbf{x}(T)$, es decir $\mathbf{X}=\left[\mathbf{x}(1)\,|\,\mathbf{x}(2)\,|\cdots|\,\mathbf{x}(T)\right]^T$, y sea $\hat{\mathbf{Y}}$ la matriz de dimensión $(T \times p)$ cuyas filas son los vectores de componentes principales $\hat{\mathbf{y}}(1), \hat{\mathbf{y}}(2), \cdots, \hat{\mathbf{y}}(T)$, es decir $\hat{\mathbf{Y}}=\left[\hat{\mathbf{y}}(1)\,|\,\hat{\mathbf{y}}(2)\,|\cdots|\,\hat{\mathbf{y}}(T)\right]^T$. Entonces, el modelo PCA dado por (3.1) se puede expresar, puesto que $\mathbf{U}$ es ortogonal, como

$$\mathbf{X} = \hat{\mathbf{Y}}\mathbf{U}^T$$

Ahora, sea $\mathbf{M}$ la matriz de dimensión $(T \times T)$ correspondiente a un filtro lineal.

El filtrado de $\mathbf{X}$ se realiza multiplicando dicha matriz por la izquierda por la matriz $\mathbf{M}$ para obtener el siguiente resultado

$$\mathbf{X}^* = \mathbf{M}\mathbf{X} = \mathbf{M}\hat{\mathbf{Y}}\mathbf{U}^T = \hat{\mathbf{Y}}^*\mathbf{U}^T$$

Las componentes principales son filtradas con el mismo filtro que se aplica a las series observadas. Con ello, se demuestra que el modelo PCA continua siendo válido. ■

El hecho de que la matriz de autovectores se mantenga sin cambios justifica que se puede realizar un SSA sobre cada una de las series originales (siempre que sean estacionarias en varianza) de las tasas de empleo sin tomar una primera diferencia con el objetivo de convertirlas en estacionarias en media.

En el SSA, las PCs son medias móviles de la serie temporal cuyas ponderaciones las proporcionan los autovectores de forma que, por ser funciones trigonométricas para las matrices Toeplitz, cada PC recoge un tipo de oscilación subyacente en la serie, desde tendencia a ruido.

Dada una muestra de una serie temporal $x_t,\ x_1, x_2, \cdots, x_T,$ se reorganiza para obtener una matriz $\mathbf{X}$ de dimensión $\left(T' \times p\right)$ cuya t-ésima fila es

$$\mathbf{x}(t) = \left(x_t, x_{t+1}, \cdots, x_{t+p-1}\right)^T$$

para $t = 1, 2, \cdots, T'$ donde $T' = T - p + 1$. La matriz muestral de covarianzas $\mathbf{C}$ verificará que $c_{ij} = c_{i+1,j+1}$ si la serie x_t es estacionaria en varianza debido a que la covarianza entre las variables i-ésima y j-ésima depende solo de $|i - j|$.

La proyección de la serie temporal x_t sobre cada autovector $\mathbf{u}_k$ de la matriz $\mathbf{C}$ es la k-ésima PC $\hat{y}_k = \mathbf{u}_k^T \mathbf{x}(t)$ pero con $T' < T$ observaciones. Sin embargo, cada componente oscilatoria de la serie temporal, asociada con un determinado autovector, se puede reconstruir mediante la fórmula (1.63) para lograr estar en fase con la misma serie temporal. Además, puesto que no existe pérdida de información en la reconstrucción, $x_t = \sum_{k=1}^{p} R_{tk}$ siendo R_{tk} las componentes reconstruidas, se pueden realizar reconstrucciones parciales con una combinación de los autovectores asociados a un conjunto de determinadas oscilaciones $K = \{k_1, k_2, \cdots, k_s\}$ mediante

$$R_{K,t} = \sum_{k \in K} R_{tk} \tag{3.3}$$

En consecuencia, el SSA se muestra como un método eficaz para sustraer el ruido (componentes oscilatorias de periodo igual o inferior al año) subyacente en una serie temporal al reconstruir esa serie con las componentes asociadas al resto de frecuencias.

3.3.3 Análisis de Componentes Independientes, ICA

El objetivo principal de este método es encontrar componentes no observables que relacionan variables aleatorias o señales. El modelo ICA admite que las variables observables, los datos, son mezclas lineales de variables latentes desconocidas, es decir, que no pueden observarse directamente y se denominan componentes independientes, ICs. El ICA está relacionado con el PCA pero se distingue de él en que asume que las componentes son estadísticamente independientes y no tienen una distribución gaussiana multivariante.

- **Caracterización del ICA**

El modelo general sin ruido está formado por el vector aleatorio $\mathbf{x} = \left(x_1, x_2, \cdots, x_p\right)^T$ cuyos elementos, las variables observables x_i, son una mezcla lineal de los elementos del vector $\mathbf{s} = \left(s_1, s_2, \cdots, s_n\right)^T$ que son las ICs s_i. En forma matricial se expresa como

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{3.4}$$

siendo $\mathbf{A}$ la matriz de mezcla cuyos elementos a_{ij} son parámetros desconocidos que se deben estimar.

Para asegurar la estimación del modelo básico de ICA definido en (3.4) es necesario realizar las siguientes suposiciones:

1) Las componentes no observadas s_i *son estadísticamente independientes.*

2) Las componentes independientes s_i *tienen media cero y varianza unidad.*

3) Las componentes independientes s_i *deben tener distribuciones no-gaussianas.*

4) La matriz de mezcla desconocida $\mathbf{A}$ *es cuadrada,* $p = n$.

5) La matriz de mezcla $\mathbf{A}$ *es de rango completo por columnas.*

6) Si los elementos de los vectores aleatorios $\mathbf{x}$ *y* $\mathbf{s}$ *son series temporales, estas deben ser estacionarias en sentido estricto.*

La estimación de la matriz $\mathbf{A}$ se realiza a partir de los datos observados $\mathbf{x}$ y de los supuestos realizados mediante la optimización de una función de contraste. Una vez estimada la matriz $\mathbf{A}$, y admitiendo que es no singular por los supuestos anteriores, se puede determinar su inversa $\mathbf{W} = \mathbf{A}^{-1}$ lo cual permite obtener las ICs mediante

$$\mathbf{s} = \mathbf{W}\mathbf{x} \tag{3.5}$$

- **Maximización de la no-gaussianidad**

Un procedimiento sencillo para la estimación de un modelo ICA es la maximización de la no-gaussianidad. Como medida de la no-gaussianidad de una variable aleatoria se puede utilizar la curtosis que corresponde al cumulante de cuarto orden. Al contrario de la mayoría de las variables aleatorias, una variable aleatoria gaussiana presenta una curtosis nula. Debido a que las ICs se pueden encontrar tanto en un sentido como en el contrario de las direcciones en las que los datos optimizan la curtosis, la no-gaussianidad puede ser medida por la optimización del valor absoluto de la curtosis. No

obstante, la curtosis no es una medida robusta de la no-gaussianidad porque es muy sensible a valores atípicos.

Con el fin de salvar los inconvenientes de la curtosis, surge como medida alternativa de la no-gaussianidad la sintropía. La sintropía J de una variable aleatoria y se define como

$$J(y) = H(y_{gauss}) - H(y) \tag{3.6}$$

donde H es la entropía diferencial e y_{gauss} es una variable aleatoria gaussiana con la misma varianza que la variable aleatoria y. Sin embargo, el cálculo de la sintropía es difícil porque requiere una estimación de la función de densidad aunque aproximaciones de la sintropía resultan muy útiles y se pueden usar para obtener métodos eficientes de estimación de modelos ICA.

La aproximación clásica de la sintropía, para una variable aleatoria de media cero y varianza unidad, está dada por

$$J(y) \approx \frac{1}{12}\mathrm{E}\left[y^3\right]^2 + \frac{1}{48}\mathrm{kurt}(y)^2 \tag{3.7}$$

No obstante, estas aproximaciones, al igual que la curtosis, no son robustas. Por ello, se desarrollan otras aproximaciones entre las que destaca la especificada por

$$J(y) \propto \left\{\mathrm{E}\left[G(y)\right] - \mathrm{E}\left[G(z)\right]\right\}^2 \tag{3.8}$$

donde las variables aleatorias z e y tienen media cero y varianza unidad siendo z gaussiana y G es una función no cuadrática. Para el caso $G(y) = y^4$ se obtiene una aproximación basada en la curtosis. La función G se debe elegir de forma que no crezca demasiado rápida para obtener estimadores más robustos que la sintropía. Las siguientes alternativas para G

$$\begin{aligned} G_1(y) &= \frac{1}{\alpha}\log\left(\cosh(\alpha y)\right) \\ G_2(y) &= -\exp\left(\frac{-y^2}{2}\right) \end{aligned} \tag{3.9}$$

donde $1 \le \alpha \le 2$ es una constante, han demostrado ser bastante útiles para aproximar la sintropía y en este trabajo se ha optado por utilizar G_1.

- **Preparación de los datos para ICA**

a) Estacionariedad

Si los elementos del vector aleatorio $\mathbf{x}(t)$ son series temporales, estas deben ser estacionarias en sentido estricto según los supuestos para la estimación del modelo ICA. Si las series de $\mathbf{x}(t)$ poseen una raíz unitaria, es decir, no son estacionarias en media, es necesario diferenciarlas para que verifiquen los supuestos requeridos. En tal caso, el modelo ICA $\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t)$ continua siendo válido, por los Teoremas 2.2 y 2.3, para las series diferenciadas con la misma matriz de mezcla, $\tilde{\mathbf{x}}(t) = \mathbf{A}\tilde{\mathbf{s}}(t)$, donde $\tilde{\mathbf{x}}(t) = \Delta\mathbf{x}(t)$ y $\tilde{\mathbf{s}}(t) = \Delta\mathbf{s}(t)$.

b) Centrado

En los supuestos para la existencia de un modelo ICA se asume que las ICs tienen media cero y, por tanto, también las variables mezcla observadas según se deduce al tomar esperanzas en la ecuación (3.4). En el caso de que las variables observadas no tengan media cero, se centran los datos observados restándoles su vector de medias.

c) Blanqueo

El blanqueo de variables consiste en transformar linealmente el vector p-dimensional $\mathbf{x}$ de las observaciones (previamente centradas), multiplicándolo por una matriz $\mathbf{Q}$, de forma que se obtenga otro vector $\mathbf{z} = \mathbf{Q}\mathbf{x}$, cuyas componentes están incorrelacionadas y tienen varianza unidad, es decir, su matriz de covarianzas es la matriz identidad.

A partir de la diagonalización ortogonal de la matriz de covarianzas, del vector aleatorio de las variables observadas, dada por

$$\mathrm{E}\left[\mathbf{x}\mathbf{x}^T\right] = \mathbf{C} = \mathbf{U}\mathbf{D}\mathbf{U}^T \tag{3.10}$$

donde $\mathbf{U}$ es una matriz ortogonal de autovectores de $\mathbf{C}$ y $\mathbf{D} = \mathrm{diag}\left(\lambda_1, \cdots, \lambda_p\right)$ es la matriz diagonal con los autovalores de $\mathbf{C}$, se puede obtener la matriz blanqueadora

$$\mathbf{Q} = \mathbf{D}^{-1/2}\mathbf{U}^T \tag{3.11}$$

siendo $\mathbf{D}^{-1/2} = \mathrm{diag}\left(\lambda_1^{-1/2}, \cdots, \lambda_p^{-1/2}\right)$. De esta forma, los elementos del vector blanco

$$\mathbf{z} = \mathbf{Q}\mathbf{x} = \mathbf{D}^{-1/2}\mathbf{U}^T\mathbf{x} \tag{3.12}$$

se corresponden con las PCs estandarizadas para las observaciones mezcla de $\mathbf{x}$.

La utilidad del blanqueo reside en que a partir de las ecuaciones (3.4) y (3.12) con

$$\mathbf{z} = \mathbf{Q}\mathbf{x} = \mathbf{Q}\mathbf{A}\mathbf{s} = \mathbf{B}^T\mathbf{s} \tag{3.13}$$

se obtiene una nueva matriz de mezcla $\mathbf{B}^T = \mathbf{QA}$ que es ortogonal

$$\mathbf{I} = \mathrm{E}\left[\mathbf{z}\mathbf{z}^T\right] = \mathbf{B}^T\,\mathrm{E}\left[\mathbf{s}\mathbf{s}^T\right]\mathbf{B} = \mathbf{B}^T\mathbf{I}\mathbf{B} = \mathbf{B}^T\mathbf{B}$$

y, por tanto, la nueva matriz de separación es ortogonal y está dada por $\mathbf{B}$.

Por otro lado, no hay que olvidar que cualquier transformación ortogonal $\mathbf{V}$ de $\mathbf{z}$, $\mathbf{y} = \mathbf{Vz}$, también proporciona un vector blanco según se demostró en (2.40). Por ello, el blanqueo proporciona las ICs salvo una transformación ortogonal.

d) Número de ICs a estimar

En ocasiones, el número de ICs a estimar, m, será menor que el número de variables observadas, $m < p$. No obstante, esto no representa ningún problema para que se verifiquen los supuestos 3) y 4) de entre los que aseguran la estimación del modelo ICA. En cualquier caso, una forma de conseguirlo es reduciendo la dimensión mediante el PCA. Con este fin, se construye la matriz blanqueadora $\mathbf{Q}$ con los m mayores autovalores de la diagonalización de la matriz $\mathbf{C} = \mathrm{E}\left[\mathbf{x}\mathbf{x}^T\right]$ dada en (3.10) obteniendo una matriz de dimensión $(m \times p)$ dada por

$$\mathbf{Q}_{m\times p} = \mathbf{D}^{-1/2}_{m\times m}\mathbf{U}^T_{m\times p} \tag{3.14}$$

donde $\mathbf{D}^{-1/2}_{m\times m} = \mathrm{diag}\left(\lambda_1^{-1/2}, \cdots, \lambda_m^{-1/2}\right)$ siendo $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$ los m mayores autovalores de la matriz $\mathbf{C}$ y $\mathbf{U}_{p\times m} = \left[\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_m\right]$ contiene los autovectores ortonormales asociados a esos primeros m autovalores.

En esta situación, tanto el vector blanco $\mathbf{z}$ como el vector $\mathbf{s}$ con las ICs tienen m variables y las ecuaciones (3.4) y (3.5), respectivamente, quedan como

$$\mathbf{x}_{p\times 1} = \mathbf{A}_{p\times m}\mathbf{s}_{m\times 1} \tag{3.15}$$

$$\mathbf{s}_{m\times 1} = \mathbf{W}_{m\times p}\mathbf{x}_{p\times 1} \tag{3.16}$$

Por tanto, las dimensiones de las matrices de mezcla $\mathbf{A}$ y de separación $\mathbf{W}$ son, respectivamente, $(p \times m)$ y $(m \times p)$. Así, la ecuación (3.13) se rescribe como

$$\mathbf{z}_{m\times 1} = \mathbf{Q}_{m\times p}\mathbf{x}_{p\times 1} = \mathbf{Q}_{m\times p}\mathbf{A}_{p\times m}\mathbf{s}_{m\times 1} = \mathbf{B}^T_{m\times m}\mathbf{s}_{m\times 1} \tag{3.17}$$

de forma que, como era de esperar, la matriz ortogonal de mezcla dada por

$$\mathbf{B}^T_{m\times m} = \mathbf{Q}_{m\times p}\mathbf{A}_{p\times m} \tag{3.18}$$

es de dimensión $(m \times m)$ y deberá ser estimada por alguno de los algoritmos que posteriormente se detallan.

A partir de (3.18), puesto que $\mathbf{B}_{m\times m}^{T}$ es ortogonal, se puede escribir

$$\mathbf{I}_{m\times m}=\mathbf{B}_{m\times m}\mathbf{B}_{m\times m}^{T}=\mathbf{B}_{m\times m}\mathbf{Q}_{m\times p}\mathbf{A}_{p\times m}=\mathbf{B}_{m\times m}\mathbf{D}_{m\times m}^{-1/2}\mathbf{U}_{m\times p}^{T}\mathbf{A}_{p\times m}$$

con lo cual, para que se verifique dicha igualdad, la matriz de mezcla $\mathbf{A}_{p\times m}$ para obtener los datos observados $\mathbf{x}_{p\times 1}$ debe ser

$$\mathbf{A}_{p\times m}=\mathbf{U}_{p\times m}\mathbf{D}_{m\times m}^{1/2}\mathbf{B}_{m\times m}^{T} \tag{3.19}$$

Por otro lado, a partir de la identidad establecida en (3.17), $\mathbf{Q}_{m\times p}\mathbf{x}_{p\times 1}=\mathbf{B}_{m\times m}^{T}\mathbf{s}_{m\times 1}$, y haciendo uso de la ecuación (3.16) se llega a $\mathbf{Q}_{m\times p}\mathbf{x}_{p\times 1}=\mathbf{B}_{m\times m}^{T}\mathbf{W}_{m\times p}\mathbf{x}_{p\times 1}$, con lo cual $\mathbf{Q}_{m\times p}=\mathbf{B}_{m\times m}^{T}\mathbf{W}_{m\times p}$ y, debido a la ortogonalidad de $\mathbf{B}_{m\times m}^{T}$, se tiene que la matriz de separación $\mathbf{W}_{m\times p}$, para obtener $\mathbf{s}_{m\times 1}$ a partir de $\mathbf{x}_{p\times 1}$, está dada por

$$\mathbf{W}_{m\times p}=\mathbf{B}_{m\times m}\mathbf{Q}_{m\times p}=\mathbf{B}_{m\times m}\mathbf{D}_{m\times m}^{-1/2}\mathbf{U}_{m\times p}^{T} \tag{3.20}$$

Una vez que se ha detallado el procedimiento para calcular las matrices de mezcla y de separación, $\mathbf{A}$ y $\mathbf{W}$ respectivamente, cuando se estiman menos ICs que variables observadas, resta establecer alguna regla para decidir el número m de ICs a estimar.

Si $\mathbf{x}$ es el vector p-dimensional con las variables observadas centradas para el que se estima un modelo ICA completo $\mathbf{x}=\mathbf{A}\mathbf{s}$, es decir, el vector $\mathbf{s}$ con las ICs es p-dimensional y la matriz de mezcla $\mathbf{A}$ tiene dimensión $(p\times p)$ y rango p, entonces la matriz de covarianzas de $\mathbf{x}$ está dada por

$$\mathbf{C}=\mathrm{E}\left[\mathbf{x}\mathbf{x}^{T}\right]=\mathbf{A}\,\mathrm{E}\left[\mathbf{s}\mathbf{s}^{T}\right]\mathbf{A}^{T}=\mathbf{A}\mathbf{I}\mathbf{A}^{T}=\mathbf{A}\mathbf{A}^{T}$$

Según el teorema de Eckart-Young, la distancia de $\mathbf{C}$ según la norma de Frobenius a la matriz más próxima de rango k (la versión truncada de la SVD con k términos $\mathbf{C}^{(k)}$) está dada por

$$\delta_k=\min_{rank(\mathbf{S})=k}\left\|\mathbf{C}-\mathbf{S}\right\|_F=\left\|\mathbf{C}-\mathbf{C}^{(k)}\right\|_F=\sqrt{\sum_{i=k+1}^{p}\sigma_i^2}=\sqrt{\sum_{i=k+1}^{p}\lambda_i^2} \tag{3.21}$$

donde σ_i es el i-ésimo valor singular de la matriz $\mathbf{C}$ verificando que $\sigma_i=\lambda_i$, siendo λ_i su i-ésimo autovalor, puesto que $\mathbf{C}$ es una matriz normal y definida positiva. De esta forma, si se estiman $m<p$ ICs, la matriz de mezcla $\mathbf{A}_{p\times m}$ estimada por (3.19), de dimensión $(p\times m)$, tiene rango m y $\mathbf{A}_{p\times m}\mathbf{A}_{m\times p}^{T}=\mathbf{U}_{p\times m}\mathbf{D}_{m\times m}\mathbf{U}_{m\times p}^{T}=\mathbf{C}^{(m)}$ es la versión truncada de la SVD de $\mathbf{C}$ con m términos cuya distancia a $\mathbf{C}$ es δ_m. En consecuencia, la representación gráfica de los pares (k,δ_k) permite determinar el número m de ICs a estimar que será el valor k^{*} a partir del cual la pendiente de esta curva, que representa la falta de aproximación, disminuye de forma significativa.

El gráfico de los pares (k, δ_k) para ICA y el de los pares (k, λ_k) para PCA, gráfico de sedimentación, tratan de establecer de forma empírica el número de componentes necesarias con el fin de reducir el error entre los valores observados y los reconstruidos a partir de dichas componentes. Sin embargo, lo realizan desde diferentes ópticas. El PCA lo intenta aumentando la variación total explicada y el ICA mejorando la aproximación de la matriz de covarianzas de los datos observados.

- **El algoritmo FastICA**

FastICA es un algoritmo de punto fijo basado en la maximización de la no-gaussianidad medida con una aproximación de la sintropía. Es un algoritmo más eficiente y rápido que los basados en el gradiente. La versión utilizada de este algoritmo para varias componentes usa el método clásico de ortogonalización simétrica, ha sido programado en MATLAB y consta de los siguientes pasos:

1. Centrar los datos observados $\mathbf{x}$ para que su media sea cero.
2. Blanquear los datos centrados para obtener el vector aleatorio $\mathbf{z}$.
3. Elegir el número de componentes independientes m a estimar.
4. Elegir, de forma aleatoria, valores iniciales para los $\mathbf{b}_i$, con $i = 1, 2, \cdots, m$, cada uno de norma unidad. Ortogonalizar la matriz $\mathbf{B} = \left[\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_m\right]^T$ según el posterior paso 6.
5. Para cada $i = 1, 2, \cdots, m$, hacer

$$\mathbf{b}_i \leftarrow \mathrm{E}\left[\mathbf{z}\, g\left(\mathbf{b}_i^T \mathbf{z}\right)\right] - \mathrm{E}\left[g'\left(\mathbf{b}_i^T \mathbf{z}\right)\right]\mathbf{b}_i$$

   donde g es la primera derivada de alguna de las funciones definidas en (3.9).

6. Realizar una ortogonalización simétrica de la matriz $\mathbf{B}$

$$\mathbf{B} \leftarrow \left(\mathbf{B}\mathbf{B}^T\right)^{-1/2} \mathbf{B}$$

7. Si la matriz $\mathbf{B}$ no converge, volver al paso 5. En caso contrario estimar la matriz de mezcla $\mathbf{A}_{p \times m}$ y la matriz de separación $\mathbf{W}_{m \times p}$ para los datos observados según (3.19) y (3.20) respectivamente.

La convergencia significa que el valor antiguo y el valor actual de los vectores $\mathbf{b}_i$, las filas de la matriz $\mathbf{B}$, deben indicar las mismas direcciones, es decir, el valor absoluto de su producto escalar debe ser (casi) igual a 1. No es necesario que los vectores converjan a un único punto puesto que $\mathbf{b}_i$ y $-\mathbf{b}_i$ definen la misma dirección.

- **El algoritmo AMUSE**

El algoritmo AMUSE es una alternativa para aquellos casos en los que las componentes independientes poseen una dependencia temporal. Este algoritmo intenta encontrar una matriz que anule tanto las covarianzas contemporáneas como las retardadas. La versión utilizada de este algoritmo se ha programado en MATLAB y se concreta en los siguientes pasos:

1. Centrar los datos observados $\mathbf{x}$ para que su media sea cero.
2. Blanquear los datos centrados para obtener el vector aleatorio $\mathbf{z}$.
3. Elegir el número de componentes independientes m a estimar.
4. Calcular

$$\bar{\mathbf{C}}_{\tau}^{\mathbf{z}} = \frac{1}{2}\left[\mathbf{C}_{\tau}^{\mathbf{z}} + \left(\mathbf{C}_{\tau}^{\mathbf{z}}\right)^{T}\right]$$

   donde $\mathbf{C}_{\tau}^{\mathbf{z}} = E\left[\mathbf{z}(t)\mathbf{z}(t-\tau)^{T}\right]$ es la matriz de covarianzas retardadas para algún retardo τ que, por simplicidad, en este trabajo se ha elegido igual a 1.

5. Diagonalizar ortogonalmente la matriz $\bar{\mathbf{C}}_{\tau}^{\mathbf{z}}$, $\bar{\mathbf{C}}_{\tau}^{\mathbf{z}} = \mathbf{B}^{T}\mathbf{S}\mathbf{B}$.
6. Las filas de la matriz ortogonal de separación $\mathbf{B}$ son los autovectores de la diagonalización anterior. Estimar la matriz de mezcla $\mathbf{A}_{p\times m}$ y la matriz de separación $\mathbf{W}_{m\times p}$ para los datos observados según (3.19) y (3.20) respectivamente.

3.4 Resultados del análisis

En esta sección se presentan los cálculos y resultados obtenidos con la metodología que se describió en la sección anterior y que se aplica a las series de las tasas de empleo de las CC.AA. Se comienza con el filtrado de dichas series mediante el SSA para eliminar el ruido presente en las mismas junto con el estudio de la existencia de raíces unitarias. Posteriormente, sobre las series filtradas y estacionarias, se realiza un PCA descriptivo y se finaliza con un ICA, sobre las mismas series, utilizando tanto el algoritmo FastICA como el AMUSE, más adaptado a datos con dependencia temporal.

3.4.1 Series filtradas de ruido y estacionarias

- **SSA, series filtradas de ruido**

El objetivo en este apartado es eliminar el ruido en las series observadas por dos motivos. El primero es adaptarse a las condiciones para la estimación del modelo ICA descrito en la sección anterior que no considera ningún término de ruido. El segundo es obtener, con PCA e ICA, las componentes oscilatorias, de periodo superior al año, comunes a las series que sean lo más nítidas posibles y por ello, además, se considera ruido todas las componentes oscilatorias de periodo igual o inferior al año (estacionalidad e irregularidad) que subyacen en las series.

Para realizar el SSA a cada una de las series originales, estas no se van a transformar por el Teorema 3.1, aunque fuera necesario, con el fin de que sean estacionarias en media. No obstante, las series deben verificar que son estacionarias en varianza puesto que la matriz de covarianzas del vector $\mathbf{x}(t)=\left(x_t, x_{t+1}, \cdots, x_{t+p}\right)^T$ es Toeplitz simétrica, es decir, la covarianza entre las variables i-ésima y j-ésima depende solo de $|i-j|$.

El estudio de la estacionariedad en varianza de las series se realiza con un test rango-media sobre el coeficiente de correlación. Para efectuar el test se divide cada serie en grupos de cuatro observaciones, por ser trimestrales, y para cada grupo se calcula una medida robusta de localización como es la mediana y una medida robusta de dispersión como es la desviación absoluta mediana. Los resultados del test, programado en MATLAB (Anexo F), se muestran en la Tabla 3.1 y en ella se aprecia que para todas las comunidades, excepto para Baleares, el intervalo de confianza al 95% para el coeficiente de correlación incluye el valor cero con lo cual no se puede rechazar que las series son estacionarias en varianza con un nivel de significación del 5%.

Comunidad Autónoma	Coeficiente correlación	p-valor	Intervalo de confianza	
			Límite inferior	Límite superior
Andalucía	0,0484	0,7825	-0,2895	0,3756
Aragón	0,2126	0,2201	-0,1298	0,5098
Asturias	0,3215	0,0596	-0,0131	0,5914
Baleares	0,7032	0,0000	0,4832	0,8397
Canarias	0,1060	0,5445	-0,2356	0,4243
Cantabria	-0,1479	0,3964	-0,4586	0,1949
Castilla y León	0,2721	0,1139	-0,0673	0,5550
Castilla - La Mancha	-0,0303	0,8630	-0,3599	0,3061
Cataluña	0,1417	0,4168	-0,2010	0,4535
C. Valenciana	0,0994	0,5698	-0,2418	0,4188
Extremadura	0,0045	0,9795	-0,3292	0,3373
Galicia	0,2020	0,2446	-0,1407	0,5015
Madrid	0,2059	0,2354	-0,1368	0,5045
Murcia	0,0119	0,9458	-0,3226	0,3438
Navarra	-0,2236	0,1966	-0,5183	0,1184
País Vasco	0,1755	0,3134	-0,1676	0,4806
La Rioja	0,2565	0,1369	-0,0839	0,5433
Total Nacional	0,1489	0,3931	-0,1939	0,4594

Tabla 3.1 Test rango-media para el coeficiente de correlación de las series originales

Un trato especial merece la serie de la tasa de empleo de Baleares. En el gráfico de dispersión rango-media de la Figura 3.1 se observan dos nubes de puntos, que aunque son paralelas al eje de abscisas, una se sitúa encima de la otra y parcialmente superpuesta lo que origina que el coeficiente de correlación sea significativo. Sin embargo, el gráfico del rango frente al tiempo de la Figura 3.2 revela que este es invariante al tiempo antes y después del año 2002 que se puede considerar como un año de transición hacia un cambio de nivel. Los valores más elevados en el rango a partir del 2002 se deben a que se acentúa el perfil estacional debido al tercer trimestre pudiendo estar motivado por la elección de Baleares como destino turístico veraniego

más seguro en el área del Mediterráneo a raíz de los hechos de septiembre de 2001. Por todo ello, se considera que la serie de Baleares también es estacionaria en varianza puesto que, como se ha probado, el rango no depende de la media.

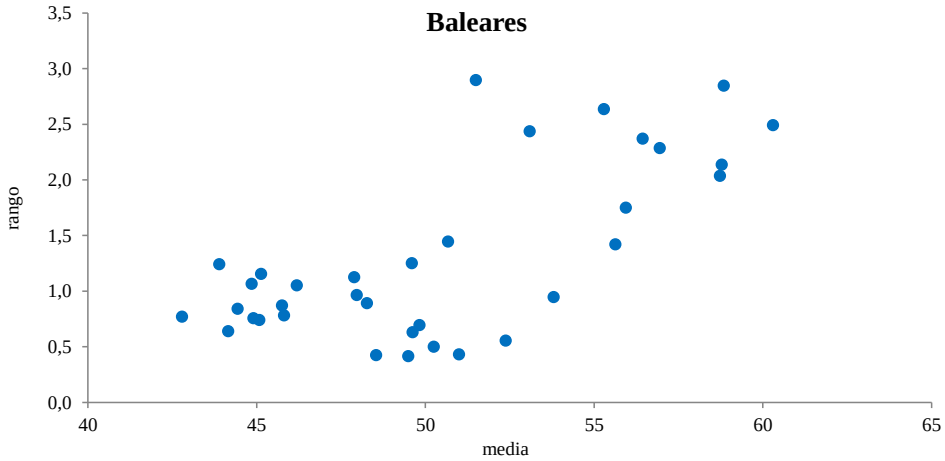

Figura 3.1 Gráfico de dispersión de los pares rango-medio de Baleares

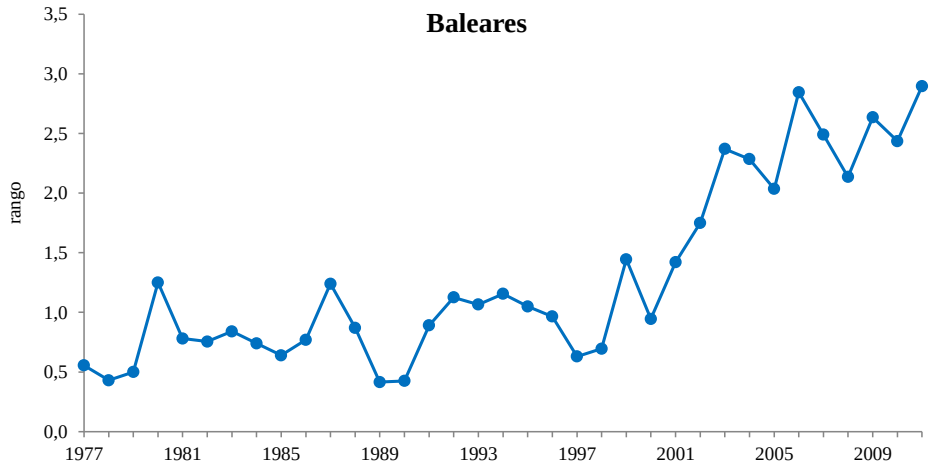

Figura 3.2 Evolución temporal del rango en la serie de Baleares

La estacionariedad en varianza se verifica para todas las series. Ello permite realizar un SSA, programado en MATLAB (Anexo F), para cada una de ellas con el fin de filtrarlas de ruido. En el SSA, la matriz de covarianzas es Toeplitz simétrica con lo cual sus autovectores son funciones trigonométricas y se corresponden con diferentes patrones oscilatorios. En este trabajo, para recoger los más diversos patrones oscilatorios se han tomado $p = 33$ retardos. Para cada serie, se tabulan sus autovalores

correspondientes y el porcentaje del total de varianza que explican y los coeficientes de los autovectores se grafican con el fin de determinar los que corresponden a ruido (oscilaciones de periodo igual o inferior al año), todo ello se recoge en el Anexo B. Con el fin de reconstruir las series con la mayor precisión pero con el mínimo ruido posible, para cada serie se han seleccionado las PCs que corresponden a autovectores cuyos patrones oscilatorios no son ruido, los autovalores correspondientes son mayores que la unidad y explican al menos un 0,1% de la variación total. En la Tabla 3.2 se presenta un resumen de los SSA realizados donde aparecen los autovectores seleccionados y el porcentaje de varianza acumulada que explican para cada comunidad. Excepto para Baleares con un 88,78%, la varianza explicada en cada SSA es superior al 95% llegando en varias comunidades a superar el 99%. El menor porcentaje de Baleares se explica por su significada estacionalidad según se analiza en la Figura B.4 y en la Tabla B.2.

Comunidad Autónoma	Autovectores seleccionados	% Varianza explicada
Andalucía	1 - 2 - 3 - 4 - 5	99,56
Aragón	1 - 2 - 3 - 4 - 5 - 6	99,21
Asturias	1 - 2 - 3 - 4 - 5 - 8	97,19
Baleares	1 - 2 - 5 - 6 - 7 - 8	88,78
Canarias	1 - 2 - 3 - 4 - 5 - 6	99,17
Cantabria	1 - 2 - 3 - 4 - 5 - 6 - 7 - 8	98,69
Castilla y León	1 - 2 - 3 - 4 - 7 - 8	98,11
Castilla - La Mancha	1 - 2 - 3 - 7	98,11
Cataluña	1 - 2 - 3 - 4	99,48
C. Valenciana	1 - 2 - 3 - 4 - 5 - 6	99,60
Extremadura	1 - 2 - 3 - 4 - 8 - 9 - 10	95,78
Galicia	1 - 2 - 3 - 4 - 5 - 8 - 9 - 10	98,10
Madrid	1 - 2 - 3 - 4	99,73
Murcia	1 - 2 - 3 - 4 - 5 - 6 - 7 - 8	99,14
Navarra	1 - 2 - 3 - 4	99,03
País Vasco	1 - 2 - 3 - 4 - 5	99,43
La Rioja	1 - 2 - 3 - 4 - 5 - 6 - 7	98,77
Total Nacional	1 - 2 - 3 - 4	99,55

Tabla 3.2 Resumen del SSA realizado para la series de las tasas de empleo.

Una vez que se han seleccionado los autovectores bajo las condiciones especificadas, se estiman las correspondientes PCs teniendo en cuenta que tienen menos observaciones que la serie original puesto que el vector de observaciones está formado por $p = 33$ retardos de la serie original. Para conseguir que estas PCs estén en fase con la serie original y tengan el mismo número de observaciones se reconstruyen con la fórmula dada en (1.63). A partir de la combinación de las PCs seleccionadas que se han reconstruido, mediante la fórmula de (3.3), se recomponen parcialmente las series originales de forma que se han filtrado de ruido, se han sustraído aquellas componentes oscilatorias que se desea que no estén presentes.

Los gráficos de las series filtradas de las tasas de empleo para cada una de las CC.AA., junto con el Total Nacional para poder efectuar comparaciones, aparecen en el Anexo C desde la Figura C.1 a la Figura C.6. En estos gráficos se han incluido las series originales y se observa como las series filtradas son mucho más suaves que ellas. Esta suavidad de las series filtradas permite distinguir más claramente oscilaciones de diferentes periodos en todas ellas y, en mayor o menor medida, se puede advertir algunas oscilaciones subyacentes comunes al conjunto de las series filtradas.

- **Test ADF, series estacionarias**

Las funciones de autocorrelación de las series filtradas, Figuras C.7 y C.8 del Anexo C, evidencian de nuevo que las observaciones están autocorrelacionadas, pero para realizar un PCA y un ICA, desde un punto de vista descriptivo, el supuesto de independencia entre las observaciones no es necesario según se explicó con anterioridad.

Las series originales son estacionarias en varianza y no existe evidencia para suponer lo contrario en las series filtradas. Además, interesa trabajar con series estacionarias en media, es decir, que no posean una raíz unitaria, tanto por los supuestos exigidos para realizar el ICA como para conseguir estimar componentes subyacentes independientes para lo cual puede interferir la presencia de una tendencia.

Las funciones de autocorrelación parcial de las series filtradas, Figuras C.9 y C.10 del Anexo C, ponen de manifiesto la posible existencia de una raíz unitaria en cada una de las series puesto que el primer coeficiente de autocorrelación parcial es casi igual a uno para todas las series filtradas. No obstante, ello se va a comprobar con la realización del test aumentado de Dickey-Fuller, ADF, para el cual se estima el modelo

$$x_t = \delta + \alpha t + \rho x_{t-1} + \sum_{j=1}^{q} \beta_j \Delta x_{t-j}$$

cuya hipótesis nula es la existencia de una raíz unitaria, $\rho = 1$, en la serie analizada x_t.

Los resultados del test ADF (función *adftest()* de MATLAB) para las series filtradas se muestran en la Tabla 3.3, donde el número de retardos que aparece es el mayor valor de q para el que β_q es significativo. Según esos resultados, ninguna serie presenta una tendencia determinista, salvo Asturias y Galicia que admiten una constante. Ahora bien, lo más relevante es que para ninguna serie se puede rechazar la presencia de una raíz unitaria, ratificando a las funciones de autocorrelación parcial, puesto que todos los valores del test son mayores que el correspondiente cuantil del 90%. Así, para que las series filtradas sean estacionarias en media, se deben diferenciar, con lo cual las componentes subyacentes de las diferencias son independientes por el Teorema 2.3.

Comunidad Autónoma	Test ADF			
	Deriva	Retardos	Valor	Cuantil 90%
Andalucía	No	4	-0,85	-1,62
Aragón	No	7	-1,29	-1,62
Asturias	Si	3	-2,11	-2,58
Baleares	No	7	-0,64	-1,62
Canarias	No	3	-0,80	-1,62
Cantabria	No	3	-1,10	-1,62
Castilla y León	No	6	-1,36	-1,62
Castilla - La Mancha	No	6	-1,31	-1,62
Cataluña	No	2	-0,73	-1,62
C. Valenciana	No	3	-1,05	-1,62
Extremadura	No	4	-1,22	-1,62
Galicia	Si	5	-2,24	-2,58
Madrid	No	2	0,62	-1,62
Murcia	No	4	-0,42	-1,62
Navarra	No	7	-1,26	-1,62
País Vasco	No	3	-1,06	-1,62
La Rioja	No	4	-0,76	-1,62
Total Nacional	No	2	-0,71	-1,62

Tabla 3.3 Resultados del test ADF para las series filtradas de ruido.

3.4.2 Análisis de Componentes Principales, PCA

El PCA se realiza sobre las series filtradas de ruido y diferenciadas, estacionarias en sentido estricto, con el fin de obtener unas componentes comunes, no correlacionadas, que describan las oscilaciones subyacentes a las tasas de empleo. Antes de estimar las PCs se debe determinar el número de éstas que se van a seleccionar. Para ello, debido a que el análisis es exploratorio, no confirmatorio, la elección del número de PCs se realiza de forma empírica combinando tres reglas: porcentaje acumulado de varianza total explicada, tamaño de las varianzas de las PCs y gráfico de sedimentación.

Nº	Autovalores	Log - autovalores	Varianza explicada (%)	Varianza explicada acumulada (%)
1	1,8822	0,63	76,17	76,17
2	0,1381	-1,98	5,59	81,76
3	0,1337	-2,01	5,41	87,17
4	0,0635	-2,76	2,57	89,73
5	0,0575	-2,86	2,33	92,06
6	0,0495	-3,00	2,01	94,07
7	0,0386	-3,25	1,56	95,63
8	0,0305	-3,49	1,23	96,86
9	0,0195	-3,94	0,79	97,65
10	0,0155	-4,17	0,63	98,28
11	0,0115	-4,46	0,47	98,75
12	0,0084	-4,78	0,34	99,09
13	0,0062	-5,09	0,25	99,34
14	0,0055	-5,21	0,22	99,56
15	0,0046	-5,37	0,19	99,75
16	0,0033	-5,72	0,13	99,88
17	0,0030	-5,81	0,12	100,00

Tabla 3.4 Autovalores de la matriz de covarianzas y varianza explicada

En la Tabla 3.4 se recogen los autovalores de la matriz de covarianzas **C** (Anexo D) de las series filtradas y diferenciadas, asimismo, se acompañan los porcentajes de varianza que explican, tanto de forma individual como acumulada. Si se desea que el porcentaje de varianza total explicada sea superior al 80%, para obtener un buen ajuste, al menos se deben seleccionar dos PCs. Al mismo tiempo, teniendo en cuenta que el valor medio de los autovalores es 0,1454 y el 70% de dicho valor es 0,1018, por el criterio del tamaño de las varianzas propuesto por Jolliffe (1972) se deberían tomar las tres primeras PCs que en conjunto explican un 87,2% de la variación total y que es coherente con el resultado de la regla anterior. Si se tiene en cuenta el gráfico de sedimentación, Figura 3.3, o el diagrama del log-autovalor, Figura 3.4, en el que se aprecian con mayor claridad los cambios de pendiente, es a partir del cuarto autovalor cuando la pendiente se suaviza y, teniendo en cuenta que los autovalores segundo y tercero son casi idénticos, se deberían seleccionar entre tres o cuatro PCs. Si se seleccionan cuatro PCs, el porcentaje de la variación total que explican es del 89,7%, apenas 5 décimas más que si se escogen las tres primeras PCs. Por todo ello, para no obtener un número excesivo de PCs pero al mismo tiempo cubrir un porcentaje elevado de la variación total, se ha estimado oportuno optar por la estimación de las tres primeras componentes principales como aproximación de las oscilaciones subyacentes comunes a las series de las tasas de empleo en las diferentes comunidades.

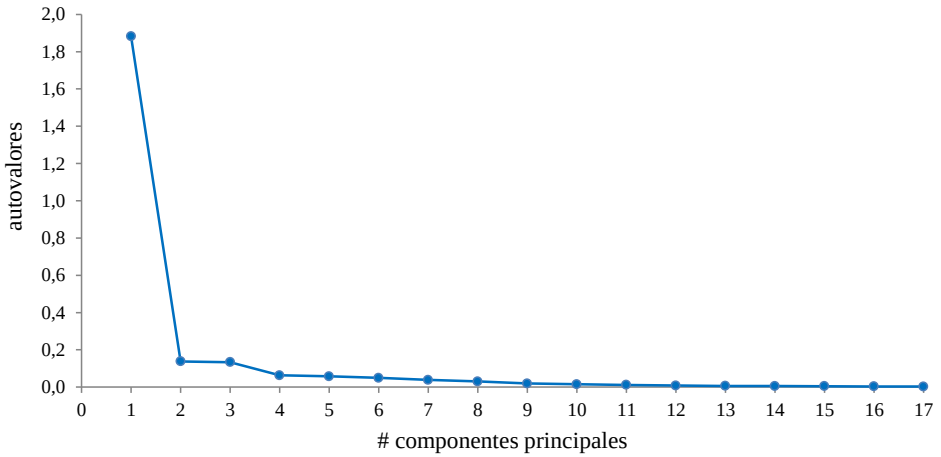

Figura 3.3 Gráfico de sedimentación

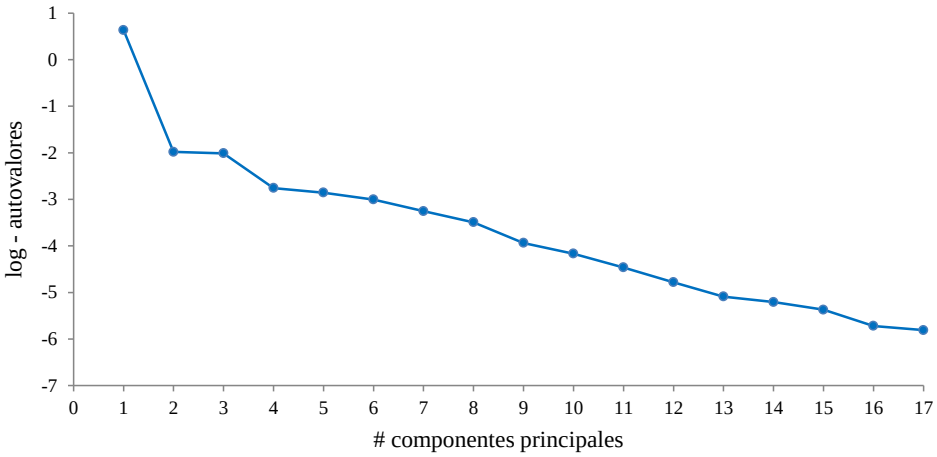

Figura 3.4 Diagrama del log-autovalor

La estimación de las PCs comienza centrando las series de trabajo, series filtradas de ruido y diferenciadas, a continuación diagonaliza ortogonalmente su matriz de covarianzas $\mathbf{C}$ para obtener la matriz ortogonal de autovectores $\mathbf{U}$ (Anexo D) y, finalmente, calcula las PCs centradas según la ecuación (3.1). Todo ello se ha realizado mediante la función *pca()*, programada en MATLAB (Anexo F), siguiendo la metodología expuesta en el apartado 3.3.1 de este capítulo.

Las PCs obtenidas a partir de datos centrados tienen media cero y sus varianzas están dadas por la primera identidad de (3.2) correspondiéndose con los mayores autovalores ordenados en sentido decreciente. Las PCs se estandarizan para poder compararlas con las ICs, que se derivarán posteriormente, puesto que, por construcción, éstas tienen varianza igual a la unidad.

La representación gráfica de las tres primeras PCs estandarizadas, que se han estimado, se encuentra en la Figura D.1 (Anexo D). En ella se observa que cada una de las tres PCs corresponde, como era de esperar, a señales cíclicas pero de diferentes periodos. Las señales de la primera PC corresponden a las oscilaciones de mayores periodos, a simple vista de cuatro o más años. Por el contrario, las oscilaciones que muestran las otras dos PCs tienen periodos menores, a simple vista entre dos y cuatro años, y similares por ser análogos sus correspondientes autovalores. No obstante, las oscilaciones, en gran parte, son de sentido contrario puesto que un máximo de la segunda PC se corresponde con un mínimo de la tercera y viceversa.

3.4.3 Análisis de Componentes Independientes, ICA

El ICA, por las mismas razones que el PCA, se realiza sobre las series filtradas de ruido y diferenciadas, estacionarias en sentido estricto, con el fin de obtener unas componentes comunes, en este caso independientes, que describan las oscilaciones subyacentes a las tasas de empleo. Antes de estimar las matrices de separación y derivar las ICs se debe determinar cuántas se van a calcular. Para ello, se puede optar por estimar el mismo número que de PCs. Sin embargo, en este trabajo se hace uso del gráfico de la falta de aproximación de $\mathbf{A}_{p\times k}\mathbf{A}_{k\times p}^{T}$ a la matriz de covarianzas $\mathbf{C}$, de los pares (k,δ_k), según lo argüido a partir del resultado (3.21). En la Tabla 3.5 se recogen los valores de δ_k, la distancia de $\mathbf{C}$ a sus aproximaciones de rango inferior $\mathbf{A}_{p\times k}\mathbf{A}_{k\times p}^{T}$.

Nº	Autovalores	δ_k
1	1,8822	0,2238
2	0,1381	0,1761
3	0,1337	0,1147
4	0,0635	0,0955
5	0,0575	0,0762
6	0,0495	0,0579
7	0,0386	0,0431
8	0,0305	0,0306
9	0,0195	0,0235
10	0,0155	0,0177
11	0,0115	0,0134
12	0,0084	0,0105
13	0,0062	0,0085
14	0,0055	0,0064
15	0,0046	0,0044
16	0,0033	0,0030
17	0,0030	

Tabla 3.5 Autovalores de la matriz $\mathbf{C}$ y distancias δ_k de la aproximación a la matriz.

En el gráfico de los pares (k, δ_k) de la Figura 3.5 se observa que a partir del tercer autovalor la pendiente desciende de forma notable. Ello significa que la distancia de la matriz de covarianzas $\mathbf{C}$ a sus aproximaciones de rango inferior $\mathbf{A}_{p\times k}\mathbf{A}_{k\times p}^{T}$ no se reduce de manera sustancial. Por ello, el número de ICs que se elige para estimar es de tres.

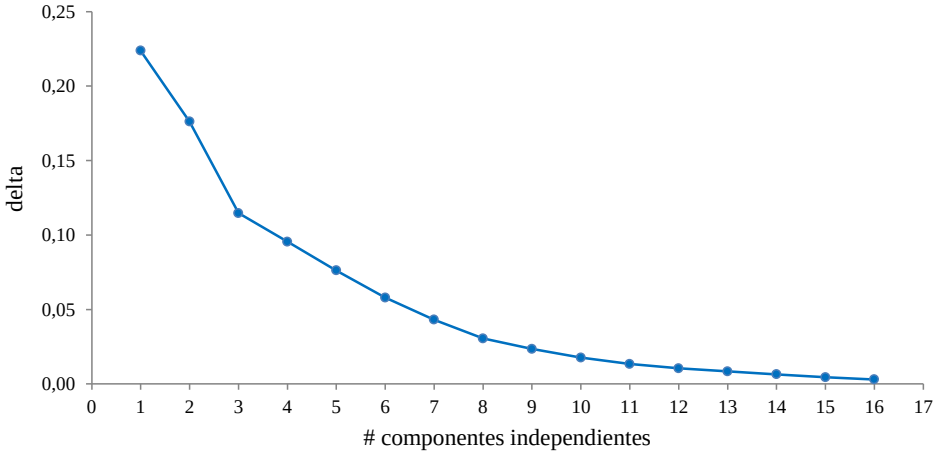

Figura 3.5 Gráfico de los valores δ_k según el número de ICs estimadas.

La estimación de las ICs comienza con el centrado de las series de trabajo para, a continuación, diagonalizar ortogonalmente su matriz de covarianzas $\mathbf{C}$ con el fin de obtener la matriz $\mathbf{Q}$ (Anexo E) y blanquear los datos centrados. A partir de los datos blanqueados se estima la matriz ortogonal de separación $\mathbf{B}$ (Anexo E) tanto con el algoritmo FastICA como con el algoritmo AMUSE. Tanto un algoritmo como otro se han programado en MATLAB, funciones *icarapid()* y *amuse()* respectivamente (Anexo F), según la metodología expuesta en el apartado 3.3.3 de este capítulo. Una vez calculada la matriz $\mathbf{B}$, se deriva (para cada uno de los dos algoritmos) la matriz de mezcla $\mathbf{A}$ (Anexo E), para obtener los datos observados a partir de las ICs, y la matriz de separación $\mathbf{W}$ (Anexo E), para obtener las ICs a partir de los datos observados, según las ecuaciones (3.19) y (3.20) respectivamente. De esta forma, se tienen dos estimaciones de las ICs a partir de la ecuación (3.16), las calculadas con el algoritmo FastICA, Figura E.1 (Anexo E), y las computadas con el algoritmo AMUSE, Figura E.2 (Anexo E).

Las ICs calculadas con el algoritmo AMUSE están ordenadas según su contribución a la varianza de las variables observadas al estar derivadas a partir de una matriz de covarianzas como en el PCA. Por el contrario, el algoritmo FastICA computa las ICs

sin ningún criterio de ordenación preestablecido. Debido a ello, para poder realizar comparaciones entre las ICs estimadas por uno y otro algoritmo, se han ordenado las deducidas por el algoritmo FastICA según la norma de las columnas de la matriz de mezcla $\mathbf{A}$ que determina la contribución de las ICs a la varianza total de las variables observadas. Posteriormente, se reordenan las columnas de las matrices $\mathbf{B}^T$, $\mathbf{A}$ y $\mathbf{W}^T$ obtenidas con el algoritmo FastICA y, como tales, son las que aparecen en el Anexo E.

De la observación de las Figuras E.1 y E.2 (Anexo E) se desprende que las ICs estimadas por ambos algoritmos son similares y se corresponden, al igual que las PCs del PCA, las primeras ICs con oscilaciones de cuatro o más años y las segundas y terceras con oscilaciones de periodos entre dos y cuatro años. Una comparación más exhaustiva entre las PCs y ambas ICs se ofrece en la siguiente sección.

3.5 Comparación de resultados PCA – ICA

En esta sección se comparan las PCs calculadas y ambas estimaciones de las ICs tanto desde un punto de vista geométrico como espectral. El motivo es que no se puede afirmar, según se prueba en el siguiente teorema, que un tipo de análisis sea mejor que otro (en el sentido de proporcionar una mejor aproximación a los datos observados con los datos reconstruidos a partir de las correspondientes componentes) medido por la suma de los cuadrados de los residuos, SCR.

Teorema 3.2 *Si se estiman igual número de PCs que de ICs determinadas a partir de los mismos autovalores de la matriz de covarianzas de los datos observados, entonces los datos reconstruidos a partir de esas PCs y de esas ICs son los mismos.*

Demostración. Sea $\mathbf{x}_{p\times 1}$ el vector de datos observados centrados siendo $\mathbf{C} = E\left[\mathbf{x}_{p\times 1}\mathbf{x}_{1\times p}^T\right] = \mathbf{U}\mathbf{D}\mathbf{U}^T$ la diagonalización de su matriz de covarianzas. Sea $\mathbf{D}_{m\times m} = \text{diag}\left(\lambda_1,\cdots,\lambda_m\right)$ donde, sin pérdida de generalidad, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$ son los m mayores autovalores de la matriz $\mathbf{C}$ y sea $\mathbf{U}_{p\times m} = \left[\mathbf{u}_1,\mathbf{u}_2,\cdots,\mathbf{u}_m\right]$ la matriz con los autovectores ortonormales asociados a esos primeros m autovalores.

El vector con las m primeras PCs está dado por $\mathbf{y}_{m\times 1} = \mathbf{U}_{m\times p}^T\mathbf{x}_{p\times 1}$ y, por tanto, el vector con los datos reconstruidos a partir de las PCs es

$$\hat{\mathbf{x}}_{p\times 1}^{PCA} = \mathbf{U}_{p\times m}\mathbf{y}_{m\times 1} = \mathbf{U}_{p\times m}\mathbf{U}_{m\times p}^T\mathbf{x}_{p\times 1}$$

Por otro lado, para estimar las ICs, en lugar de utilizar la matriz blanqueadora dada por (3.14) se va a especificar una más general definida por

$$\mathbf{Q}_{m\times p} = \mathbf{V}_{m\times m}\mathbf{D}^{-1/2}_{m\times m}\mathbf{U}^{T}_{m\times p}$$

donde $\mathbf{V}_{m\times m}$ es una matriz ortogonal, $\mathbf{V}_{m\times m}\mathbf{V}^{T}_{m\times m} = \mathbf{I}_{m\times m}$.

Si $\mathbf{B}_{m\times m}$ es la matriz ortogonal de separación (obtenida por algún algoritmo) tal que

$$\mathbf{s}_{m\times 1} = \mathbf{B}_{m\times m}\mathbf{z}_{m\times 1}$$

entonces, se puede afirmar que las ICs son una transformación ortogonal, una rotación, de las PCs estandarizadas.

Ahora bien, la matriz de mezcla $\mathbf{A}_{p\times m}$, para obtener los datos observados $\mathbf{x}_{p\times 1}$, dada en (3.19) se transforma en

$$\mathbf{A}_{p\times m} = \mathbf{U}_{p\times m}\mathbf{D}^{1/2}_{m\times m}\mathbf{V}^{T}_{m\times m}\mathbf{B}^{T}_{m\times m}$$

y la matriz de separación $\mathbf{W}_{m\times p}$, para obtener $\mathbf{s}_{m\times 1}$ a partir de $\mathbf{x}_{p\times 1}$, dada en (3.20) se convierte en

$$\mathbf{W}_{m\times p} = \mathbf{B}_{m\times m}\mathbf{V}_{m\times m}\mathbf{D}^{-1/2}_{m\times m}\mathbf{U}^{T}_{m\times p}$$

Por todo ello, y sustituyendo (3.16) en (3.15), el vector con los datos reconstruidos a partir de las ICs es

$$\begin{aligned}
\hat{\mathbf{x}}^{ICA}_{p\times 1} &= \mathbf{A}_{p\times m}\mathbf{s}_{m\times 1} = \mathbf{A}_{p\times m}\mathbf{W}_{m\times p}\mathbf{x}_{p\times 1} \\
&= \mathbf{U}_{p\times m}\mathbf{D}^{1/2}_{m\times m}\mathbf{V}^{T}_{m\times m}\mathbf{B}^{T}_{m\times m}\mathbf{B}_{m\times m}\mathbf{V}_{m\times m}\mathbf{D}^{-1/2}_{m\times m}\mathbf{U}^{T}_{m\times p}\mathbf{x}_{p\times 1} \\
&= \mathbf{U}_{p\times m}\mathbf{D}^{1/2}_{m\times m}\mathbf{V}^{T}_{m\times m}\mathbf{V}_{m\times m}\mathbf{D}^{-1/2}_{m\times m}\mathbf{U}^{T}_{m\times p}\mathbf{x}_{p\times 1} \\
&= \mathbf{U}_{p\times m}\mathbf{D}^{1/2}_{m\times m}\mathbf{D}^{-1/2}_{m\times m}\mathbf{U}^{T}_{m\times p}\mathbf{x}_{p\times 1} \\
&= \mathbf{U}_{p\times m}\mathbf{U}^{T}_{m\times p}\mathbf{x}_{p\times 1} = \hat{\mathbf{x}}^{PCA}_{p\times 1}
\end{aligned}$$

Y así, queda demostrado el teorema. ■

Es evidente, que si los dos modelos de análisis, PCA e ICA, tienen el mismo número de componentes determinadas por los mismos autovalores pero en uno de ellos se estima una componente más, entonces dicho modelo será mejor por tener menor SCR. Sin embargo, lo interesante no es mejorar la aproximación de los datos observados con los reconstruidos añadiendo componentes en un modelo de análisis, sino comparar ambos modelos o, en el caso del ICA, diferentes algoritmos con el mismo número de componentes explicando la geometría y el espectro de sus componentes como realizan Sebastiao y Oliveira (2009).

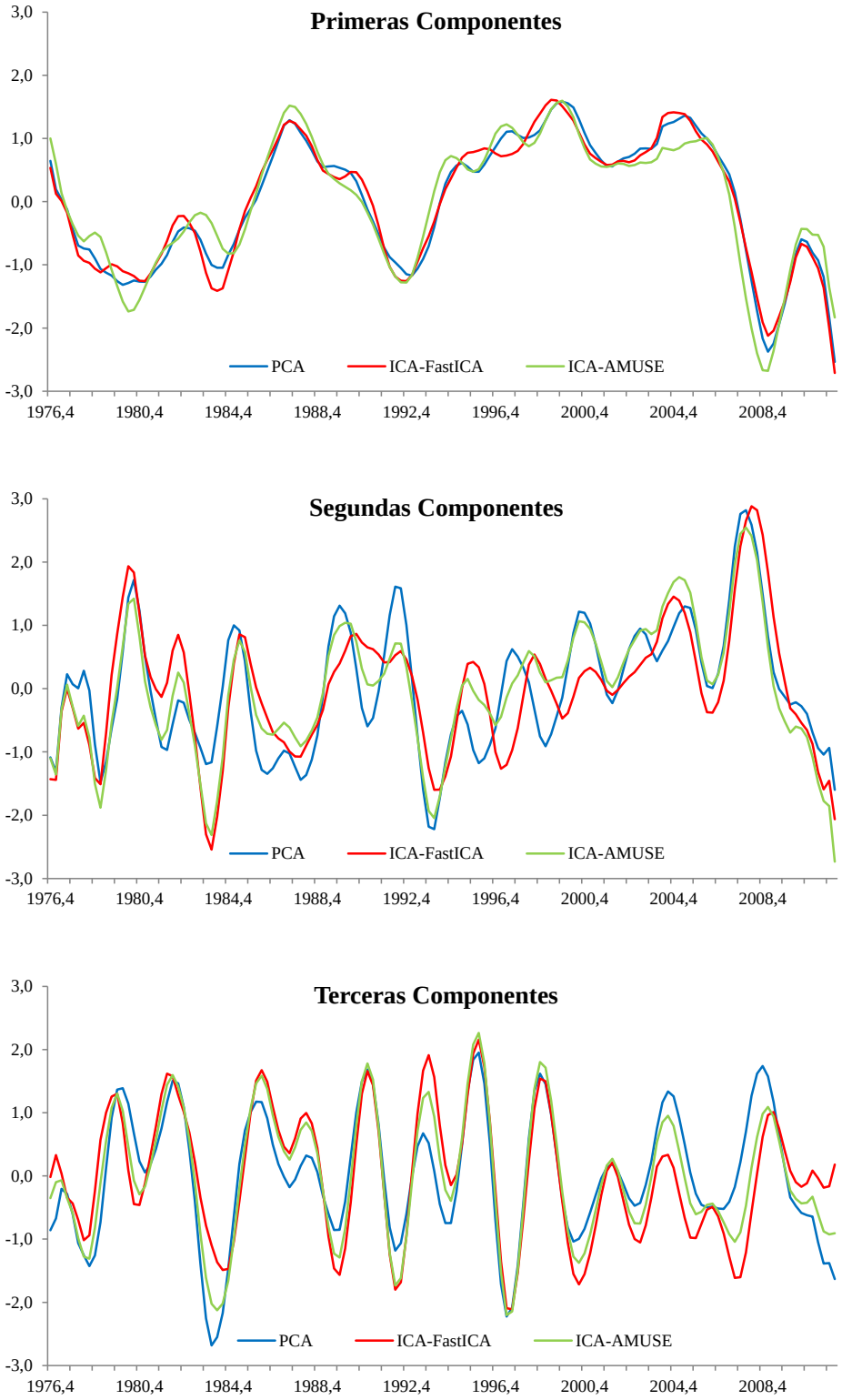

Figura 3.6 Coordenadas de las series de las tres primeras componentes estimadas mediante PCA, ICA-FastICA e ICA-AMUSE.

La semejanza existente entre las PCs, ICs-FastICA e ICs-AMUSE se evidencia en la Figura 3.6, donde, como ya se ha mencionado, las componentes estimadas por cada técnica se han ordenado según su contribución a la varianza total de las variables observadas. No obstante, existe una mayor semejanza entre las primeras componentes (representando oscilaciones de periodos superiores a cuatro años) y entre las terceras componentes que entre las segundas componentes (ambas representando oscilaciones de periodos entre dos y cuatro años).

Para analizar las coordenadas de las series temporales de las componentes se comparan las correlaciones entre los pares de componentes de las dos técnicas, PCA e ICA, cuyos valores aparecen en la Tabla 3.6. En dicha tabla, se aprecia que las correlaciones entre las componentes de PCA e ICA son mejores para las calculadas con el algoritmo AMUSE que con el algoritmo FastICA. En efecto, además de obtener mayores correlaciones en la diagonal entre PCA e ICA-AMUSE, fuera de la diagonal (donde las correlaciones deberían ser pequeñas) se alcanza una elevada y significativa correlación entre IC2-FastICA y PC3 y entre IC3-FastICA y PC2 que, por sus valores absolutos elevados, puede llevar a confundir las segundas con las terceras componentes entre las diferentes técnicas. Sin embargo, esta hipotética confusión no tiene lugar entre las componentes de PCA e ICA-AMUSE ya que cada una de las PCs tiene elevada correlación con solo una de las ICs y viceversa. Por otro lado, como era de esperar, cada una de las ICs-FastICA está correlacionada de forma significativa con solo una de las ICs-AMUSE y viceversa.

		PCA			**FastICA**		
		PC1	PC2	PC3	IC1	IC2	IC3
FastICA	IC1	0,987	-0,030	0,157			
	IC2	-0,080	0,757	0,649			
	IC3	-0,138	-0,653	0,744			
AMUSE	IC1	0,960	-0,251	-0,122	0,936	-0,346	-0,060
	IC2	0,279	0,884	0,374	0,308	0,890	-0,337
	IC3	0,014	-0,394	0,919	0,170	0,298	0,939

Tabla 3.6 Coeficientes de correlación entre las series de las componentes

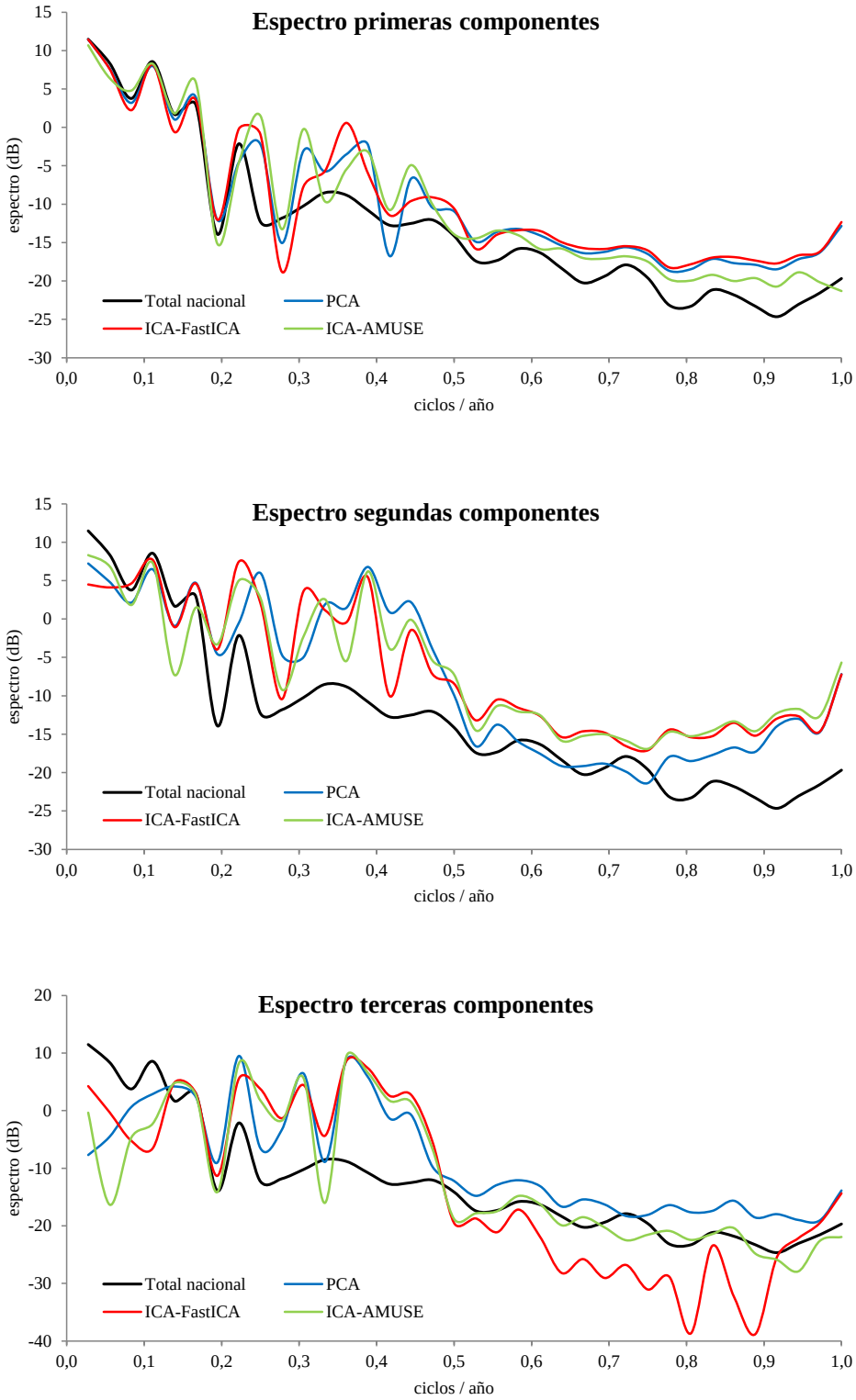

Figura 3.7 Espectro de las series de las tres primeras componentes estimadas mediante PCA, ICA-FastICA e ICA-AMUSE comparadas con el espectro del Total nacional.

Otro procedimiento para estudiar las series temporales correspondientes a las componentes es con el análisis de su espectro. En la Figura 3.7 se apreciar la semejanza de los espectros de cada grupo de componentes (primeras, segundas y terceras) calculadas por los diferentes métodos (PCA, ICA-FastICA e ICA-AMUSE). El espectro de cada grupo de componentes posee una estructura similar al espectro de la serie del Total nacional, aunque cada grupo resalta unos diferentes picos de frecuencias, lo cual pone de manifiesto la semejanza entre las componentes y el Total nacional.

El espectro de las primeras componentes resalta los principales picos de frecuencias del Total nacional, que se corresponden con las oscilaciones de 9, 6 y 4,5 años, y, además, destaca los picos de frecuencias propios a las oscilaciones de alrededor de los 3 años. Por su parte, el espectro de las segundas componentes aporta información para los dos primeros picos de frecuencias del Total nacional y, sobre todo, enfatiza el tercer pico de frecuencia (oscilaciones de 4,5 años) y los picos de frecuencias correspondientes a las oscilaciones entre 2 y 3,5 años. Finalmente, el espectro de las terceras componentes es similar al de las segundas excepto que no enfatiza el primer pico de frecuencias del Total nacional, oscilaciones de 9 años.

		PCA			**FastICA**		
		PC1	PC2	PC3	IC1	IC2	IC3
FastICA	IC1	0,855	-0,097	0,510			
	IC2	0,022	-0,754	-0,657			
	IC3	0,037	0,653	-0,756			
AMUSE	IC1	0,680	-0,657	-0,325	0,959	-0,193	0,206
	IC2	0,078	0,916	0,394	-0,244	-0,934	0,260
	IC3	0,004	-0,388	0,921	0,142	-0,300	-0,943

Tabla 3.7 Cosenos de los ángulos entre los vectores de las componentes.

Una última comparación entre las diferentes técnicas se realiza con los cosenos de los ángulos entre los coeficientes de los vectores de las componentes. La Tabla 3.7 revela que las ICs ordenadas de FastICA son menos ortogonales que las de AMUSE a las PCs, o datos blanqueados, puesto que los elementos de la diagonal, de las matrices de cosenos, son menores en valor absoluto. Y, como era de esperar, la matriz de cosenos entre FastICA y AMUSE es, en valor absoluto, muy parecida a la matriz identidad.

3.6 Conclusiones

En este trabajo se han descrito, con cierto detalle, dos técnicas, PCA e ICA, con el objetivo de obtener las componentes oscilatorias comunes subyacentes a un grupo de series temporales, las tasas de empleo de las CC.AA. El objetivo del PCA es reducir la dimensión de un conjunto de datos mediante la estimación de un grupo de componentes principales no correlacionadas que explique la mayor varianza posible del conjunto de datos. Por su parte, el objetivo del ICA es estimar, a partir de los datos observados, un conjunto de componentes que son independientes, no-gaussianas y que generan los datos observados.

Se ha probado que el PCA se puede usar con series temporales siempre que el objetivo sea descriptivo y no inferencial puesto que los datos observados están correlacionados y por tanto no son independientes. De igual forma, la técnica SSA, método especial del PCA aplicado a una sola serie temporal, se ha demostrado que, siempre y cuando la serie sea estacionaria en varianza aunque no necesariamente en media según se desprende del Teorema 3.1, es muy eficaz para filtrar series temporales sin recurrir a filtros fijos o basados en modelos. En concreto, el SSA ha servido para eliminar de las series de las tasas de empleo las oscilaciones de periodo igual o inferior al año, es decir, corregir de estacionalidad e irregularidad. Además, las series han debido ser diferenciadas tanto para conseguir estacionariedad como obtener unas componentes subyacentes que se correspondan, en su totalidad, con señales oscilatorias.

La determinación del número de PCs a estimar se ha realizado de forma empírica combinando tres reglas: porcentaje acumulado de varianza total explicada, tamaño de las varianzas de las PCs y gráfico de sedimentación, de las diversas que existen, puesto que el análisis de las series de las tasas de empleo ha sido exploratorio y no confirmatorio. Sin embargo, para determinar el número de ICs a estimar, cuando su número es menor al de series observadas, no se ha encontrado en la literatura ninguna regla. Por ello, se ha propuesto, motivando su utilización, el gráfico de la falta de aproximación de la matriz de covarianzas de los datos observados, los pares $\left(k, \delta_k\right)$ donde δ_k está dado por (3.21).

Para estimar las ICs se han utilizado dos algoritmos. El primero de ellos, el algoritmo FastICA, obtiene las ICs maximizando la no-gaussianidad mediante una aproximación de la sintropía. El segundo, el algoritmo AMUSE, para superar los problemas que presentan los algoritmos basados en medidas de no-gaussianidad cuando las ICs son gaussianas, utiliza una matriz de autocovarianzas retardada para obtener

las correspondientes ICs. Ambos algoritmos parten de los datos blanqueados, es decir, de las PCs estandarizadas, para estimar las ICs. De esta forma, las ICs son una transformación ortogonal, una rotación, de las PCs.

Finalmente, se han estimado tres PCs y tres ICs (utilizando FastICA y AMUSE). La comparación de los resultados, entre PCA e ICA así como entre FastICA y AMUSE, se ha efectuado por métodos geométricos y espectrales. Ello es debido a que, por el Teorema 3.2, la comparación en términos de distancia, de los datos reconstruidos a los observados, medida por la SCR no es posible porque las diferentes técnicas obtienen los mismos datos reconstruidos.

Las series de las primeras y terceras componentes estimadas tanto para PCA, ICA-FastICA e ICA-AMUSE son semejantes y en menor medida ocurre para las segundas componentes. Sin embargo, no hay que olvidar que, aunque las PCs están no correlacionadas, las ICs son independientes siendo la independencia una cualidad estadística más deseable ya que evita la falta de identificación cuando las componentes son gaussianas. Por otra parte, aunque el espectro de la serie del Total nacional es semejante a los espectros de las componentes, estos resaltan más picos de frecuencias porque el Total es un promedio de medidas relativas como son los ratios.

Debido a que las ICs son rotaciones de las PCs, del análisis de las matrices de correlaciones entre PCs e ICs-FastICA y PCs e ICs-AMUSE se desprende que, para series temporales, el algoritmo AMUSE es mejor que el FastICA porque su correspondiente matriz se parece más, en valor absoluto, a la matriz identidad. Así mismo, las ICs ordenadas de FastICA son menos ortogonales que las de AMUSE a las PCs puesto que su correspondiente matriz de cosenos se parece menos, en valor absoluto, a la matriz identidad de forma que este criterio también señala al algoritmo AMUSE como mejor que el FastICA cuando se trabaja con series temporales.

En consecuencia, por todo lo expuesto, si los datos observados se corresponden con series temporales es preferible estimar un modelo ICA para obtener componentes subyacentes comunes y utilizar el algoritmo AMUSE.

Finalmente, indicar que una posible línea de investigación futura sería realizar un trabajo similar al presentado pero añadiendo un término de ruido al modelo ICA.

Anexo A

Series Observadas de las Tasas de Empleo: Figuras

En este Anexo se presentan las siguientes figuras:

- Figura A.1 Tasas de empleo de Andalucía, Aragón y Asturias.
- Figura A.2 Tasas de empleo de Baleares, Canarias y Cantabria.
- Figura A.3 Tasas de empleo de Castilla y León, Castilla – La Mancha y Cataluña.
- Figura A.4 Tasas de empleo de C. Valenciana, Extremadura y Galicia.
- Figura A.5 Tasas de empleo de Madrid, Murcia y Navarra.
- Figura A.6 Tasas de empleo de País Vasco, La rioja y Total Nacional.
- Figura A.7 Funciones de autocorrelación de Andalucía, Aragón, Asturias, Baleares, Canarias, Cantabria, Castilla y León, Castilla – La Mancha y Cataluña.
- Figura A.8 Funciones de autocorrelación de C. Valenciana, Extremadura, Galicia, Madrid, Murcia, Navarra, País Vasco, La rioja y Total Nacional.
- Figura A.9 Funciones de autocorrelación parcial de Andalucía, Aragón, Asturias, Baleares, Canarias, Cantabria, Castilla y León, Castilla – La Mancha y Cataluña.
- Figura A.10 Funciones de autocorrelación parcial de C. Valenciana, Extremadura, Galicia, Madrid, Murcia, Navarra, País Vasco, La rioja y Total Nacional.

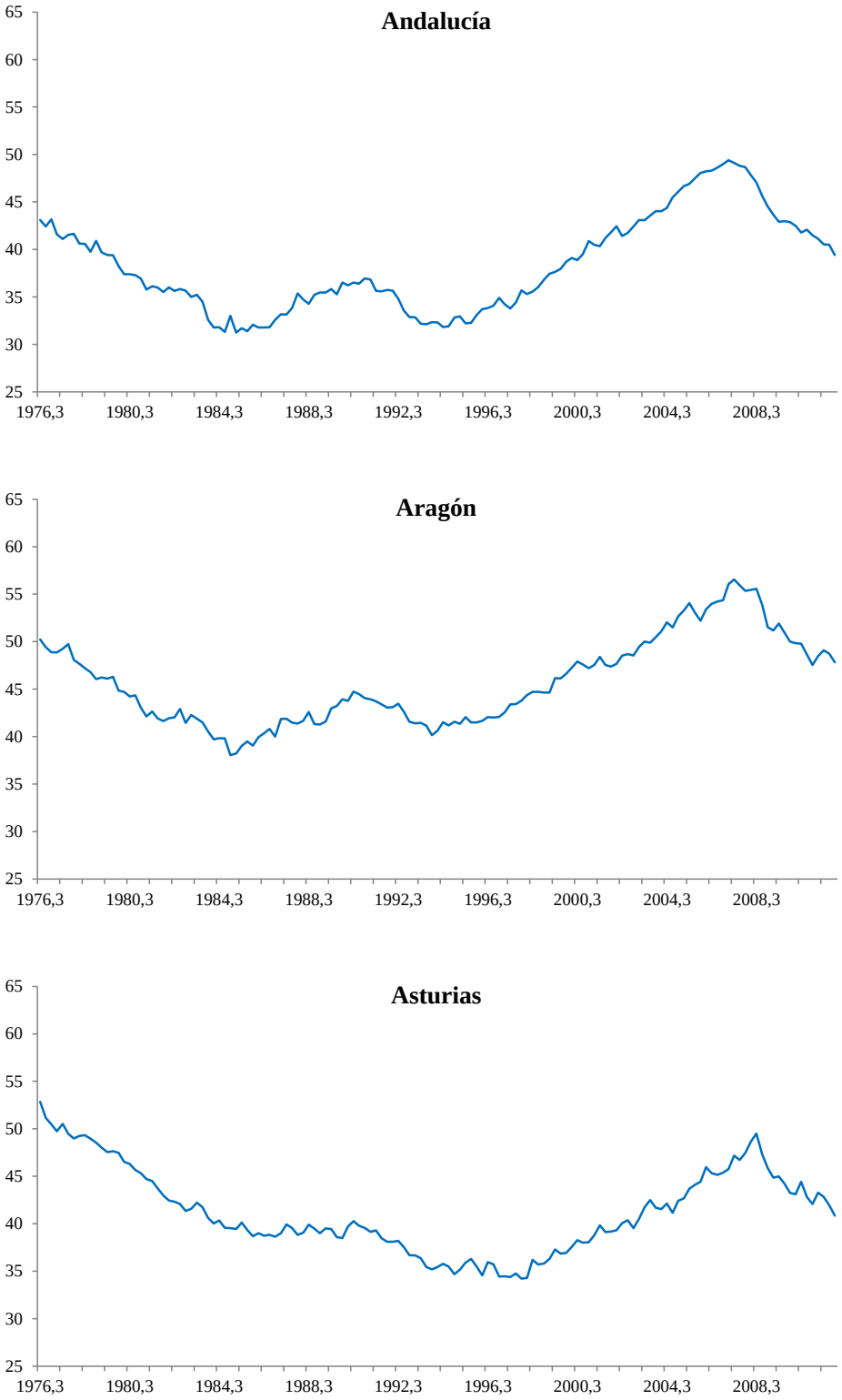

Figura A.1 Tasas de empleo de Andalucía, Aragón y Asturias.

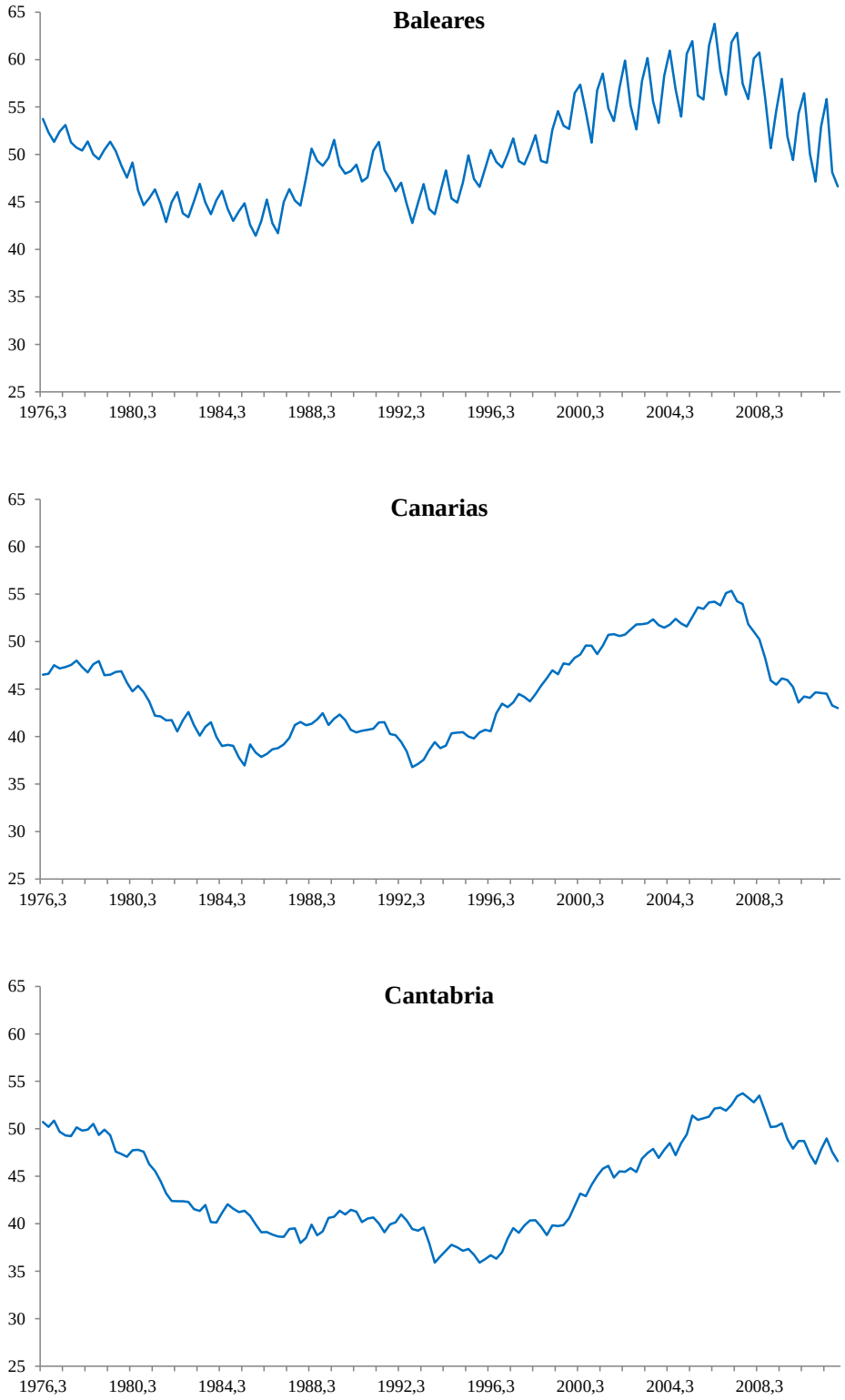

Figura A.2 Tasas de empleo de Baleares, Canarias y Cantabria.

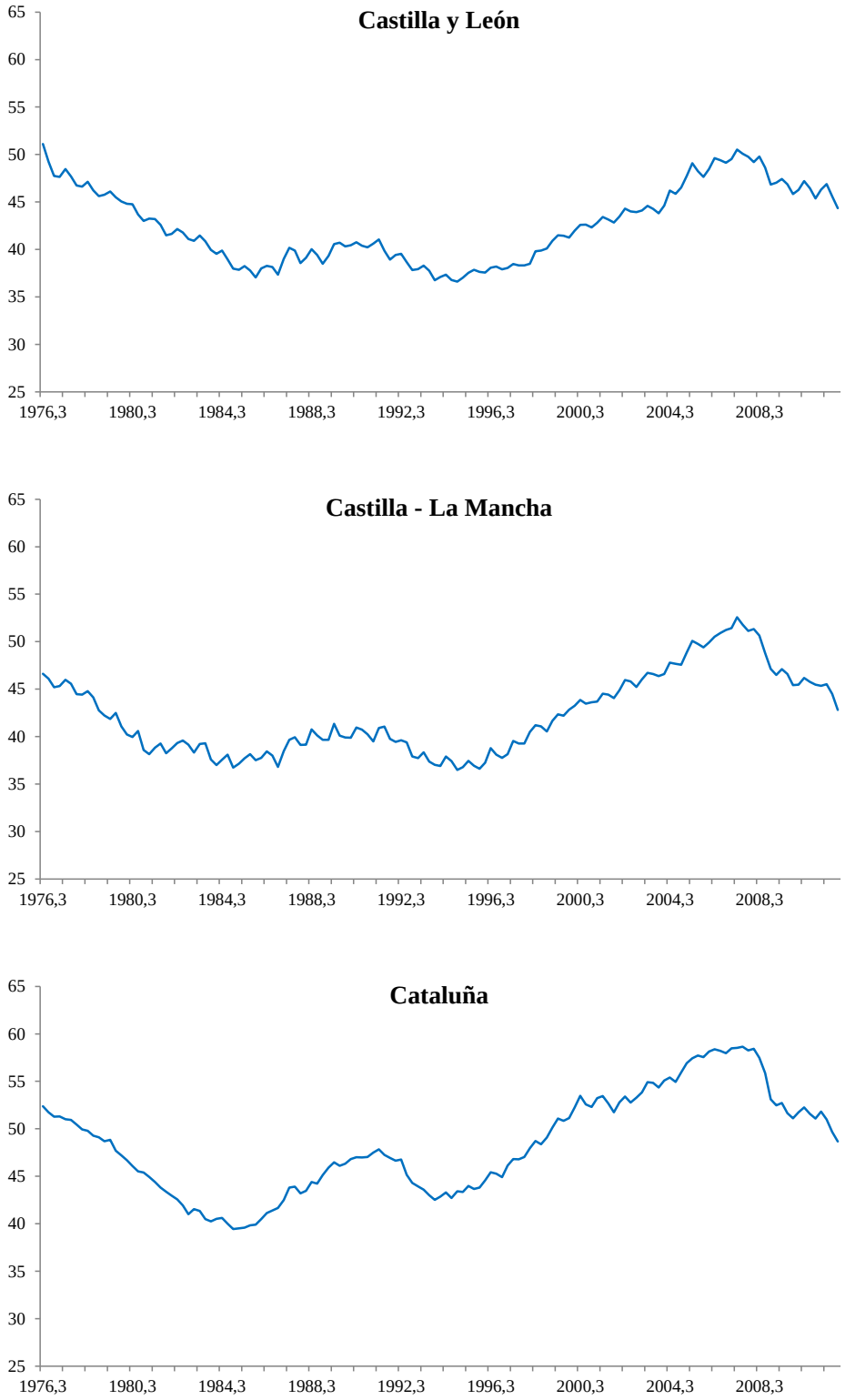

Figura A.3 Tasas de empleo de Castilla y León, Castilla – La Mancha y Cataluña.

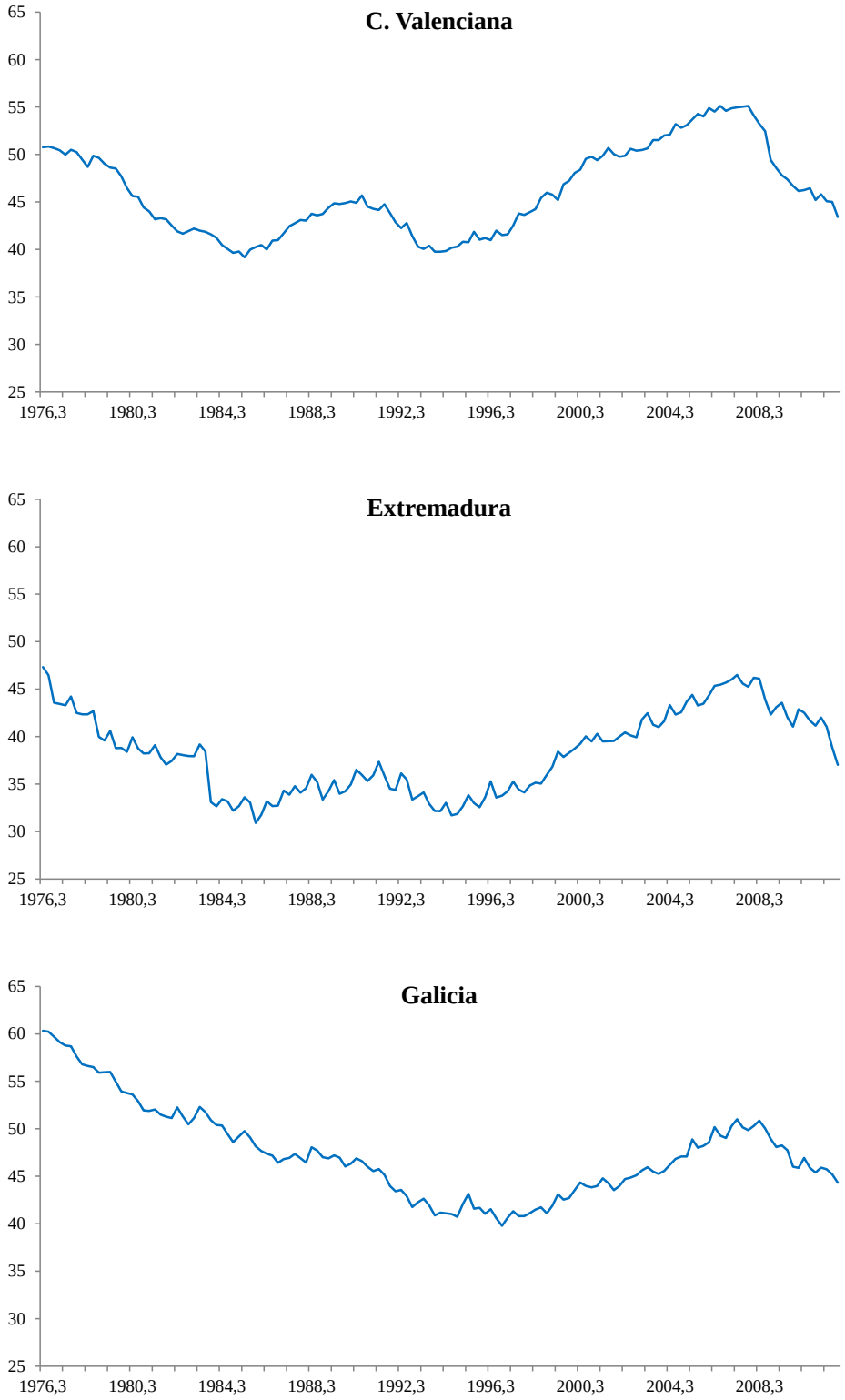

Figura A.4 Tasas de empleo de C. Valenciana, Extremadura y Galicia.

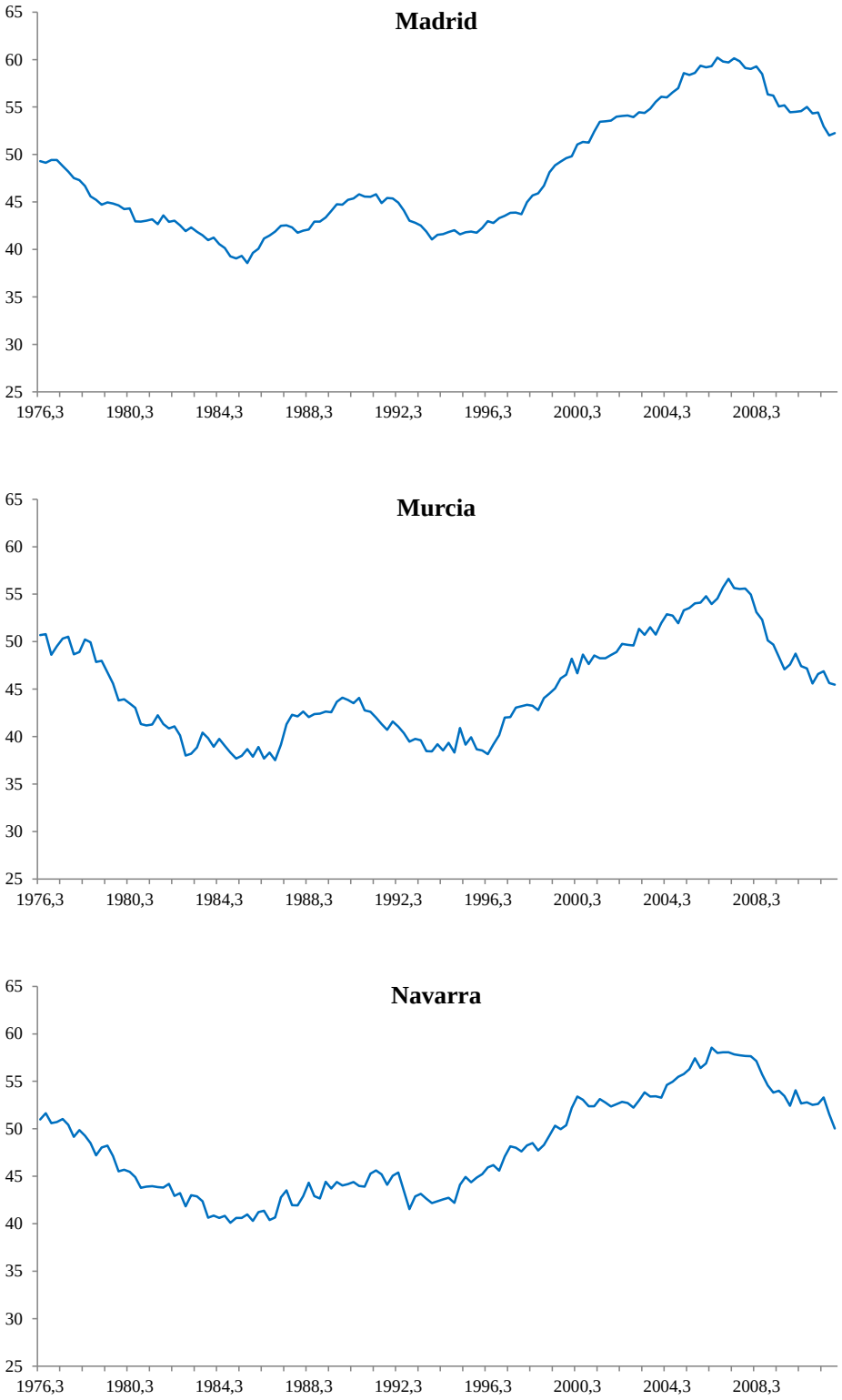

Figura A.5 Tasas de empleo de Madrid, Murcia y Navarra.

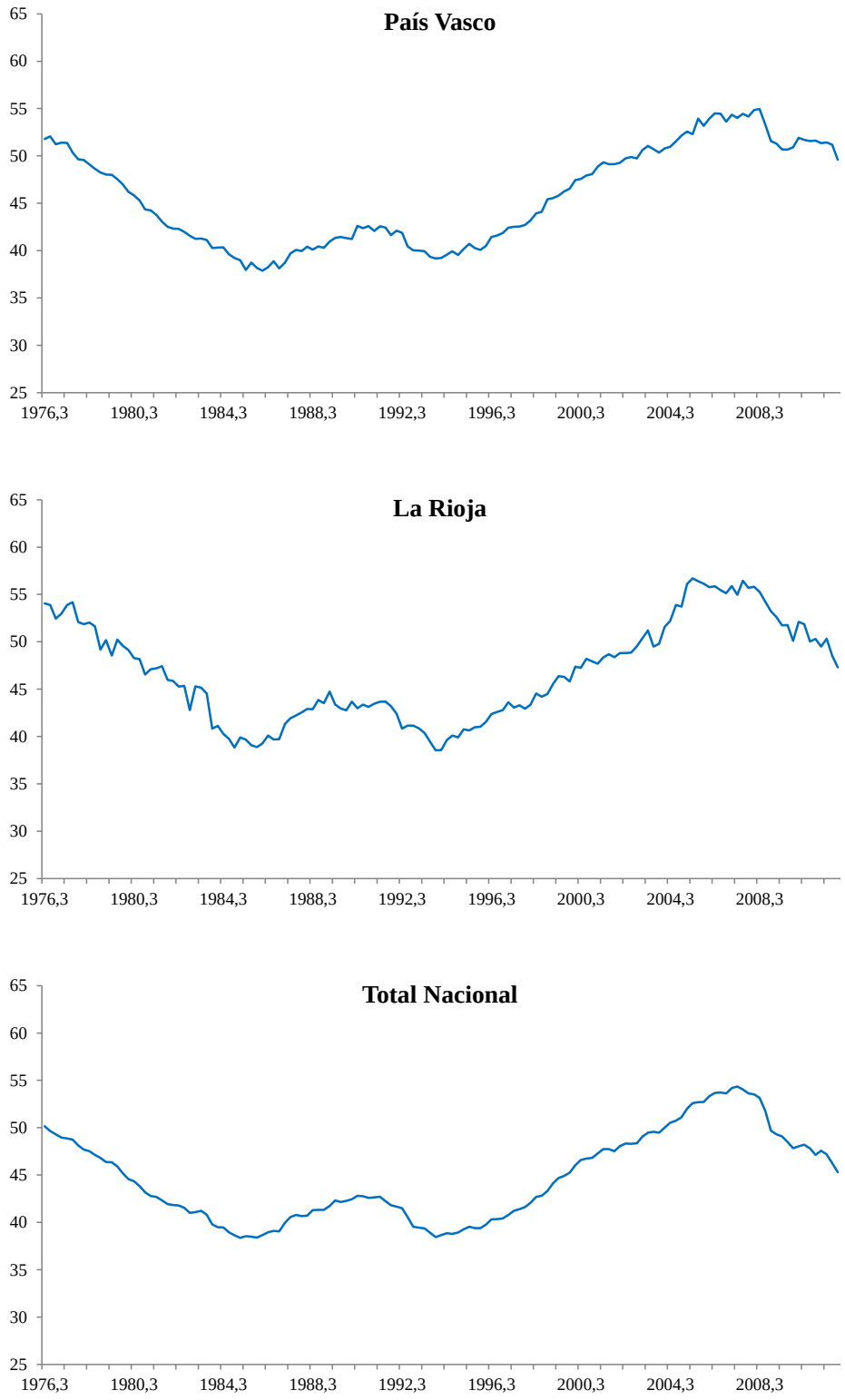

Figura A.6 Tasas de empleo de País Vasco, La Rioja y Total Nacional.

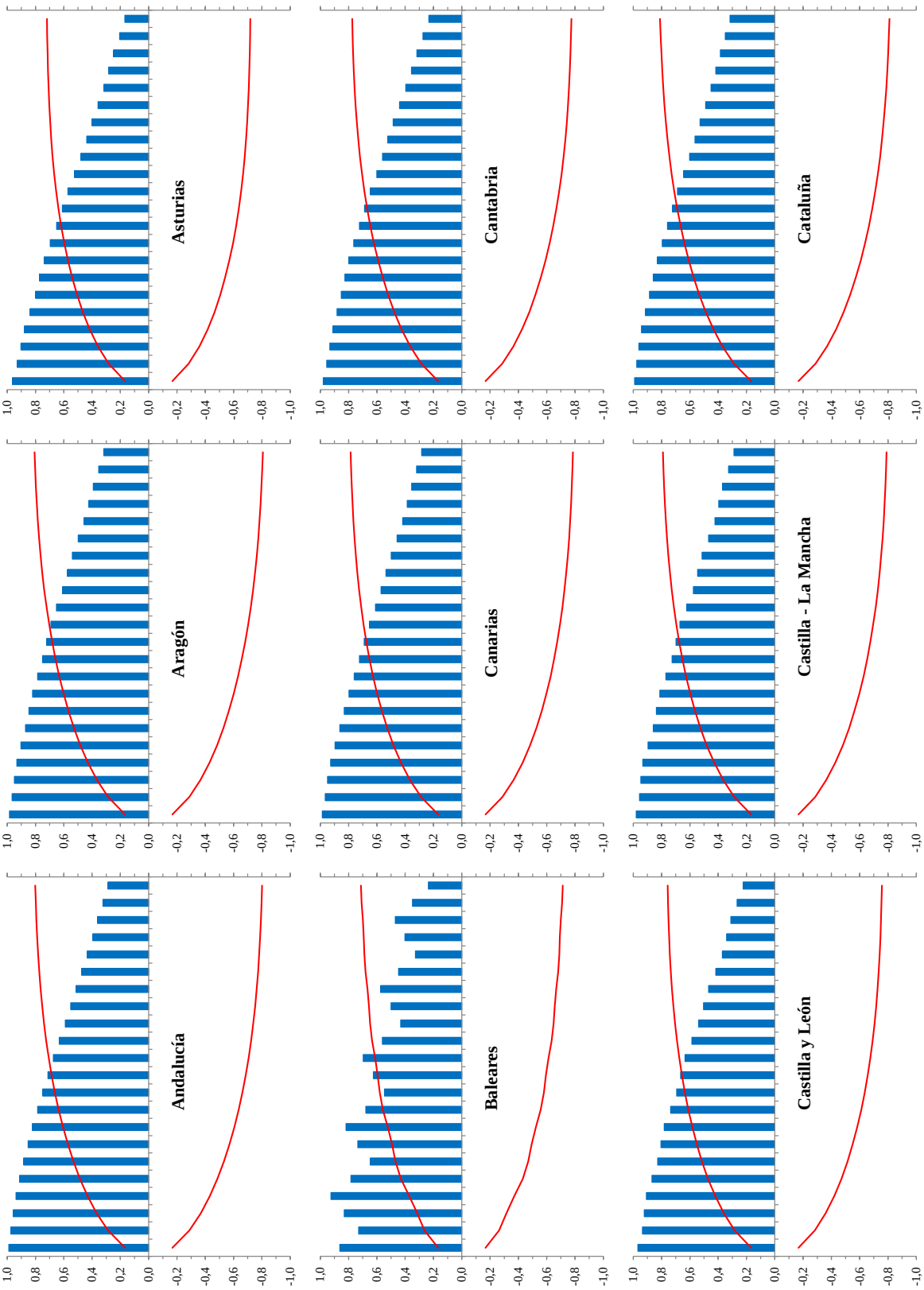

Figura A.7 Funciones de autocorrelación de Andalucía, Aragón, Asturias, Baleares, Canarias, Cantabria, Castilla y León, Castilla – La Mancha y Cataluña.

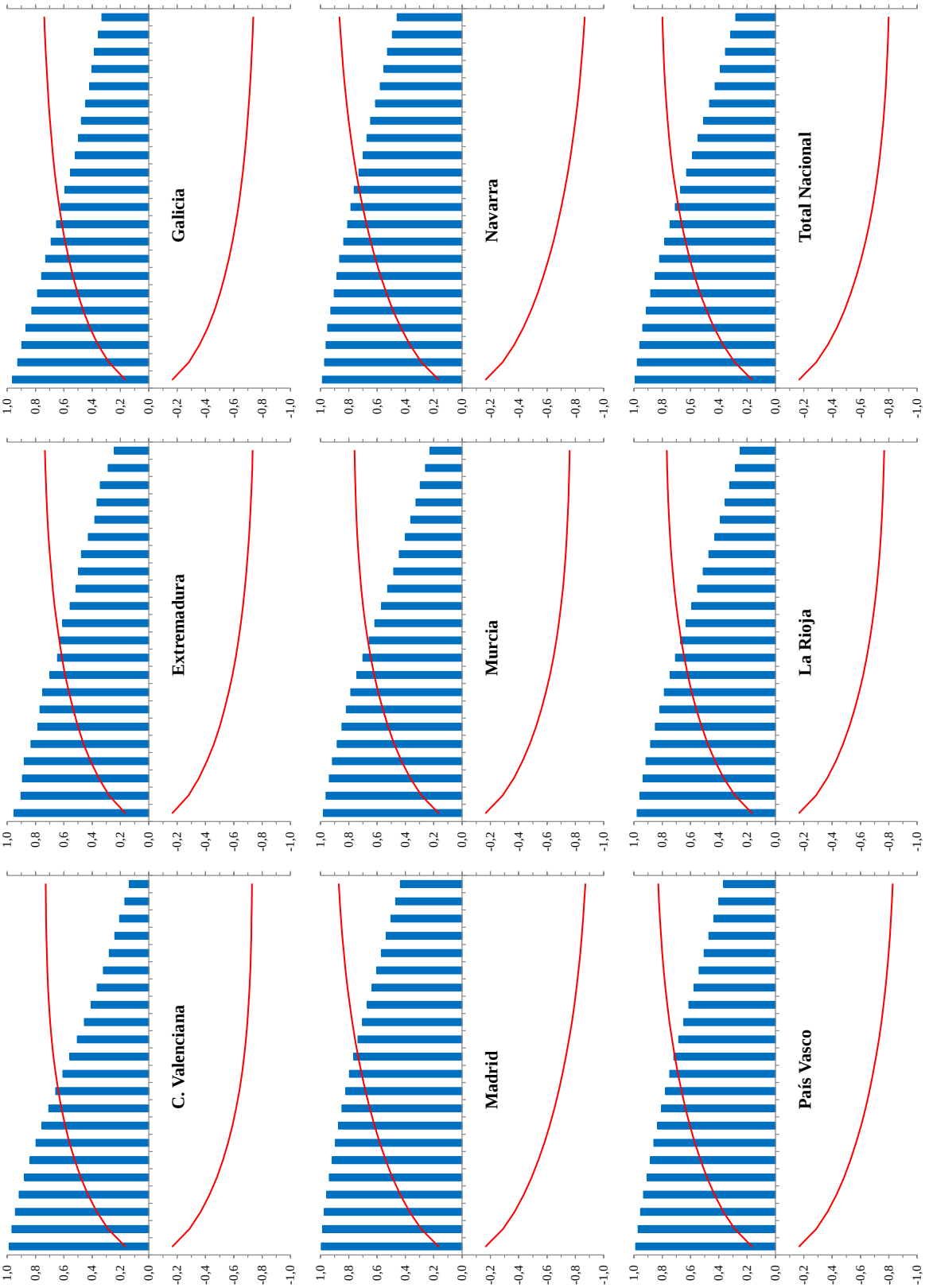

Figura A.8 Funciones de autocorrelación de C. Valenciana, Extremadura, Galicia, Madrid, Murcia, Navarra, País Vasco, La rioja y Total Nacional.

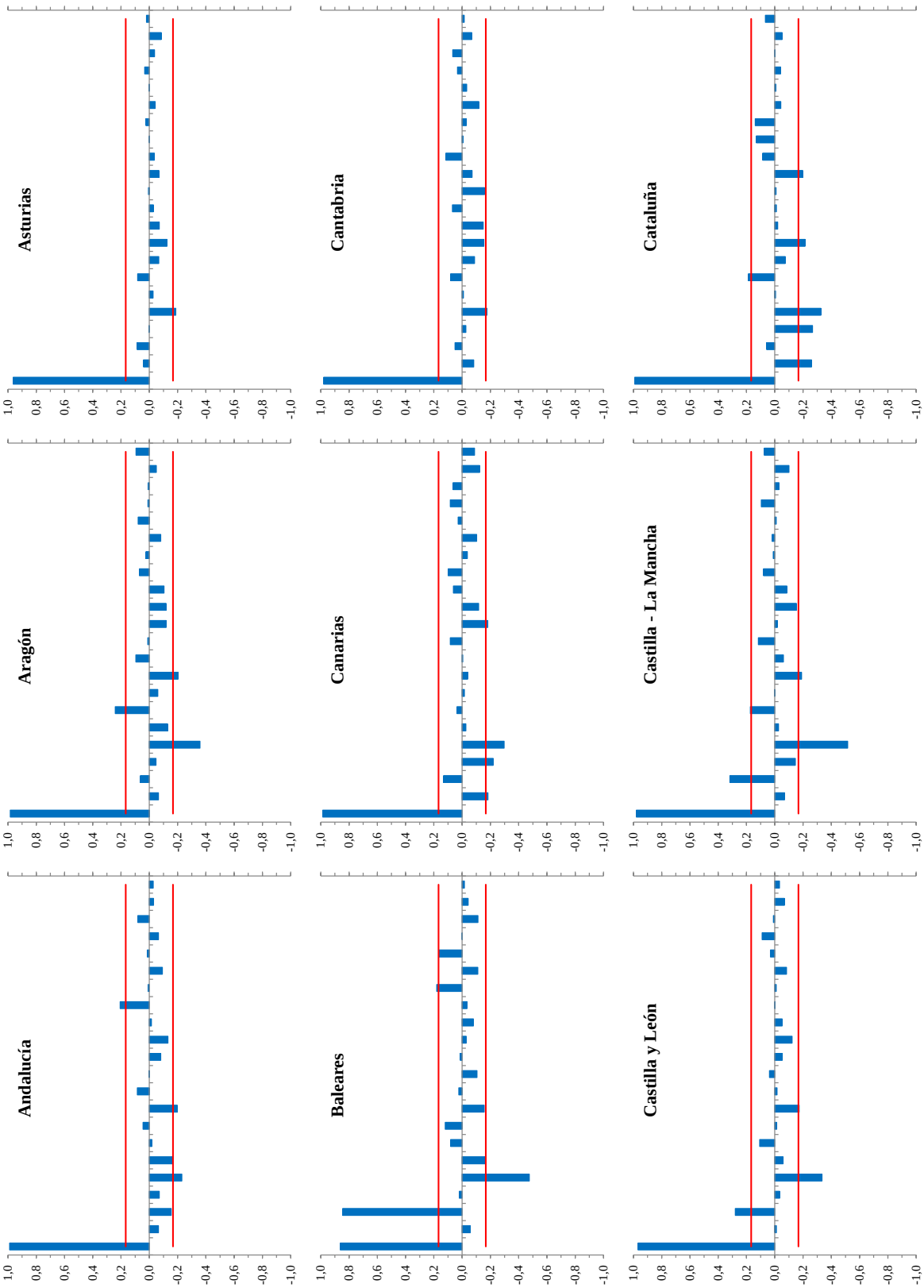

Figura A.9 Funciones de autocorrelación parcial de Andalucía, Aragón, Asturias, Baleares, Canarias, Cantabria, Castilla y León, Castilla – La Mancha y Cataluña.

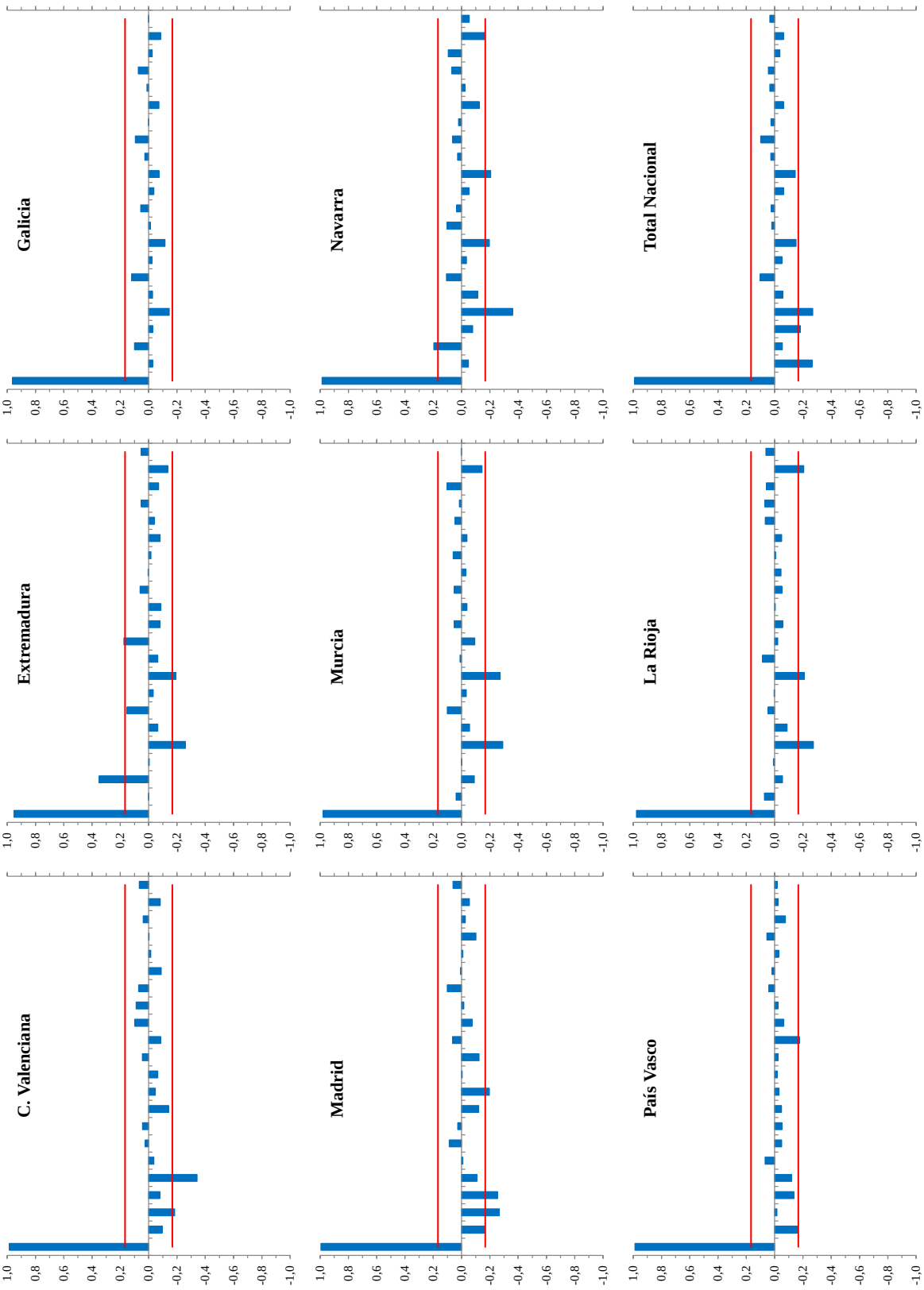

Figura A.10 Funciones de autocorrelación de C. Valenciana, Extremadura, Galicia, Madrid, Murcia, Navarra, País Vasco, La rioja y Total Nacional.

Anexo B

Análisis Espectral Singular, SSA: Figuras y Tablas

En este Anexo se presentan las siguientes figuras:

- Figura B.1 Gráficos de los 10 primeros autovectores del SSA de Andalucía.
- Figura B.2 Gráficos de los 10 primeros autovectores del SSA de Aragón.
- Figura B.3 Gráficos de los 10 primeros autovectores del SSA de Asturias.
- Figura B.4 Gráficos de los 10 primeros autovectores del SSA de Baleares.
- Figura B.5 Gráficos de los 10 primeros autovectores del SSA de Canarias.
- Figura B.6 Gráficos de los 10 primeros autovectores del SSA de Cantabria.
- Figura B.7 Gráficos de los 10 primeros autovectores del SSA de Castilla y León.
- Figura B.8 Gráficos de los 10 primeros autovectores del SSA de Castilla-La Mancha.
- Figura B.9 Gráficos de los 10 primeros autovectores del SSA de Cataluña.
- Figura B.10 Gráficos de los 10 primeros autovectores del SSA de C. Valenciana.
- Figura B.11 Gráficos de los 10 primeros autovectores del SSA de Extremadura.
- Figura B.12 Gráficos de los 10 primeros autovectores del SSA de Galicia.
- Figura B.13 Gráficos de los 10 primeros autovectores del SSA de Madrid.
- Figura B.14 Gráficos de los 10 primeros autovectores del SSA de Murcia.
- Figura B.15 Gráficos de los 10 primeros autovectores del SSA de Navarra.
- Figura B.16 Gráficos de los 10 primeros autovectores del SSA de País Vasco.
- Figura B.17 Gráficos de los 10 primeros autovectores del SSA de La Rioja.
- Figura B.18 Gráficos de los 10 primeros autovectores del SSA de Total Nacional.

Y, también, las siguientes tablas:

- Tabla B.1 Autovalores y varianza explicada del SSA de Andalucía, Aragón y Asturias.
- Tabla B.2 Autovalores y varianza explicada del SSA de Baleares, Canarias y Cantabria.

- Tabla B.3 Autovalores y varianza explicada del SSA de Castilla y León, Castilla – La Mancha y Cataluña.
- Tabla B.4 Autovalores y varianza explicada del SSA de C. Valenciana, Extremadura y Galicia.
- Tabla B.5 Autovalores y varianza explicada del SSA de Madrid, Murcia y Navarra.
- Tabla B.6 Autovalores y varianza explicada del SSA de País Vasco, La Rioja y Total Nacional.

Los autovalores que aparecen en negrita en las tablas se corresponden con los seleccionados para la reconstrucción de cada serie y el total corresponde a la suma del porcentaje de varianza explicada por cada uno de ellos.

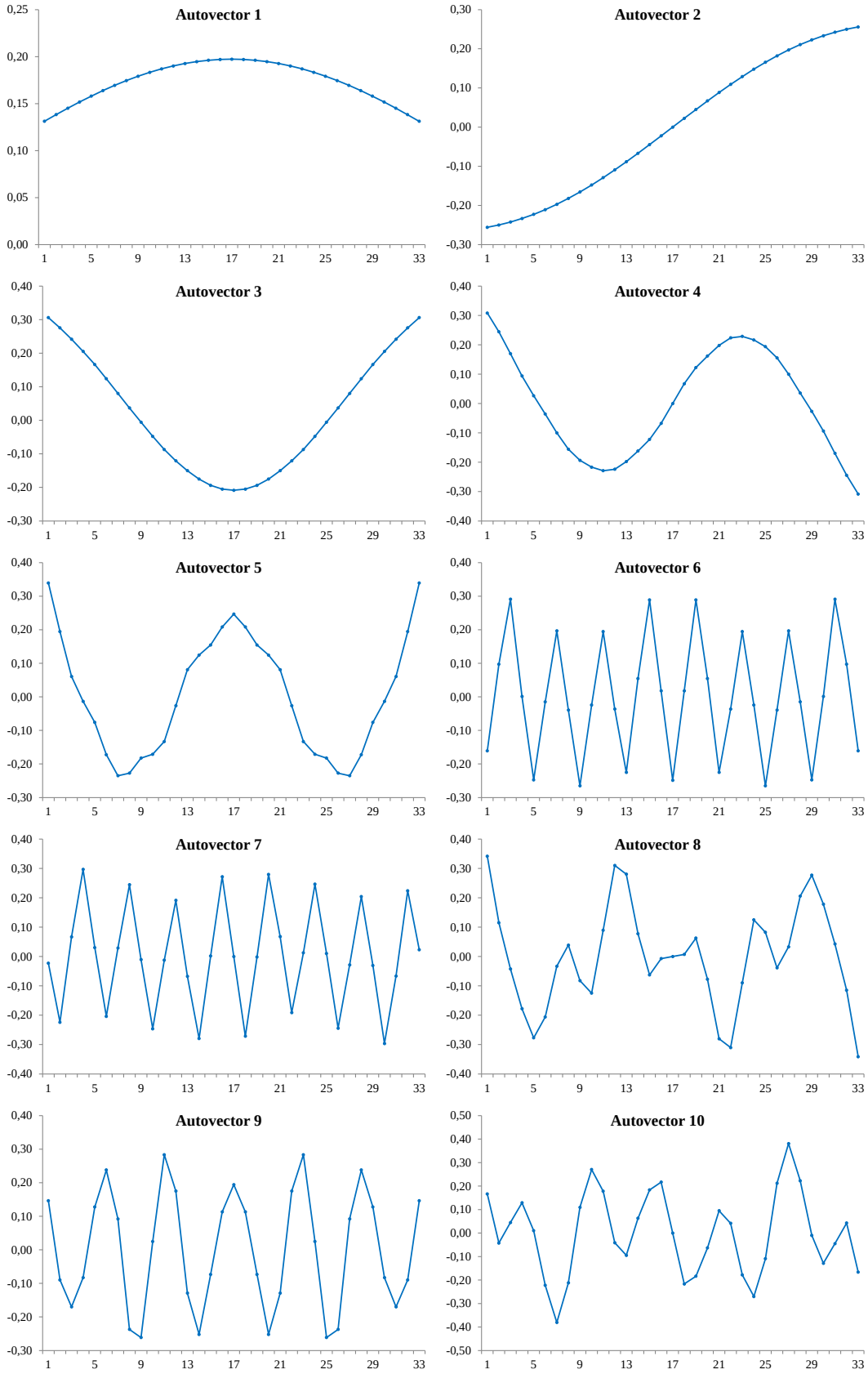

Figura B.1 Gráficos de los 10 primeros autovectores del SSA de Andalucía.

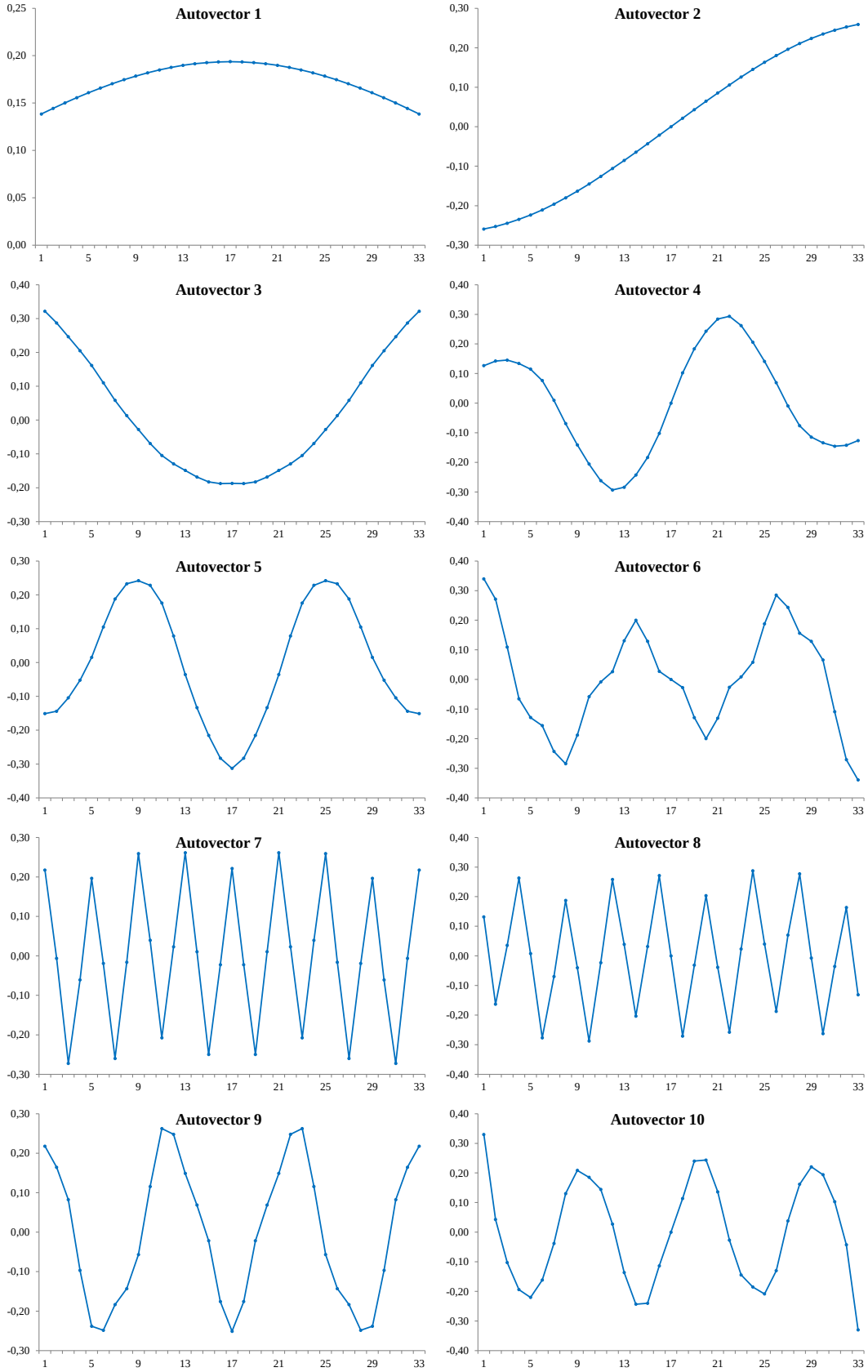

Figura B.2 Gráficos de los 10 primeros autovectores del SSA de Aragón.

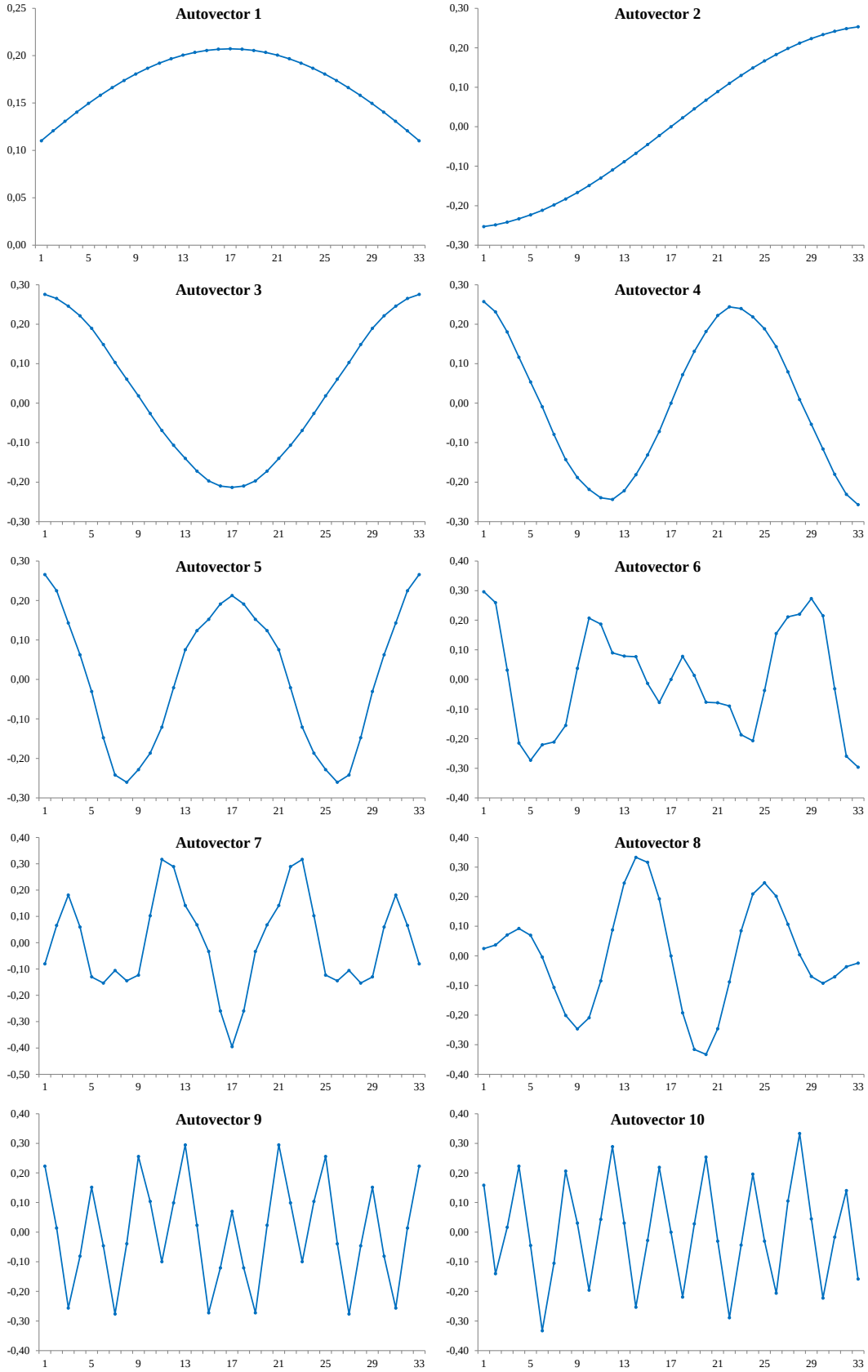

Figura B.3 Gráficos de los 10 primeros autovectores del SSA de Asturias.

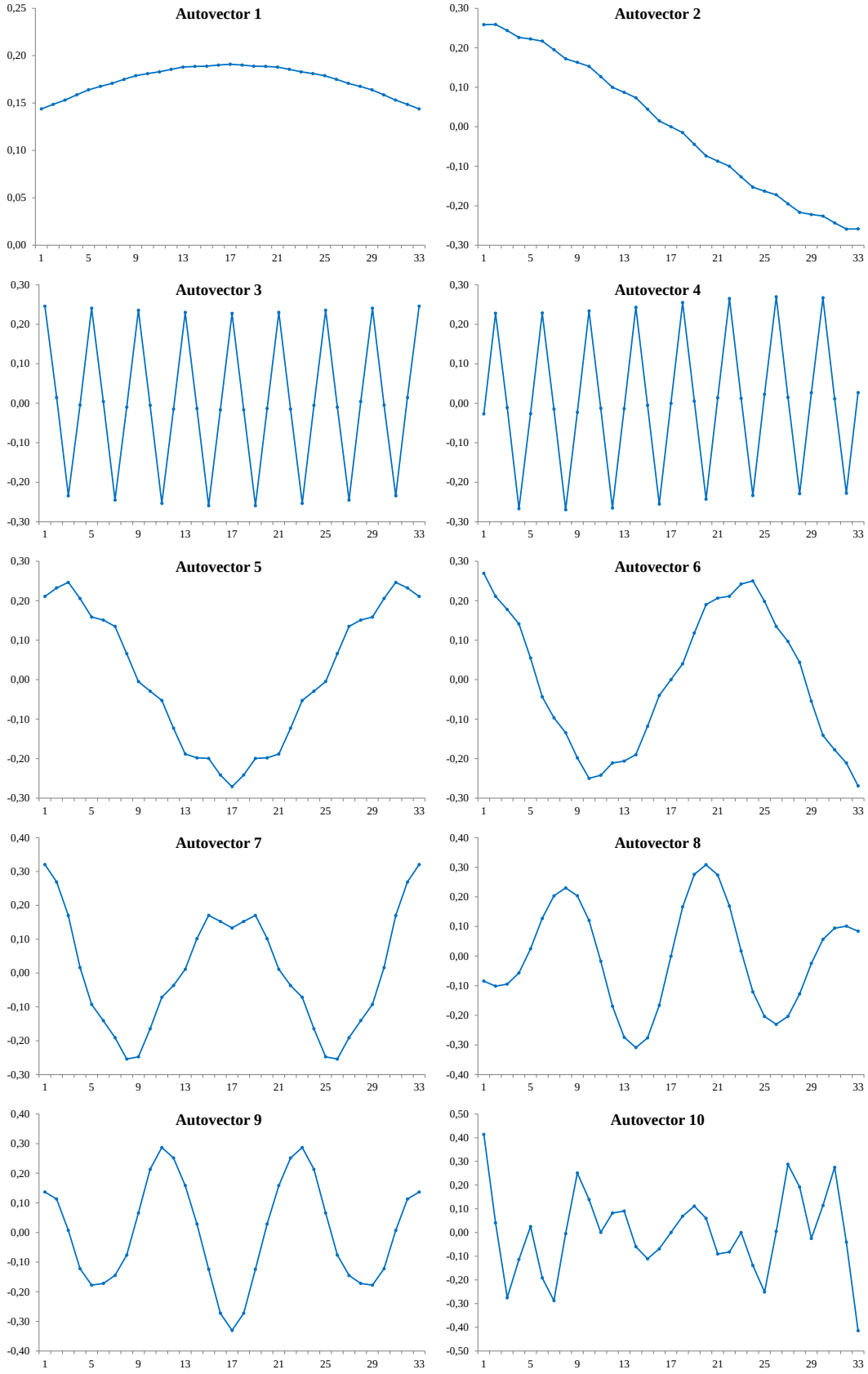

Figura B.4 Gráficos de los 10 primeros autovectores del SSA de Baleares.

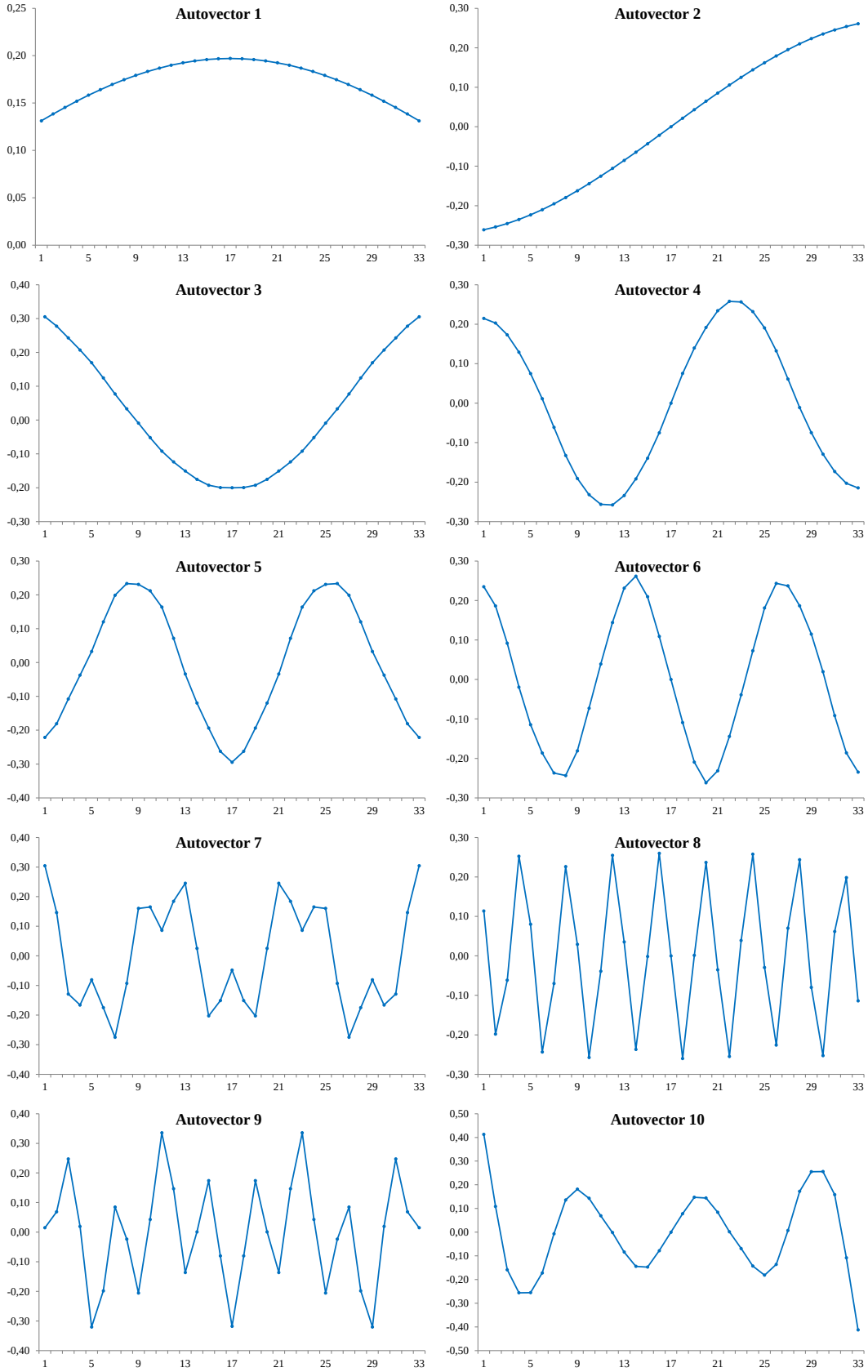

Figura B.5 Gráficos de los 10 primeros autovectores del SSA de Canarias.

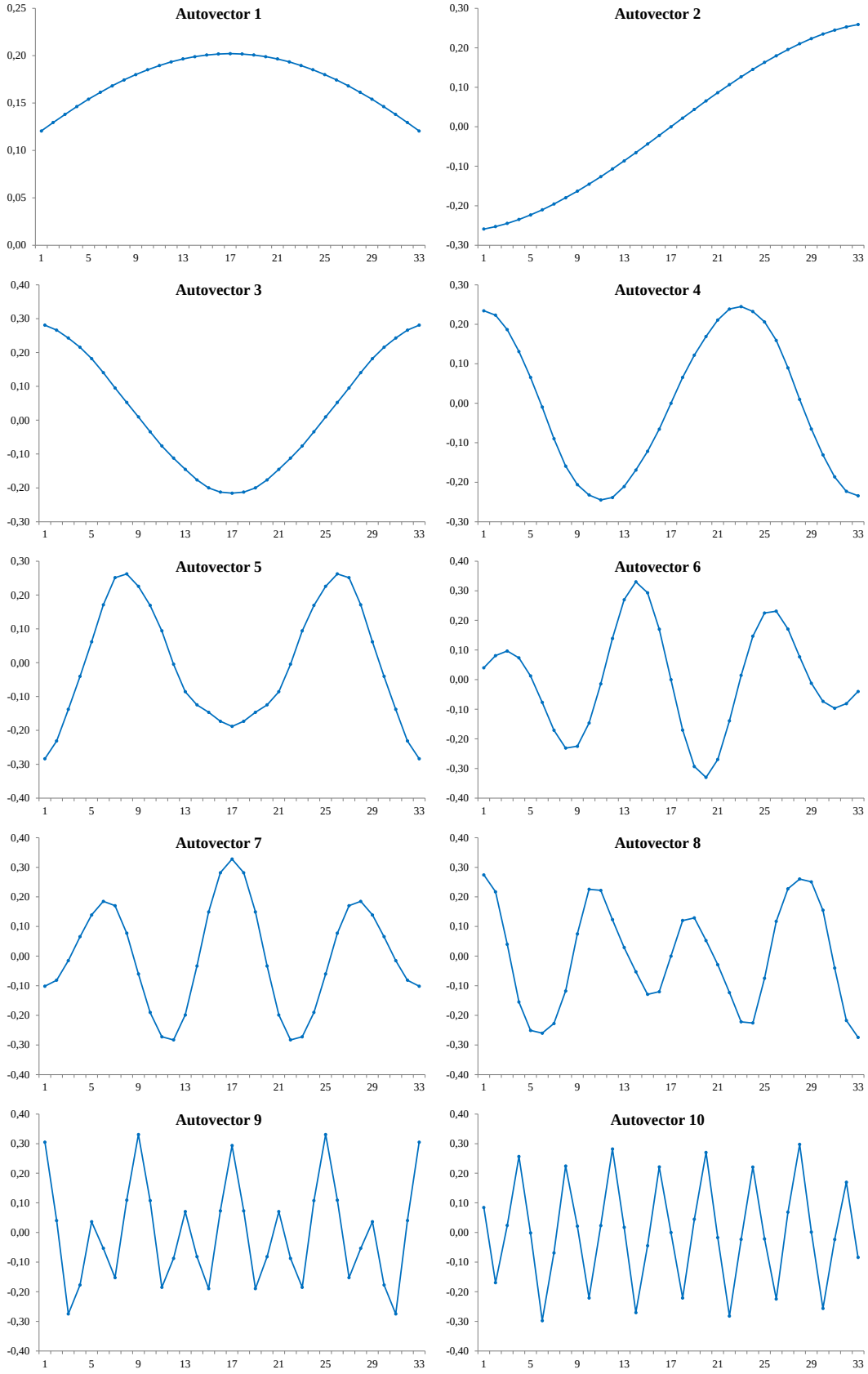

Figura B.6 Gráficos de los 10 primeros autovectores del SSA de Cantabria.

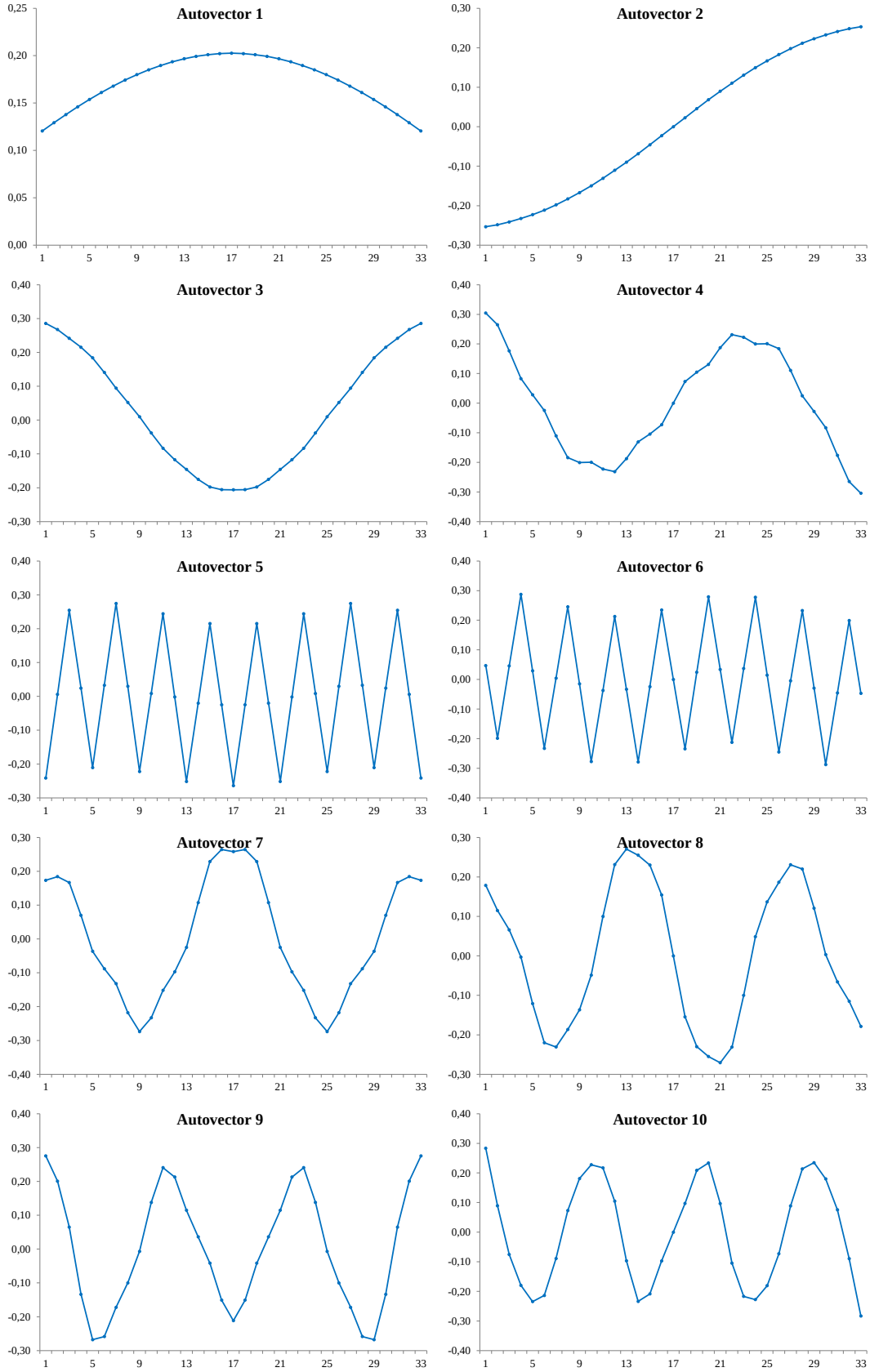

Figura B.7 Gráficos de los 10 primeros autovectores del SSA de Castilla y León.

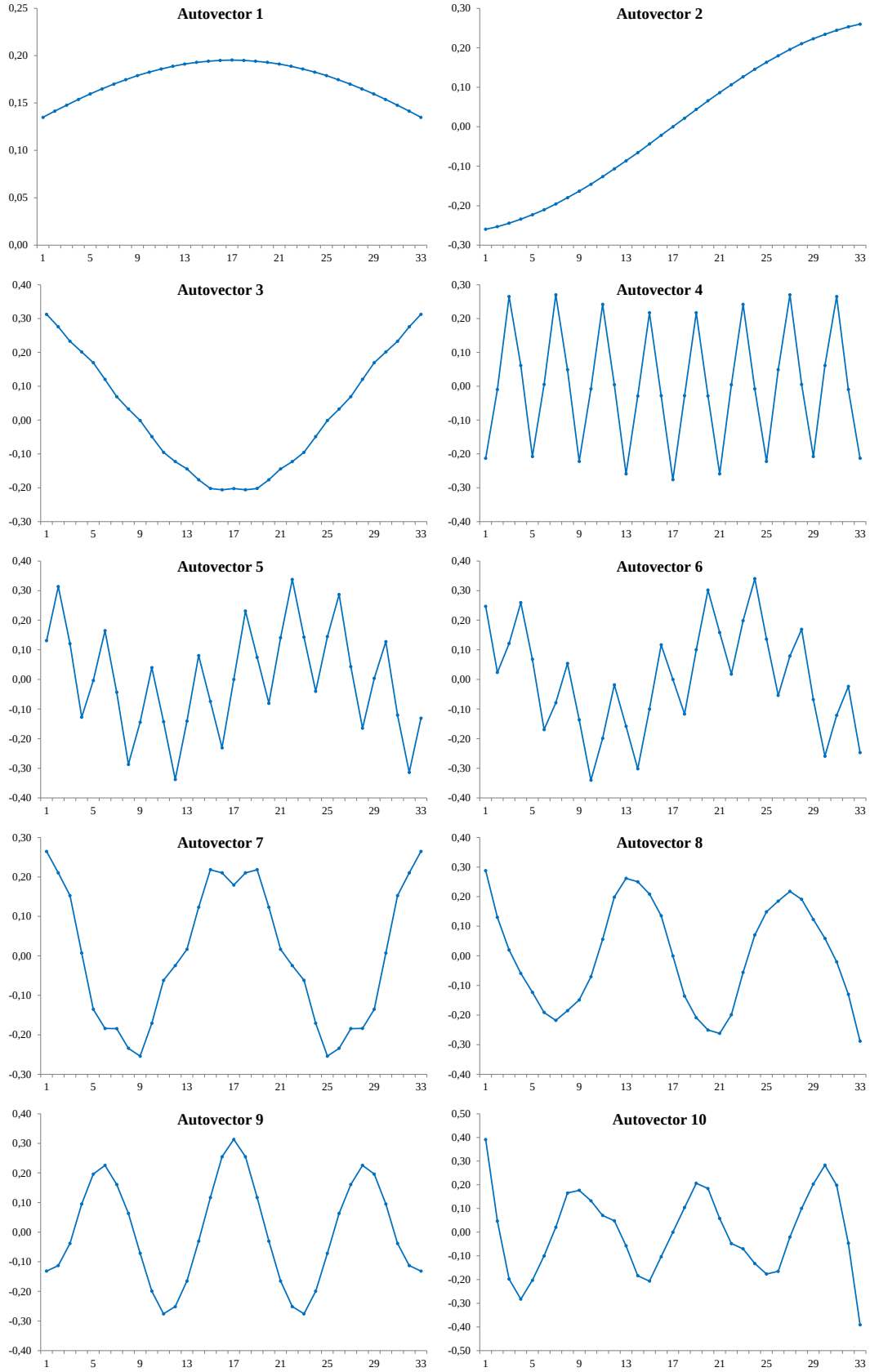

Figura B.8 Gráficos de los 10 primeros autovectores del SSA de Castilla-La Mancha.

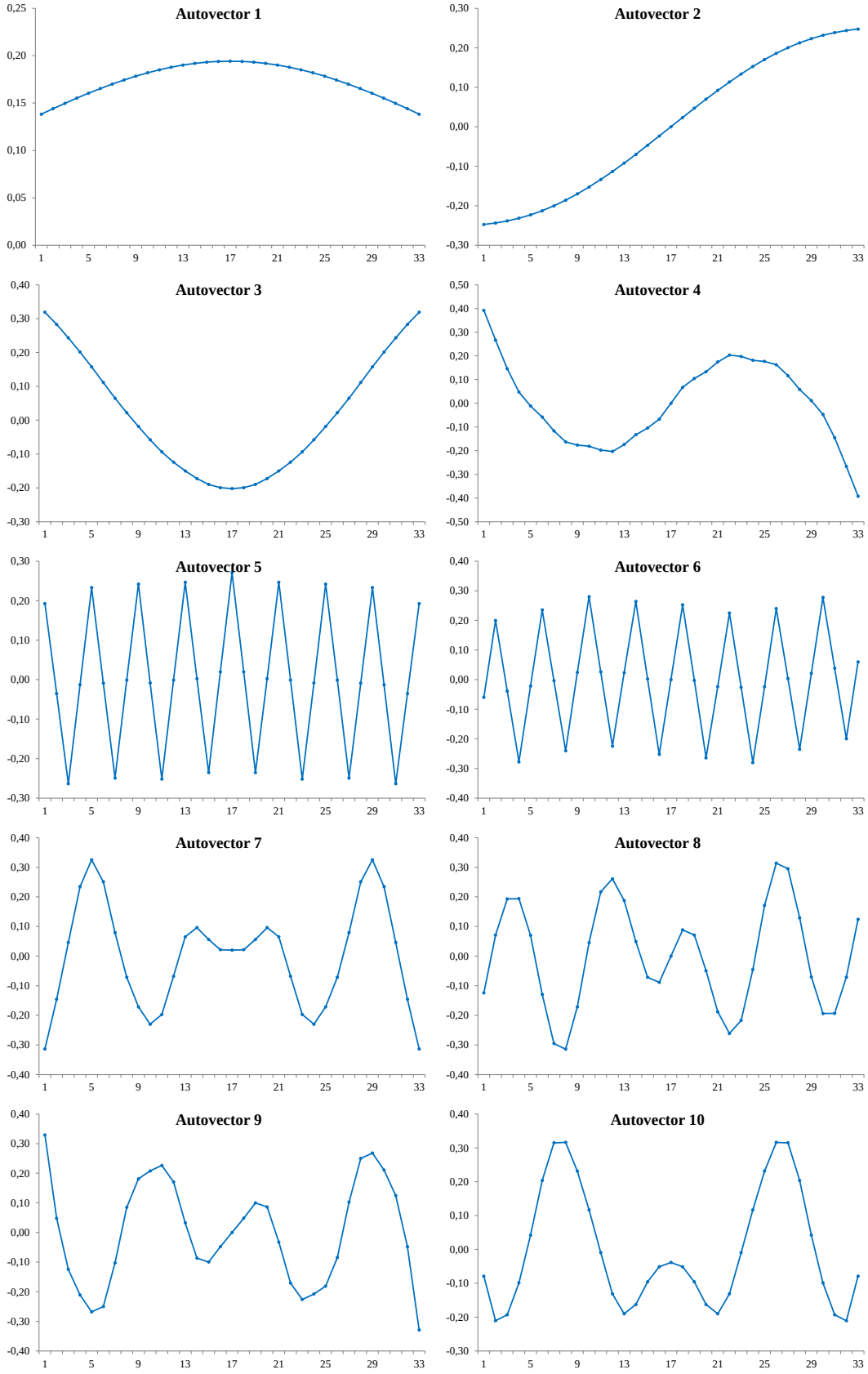

Figura B.9 Gráficos de los 10 primeros autovectores del SSA de Cataluña.

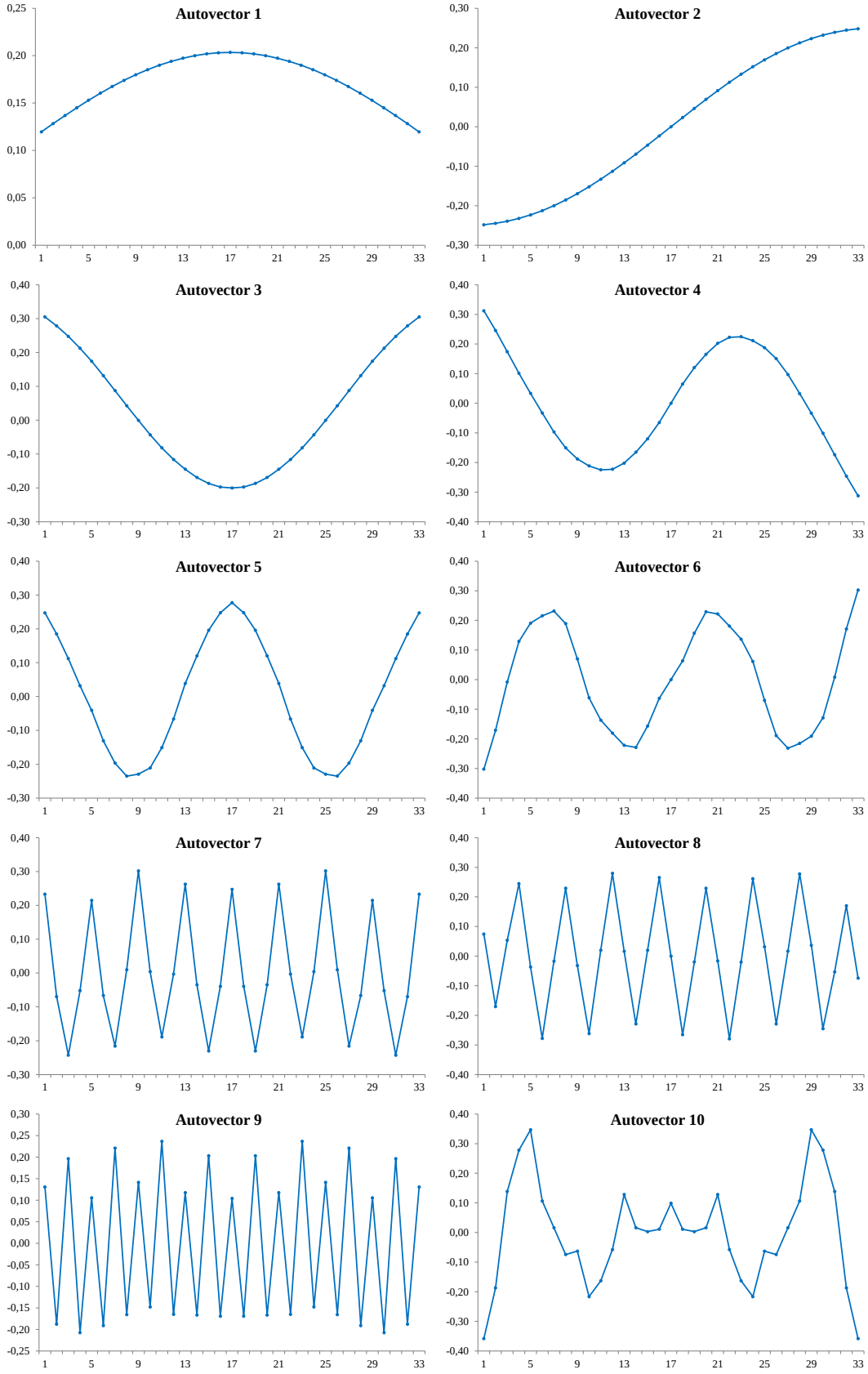

Figura B.10 Gráficos de los 10 primeros autovectores del SSA de C. Valenciana.

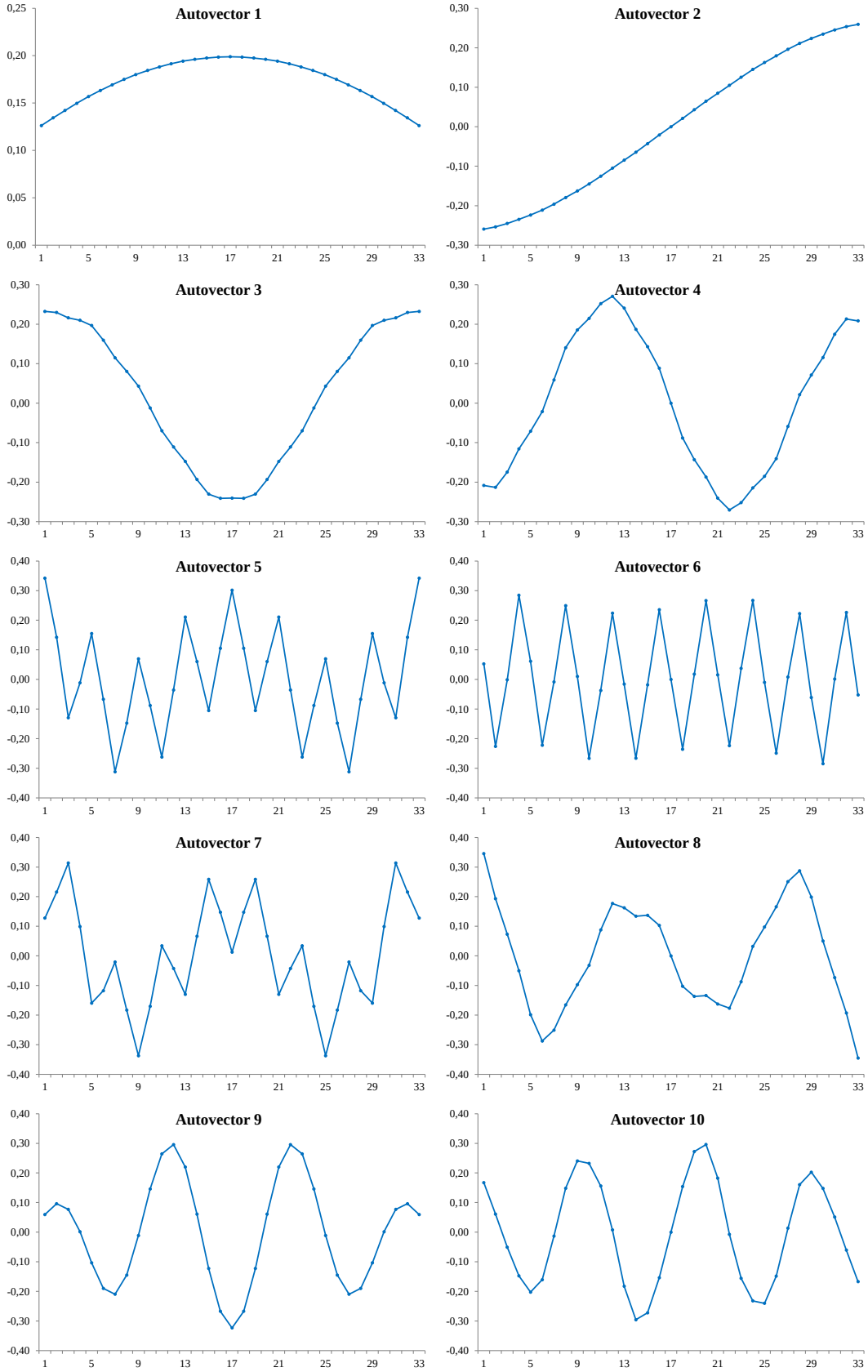

Figura B.11 Gráficos de los 10 primeros autovectores del SSA de Extremadura.

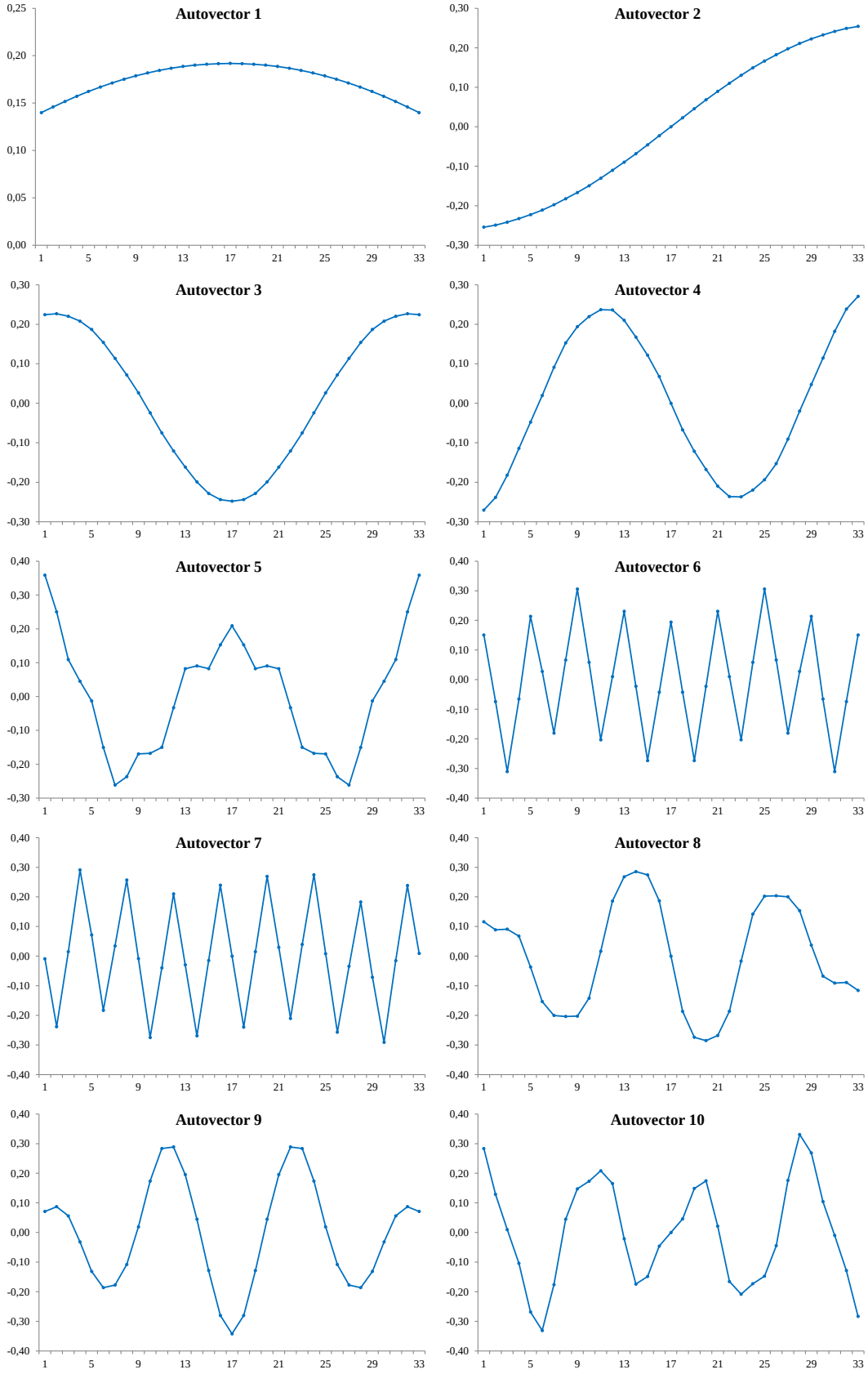

Figura B.12 Gráficos de los 10 primeros autovectores del SSA de Galicia.

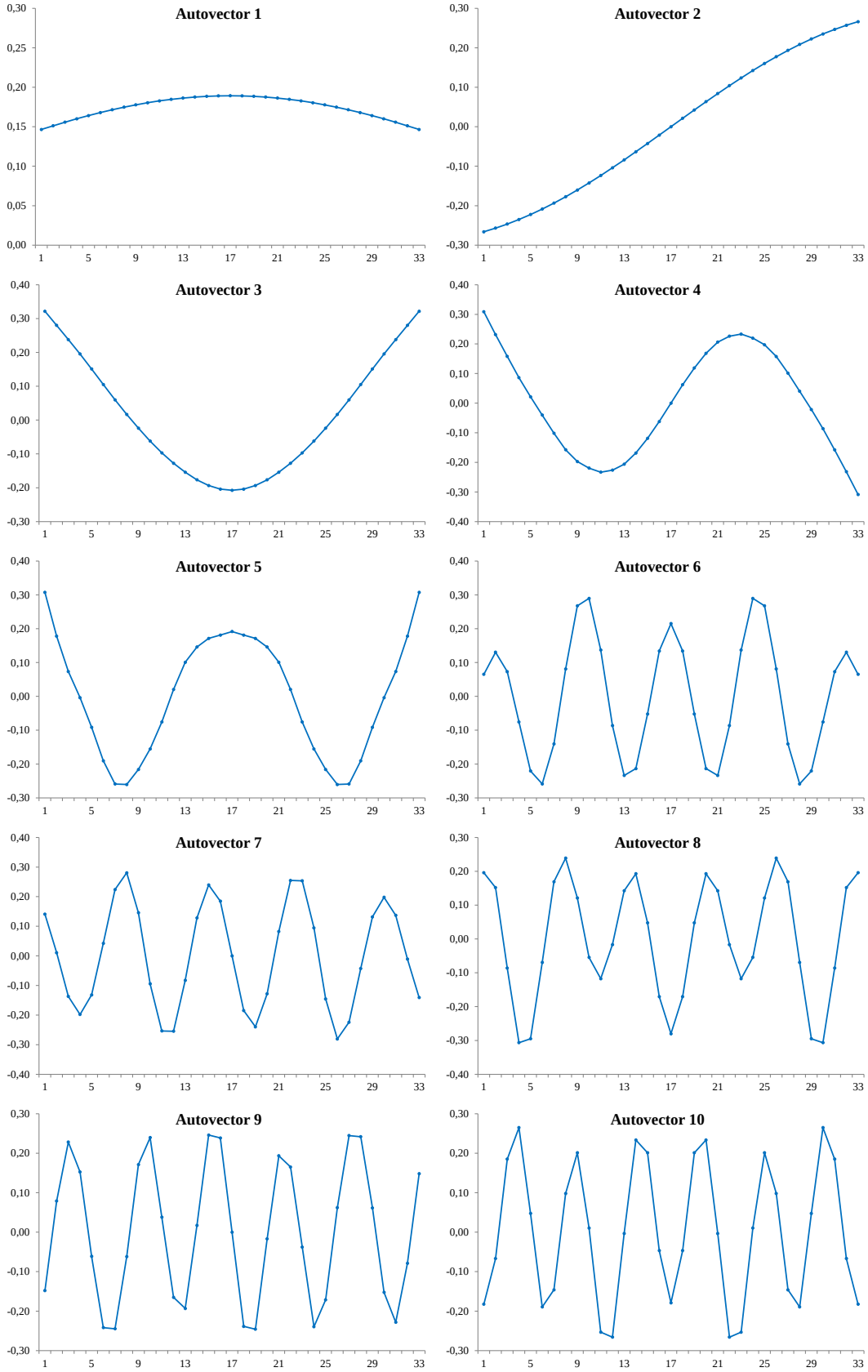

Figura B.13 Gráficos de los 10 primeros autovectores del SSA de Madrid.

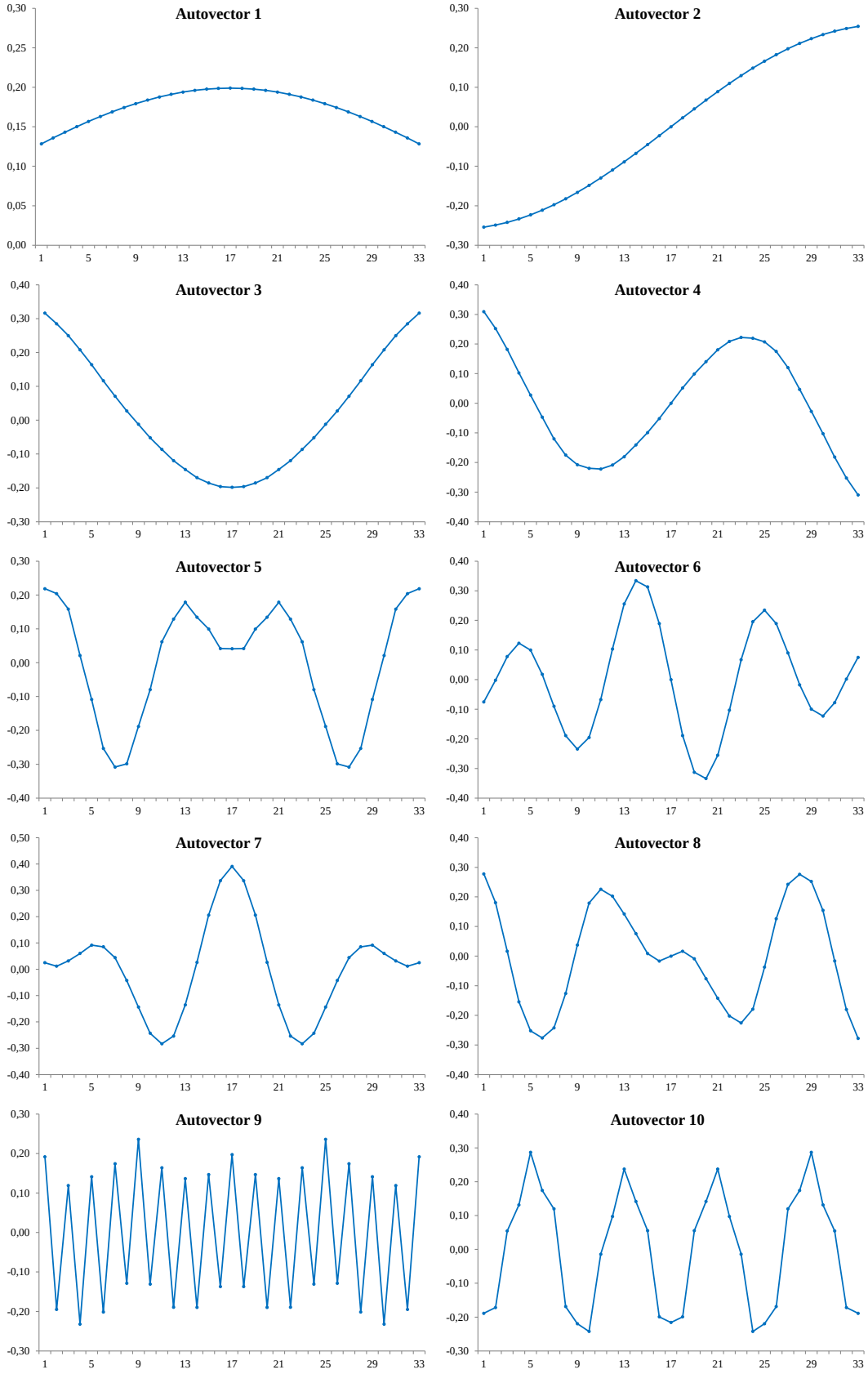

Figura B.14 Gráficos de los 10 primeros autovectores del SSA de Murcia.

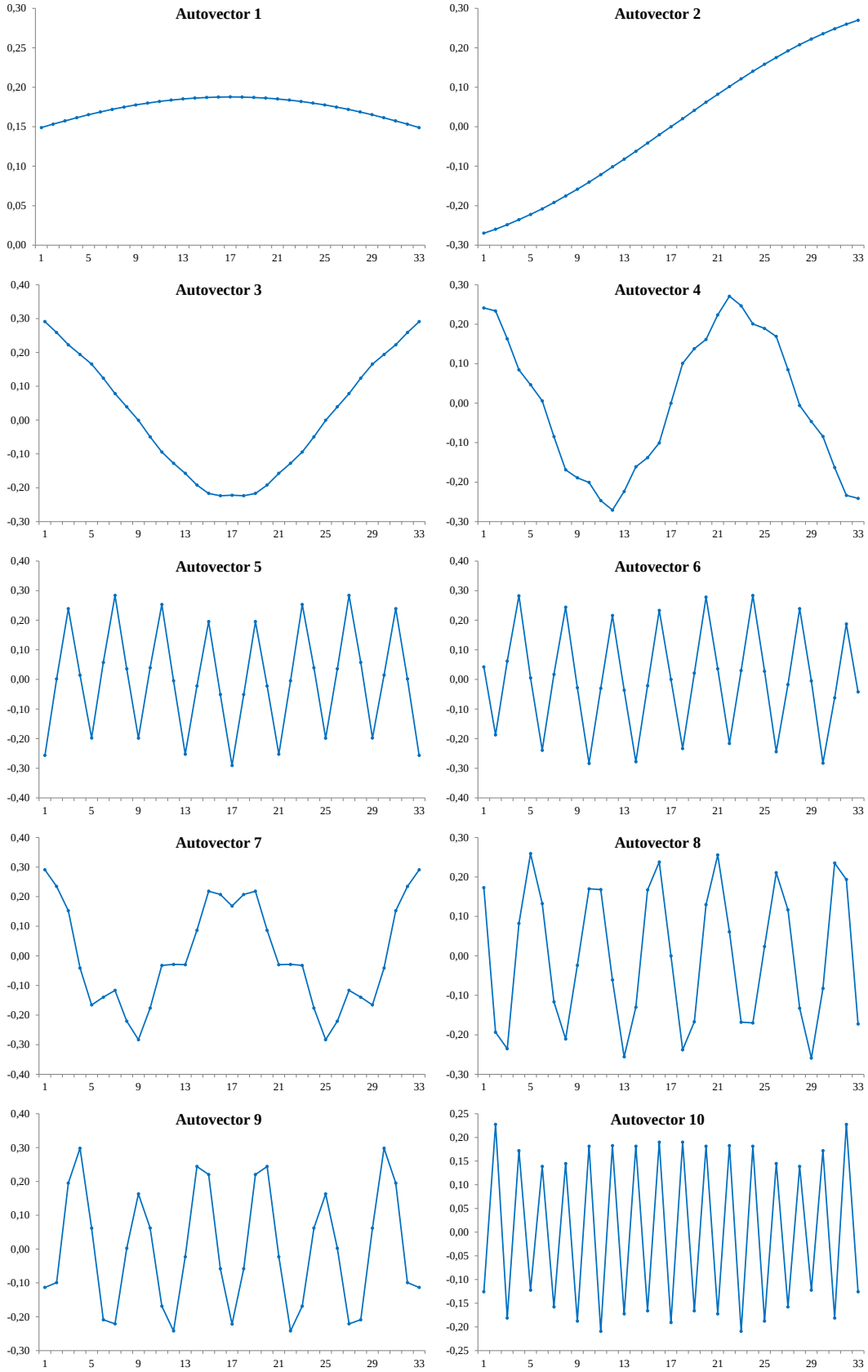

Figura B.15 Gráficos de los 10 primeros autovectores del SSA de Navarra.

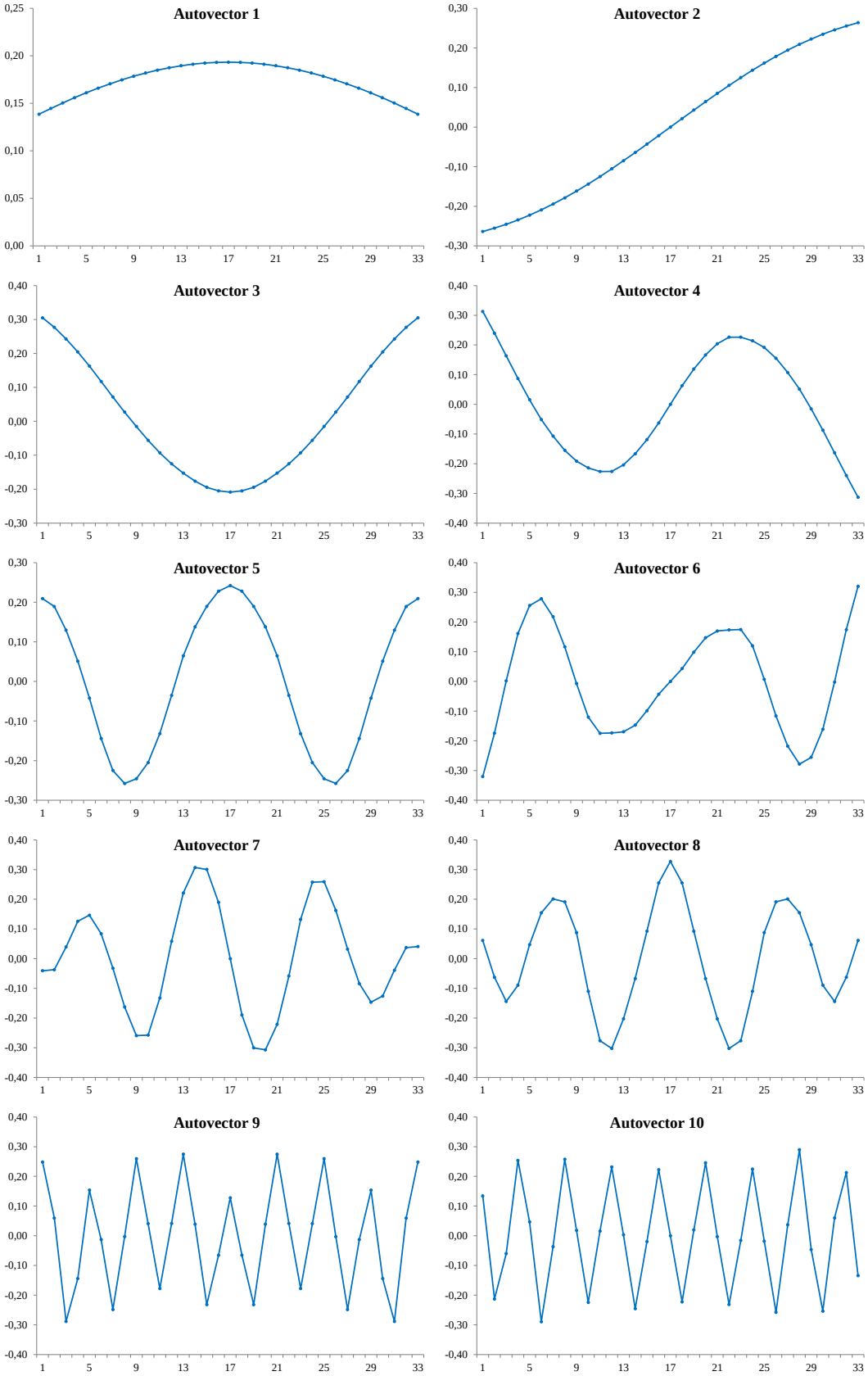

Figura B.16 Gráficos de los 10 primeros autovectores del SSA de País Vasco.

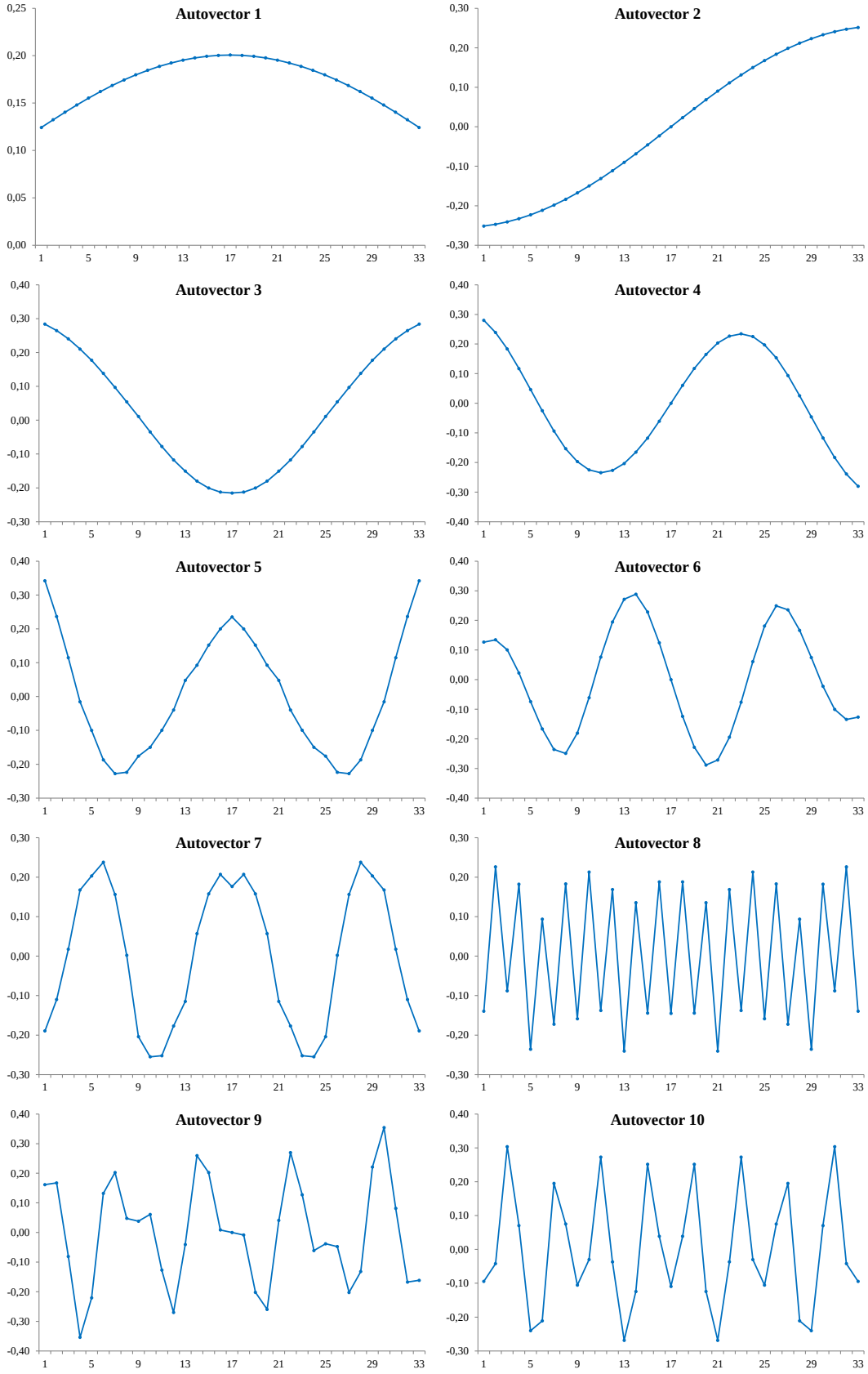

Figura B.17 Gráficos de los 10 primeros autovectores del SSA de La Rioja.

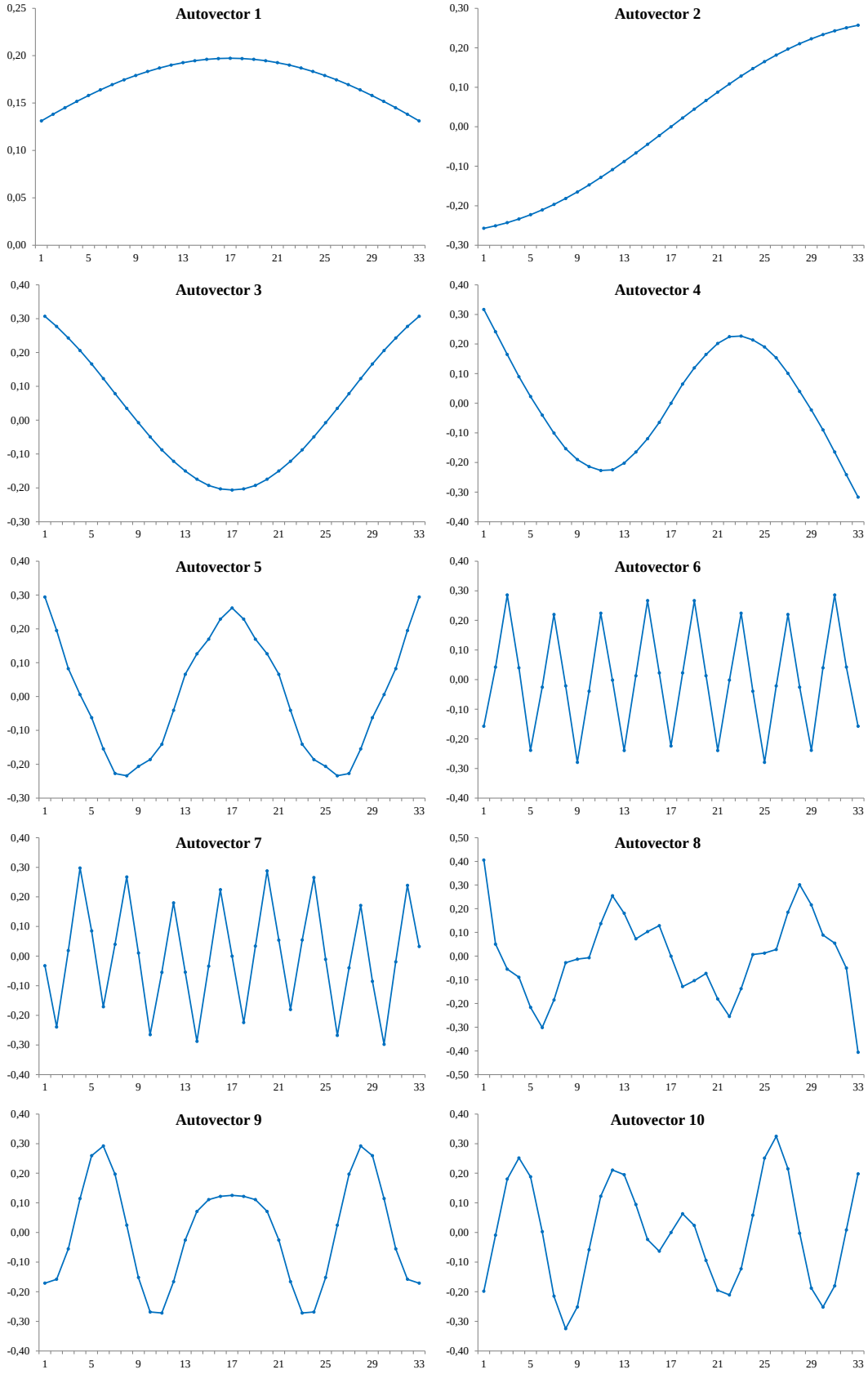

Figura B.18 Gráficos de los 10 primeros autovectores del SSA de Total Nacional.

Nº	Andalucía		Aragón		Asturias	
	Autovalor	% Varianza	Autovalor	% Varianza	Autovalor	% Varianza
1	**587,225**	**74,522**	**485,169**	**77,236**	**278,881**	**64,096**
2	**172,077**	**21,838**	**118,422**	**18,852**	**119,792**	**27,532**
3	**19,798**	**2,513**	**12,403**	**1,975**	**14,089**	**3,238**
4	**4,264**	**0,541**	**2,894**	**0,461**	**5,669**	**1,303**
5	**1,181**	**0,150**	**2,610**	**0,415**	**2,894**	**0,665**
6	0,643	0,082	**1,686**	**0,268**	1,750	0,402
7	0,619	0,079	1,291	0,205	1,555	0,357
8	0,350	0,044	1,196	0,190	**1,539**	**0,354**
9	0,267	0,034	0,681	0,108	1,468	0,337
10	0,228	0,029	0,436	0,069	1,350	0,310
11	0,134	0,017	0,181	0,029	0,811	0,186
12	0,132	0,017	0,112	0,018	0,630	0,145
	% Total	**99,56**	% Total	**99,21**	% Total	**97,19**

Tabla B.1 Autovalores y varianza explicada del SSA de Andalucía, Aragón y Asturias.

Nº	Baleares		Canarias		Cantabria	
	Autovalor	% Varianza	Autovalor	% Varianza	Autovalor	% Varianza
1	**648,393**	**69,935**	**633,117**	**74,427**	**477,050**	**68,930**
2	**138,357**	**14,923**	**182,740**	**21,482**	**171,151**	**24,730**
3	48,315	5,211	**16,959**	**1,994**	**16,602**	**2,399**
4	45,292	4,885	**5,119**	**0,602**	**6,739**	**0,974**
5	**20,452**	**2,206**	**3,430**	**0,403**	**3,773**	**0,545**
6	**10,234**	**1,104**	**2,216**	**0,261**	**2,780**	**0,402**
7	**3,579**	**0,386**	1,534	0,180	**2,765**	**0,400**
8	**2,086**	**0,225**	1,283	0,151	**2,124**	**0,307**
9	2,024	0,218	1,229	0,144	1,603	0,232
10	1,415	0,153	0,652	0,077	1,419	0,205
11	1,135	0,122	0,366	0,043	1,297	0,187
12	1,112	0,120	0,353	0,042	0,701	0,101
	% Total	**88,78**	% Total	**99,17**	% Total	**98,69**

Tabla B.2 Autovalores y varianza explicada del SSA de Baleares, Canarias y Cantabria.

Nº	Castilla y León		Castilla - La Mancha		Cataluña	
	Autovalor	% Varianza	Autovalor	% Varianza	Autovalor	% Varianza
1	**287,193**	**68,456**	**407,381**	**75,709**	**723,399**	**76,671**
2	**104,704**	**24,958**	**108,481**	**20,160**	**183,596**	**19,459**
3	**13,108**	**3,124**	**10,983**	**2,041**	**28,501**	**3,021**
4	**4,115**	**0,981**	2,827	0,525	**3,148**	**0,334**
5	2,277	0,543	2,802	0,521	1,052	0,112
6	2,120	0,505	2,515	0,467	0,991	0,105
7	**1,371**	**0,327**	**1,089**	**0,202**	0,444	0,047
8	**1,104**	**0,263**	0,390	0,072	0,342	0,036
9	0,759	0,181	0,305	0,057	0,294	0,031
10	0,428	0,102	0,242	0,045	0,276	0,029
11	0,328	0,078	0,138	0,026	0,239	0,025
12	0,269	0,064	0,087	0,016	0,204	0,022
	% Total	**98,11**	% Total	**98,11**	% Total	**99,48**

Tabla B.3 Autovalores y varianza explicada del SSA de Castilla y León, Castilla – La Mancha y Cataluña.

Nº	C. Valenciana		Extremadura		Galicia	
	Autovalor	% Varianza	Autovalor	% Varianza	Autovalor	% Varianza
1	**468,431**	**68,989**	**352,036**	**68,333**	**370,284**	**74,079**
2	**175,582**	**25,859**	**113,898**	**22,109**	**88,335**	**17,672**
3	**24,184**	**3,562**	**14,026**	**2,723**	**16,121**	**3,225**
4	**5,290**	**0,779**	**8,477**	**1,645**	**8,497**	**1,700**
5	**1,801**	**0,265**	5,259	1,021	**2,881**	**0,576**
6	**0,997**	**0,147**	4,693	0,911	2,001	0,400
7	0,557	0,082	4,332	0,841	1,940	0,388
8	0,513	0,076	**1,923**	**0,373**	**1,487**	**0,297**
9	0,428	0,063	**1,555**	**0,302**	**1,445**	**0,289**
10	0,204	0,030	**1,528**	**0,297**	**1,302**	**0,260**
11	0,123	0,018	0,962	0,187	0,857	0,171
12	0,119	0,018	0,868	0,168	0,671	0,134
	% Total	**99,60**	% Total	**95,78**	% Total	**98,10**

Tabla B.4 Autovalores y varianza explicada del SSA de C. Valenciana, Extremadura y Galicia.

Nº	Madrid		Murcia		Navarra	
	Autovalor	% Varianza	Autovalor	% Varianza	Autovalor	% Varianza
1	**991,408**	**82,877**	**636,549**	**72,468**	**745,742**	**83,476**
2	**184,466**	**15,421**	**199,448**	**22,706**	**125,855**	**14,088**
3	**15,005**	**1,254**	**23,430**	**2,667**	**9,756**	**1,092**
4	**2,118**	**0,177**	**4,376**	**0,498**	**3,369**	**0,377**
5	0,564	0,047	**2,007**	**0,228**	2,287	0,256
6	0,381	0,032	**1,787**	**0,203**	2,115	0,237
7	0,380	0,032	**1,764**	**0,201**	1,429	0,160
8	0,251	0,021	**1,495**	**0,170**	0,455	0,051
9	0,234	0,020	1,247	0,142	0,454	0,051
10	0,199	0,017	1,014	0,115	0,263	0,029
11	0,181	0,015	0,853	0,097	0,244	0,027
12	0,114	0,010	0,680	0,077	0,136	0,015
	% Total	**99,73**	% Total	**99,14**	% Total	**99,03**

Tabla B.5 Autovalores y varianza explicada del SSA de Madrid, Murcia y Navarra.

Nº	País Vasco		La Rioja		Total Nacional	
	Autovalor	% Varianza	Autovalor	% Varianza	Autovalor	% Varianza
1	**622,831**	**78,484**	**563,191**	**69,736**	**502,116**	**74,845**
2	**149,030**	**18,779**	**192,894**	**23,885**	**146,639**	**21,858**
3	**13,196**	**1,663**	**26,709**	**3,307**	**15,881**	**2,367**
4	**2,911**	**0,367**	**8,811**	**1,091**	**3,237**	**0,483**
5	**1,125**	**0,142**	**3,187**	**0,395**	0,848	0,126
6	0,755	0,095	**1,501**	**0,186**	0,364	0,054
7	0,473	0,060	**1,400**	**0,173**	0,356	0,053
8	0,470	0,059	1,009	0,125	0,265	0,040
9	0,345	0,044	0,986	0,122	0,156	0,023
10	0,326	0,041	0,980	0,121	0,140	0,021
11	0,232	0,029	0,960	0,119	0,106	0,016
12	0,205	0,026	0,883	0,109	0,097	0,014
	% Total	**99,43**	% Total	**98,77**	% Total	**99,55**

Tabla B.6 Autovalores y varianza explicada del SSA de País Vasco, La Rioja y Total Nacional.

Anexo C

Series Filtradas de Ruido: Figuras

En este Anexo se presentan las siguientes figuras:

- Figura C.1 Series filtradas de ruido de Andalucía, Aragón y Asturias.
- Figura C.2 Series filtradas de ruido de Baleares, Canarias y Cantabria.
- Figura C.3 Series filtradas de ruido de Castilla y León, Castilla – La Mancha y Cataluña.
- Figura C.4 Series filtradas de ruido de C. Valenciana, Extremadura y Galicia.
- Figura C.5 Series filtradas de ruido de Madrid, Murcia y Navarra.
- Figura C.6 Series filtradas de ruido de País Vasco, La rioja y Total Nacional.
- Figura C.7 Funciones de autocorrelación series filtradas: Andalucía, Aragón, Asturias, Baleares, Canarias, Cantabria, Castilla y León, Castilla – La Mancha y Cataluña.
- Figura C.8 Funciones de autocorrelación series filtradas: C. Valenciana, Extremadura, Galicia, Madrid, Murcia, Navarra, País Vasco, La rioja y Total Nacional.
- Figura C.9 Funciones de autocorrelación parcial series filtradas: Andalucía, Aragón, Asturias, Baleares, Canarias, Cantabria, Castilla y León, Castilla – La Mancha y Cataluña.
- Figura C.10 Funciones de autocorrelación parcial series filtradas: C. Valenciana, Extremadura, Galicia, Madrid, Murcia, Navarra, País Vasco, La rioja y Total Nacional.

En las Figuras C.1 a C.6, las series filtradas (azul) se han superpuesto sobre las series originales (rojo).

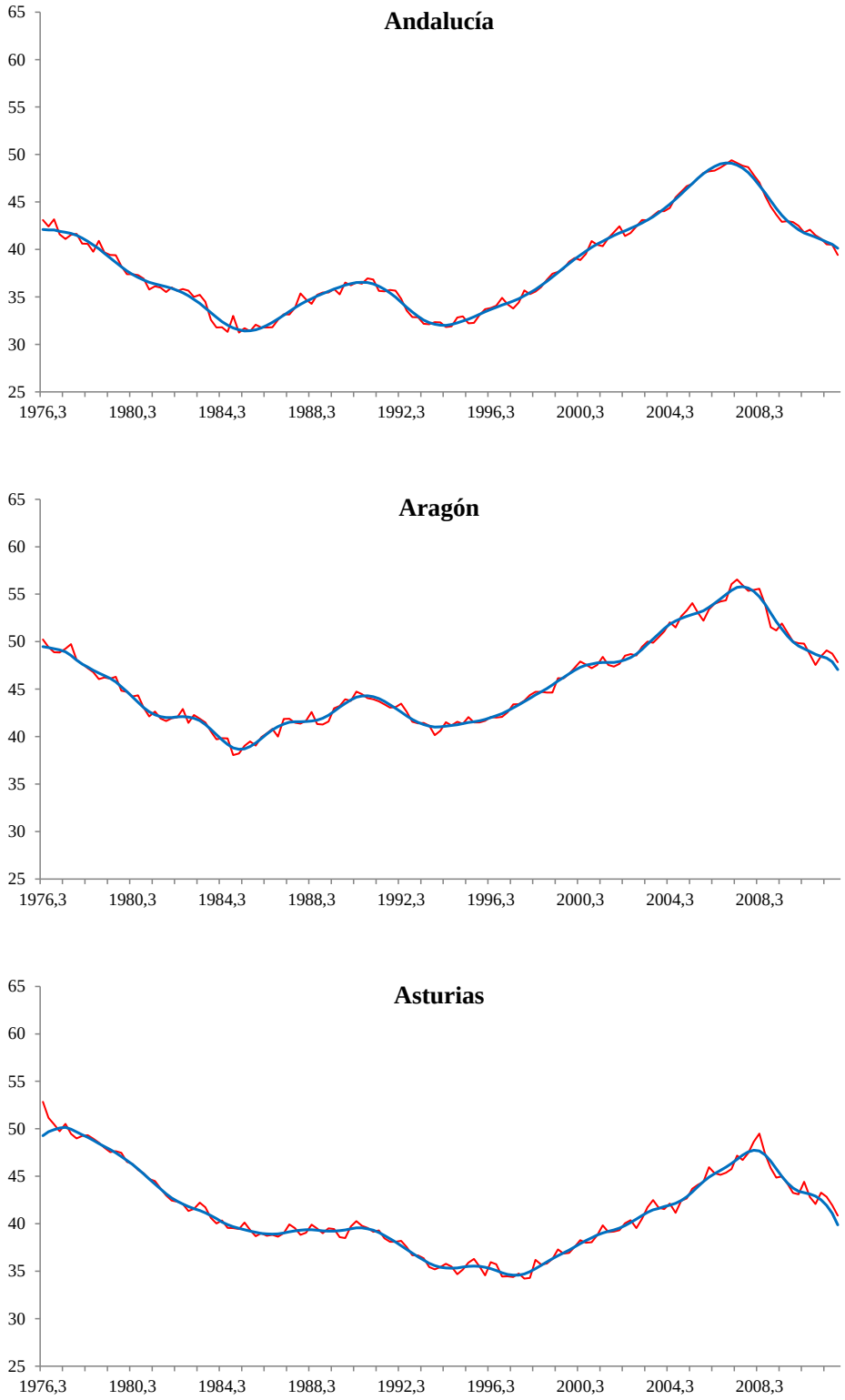

Figura C.1 Series filtradas de ruido de Andalucía, Aragón y Asturias.

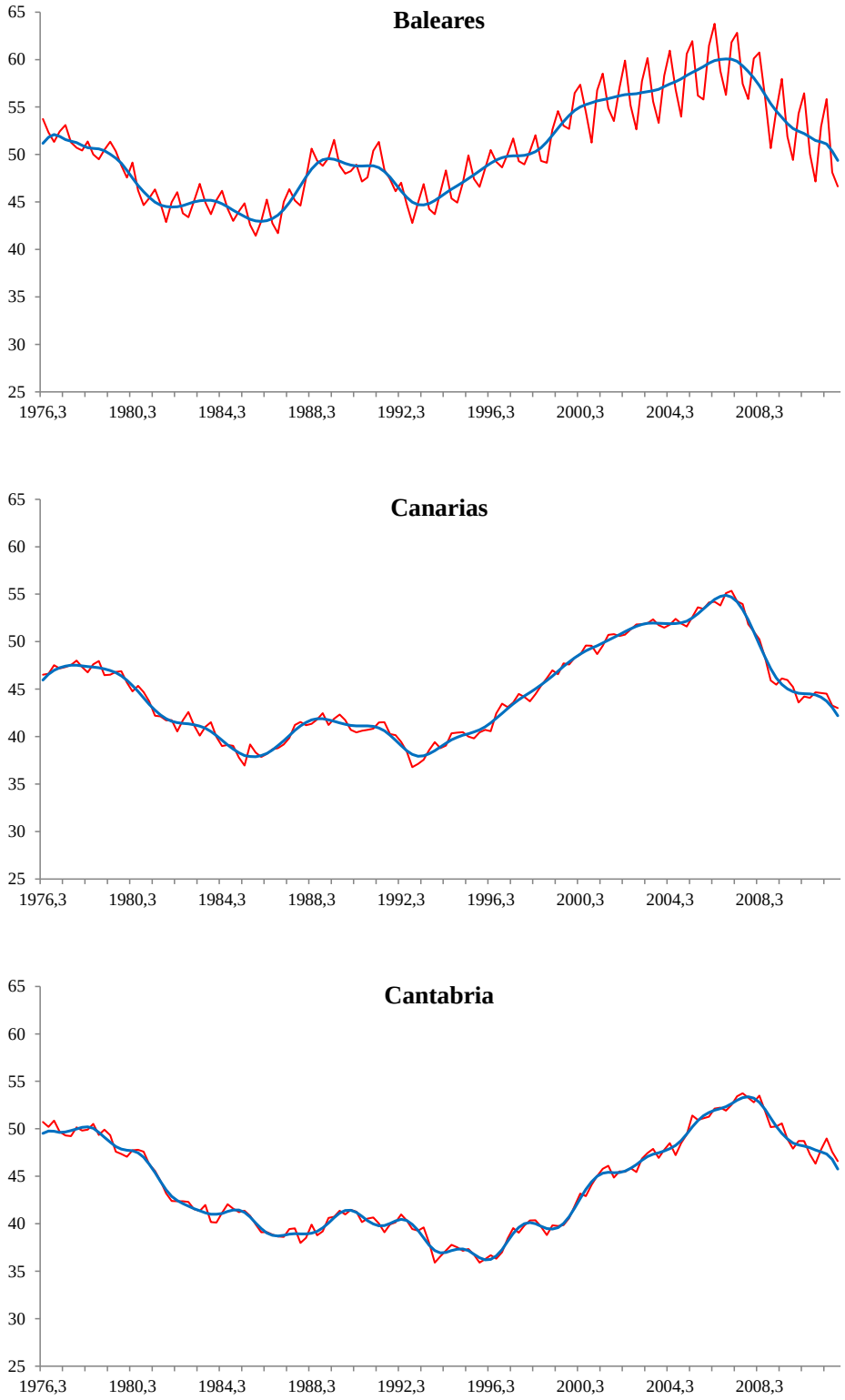

Figura C.2 Series filtradas de ruido de Baleares, Canarias y Cantabria.

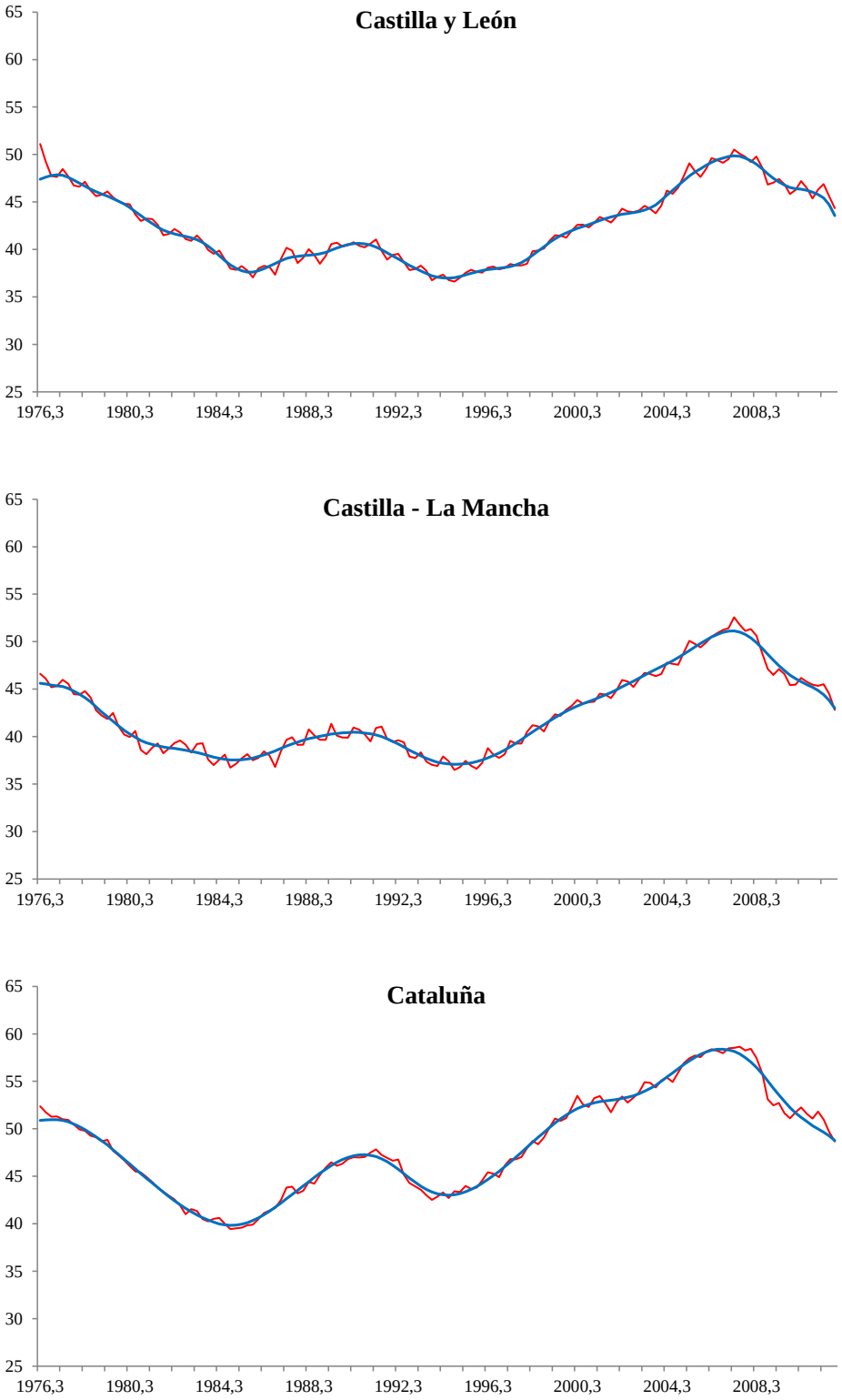

Figura C.3 Series filtradas de ruido de Castilla y León, Castilla – La Mancha y Cataluña.

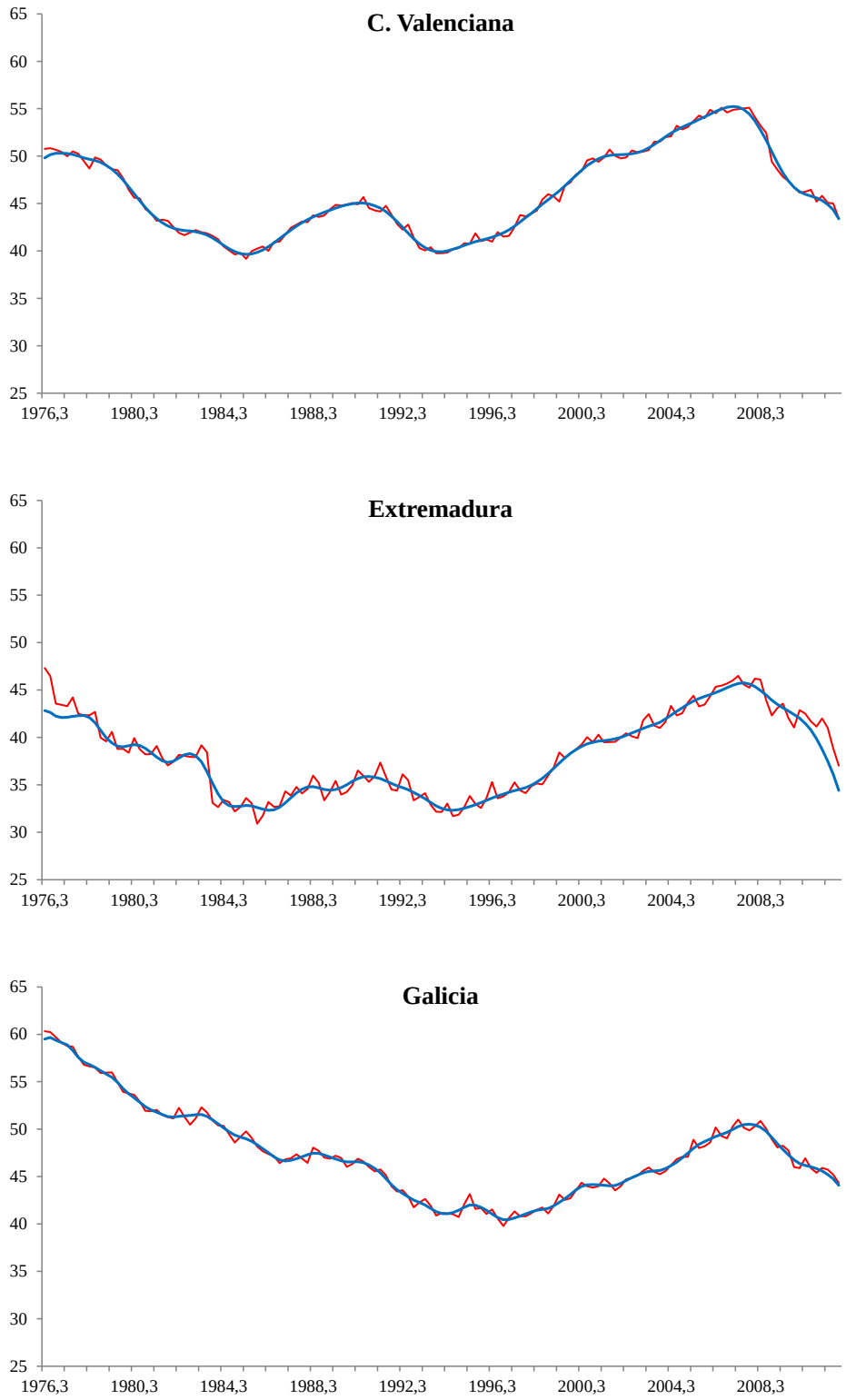

Figura C.4 Series filtradas de ruido de C. Valenciana, Extremadura y Galicia.

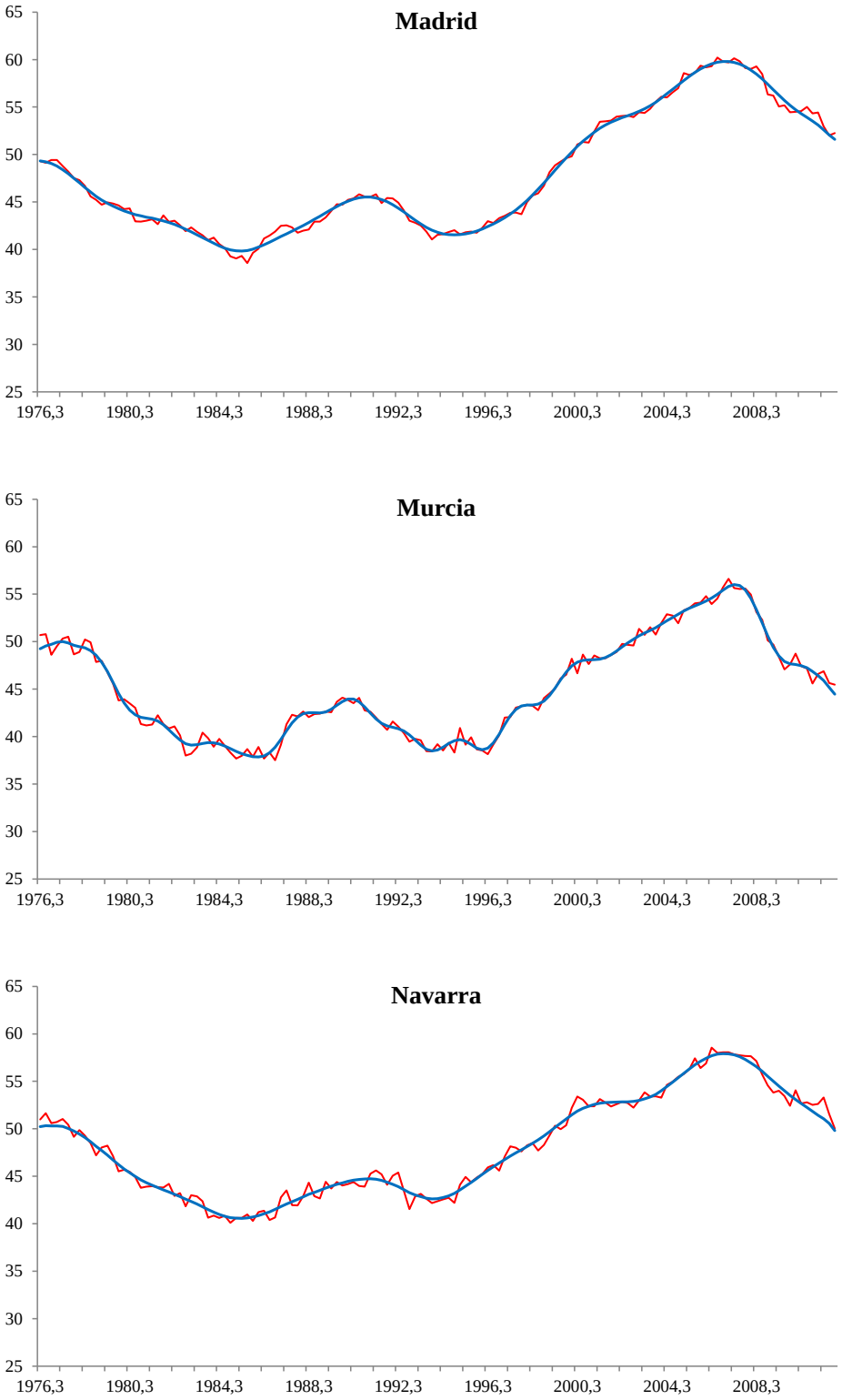

Figura C.5 Series filtradas de ruido de Madrid, Murcia y Navarra.

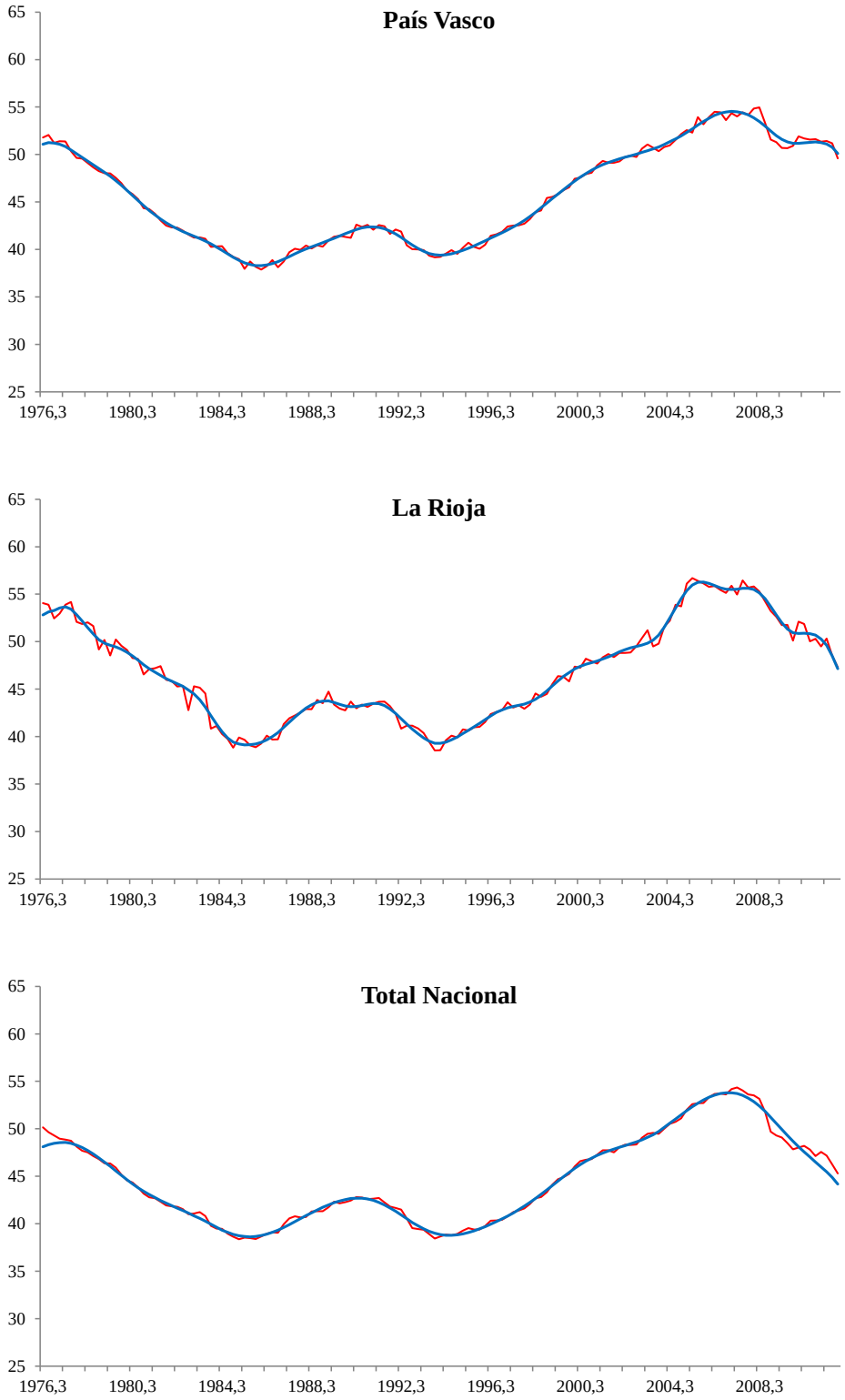

Figura C.6 Series filtradas de ruido de País Vasco, La rioja y Total Nacional.

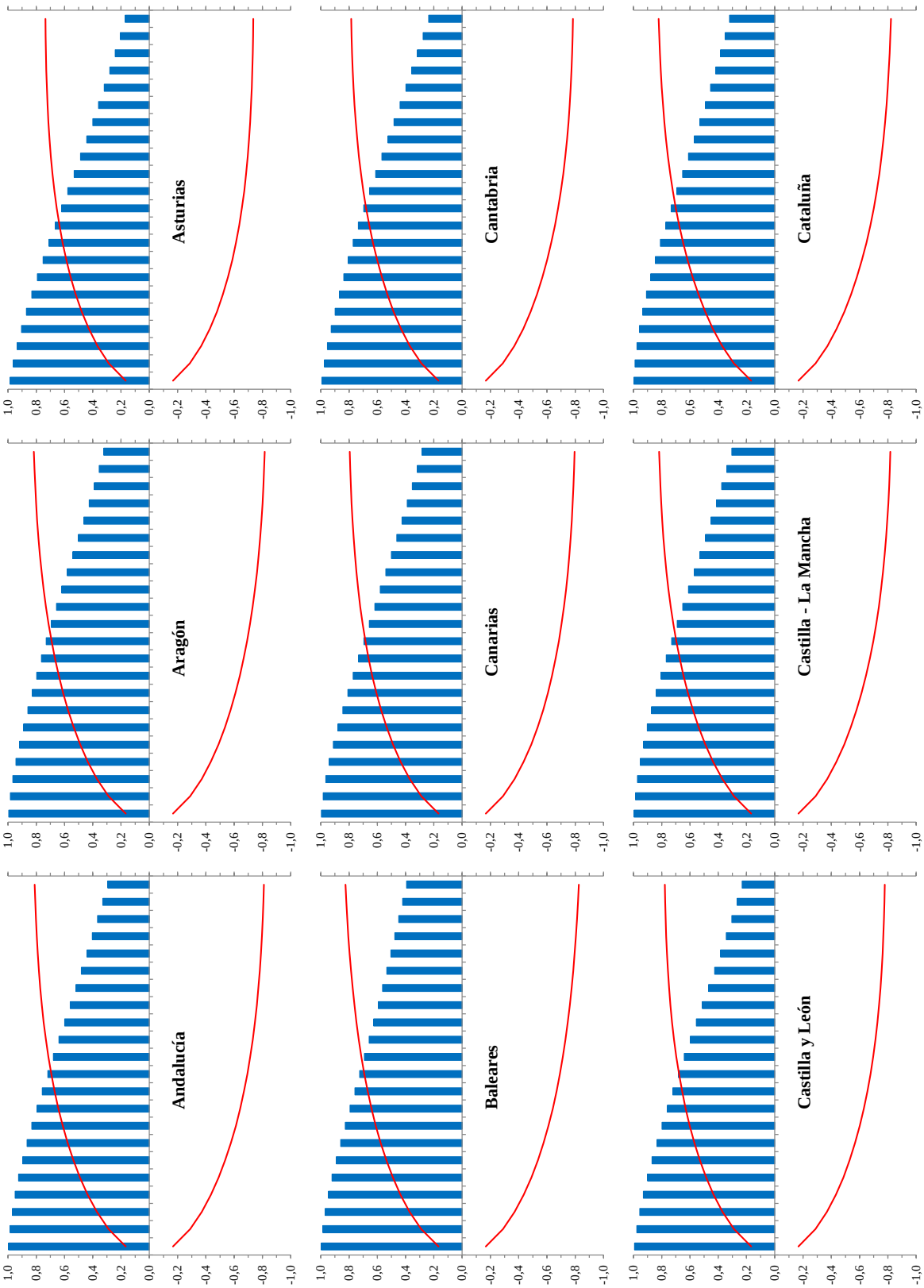

Figura C.7 Funciones de autocorrelación series filtradas: Andalucía, Aragón, Asturias, Baleares, Canarias, Cantabria, Castilla y León, Castilla – La Mancha y Cataluña

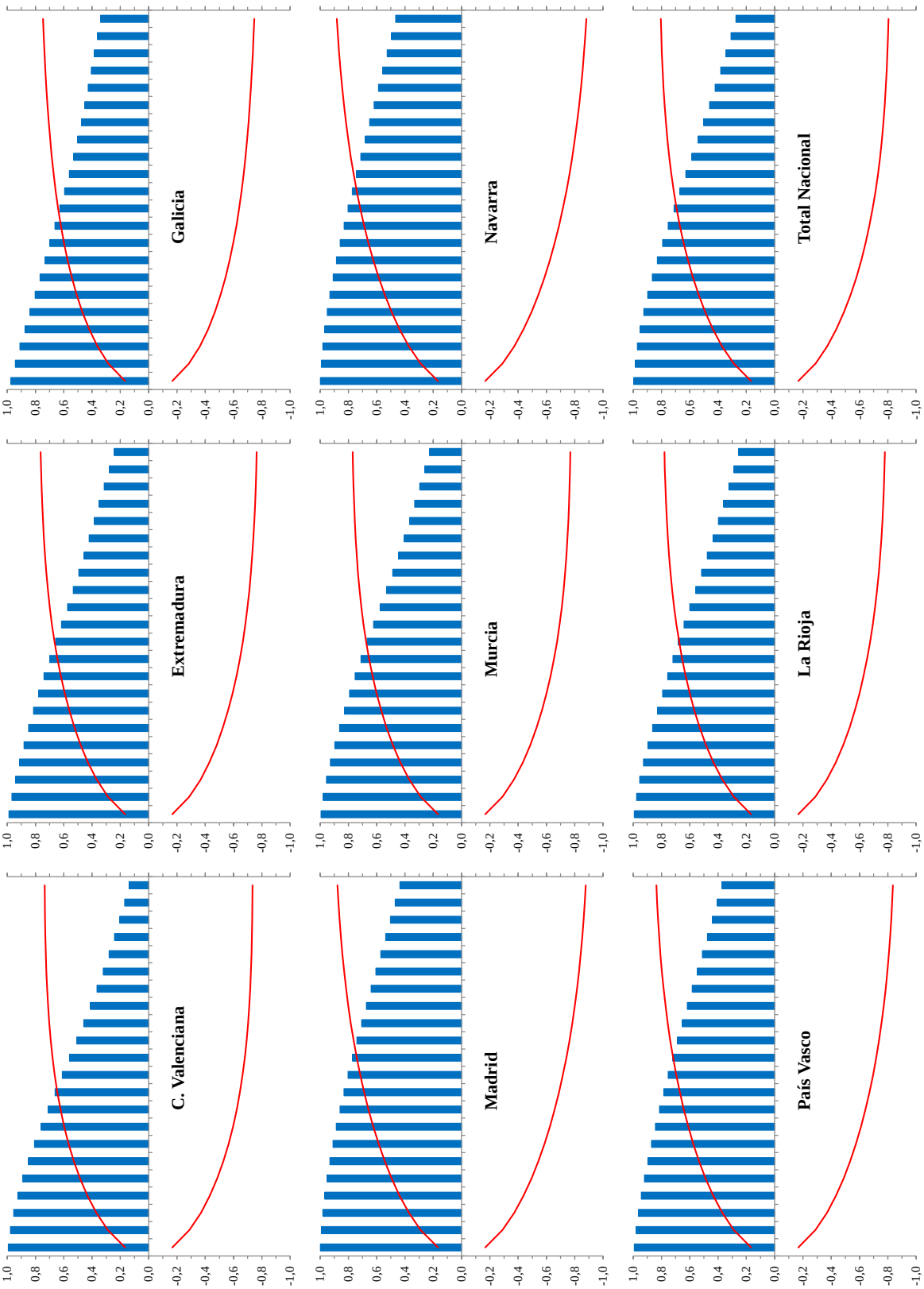

Figura C.8 Funciones de autocorrelación series filtradas: C. Valenciana, Extremadura, Galicia, Madrid, Murcia, Navarra, País Vasco, La rioja y Total Nacional.

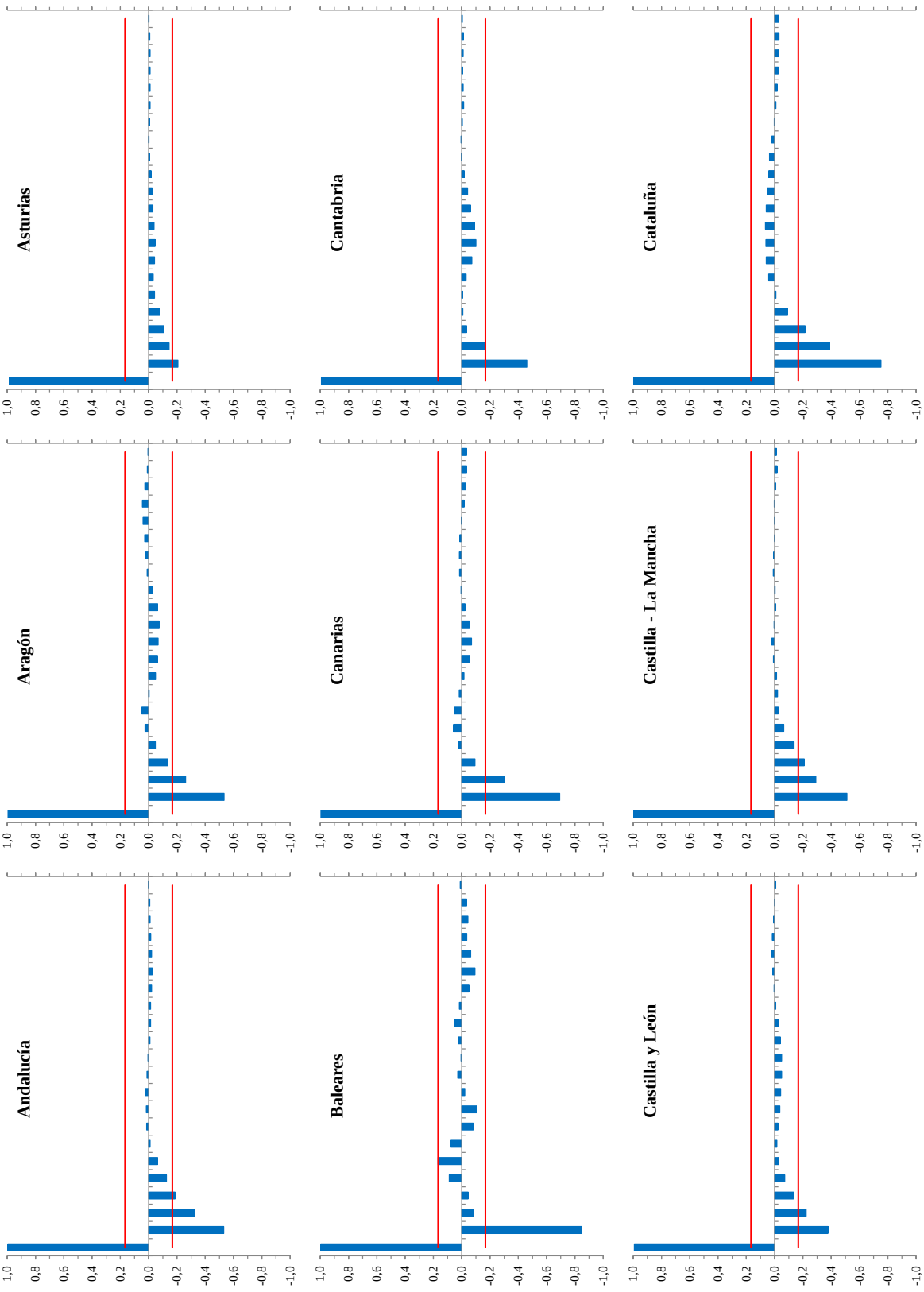

Figura C.9 FAC parcial series filtradas: Andalucía, Aragón, Asturias, Baleares, Canarias, Cantabria, Castilla y León, Castilla – La Mancha y Cataluña.

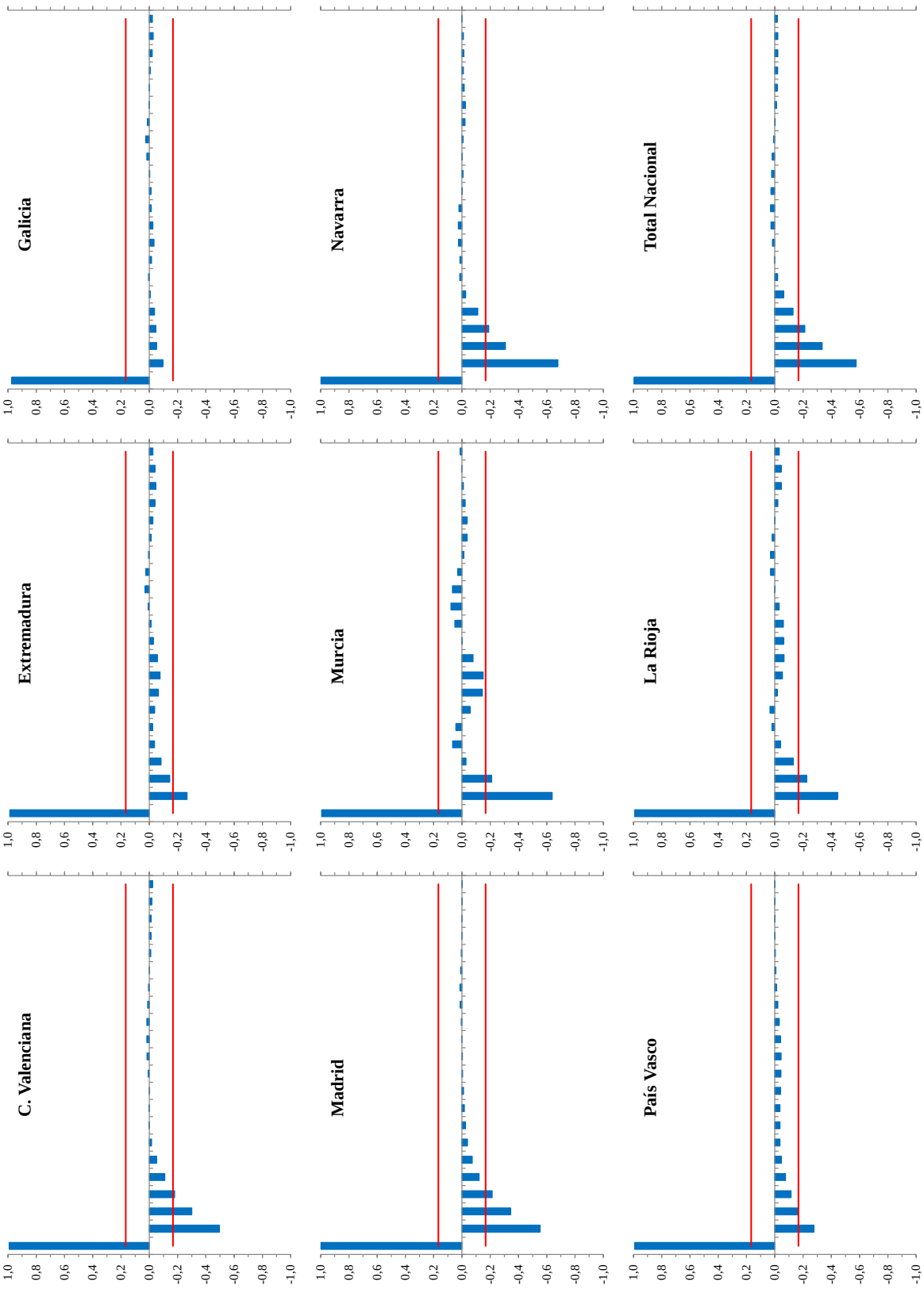

Figura C.10 FAC parcial series filtradas: C. Valenciana, Extremadura, Galicia, Madrid, Murcia, Navarra, País Vasco, La rioja y Total Nacional.

Anexo D

Análisis de Componentes Principales, PCA: Matrices y Figuras

En este Anexo se presentan las siguientes matrices:

- Matriz de covarianzas $\mathbf{C}$ de las series filtradas y diferenciadas.
- Matriz ortogonal $\mathbf{U}$ con los autovectores de la diagonalización de $\mathbf{C}$.

Y, también, la siguiente figura:

- Figura D.1 Series de las tres primeras componentes principales estimadas.

Matriz de covarianzas **C** de las series filtradas y diferenciadas.

0,133	0,115	0,090	0,123	0,126	0,097	0,097	0,100	0,127	0,140	0,110	0,081	0,121	0,143	0,107	0,103	0,135
0,115	0,130	0,092	0,107	0,108	0,086	0,094	0,095	0,116	0,134	0,107	0,081	0,109	0,132	0,097	0,094	0,121
0,090	0,092	0,108	0,089	0,086	0,098	0,086	0,080	0,090	0,107	0,097	0,084	0,087	0,106	0,078	0,082	0,110
0,123	0,107	0,089	0,182	0,159	0,082	0,096	0,096	0,125	0,154	0,102	0,088	0,108	0,152	0,108	0,105	0,137
0,126	0,108	0,086	0,159	0,187	0,094	0,097	0,097	0,123	0,157	0,103	0,074	0,104	0,162	0,104	0,105	0,129
0,097	0,086	0,098	0,082	0,094	0,190	0,079	0,085	0,105	0,116	0,097	0,079	0,096	0,149	0,085	0,089	0,097
0,097	0,094	0,086	0,096	0,097	0,079	0,090	0,079	0,095	0,109	0,099	0,070	0,092	0,106	0,083	0,086	0,117
0,100	0,095	0,080	0,096	0,097	0,085	0,079	0,089	0,103	0,112	0,096	0,070	0,099	0,121	0,087	0,082	0,105
0,127	0,116	0,090	0,125	0,123	0,105	0,095	0,103	0,143	0,144	0,108	0,073	0,124	0,147	0,113	0,105	0,133
0,140	0,134	0,107	0,154	0,157	0,116	0,109	0,112	0,144	0,173	0,117	0,092	0,127	0,171	0,118	0,115	0,146
0,110	0,107	0,097	0,102	0,103	0,097	0,099	0,096	0,108	0,117	0,180	0,081	0,112	0,118	0,101	0,088	0,142
0,081	0,081	0,084	0,088	0,074	0,079	0,070	0,070	0,073	0,092	0,081	0,097	0,078	0,101	0,068	0,072	0,093
0,121	0,109	0,087	0,108	0,104	0,096	0,092	0,099	0,124	0,127	0,112	0,078	0,128	0,129	0,103	0,099	0,126
0,143	0,132	0,106	0,152	0,162	0,149	0,106	0,121	0,147	0,171	0,118	0,101	0,129	0,247	0,120	0,119	0,139
0,107	0,097	0,078	0,108	0,104	0,085	0,083	0,087	0,113	0,118	0,101	0,068	0,103	0,120	0,100	0,089	0,118
0,103	0,094	0,082	0,105	0,105	0,089	0,086	0,082	0,105	0,115	0,088	0,072	0,099	0,119	0,089	0,098	0,113
0,135	0,121	0,110	0,137	0,129	0,097	0,117	0,105	0,133	0,146	0,142	0,093	0,126	0,139	0,118	0,113	0,197

Matriz ortogonal $\mathbf{U}$ con los autovectores de la diagonalización de $\mathbf{C}$.

0,254	-0,050	0,061	-0,229	0,058	0,059	0,024	0,040	0,266	-0,289	-0,267	-0,195	-0,620	0,422	-0,101	0,176	0,017
0,236	0,028	0,102	-0,232	0,198	-0,254	0,322	-0,387	-0,391	-0,306	0,266	0,003	0,185	0,204	-0,236	-0,106	0,248
0,202	0,250	0,055	0,319	0,320	-0,046	0,141	-0,246	0,019	0,606	-0,329	-0,041	-0,103	0,081	-0,288	-0,170	-0,019
0,264	-0,470	-0,005	0,368	0,061	0,079	0,000	0,503	-0,361	0,113	0,114	-0,273	0,060	0,221	-0,061	0,038	0,150
0,265	-0,470	-0,153	0,262	-0,292	0,215	0,240	-0,333	0,391	-0,133	-0,134	0,229	0,157	-0,069	-0,053	-0,143	0,152
0,223	0,561	-0,545	0,168	-0,088	0,460	0,014	0,052	-0,140	-0,205	0,047	-0,048	0,073	0,106	0,020	0,013	0,070
0,204	0,061	0,153	0,069	0,103	0,040	-0,002	-0,244	0,167	0,200	0,338	-0,222	-0,094	0,082	0,743	-0,011	0,237
0,207	0,052	0,019	-0,153	0,028	-0,115	0,112	0,082	0,105	0,133	-0,261	0,172	0,526	0,302	0,172	0,598	-0,126
0,257	-0,039	0,001	-0,396	0,025	0,201	-0,004	0,144	-0,207	0,244	-0,130	0,333	-0,200	-0,425	0,010	0,153	0,489
0,292	-0,165	-0,077	-0,040	0,095	0,062	0,202	-0,172	-0,338	-0,100	-0,146	-0,155	-0,135	-0,392	0,232	0,079	-0,629
0,240	0,321	0,410	0,179	-0,684	-0,207	0,246	0,142	-0,051	0,024	0,082	-0,031	-0,120	-0,109	-0,073	0,070	-0,031
0,178	0,157	0,023	0,381	0,426	-0,372	0,049	0,301	0,198	-0,396	-0,046	0,309	-0,058	-0,257	0,127	-0,015	0,084
0,239	0,093	0,117	-0,362	0,073	0,062	0,069	0,297	0,308	-0,041	-0,151	-0,444	0,377	-0,195	-0,042	-0,435	0,009
0,310	-0,064	-0,531	-0,122	-0,225	-0,596	-0,387	-0,061	0,040	0,157	0,051	-0,077	-0,055	-0,015	-0,020	-0,076	-0,003
0,218	-0,008	0,084	-0,178	-0,019	0,109	-0,037	0,207	-0,047	0,117	0,164	0,564	-0,059	0,360	0,134	-0,473	-0,348
0,213	-0,010	0,011	-0,044	0,180	0,159	-0,035	0,015	0,349	0,134	0,646	0,004	-0,025	-0,172	-0,399	0,308	-0,226
0,281	0,039	0,402	0,134	0,000	0,171	-0,738	-0,242	-0,119	-0,200	-0,133	0,023	0,163	-0,035	-0,088	0,023	-0,013

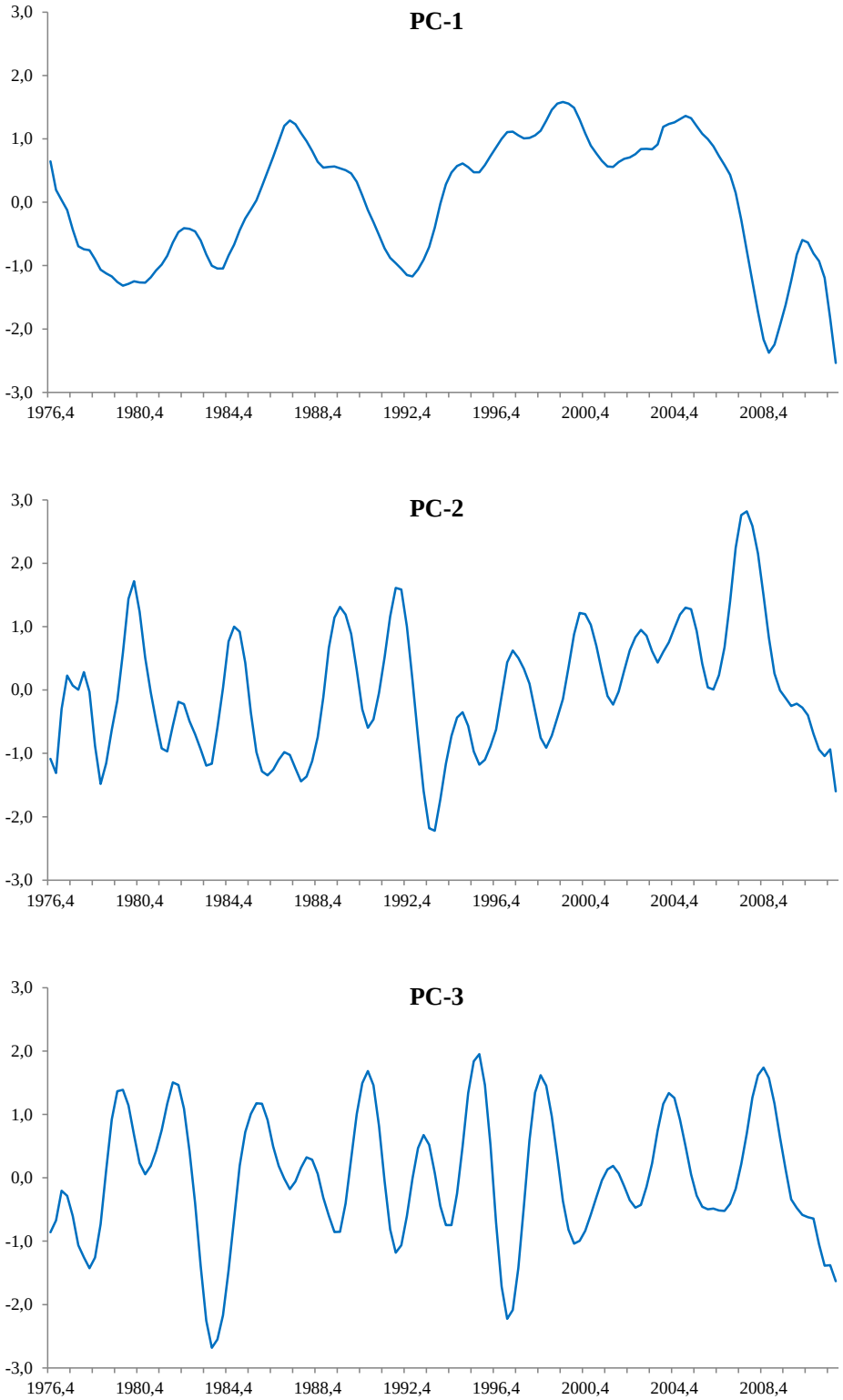

Figura D.1 Series de las tres primeras componentes principales estimadas.

Anexo E

Análisis de Componentes Independientes, ICA: Matrices y Figuras

En este Anexo se presentan las siguientes matrices:

- Matriz traspuesta $\mathbf{Q}^T$ de blanqueo de los datos centrados.
- Matriz ortogonal de separación $\mathbf{B}$ para obtener las ICs a partir de los datos blanqueados.
- Matriz de mezcla $\mathbf{A}$ para obtener los datos observados a partir de las ICs.
- Matriz traspuesta de separación $\mathbf{W}^T$ para obtener las ICs a partir de los datos observados.

Y, también, las siguientes figuras:

- Figura E.1 Series de las tres primeras componentes independientes estimadas con el algoritmo FastICA.
- Figura E.2 Series de las tres primeras componentes independientes estimadas con el algoritmo AMUSE.

- Matriz traspuesta $\mathbf{Q}^T$ de blanqueo de los datos centrados.

0,185	-0,136	0,168
0,172	0,075	0,278
0,147	0,674	0,149
0,192	-1,265	-0,013
0,193	-1,266	-0,418
0,162	1,510	-1,491
0,149	0,164	0,417
0,151	0,141	0,052
0,188	-0,106	0,001
0,213	-0,445	-0,210
0,175	0,863	1,123
0,130	0,421	0,064
0,174	0,249	0,320
0,226	-0,173	-1,451
0,159	-0,021	0,230
0,155	-0,026	0,031
0,204	0,106	1,098

- Matriz ortogonal de separación **B** para obtener las ICs a partir de los datos blanqueados.

FastICA			AMUSE		
0,987	-0,030	0,157	0,960	0,251	0,122
0,080	-0,757	-0,649	-0,279	0,884	0,374
0,138	0,653	-0,744	-0,014	-0,394	0,919

- Matriz de mezcla **A** para obtener los datos observados a partir de las ICs.

FastICA			AMUSE		
0,348	0,027	0,019	0,336	0,089	0,033
0,325	-0,006	0,024	0,304	0,114	0,035
0,274	-0,061	0,084	0,240	0,167	-0,014
0,362	0,162	-0,063	0,391	-0,054	0,072
0,355	0,198	-0,022	0,399	-0,074	0,023
0,264	-0,004	0,327	0,265	0,195	-0,261
0,284	-0,031	0,012	0,256	0,119	0,046
0,281	0,003	0,047	0,267	0,099	0,003
0,349	0,039	0,039	0,343	0,086	0,011
0,392	0,097	0,036	0,403	0,047	0,004
0,345	-0,161	0,012	0,268	0,254	0,096
0,240	-0,030	0,065	0,218	0,123	-0,012
0,329	-0,028	0,036	0,301	0,138	0,030
0,390	0,178	0,188	0,438	0,025	-0,163
0,301	0,006	0,017	0,285	0,093	0,034
0,290	0,023	0,035	0,281	0,080	0,009
0,402	-0,076	-0,047	0,348	0,175	0,135

- Matriz traspuesta de separación $\mathbf{W}^T$ para obtener las ICs a partir de los datos observados.

FastICA			AMUSE		
0,213	0,009	-0,188	0,191	-0,006	0,210
0,211	-0,224	-0,134	0,112	0,219	0,229
0,148	-0,595	0,349	-0,046	0,693	-0,126
0,226	0,981	-0,790	0,504	-1,069	0,489
0,163	1,245	-0,489	0,555	-1,222	0,117
-0,120	-0,162	2,118	-0,041	0,822	-1,962
0,207	-0,383	-0,183	0,050	0,343	0,321
0,153	-0,129	0,074	0,103	0,186	-0,005
0,189	0,094	-0,044	0,207	-0,041	0,046
0,190	0,490	-0,105	0,342	-0,413	-0,015
0,323	-1,367	-0,248	-0,186	1,232	0,695
0,125	-0,350	0,246	0,011	0,432	-0,106
0,215	-0,382	-0,052	0,065	0,389	0,199
0,001	1,091	0,999	0,438	-0,633	-1,263
0,194	-0,121	-0,163	0,130	0,112	0,222
0,159	0,012	-0,018	0,152	0,032	0,041
0,371	-0,777	-0,720	0,035	0,562	0,971

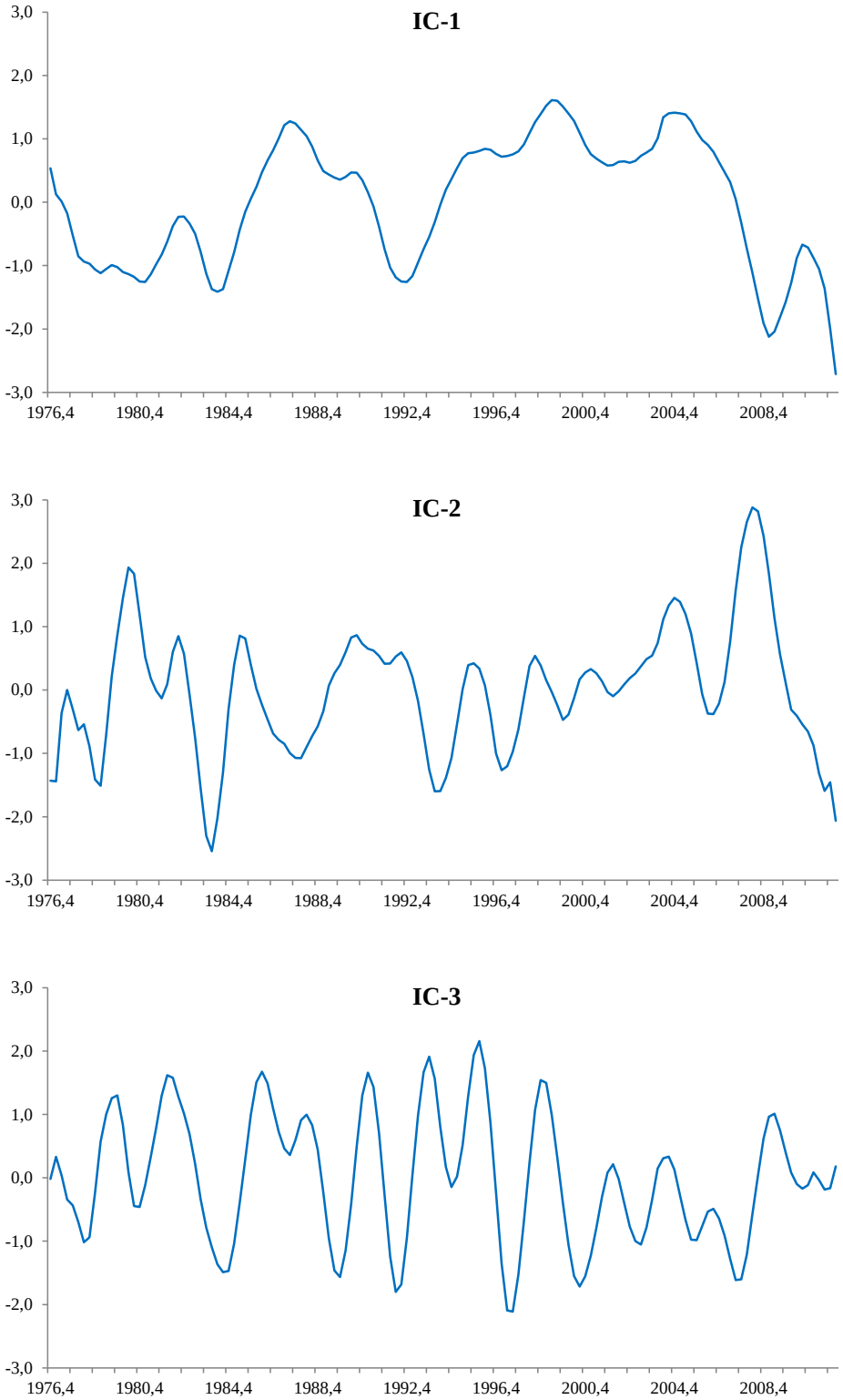

Figura E.1 Series de las tres primeras componentes independientes estimadas con el algoritmo FastICA.

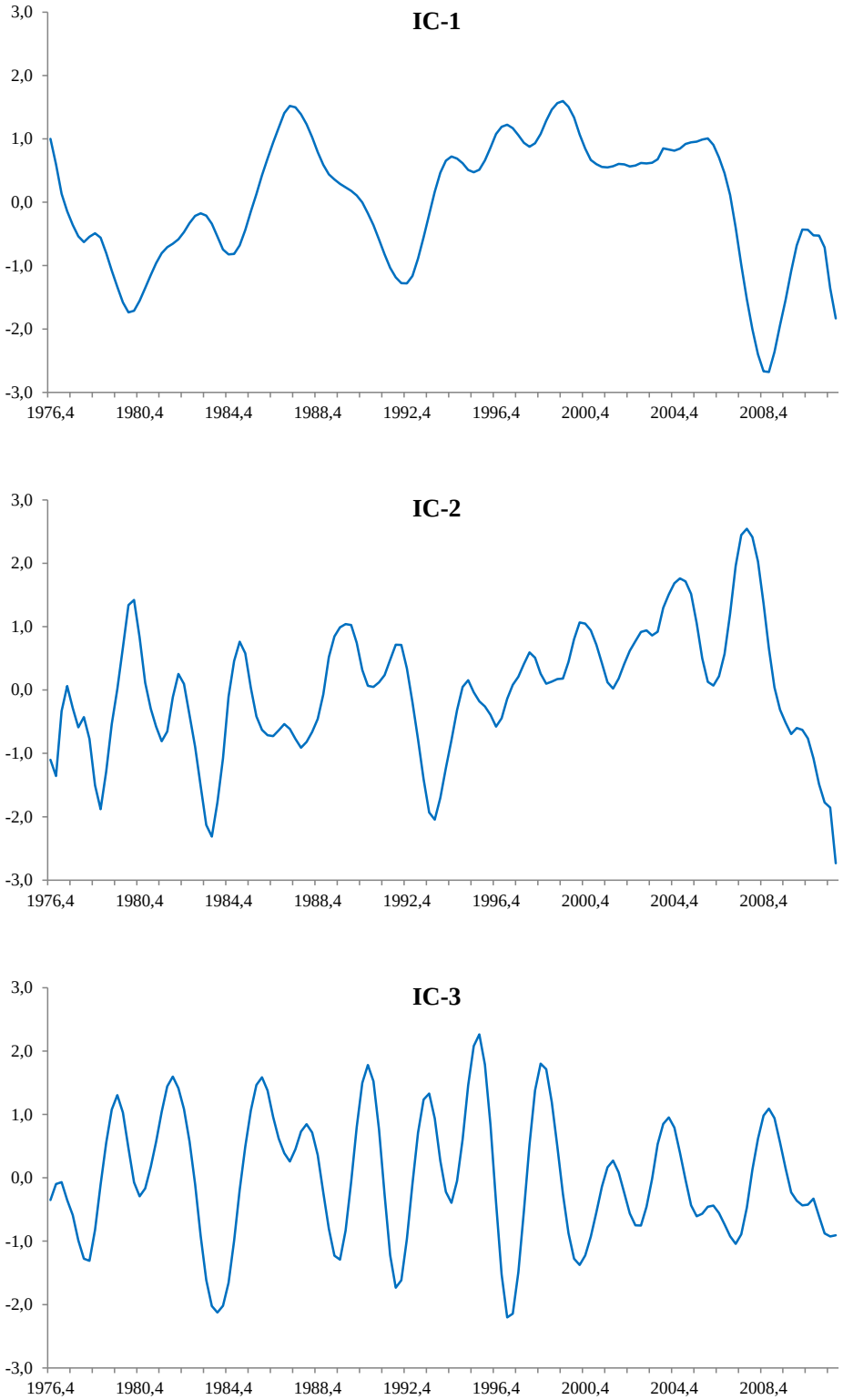

Figura E.2 Series de las tres primeras componentes independientes estimadas con el algoritmo AMUSE.

Anexo F

Código de las funciones programadas en MATLAB

En este Anexo se presentan las siguientes funciones programadas en MATLAB:

- Función *testrm()*.
- Función *ssa()*.
- Función *pca()*.
- Función *icarapid()*.
- Función *amuse()*.

- **Función *testrm()***

Esta función se ha diseñado para realizar el test rango-media sobre el coeficiente de correlación tomando como medida robusta de localización la mediana y como medida robusta de dispersión la desviación absoluta mediana.

```
function [r pval rinf rsup rango media] = testrm(X,p,alfa)
% Test rango/media con el coeficiente de correlación para analizar la
% estacionariedad en varianza de varias series temporales
%
% Sintaxis:     [r pval rinf rsup rango media] = testrm(X,p,alfa)
%
% Parámetros de entrada:
% X:    Matriz (n*k) con las series temporales en columnas
% p:    Número de elementos por grupo (por defecto 10)
% alfa: Nivel de significación del intervalo de confianza
%       (por defecto 0.05)
%
% Salida:
% r:     Coeficiente de correlación entre del rango/media
% pval:  p-valor para contrastar la hipótesis de no correlación
% rinf:  Límite inferior del intervalo de confianza del coeficiente
% rsup:  Límite superior del intervalo de confianza del coeficiente
% rango: Matriz con los rangos de los grupos para cada serie
% media: Matriz con las medias de los grupos para cada serie

% Dimensiones
[n k] = size(X);

% Vslores por defecto
if nargin < 3, alfa = 0.05; end
if nargin < 2, p = 10; end

% Nº de grupos de datos
m = fix(n/p);

% Rango/media de los grupos
rango = zeros(m,k);
media = zeros(m,k);
for j=1:k
    for i=1:m
        rango(i,j) = mad(X((i-1)*p+1:i*p,j),1);
        media(i,j) = median(X((i-1)*p+1:i*p,j));
    end
end

% Test correlación rango/media
r = zeros(k,1); pval = zeros(k,1);
rinf = zeros(k,1); rsup = zeros(k,1);
for j=1:k
    [R PV RI RS] = corrcoef(rango(:,j),media(:,j),'alpha',alfa);
    r(j) = R(1,2);
    pval(j) = PV(1,2);
```

```
    rinf(j) = RI(1,2);
    rsup(j) = RS(1,2);
end
```

- **Función *ssa()***

Esta función se ha diseñado para realizar un Análisis Espectral Singular, SSA, sobre una serie temporal según se describe en el apartado 3.3.2 del Capítulo 3.

```
function [PC RC S U D] = ssa(x,p)
% Análisis espectral singular (SSA)
%
% Sintaxis:     [PC RC S U D] = ssa(x,p)
%
% Parámetros de entrada:
% x:  Vector de dimensión (n*1) con la serie temporal
% p:  Número de desfases de la serie x que se incluyen en la matriz de
%     datos X de dimensión (np*p) con np=n-p+1.
%
% Salida:
% PC: Componentes principales de la matriz de datos X
% RC: Componentes reconstruidas con la misma longitud que la serie x
% S:  Matriz de covarianzas de X con estructura Toeplitz
% U:  Matriz con los autovectores en columnas de la matriz S
% D:  Matriz diagonal con los autovalores de la matriz S

% Tamaños
[n, ~] = size(x);
np = n-p+1;

% Matriz X reorganizada
X = zeros(np,p);
for k=1:p
    X(:,k) = x(k:end-p+k);
end

% Matriz de covarianzas con estructura de Toeplitz
C = corr(X);
S = zeros(p,p);
for k=1:p
    cte = mean(diag(C(1:p-k+1,k:p)));
    for i=1:p-k+1
        S(i,i+k-1) = cte;
        S(i+k-1,i) = cte;
    end
end

% Diagonalización de S
[U D] = eigs(S,p);
D = abs(D);

% Componentes principales
```

```
PC = X*U;

% Componentes reconstruidas
RC = zeros(n,p);
for t=1:n
    if 1<=t && t<=p-1
        j_inf = 1; j_sup = t;
    elseif p<=t && t<=np
        j_inf = 1; j_sup = p;
    else
        j_inf = t-n+p; j_sup = p;
    end
    nsum = j_sup-j_inf+1;
    for k=1:p
        for j=j_inf:j_sup
            RC(t,k) = RC(t,k)+(1/nsum)*PC(t-j+1,k)*U(j,k);
        end
    end
end
```

- **Función *pca()***

Esta función se ha diseñado para realizar un Análisis de Componentes Principales, PCA, sobre un conjunto de datos observados según se describe en el apartado 3.3.1 del Capítulo 3.

```
function [PC PO PZ Y S U D] = pca(X,m)
% Análisis de Componentes Principales PCA
%
% Sintaxis:     [PC PO PZ Y S U D] = pca(X,m)
%
% Parámetros de entrada:
% X: Matriz de dimensión (n*p) con los datos originales
% m: Número de Componentes Principales a estimar (por defecto p)
%
% Salida:
% PC: Componentes principales, matriz de dimensión (n*m)
% PO: Componentes principales centradas, matriz de dimensión (n*m)
% PZ: Componentes principales tipificadas, matriz de dimensión (n*m)
% Y:  Datos centrados, matriz de dimensión (n*p)
% S:  Matriz de covarianzas de X de dimensión (p*p)
% U:  Matriz con los autovectores de dimensión (p*m)
% D:  Matriz diagonal con los autovalores de dimensión (m*m)

% Dimensiones
[n p] = size(X);

% Valores por defecto
if nargin < 2, m = p; end

% Chequeo de los datos
if ~isreal(X)
```

```
  error('La matriz de datos tiene parte imaginaria.');
end

% Chequeo del n° de componentes
if (m < 1) || (m > p)
    error('N° de componentes a estimar no válido.');
end

% Datos centrados en media: Y
M = kron(mean(X),ones(n,1));
Y = X - M;

% Matriz de covarianzas: S
S = cov(Y,1);

% Diagonalización de la matriz de covarianzas
[U D] = eigs(S,m);
U = U.*kron(sign(sum(U)),ones(p,1));

% Componentes principales centradas: PO
PO = Y * U;

% Componentes principales tipificadas: PZ
PZ = PO./repmat(sqrt(diag(D))',n,1);

% Componentes principales: PC
PC = PO + M * U;
```

- **Función *icarapid()***

Esta función se ha diseñado para realizar un Análisis de Componentes Independientes, ICA, sobre un conjunto de datos observados mediante el algoritmo FastICA según se describe en el apartado 3.3.3 del Capítulo 3.

```
function [IC IO Y Z A W Q B] = icarapid(X,m,g)
% ICA con el algoritmo FastICA con ortogonalización simétrica.
%
% Sintaxis:     [IC IO Y Z A W Q B] = icarapid(X,m,g)
%
% Parámetros de entrada:
% X: Matriz de dimensión (n*p) con los datos originales
% m: Número de Componentes Independientes a estimar (por defecto p)
% g: Función no lineal usada en el algoritmo de punto fijo:
%    1: g(u) = tanh(u) (por defecto)
%    2: g(u) = u*exp(-u^2/2)
%
% Salida:
% IC: Componentes independientes, matriz de dimensión (n*m)
% IO: Componentes independientes centradas, matriz de dimensión (n*m)
% Y:  Datos centrados, matriz de dimensión (n*p)
% Z:  Datos blanqueados, matriz de dimensión (n*m)
% A:  Matriz de mezcla de dimensión (p*m)
```

```
% W:  Matriz de separación de dimensión (m*p)
% Q:  Matriz de blanqueo de dimensión (m*p)
% B:  Matriz ortogonal de separación de datos blanqueados (m*m)

% Dimensiones
[n p] = size(X);

% Valores por defecto
if nargin < 3, g = 1; end
if nargin < 2, m = p; end

% Chequeo de los datos
if ~isreal(X)
  error('La matriz de datos tiene parte imaginaria.');
end

% Chequeo del n° de componentes
if (m < 1) || (m > p)
    error('N° de componentes a estimar no válido.');
end

% Chequeo de la función no lineal
if (g < 1) || (g > 2)
    error('Valor incorrecto para la función no lineal.');
end

% Error de estimación
res = 0.0001;
maxiter = 100000;

% Datos centrados en medias: Y
M = kron(mean(X),ones(n,1));
Y = X - M;

% Datos blanqueados: Z
Cy = cov(Y,1);
[U D] = eigs(Cy,m);
Q = sqrt(D) \ U';
Z = Y * Q';

% Matriz ortogonal B inicial
B = toeplitz((1:m),(1:m));
B = real(inv(B*B')^0.5)*B;
Bo = B;

% Estimación matriz ortogonal de separación de datos blanqueados: B
con = 1;
iter = 0;
while con && (iter < maxiter)
    S = B * Z';
    switch g
        case 1
            g_S = tanh(S);
            dg_S = 1 - tanh(S).^2;
        case 2
            g_S =  S.*exp(-S.^2 / 2);
            dg_S = (1 - S.^2).*exp(-S.^2 / 2);
```

```
    end
    B = (g_S*Z)/n - (mean(dg_S,2)*ones(1,size(B,2))).*B;
    B = real(inv(B*B')^0.5)*B;
    if (1-min(abs(diag(Bo*B')))) < res
        con = 0;
    else
        Bo = B;
    end
    iter = iter+1;
end
fprintf(1,'%d iteraciones realizadas\n',iter)

% Matriz de mezcla: A
A = U * sqrt(D) * B';

% Matriz de separación: W
W = B * Q;

% Componentes independientes centradas: IO
IO = Z * B';

% Componentes independientes: IC
IC = Y * W' + M * W';
```

- **Función *amuse()***

Esta función se ha diseñado para realizar un Análisis de Componentes Independientes, ICA, sobre un conjunto de datos observados mediante el algoritmo AMUSE según se describe en el apartado 3.3.3 del Capítulo 3.

```
function [IC IO Y Z A W Q B] = amuse(X,m,tau)
% ICA para series temporales: algoritmo AMUSE
%
% Sintaxis:     [IC IO Y Z A W Q B] = amuse(X,m,tau)
%
% Parámetros de entrada:
% X:   Matriz de dimensión (n*p) con los datos originales
% m:   Número de Componentes Independientes a estimar (por defecto p)
% tau: Retardo para la matriz de covarianzas retardadas(por defecto 1)
%
% Salida:
% IC: Componentes independientes, matriz de dimensión (n*m)
% IO: Componentes independientes centradas, matriz de dimensión (n*m)
% Y:  Datos centrados, matriz de dimensión (n*p)
% Z:  Datos blanqueados, matriz de dimensión (n*m)
% A:  Matriz de mezcla de dimensión (p*m)
% W:  Matriz de separación de dimensión (m*p)
% Q:  Matriz de blanqueo de dimensión (m*p)
% B:  Matriz ortogonal de separación de datos blanqueados (m*m)

% Dimensiones
[n p] = size(X);
```

```
% Valores por defecto
if nargin < 3, tau = 1; end
if nargin < 2, m = p; end

% Chequeo de los datos
if ~isreal(X)
  error('La matriz de datos tiene parte imaginaria.');
end

% Chequeo del n° de componentes
if (m < 1) || (m > p)
    error('N° de componentes a estimar no válido.');
end

% Chequeo del retardo
if (tau < 1) || (tau > n-m)
    error('Valor incorrecto para el retardo.');
end

% Datos centrados en medias: Y
M = kron(mean(X),ones(n,1));
Y = X - M;

% Datos blanqueados: Z
Cy = cov(Y,1);
[U D] = eigs(Cy,m);
Q = sqrt(D) \ U';
Z = Y * Q';

% Matriz covarianzas retardadas de Z
Cz_tau = Z(1+tau:end,:)' * Z(1:end-tau,:) / (n-tau);
Cz = 0.5 * (Cz_tau + Cz_tau');

% Matriz ortogonal de separación de datos blanqueados: B
[V, ~] = eigs(Cz,m);
V = V.*kron(sign(sum(V)),ones(m,1));
B = V';

% Matriz de mezcla: A
A = U * sqrt(D) * B';

% Matriz de separación: W
W = B * Q;

% Componentes independientes centradas: IO
IO = Z * B';

% Componentes independientes: IC
IC = Y * W' + M * W';
```

Referencias

1. Allen, M. R. y Smith, L. A. (1996). Monte Carlo SSA: Detecting irregular oscillations in the presence of colored noise. *Journal of Climate*, 9, pág. 3373-3404.
2. Anderson, T. W. (1963). Asymptotic theory for principal component analysis. *Annals of Mathematical Statistics*, 34, pág. 122-148.
3. Back, A. D. y Weigend, A. S. (1997). A first application of independent component analysis to extracting structure from stock returns. *International Journal of Neural Systems*, 8 (5), pág. 473-484.
4. Barnett, T. P. (1983). Interaction of the monsoon and pacific trade wind system of interannual time scales. Part I: The equatorial case. *Monthly Weather Review*, 111, pág. 756-773.
5. Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Statistical Psychology*, 3, pág. 77-85.
6. Bell, A.J. y Sejnowski, T.J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7, pág. 1129-1159.
7. Benzi, R., Deidda, R. y Marrocu, M. (1997). Characterization of temperature and precipitation fields over Sardinia with PCA and SSA. *International Journal of Climatology*, 17, pág. 1231-1262.
8. Besse, P. y Ferré, L. (1993). Sur l'usage de la validation croisée en analyse en composantes principales. *Rev. Statistique Appliquée*, 41, pág. 71-76.
9. Bógalo, J. y Quilis, E.M. (2003). Estimación del ciclo económico mediante filtros de Butterworth. *Boletín Trimestral de Coyuntura del INE*, 87.
10. Brillinger, D. R. (1981). *Time Series: Data Analysis and Theory.* Holden-Day.
11. Cardoso, J.F. (1991). Super-symmetric decomposition of the fourth-order cumulant tensor. Blind identification of more sources than sensors. En *Proc. ICASSP'91*, pág. 3109-3112.
12. Cardoso, J.F. (1997). Infomax and maximum likelihood for source separation. *IEEE Letters on Signal Processing*, 4, pág. 112-114.

13. Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1, pág. 245-276.

14. Comon, P. (1994). Independent Component Analysis. A new concept? *Signal Processing*, 36, pág. 287-314.

15. Cover, T.M. y Thomas J.A. (1991). *Elements of Information Theory*. Wiley and Sons.

16. Delfosse, N. y Loubaton, P. (1995). Adaptive blind separation of independent sources: a deflation approach. *Signal Processing*, 45, pág. 59-83.

17. Eastment, H. T. y Krzanowski, W. J. (1982). Cross-validatory choice of the number of components from a PCA. *Technometrics*, 24, pág. 73-77.

18. Elsner, J. B. y Tsonis, A. A. (1996). *Singular Spectrum Analysis: A New Tool in Time Series Analysis*. Plenun Press.

19. Ghil, M. y Vautard, R. (1991). Interdecadal oscillations and the warming trend in global temperature time series. *Nature*, 35, pág. 324-327.

20. Girschick, M. A. (1939). On the sampling theory of roots of determinantal equations. *Annals of Mathematical Statistics*, 10, pág. 203-224.

21. Hasselman, K. (1988). PIPs and POPs: The reduction of complex dynamical systems using principal iteration and oscillation patterns. *Journal of Geophysical Research*, 93, pág. 11015-11021.

22. Horel, J. D. (1984). Complex principal component analysis: Theory and examples. *Journal of Climate and Applied Meteorology*, 23, pág. 1660-1673.

23. Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24 (7), pág. 498-520.

24. Huber, P. (1985). Projection pursuit. *The Annals of Statistics*, 13(2), pág. 435-475.

25. Hyvärinen, A. (1998a). New approximations of differential entropy for independent component analysis and projection pursuit. *Advances in Neural Information Processing Systems*, 10, pág. 273-279. MIT Press.

26. Hyvärinen, A. (1998b). Independent component analysis for time-dependent stochastic processes. En *Proc. Int. Conf. on Artificial Neural Networks (ICANN'98)*, pág. 135-140.

27. Hyvärinen, A. (1999a). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. on Neural Networks*, 10 (3), pág. 626-634.

28. Hyvärinen, A. (1999b). Survey on independent component analysis. *Neural Computing Surveys*, 2, pág. 94-128.

29. Hyvärinen, A. y Oja, E. (1997). A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9 (7), pág. 1483-1492.

30. Hyvärinen, A. y Oja, E. (2000). Independent component analysis: Algorithms and Applications. *Neural Networks*, 13 (4-5), pág. 411-430.

31. Hyvärinen, A., Karhunen, J. y Oja, E. (2001). *Independent Component Analysis*. John Wiley and Sons.

32. Johnson, R. A. y Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*. Pearson Education.

33. Jolliffe, I. T. (1970). *Redundant Variables in Multivariate Analysis*. Tesis doctoral. University of Sussex.

34. Jolliffe, I. T. (1972). Discarding variables in a principal component analysis 1: Artificial data. *Applied Statistic*, 21, pág. 160-173.

35. Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer-Verlag.

36. Jones, M. y Sibson, R. (1987). What is projection pursuit? *Journal of the Royal Statistical Society, Ser. A*, 150, pág. 1-36.

37. Jutten, C. y Herault, J. (1991). Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24, pág. 1-10.

38. Kaiser, H.F. (1960). The application of electronic computers to factor analysis. *Educ. Psychol. Meas.*, 20, pág. 141–151.

39. Karhunen, J., Oja, E., Wang, L., Vigário, R. y Joutsensalo, J. (1997). A class of neural networks for independent component analysis. *IEEE Trans. On Neural Networks*, 8 (3), pág. 486-504.

40. Kim, K. Y. y Wu, Q. (1999). A comparison study of EOF techniques: analysis of nonstationary data with periodic statistics. *Journal of Climate*, 12, pág. 185-199.

41. Lawley, D. N. (1963). On testing a set of correlation coefficients for equality. *Annals of Mathematical Statistics*, 34, pág. 149-151.

42. Mardia, K. V., Kent, J. T. y Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press.

43. Meyer, C. D. (2000). *Matrix Analysis and Applied Linear Algebra*. SIAM.

44. Nadal, J.P. y Parga, N. (1994). Non-linear neurons in the low noise limit: a factorial code maximizes information transfer. *Network*, 5, pág. 565-581.

45. Pearlmutter, B.A. y Parra, L.C. (1997). Maximum likelihood blind source separation: A context-sensitive generalization of ICA. *Advances in Neural Information Processing Systems*, vol. 9, pág. 613-619.

46. Pearson, K. (1901). On line and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2, pág. 559-572.

47. Peña, D. (1991). *Estadística. Modelos y Métodos*. Volumen I. Alianza Editorial.

48. Peña, D. (2002). *Análisis de Datos Multivariantes*. Mc Graw Hill.

49. Pham, D.T., Garrat, P. y Jutten, C. (1992). Separation of a mixture of independent sources through a maximum likelihood approach. En *Proc. EUSIPCO*, pág. 771-774.

50. Plaut, G. y Vautard, R. (1994). Spells of low-frequency oscillations and weather regimes in the Northern Hemisphere. *Journal Atmospheric Sciences*, 51, pág. 210-236.

51. Rassmusson, E. M., Arkin, P. A., Chen, W. Y. y Jalickee, J. B. (1981). Biennial variations in surface temperature over the United States as revealed by singular decomposition. *Monthly Weather Review*, 109, pág. 587-598.

52. Ríos, S. (1985). *Métodos Estadísticos*. Ediciones del Castillo.

53. Sebastiao, F. y Oliveira, I. (2009). Estudo de Séries Temporais na Análise em Componentes Principais e na Análise em Componentes Independentes. *Actas do XVI Congresso Anual da SPE*, pág. 1-8.

54. Stoffer, D. S. (1999). Detecting common signals in multiple time series using the spectral envelope. *JASA*, 94, pág. 1341-1356.

55. Tong, L., Soon, V. C., Huang, Y. F. y Liu, R. (1990). Amuse: a new blind identification algorithm. *IEEE International Symposium on Circuits and Systems*, 3, pág. 1784-1787.

56. Tong, L., Soon, V. C., Huang, Y. F. y Liu, R. (1991). Indeterminancy and identifiability of blind separation. *IEEE Transactions on Circuits and Systems*, 38, pág. 499-509.

57. Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, 41, pág. 321-327.

58. Von Storch, H., Bruns, T., Fischer-Bruns, I. y Hasselman, K. (1988). Principal oscillation pattern analysis of the 30 to 60 day oscillation in general circulation model equatorial troposphere. *J. Geophysical Research*, 93, pág. 11022-11036.

59. Wold, S. (1978). Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics*, 20, pág. 771-774.

60. Zwiers, F. W. (1999). The detection of climate change. En *Anthropogenic Climate Change*, eds. H. von Storch y G. Flöser, pág. 161-206. Springer.