Universidad de Sevilla

Facultad de Matemáticas

ANÁLISIS DE INFLUENCIA EN COMPONENTES PRINCIPALES

Memoria dirigida por:

Prof. Dr. D. Juan Muñoz Pichardo
Prof. Dr. D. Rafael Pino Mejías

Memoria presentada por
Alicia Enguix González
para optar al grado de
Doctor en Ciencias Matemáticas.

Vº Bº del Director

Prof. Dr. D. Juan Muñoz Pichardo

Vº Bº del Director

Prof. Dr. D. Rafael Pino Mejías

Fdo.: Alicia Enguix González

Sevilla, Mayo de 2001

ANÁLISIS DE INFLUENCIA

EN COMPONENTES PRINCIPALES

UNIVERSIDAD DE SEVILLA
NEGOCIADO DE TESIS

Queda registrado este Título de Doctor al folio 54 número 24 del libro correspondiente.
Sevilla, 25 MAYO 2001
El Jefe del Negociado.

P. D.

UNIVERSIDAD DE SEVILLA

Depositado en el D.E. de Estadística e I.O. de la F. de Matemáticas de esta Universidad desde el día 26 - mayo hasta el día 10 de junio de 2001

Sevilla 16 de junio de 2001

EL DIRECTOR DEL Departamento

A mi familia

Agradecimientos

Quisiera expresar mi más sincera gratitud a los directores de esta memoria, quienes han sido mis maestros en el trayecto mi carrera docente e investigadora, por su excelente labor de dirección, paciencia, apoyo y sabios consejos.

También quiero dar las gracias a todos los miembros del departamento de Estadística e Investigación Operativa, en particular a su director, Don Rafael Infante Macías por la confianza que ha depositado en mi, a Don Juan Luis Moreno Rebollo por su gran colaboración desinteresada y sus incesantes ánimos, a Doña María Dolores Jiménez Gamero, por sus interesantes consejos científicos, a Don Joaquín Muñoz García y Doña Teresa Gómez Gómez por su constante preocupación, a Don David Gutiérrez Rubio, por su generosidad, continuo apoyo moral y sus valiosos comentarios; y en general al resto de mis compañeros, por su preocupación y paciencia.

Mis reconocimientos al servicio de la Biblioteca de la Facultad de Matemáticas por su gran amabilidad y eficiencia.

Y por último, quisiera agradecer a mi familia su enorme paciencia, sacrificio y comprensión.

PRÓLOGO

Cualquier análisis estadístico tiene como objetivo básico la obtención de conclusiones fiables, a partir de los datos de las variables analizadas. Así, el papel que desempeñan las observaciones es de gran importancia para el desarrollo del estudio. Por ello, es conveniente realizar un análisis previo sobre la repercusión que pueden ejercer los datos sobre los resultados obtenidos a partir de los mismos. El Análisis de Influencia abarca un conjunto de métodos que pretenden detectar observaciones que puedan provocar grandes alteraciones en las conclusiones finales del análisis realizado.

El objetivo principal de esta memoria es la obtención de diagnósticos de influencia en una técnica clásica, el Análisis de Componentes Principales, basados en el sesgo condicionado, herramienta introducida recientemente en el Análisis de Influencia.

Los parámetros de mayor interés en el Análisis de Componentes Principales son los autovalores y autovectores de la matriz de covarianzas, y por ello, tradicionalmente, han sido el centro del desarrollo de los análisis de influencia dentro de esta técnica y también lo serán en este trabajo.

En el primer capítulo de la memoria se recogen los conceptos fundamentales del Análisis de Influencia y las herramientas más comúnmente utilizadas en él: las funciones de influencia y sus versiones muestrales. A continuación, se recoge el concepto del sesgo condicionado y su utilidad para analizar el efecto que ejercen las observaciones en los estadísticos de interés en un análisis. En este mismo capítulo, también se presentan los fundamentos básicos del Análisis de Componentes Principales, con el objetivo de centrar la notación a utilizar a lo largo del trabajo. Y se finaliza realizando una síntesis de las referencias bibliográficas sobre el Análisis de Influencia en la técnica objeto de esta memoria.

En el segundo capítulo, con el objetivo de relacionar los estudios basados en el sesgo condicionado con los realizados desde otras técnicas, se recogen los resultados sobre funciones de influencia en los parámetros y estadísticos de interés. En él se realizan algunas aportaciones, en especial para el caso de influencia conjunta por distintas observaciones, sobre un parámetro o estadístico, apartado del Análisis de Influencia poco desarrollado en la literatura.

En los dos capítulos siguientes se calcula, respectivamente, una aproximación del sesgo condicionado de los estadísticos de interés en el Análisis de Componentes Principales, autovalores y autovectores de la matriz de covarianzas muestrales, bajo hipótesis de normalidad. Para ello, se utiliza el desarrollo en serie del estadístico correspondiente en función de parámetros poblacionales y las componentes de la matriz de covarianzas muestrales. Esto permite obtener una aproximación del sesgo condicionado para ambos estadísticos. Será necesario obtener una serie de resultados previos sobre los sesgos condicionados de cada uno de los términos considerados del desarrollo en serie.

En el capítulo 5, se proponen estimaciones para el sesgo condicionado de un autovalor y un autovector de la matriz de covarianzas muestrales y se sugieren distintas medidas de influencia conjunta, para autovalores, autovectores, la suma de autovalores (parámetro que desempeña un papel importante dentro del Análisis de Componentes Principales) y para un conjunto de autovectores. Finalmente, en este capítulo, se proporcionan distintas aplicaciones prácticas para ilustrar el análisis de influencia llevado a cabo con diagnósticos basados en el sesgo condicionado.

Índice

Capítulo 1

INTRODUCCIÓN

El objetivo de todo análisis estadístico es obtener conclusiones fiables a partir de los datos resultantes de una experimentación. Por tanto, la fiabilidad de las observaciones del proceso es de especial interés, ya que el análisis se realiza sobre codificaciones del fenómeno natural en estudio, y las técnicas estadísticas que se apliquen pueden verse fuertemente afectadas por algunas de las observaciones realizadas. Este problema ha originado un gran número de métodos enfocados bien al desarrollo de nuevas técnicas que no se vean influenciadas excesivamente por la modelización del fenómeno natural (Estadística Robusta), bien al análisis de la calidad de los datos (Análisis de Observaciones Atípicas), o bien al estudio de aquellas observaciones que afectan considerablemente a los resultados del análisis. En este tercer enfoque se han propuesto un conjunto de métodos englobados en lo que genéricamente se conoce como el **Análisis de Influencia**.

En esta memoria, se aborda el problema de la influencia en una técnica clásica: el Análisis de Componentes Principales.

Con objeto de introducir adecuadamente los conceptos necesarios para el desarrollo de las técnicas recogidas en esta memoria y realizar una síntesis del estado actual del problema considerado, en el presente capítulo se comienza con una introducción al Análisis de Influencia (sección 1.1). Posteriormente, se recogen los conceptos y notación básicos del Análisis de Componentes Principales (sección 1.2), para culminar con una revisión del problema de la influencia en esta técnica, en la sección 1.3.

1.1 El problema de la influencia

Uno de los problemas con los que generalmente se enfrenta el estadístico al realizar un estudio es analizar el comportamiento de las observaciones

frente al modelo estadístico, así como la fiabilidad y precisión de las mismas.

Ello puede ser realizado, entre otros, bajo el enfoque del Análisis de Influencia, es decir, el análisis de aquellas observaciones con un efecto considerable sobre los resultados y conclusiones de las técnicas estadísticas aplicadas en el estudio.

Desde el punto de vista práctico, el Análisis de Influencia se ha basado, principalmente, en la comparación de los resultados, a través de algún estadístico de interés, obtenidos al considerar u omitir las observaciones que pueden provocar conclusiones poco fiables. Así, como recogen textualmente Kotz y Johnson [38]:

> "Las observaciones son consideradas como influyentes si su omisión de los datos da lugar a cambios sustanciales en rasgos importantes del análisis."

La influencia debe analizarse en distintos estadísticos de interés en el modelo estadístico bajo estudio, pudiendo ocurrir que, para algunos de ellos, la observación sea altamente influyente y no para otros. Por lo tanto, para hablar de observaciones influyentes hay que concretar el estadístico afectado. De forma general se puede decir que una observación debe considerarse como **influyente** si, bien de forma individual o bien conjuntamente con otras, tiene un mayor impacto que el resto de las observaciones sobre los valores de varias estimaciones y/o estadísticos.

El Análisis de Influencia surge a partir de los estudios sobre la sensibilidad de las técnicas estadísticas dentro del área de la robustez y de los métodos de análisis de la calidad de los datos a través de las técnicas de detección de observaciones atípicas. Hampel [23][1], [24][2], [25] introdujo, con el objetivo de estudiar la robustez de los estimadores, el concepto de función de influencia. Entre los primeros trabajos realizados sobre el Análisis de Influencia mediante la función de influencia, se encuentran los de Mallows [42][3], [43][4] y [44][5] en los que se introdujeron distintas versiones muestrales y se obtuvo la función de influencia para el coeficiente de correlación poblacional en el caso bivariante. Esto fue aplicado por Devlin y otros [17] a la detección de observaciones atípicas y a la estimación robusta.

En general, el estudio de la influencia se ha desarrollado principalmente en el Modelo de Regresión Lineal. En esta línea argumental, cabe citar

[1] Referencias en Hampel [25], Radhakrishnan y Kshirsagar [55], Gnanadesikan [22].

[2] Referencias en Devlin y otros [17], Gnanadesikan [22], Radhakrishnan y Kshirsagar [55].

[3] Referencia en Gnanadesikan [22].

[4] Referencia en Cook y Weisberg [13].

[5] Referencias en Campbell [10], Devlin y otros [17].

los trabajos de Belsley y otros [4], Cook y Weisberg [13], Atkinson [2] y Chatterjee y Hadi [12].

A través de los numerosos trabajos sobre el Análisis de Influencia en técnicas estadísticas, se puede observar la necesidad de contemplar adecuadamente tres rasgos de interés: los estadísticos y/o estimaciones sobre los que se centra el análisis, el carácter de influencia individual o conjunta de cada observación y el impacto o efecto sobre los resultados considerados en relación al resto de las observaciones.

Para medir o cuantificar el impacto antes citado, se ha seguido, generalmente, el enfoque de Cook y Weisberg [13] que, textualmente, describe como sigue:

> "La idea básica en el Análisis de Influencia es muy simple. Se introducen pequeñas perturbaciones en la formulación del problema y, entonces, se calcula cuánto cambian los resultados del análisis por la perturbación."

En la literatura, el esquema de perturbación más extendido es el de la omisión de las observaciones a las que se le pretende estudiar su influencia, cuantificando la diferencia entre los resultados obtenidos para el modelo postulado inicialmente y el modelo perturbado.

Cook [14] introduce la influencia local como un método alternativo para el estudio de la influencia. Consiste en la evaluación de la influencia, de forma local, al considerar un esquema de perturbación. En esta línea se pueden citar los trabajos de Beckman, Nachtsheim y Cook [3], Lawrence [41], Thomas y Cook [69], Lei Shi [59].

Muñoz Pichardo y otros [50], introducen un enfoque distinto para el estudio de influencia, inicialmente en el campo del Modelo Lineal General, mediante el concepto del sesgo condicionado. Este enfoque se relacionó posteriormente con la influencia local (Muñoz Pichardo y otros [51]), y se generalizó para el caso del Modelo Lineal General Multivariante (Muñoz Pichardo y otros [52]). Más tarde se ha aplicado en otros campos como en estimaciones bootstrap (Jiménez [33], Jiménez y otros [34]) y en el muestreo en poblaciones finitas (Moreno y otros [48]).

1.1.1 Funciones de influencia

En gran medida, el Análisis de Influencia se ha desarrollado a partir de la función de influencia y de sus distintas versiones muestrales. En este apartado se recogen los conceptos básicos sobre este tema.

Sea $\mathcal{F}_p$ el conjunto de funciones de distribución de variables aleatorias p-dimensionales. Dado $\underline{x} = (x_1, ..., x_p)' \in \mathbb{R}^p$, se denota por $\delta_{\underline{x}}$ a la función

de distribución de la variable degenerada en $\underline{x}$, es decir, si $\underline{t} = (t_1, ..., t_p)' \in \mathbb{R}^p$,

$$\delta_{\underline{x}}(\underline{t}) = \begin{cases} 1 & si \quad t_i \geq x_i \qquad \forall\, i = 1, ..., p \\ 0 & en\ caso\ contrario. \end{cases}$$

Hampel [23], [24], [25] define la función de influencia de un funcional real que se recoge a continuación.

Definición 1.1.1 *Sea T un funcional, $T : G \subset \mathcal{F}_p \longrightarrow \mathbb{R}$, donde G es un conjunto tal que $\forall\ F \in G$, $\forall\ \varepsilon \in (0,1)$, $\forall\ \underline{x} \in \mathbb{R}^p$, la mixtura de F y $\delta_{\underline{x}}$ dada por*

$$F_{\varepsilon}^{\underline{x}} = (1 - \varepsilon) F + \varepsilon \delta_{\underline{x}} \in G. \tag{1.1}$$

*La **función de influencia** de T en F se define por*

$$I(\underline{x}; T, F) = \lim_{\varepsilon \to 0^+} \frac{T\left(F_{\varepsilon}^{\underline{x}}\right) - T(F)}{\varepsilon}. \tag{1.2}$$

para todo $\underline{x}$ del espacio muestral en el que exista el límite.

Aunque Hampel [25] define la función de influencia ("curva de influencia", originalmente), como el límite direccional por la derecha dado en (1.2), en la bibliografía se ha sustituido por el límite no direccional (Devlin y otros [17], Campbell [10], etc.). La utilidad de esta variante de la definición originaria es necesaria para la justificación de la función de influencia muestral, versión muestral de la misma, definida posteriormente. No obstante, la mixtura (1.1) con ε un valor negativo próximo a cero únicamente es una función de distribución si $\underline{x}$ es un punto de discontinuidad de F.

Hampel [25] comenta la utilidad de los desarrollos en serie de Taylor, del funcional bajo estudio, en el cálculo de su función de influencia, ya que ésta viene dada por las derivadas parciales, las cuales constituyen el segundo término del desarrollo en serie.

Nota 1.1.1 *Siempre que exista el límite dado en (1.2), se puede decir que la función de influencia representa, esencialmente, la primera derivada direccional de un funcional.*

A partir de esta idea, algunos autores proporcionan interpretaciones de la función de influencia:

- *Hampel y otros [26]:*

> *"La función de influencia describe el efecto de una contaminación infinitesimal en el punto $\underline{x}$ sobre T, tipificado por la masa de la contaminación".*

- *Cook y Weisberg [13]:*

> *"La función de influencia es una medida de la influencia en T al añadir una observación $\underline{x}$ cuando el tamaño muestral tiende a infinito".*

- *Critchley [15]:*

> *"La función de influencia es la tasa de cambio de T cuando F experimenta una transformación infinitesimal en la dirección de $\underline{x}$".*

En la práctica, el objetivo del Análisis del Influencia es evaluar la influencia que ejerce una observación sobre un estadístico determinado y, por ello, son necesarias versiones muestrales de la función de influencia. Existen diversas versiones, principalmente propuestas por Mallows [44], las cuales se definen a continuación.

Las primeras de ellas, función de influencia empírica y función de influencia empírica con omisión, están motivadas por el hecho de que la función de distribución empírica es una función de distribución y por tanto es posible calcular la función de influencia a través de ella.

En general, en esta memoria, dada $\underline{X}_1, ..., \underline{X}_n$ una muestra aleatoria del vector aleatorio $\underline{X}$, la función de distribución empírica basada en la realización muestral $\underline{x}_1, ..., \underline{x}_n$, se denotará por $\widehat{F}_n$. Al fundamentarse algunas versiones muestrales de la función de influencia en la omisión de las observaciones, se denotará con el superíndice (i) en un estadístico, cuando el cálculo del mismo se realice bajo la omisión del i-ésimo elemento de la muestra, $\underline{X}_i$ y por $\widehat{F}_{n-1}^{(i)}$ la función de distribución empírica bajo tal omisión.

Definición 1.1.2 *En las condiciones de la definición 1.1.1, sean $\underline{X}_1, \underline{X}_2, ..., \underline{X}_n$ una muestra aleatoria procedente de una variable aleatoria $\underline{X}$ p-dimensional y $\underline{x}_1, \underline{x}_2, ..., \underline{x}_n$ una realización muestral de la misma.*

Se define la ***función de influencia empírica*** *de T en $\underline{x}_i$, $i = 1, ..., n$, como*

$$FIE\left(\underline{x}_i; T\left(\widehat{F}_n\right)\right) = I\left(\underline{x}_i; T, \widehat{F}_n\right).$$

Se define la ***función de influencia empírica con omisión*** *de T en $\underline{x}_i$, $i = 1, ..., n$, como*

$$FIE_{(i)}\left(\underline{x}_i; T\left(\widehat{F}_n\right)\right) = I\left(\underline{x}_i; T, \widehat{F}_{n-1}^{(i)}\right).$$

Así, un funcional T aplicado a una función de distribución empírica es un estadístico calculado con la muestra correspondiente:

$$T\left(\widehat{F}_n\right) = T; \qquad T\left(\widehat{F}_{n-1}^{(i)}\right) = T^{(i)}.$$

La siguiente versión muestral, fue utilizada en primer lugar por Mallows [44], aunque su nombre se debe a Devlin y otros [17]. Se basa en el hecho de que, para tamaño muestral suficientemente grande, la función de influencia muestral se puede considerar como una aproximación de la función de influencia eliminando el límite y tomando como función de distribución, la función de distribución empírica, $\widehat{F}_n$, y $\varepsilon = -\dfrac{1}{n-1}$. De este modo la perturbación de $\widehat{F}_n$ es la función de distribución empírica de una muestra de tamaño $n-1$, obtenida al eliminar la i-ésima observación, $\widehat{F}_{n-1}^{(i)}$,

$$\left[1 - \left(-\frac{1}{n-1}\right)\right] \widehat{F}_n + \left(-\frac{1}{n-1}\right) \delta_{\underline{x}_i} = \frac{n\widehat{F}_n - \delta_{\underline{x}_i}}{n-1} = \widehat{F}_{n-1}^{(i)}. \qquad (1.3)$$

Al ser $\underline{x}_i$ un punto de discontinuidad de $\widehat{F}_n$, es válida la extensión del límite dado en (1.2) a valores negativos de ε.

La definición que se expone a continuación, es algo más general del planteamiento realizado previamente, al definirse para un estadístico cualquiera, sin ser necesaria la existencia de un funcional que lo determine.

Definición 1.1.3 *Sean $\underline{X}_1, \underline{X}_2, ..., \underline{X}_n$ una muestra aleatoria procedente de una variable aleatoria $\underline{X}$ y $\underline{x}_1, \underline{x}_2, ..., \underline{x}_n$ una realización muestral de la misma. Se define la* ***función de influencia muestral*** *de un estadístico $T = T(\underline{X}_1, ..., \underline{X}_n)$ como*

$$FIM(\underline{x}_i; T) = (n-1)\left(T - T^{(i)}\right).$$

La función de influencia muestral, en definitiva, realiza una comparación de los estadísticos calculados sobre la muestra completa y omitiendo una de las observaciones. Esta idea ya se había utilizado anteriormente por Hampel [25], para estudiar la influencia que ejerce una observación sobre un estadístico.

Existen otras versiones muestrales, como la curva de sensibilidad (Andrews y otros [1][6]), utilizada inicialmente en estimadores de localización robustos, que consiste en la comparación de los estadísticos obtenidos con las observaciones de una muestra y los obtenidos al añadir una nueva observación a la muestra anterior.

[6] Referencia en Gnanadesikan [22].

1.1.2 Funciones de influencia conjunta

Las definiciones 1.1.1, 1.1.2 y 1.1.3 se pueden generalizar para un conjunto de puntos u observaciones. Para ello, es necesario extender la perturbación de la función de distribución dada en (1.1). La extensión de la función de influencia se debe a Fung [20].

Definición 1.1.4 *Sea T un funcional, $T : G \subset \mathcal{F}_p \longrightarrow \mathbb{R}$, donde G es un conjunto tal que $\forall\ F \in G$, $\forall\ \varepsilon \in (0,1)$, $\forall\ \underline{x}_1, ..., \underline{x}_r \in \mathbb{R}^p$, la mixtura de F y $\delta_{\underline{x}_i}$, $i = 1, ..., r$, dada por*

$$F_\varepsilon^{\mathbf{x}_I} = (1 - r\varepsilon)\, F + \varepsilon \sum_{i=1}^{r} \delta_{\underline{x}_i} \in G, \tag{1.4}$$

donde $\mathbf{x}_I = \{\underline{x}_1, ..., \underline{x}_r\}$.

*La **función de influencia conjunta** de T en F se define por*

$$I(\mathbf{x}_I; T, F) = \lim_{\varepsilon \to 0^+} \frac{T\left(F_\varepsilon^{\mathbf{x}_I}\right) - T(F)}{\varepsilon},$$

para todo conjunto de observaciones $\mathbf{x}_I$ del espacio muestral en el que exista el límite.

Se puede comprobar que para $F \in \mathcal{F}_p$, si ε es un valor negativo próximo a cero, $F_\varepsilon^{\mathbf{x}_I} \in \mathcal{F}_p$ únicamente si todos los puntos de $\mathbf{x}_I$ son puntos de discontinuidad para F.

Para extender la definición de la función de influencia muestral, se denotará por $\widehat{F}_{n-r}^{(I)}$ la función de distribución empírica construida a partir de una muestra de tamaño n tras la omisión de r observaciones con índices en I, y en general a los estadísticos, T, calculados bajo la omisión de observaciones con índices en I, por el superíndice (I), $T^{(I)}$.

Definición 1.1.5 *Sea $\underline{X}_1, ..., \underline{X}_n$ una muestra aleatoria de un vector $\underline{X}$ y $\underline{x}_1, ..., \underline{x}_n$ una realización muestral. Si $I \subset \{1, 2, ..., n\}$ tal que $card(I) = r$, y $\mathbf{x}_I = \{\underline{x}_i \ / \ i \in I\}$, se define la **función de influencia muestral conjunta** de un estadístico T como*

$$FIM(\mathbf{x}_I; T) = (n - r)\left[T\left(\widehat{F}_n\right) - T\left(\widehat{F}_{n-r}^{(I)}\right)\right].$$

Teniendo en cuenta que $T\left(\widehat{F}_n\right) = T$ y $T\left(\widehat{F}_{n-r}^{(I)}\right) = T^{(I)}$, son estadísticos, una definición más general para la función de influencia muestral conjunta viene dada por

$$FIM(\mathbf{x}_I; T) = (n - r)\left(T - T^{(I)}\right)$$

sin ser necesario que exista un funcional que determine el estadístico T.

1.1.3 Sesgo condicionado

Una técnica alternativa a las funciones de influencia para llevar a cabo el Análisis de Influencia, se basa en el concepto de sesgo condicionado, introducido por Muñoz Pichardo y otros [50].

Definición 1.1.6 *Sea* $\underline{X}_1, ..., \underline{X}_n$ *una muestra aleatoria de un vector aleatorio* $\underline{X}$, $\underline{x}_1, ..., \underline{x}_n$ *una realización muestral de la misma y* $T = T(\underline{X}_1, ..., \underline{X}_n)$ *un estadístico. Sea* $I \subset \{1, 2, ..., n\}$ *tal que* $card(I) = r$, $\mathbf{X}_I = \{\underline{X}_i \ / \ i \in I\}$ *y* $\mathbf{x}_I = \{\underline{x}_i \ / \ i \in I\}$. *El sesgo condicionado de* T *dado el conjunto de observaciones* $\mathbf{x}_I$ *se define como*

$$S(\mathbf{x}_I; T) = E[T \mid \mathbf{X}_I = \mathbf{x}_I] - E[T]$$

En particular, para $I = \{i\}$, el sesgo condicionado, $S(\underline{x}_i; T)$, se puede interpretar como el efecto medio que la realización del i-ésimo elemento de la muestra produce sobre el estadístico T. Por ello, en general, $S(\mathbf{x}_I; T)$ puede considerarse como una herramienta de evaluación de la influencia conjunta que ejercen sobre T las observaciones que constituyen $\mathbf{x}_I$.

El sesgo condicionado depende de la distribución del estadístico T y de los valores observados con índices en I. Por tanto, al contrario de la función de influencia muestral, que se define a partir de toda la muestra, el sesgo condicionado mide la influencia del valor observado sobre el estadístico, en términos de la esperanza de su distribución muestral, y por tanto, es independiente de cualquier realización muestral concreta.

Sin embargo, la dependencia de la distribución del estadístico T, provoca que, en general, el sesgo condicionado sea un parámetro poblacional desconocido que es necesario estimar. Con este objetivo, Muñoz Pichardo y otros [50] enuncian el siguiente resultado.

Teorema 1.1.1 *Sea* $\underline{X}_1, ..., \underline{X}_n$ *una muestra aleatoria de un vector aleatorio,* $\underline{X}$, *cuya distribución depende de un parámetro desconocido* $\theta \in \Theta \subset \mathbb{R}^k$. *Sea* $I \subset \{1, 2, ..., n\}$ *tal que* $card(I) = r$, $\mathbf{X}_I = \{\underline{X}_i \ / \ i \in I\}$ *y* $\mathbf{x}_I = \{\underline{x}_i \ / \ i \in I\}$. *Sea* $\widehat{\theta}$ *un estimador insesgado de* θ *y sea* $\widehat{\theta}^{(I)}$ *el estimador obtenido bajo la omisión de las observaciones con índices en* I. *Entonces,*

$$E\left[\widehat{\theta} - \widehat{\theta}^{(I)} \mid \mathbf{X}_I = \mathbf{x}_I\right] = S\left(\mathbf{x}_I; \widehat{\theta}\right).$$

En las condiciones del teorema 1.1.1, Muñoz Pichardo y otros [50] sugieren como estimador del sesgo condicionado

$$\widehat{S}\left(\mathbf{x}_I; \widehat{\theta}\right) = \widehat{\theta} - \widehat{\theta}^{(I)}.$$

El concepto del sesgo condicionado puede aplicarse al Análisis de Influencia de cualquier estadístico en cualquier modelo. Su definición genérica ha permitido su aplicación, además de al Modelo Lineal General (Muñoz Pichardo y otros [50], [51]) y su extensión al caso multivariante (Muñoz Pichardo y otros [52]), a diversos campos como las técnicas de remuestreo (Jiménez [33], Jiménez y otros [34]) y el muestreo en poblaciones finitas (Moreno y otros [48]).

1.1.4 Diagnósticos escalares de influencia

Tanto las funciones de influencia como el sesgo condicionado tienen la misma dimensión que el parámetro o estadístico analizado. Por ello, en el caso en el que éste sea un vector o una matriz, para evaluar y comparar la influencia que ejerce una observación o un conjunto de observaciones sobre el estadístico, es necesario el uso de diagnósticos de influencia de tipo escalar. Un tipo de medidas unidimensionales utilizadas se obtienen a partir de normas vectoriales o matriciales, según el caso.

Con este objetivo, Cook y Weisberg [13], Belsley y otros [4], proponen una serie de normas. Entre ellas destaca el uso de las $(\mathbf{Q}, c)$-normas, donde $\mathbf{Q}$ es una matriz de dimensión $q \times q$ simétrica definida positiva y c un escalar positivo. Este tipo de normas se define para un vector q-dimensional, $\underline{\varphi}$, de la forma

$$\|\underline{\varphi}\|_{(\mathbf{Q},c)} = \frac{1}{c}\underline{\varphi}'\mathbf{Q}\underline{\varphi}.$$

Muñoz Pichardo y otros [52], utilizan una generalización de la anterior, la $(\mathbf{Q}, \mathbf{C})$-norma, donde $\mathbf{Q}$ una matriz simétrica definida positiva de dimensión $q \times q$ y $\mathbf{C}$ una matriz simétrica no singular de dimensión $d \times d$. Esta norma matricial se define, para una matriz $\mathbf{M}$ de dimensión $q \times d$ de la forma

$$\|\mathbf{M}\|_{(\mathbf{Q},c)} = \left[tr\left(\mathbf{M}'\mathbf{Q}\mathbf{M}\mathbf{C}^{-1}\right)\right]^{\frac{1}{2}}.$$

1.2 El Análisis de Componentes Principales

Con el objetivo de centrar la notación a utilizar en este trabajo, en el presente apartado se introducen algunos conceptos y resultados básicos dentro del Análisis de Componentes Principales (ACP).

Esta técnica fue inicialmente descrita por Pearson [54] y posteriormente desarrollada por Hotelling [28], y tiene como objetivo esencial la reducción de la dimensión del vector de variables bajo estudio.

El Análisis de Componentes Principales consiste en obtener una rotación de los ejes de coordenadas, donde se representan las variables originales, a un nuevo sistema de coordenadas que resalta las propiedades optimales de la variación de las variables.

El estudio de componentes principales está directamente relacionado con la teoría de autovalores y autovectores de matrices simétricas semidefinidas positivas, al ser las componentes principales combinaciones lineales de las variables originales, dadas por los distintos autovectores de la matriz de covarianzas o de correlación del vector multidimensional que describen. Además, la varianza de cada componente principal es el autovalor asociado al autovector correspondiente.

La utilidad de las componentes principales es básicamente descriptivo, más que inferencial. Las componentes principales proporcionan información descriptiva valiosa para una amplia variedad de datos. Pero la inferencia proporciona también algunas ideas útiles. Este campo está fundamentalmente desarrollado bajo condiciones de normalidad que serán las asumidas en este trabajo.

En los siguientes subapartados, se proporciona una serie de resultados útiles en el contexto del Análisis de Componentes Principales, que se recogen en libros clásicos de Análisis Multivariante (Seber [58], Muirhead [49]), o en algunos específicos del Análisis de Componentes Principales (Jolliffe [35], Flury [19], Jackson [30]).

1.2.1 Componentes principales poblacionales

Componentes principales poblacionales basadas en la matriz de covarianzas

Definición 1.2.1 *Sea $\underline{X}$ un vector aleatorio p-dimensional de media $\underline{\mu}$ y matriz de covarianzas Σ. Sean $\underline{\alpha}_1, \underline{\alpha}_2, ..., \underline{\alpha}_p$ autovectores de Σ, ortonormales y $\mathbf{A} = \left[\underline{\alpha}_1\ \underline{\alpha}_2\ \cdots\ \underline{\alpha}_p\right]'$. Sea $\underline{Y} = (Y_1, ..., Y_p)' = \mathbf{A}\left(\underline{X} - \underline{\mu}\right)$. Entonces se define la k-ésima* ***componente principal (poblacional)*** *de $\underline{X}$ como la k-ésima componente del vector $\underline{Y}$.*

Por lo tanto la expresión de la k-ésima componente principal de $\underline{X}$ es

$$Y_k = \underline{\alpha}_k' \left(\underline{X} - \underline{\mu}\right),$$

y el vector $\underline{Y}$ constituye el vector de componentes principales de $\underline{X}$.

Nota 1.2.1 *A partir de la definición 1.2.1 se pueden realizar las siguientes comentarios:*

- *Al considerar los autovectores unitarios, la k-ésima componente principal de* $\underline{X}$ *se pueden interpretar como la proyección ortogonal de* $\underline{X} - \underline{\mu}$ *en la dirección de* $\underline{\alpha}_k$. *Por lo tanto, la transformación realizada por las componentes principales produce una rotación de los ejes de coordenadas de las variables originales, llevando el vector* $\underline{X}$ *a un nuevo sistema de coordenadas.*

  En la distribución normal, donde las superficies de densidad constante son elipsoides,

$$\left(\underline{X} - \underline{\mu}\right)' \boldsymbol{\Sigma}^{-1} \left(\underline{X} - \underline{\mu}\right) = c,$$

  con c una constante positiva, la rotación lleva los ejes originales a los ejes de estos elipsoides.

- *Si todos los autovalores de* $\boldsymbol{\Sigma}$ *son simples, los autovectores unitarios asociados son únicos salvo signo, por lo que las componentes principales son únicas salvo signo. En cambio, si algún autovalor de* $\boldsymbol{\Sigma}$ *es múltiple, existe un subespacio de autovectores de dimensión estrictamente superior a uno asociado a él, por lo que por cada autovalor múltiple, existen infinitas componentes principales y como consecuencia, infinitos posibles vectores de componentes principales.*

Teorema 1.2.1 *Sea* $\underline{X}$ *un vector aleatorio p-dimensional de media* $\underline{\mu}$ *y matriz de covarianzas* $\boldsymbol{\Sigma}$. *Sean* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ *los autovalores de* $\boldsymbol{\Sigma}$ *e* $\underline{Y} = (Y_1, ..., Y_p)'$ *el vector de componentes principales de* $\underline{X}$.

Entonces, se verifica que

1. $E\left(\underline{Y}\right) = \underline{0}$.
2. $var\left(\underline{Y}\right) = diag\left\{\lambda_1, \lambda_2, ..., \lambda_p\right\}$.

Teorema 1.2.2 *Sea* $\underline{X}$ *un vector aleatorio p-dimensional de media* $\underline{\mu}$ *y matriz de covarianzas* $\boldsymbol{\Sigma}$. *Sean* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ *los autovalores de* $\boldsymbol{\Sigma}$, $\underline{\alpha}_1, \underline{\alpha}_2, ..., \underline{\alpha}_p$ *autovectores ortonormales asociados, respectivamente,* $\boldsymbol{\Lambda} = diag\left\{\lambda_1, \lambda_2, ..., \lambda_p\right\}$, *e* $\underline{Y} = (Y_1, ..., Y_p)'$ *vector de componentes principales de* $\underline{X}$.

Se verifican las siguientes propiedades:

1. $arg \max_{\substack{\underline{a} \in \mathbb{R}^p \\ \|\underline{a}\| = 1}} var\left(\underline{a}'\underline{X}\right) = \underline{\alpha}_1 \quad y \quad var\left(\underline{\alpha}_1'\underline{X}\right) = \lambda_1$.
2. $arg \max_{\underline{a} \in C_k} var\left(\underline{a}'\underline{X}\right) = \underline{\alpha}_k \quad y \quad var\left(\underline{\alpha}_k'\underline{X}\right) = \lambda_k$, *con*

$$C_k = \left\{\underline{a} \in \mathbb{R}^p / \ \|\underline{a}\| = 1 \ y \ \underline{a}'\underline{\alpha}_j = 0, \quad j = 1, ..., k-1\right\}, \ k = 1, ..., p.$$

3. $tr(\mathbf{\Lambda}) = tr(\mathbf{\Sigma})$ *y* $\det(\mathbf{\Sigma}) = \det(\mathbf{\Lambda})$.

A partir del teorema 1.2.2 se pueden realizar el siguiente comentario.

Nota 1.2.2 *Algunas medidas de la variabilidad total de un vector aleatorio vienen dadas por la traza o por el determinante de la matriz de covarianzas. Ésta última tiene la desventaja de ser muy sensible a autovalores pequeños. Por lo que la traza puede ser más adecuada.*

Considerando como medida de la variabilidad de $\underline{X}$, $tr(\mathbf{\Sigma})$, *la propiedad* **3** *del teorema 1.2.2 indica la igualdad de variabilidad total entre las variables originales y sus componentes principales. Por lo tanto, la razón*

$$\frac{\lambda_j}{\sum_{l=1}^{p} \lambda_l} = \frac{var(Y_j)}{tr(\mathbf{\Sigma})}$$

mide la contribución de la j*-ésima componente principal en la variación total de* $\underline{X}$. *Así, el número,* k, *de componentes principales a escoger para reducir la dimensión de un problema, puede ser determinado por la razón*

$$\frac{\sum_{j=1}^{k} \lambda_j}{tr(\mathbf{\Sigma})}.$$

Cuando esta razón es suficientemente próxima a la unidad, las k *primeras componentes principales describen un alto porcentaje de la variabilidad del problema inicial.*

Componentes principales poblacionales basadas en la matriz de correlación

Todo el desarrollo expuesto previamente se puede realizar sustituyendo la matriz de covarianzas del vector aleatorio $\underline{X}$ por su matriz de correlación. Esto es equivalente a desarrollar el estudio anterior sobre la tipificación de las componentes de dicho vector, $X_i^* = \dfrac{X_i - \mu_i}{\sigma_i}$.

Los análisis basados en la matriz de covarianzas o la matriz de correlación, son diferentes al ser distintos los autovalores de ambas matrices y por lo tanto puede ser diferente el número de componentes principales elegidas para reducir la dimensionalidad.

La elección entre el análisis de componentes principales, en función de la matriz de covarianzas, o de la matriz de correlación, dependerá de las características del problema. La desventaja de la utilización de Σ, estriba en la dependencia de las unidades de medida, que es salvada al considerar la matriz de correlación. Pero el desarrollo teórico y la interpretación de resultados se hace mucho más complejo con el segundo método.

1.2.2 Componentes principales muestrales

Componentes principales muestrales basadas en la matriz de covarianzas

En la práctica, el vector de medias $\underline{\mu}$ y la matriz de covarianzas Σ son parámetros desconocidos de la distribución de un vector $\underline{X}$, por lo que deben ser estimados mediante una muestra. Sea, por tanto, $\underline{X}_1, \underline{X}_2, ..., \underline{X}_n$ una muestra aleatoria del vector aleatorio $\underline{X}$.

Generalmente se consideran como estimador de la media poblacional, la media muestral, $\overline{\underline{X}}$, y como estimador de la matriz de covarianzas, o bien

$$\widehat{\Sigma} = \frac{1}{n}\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}},$$

o bien, el estimador insesgado de Σ

$$\mathbf{S} = \frac{1}{n-1}\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}},$$

donde $\mathbf{X} = \begin{bmatrix} \underline{X}_1 & \cdots & \underline{X}_n \end{bmatrix}'$, $\mathbf{1}_n = (1, ..., 1)' \in \mathbb{R}^n$, $\overline{\mathbf{X}} = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'\mathbf{X}$ y $\widetilde{\mathbf{X}} = \mathbf{X} - \overline{\mathbf{X}}$.

Como estimadores de los autovalores y autovectores poblacionales se toman los autovalores y autovectores del estimador de Σ correspondiente. Se utilizará la notación $\widehat{\lambda}_1, \widehat{\lambda}_2, ..., \widehat{\lambda}_p$ y $\widehat{\underline{\alpha}}_1, \widehat{\underline{\alpha}}_2, ..., \widehat{\underline{\alpha}}_p$, respectivamente, para $\widehat{\Sigma}$ y $\widetilde{\lambda}_1, \widetilde{\lambda}_2, ..., \widetilde{\lambda}_p$ y $\widetilde{\underline{\alpha}}_1, \widetilde{\underline{\alpha}}_2, ..., \widetilde{\underline{\alpha}}_p$, respectivamente, para $\mathbf{S}$. En algunas ocasiones se podrán enunciar resultados de forma conjunta para ambos métodos de estimación y para ello se utilizará $\widehat{\widehat{\Sigma}}$ para denotar a cualquiera de los dos estimadores de Σ, $\widehat{\Sigma}$ o $\mathbf{S}$ y $\widehat{\widehat{\lambda}}_1, \widehat{\widehat{\lambda}}_2, ..., \widehat{\widehat{\lambda}}_p$ y $\widehat{\widehat{\underline{\alpha}}}_1, \widehat{\widehat{\underline{\alpha}}}_2, ..., \widehat{\widehat{\underline{\alpha}}}_p$ para denotar los estimadores de $\lambda_1, \lambda_2, ..., \lambda_p$ y $\underline{\alpha}_1, \underline{\alpha}_2, ..., \underline{\alpha}_p$ respectivamente.

La relación existente entre los autovalores de ambas estimaciones la proporciona una constante multiplicativa y los autovectores coinciden en ambos casos:

$$\widetilde{\lambda}_j = \frac{n}{n-1}\widehat{\lambda}_j, \qquad \widehat{\underline{\alpha}}_j = \widetilde{\underline{\alpha}}_j. \tag{1.5}$$

En la estimación de los autovalores y los autovectores se plantea un problema de correspondencia respecto al orden de los autovalores poblacionales y muestrales, que se comentará en el capítulo 3 de este trabajo. Parece lógico estimar cada autovalor poblacional, por el autovalor muestral correspondiente en orden. Esta estimación conlleva a la estimación de los autovectores poblacionales por los autovectores muestrales asociados al autovalor correspondiente. Pero, puede ocurrir que, para tamaños muestrales no muy grandes, y especialmente ante la presencia de autovalores poblacionales próximos, se estimen los autovectores asociados de forma inadecuada. Esto puede llevar a equívocos en el Análisis de Influencia, como se verá en posteriores capítulos, problema al que se abordará en este trabajo.

Definición 1.2.2 *Sea $\underline{X}$ un vector aleatorio p-dimensional de media $\underline{\mu}$ y matriz de covarianzas $\boldsymbol{\Sigma}$. Sea $\underline{X}_1, \underline{X}_2, ..., \underline{X}_n$ una muestra aleatoria del vector aleatorio $\underline{X}$. Sean $\widehat{\underline{\alpha}}_1, \widehat{\underline{\alpha}}_2, ..., \widehat{\underline{\alpha}}_p$ autovectores de $\widehat{\boldsymbol{\Sigma}}$, ortonormales y $\widehat{\mathbf{A}} = \left[\widehat{\underline{\alpha}}_1\ \widehat{\underline{\alpha}}_2\ \cdots\ \widehat{\underline{\alpha}}_p\right]'$. Entonces, se define la k-ésima* ***componente principal muestral*** *como la k-ésima componente del vector*

$$\underline{Y}_i = (Y_{i1}, ..., Y_{ip})' = \widehat{\mathbf{A}}\left(\underline{X}_i - \overline{\underline{X}}\right).$$

Nota 1.2.3 *A la definición de la k-ésima componente principal muestral de $\underline{X}_i$ se le pueden dar distintas interpretaciones:*

- *La k-ésima componente principal es la proyección ortogonal de $\underline{X}_i$ en la dirección de $\widehat{\underline{\alpha}}_k$.*
- *Las componentes principales muestrales se pueden considerar como estimaciones de las componentes principales poblacionales, en el caso en el que los autovalores de $\boldsymbol{\Sigma}$ sean simples. En caso contrario, para un autovalor de multiplicidad mayor que la unidad, se tendrían distintas estimaciones.*
- *Bajo hipótesis de normalidad, si se considera la familia de elipsoides concéntricos*

  $$\left(\underline{X}_i - \overline{\underline{X}}\right)' \widehat{\boldsymbol{\Sigma}}^{-1} \left(\underline{X}_i - \overline{\underline{X}}\right) = c,$$

  con $c > 0$, la componente principal de $\underline{X}_i$ es la proyección de $\underline{X}_i$ en los ejes principales de la familia.

Resultados semejantes a los enunciados en los teoremas 1.2.1 y 1.2.2, para las componentes principales poblacionales, pueden ser formulados para sus versiones muestrales.

Al igual que en el caso poblacional, se pueden calcular las componentes principales muestrales a partir de la matriz de correlación muestral del vector aleatorio $\underline{X}$ a partir de las variables tipificadas $X_{ij}^{*} = \dfrac{X_{ij} - \widehat{\mu}_j}{\widehat{\sigma}_j}$.

Inferencia sobre las componentes principales basadas en la matriz de covarianzas

En esta memoria, el objetivo fundamental es la evaluación de la influencia mediante el sesgo condicionado, en los estadísticos de interés en el Análisis de Componentes Principales: autovalores y autovectores del estimador de $\boldsymbol{\Sigma}$. En este estudio será necesario el uso de algunos resultados inferenciales de las componentes principales bajo la hipótesis de normalidad de las variables estudiadas.

Teorema 1.2.3 *Sea $\underline{X}_1, \underline{X}_2, ..., \underline{X}_n$ una muestra aleatoria de un vector aleatorio $\underline{X} \sim \mathcal{N}_p\left(\underline{\mu}, \boldsymbol{\Sigma}\right)$. Entonces,*

1. *$\overline{\underline{X}}$ y $\widehat{\boldsymbol{\Sigma}}$ son los estimadores de máxima verosimilitud de $\underline{\mu}$ y $\boldsymbol{\Sigma}$ respectivamente.*

2. *Si $n - 1 \geq p$, se verifica, con probabilidad uno, que:*

   - *$\widehat{\widehat{\boldsymbol{\Sigma}}}$ es una matriz simétrica definida positiva.*
   - *$\widehat{\widehat{\lambda}}_1, \widehat{\widehat{\lambda}}_2, ..., \widehat{\widehat{\lambda}}_p$ son distintos y no nulos.*
   - *Los autovectores $\widehat{\widehat{\underline{\alpha}}}_1$ $\widehat{\widehat{\underline{\alpha}}}_2$ ... $\widehat{\widehat{\underline{\alpha}}}_p$ son únicos (salvo signo).*
   - *Las componentes principales muestrales son únicas (salvo signo).*

3. *Si $n \leq p$, con probabilidad uno, se verifica que $\widehat{\widehat{\boldsymbol{\Sigma}}}$ es una matriz singular y, por lo tanto, tiene autovalores nulos.*

Teorema 1.2.4 *Sea $\underline{X}_1, \underline{X}_2, ..., \underline{X}_n$ una muestra aleatoria de un vector aleatorio $\underline{X} \sim \mathcal{N}_p\left(\underline{\mu}, \boldsymbol{\Sigma}\right)$. Entonces,*

1. *Los autovalores y autovectores de $\widehat{\widehat{\boldsymbol{\Sigma}}}$ son asintóticamente normales.*

2. *$\left(\widehat{\widehat{\lambda}}_1, \widehat{\widehat{\lambda}}_2, ..., \widehat{\widehat{\lambda}}_p\right)$ y $\left(\widehat{\widehat{\underline{\alpha}}}_1, \widehat{\widehat{\underline{\alpha}}}_2, ..., \widehat{\widehat{\underline{\alpha}}}_p\right)$ son asintóticamente independientes.*

3. $\sqrt{\frac{n}{2}}\frac{\left(\widehat{\lambda}_k - \lambda_k\right)}{\lambda_k}$, $k = 1, ..., p$, *son asintóticamente independientes e idénticamente distribuidas según una ley* $\mathcal{N}(0,1)$.

4. $\sqrt{\frac{n-1}{2}}\frac{\left(\widetilde{\lambda}_k - \lambda_k\right)}{\lambda_k}$, $k = 1, ..., p$, *son asintóticamente independientes e idénticamente distribuidas según una ley* $\mathcal{N}(0,1)$.

Teorema 1.2.5 *Sea* $\underline{X}_1, \underline{X}_2, ..., \underline{X}_n$ *una muestra aleatoria de un vector aleatorio* $\underline{X} \sim \mathcal{N}_p(\underline{\mu}, \boldsymbol{\Sigma})$. *Sean* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ *los autovalores de* $\boldsymbol{\Sigma}$ *y* $\underline{\alpha}_1, \underline{\alpha}_2, ..., \underline{\alpha}_p$ *autovectores ortonormales asociados, respectivamente. Sean* $\widetilde{\lambda}_1 \geq \widetilde{\lambda}_2 \geq \cdots \geq \widetilde{\lambda}_p$ *los autovalores de* $\mathbf{S}$, *con autovectores asociados* $\underline{\widetilde{\alpha}}_1, \underline{\widetilde{\alpha}}_2, ..., \underline{\widetilde{\alpha}}_p$, *respectivamente. Si* λ_k *es simple, para tamaño muestral suficientemente grande, se verifica que*

$$\begin{aligned}
E\left(\widetilde{\lambda}_k\right) &= \lambda_k - \frac{1}{n-1}\sum_{j\neq k}\frac{\lambda_j\lambda_k}{\lambda_j - \lambda_k} + O\left(n^{-2}\right), \\
var\left(\widetilde{\lambda}_k\right) &= \frac{2\lambda_k^2}{n-1}\left[1 - \frac{1}{n-1}\sum_{j\neq k}\frac{\lambda_j^2}{\left(\lambda_j - \lambda_k\right)^2}\right] + O\left(n^{-3}\right), \\
cov\left(\widetilde{\lambda}_j, \widetilde{\lambda}_k\right) &= O\left(n^{-2}\right) \qquad si\ j \neq k, \\
E\left(\underline{\widetilde{\alpha}}_k\right) &= \underline{\alpha}_k + O\left(n^{-1}\right), \\
var\left(\underline{\widetilde{\alpha}}_k\right) &= \frac{\lambda_k}{n-1}\sum_{j\neq k}\frac{\lambda_j}{\left(\lambda_j - \lambda_k\right)^2}\underline{\alpha}_j\underline{\alpha}_j' + O\left(n^{-2}\right), \\
cov\left(\underline{\widetilde{\alpha}}_j, \underline{\widetilde{\alpha}}_k\right) &= -\frac{1}{n-1}\frac{\lambda_j\lambda_k}{\left(\lambda_j - \lambda_k\right)^2}\underline{\alpha}_j\underline{\alpha}_k' + O\left(n^{-2}\right), \qquad si\ j \neq k.
\end{aligned} \tag{1.6}$$

Los momentos recogidos en el teorema 1.2.5, relativos a los autovalores de $\mathbf{S}$, fueron obtenidos por Lawley [40], a partir de un desarrollo en serie de los autovalores muestrales en función de los autovalores poblacionales.

Teniendo en cuenta la relación existente entre los autovalores de $\widehat{\boldsymbol{\Sigma}}$ y $\mathbf{S}$, dada en (1.5), se derivan directamente los momentos de los autovalores de $\widehat{\boldsymbol{\Sigma}}$.

Otro aspecto de interés dentro de la inferencia en el Análisis de Componentes Principales es la falta de unicidad de las componentes principales, tal como se ha comentado en la nota 1.2.1. En el caso de un autovalor simple, λ_k, existen dos componentes principales, la construida con un autovector unitario asociado a λ_k, $\underline{\alpha}_k$, y la obtenida con el opuesto en signo, $-\underline{\alpha}_k$. En el caso

en el que el autovalor sea múltiple, existen infinitas componentes principales asociadas.

A veces, es conveniente plantear un contraste de hipótesis para detectar esta última situación. En la bibliografía se encuentran planteados distintos contrastes de hipótesis en este sentido, como en los trabajos de Lawley [40] y James [31].

1.3 Análisis de Influencia en el ACP

El Análisis de Componentes Principales es muy sensible ante la presencia de ciertas observaciones extremas, como muestra Huber [29] en un ejemplo en el que una observación determina una componente, que además es la de mayor varianza. En tal caso, las conclusiones obtenidas mediante el Análisis de Componentes Principales, pueden ser incorrectas. Por ello es necesario, o bien utilizar estimadores robustos de la matriz considerada en el cálculo de las componentes principales, o bien, seguir la línea del Análisis de Influencia.

En el Análisis de Componentes Principales los parámetros o estadísticos de interés son fundamentalmente los autovalores y autovectores de la matriz de covarianzas o correlación poblacional o muestral según el caso. Por ello el Análisis de Influencia se centra principalmente en ellos.

1.3.1 Funciones de influencia de autovalores y autovectores

En la bibliografía se encuentran distintos trabajos, tanto desde el punto de vista teórico como desde el punto de vista práctico, en los que se desarrollan y aplican las funciones de influencia y sus versiones muestrales al Análisis de Componentes Principales.

Desde el punto de vista poblacional, para autovalores poblaciones simples de la matriz de covarianzas, se pueden destacar diversos trabajos. Radhakrishnan y Kshirsagar [55] obtienen las funciones de influencia de los estadísticos de interés. Critchley [15] aporta además las versiones muestrales de las funciones de influencia. Tanaka [62] extiende los resultados anteriores al caso de autovalores múltiples. Utilizando la matriz de correlación existen pocos trabajos desde el punto de vista poblacional, debido a la complejidad de las expresiones de la función de influencia, como los de Calder [8][7] y [9][8]. Desde el punto de vista muestral, se pueden citar los trabajos de Pack y otros [53], Brooks [7] y Mertens [46].

[7] Referencia en Mertens [46].

[8] Referencias en Pack y otros [53], Brooks [7].

Desde otros puntos de vista, pero que derivan en la obtención de funciones de influencia, Tanaka y Tarumi, paralelamente a los primeros trabajos relativos a las funciones de influencia en el Análisis de Componentes Principales, desarrollan una teoría muy general y de gran aplicación para la evaluación de la influencia de observaciones tanto de forma individual como conjunta. Basándose en la teoría de perturbación de autovalores de matrices simétricas, abordan diversos campos: Análisis de Correspondencia (Tanaka [61]); Análisis de Correspondencia, Correlaciones Canónicas y Análisis de Componentes Principales (Tarumi [67]); Análisis Discriminante de variables categóricas (Tanaka y Tarumi [65]).

Tarumi y Tanaka [68] presentan un programa informático para llevar a cabo Análisis de Influencia en métodos estadísticos multivariantes, entre otros, en el Análisis de Componentes Principales: mediante la ponderación de observaciones individual o conjuntamente, mediante la omisión de observaciones individual o conjuntamente, utilizando técnicas de computación directa (recalculando los datos modificados) y utilizando la teoría de la perturbación tanto para autovalores simples como múltiples.

Tanaka y Tarumi [66] muestran la importancia de la hipótesis de simplicidad o multiplicidad de los autovalores en el Análisis de Influencia. Con un ejemplo pone de manifiesto cómo, bajo hipótesis de simplicidad la influencia en algunas observaciones es sensiblemente más elevada que bajo hipótesis de multiplicidad de autovalores.

El análisis de influencia conjunto se encuentra poco desarrollado en este campo, apareciendo únicamente alguna referencia como en el trabajo de Critchley [15].

Un problema en el Análisis de Componentes Principales es el del orden de las estimaciones de los autovalores, comentado anteriormente. Al utilizar la función de influencia muestral como herramienta de evaluación de la influencia, puede ocurrir que al comparar los autovalores muestrales calculados mediante todos los elementos de la muestra y omitiendo uno de ellos, éstos no sean los estimadores más adecuados para el mismo autovalor poblacional y provoque, principalmente en el autovector, la detección errónea de observaciones altamente influyentes.

Pack y otros [53] presentan un ejemplo práctico de esta situación. Algunos autores, Pack y otros [53], Brooks [7], proponen la evaluación de la función de influencia empírica y de la función de influencia muestral, que en general son similares, para la detección de este tipo de situaciones, para corregir la evaluación de la influencia.

Por otro lado, se plantea la evaluación adecuada de la influencia sobre los estadísticos de interés. Las funciones de influencia de los autovalores son unidimensionales, aunque puede tomar tanto valores positivos como nega-

tivos. La evaluación de la influencia mediante ésta se suele realizar tomando el valor absoluto de la función de influencia.

La función de influencia de un autovector es un vector de la misma dimensión, por lo que para la evaluación de la influencia es necesario considerar alguna estrategia para transformarlas en diagnósticos escalares, como utilizar la norma euclídea de las versiones muestrales de la función de influencia o estudiando el ángulo formado por el autovector muestral calculado mediante todas las observaciones y el obtenido al omitir una de ellas (Pack y otros [53], Brooks [7]). Critchley [15] hace referencia al uso de normas utilizadas en la regresión por Cook y Weisberg [13]. Existen, además, algunas propuestas de normalizaciones de los diagnósticos basados en las funciones de influencia (Mertens [46]).

Las funciones de influencia de los parámetros de interés en el Análisis de Componentes Principales son de importancia en esta memoria, pues se puede establecer cierta relación con el sesgo condicionado de las estimaciones de éstos. Por ello, en el siguiente capítulo, se expone con más detalle los trabajos publicados al respecto.

1.3.2 Influencia en espacios generados por componentes principales

A veces el interés no se centra en cada componente principal, sino en el espacio generado por las componentes principales dominantes, es decir, las de mayor varianza, como ocurre cuando el objetivo del Análisis de Componentes Principales es la reducción de la dimensión del vector de variables bajo estudio. Por ello, Tanaka [62] desarrolla un conjunto de medidas de influencia para evaluar la influencia sobre el espacio generado por las componentes principales. La posibilidad de construir dichas medidas a partir de las funciones de influencia de autovalores y autovectores no es recomendable debido a la inestabilidad y tendencia a contener grandes errores computacionales en el caso de existencia de autovalores cercanos. Por ello, Tanaka [62] considera la descomposición espectral

$$\boldsymbol{\Sigma} = \mathbf{A}_1'\boldsymbol{\Lambda}_1\mathbf{A}_1 + \mathbf{A}_2'\boldsymbol{\Lambda}_2\mathbf{A}_2,$$

donde

$$\boldsymbol{\Lambda}_1 = diag\left\{\lambda_1, \ldots, \lambda_k\right\}, \qquad \boldsymbol{\Lambda}_2 = diag\left\{\lambda_{k+1}, \ldots, \lambda_p\right\},$$
$$\mathbf{A}_1 = \left[\underline{\alpha}_1, \ldots, \underline{\alpha}_k\right]', \qquad \mathbf{A}_2 = \left[\underline{\alpha}_{k+1}, \ldots, \underline{\alpha}_p\right]'.$$

Para analizar la influencia sobre el subespacio generado por las k primeras componentes principales, se basa en la estabilidad de la descomposición. Para ello considera las siguientes matrices

$$\mathbf{T}_1 = \mathbf{A}_1'\mathbf{\Lambda}_1\mathbf{A}_1 \qquad \text{y} \qquad \mathbf{T}_2 = \mathbf{A}_2'\mathbf{A}_2.$$

La estabilidad dependerá directamente de la matriz $\mathbf{T}_1$, primera parte de la descomposición, e indirectamente de $\mathbf{T}_2$, operador proyección sobre el subespacio $L(\mathbf{A}_1)$, generado por los autovectores asociados a los k autovalores de interés.

La influencia del subespacio generado por las primeras componentes principales se evalúa mediante las funciones de influencia de ambas matrices, cuya expresión también se recoge en el trabajo de Tanaka [62]. Desde el punto de vista práctico, el análisis se realiza mediante aproximaciones basadas en desarrollos en serie. Resultados de este trabajo fueron determinados con mayor precisión por Tanaka y Castaño [64], al añadir términos de orden cuadrático a las anteriores, que se obtenían solamente con términos lineales.

Bénasséni [6] estudió un enfoque alternativo para cuantificar la sensibilidad del espacio generado por las componentes principales dominantes, mediante un desarrollo infinitesimal de la matriz de coeficientes. Entre otros coeficientes, analiza el desarrollo del coeficiente RV de Escoufier (Escoufier[18], Robert y Escoufier [57]).

Existen otros enfoques para llevar a cabo el Análisis de Influencia en el Análisis de Componentes Principales, como los recogidos en los trabajos de Krzanowski [39], Bénasséni [5], Daudin y otros [16], y Shi [59].

Capítulo 2

FUNCIONES DE INFLUENCIA EN EL ACP

2.1 Introducción

En el Análisis de Componentes Principales, al igual que en otras técnicas estadísticas, el Análisis de Influencia se ha desarrollado fundamentalmente desde la perspectiva de las funciones de influencia. Los parámetros de interés, como ya se ha comentado, son los autovalores y autovectores de la matriz de covarianzas, o de la matriz de correlación, según el procedimiento utilizado.

En este capítulo, en primer lugar, se realiza una síntesis de los estudios de influencia publicados hasta el momento, en la línea de las funciones de influencia.

En las componentes basadas en la matriz de correlación, las expresiones de las funciones de influencia de los autovalores y autovectores son más complejas que para la matriz de covarianzas. En este capítulo se obtienen unas expresiones algo más sencillas que las recogidas en la bibliografía, lo que permiten su interpretación y comparación con la función de influencia de estos parámetros relativos a la matriz de covarianzas.

Los trabajos publicados en el campo de la influencia suelen considerar $\widehat{\boldsymbol{\Sigma}}$ como estimación de la matriz de covarianzas poblacionales para las componentes principales muestrales basadas en dicha matriz. Esto se debe a su obtención directa a partir de la función de influencia. En cambio, en la práctica, el estimador más utilizado es $\mathbf{S}$ por ser insesgado de $\boldsymbol{\Sigma}$. Resulta conveniente, por tanto, utilizar una medida de influencia coherente con el método aplicado en el cálculo de las componentes principales. En el presente capítulo se muestra que no es posible determinar un funcional, similar al que proporciona la matriz de covarianzas, tal que evaluado en cualquier función

de distribución empírica asigne la matriz de covarianzas muestrales, $\mathbf{S}$, de la muestra de la que procede.

No obstante, tiene sentido calcular la función de influencia muestral de los autovalores y autovectores de $\mathbf{S}$, cuyas expresiones se aportan en el presente capítulo.

Por otro lado, en algunas ocasiones también es conveniente, desde el punto de vista práctico, el estudio de la influencia conjunta ejercida por varias observaciones. En la bibliografía se encuentran algunas indicaciones del camino a seguir en tal caso, aunque se halla poco desarrollado. En el último apartado del capítulo, se generaliza al caso múltiple, la expresión de las funciones de influencia de los parámetros de interés, tanto desde el punto de vista poblacional como muestral.

En consecuencia, este capítulo se estructura en tres secciones. La primera de ellas recoge diversos resultados enfocados a la obtención de la función de influencia de los autovalores y autovectores de la matriz de covarianzas, así como de la matriz de correlación. La segunda sección, tiene como objetivo mostrar distintas versiones muestrales de la misma. En la tercera sección se realiza una extensión del análisis de influencia para el caso múltiple, en el que se pretende determinar mediante la generalización de la función de influencia, el efecto que producen simultáneamente varias observaciones.

En este capítulo se utilizarán los funcionales $\underline{\mu}$ y $\boldsymbol{\Sigma}$, que se definen a continuación.

Si $\mathcal{F}_p^j$ es el conjunto de las funciones de distribución p-dimensionales con momento de orden j, finito, y $\mathcal{S}_p$ el conjunto de las matrices simétricas de dimensión p,

$$\underline{\mu} : \mathcal{F}_p^1 \longrightarrow \mathbb{R}^p, \qquad \boldsymbol{\Sigma} : \mathcal{F}_p^2 \longrightarrow \mathcal{S}_p,$$

son tales que, si $F \in \mathcal{F}_p^2$ y $\underline{X}$ es una variable aleatoria con función de distribución F,

$$\underline{\mu}(F) = E[\underline{X}], \qquad \boldsymbol{\Sigma}(F) = var(\underline{X}).$$

2.2 Funciones de influencia

Dado que el Análisis de Componentes Principales puede basarse en la matriz de covarianzas o en la matriz de correlación, es necesario distinguir entre ambos, para la obtención de las funciones de influencia de los parámetros de interés.

2.2.1 Estudio basado en la matriz de covarianzas

En los primeros años de la década de los ochenta, se siguieron diversos procedimientos, prácticamente simultáneos y poco conexos, aunque sí complementarios, para la obtención de las funciones de influencia de autovalores simples de la matriz de covarianzas y sus autovectores asociados.

El primero de ellos lo propusieron Radhakrishnan y Kshirsagar [55], con una sencilla demostración, basada en la expresión de $\underline{\mu}$ y $\boldsymbol{\Sigma}$ tras perturbar la función de distribución, en función de la media y matriz de covarianzas de la distribución original.

En segundo lugar, Tanaka [61], aunque no da explícitamente el valor de la función de influencia, obtiene el desarrollo en serie de los autovalores y autovectores de matrices simétricas, de los cuales se puede deducir dicha función. Los resultados dados por Tanaka son bastante generales dentro de la teoría de perturbación de autovalores en matrices reales simétricas, teoría que pretende determinar la variación que sufren los autovalores y los autovectores de matrices simétricas, tras realizar una perturbación simétrica a la matriz correspondiente. Simultáneamente, Critchley [15], también basándose en la teoría de perturbación, aportó las expresiones de los coeficientes del desarrollo para el caso concreto del Análisis de Componentes Principales, y de las funciones de influencia de los parámetros de interés, llegando a una expresión para el autovector más simple que la de Radhakrishnan y Kshirsagar [55]. La obtención de los coeficientes de los desarrollos en serie, además de proporcionar directamente el valor de la función de influencia, son de gran utilidad en el cálculo de las versiones muestrales. En este apartado se exponen, principalmente, los resultados obtenidos por Tanaka [61], por ser más generales y, en consecuencia, de mayor utilidad para las aportaciones realizadas en este capítulo, y se obtendrán como casos particulares los proporcionados por Critchley [15].

Tanaka y Tarumi [65] obtuvieron, también de forma general, la expresión de los coeficientes de los desarrollos para los autovalores y autovectores asociados, en el caso de que el autovalor de interés sea múltiple.

Dentro de la Teoría de la Perturbación, conviene también mencionar a autores como Wilkinson [71], Rellich [56], Sibson [60], y Kato [36], que desarrollan la base necesaria que permite dar coherencia al Análisis de Influencia en éste y otros campos.

En primer lugar, se enuncian dos resultados generales que fundamentan el cálculo de las funciones de influencia de autovalores y autovectores de la matriz de covarianzas. El primero de ellos, debido a Rellich [56], asegura la existencia de desarrollos en serie convergentes de los autovalores y autovectores de una matriz perturbada, cuya perturbación se puede expresar según

un desarrollo en serie convergente. El segundo de ellos, debido a Tanaka [61], proporciona los coeficientes de dichos desarrollos. En realidad, la versión que se expone de éste, es posterior, Tanaka y Castaño [64], en la que se utiliza una notación más simple.

Teorema 2.2.1 *Sean* $\mathbf{B} \in \mathcal{S}_p$, $\lambda_1 \geq \cdots \geq \lambda_p$ *los autovalores de* $\mathbf{B}$ *y* $\underline{\alpha}_1, ..., \underline{\alpha}_k$ *autovectores ortonormales asociados, respectivamente. Dado* $\varepsilon \in \mathbb{R}$, *sea* $\mathbf{B}(\varepsilon) \in \mathcal{S}_p$ *una perturbación de la matriz* $\mathbf{B}$, *que se puede expresar según un desarrollo en serie de potencias convergente de la forma*

$$\mathbf{B}(\varepsilon) = \mathbf{B} + \varepsilon \mathbf{B}_1 + \frac{1}{2}\varepsilon^2 \mathbf{B}_2 + O\left(\varepsilon^3\right),$$

donde $\mathbf{B}_1$, $\mathbf{B}_2 \in \mathcal{S}_p$. *Entonces, existe un conjunto de autovectores de* $\mathbf{B}(\varepsilon)$, $\underline{\alpha}_k(\varepsilon)$, $k = 1, ..., p$, *ortonormales, tales que, ellos y sus autovalores asociados,* $\lambda_k(\varepsilon)$, $k = 1, ..., p$, *respectivamente, se pueden expresar según desarrollos en serie de potencias convergentes de la forma*

$$\lambda_k(\varepsilon) \equiv \lambda_k + \varepsilon \nu_k + \frac{1}{2}\varepsilon^2 \pi_k + O\left(\varepsilon^3\right), \qquad k = 1, ..., p,$$

$$y \quad \underline{\alpha}_k(\varepsilon) \equiv \underline{\alpha}_k + \varepsilon \underline{\beta}_k + \frac{1}{2}\varepsilon^2 \underline{\gamma}_k + O\left(\varepsilon^3\right), \qquad k = 1, ..., p.$$

A partir del teorema 2.2.1 se pueden realizar el siguiente comentario.

Nota 2.2.1 *Para valores de* ε *suficientemente pequeños, se mantiene el orden entre los autovalores de* $\mathbf{B}(\varepsilon)$ *respecto a los de* $\mathbf{B}$, *es decir,* $\lambda_1(\varepsilon) \geq \cdots \geq \lambda_p(\varepsilon)$ *y los signos de las componentes no nulas de* $\underline{\alpha}_k(\varepsilon)$ *coinciden con los de* $\underline{\alpha}_k$.

Teorema 2.2.2 *En las condiciones del teorema 2.2.1, si* λ_k *es un autovalor simple de* $\mathbf{B}$, *entonces,*

$$\nu_k = a_{kk}^{(1)},$$

$$\underline{\beta}_k = -\sum_{j \neq k} \frac{a_{jk}^{(1)}}{\lambda_j - \lambda_k} \underline{\alpha}_j,$$

$$\pi_k = a_{kk}^{(2)} - 2\sum_{j \neq k} \frac{\left(a_{jk}^{(1)}\right)^2}{\lambda_j - \lambda_k},$$

$$\underline{\gamma}_k = -\sum_{j \neq k} \frac{1}{\lambda_j - \lambda_k} \left\{ a_{jk}^{(2)} - 2\sum_{l \neq k} \frac{1}{\lambda_l - \lambda_k} a_{kl}^{(1)} \left(a_{jl}^{(1)} - a_{kk}^{(1)} \delta_{jl} \right) \right\} \underline{\alpha}_j -$$

$$- \left\| \underline{\beta}_k \right\|^2 \underline{\alpha}_k,$$

donde δ_{jl} *es la delta de Kronecker y* $a_{jl}^{(s)} = \underline{\alpha}_j' \mathbf{B}_s \underline{\alpha}_l, \quad j,l = 1,...,p;\ s = 1,2.$

Nota 2.2.2 *Sibson [60] da una expresión alternativa para los coeficientes* $\underline{\beta}_k$ *y* π_k *de los desarrollos anteriores:*

$$\begin{aligned} \underline{\beta}_k &= -\left(\mathbf{B}-\lambda_k \mathbf{I}_p\right)^+ \mathbf{B}_1 \underline{\alpha}_k \\ y \qquad \pi_k &= \underline{\alpha}_k' \left[\mathbf{B}_2 - 2\mathbf{B}_1 \left(\mathbf{B}-\lambda_k \mathbf{I}_p\right)^+ \mathbf{B}_1\right] \underline{\alpha}_k, \end{aligned}$$

donde $\mathbf{I}_p$ *es la matriz identidad de dimensión* p *y* $\mathbf{M}^+$ *representa la inversa de Moore-Penrose de una matriz* $\mathbf{M}$.

Los teoremas 2.2.1 y 2.2.2 proporcionan la base necesaria para calcular las funciones de influencia de los autovalores simples y los autovectores unitarios asociados. Para ello basta expresar la matriz de covarianzas de la perturbación de la función de distribución dada en (1.1) mediante un desarrollo en serie convergente.

Teorema 2.2.3 *Sean* $F \in \mathcal{F}_p^2$, $\underline{x} \in \mathbb{R}^p$ *y* $\varepsilon \in \mathbb{R}$. *Si* $F_{\varepsilon}^{\underline{x}} \in \mathcal{F}_p$, *entonces,*

1. $\underline{\mu}\left(F_{\varepsilon}^{\underline{x}}\right) = \underline{\mu}\left(F\right) + \varepsilon\left(\underline{x} - \underline{\mu}\left(F\right)\right).$

2. $\boldsymbol{\Sigma}\left(F_{\varepsilon}^{\underline{x}}\right) = \boldsymbol{\Sigma}\left(F\right) + \varepsilon\left\{\left(\underline{x} - \underline{\mu}\left(F\right)\right)\left(\underline{x} - \underline{\mu}\left(F\right)\right)' - \boldsymbol{\Sigma}\left(F\right)\right\} - \\ -\varepsilon^2\left(\underline{x} - \underline{\mu}\left(F\right)\right)\left(\underline{x} - \underline{\mu}\left(F\right)\right)'.$

A partir del apartado **2** del teorema 2.2.3, se deduce directamente la expresión de la función de influencia de $\boldsymbol{\Sigma}$.

Corolario 2.2.4 *Dada* $F \in \mathcal{F}_p^2$ *y* $\underline{x} \in \mathbb{R}^p$, *se verifica que*

$$I\left(\underline{x}; \boldsymbol{\Sigma}, F\right) = \left(\underline{x} - \underline{\mu}\left(F\right)\right)\left(\underline{x} - \underline{\mu}\left(F\right)\right)' - \boldsymbol{\Sigma}\left(F\right).$$

Utilizando el resultado dado en el apartado **2** del teorema 2.2.3 y el teorema 2.2.2, se puede determinar los primeros coeficientes de los desarrollos en serie de los autovalores y autovectores de la matriz de covarianzas poblacionales, obtenidos tras la perturbación (1.1), en función de parámetros de la distribución original. Así se obtiene el siguiente teorema.

Teorema 2.2.5 *Sea* $F \in \mathcal{F}_p^2$ *y sean* $\underline{\mu}$ *y* $\boldsymbol{\Sigma}$ *respectivamente el vector de medias y la matriz de covarianzas asociados a dicha distribución. Sean* $\lambda_1 \geq \cdots \geq \lambda_p$ *los autovalores de* $\boldsymbol{\Sigma}$ *y* $\underline{\alpha}_1, ..., \underline{\alpha}_p$ *autovectores ortonormales asociados respectivamente. Dados* $\underline{x} \in \mathbb{R}^p$ *y* $\varepsilon \in \mathbb{R}$, *si* $F_{\varepsilon}^{\underline{x}} \in \mathcal{F}_p$, *existe un conjunto de*

autovectores de $\Sigma\left(F_{\varepsilon}^{\underline{x}}\right)$, $\underline{\alpha}_k(\varepsilon)$, $k = 1, ..., p$, *ortonormales, tales que, ellos y sus autovalores asociados,* $\lambda_k(\varepsilon)$, $k = 1, ..., p$, *respectivamente, se pueden expresar según desarrollos en serie de potencias convergentes de la forma*

$$\lambda_j(\varepsilon) \equiv \lambda_j + \varepsilon\nu_j + \frac{1}{2}\varepsilon^2\pi_j + O(\varepsilon^3),$$
$$\underline{\alpha}_j(\varepsilon) \equiv \underline{\alpha}_j + \varepsilon\underline{\beta}_j + \frac{1}{2}\varepsilon^2\underline{\gamma}_j + O(\varepsilon^3).$$

Además, si λ_k *es simple, entonces,*

$$\nu_k = y_k^2 - \lambda_k, \tag{2.1}$$
$$\underline{\beta}_k = -y_k \sum_{j\neq k} \frac{y_j}{\lambda_j - \lambda_k}\underline{\alpha}_j, \tag{2.2}$$
$$\pi_k = -2y_k^2(1 + b_{(k)}(2,1)), \tag{2.3}$$
$$\underline{\gamma}_k = -y_k^2 b_{(k)}(2,2)\underline{\alpha}_k - 2b_{(k)}(2,1)\underline{\beta}_k - 2y_k^3 \sum_{j\neq k} \frac{y_j}{(\lambda_j - \lambda_k)^2}\underline{\alpha}_j, \tag{2.4}$$

donde

$$y_k = \underline{\alpha}_k'\left(\underline{x} - \underline{\mu}\right),$$
$$y \quad b_{(k)}(r,s) = \sum_{j\neq k} \frac{y_j^r}{(\lambda_j - \lambda_k)^s}.$$

Nota 2.2.3 *En el teorema 2.2.5, para* ε *suficientemente pequeño, queda asegurado que el orden de los autovalores en la distribución perturbada es el mismo que los correspondientes a la distribución original y que los signos de las componentes de los autovectores, coinciden. Esta cuestión será de importancia en el Análisis de Influencia, ya que surge el problema de la detección del autovalor y del autovector unitario de la distribución original asociados a cada autovalor y autovector de la distribución perturbada.*

Como consecuencia, se obtienen las funciones de influencia de los autovalores simples y autovectores unitarios asociados de la matriz de covarianzas de un vector aleatorio.

Corolario 2.2.6 *Sea* $F \in \mathcal{F}_p^2$ *y sean* $\underline{\mu}$ *y* Σ *respectivamente el vector de medias y la matriz de covarianzas asociados a dicha distribución. Sean* $\lambda_1 \geq \cdots \geq \lambda_p$ *los autovalores de* Σ *y* $\underline{\alpha}_1, ..., \underline{\alpha}_p$ *autovectores ortonormales asociados respectivamente. Si* λ_k *es un autovalor simple, entonces,*

$$I\left(\underline{x}, \lambda_k, F\right) = y_k^2 - \lambda_k \tag{2.5}$$
$$y \quad I\left(\underline{x}, \underline{\alpha}_k, F\right) = -y_k \sum_{j\neq k} \frac{y_j}{\lambda_j - \lambda_k}\underline{\alpha}_j. \tag{2.6}$$

Critchley [15] obtuvo, por un camino paralelo, las expresiones de los coeficientes dados en el teorema 2.2.5 y de las funciones de influencia dadas en el corolario 2.2.6, específicamente para el Análisis de Influencia en Componentes Principales, aunque se pueden considerar como casos particulares de los resultados dados por Tanaka [61].

Nota 2.2.4 *A partir del corolario 2.2.6 se pueden hacer los siguientes comentarios:*

- *La función influencia del k-ésimo autovalor en un punto del espacio* $\mathbb{R}^p$, *sólo depende de la desviación de la k-ésima componente principal en dicho punto respecto al citado autovalor. Es de destacar, que al ser nulo el valor esperado de una componente principal, la varianza coincide con el momento de orden dos, y por tanto, la función de influencia compara el valor del cuadrado de la k-ésima componente principal de un punto, con su valor esperado.*

- $I(\underline{x}, \lambda_k, F)$ *y* $I(\underline{x}, \underline{\alpha}_k, F)$ *son no acotados, por lo que existen zonas del espacio altamente influyentes.*

- *La función de influencia de un autovector es un vector de la misma dimensión, por lo que es necesario utilizar normas vectoriales para evaluar la influencia que ejerce, sobre un autovector, una perturbación de la distribución original en cierta dirección. No obstante, se puede interpretar el caso de la función de influencia nula. Esto sucede cuando ocurre alguna de las siguientes situaciones:*

  - *El valor de la k-ésima componente principal de* $\underline{x}$ *es nula: Este hecho se puede interpretar de distintas formas:*

    * *El valor de la k-ésima componente principal de* $\underline{x}$ *coincide con su valor esperado.*
    * *El punto* $\underline{x}$ *coincide con la media poblacional,* $\underline{\mu}$, *o bien,* $\underline{x} - \underline{\mu}$ *es ortogonal a* $\underline{\alpha}_k$. *Esto último ocurre cuando la proyección de* $\underline{x}$ *sobre* $\underline{\alpha}_k$ *coincide con la de* $\underline{\mu}$.

  - *El valor del resto de las componentes principales es nulo. Además de interpretaciones análogas al caso anterior, se deduce que si el punto* $\underline{x}$ *no es la media poblacional, al ser el vector* $\underline{x} - \underline{\mu}$ *ortogonal a cada uno de los autovectores* $\underline{\alpha}_j$, *con* $j \neq k$, *entonces es paralelo a* $\underline{\alpha}_k$. *Así,* $y_k^2 = \left\| \underline{x} - \underline{\mu} \right\|^2$.

- *La función de influencia de un autovector, $\underline{\alpha}_k$, es ortogonal a dicho autovector, ya que, como se observa en (2.6), no tiene componente en dicha dirección.*

- *Fijado el autovector $\underline{\alpha}_k$, la elección del signo del resto de los autovectores no influye en el valor de $I(\underline{x}, \lambda_k, F)$ y $I(\underline{x}, \underline{\alpha}_k, F)$, ya que si se cambia $\underline{\alpha}_j$ por $-\underline{\alpha}_j$, y_j cambiará por $-y_j$ y la componente correspondiente quedará invariante. Por otro lado, si se sustituye el vector $\underline{\alpha}_k$ por $-\underline{\alpha}_k$, variará el sentido del vector que define la función de influencia, pero no su módulo.*

Tanaka y Tarumi [65] extendieron los resultados del corolario 2.2.5, proporcionando las expresiones de los coeficientes hasta segundo orden de los desarrollos de los autovalores y autovectores para el caso de autovalores múltiples, bajo la restricción de simplicidad de autovalores para la matriz obtenida tras la perturbación.

2.2.2 Estudio basado en la matriz de correlación

En algunas ocasiones es preferible el uso de la matriz de correlación en el cálculo de componentes principales, con objeto de evitar el problema de la dependencia de las unidades de medida. Por ello, el Análisis de Influencia también se ha tratado de desarrollar cuando la matriz utilizada es la de correlación, encontrando el problema de que, aunque la técnica utilizada en el caso de matriz de covarianzas sería válida (bastaría tomar la matriz de correlación en el teorema 2.2.2), la complejidad de las expresiones aumenta considerablemente. Por ello, los estudios hechos desde esta perspectiva se limitan a indicar el camino a seguir en el cálculo de las funciones de influencia. A continuación se recogen algunos resultados que se encuentran en la bibliografía.

Critchley [15] plantea el desarrollo análogo de las funciones de influencia en los parámetros de las componentes principales calculadas mediante la matriz de correlación $\mathbf{P}$, utilizando la relación existente con la matriz $\mathbf{\Sigma}$. Calder [9], partiendo del planteamiento de Critchley [15], proporciona la función de influencia de la matriz $\mathbf{P}$, con el objetivo de obtener la función de influencia sobre autovalores y autovectores de dicha matriz.

Teorema 2.2.7 *Si $\mathbf{P}$ es la matriz de correlación de una variable aleatoria p-dimensional con función de distribución F, media $\underline{\mu}$ y matriz de covarianzas $\mathbf{\Sigma} = (\sigma_{lm})$, entonces,*

$$I(\underline{x}; \mathbf{P}, F) = -\frac{1}{2}\widetilde{\mathbf{X}}^2_{diag}\mathbf{P} - \frac{1}{2}\mathbf{P}\widetilde{\mathbf{X}}^2_{diag} + \widetilde{\underline{x}}\widetilde{\underline{x}}',$$

donde

$$\begin{aligned} \widetilde{\underline{x}} &= \boldsymbol{\Sigma}_{diag}^{-\frac{1}{2}}\left(\underline{x}-\underline{\mu}\right) = \left(\widetilde{x}_1,...,\widetilde{x}_p\right)', \\ \boldsymbol{\Sigma}_{diag} &= diag\left\{\sigma_{11},\sigma_{22},...,\sigma_{pp}\right\} \\ y \qquad \widetilde{\mathbf{X}}_{diag} &= diag\left\{\widetilde{x}_r : r=1,...p\right\}. \end{aligned}$$

Nota 2.2.5 *En el caso en el que la variable aleatoria esté tipificada, la matriz de covarianzas y la matriz de correlación coinciden. En cambio la expresión de las funciones de influencia difieren en ambas. Esto se debe al hecho de que los funcionales que definen a las matrices anteriores son distintos y por tanto, la perturbación de una matriz de covarianzas sigue siendo una matriz de covarianzas pero no tiene porqué corresponder a una variable tipificada. Así la matriz de covarianzas dejaría de coincidir con la matriz de correlación.*

A partir de los teoremas 2.2.2 y 2.2.7, se puede obtener la función de influencia de un autovalor y de un autovector de la matriz de correlación, como se muestra en el siguiente corolario (Calder [9]).

Corolario 2.2.8 *En las condiciones del teorema 2.2.7, si* $\lambda_1 \geq \cdots \geq \lambda_p$ *son los autovalores de* $\mathbf{P}$ *y* $\underline{\alpha}_1,...,\underline{\alpha}_p$, *autovectores ortonormales asociados, respectivamente, entonces, las funciones de influencia de un autovalor simple de* $\mathbf{P}$, λ_k, *y de su autovector asociado,* $\underline{\alpha}_k$, *vienen dadas por*

$$\begin{aligned} I\left(\underline{x};\lambda_k,F\right) &= \underline{\alpha}_k'\left(-\frac{1}{2}\widetilde{\mathbf{X}}_d^2\mathbf{P}-\frac{1}{2}\mathbf{P}\widetilde{\mathbf{X}}_d^2+\widetilde{\underline{x}}\widetilde{\underline{x}}'\right)\underline{\alpha}_k \\ y \qquad I\left(\underline{x};\underline{\alpha}_k,F\right) &= -\sum_{j\neq k}\frac{\underline{\alpha}_j' I(\underline{x};\mathbf{P},F)\underline{\alpha}_k}{\lambda_j-\lambda_k}\underline{\alpha}_j. \end{aligned}$$

Las expresiones dadas en el corolario 2.2.8, se pueden simplificar tal como se recogen en el siguiente corolario.

Corolario 2.2.9 *En las condiciones del corolario 2.2.8, se verifica que*

$$I(\underline{x};\lambda_k,F) = \widetilde{y}_k^2-\lambda_k\sum_{j=1}^{p}\alpha_{kj}^2\widetilde{x}_j^2, \tag{2.7}$$

$$I(\underline{x};\underline{\alpha}_k,F) = -\sum_{j\neq k}\frac{\widetilde{y}_j\widetilde{y}_k-\lambda_k\sum_{l=1}^{p}\alpha_{jl}\alpha_{kl}\widetilde{x}_l^2}{\lambda_j-\lambda_k}\underline{\alpha}_j, \tag{2.8}$$

donde $\widetilde{y}_k = \underline{\alpha}_k'\widetilde{\underline{x}}$ *y* α_{kj} *es la* j*-ésima componente de* $\underline{\alpha}_k$.

Demostración

Desarrollando las expresiones dadas en el corolario 2.2.8, para un autovalor simple, λ_k,

$$\begin{aligned} I(\underline{x};\lambda_k,F) &= -\frac{1}{2}\underline{\alpha}_k'\widetilde{\mathbf{X}}_d^2\mathbf{R}\underline{\alpha}_k - \frac{1}{2}\underline{\alpha}_k'\mathbf{R}\widetilde{\mathbf{X}}_d^2\underline{\alpha}_k + \underline{\alpha}_k'\widetilde{\underline{x}}\widetilde{\underline{x}}'\underline{\alpha}_k = \\ &= -\frac{1}{2}\lambda_k\underline{\alpha}_k'\widetilde{\mathbf{X}}_d^2\underline{\alpha}_k - \frac{1}{2}\lambda_k\underline{\alpha}_k'\widetilde{\mathbf{X}}_d^2\underline{\alpha}_k + \widetilde{y}_k^2 = \\ &= \widetilde{y}_k^2 - \lambda_k\sum_{j=1}^{p}\alpha_{kj}^2\widetilde{x}_j^2. \end{aligned}$$

Y para un autovector unitario asociado a λ_k,

$$\begin{aligned} I(\underline{x};\underline{\alpha}_k,F) &= -\sum_{j\neq k}\frac{-\frac{1}{2}\underline{\alpha}_j'\widetilde{\mathbf{X}}_d^2\mathbf{R}\underline{\alpha}_k - \frac{1}{2}\underline{\alpha}_j'\mathbf{R}\widetilde{\mathbf{X}}_d^2\underline{\alpha}_k + \underline{\alpha}_j'\widetilde{\underline{x}}\widetilde{\underline{x}}'\underline{\alpha}_k}{\lambda_j-\lambda_k}\underline{\alpha}_j = \\ &= -\sum_{j\neq k}\frac{-\frac{1}{2}\lambda_k\underline{\alpha}_j'\widetilde{\mathbf{X}}_d^2\underline{\alpha}_k - \frac{1}{2}\lambda_k\underline{\alpha}_j'\widetilde{\mathbf{X}}_d^2\underline{\alpha}_k + \widetilde{y}_j\widetilde{y}_k}{\lambda_j-\lambda_k}\underline{\alpha}_j = \\ &= -\sum_{j\neq k}\frac{\widetilde{y}_j\widetilde{y}_k - \lambda_k\sum_{l=1}^{p}\alpha_{jl}\alpha_{kl}\widetilde{x}_l^2}{\lambda_j-\lambda_k}\underline{\alpha}_j. \end{aligned}$$

■

Las expresiones (2.7) y (2.8) posibilitan la interpretación de las funciones de influencia de los autovalores y autovectores de la matriz de correlación y su comparación con las obtenidas al utilizar la matriz de covarianzas.

Nota 2.2.6 *A partir de las expresiones (2.7) y (2.8) se pueden realizar los siguientes comentarios:*

- *Para variables tipificadas, las expresiones de las funciones de influencia de autovalores simples de la matriz de correlación y autovectores unitarios asociados, no coinciden con las obtenidas en el estudio basado en la matriz de covarianzas, como se puede observar comparando las expresiones (2.5) y (2.6) con (2.7) y (2.8), respectivamente. Este hecho se deriva de los mismos motivos justificados en la nota 2.2.5.*

  Si se considera una variable tipificada, las funciones de influencia de un autovalor difieren en las ponderaciones $\widetilde{x}_j^2$ de cada uno de los cuadrados

de las componentes del autovector $\underline{\alpha}_j$. Si estas ponderaciones fueran la unidad, las funciones de influencia serían equivalentes. Esto pone de relieve la posibilidad de que existan puntos con valores altos para la función de influencia de un autovalor de la matriz de covarianzas, pero no para la de correlación y viceversa.

En el caso de la función de influencia de un autovector, aparece un elemento corrector en cada componente, $-\lambda_k \sum_{l=1}^{p} \alpha_{jl}\alpha_{kl}\widetilde{x}_l^2$, el cual se anula, en particular, para la media muestral.

2.3 Versiones muestrales basadas en la matriz de covarianzas

En realidad, el objetivo del Análisis de Influencia es práctico, tal como se ha comentado en el capítulo anterior. Por esta razón, será necesario recurrir a alguna versión muestral de la función de influencia correspondiente: la función de influencia empírica, función de influencia empírica con omisión o función de influencia muestral.

Desde el punto de vista muestral, la matriz a utilizar en el cálculo de las componentes principales, es un estimador de la matriz utilizada desde el punto de vista poblacional: la matriz de correlación o la matriz de covarianzas. En el primer caso se considera como estimador, la matriz de correlaciones muestrales y en el segundo caso, existen dos estimadores comúnmente usados: $\widehat{\mathbf{\Sigma}}$ y $\mathbf{S}$. Las versiones muestrales habituales se obtienen a partir de la función de distribución empírica, la cual tiene como matriz de covarianzas asociada, la matriz $\widehat{\mathbf{\Sigma}}$. Por ello, los estudios realizados desde el punto de vista muestral de las funciones de influencia en el Análisis de Componentes Principales utilizan dicho estimador. Pero en la práctica, el estimador más utilizado, es la matriz $\mathbf{S}$. En este apartado, además de recopilar las ideas y resultados más importantes de los trabajos publicados sobre las funciones de influencia en el Análisis de Componentes Principales, se aportan las funciones de influencia muestrales de la matriz $\mathbf{S}$, sus autovalores y autovectores.

Las versiones muestrales de las funciones de influencia se desarrollan desde el punto de vista de las realizaciones muestrales, más que desde la propia muestra. No obstante, por simplicidad se seguirá la misma notación para los estimadores de $\mathbf{\Sigma}$, $\widehat{\mathbf{\Sigma}}$ y $\mathbf{S}$, sus autovalores y autovectores, que para sus estimaciones con la realización muestral.

Dada $\underline{X}_1, ..., \underline{X}_n$ una muestra aleatoria del vector aleatorio $\underline{X}$, de media $\underline{\mu}$ y matriz de covarianzas $\mathbf{\Sigma}$, y $\underline{x}_1, ..., \underline{x}_n$ una realización muestral, para la

función de distribución empírica, $\widehat{F}_n$, se verifica que

$$\underline{\mu}\left(\widehat{F}_n\right) = \overline{\underline{x}},$$
$$\Sigma\left(\widehat{F}_n\right) = \widehat{\Sigma},$$

donde $\overline{\underline{x}}$ es la media de la realización muestral, lo que permite obtener de forma directa la expresión de las versiones muestrales de la función de influencia.

2.3.1 Estudio basado en la matriz $\widehat{\Sigma}$

Sean $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq ... \geq \widehat{\lambda}_p$, los p autovalores de $\widehat{\Sigma}$ y $\underline{\widehat{\alpha}}_1, \underline{\widehat{\alpha}}_2, ..., \underline{\widehat{\alpha}}_p$ autovectores ortonormales asociados respectivamente. Para cada $i = 1, ..., n$, se denota por $\underline{\widehat{y}}_i = (\widehat{y}_{i1}, ..., \widehat{y}_{ip})$ al vector de componentes principales de $\underline{x}_i$, con

$$\widehat{y}_{ij} = \underline{\widehat{\alpha}}_j' \left(\underline{x}_i - \overline{\underline{x}}\right), \qquad j = 1, ..., p.$$

Se consideran las siguientes versiones muestrales de los parámetros poblacionales recogidos en (2.1), (2.2), (2.3), y (2.4), para un autovalor simple, $\widehat{\lambda}_k$.

$$\begin{aligned}
\widehat{\nu}_{ik} &= \widehat{y}_{ik}^2 - \widehat{\lambda}_k, \\
\underline{\widehat{\beta}}_{ik} &= -\widehat{y}_{ik} \sum_{j \neq k} \frac{\widehat{y}_{ij}}{\widehat{\lambda}_j - \widehat{\lambda}_k} \underline{\widehat{\alpha}}_j, \\
\widehat{\pi}_{ik} &= -2\widehat{y}_{ik}^2 \left(1 + \widehat{b}_{i(k)}(2,1)\right), \\
\underline{\widehat{\gamma}}_{ik} &= -\widehat{y}_{ik}^2 \, \widehat{b}_{i(k)}(2,2) \, \underline{\widehat{\alpha}}_k - 2 \, \widehat{b}_{i(k)}(2,1) \, \underline{\widehat{\beta}}_{ik} - 2\widehat{y}_{ik}^3 \sum_{j \neq k} \frac{\widehat{y}_{ij}}{(\widehat{\lambda}_j - \widehat{\lambda}_k)^2} \underline{\widehat{\alpha}}_j,
\end{aligned}$$

donde $\widehat{b}_{i(k)}(r,s) = \sum_{j \neq k} \dfrac{\widehat{y}_{ij}^r}{(\widehat{\lambda}_j - \widehat{\lambda}_k)^s}$.

A continuación, se recogen las versiones muestrales de las funciones de influencia y los desarrollos utilizados en la obtención de cada una de ellas, que fueron dadas por Critchley [15].

En primer lugar, tomando $F = \widehat{F}_n$ en (2.5) y (2.6) se obtienen las expresiones de la función de influencia empírica.

Teorema 2.3.1 *Sea $\underline{X}$ un vector aleatorio con matriz de covarianzas Σ. Sea $\underline{X}_1, ..., \underline{X}_n$ una muestra aleatoria del vector $\underline{X}$ y $\underline{x}_1, ..., \underline{x}_n$ una realización*

muestral de la misma. La función de influencia empírica de $\widehat{\Sigma}$ *en* $\underline{x}_i$, $i = 1, ..., n$, *viene dada por*

$$FIE\left(\underline{x}_i; \widehat{\Sigma}\right) = (\underline{x}_i - \overline{\underline{x}})(\underline{x}_i - \overline{\underline{x}})' - \widehat{\Sigma}.$$

Además, si $\widehat{\lambda}_k$ *es un autovalor simple de* $\widehat{\Sigma}$, *y* $\widehat{\underline{\alpha}}_k$ *es un autovector unitario asociado, las funciones de influencia empíricas en* $\underline{x}_i$, $i = 1, ..., n$, *son*

$$FIE(\underline{x}_i; \widehat{\lambda}_k) = \widehat{\nu}_{ik} = \widehat{y}_{ik}^2 - \widehat{\lambda}_k, \tag{2.9}$$

$$y \qquad FIE(\underline{x}_i; \widehat{\underline{\alpha}}_k) = \widehat{\underline{\beta}}_{ik} = -\widehat{y}_{ik} \sum_{j \neq k} \frac{\widehat{y}_{ij}}{\widehat{\lambda}_j - \widehat{\lambda}_k} \widehat{\underline{\alpha}}_j. \tag{2.10}$$

Los comentarios realizados en la nota 2.2.4, se puede particularizar para el caso $F = \widehat{F}_n$, lo que permite interpretar las funciones de influencia empíricas dadas en (2.9) y (2.10).

Otras versiones muestrales de la función de influencia son la función de influencia empírica con omisión y la función de influencia muestral. Ambas se basan en la omisión de observaciones.

La omisión de una observación es equivalente a introducir una perturbación dada por $\varepsilon = -\dfrac{1}{n-1}$ sobre la función de distribución $F = \widehat{F}_n$, como se vio en (1.3). Esto presenta dos problemas en el campo de las componentes principales. En primer lugar, desde el punto de vista computacional, para analizar la influencia que ejerce cada observación sobre un autovalor o un autovector, es necesaria la resolución de un problema de autovalores. En segundo lugar, al resolver un problema de autovalores, se obtienen p autovalores en los que la única distinción posible entre ellos es el orden que guardan. Al realizar una perturbación, omisión en este caso, sobre la matriz de la que provienen, se obtiene un conjunto de autovalores que vienen asociados a los de la matriz original. Pero no se tiene asegurado que se mantenga el orden entre ellos ni se tiene determinado de forma única un autovector unitario.

A este problema práctico, se le puede dar solución utilizando aproximaciones de las funciones de influencia empíricas y de las funciones de influencia muestrales, basadas en desarrollos en serie en función de estimaciones calculados con la muestra completa. Para ello, llevando a cabo una perturbación de tipo omisión y aplicando el teorema 2.2.3, se tiene que

$$\widehat{\Sigma}^{(i)} = \Sigma\left(\widehat{F}_{n-1}^{(i)}\right) = \widehat{\Sigma} - \frac{1}{n-1}\left[(\underline{x}_i - \overline{\underline{x}})(\underline{x}_i - \overline{\underline{x}})' - \widehat{\Sigma}\right] - \\ - \frac{1}{(n-1)^2}(\underline{x}_i - \overline{\underline{x}})(\underline{x}_i - \overline{\underline{x}})'. \tag{2.11}$$

Y aplicando el teorema 2.2.5 se obtienen los desarrollos en serie convergente de los autovalores ya autovectores de $\widehat{\boldsymbol{\Sigma}}^{(i)}$.

Teorema 2.3.2 *Sea $\underline{X}$ un vector aleatorio p-dimensional, con matriz de covarianzas $\boldsymbol{\Sigma}$. Sea $\underline{X}_1, ..., \underline{X}_n$ una muestra aleatoria del vector $\underline{X}$ y $\underline{x}_1, ..., \underline{x}_n$ una realización muestral de la misma. Sean $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_p$ los autovalores de $\widehat{\boldsymbol{\Sigma}}$ y $\underline{\widehat{\alpha}}_1, ..., \underline{\widehat{\alpha}}_p$ autovectores ortonormales asociados, respectivamente. Entonces, para cada $i = 1, ..., n$, existe un conjunto de autovectores ortonormales de $\widehat{\boldsymbol{\Sigma}}^{(i)}$, $\underline{\widehat{\alpha}}_j^{(i)}$, $j = 1, ..., p$, tales que, ellos y sus autovalores asociados, $\widehat{\lambda}_j^{(i)}$, $j = 1, ..., p$, se pueden expresar en desarrollos en serie de potencias convergentes. En particular, si $\widehat{\lambda}_k$ es un autovalor simple de $\widehat{\boldsymbol{\Sigma}}$,*

$$\widehat{\lambda}_k^{(i)} = \widehat{\lambda}_k - \frac{1}{n-1}\widehat{\nu}_{ik} + \frac{1}{2}\frac{1}{(n-1)^2}\widehat{\pi}_{ik} + O\left(n^{-3}\right) \tag{2.12}$$

y

$$\underline{\widehat{\alpha}}_k^{(i)} = \underline{\widehat{\alpha}}_k - \frac{1}{n-1}\underline{\widehat{\beta}}_{ik} + \frac{1}{2}\frac{1}{(n-1)^2}\underline{\widehat{\gamma}}_{ik} + O\left(n^{-3}\right). \tag{2.13}$$

Nota 2.3.1 *A partir de (2.12) y (2.13), se observa que, para tamaño muestral suficientemente grande, $\widehat{\lambda}_1^{(i)} \geq \cdots \geq \widehat{\lambda}_p^{(i)}$, y los signos de las componentes no nulas de $\underline{\widehat{\alpha}}_k$ son los mismos que los de las componentes de $\underline{\widehat{\alpha}}_k^{(i)}$.*

Como consecuencia y utilizando la relación existente entre las estimaciones que intervienen, calculadas con y sin la omisión de observaciones, las funciones de influencia empíricas con omisión, se pueden expresar según se indica en el siguiente corolario (Critchley [15]).

Corolario 2.3.3 *Sea $\underline{X}$ un vector aleatorio con matriz de covarianzas $\boldsymbol{\Sigma}$. Sea $\underline{X}_1, ..., \underline{X}_n$ una muestra aleatoria del vector $\underline{X}$ y $\underline{x}_1, ..., \underline{x}_n$ una realización muestral. Si $\widehat{\lambda}_k$ es un autovalor simple de $\widehat{\boldsymbol{\Sigma}}$, entonces, para cada $i = 1, ..., n$,*

$$FIE_{(i)}\left(\underline{x}_i; \widehat{\lambda}_k\right) = \widehat{\nu}_{ik} - \frac{1}{n-1}\left(\widehat{\pi}_{ik} - \widehat{\nu}_{ik}\right) + O(n^{-2}) \tag{2.14}$$

y

$$FIE_{(i)}\left(\underline{x}_i; \underline{\widehat{\alpha}}_k\right) = \underline{\widehat{\beta}}_{ik} - \frac{1}{n-1}\left(\underline{\widehat{\gamma}}_{ik} - \underline{\widehat{\beta}}_{ik}\right) + O(n^{-2}). \tag{2.15}$$

Los desarrollos en serie (2.12) y (2.13) son útiles para calcular una aproximación de la función de influencia muestral de los autovalores y autovectores de $\widehat{\boldsymbol{\Sigma}}$. Para ello se enuncia el siguiente resultado.

Corolario 2.3.4 *Sea $\underline{X}$ un vector aleatorio con matriz de covarianzas Σ. Sea $\underline{X}_1, ..., \underline{X}_n$ una muestra aleatoria del vector $\underline{X}$ y $\underline{x}_1, ..., \underline{x}_n$ una realización muestral. La función de influencia muestral de la matriz $\widehat{\Sigma}$, en $\underline{x}_i$, $i = 1, ..., n$, viene dada por*

$$FIM(\underline{x}_i, \widehat{\Sigma}) = (\underline{x}_i - \overline{\underline{x}})(\underline{x}_i - \overline{\underline{x}})' - \widehat{\Sigma} + \frac{1}{n-1}(\underline{x}_i - \overline{\underline{x}})(\underline{x}_i - \overline{\underline{x}}).$$

Si $\widehat{\lambda}_k$ es un autovalor simple de la matriz $\widehat{\Sigma}$, y $\underline{\widehat{\alpha}}_k$ un autovector unitario asociado, las funciones de influencia muestrales de $\widehat{\lambda}_k$ y $\underline{\widehat{\alpha}}_k$, en la observación $\underline{x}_i$, $i = 1, ..., n$, vienen dadas por

$$FIM\left(\underline{x}_i, \widehat{\lambda}_k\right) = \widehat{\nu}_{ik} - \frac{1}{2}\frac{1}{n-1}\widehat{\pi}_{ik} + O\left(n^{-2}\right) \tag{2.16}$$

$$y \qquad FIM\left(\underline{x}_i, \underline{\widehat{\alpha}}_k\right) = \underline{\widehat{\beta}}_{ik} - \frac{1}{2}\frac{1}{n-1}\underline{\widehat{\gamma}}_{ik} + O\left(n^{-2}\right). \tag{2.17}$$

Nota 2.3.2 *En relación a la incertidumbre en la correspondencia de los autovalores calculados con o sin la omisión de observaciones, Brooks [7] propone la comparación de la función de influencia empírica, que no presenta dicho problema, con la función de influencia muestral, supuesto que en la correspondencia se mantiene el orden.*

Si este supuesto fuera cierto, las conclusiones deben ser similares, ya que la función de influencia empírica se puede considerar como una aproximación de la función de influencia muestral. Por lo tanto, si las conclusiones obtenidas en ambos estudios difieren, se debe recalcular la función de influencia muestral tras modificar adecuadamente el orden de los autovalores. De esta forma, se supedita el Análisis de Influencia mediante la función de influencia muestral, al cálculo previo de la función de influencia empírica.

Un razonamiento similar al propuesto por Brooks [7] se podría realizar para solucionar el problema de la elección adecuada del sentido del autovector calculado tras la omisión de una observación.

Ambos problemas se pueden eludir mediante el uso de aproximaciones basadas en las expresiones (2.16) y (2.17), evitando así la comparación de la función de influencia empírica y la función de influencia muestral.

Nota 2.3.3 *Comparando las expresiones (2.9), (2.14), y (2.16) y, por otro lado, (2.10), (2.15) y (2.17), se observa que las tres versiones muestrales coinciden asintóticamente, tanto para el autovalor como para el autovector, ya que omitiendo los términos de orden n^{-1} e inferiores, los tres casos se reducen a la función de influencia empírica.*

Además, en la función de influencia empírica con omisión y en la función de influencia muestral de un autovalor simple de $\widehat{\Sigma}$, intervienen términos

con factores $\left(\widehat{\lambda}_j - \widehat{\lambda}_k\right)^{-1}$ *lo que refleja, para tamaños muestrales no muy grandes, que los valores de la función de influencia muestral pueden ser sensiblemente más elevados que los de la función de influencia empírica, cuando hay autovalores próximos a* $\widehat{\lambda}_k$.

2.3.2 Estudio basado en la matriz S

El Análisis de Influencia en el Análisis de Componentes Principales, se suele llevar a cabo a partir de $\widehat{\boldsymbol{\Sigma}}$ ya que las versiones muestrales de las funciones de influencia se derivan de la función de influencia de $\boldsymbol{\Sigma}$, y de los desarrollos en serie basados en ésta. En cambio, en la práctica, el estimador de $\boldsymbol{\Sigma}$ que suele utilizarse es $\mathbf{S}$. Así, es necesario estudiar la influencia en los estadísticos de interés calculados de esta forma. Las conclusiones que se obtienen mediante el análisis de influencia a través de la matriz $\mathbf{S}$, deberán ser similares a las que se obtienen con la matriz $\widehat{\boldsymbol{\Sigma}}$, ya que la relación entre ambas estimaciones se establece mediante una constante multiplicativa, tanto en la matriz como en los autovalores. En el caso de los autovectores, dicha relación es de igualdad.

La dificultad en plantear las versiones muestrales a partir de un funcional definido en el conjunto de las funciones de distribución parte de lo siguiente. Un funcional de este tipo debe cumplir que:

- para la verdadera función de distribución, se obtenga la matriz de covarianzas ($T(F) = \boldsymbol{\Sigma}$), y
- para la función de distribución empírica asociada a una muestra de dicha variable, se obtenga la estimación deseada ($T\left(\widehat{F}_n\right) = \mathbf{S}$).

En general, dada una función de distribución empírica $\widehat{F}$, ésta no determina de forma unívoca las observaciones que participan en su construcción. Así, si $\widehat{F}(\underline{x}_0) = \frac{k}{m}$ también se verifica que $\widehat{F}(\underline{x}_0) = \frac{a \cdot k}{a \cdot m}$ para todo entero positivo a. Si $\boldsymbol{\Sigma}_c$ fuera un funcional $\boldsymbol{\Sigma}_c : \mathcal{F}_p^2 \longrightarrow \mathcal{S}_p$ tal que a toda función de distribución empírica le asocia la matriz de covarianzas muestrales de la muestra de la que procede, se verificaría que

$$\boldsymbol{\Sigma}_c\left(\widehat{F}\right) = \frac{m}{m-1}\boldsymbol{\Sigma}\left(\widehat{F}\right)$$

cuando $\widehat{F}$ fuera una función de distribución empírica de una muestra de tamaño m; y

$$\boldsymbol{\Sigma}_c\left(\widehat{F}\right) = \frac{a \cdot m}{a \cdot m - 1}\boldsymbol{\Sigma}\left(\widehat{F}\right)$$

cuando $\widehat{F}$ fuera una función de distribución empírica de una muestra de tamaño $a \cdot m$.

$\boldsymbol{\Sigma}\left(\widehat{F}\right)$ coincide en ambos casos, pero no $\boldsymbol{\Sigma}_c\left(\widehat{F}\right)$. Por lo tanto no es posible determinar un funcional que asigne a toda función de distribución empirica la matriz de covarianzas muestrales de la realización muestral de la que procede.

Por ello, no es posible realizar el Análisis de Influencia desde el punto de vista teórico a través de un funcional, al menos de forma global. Así, no tiene sentido el uso de función de influencia empírica, aunque sí de la función de influencia muestral, definida a través de la diferencia de los estadísticos de interés calculados mediante todas las observaciones de la muestra y tras la omisión de una de ellas. Por ejemplo,

$$FIM\left(\underline{x}_i; \mathbf{S}\right) = (n-1)\left[\mathbf{S} - \mathbf{S}^{(i)}\right].$$

El cálculo de las funciones de influencia muestrales, directamente a través de su definición, conlleva la resolución de un número elevado de problemas de autovalores, tantos como observaciones en la muestra. Para disminuir el número de operaciones a realizar, a continuación se hallan aproximaciones en función de estadísticos calculados mediante la muestra completa. Por lo tanto es necesario expresar los estadísticos obtenidos tras la omisión de una de las observaciones, en función de estadísticos calculados con la muestra completa. Con estos desarrollos, de nuevo, se evita el problema de la ordenación de los autovalores y de la elección del autovector unitario tras la omisión de una observación.

Teorema 2.3.5 *Sea $\underline{X}$ un vector aleatorio p-dimensional, con matriz de covarianzas $\boldsymbol{\Sigma}$. Sea $\underline{X}_1, ..., \underline{X}_n$ una muestra aleatoria del vector $\underline{X}$ y $\underline{x}_1, ..., \underline{x}_n$ una realización muestral de la misma. Sean $\widetilde{\lambda}_1 \geq \cdots \geq \widetilde{\lambda}_p$ los autovalores de $\mathbf{S}$ y $\underline{\widetilde{\alpha}}_1, ..., \underline{\widetilde{\alpha}}_p$ autovectores ortonormales asociados, respectivamente. Entonces, para $i = 1, ..., n$,*

$$\begin{aligned}\mathbf{S}^{(i)} = \mathbf{S} &- \frac{1}{n-2}\left[\left(\underline{x}_i - \overline{\underline{x}}\right)\left(\underline{x}_i - \overline{\underline{x}}\right)' - \mathbf{S}\right] - \\ &- \frac{1}{(n-2)(n-1)}\left(\underline{x}_i - \overline{\underline{x}}\right)\left(\underline{x}_i - \overline{\underline{x}}\right)'. \end{aligned} \tag{2.18}$$

Además, para cada $i = 1, ..., n$, existe un conjunto de autovectores ortonormales de $\mathbf{S}^{(i)}$, $\underline{\widetilde{\alpha}}_j^{(i)}$, $j = 1, ..., p$, tales que, ellos y sus autovalores asociados, $\widetilde{\lambda}_j^{(i)}$, $j = 1, ..., p$, se pueden expresar según desarrollos en serie de potencias

convergentes. En particular, si $\widetilde{\lambda}_k$ *es un autovalor simple de* $\mathbf{S}$,

$$\begin{aligned}
\widetilde{\lambda}_k^{(i)} &= \widetilde{\lambda}_k - \frac{1}{n-2}\widetilde{\nu}_{ik} + \frac{1}{2}\frac{1}{(n-2)^2}\widetilde{\pi}_{ik} + O\left(n^{-3}\right), \\
\underline{\widetilde{\alpha}}_k^{(i)} &= \underline{\widetilde{\alpha}}_k - \frac{1}{n-2}\underline{\widetilde{\beta}}_{ik} + \frac{1}{2}\frac{1}{(n-2)^2}\underline{\widetilde{\gamma}}_{ik} + O\left(n^{-3}\right),
\end{aligned} \tag{2.19}$$

donde

$$\begin{aligned}
\widetilde{\nu}_{ik} &= \widetilde{y}_{ik}^2 - \widetilde{\lambda}_k, \\
\underline{\widetilde{\beta}}_{ik} &= -\sum_{j\neq k}\frac{\widetilde{y}_{ij}\widetilde{y}_{ik}}{\widetilde{\lambda}_j - \widetilde{\lambda}_k}\underline{\widetilde{\alpha}}_j, \\
\widetilde{\pi}_{ik} &= -2\widetilde{y}_{ik}^2\left(1 + \widetilde{b}_{i(k)}(2,1)\right), \\
\underline{\widetilde{\gamma}}_{ik} &= -\widetilde{y}_{ik}^2\widetilde{b}_{i(k)}(2,2)\underline{\widetilde{\alpha}}_k - 2\widetilde{b}_{i(k)}(2,1)\underline{\widetilde{\beta}}_{ik} - 2\widetilde{y}_{ik}^3\sum_{j\neq k}\frac{\widetilde{y}_{ij}}{(\widetilde{\lambda}_j - \widetilde{\lambda}_k)^2}\underline{\widetilde{\alpha}}_j
\end{aligned}$$

y $\quad \widetilde{b}_{i(k)}(r,s) = \sum_{j\neq k}\frac{\widetilde{y}_{ij}^r}{(\widetilde{\lambda}_j - \widetilde{\lambda}_k)^s}.$

Demostración

De la expresión (2.11), se deduce que

$$\begin{aligned}
\mathbf{S}^{(i)} &= \frac{n-1}{n-2}\widehat{\mathbf{\Sigma}}^{(i)} = \frac{n-1}{n-2}\widehat{\mathbf{\Sigma}} - \frac{1}{n-2}\left[(\underline{x}_i - \overline{\underline{x}})(\underline{x}_i - \overline{\underline{x}})' - \widehat{\mathbf{\Sigma}}\right] - \\
&\quad - \frac{1}{(n-2)(n-1)}(\underline{x}_i - \overline{\underline{x}})(\underline{x}_i - \overline{\underline{x}})' \\
&= \mathbf{S} + \frac{1}{(n-2)n}\mathbf{S} - \frac{1}{n-2}(\underline{x}_i - \overline{\underline{x}})(\underline{x}_i - \overline{\underline{x}})' + \frac{n-1}{(n-2)n}\mathbf{S} - \\
&\quad - \frac{1}{(n-2)(n-1)}(\underline{x}_i - \overline{\underline{x}})(\underline{x}_i - \overline{\underline{x}})' = \\
&= \mathbf{S} - \frac{1}{n-2}\left[(\underline{x}_i - \overline{\underline{x}})(\underline{x}_i - \overline{\underline{x}})' - \mathbf{S}\right] - \\
&\quad - \frac{1}{(n-2)(n-1)}(\underline{x}_i - \overline{\underline{x}})(\underline{x}_i - \overline{\underline{x}})',
\end{aligned}$$

que también se puede expresar como

$$\begin{aligned}
\mathbf{S}^{(i)} &= \mathbf{S} - \frac{1}{n-2}\left[(\underline{x}_i - \overline{\underline{x}})(\underline{x}_i - \overline{\underline{x}})' - \mathbf{S}\right] - \\
&\quad - \frac{1}{(n-2)^2}(\underline{x}_i - \overline{\underline{x}})(\underline{x}_i - \overline{\underline{x}})' + O\left(n^{-3}\right),
\end{aligned}$$

Y como consecuencia, utilizando el teorema 2.2.2, para $\varepsilon = -\dfrac{1}{n-2}$, $\mathbf{B} = \mathbf{S}$, $\mathbf{B}_1 = (\underline{x}_i - \overline{\underline{x}})(\underline{x}_i - \overline{\underline{x}})' - \mathbf{S}$ y $\mathbf{B}_2 = -2(\underline{x}_i - \overline{\underline{x}})(\underline{x}_i - \overline{\underline{x}})'$, $\widetilde{\lambda}_k^{(i)}$ y $\underline{\widetilde{\alpha}}_k^{(i)}$ se pueden expresar según un desarrollo en serie convergente en función de los elementos de la muestra completa,

$$\widetilde{\lambda}_k^{(i)} = \widetilde{\lambda}_k - \frac{1}{n-2}\widetilde{\nu}_{ik} + \frac{1}{2}\frac{1}{(n-2)^2}\widetilde{\pi}_{ik} + O\left(n^{-3}\right)$$

$$\text{y} \quad \underline{\widetilde{\alpha}}_k^{(i)} = \underline{\widetilde{\alpha}}_k - \frac{1}{n-2}\underline{\widetilde{\beta}}_{ik} + \frac{1}{2}\frac{1}{(n-2)^2}\underline{\widetilde{\gamma}}_{ik} + O\left(n^{-3}\right),$$

donde los coeficientes se obtienen únicamente utilizando las propiedades de autovalores y autovectores de $\mathbf{S}$. Al ser los elementos que intervienen equivalentes a los del teorema 2.3.2 sustituyendo $\widehat{\boldsymbol{\Sigma}}$ por $\mathbf{S}$, las expresiones de los coeficientes que se obtienen son análogas.

■

En el siguiente corolario se exponen las funciones de influencia muestrales de la matriz $\mathbf{S}$, sus autovalores simples y autovectores unitarios asociados. Su resultado se obtiene directamente del teorema 2.3.5.

Corolario 2.3.6 *Sean $\underline{X}$ un vector aleatorio con matriz de covarianzas $\boldsymbol{\Sigma}$, $\underline{X}_1, ..., \underline{X}_n$ una muestra aleatoria del vector $\underline{X}$ y $\underline{x}_1, ..., \underline{x}_n$ una realización muestral de la misma. Entonces, la función influencia muestral de $\mathbf{S}$ en $\underline{x}_i$, $i = 1, ..., n$, viene dada por*

$$FIM(\underline{x}_i; \mathbf{S}) = \left[(\underline{x}_i - \overline{\underline{x}})(\underline{x}_i - \overline{\underline{x}})' - \mathbf{S}\right] + \frac{1}{n-2}\left[2(\underline{x}_i - \overline{\underline{x}})(\underline{x}_i - \overline{\underline{x}})' - \mathbf{S}\right]$$

Además, si $\widetilde{\lambda}_k$ es un autovalor simple de $\mathbf{S}$, y $\underline{\widetilde{\alpha}}_k$ un autovector unitario asociado, entonces, para cada $i = 1, ..., n$,

$$FIM\left(\underline{x}_i; \widetilde{\lambda}_k\right) = \widetilde{\nu}_{ik} + \frac{1}{n-2}\left[\widetilde{\nu}_{ik} - \frac{1}{2}\widetilde{\pi}_{ik}\right] + O\left(n^{-2}\right)$$

$$y \quad FIM\left(\underline{x}_i; \underline{\widetilde{\alpha}}_k\right) = \underline{\widetilde{\beta}}_{ik} + \frac{1}{n-2}\left[\underline{\widetilde{\beta}}_{ik} - \frac{1}{2}\underline{\widetilde{\gamma}}_{ik}\right] + O\left(n^{-2}\right).$$

Nota 2.3.4 *La función de influencia muestral detecta las mismas observaciones altamente influyentes para el análisis de componentes principales llevado a cabo a través de $\mathbf{S}$ que de $\widehat{\boldsymbol{\Sigma}}$, ya que*

$$\begin{aligned}\frac{1}{n-1}FIM\left(\underline{x}_i; \widehat{\lambda}_k\right) &= \frac{n-1}{n}\widetilde{\lambda}_k - \frac{n-2}{n-1}\widetilde{\lambda}_k^{(i)} = \\ &= \frac{n-2}{n-1}\left(\frac{n-1}{n}\widetilde{\lambda}_k - \widetilde{\lambda}_k^{(i)}\right) = \\ &= \frac{n-2}{(n-1)^2}FIM\left(\underline{x}_i; \widetilde{\lambda}_k\right) - \frac{n-2}{(n-1)n}\widetilde{\lambda}_k,\end{aligned}$$

y, análogamente,

$$\frac{1}{n-1}FIM\left(\underline{x}_i;\widehat{\underline{\alpha}}_k\right)=\frac{n-2}{(n-1)^2}FIM\left(\underline{x}_i;\widetilde{\underline{\alpha}}_k\right)-\frac{n-2}{(n-1)n}\widetilde{\underline{\alpha}}_k,$$

donde se tiene una relación de equivalencia salvo constantes aditivas y multiplicativas, que no dependen de la observación considerada.

2.4 Extensión al análisis de influencia conjunta

En el Análisis de Influencia, las técnicas de diagnóstico aplicadas al estudio individual de cada observación se han de complementar con el estudio de influencia conjunta de varias observaciones. Textualmente Kotz y Johnson [38] recogen:

> "Las observaciones pueden ser influyentes individual o conjuntamente con una o más observaciones. Sin embargo, no se da siempre el caso de que observaciones conjuntamente influyentes, sean individualmente influyentes."

Y se podría añadir también el caso contrario, es decir, una observación individualmente influyente puede no serlo al considerarla conjuntamente con otras.

En este apartado se calculan las distintas versiones muestrales de la función de influencia para un conjunto de observaciones, tanto de las estimaciones consideradas anteriormente, de la matriz de covarianzas, como de sus autovalores simples y autovectores asociados. En el caso de la función de influencia empírica, para la estimación $\widehat{\boldsymbol{\Sigma}}$, se llega a la propiedad de aditividad de las mismas, como comenta Critchley [15].

Con el objetivo de seguir un camino paralelo al caso de influencia individual, se obtiene, en primer lugar, la función de influencia conjunta de autovalores y autovectores de las matrices $\widehat{\boldsymbol{\Sigma}}$ y $\mathbf{S}$. Para ello, previamente, mediante el siguiente teorema, se proporciona una generalización de los desarrollos del vector de medias y de la matriz de covarianzas recogidos en el teorema 2.2.3.

Teorema 2.4.1 *Sea* $F\in\mathcal{F}_p^2$, $\varepsilon\in\mathbb{R}$, $\underline{x}_1,...,\underline{x}_r\in\mathbb{R}^p$, $I=\{1,...,r\}$ *y* $\mathbf{x}_I=\{\underline{x}_1,...,\underline{x}_r\}$. *Si* $F_\varepsilon^{\mathbf{x}_I}\in\mathcal{F}_p$, *entonces,*

1. $\underline{\mu}\left(F_\varepsilon^{\mathbf{x}_I}\right)=\underline{\mu}\left(F\right)+\varepsilon\sum_{i\in I}\left(\underline{x}_i-\underline{\mu}\left(F\right)\right)$.

2. $$\boldsymbol{\Sigma}\left(F_{\varepsilon}^{\mathbf{x}_I}\right) = \boldsymbol{\Sigma}(F) + \varepsilon \sum_{i\in I}\left[\left(\underline{x}_i - \underline{\mu}(F)\right)\left(\underline{x}_i - \underline{\mu}(F)\right)' - \boldsymbol{\Sigma}(F)\right] - \\ -\varepsilon^2 \sum_{i\in I}\left(\underline{x}_i - \underline{\mu}(F)\right)\sum_{j\in I}\left(\underline{x}_j - \underline{\mu}(F)\right)'.$$

Demostración

1. Se obtiene directamente aplicando la linealidad de la esperanza.

2. Teniendo en cuenta la expresión de la función de distribución $F_{\varepsilon}^{\mathbf{x}_I}$, (1.4), y aplicando propiedades de la matriz de covarianzas asociada a dicha distribución, se tiene que

$$\boldsymbol{\Sigma}\left(F_{\varepsilon}^{\mathbf{x}_I}\right) = (1-r\varepsilon)\left[\boldsymbol{\Sigma}(F) + \varepsilon^2 \sum_{i\in I}\left(\underline{x}_i - \underline{\mu}(F)\right)\sum_{j\in I}\left(\underline{x}_j - \underline{\mu}(F)\right)'\right] + \\ +\varepsilon \sum_{i\in I}\left[\underline{x}_i - \underline{\mu}(F) - \varepsilon\sum_{j\in I}\left(\underline{x}_j - \underline{\mu}(F)\right)\right]\cdot \\ \cdot\left(\underline{x}_i - \underline{\mu}(F) - \varepsilon\sum_{j\in I}\left(\underline{x}_j - \underline{\mu}(F)\right)\right)' = \\ = \boldsymbol{\Sigma}(F) + \varepsilon\left[-r\boldsymbol{\Sigma}(F) + \sum_{i\in I}\left(\underline{x}_i - \underline{\mu}(F)\right)\left(\underline{x}_i - \underline{\mu}(F)\right)'\right] + \\ +\varepsilon^2\left[\sum_{i\in I}\left(\underline{x}_i - \underline{\mu}(F)\right)\sum_{j\in I}\left(\underline{x}_j - \underline{\mu}(F)\right)' - \right. \\ -\sum_{i\in I}\left(\underline{x}_i - \underline{\mu}(F)\right)\sum_{j\in I}\left(\underline{x}_j - \underline{\mu}(F)\right)' - \\ \left. -\sum_{i\in I}\left(\underline{x}_i - \underline{\mu}(F)\right)\sum_{j\in I}\left(\underline{x}_j - \underline{\mu}(F)\right)'\right] + \\ +\varepsilon^3\left[-r\sum_{i\in I}\left(\underline{x}_i - \underline{\mu}(F)\right)\sum_{j\in I}\left(\underline{x}_j - \underline{\mu}(F)\right)' + \right. \\ \left. +\sum_{i\in I}\sum_{j\in I}\left(\underline{x}_j - \underline{\mu}(F)\right)\sum_{l\in I}\left(\underline{x}_l - \underline{\mu}(F)\right)'\right].$$

Y, agrupando los términos en función de la potencia de ε, se obtiene el resultado recogido en el apartado **2**.

■

Como consecuencia del corolario 2.2.4 y del teorema 2.4.1, se tiene que

$$\boldsymbol{\Sigma}\left(F_{\varepsilon}^{\mathbf{x}_I}\right) = \boldsymbol{\Sigma}(F) + \varepsilon \sum_{i \in I} I\left(\underline{x}_i; \boldsymbol{\Sigma}(F)\right) + O\left(\varepsilon^2\right),$$

de donde se deduce la propiedad de aditividad de la función de influencia del funcional $\boldsymbol{\Sigma}$. Como consecuencia, utilizando el teorema 2.2.2, se obtiene la misma propiedad para sus autovalores y autovectores.

Corolario 2.4.2 *Sea $F \in \mathcal{F}_p^2$, $\underline{x}_1, ..., \underline{x}_r \in \mathbb{R}^p$, $I = \{1, ..., r\}$ y $\mathbf{x}_I = \{\underline{x}_1, ..., \underline{x}_r\}$. Las funciones de influencia de $\boldsymbol{\Sigma}$ en F, de sus autovalores simples, λ_k, y de sus autovectores unitarios asociados, $\underline{\alpha}_k$, son aditivas es decir,*

$$\begin{aligned} I\left(\mathbf{x}_I; \boldsymbol{\Sigma}, F\right) &= \sum_{i \in I} I\left(\underline{x}_i; \boldsymbol{\Sigma}, F\right), \\ I\left(\mathbf{x}_I; \lambda_k, F\right) &= \sum_{i \in I} I\left(\underline{x}_i; \lambda_k, F\right) \\ y \qquad I\left(\mathbf{x}_I; \underline{\alpha}_k, F\right) &= \sum_{i \in I} I\left(\underline{x}_i; \underline{\alpha}_k, F\right). \end{aligned}$$

A continuación, se obtienen las distintas versiones muestrales tanto para las estimaciones de la matriz de covarianzas, como para sus autovalores simples y autovectores unitarios asociados.

Se puede extender la definición de la función de influencia empírica tomando en la función de influencia para $F = \widehat{F}_n$, y, a partir del corolario 2.4.2, se deduce la aditividad de dichas versiones muestrales.

Corolario 2.4.3 *Sea $\underline{X}$ un vector aleatorio con matriz de covarianzas $\boldsymbol{\Sigma}$. Sea $\underline{X}_1, ..., \underline{X}_n$ una muestra aleatoria del vector $\underline{X}$ y $\underline{x}_1, ..., \underline{x}_n$ una realización muestral de la misma. Las funciones de influencia empíricas de $\widehat{\boldsymbol{\Sigma}}$, de sus autovalores simples y sus autovectores unitarios asociados, son aditivas sobre el conjunto de las r observaciones consideradas.*

Nota 2.4.1 *La propiedad de aditividad de las funciones de influencia empíricas permite el análisis de influencia conjunta a partir de las funciones de influencia en cada una de las observaciones. De esta forma, se facilita la detección de conjuntos de observaciones altamente influyentes.*

Por otro lado, la aditividad también muestra la posibilidad de enmascarar la influencia de observaciones que, individualmente, puedan ser altamente influyentes, pero no lo sean conjuntamente. Esto puede ocurrir debido a compensaciones.

El análisis en los autovectores es más complejo, al ser necesario el uso de medidas unidimensionales para la evaluación de la influencia ejercida por una o más observaciones.

Se ha visto, en los apartados anteriores, cómo la posibilidad de expresar los estadísticos de interés calculados al omitir una observación en función del estadístico calculado con todas ellas, es útil computacionalmente en el cálculo de las funciones de influencia muestrales, para reducir la magnitud de los cálculos a realizar, además de evitar el problema planteado sobre la ordenación de los autovalores y la elección del sentido del autovector asociado. Por ello, es necesario encontrar el desarrollo en serie de las estimaciones de la matriz de covarianzas utilizadas previamente, $\widehat{\boldsymbol{\Sigma}}$ y $\mathbf{S}$, sus autovalores y autovectores, tras la omisión de un conjunto de observaciones.

Aplicando el teorema 2.4.1 para $\varepsilon = -\dfrac{1}{n-r}$ y $F = \widehat{F}_n$, se obtienen los desarrollos de $\widehat{\boldsymbol{\Sigma}}^{(I)}$ en función de estadísticos calculados a través de la muestra completa.

Teorema 2.4.4 *Sea $\underline{X}$ un vector aleatorio con matriz de covarianzas $\boldsymbol{\Sigma}$. Sea $\underline{X}_1, ..., \underline{X}_n$ una muestra aleatoria del vector $\underline{X}$ y $\underline{x}_1, ..., \underline{x}_n$ una realización muestral de la misma. Si $I \subset \{1, ..., n\}$ tal que $card\,(I) = r$, entonces,*

$$\overline{\underline{x}}^{(I)} = \overline{\underline{x}} - \frac{1}{n-r} \sum_{i \in I} (\underline{x}_i - \overline{\underline{x}}),$$

$$\widehat{\boldsymbol{\Sigma}}^{(I)} = \widehat{\boldsymbol{\Sigma}} - \frac{1}{n-r} \sum_{i \in I} \widehat{\boldsymbol{\Sigma}}_{1,i} - \frac{1}{(n-r)^2} \sum_{i \in I} (\underline{x}_i - \overline{\underline{x}}) \sum_{l \in I} (\underline{x}_l - \overline{\underline{x}})', \quad (2.20)$$

donde

$$\widehat{\boldsymbol{\Sigma}}_{1,i} = (\underline{x}_i - \overline{\underline{x}})(\underline{x}_i - \overline{\underline{x}})' - \widehat{\boldsymbol{\Sigma}}.$$

Como consecuencia, se obtiene la expresión de $\mathbf{S}^{(I)}$ en función de $\mathbf{S}$, la cual se recoge en el siguiente corolario.

Corolario 2.4.5 *En las condiciones del teorema 2.4.4, se verifica que*

$$\mathbf{S}^{(I)} = \mathbf{S} - \frac{1}{n-r-1} \sum_{i \in I} \left[(\underline{x}_i - \overline{\underline{x}})(\underline{x}_i - \overline{\underline{x}})' - \mathbf{S} \right] - \\ - \frac{1}{(n-r-1)(n-r)} \sum_{i \in I} (\underline{x}_i - \overline{\underline{x}}) \sum_{l \in I} (\underline{x}_l - \overline{\underline{x}})'.$$

Para obtener el desarrollo de los autovalores y autovectores de $\mathbf{S}^{(I)}$ en función de los de $\mathbf{S}$, es conveniente escribir el desarrollo de $\mathbf{S}^{(I)}$ según la siguiente expresión:

$$\mathbf{S}^{(I)} = \mathbf{S} - \frac{1}{n-r-1}\sum_{i\in I}\mathbf{S}_{1,i} - \\ -\frac{1}{(n-r-1)^2}\sum_{i\in I}(\underline{x}_i - \overline{\underline{x}})\sum_{l\in I}(\underline{x}_l - \overline{\underline{x}})' + O\left(n^{-3}\right), \tag{2.21}$$

donde

$$\mathbf{S}_{1,i} = (\underline{x}_i - \overline{\underline{x}})(\underline{x}_i - \overline{\underline{x}})' - \mathbf{S}.$$

A partir del teorema 2.4.4 y del corolario 2.4.5, se obtienen las funciones de influencia muestrales de $\widehat{\mathbf{\Sigma}}$ y $\mathbf{S}$.

Corolario 2.4.6 *En las condiciones del teorema 2.4.4, se verifica que*

$$FIM\left(\mathbf{x}_I; \widehat{\mathbf{\Sigma}}\right) = \sum_{i\in I}\widehat{\mathbf{\Sigma}}_{1,i} + \frac{1}{n-r}\sum_{i\in I}(\underline{x}_i - \overline{\underline{x}})\sum_{l\in I}(\underline{x}_l - \overline{\underline{x}})',$$

$$FIM\left(\mathbf{x}_I; \mathbf{S}\right) = \sum_{i\in I}\mathbf{S}_{1,i} + \frac{1}{n-r-1}\left[\sum_{i\in I}\mathbf{S}_{1,i} + \sum_{i\in I}(\underline{x}_i - \overline{\underline{x}})\sum_{l\in I}(\underline{x}_l - \overline{\underline{x}})'\right].$$

Los coeficientes de los desarrollos en serie de los autovalores y autovectores de $\widehat{\mathbf{\Sigma}}^{(I)}$ y $\mathbf{S}^{(I)}$ se pueden presentar de forma conjunta unificando la notación, tal como se indicó en el primer capítulo de esta memoria, denotando por $\widehat{\widetilde{\mathbf{\Sigma}}}$ a cualquiera de las dos estimaciones habituales de la matriz de covarianzas poblacionales y considerando la constante

$$a = \begin{cases} (n-r)^{-1} & \text{si } \widehat{\widetilde{\mathbf{\Sigma}}} = \widehat{\mathbf{\Sigma}} \\ (n-r-1)^{-1} & \text{si } \widehat{\widetilde{\mathbf{\Sigma}}} = \mathbf{S}. \end{cases}$$

De esta forma, teniendo en cuenta los desarrollos de $\widetilde{\mathbf{\Sigma}}$ y $\mathbf{S}$, (2.20) y (2.21), respectivamente, se puede demostrar el siguiente teorema.

Teorema 2.4.7 *Sea $\underline{X}$ un vector aleatorio p-dimensional, con matriz de covarianzas $\mathbf{\Sigma}$. Sea $\underline{X}_1, ..., \underline{X}_n$ una muestra aleatoria del vector $\underline{X}$ y $\underline{x}_1, ..., \underline{x}_n$ una realización muestral de la misma. Sean $\widehat{\widetilde{\lambda}}_1 \geq \cdots \geq \widehat{\widetilde{\lambda}}_p$ los autovalores de $\widehat{\widetilde{\mathbf{\Sigma}}}$ y $\widehat{\widetilde{\underline{\alpha}}}_1, ..., \widehat{\widetilde{\underline{\alpha}}}_p$ autovectores ortonormales asociados, respectivamente. Sea*

$I \subset \{1,...,n\}$ tal que $card\,(I) = r$, y $\mathbf{x}_I = \{\underline{x}_i \ / \ i \in I\}$. Entonces, existe un conjunto de autovectores ortonormales de $\widehat{\widehat{\mathbf{\Sigma}}}^{(I)}$, $\widehat{\widehat{\underline{\alpha}}}_j^{(I)}$, $j = 1,...,p$, tales que, ellos y sus autovalores asociados, $\widehat{\widehat{\lambda}}_j^{(I)}$, $j = 1,...,p$, se pueden expresar según desarrollos en serie de potencias convergentes. En particular, si $\widehat{\widehat{\lambda}}_k$ es un autovalor simple de $\widehat{\widehat{\mathbf{\Sigma}}}$,

$$\begin{aligned}
\widehat{\widehat{\lambda}}_k^{(I)} &\equiv \widehat{\widehat{\lambda}}_k - a\widehat{\widehat{\nu}}_{I,k} + \frac{a^2}{2}\widehat{\widehat{\pi}}_{I,k} + O\left(n^{-3}\right), \\
\widehat{\widehat{\underline{\alpha}}}_k^{(I)} &\equiv \widehat{\widehat{\underline{\alpha}}}_k - a\widehat{\widehat{\underline{\beta}}}_{I,k} + \frac{a^2}{2}\widehat{\widehat{\underline{\gamma}}}_{I,k} + O\left(n^{-3}\right),
\end{aligned}$$

donde, si $\widehat{\widehat{y}}_{ij} = \widehat{\widehat{\underline{\alpha}}}_j'\left(\underline{x}_i - \overline{\underline{x}}\right)$,

$$\widehat{\widehat{\nu}}_{I,k} = \sum_{i\in I}\left(\widehat{\widehat{y}}_{ik}^2 - \widehat{\widehat{\lambda}}_k\right), \tag{2.22}$$

$$\widehat{\widehat{\underline{\beta}}}_{I,k} = -\sum_{i\in I}\sum_{j\neq k}\frac{\widehat{\widehat{y}}_{ij}\widehat{\widehat{y}}_{ik}}{\widehat{\widehat{\lambda}}_j - \widehat{\widehat{\lambda}}_k}\widehat{\widehat{\underline{\alpha}}}_j, \tag{2.23}$$

$$\widehat{\widehat{\pi}}_{I,k} = -2\left[\sum_{i\in I}\sum_{j\neq k}\frac{\left(\widehat{\widehat{y}}_{ij}\widehat{\widehat{y}}_{ik}\right)^2}{\widehat{\widehat{\lambda}}_j - \widehat{\widehat{\lambda}}_k} + \left(\sum_{i\in I}\widehat{\widehat{y}}_{ik}\right)^2\right], \tag{2.24}$$

$$\begin{aligned}
\widehat{\widehat{\underline{\gamma}}}_{I,k} = 2\sum_{j\neq k}\Bigg[&\frac{1}{\widehat{\widehat{\lambda}}_j - \widehat{\widehat{\lambda}}_k}\sum_{i_1\in I}\widehat{\widehat{y}}_{i_1 j}\sum_{i_2\in I}\widehat{\widehat{y}}_{i_2 k} - \frac{1}{\widehat{\widehat{\lambda}}_j - \widehat{\widehat{\lambda}}_k}\sum_{i\in I}\widehat{\widehat{y}}_{ij}\widehat{\widehat{y}}_{ik} + \\
&+\sum_{l\neq k}\frac{1}{\left(\widehat{\widehat{\lambda}}_j - \widehat{\widehat{\lambda}}_k\right)\left(\widehat{\widehat{\lambda}}_l - \widehat{\widehat{\lambda}}_k\right)}\sum_{i_1\in I}\widehat{\widehat{y}}_{i_1 l}\widehat{\widehat{y}}_{i_1 k}\sum_{i_2\in I}\widehat{\widehat{y}}_{i_2 j}\widehat{\widehat{y}}_{i_2 l} - \\
&- \frac{1}{\left(\widehat{\widehat{\lambda}}_j - \widehat{\widehat{\lambda}}_k\right)^2}\sum_{i_1\in I}\widehat{\widehat{y}}_{i_1 j}\widehat{\widehat{y}}_{i_1 k}\sum_{i_2\in I}\widehat{\widehat{y}}_{i_2 k}^2\Bigg]\widehat{\widehat{\underline{\alpha}}}_j - \\
&- \sum_{j\neq k}\frac{1}{\left(\widehat{\widehat{\lambda}}_j - \widehat{\widehat{\lambda}}_k\right)^2}\left[\sum_{i\in I}\widehat{\widehat{y}}_{ij}\widehat{\widehat{y}}_{ik}\right]^2\widehat{\widehat{\underline{\alpha}}}_k.
\end{aligned} \tag{2.25}$$

Demostración

Aplicando el teorema 2.2.2 a los desarrollos (2.20), (2.21) se tiene que los coeficientes del desarrollo de los autovalores y autovectores son

$$\widehat{\widehat{\nu}}_{I,k} = \widehat{\widehat{\underline{\alpha}}}_k' \sum_{i\in I} \left[(\underline{x}_i - \overline{\underline{x}})(\underline{x}_i - \overline{\underline{x}})' - \widehat{\widehat{\mathbf{\Sigma}}} \right] \widehat{\widehat{\underline{\alpha}}}_k = \sum_{i\in I} \left(\widehat{\widehat{y}}_{ik}^2 - \widehat{\widehat{\lambda}}_k \right).$$

Por otro lado,

$$\begin{aligned} \widehat{\widehat{\underline{\beta}}}_{I,k} &= -\sum_{j\neq k} \widehat{\widehat{\underline{\alpha}}}_j' \sum_{i\in I} \left[(\underline{x}_i - \overline{\underline{x}})(\underline{x}_i - \overline{\underline{x}})' - \widehat{\widehat{\mathbf{\Sigma}}} \right] \widehat{\widehat{\underline{\alpha}}}_k \frac{1}{\widehat{\widehat{\lambda}}_j - \widehat{\widehat{\lambda}}_k} \widehat{\widehat{\underline{\alpha}}}_j = \\ &= -\sum_{i\in I} \sum_{j\neq k} \frac{\widehat{\widehat{y}}_{ij}\widehat{\widehat{y}}_{ik}}{\widehat{\widehat{\lambda}}_j - \widehat{\widehat{\lambda}}_k} \widehat{\widehat{\underline{\alpha}}}_j. \end{aligned}$$

En cuanto a $\widehat{\widehat{\pi}}_{I,k}$,

$$\begin{aligned} \widehat{\widehat{\pi}}_{I,k} &= \widehat{\widehat{\underline{\alpha}}}_k' \left[-2 \sum_{i\in I} (\underline{x}_i - \overline{\underline{x}}) \sum_{l\in I} (\underline{x}_l - \overline{\underline{x}})' \right] \widehat{\widehat{\underline{\alpha}}}_k - \\ &\quad -2 \sum_{i\in I} \sum_{j\neq k} \frac{\left(\widehat{\widehat{y}}_{ij}\widehat{\widehat{y}}_{ik} \right)^2}{\widehat{\widehat{\lambda}}_j - \widehat{\widehat{\lambda}}_k} = \\ &= -2 \left[\sum_{i\in I} \sum_{j\neq k} \frac{\left(\widehat{\widehat{y}}_{ij}\widehat{\widehat{y}}_{ik} \right)^2}{\widehat{\widehat{\lambda}}_j - \widehat{\widehat{\lambda}}_k} + \left(\sum_{i\in I} \widehat{\widehat{y}}_{ik} \right)^2 \right]. \end{aligned}$$

Y finalmente,

$$\begin{aligned} \widehat{\widehat{\underline{\gamma}}}_{I,k} &= 2 \sum_{j\neq k} \frac{1}{\widehat{\widehat{\lambda}}_j - \widehat{\widehat{\lambda}}_k} \sum_{i\in I} \widehat{\widehat{y}}_{ij} \sum_{l\in I} \widehat{\widehat{y}}_{lk} + \\ &\quad +2 \sum_{j\neq k} \sum_{l\neq k} \frac{1}{\left(\widehat{\widehat{\lambda}}_j - \widehat{\widehat{\lambda}}_k \right)\left(\widehat{\widehat{\lambda}}_l - \widehat{\widehat{\lambda}}_k \right)} \sum_{i\in I} \widehat{\widehat{y}}_{il}\widehat{\widehat{y}}_{ik} \left[\sum_{i\in I} \widehat{\widehat{y}}_{ij}\widehat{\widehat{y}}_{il} - \widehat{\widehat{\lambda}}_j \delta_{jl} \right] \widehat{\widehat{\underline{\alpha}}}_j - \\ &\quad -2 \sum_{j\neq k} \sum_{l\neq k} \frac{1}{\left(\widehat{\widehat{\lambda}}_j - \widehat{\widehat{\lambda}}_k \right)\left(\widehat{\widehat{\lambda}}_l - \widehat{\widehat{\lambda}}_k \right)} \sum_{i\in I} \widehat{\widehat{y}}_{ik}\widehat{\widehat{y}}_{il} \sum_{i\in I} \widehat{\widehat{y}}_{ik}^2 \delta_{jl} \widehat{\widehat{\underline{\alpha}}}_j - \end{aligned}$$

$$
\begin{aligned}
&-\sum_{j\neq k}\left[\sum_{i\in I}\frac{\widehat{\widehat{y}}_{ij}\widehat{\widehat{y}}_{ik}}{\widehat{\widehat{\lambda}}_j-\widehat{\widehat{\lambda}}_k}\right]^2\widehat{\widehat{\underline{\alpha}}}_k=\\
&=2\sum_{j\neq k}\frac{1}{\widehat{\widehat{\lambda}}_j-\widehat{\widehat{\lambda}}_k}\sum_{i\in I}\widehat{\widehat{y}}_{ij}\sum_{l\in I}\widehat{\widehat{y}}_{lk}+\\
&+2\sum_{j\neq k}\sum_{l\neq k}\frac{1}{\left(\widehat{\widehat{\lambda}}_j-\widehat{\widehat{\lambda}}_k\right)\left(\widehat{\widehat{\lambda}}_l-\widehat{\widehat{\lambda}}_k\right)}\sum_{i\in I}\widehat{\widehat{y}}_{il}\widehat{\widehat{y}}_{ik}\left[\sum_{i\in I}\widehat{\widehat{y}}_{ij}\widehat{\widehat{y}}_{il}-\widehat{\widehat{\lambda}}_j\delta_{jl}\right]\widehat{\widehat{\underline{\alpha}}}_j-\\
&-2\sum_{j\neq k}\frac{1}{\left(\widehat{\widehat{\lambda}}_j-\widehat{\widehat{\lambda}}_k\right)^2}\sum_{i\in I}\widehat{\widehat{y}}_{ik}\widehat{\widehat{y}}_{ij}\sum_{i\in I}\left(\widehat{\widehat{y}}^2_{ik}-\widehat{\widehat{\lambda}}_k\right)\widehat{\widehat{\underline{\alpha}}}_j-\\
&-\sum_{j\neq k}\frac{1}{\left(\widehat{\widehat{\lambda}}_j-\widehat{\widehat{\lambda}}_k\right)^2}\left[\sum_{i\in I}\widehat{\widehat{y}}_{ij}\widehat{\widehat{y}}_{ik}\right]^2\widehat{\widehat{\underline{\alpha}}}_k
\end{aligned}
$$

de donde, agrupando convenientemente, se obtiene la expresión (2.25). ■

Como consecuencia del resultado del teorema 2.4.7, se puede obtener una aproximación de la función de influencia muestral de los autovalores y autovectores de las estimaciones de $\boldsymbol{\Sigma}$ para un conjunto de observaciones. Para ello se denotará por $\widehat{\nu}_{I,k}$, $\widehat{\pi}_{I,k}$, $\widehat{\underline{\beta}}_{I,k}$ y $\widehat{\underline{\gamma}}_{I,k}$ a los coeficientes $\widehat{\widehat{\nu}}_{I,k}$, $\widehat{\widehat{\pi}}_{I,k}$, $\widehat{\widehat{\underline{\beta}}}_{I,k}$ y $\widehat{\widehat{\underline{\gamma}}}_{I,k}$ del teorema 2.4.7 para la matriz $\widehat{\boldsymbol{\Sigma}}$, y por $\widetilde{\nu}_{I,k}$, $\widetilde{\pi}_{I,k}$, $\widetilde{\underline{\beta}}_{I,k}$ y $\widetilde{\underline{\gamma}}_{I,k}$ a los mismos coeficientes para la matriz $\mathbf{S}$.

Corolario 2.4.8 *Sean $\underline{X}$ un vector aleatorio con matriz de covarianzas $\boldsymbol{\Sigma}$, $\underline{X}_1,\ldots,\underline{X}_n$ una muestra aleatoria del vector $\underline{X}$ y $\underline{x}_1,\ldots,\underline{x}_n$ una realización muestral de la misma. Si $\widehat{\lambda}_k$ es un autovalor simple de la matriz $\widehat{\boldsymbol{\Sigma}}$, y $\widehat{\underline{\alpha}}_k$ un autovector unitario asociado,*

$$
\begin{aligned}
FIM\left(\mathbf{x}_I;\widehat{\lambda}_k\right) &= \widehat{\nu}_{Ik}-\frac{1}{2}\frac{1}{n-r}\widehat{\pi}_{Ik}+O\left(n^{-2}\right)\\
y\qquad FIM\left(\mathbf{x}_I;\widehat{\underline{\alpha}}_k\right) &= \widehat{\underline{\beta}}_{Ik}-\frac{1}{2}\frac{1}{n-r}\widehat{\underline{\gamma}}_{Ik}+O\left(n^{-2}\right).
\end{aligned}
$$

Corolario 2.4.9 *Sean $\underline{X}$ un vector aleatorio con matriz de covarianzas $\boldsymbol{\Sigma}$, $\underline{X}_1,\ldots,\underline{X}_n$ una muestra aleatoria del vector $\underline{X}$ y $\underline{x}_1,\ldots,\underline{x}_n$ una realización muestral de la misma. Si $\widetilde{\lambda}_k$ es un autovalor simple de $\mathbf{S}$, y $\widetilde{\underline{\alpha}}_k$ un autovector*

unitario asociado, entonces,

$$FIM\left(\mathbf{x}_I;\widetilde{\lambda}_k\right) = \widetilde{\nu}_{Ik} + \frac{1}{n-r-1}\left[\widetilde{\nu}_{Ik} - \frac{1}{2}\widetilde{\pi}_{Ik}\right] + O\left(n^{-2}\right)$$

$$y \qquad FIM\left(\mathbf{x}_I;\underline{\widetilde{\alpha}}_k\right) = \underline{\widetilde{\beta}}_{Ik} + \frac{1}{n-r-1}\left[\underline{\widetilde{\beta}}_{Ik} - \frac{1}{2}\underline{\widetilde{\gamma}}_{Ik}\right] + O\left(n^{-2}\right).$$

Nota 2.4.2 *Los comentarios realizados en las notas 2.3.1, 2.3.3 y 2.3.4, se pueden generalizar al análisis de influencia conjunta.*

Además, dado que para un tamaño muestral suficientemente grande, los términos dominantes de $FIM\left(\mathbf{x}_I;\widehat{\widetilde{\lambda}}_k\right)$ *y* $FIM\left(\mathbf{x}_I;\underline{\widehat{\widetilde{\alpha}}}_k\right)$, *son, respectivamente,* $\widehat{\widetilde{\nu}}_{Ik}$ *y* $\underline{\widehat{\widetilde{\beta}}}_{Ik}$, *teniendo en cuenta las expresiones (2.22) y (2.23), se pueden escribir de la forma*

$$\widehat{\widetilde{\nu}}_{Ik} = \sum_{i\in I}\widetilde{\nu}_{ik} \quad y \quad \underline{\widehat{\widetilde{\beta}}}_{Ik} = \sum_{i\in I}\underline{\widetilde{\beta}}_{ik}.$$

Por ello, se pueden detectar conjuntos de observaciones altamente influyentes a partir de un análisis individual.

Capítulo 3

SESGO CONDICIONADO DE UN AUTOVALOR

3.1 Introducción

Con objeto de abordar el Análisis de Influencia en el Análisis de Componentes Principales según la línea marcada por Muñoz Pichardo y otros [50], en este capítulo se obtiene una aproximación del sesgo condicionado de los autovalores de los estimadores de la matriz de covarianzas utilizados habitualmente, $\mathbf{S}$ y $\widehat{\Sigma}$.

Al basarse el sesgo condicionado en el valor esperado del estadístico en cuestión, será necesario suponer una hipótesis distribucional previa. Este trabajo se centra en la hipótesis de normalidad multivariante, condición es muy habitual en la práctica, en experiencias biológicas, psicológicas, clínicas, agrícolas, etc. Otra restricción del estudio, afecta a los autovalores analizados de la matriz de covarianzas, que deben ser simples y no nulos, condición también necesaria en el Análisis de Influencia llevado a cabo mediante funciones de influencia. Los resultados obtenidos, al igual que en otros campos, se relacionarán con las funciones de influencia de los parámetros correspondientes.

Lawley [40] calculó los primeros términos del desarrollo en serie del valor esperado de un autovalor de $\mathbf{S}$, basándose en el desarrollo en serie de dicho autovalor, en función de parámetros poblacionales. Para encontrar los valores del sesgo condicionado de un autovalor y un autovector de $\mathbf{S}$, se sigue un razonamiento similar, obteniendo así, aproximaciones de ellos. Los términos de estos desarrollos poseen factores del tipo $S_{lm} - \sigma_{lm}$, donde S_{lm} y σ_{lm} representan, respectivamente, la covarianza muestral y poblacional entre dos variables, por lo que es necesario calcular, previamente, los sesgos

condicionados de productos entre tales términos.

En la sección 3.2, se justifica el número mínimo de términos considerado en las aproximaciones del sesgo condicionado de los estadísticos de interés. En la siguiente sección, se presenta una serie de resultados previos, necesarios para el cálculo del sesgo condicionado de autovalores y autovectores del estimador de la matriz de covarianzas. En la última sección de este capítulo, se calcula una aproximación del sesgo condicionado de los autovalores de $\mathbf{S}$ y $\widehat{\mathbf{\Sigma}}$, y se acompaña con una serie de comentarios e interpretaciones de dicha herramienta en el Análisis de Influencia.

3.2 Consideraciones previas

En la práctica, si la suma de una serie convergente no es conocida, ésta se suele truncar y aproximar mediante una suma finita, basándose en la idea de que, al ser la serie convergente, el resto que converge a cero. No obstante, conviene fijar, justificadamente, el orden mínimo necesario para llevar a cabo el Análisis de Influencia, no sólo por elegancia del trabajo en sí, sino por coherencia interna del mismo.

En primer lugar, en las funciones de influencia muestrales recogidas en el capítulo anterior se proporcionan los coeficientes de términos de orden superior a n^{-2}. Esto es equivalente a considerar los términos de orden superior a n^{-3} en los desarrollos obtenidos en el cálculo del sesgo condicionado, ya que, en general, si $\underline{X}$ es una variable aleatoria, $\underline{X}_i$ es un elemento de una muestra, $\underline{x}_i$ es el valor de la realización muestral de $\underline{X}_i$ y T es un estadístico, entonces

$$\begin{aligned} FIM\left(\underline{x}_i;T\right) &= (n-1)\left(T-T^{(i)}\right) \qquad \text{y} \\ S\left(\underline{x}_i;T\right) &= E\left[T \mid \underline{X}=\underline{x}_i\right]-E\left[T\right]. \end{aligned}$$

Tanto la función de influencia muestral como el sesgo condicionado comparan el estadístico T en dos situaciones distintas, pero en la función de influencia muestral, esta comparación aparece multiplicada por $n-1$. Por lo que el orden de la función de influencia muestral, respecto al tamaño muestral, es una unidad superior al del sesgo condicionado. Por ello conviene hallar los coeficientes de los términos de orden n^{-2} y superiores, en las expresiones que se obtengan a lo largo de este capítulo.

Por otra parte, otro argumento que justifica y conduce a la propuesta anterior está basada en el análisis detallado del cambio producido en los

estadísticos por perturbaciones del modelo. En particular, por venir caracterizado un vector por su módulo y ángulo, la variación sufrida tras una perturbación se puede medir en función del cambio observado en alguna de estas dos características.

El módulo de los autovectores se ha fijado de antemano con el valor 1, por lo que el ángulo determinado por los autovectores, o equivalentemente, su coseno, puede dar una idea de la variación sufrida. Para justificar la necesidad de los términos de orden n^{-2}, por simplicidad, basta mostrarlo para el caso de influencia de observaciones individuales.

Si se denota por $\left(\widetilde{\underline{\alpha}}_k, \widetilde{\underline{\alpha}}_k^{(i)}\right)$ el ángulo formado por $\widetilde{\underline{\alpha}}_k$ y $\widetilde{\underline{\alpha}}_k^{(i)}$, utilizando la expresión (2.19), que también es válida para los autovectores del estimador $\mathbf{S}$, su coseno se puede expresar por

$$\begin{aligned}\cos\left(\left(\widetilde{\underline{\alpha}}_k, \widetilde{\underline{\alpha}}_k^{(i)}\right)\right) &= \widetilde{\underline{\alpha}}_k' \widetilde{\underline{\alpha}}_k^{(i)} = \\ &= \widetilde{\underline{\alpha}}_k' \widetilde{\underline{\alpha}}_k + \frac{1}{n-2} \widetilde{\underline{\alpha}}_k' \widetilde{\underline{\beta}}_{ik} + \frac{1}{2} \frac{1}{(n-2)^2} \widetilde{\underline{\alpha}}_k' \widetilde{\underline{\gamma}}_{ik} + O(n^{-3}).\end{aligned}$$

Dado que

$$\widetilde{\underline{\alpha}}_k' \widetilde{\underline{\beta}}_{ik} = -\widetilde{y}_{ik} \sum_{j \neq k} \frac{\widetilde{y}_{ij}}{\widetilde{\lambda}_j - \widetilde{\lambda}_k} \widetilde{\underline{\alpha}}_k' \widetilde{\underline{\alpha}}_j = 0 \qquad \text{y}$$

$$\begin{aligned}\widetilde{\underline{\alpha}}_k' \widetilde{\underline{\gamma}}_{ik} &= -\widetilde{y}_{ik}^2 \, \widetilde{b}_{i(k)}(2,2) \, \widetilde{\underline{\alpha}}_k' \widetilde{\underline{\alpha}}_k - 2 \, \widetilde{b}_{i(k)}(2,1) \, \widetilde{\underline{\alpha}}_k' \widetilde{\underline{\beta}}_{ik} - 2 \, \widetilde{y}_{ik}^3 \sum_{j \neq k} \frac{\widetilde{y}_{ij}}{(\widetilde{\lambda}_j - \widetilde{\lambda}_k)^2} \widetilde{\underline{\alpha}}_k' \widetilde{\underline{\alpha}}_j + \\ &\quad + O(n^{-3}) = -\widetilde{y}_{ik}^2 \sum_{j \neq k} \frac{\widetilde{y}_{ij}^2}{(\widetilde{\lambda}_j - \widetilde{\lambda}_k)^2} + O(n^{-3}),\end{aligned}$$

se obtiene que

$$\cos\left(\left(\widetilde{\underline{\alpha}}_k, \widetilde{\underline{\alpha}}_k^{(i)}\right)\right) = 1 - \frac{1}{2\,(n-2)^2} \widetilde{y}_{ik}^2 \sum_{j \neq k} \frac{\widetilde{y}_{ij}^2}{(\widetilde{\lambda}_j - \widetilde{\lambda}_k)^2} + O(n^{-3}). \qquad (3.1)$$

Si se omiten los términos de orden n^{-2} e inferiores, entonces se tiene la aproximación $\cos\left(\left(\widetilde{\underline{\alpha}}_k, \widetilde{\underline{\alpha}}_k^{(i)}\right)\right) \simeq 1$ y ninguna observación se podría considerar influyente sobre un autovector a través de medidas obtenidas con dicha precisión.

3.3 Resultados previos

A continuación se muestran algunos resultados precisos para hallar el sesgo condicionado de un autovalor de $\mathbf{S} = (S_{lm})$ y de sus autovectores asociado. El cálculo de una aproximación de éste se realizará a partir del desarrollo en serie de un autovalor de la matriz de covarianzas muestrales, proporcionado por Lawley [40] y el cual se recoge en el siguiente teorema.

Teorema 3.3.1 *Sea $\underline{X}$ vector aleatorio p-dimensional con matriz de covarianzas $\mathbf{\Lambda} = diag\{\lambda_1, ..., \lambda_p\}$, siendo $\lambda_1, ..., \lambda_p$ distintos y no nulos. Sea $\underline{X}_1, ..., \underline{X}_n$ una muestra aleatoria del vector $\underline{X}$ y $\mathbf{T} = (T_{jl})$ la matriz de covarianzas muestrales con autovalores $\widetilde{\lambda}_1 \geq \widetilde{\lambda}_2 \geq \cdots \geq \widetilde{\lambda}_p$. Entonces, para tamaño muestral suficientemente grande, $\widetilde{\lambda}_k$ se puede expresar según el desarrollo en serie convergente*

$$\widetilde{\lambda}_k = \lambda_k + (T_{kk} - \lambda_k) - \sum_{j \neq k} \frac{T_{kj}^2}{\lambda_j - \lambda_k} - (T_{kk} - \lambda_k) \sum_{j \neq k} \frac{T_{kj}^2}{(\lambda_j - \lambda_k)^2} + \\ + \sum_{j \neq k} \sum_{l \neq k} \frac{T_{kj} (T_{jl} - \lambda_j \delta_{jl}) T_{lk}}{(\lambda_j - \lambda_k)(\lambda_l - \lambda_k)} + R_k, \tag{3.2}$$

donde $k = 1, ..., p$ y R_k representa los términos formados por productos de cuatro o más factores del tipo $T_{jl} - \lambda_j \delta_{jl}$.

Nota 3.3.1 *El desarrollo en serie de $\widetilde{\lambda}_k$ dado en (3.2) es válido también si existen autovalores poblaciones múltiples, siempre que λ_k sea simple. Esto se justifica mediante la analiticidad de los autovalores muestrales en tales condiciones (Girshick [21]).*

En el teorema 3.3.1 se parte del autovalor muestral de un vector aleatorio de componentes incorreladas, lo cual no supone restricción, ya que los autovalores son invariantes ante transformaciones ortogonales. La transformación ortogonal que se realizará, será $\mathbf{T} = \mathbf{ASA}'$, con $\mathbf{A}$ la matriz de autovectores de $\mathbf{\Sigma}$.

Para obtener una aproximación del sesgo condicionado de un autovalor de $\mathbf{S}$, será necesario hallar esta característica sobre productos de factores del tipo $T_{lm} - \delta_{lm}\lambda_l$. Los cálculos se realizarán para productos de términos del tipo $S_{lm} - \sigma_{lm}$ y posteriormente se particularizarán para variables incorreladas, cuando sea necesario.

En primer lugar, se pasa a describir la notación a utilizar en esta y otras secciones. Dado un conjunto de índices I, se denotará con el superíndice I

en un estadístico, a aquél calculado únicamente con las observaciones de la muestra con índices en I, y con el superíndice (I) a aquél calculado únicamente con el resto de las observaciones de la muestra.

Por otro lado, se considera el conjunto

$$\mathcal{C}_2(j_1,...,j_q) = \{\text{subconjuntos de 2 elementos de } \{j_1,...,j_q\}\}.$$

Dado un conjunto A, si $\underline{j}_l \in A$, se denota por $\overline{\underline{j}}_l$ a

$$\overline{\underline{j}}_l = A \setminus \left\{\underline{j}_l\right\}.$$

Además, se denotarán las variables aleatorias

$$\begin{aligned} B_1(l,m) &= \frac{n-r-1}{n-1}\left(S_{lm}^{(I)} - \sigma_{lm}\right), \\ B_2(l,m) &= \frac{1}{n-1}\left[-r\sigma_{lm} + (r-1)\,S_{lm}^{I}\right] \qquad \text{y} \\ B_3(l,m) &= \frac{r(n-r)}{n(n-1)}\left(\overline{X}_l^I - \overline{X}_l^{(I)}\right)\left(\overline{X}_m^I - \overline{X}_m^{(I)}\right); \end{aligned}$$

y las constantes

$$\begin{aligned} \sigma(l,m) &= \sigma_{lm}, \qquad \mu(l_1,...,l_q) = \prod_{j=1}^{q}\mu_{l_j}, \qquad a(l_1,...,l_q) = \prod_{j=1}^{q}\left(\overline{x}_{l_j}^I - \mu_{l_j}\right), \\ b_2(l,m) &= -r\sigma_{lm} + (r-1)\,s_{lm}^I, \\ c(l,m) &= b_2(l,m) + \frac{r(n-r)}{n}a(l,m) + \frac{r}{n}\sigma(l,m). \end{aligned}$$

donde $\overline{x}_l^I$, s_{lm}^I representan el valor de los estadísticos $\overline{X}_l$, S_{lm} para los elementos de una realización muestral con índices en I.

Y para ciertas constantes fijadas, a_{l_j}, $j = 1,...,q$, se denota por

$$p(l_1,...,l_q) = \prod_{j=1}^{q}\left(a_{l_j} - \mu_{l_j}\right).$$

En primer lugar, se proporciona un resultado recogido por Miller [47]:

Lema 3.3.2 *Sea* $\underline{X} = (X_1, ..., X_p)' \sim \mathcal{N}_p(\underline{\mu}, \mathbf{\Sigma})$ *con* $\underline{\mu} = (\mu_1, ..., \mu_p)'$ *y* $\mathbf{\Sigma} = (\sigma_{lm})$. *Para* $l_j = 1, ..., p$, $j = 1, ..., 4$, *se verifica que*

$$E\left[\prod_{j=1}^{4} X_{l_j}\right] = \sum_{j \in \{l_2, l_3, l_4\}} \sigma(l_1, j)\, \sigma\left(\overline{j}\right) + \sum_{\underline{j} \in\, C_2(l_1, l_2, l_3, l_4)} \sigma\left(\underline{j}\right) \mu\left(\overline{j}\right) + \mu(l_1, l_2, l_3, l_4).$$

Utilizando las propiedades de la distribución normal y aplicando el lema 3.3.2 se obtiene el siguiente resultado.

Lema 3.3.3 *Sea* $\underline{X} = (X_1, ..., X_p)' \sim \mathcal{N}_p(\underline{\mu}, \mathbf{\Sigma})$ *con* $\underline{\mu} = (\mu_1, ..., \mu_p)'$ *y* $\mathbf{\Sigma} = (\sigma_{lm})$. *Sea* $\underline{X}_1, ..., \underline{X}_n$ *una muestra aleatoria del vector* $\underline{X}$. *Entonces, para* $l_j = 1, ..., p$, *y para cualesquiera valores reales* a_{l_j}, *con* $j = 1, ..., 4$, *se verifica que*

1. $$E\left[\prod_{j=1}^{2}\left(a_{l_j} - \overline{X}_{l_j}^{(I)}\right)\right] = p(l_1, l_2) + \frac{1}{n-r}\sigma(l_1, l_2),$$

2. $$E\left[\prod_{j=1}^{4}\left(a_{l_j} - \overline{X}_{l_j}^{(I)}\right)\right] = p(l_1, l_2, l_3, l_4) + \frac{1}{n-r}\sum_{\underline{j} \in\, C_2(l_1, l_2, l_3, l_4)} \sigma\left(\underline{j}\right) p\left(\overline{j}\right) \sigma\left(\overline{j}\right) + \\ + \frac{1}{(n-r)^2}\sum_{j \in \{l_2, l_3, l_4\}} \sigma(l_1, j).$$

Wishart [72] recoge el siguiente resultado sobre el valor esperado del producto de factores del tipo $S_{lm} - \sigma_{lm}$:

Lema 3.3.4 *Sea* $\underline{X} = (X_1, ..., X_p)' \sim \mathcal{N}_p(\underline{\mu}, \mathbf{\Sigma})$ *con* $\mathbf{\Sigma} = (\sigma_{lm})$. *Sea* $\underline{X}_1, ..., \underline{X}_n$ *una muestra aleatoria del vector* $\underline{X}$ *y* $\mathbf{S} = (S_{lm})$ *la matriz de covarianzas muestrales. Entonces, para* $a, b, l, m, u, v, w, z = 1, ..., p$,

1. $$E\left[(S_{lm} - \sigma_{lm})(S_{uv} - \sigma_{uv})\right] = \frac{1}{n-1}\sum_{j \in \{u, v\}} \sigma(l, j)\, \sigma\left(m, \overline{j}\right).$$

2. $$E\left[(S_{lm} - \sigma_{lm})(S_{uv} - \sigma_{uv})(S_{wz} - \sigma_{wz})\right] = \\ = \frac{1}{(n-1)^2}\sum_{\underline{j} \in \{\{u, v\}, \{w, z\}\}} \sum_{j_1 \in \underline{j}} \sum_{j_2 \in \overline{\underline{j}}} \sigma(l, j_1)\, \sigma(m, j_2)\, \sigma\left(\overline{j_1}, \overline{j_2}\right).$$

3. $E\left[\left(S_{ab}-\sigma_{ab}\right)\left(S_{lm}-\sigma_{lm}\right)\left(S_{uv}-\sigma_{uv}\right)\left(S_{wz}-\sigma_{wz}\right)\right] =$

$$= \frac{1}{(n-1)^2} \sum_{\underline{j}\in\{\{l,m\},\{u,v\},\{w,z\}\}} \sum_{j_1\in\underline{j}} \sigma\left(a,j_1\right)\sigma\left(b,\overline{j}_1\right)\cdot$$

$$\cdot \sum_{j_3\in\underline{j}_2/\overline{j}=\{\underline{j}_2,\overline{j}_2\},\overline{j}_2=\{j_4,\overline{j_4}\}} \sigma\left(j_3,j_4\right)\sigma\left(\overline{j_3},\overline{j_4}\right)+O\left(n^{-3}\right).$$

En la bibliografía existen referencias respecto a los momentos centrados de productos de más de cuatro factores del tipo $S_{lm}-\sigma_{lm}$, los cuales son de orden n^{-3} o inferior (Lawley [40]).

Para el cálculo del sesgo condicionado de un autovalor de $\mathbf{S}$ dado un conjunto de observaciones $\mathbf{x}_I$, también es necesario obtener la esperanza del producto de factores del tipo $S_{lm}-\sigma_{lm}$, condicionado al valor de los elementos correspondientes de la muestra. Por ello, es conveniente descomponer las componentes de $\mathbf{S}$ según estadísticos calculados solamente con elementos de la muestra de $\mathbf{X}_I$ y estadísticos calculados únicamente con el resto de los elementos.

Muñoz Pichardo y otros [52], en el contexto del Modelo Lineal General, realiza una descomposición de este tipo, la cual se adaptará a la matriz de covarianzas muestrales.

Sea el modelo lineal general multivariante (MLGM)

$$\mathbf{X}=\mathbf{ZB}+\mathbf{E},$$

con $\mathbf{X}\in\mathcal{M}_{n\times q}$, $\mathbf{Z}\in\mathcal{M}_{n\times p}$, $\mathbf{B}\in\mathcal{M}_{p\times q}$ y $\mathbf{E}\in\mathcal{M}_{n\times q}$ y donde $p\leq n$, $\mathbf{X}$ es la matriz de valores del vector respuesta, $\mathbf{Z}$ es una matriz de constantes conocidas de rango $rg\left(\mathbf{Z}\right)\leq p$, $\mathbf{B}$ es la matriz de parámetros desconocidos y $\mathbf{E}=\left(\varepsilon_{ik}\right)$ es la matriz de errores que verifica

$$\begin{aligned} E\left(\mathbf{E}\right) &= \Theta \qquad \text{y} \\ cov\left(\varepsilon_{ik},\varepsilon_{i'k'}\right) &= \sigma_{kk'}\delta_{ii'}, \qquad 1\leq k,k'\leq q, \end{aligned}$$

con Θ la matriz nula y $\delta_{ii'}$ la delta de Kronecker.

Sean las matrices $\mathbf{\Sigma}=\left(\sigma_{kk'}\right)$, $\mathbf{V}=\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-}\mathbf{Z}'$, $\mathbf{M}=\mathbf{I}_n-\mathbf{V}$ y $\mathbf{T}=\dfrac{1}{n-rg\left(\mathbf{Z}\right)}\mathbf{X}'\mathbf{MX}$, y se consideran las descomposiciones

$$\mathbf{X}=\begin{bmatrix}\mathbf{X}_{(I)}\\ \mathbf{X}_I\end{bmatrix}, \quad \mathbf{Z}=\begin{bmatrix}\mathbf{Z}_{(I)}\\ \mathbf{Z}_I\end{bmatrix},$$

donde $\mathbf{X}_{(I)} \in \mathcal{M}_{(n-r)\times q}$, $\mathbf{X}_I \in \mathcal{M}_{r\times q}$, $\mathbf{Z}_{(I)} \in \mathcal{M}_{(n-r)\times p}$, $\mathbf{Z}_I \in \mathcal{M}_{r\times p}$; y

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_{(I)} & \mathbf{V}_0 \\ \mathbf{V}_0' & \mathbf{V}_I \end{bmatrix}, \quad \mathbf{M} = \begin{bmatrix} \mathbf{M}_{(I)} & \mathbf{M}_0 \\ \mathbf{M}_0' & \mathbf{M}_I \end{bmatrix},$$

donde $\mathbf{V}_{(I)}, \mathbf{M}_{(I)} \in \mathcal{M}_{(n-r)\times(n-r)}$, $\mathbf{V}_0, \mathbf{M}_0 \in \mathcal{M}_{(n-r)\times r}$ y $\mathbf{V}_I, \mathbf{M}_I \in \mathcal{M}_{r\times r}$.

De esta descomposición se puede obtener la expresión

$$\mathbf{T} = \frac{1}{n - rg(\mathbf{X})} \left[\mathbf{X}_{(I)}'\mathbf{M}_{(I)}\mathbf{X}_{(I)} + \mathbf{X}_{(I)}'\mathbf{M}_0\mathbf{X}_I + \mathbf{X}_I'\mathbf{M}_0'\mathbf{X}_{(I)} + \mathbf{X}_I'\mathbf{M}_I\mathbf{X}_I\right]. \tag{3.3}$$

Muñoz Pichardo *y otros* [52] demuestran el siguiente resultado:

Teorema 3.3.5 *Sea* $\underline{X} = (X_1, ..., X_p)' \sim \mathcal{N}_p(\underline{\mu}, \mathbf{\Sigma})$ *con* $\underline{\mu} = (\mu_1, ..., \mu_p)'$ *y* $\mathbf{\Sigma} = (\sigma_{lm})$. *Sea* $\underline{X}_1, ..., \underline{X}_n$ *una muestra aleatoria del vector* $\underline{X}$. *En el MLGM, se verifica que*

$$S(\mathbf{x}_I; \mathbf{T}) = \frac{1}{n - rg(\mathbf{X})} \left\{[\mathbf{x}_I - \mathbf{Z}_I\mathbf{B}]'[\mathbf{I}_r - \mathbf{V}_I][\mathbf{x}_I - \mathbf{Z}_I\mathbf{B}] - (tr(\mathbf{I}_r - \mathbf{V}_I))\mathbf{\Sigma}\right\}.$$

Es posible modelizar el problema de la estimación en poblaciones normales a través del MLGM. Un vector aleatorio $\underline{X}$ se puede expresar como

$$\underline{X} = \underline{\mu} + \underline{\varepsilon}, \qquad \underline{\varepsilon} \sim \mathcal{N}_p(\underline{0}, \mathbf{\Sigma}),$$

y por tanto, una muestra aleatoria $\underline{X}_1, ..., \underline{X}_n$ de $\underline{X} \sim \mathcal{N}_p(\underline{\mu}, \mathbf{\Sigma})$, se puede expresar según

$$\mathbf{X} = \left[\underline{X}_1 \ \underline{X}_2 \ \cdots \ \underline{X}_n\right]' = \underline{1}_n\underline{\mu}' + \mathbf{E}, \tag{3.4}$$

donde $\underline{1}_n$ es el vector n-dimensional de componentes unitarias.

En este caso: $\mathbf{V} = \frac{1}{n}\underline{1}_n\underline{1}_n'$, $\mathbf{M} = \mathbf{I}_n - \frac{1}{n}\underline{1}_n\underline{1}_n'$ y

$$\mathbf{T} = \frac{1}{n-1}\mathbf{X}'\mathbf{M}\mathbf{X} = \frac{1}{n-1}\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}} = \mathbf{S}.$$

A partir de la descomposición y la modelización anteriores, las componentes de la matriz $\mathbf{S}$ se pueden expresar también en términos de la descomposición correspondiente a dos conjuntos de observaciones.

Lema 3.3.6 *Sea* $\underline{X} = (X_1, ..., X_p)'$ *y* $var(\underline{X}) = \Sigma = (\sigma_{lm})$. *Sea* $\underline{X}_1, ..., \underline{X}_n$ *una muestra aleatoria del vector* $\underline{X}$. *Sea* $I \subset \{1, ..., n\}$ *tal que* $card(I) = r$. *Entonces, para* $l, m = 1, ..., p$,

$$S_{lm} = \frac{1}{n-1}\left[(n-r-1)\, S_{lm}^{(I)} + (r-1)\, S_{lm}^{I} + (n-1)\, B_3(l,m)\right] \tag{3.5}$$

y

$$S_{lm} - \sigma_{lm} = \sum_{i=1}^{3} B_i(l,m)\,. \tag{3.6}$$

Demostración

Adaptando la expresión (3.3) para el modelo dado en (3.4) y desarrollando, se tiene que

$$\begin{aligned}
\mathbf{S} &= \frac{1}{n-1}\left[\mathbf{X}'_{(I)}\left(\mathbf{I}_{n-r} - \mathbf{Z}_{(I)}(\mathbf{Z}'\mathbf{Z})^{-}\mathbf{Z}'_{(I)}\right)\mathbf{X}_{(I)} - \mathbf{X}'_{(I)}\mathbf{Z}_{(I)}(\mathbf{Z}'\mathbf{Z})^{-}\mathbf{Z}'_{I}\mathbf{X}_{I} + \right.\\
&\quad \left. + \mathbf{X}'_{I}\mathbf{Z}_{I}(\mathbf{Z}'\mathbf{Z})^{-}\mathbf{Z}'_{(I)}\mathbf{X}_{(I)} + \mathbf{X}'_{I}\left(\mathbf{I}_{r} - \mathbf{Z}_{I}(\mathbf{Z}'\mathbf{Z})^{-}\mathbf{Z}'_{I}\right)\mathbf{X}_{I}\right]\\
&= \frac{1}{n-1}\left[\mathbf{X}'_{(I)}\left(\mathbf{I}_{n-r} - \frac{1}{n}\mathbf{1}_{n-r}\mathbf{1}'_{n-r}\right)\mathbf{X}_{(I)} - \frac{1}{n}\mathbf{X}'_{(I)}\mathbf{1}_{n-r}\mathbf{1}'_{r}\mathbf{X}_{I} + \right.\\
&\quad \left. + \frac{1}{n}\mathbf{X}'_{I}\mathbf{1}_{r}\mathbf{1}'_{n-r}\mathbf{X}_{(I)} + \mathbf{X}'_{I}\left(\mathbf{I}_{r} - \frac{1}{n}\mathbf{1}_{r}\mathbf{1}'_{r}\right)\mathbf{X}_{I}\right] =\\
&= \frac{1}{n-1}\left[\mathbf{X}'_{(I)}\mathbf{X}_{(I)} - \frac{1}{n}\mathbf{X}'_{(I)}\mathbf{1}_{n-r}\mathbf{1}'_{n-r}\mathbf{X}_{(I)} - \frac{1}{n}\mathbf{X}'_{(I)}\mathbf{1}_{n-r}\mathbf{1}'_{r}\mathbf{X}_{I} + \right.\\
&\quad \left. + \frac{1}{n}\mathbf{X}'_{I}\mathbf{1}_{r}\mathbf{1}'_{n-r}\mathbf{X}_{(I)} + \mathbf{X}'_{I}\mathbf{X}_{I} - \frac{1}{n}\mathbf{X}'_{I}\mathbf{1}_{r}\mathbf{1}'_{r}\mathbf{X}_{I}\right]\\
&= \frac{1}{n-1}\left[(n-r-1)\,\mathbf{S}^{(I)} + (n-r)\left(1 - \frac{n-r}{n}\right)\overline{\underline{X}}^{(I)}\left(\overline{\underline{X}}^{(I)}\right)' + \right.\\
&\quad + (r-1)\,\mathbf{S}^{I} + r\left(1 - \frac{r}{n}\right)\overline{\underline{X}}^{I}\left(\overline{\underline{X}}^{I}\right)' -\\
&\quad \left. - \frac{r(n-r)}{n}\left(\overline{\underline{X}}^{(I)}\left(\overline{\underline{X}}^{I}\right)' + \overline{\underline{X}}^{I}\left(\overline{\underline{X}}^{(I)}\right)'\right)\right]\\
&= \frac{1}{n-1}\left[(n-r-1)\,\mathbf{S}^{(I)} + (r-1)\,\mathbf{S}^{I} + \right.\\
&\quad \left. + \frac{r(n-r)}{n}\left(\overline{\underline{X}}^{(I)} - \overline{\underline{X}}^{I}\right)\left(\overline{\underline{X}}^{(I)} - \overline{\underline{X}}^{I}\right)'\right].
\end{aligned}$$

La componente (l, m) de $\mathbf{S}$ proporciona la expresión (3.5). Además, para obtener el resultado (3.6) basta restar σ_{lm}, agrupar convenientemente y utilizar la notación indicada al principio de la sección.

■

Por lo tanto se puede enunciar el siguiente teorema, cuya demostración se basa en la modelización (3.4) y el teorema 3.3.5.

Teorema 3.3.7 *Sean* $\underline{X} = (X_1, ..., X_p)' \sim \mathcal{N}_p(\underline{\mu}, \mathbf{\Sigma})$, $\underline{X}_1, ..., \underline{X}_n$ *una muestra aleatoria del vector* $\underline{X}$ *y* $\mathbf{S}$ *la matriz de covarianzas muestrales. Si* $I \subset \{1, ..., n\}$ *tal que card* $(I) = r$ *y* $\mathbf{x}_I \in \mathcal{M}_{r\times p}$, *entonces, para* $l, m = 1, ..., p$,

1. $$S(\mathbf{x}_I; \mathbf{S}) = \frac{1}{n-1}\left\{\left[\mathbf{x}_I - \underline{1}_r\underline{\mu}'\right]'\left[\mathbf{I}_r - \frac{1}{n}\underline{1}_r\underline{1}_r'\right]\left[\mathbf{x}_I - \underline{1}_r\underline{\mu}'\right] - \right.$$
$$\left. - r\left(1 - \frac{1}{n}\right)\mathbf{\Sigma}\right\}.$$

2. $$S(\mathbf{x}_I; S_{lm}) = \frac{1}{n-1}\left[-\frac{r(n-1)}{n}\sigma_{lm} + \sum_{i\in I}(x_{il} - \mu_l)(x_{im} - \mu_m) - \right.$$
$$\left. - \frac{r^2}{n}\left(\overline{x}_l^I - \mu_l\right)\left(\overline{x}_m^I - \mu_m\right)\right].$$

El siguiente resultado será útil, posteriormente, para simplificar algunos cálculos.

Lema 3.3.8 *Sea* $\underline{X} = (X_1, ..., X_p)'$ *un vector aleatorio p-dimensional de vector de medias* $\underline{\mu} = (\mu_1, ..., \mu_p)'$ *y matriz de covarianzas* $\mathbf{\Sigma}$. *Sea* $\underline{X}_1, ..., \underline{X}_n$ *una muestra aleatoria del vector* $\underline{X}$. *Entonces, si* $I \subset \{1, ..., n\}$ *tal que card* $(I) = r$, *para* $l, m = 1, ..., p$,

$$(r-1)S_{lm}^I + r\left(\overline{X}_l^I - \mu_l\right)\left(\overline{X}_m^I - \mu_m\right) = \sum_{i\in I}(X_{il} - \mu_l)(X_{im} - \mu_m). \quad (3.7)$$

Demostración

Dado que $S_{lm}^I = \frac{1}{r-1}\sum_{i\in I}\left(X_{il} - \overline{X}_l^I\right)\left(X_{im} - \overline{X}_m^I\right)$, entonces incluyendo en el primer binomio del sumatorio $\pm\mu_l$ y en el segundo $\pm\mu_m$, se obtiene que

$$\begin{aligned}(r-1)S_{lm}^I &= \sum_{i\in I}(X_{il} - \mu_l)(X_{im} - \mu_m) + \sum_{i\in I}(X_{il} - \mu_l)\left(\mu_m - \overline{X}_m^I\right) + \\ &+ \sum_{i\in I}\left(\mu_l - \overline{X}_l^I\right)(X_{im} - \mu_m) + \sum_{i\in I}\left(\mu_l - \overline{X}_l^I\right)\left(\mu_m - \overline{X}_m^I\right) \\ &= \sum_{i\in I}(X_{il} - \mu_l)(X_{im} - \mu_m) - r\left(\overline{X}_l^I - \mu_l\right)\left(\overline{X}_m^I - \mu_m\right),\end{aligned}$$

de donde se deduce el resultado (3.7).

■

En la expresión (3.6), se observa que para el cálculo de las esperanzas condicionadas del producto de dos factores del tipo $S_{lm} - \sigma_{lm}$, será necesario conocer las esperanzas condicionadas de productos de factores del tipo $B_j(l,m)$.

Lema 3.3.9 *Sea* $\underline{X} = (X_1,...,X_p)' \sim \mathcal{N}_p(\underline{\mu}, \boldsymbol{\Sigma})$ *con* $\underline{\mu} = (\mu_1,...,\mu_p)'$ *y* $\boldsymbol{\Sigma} = (\sigma_{lm})$. *Sea* $\underline{X}_1,...,\underline{X}_n$ *una muestra aleatoria del vector* $\underline{X}$. *Sea* $I = \{i_1,...,i_r\} \subset \{1,...,n\}$, $\mathbf{X}_I = [\underline{X}_{i_1}...\underline{X}_{i_r}]$ *y* $\mathbf{x}_I \in \mathcal{M}_{p\times r}$. *Entonces,*

1. $$E[B_1(l,m) \mid \mathbf{X}_I = \mathbf{x}_I] = 0,$$
$$E[B_2(l,m) \mid \mathbf{X}_I = \mathbf{x}_I] = \frac{1}{n-1} b_2(l,m),$$
$$E[B_3(l,m) \mid \mathbf{X}_I = \mathbf{x}_I] = \frac{r(n-r)}{n(n-1)}\left[a(l,m) + \frac{1}{n-r}\sigma(l,m)\right].$$

2. $$E[B_1(l,m)B_1(u,v) \mid \mathbf{X}_I = \mathbf{x}_I] = \frac{n-r-1}{(n-1)^2} \sum_{j\in\{u,v\}} \sigma(l,j)\,\sigma(m,\overline{j}),$$
$$E[B_1(l,m)B_3(u,v) \mid \mathbf{X}_I = \mathbf{x}_I] = 0,$$
$$E[B_3(l,m)B_3(u,v) \mid \mathbf{X}_I = \mathbf{x}_I] = \frac{r^2}{n^2(n-1)^2}\left[(n-r)^2 a(l,m,u,v) + \right.$$
$$\left. + (n-r) \sum_{\underline{j}\in\ C_2(l,m,u,v)} a(\underline{j})\,\sigma(\overline{j}) + \sum_{j\in\{m,u,v\}} \sigma(l,j)\,\sigma(\overline{j})\right].$$

Demostración

1. Por ser $\mathbf{S}^{(I)}$ insesgado de $\boldsymbol{\Sigma}$, se tiene que

$$E[B_1(l,m) \mid \mathbf{X}_I = \mathbf{x}_I] = \frac{n-r-1}{n-1} E\left[S_{lm}^{(I)} - \sigma_{lm}\right] = 0.$$

Además,

$$E[B_2(l,m) \mid \mathbf{X}_I = \mathbf{x}_I] = E\left[\frac{1}{n-1} b_2(l,m) \mid \mathbf{X}_I = \mathbf{x}_I\right] = \frac{1}{n-1} b_2(l,m).$$

Y por otro lado, haciendo uso del resultado dado en el apartado **1** del lema 3.3.3,

$$E\left[B_3\left(l,m\right) \mid \mathbf{X}_I = \mathbf{x}_I\right] = \frac{r\left(n-r\right)}{n\left(n-1\right)} E\left[\left(\overline{x}_l^I - \overline{X}_l^{(I)}\right)\left(\overline{x}_m^I - \overline{X}_m^{(I)}\right)\right] =$$
$$= \frac{r\left(n-r\right)}{n\left(n-1\right)}\left[a\left(l,m\right) + \frac{1}{n-r}\sigma\left(l,m\right)\right].$$

2. Por el lema 3.3.4, se tiene que

$$E\left[B_1\left(l,m\right)B_1\left(u,v\right) \mid \mathbf{X}_I = \mathbf{x}_I\right] =$$
$$= \left(\frac{n-r-1}{n-1}\right)^2 E\left[\left(S_{lm}^{(I)} - \sigma_{lm}\right)\left(S_{uv}^{(I)} - \sigma_{uv}\right)\right] =$$
$$= \frac{n-r-1}{\left(n-1\right)^2}\sum_{j\in\{u,v\}} \sigma\left(l,j\right)\sigma\left(m,\overline{j}\right).$$

Como el vector de medias muestrales y la matriz de covarianzas muestrales son independientes,

$$E\left[B_1\left(l,m\right)B_3\left(u,v\right) \mid \mathbf{X}_I = \mathbf{x}_I\right] =$$
$$= \frac{r\left(n-r\right)\left(n-r-1\right)}{n\left(n-1\right)^2} E\left[\left(S_{lm}^{(I)} - \sigma_{lm}\right)\left(\overline{X}_u^I - \overline{X}_u^{(I)}\right)\left(\overline{X}_v^I - \overline{X}_v^{(I)}\right)\right] =$$
$$= \frac{r\left(n-r\right)\left(n-r-1\right)}{n\left(n-1\right)^2} E\left[S_{lm}^{(I)} - \sigma_{lm}\right] E\left[\left(\overline{X}_u^I - \overline{X}_u^{(I)}\right)\left(\overline{X}_v^I - \overline{X}_v^{(I)}\right)\right] =$$
$$= 0.$$

Por último, haciendo uso del resultado dado en el apartado **2** del lema 3.3.3,

$$E\left[B_3\left(l,m\right)B_3\left(u,v\right) \mid \mathbf{X}_I = \mathbf{x}_I\right] =$$
$$= \left[\frac{r\left(n-r\right)}{n\left(n-1\right)}\right]^2 E\left[\prod_{j\in\{l,m,u,v\}}\left(\overline{x}_j^I - \overline{X}_j^{(I)}\right)\right] =$$
$$= \frac{r^2\left(n-r\right)^2}{n^2\left(n-1\right)^2} a\left(l,m,u,v\right) + \frac{r^2\left(n-r\right)}{n^2\left(n-1\right)^2}\sum_{\underline{j}\in\ C_2(l,m,u,v)} a\left(\underline{j}\right)\sigma\left(\overline{\underline{j}}\right) +$$

$$+\frac{r^2}{n^2(n-1)^2}\sum_{j\in\{m,u,v\}}\sigma(l,j)\,\sigma\left(\overline{j}\right).$$

■

A continuación se proporciona el valor esperado condicionado a un conjunto de valores de $\underline{X}_i$ con $i \in I$, de dos factores del tipo $S_{lm} - \sigma_{lm}$.

Teorema 3.3.10 *Sea* $\underline{X} = (X_1, ..., X_p)' \sim \mathcal{N}_p(\underline{\mu}, \boldsymbol{\Sigma})$ *con* $\underline{\mu} = (\mu_1, ..., \mu_p)'$ *y* $\boldsymbol{\Sigma} = (\sigma_{lm})$. *Sea* $\underline{X}_1, ..., \underline{X}_n$ *una muestra aleatoria del vector* $\underline{X}$. *Sea* $I = \{i_1, ..., i_r\} \subset \{1, ..., n\}$, $\mathbf{X}_I = \left[\underline{X}_{i_1} ... \underline{X}_{i_r}\right]$ *y* $\mathbf{x}_I \in \mathcal{M}_{p\times r}$. *Entonces, para* $l, m, u, v = 1, ..., p$,

$$E\left[(S_{lm} - \sigma_{lm})(S_{uv} - \sigma_{uv}) \mid \mathbf{X}_I = \mathbf{x}_I\right] =$$

$$= \frac{1}{(n-1)^2}\left\{\frac{(n-r)(n^2-n-r)}{n^2}\sum_{j\in\{u,v\}}\sigma(l,j)\,\sigma\left(m,\overline{j}\right) + \right.$$

$$\left. + c(l,m)\cdot c(u,v) + \frac{r^2(n-r)}{n^2}\sum_{j_1\in\{l,m\}}\sum_{j_2\in\{u,v\}} a(j_1,j_2)\,\sigma\left(\overline{j}_1,\overline{j}_2\right)\right\}. \tag{3.8}$$

Demostración

Usando la expresión (3.6), $(S_{lm} - \sigma_{lm})(S_{uv} - \sigma_{uv})$ se puede descomponer en términos de índices en I y del resto de los índices (I),

$$(S_{lm} - \sigma_{lm})(S_{uv} - \sigma_{uv}) = \sum_{j_1=1}^{3} B_{j_1}(l,m)\sum_{j_2=1}^{3} B_{j_2}(u,v).$$

Teniendo en cuenta los distintos resultados obtenidos en el lema 3.3.9, se obtiene que

$$E\left[(S_{lm} - \sigma_{lm})(S_{uv} - \sigma_{uv}) \mid \mathbf{X}_I = \mathbf{x}_I\right] =$$

$$= \frac{n-r-1}{(n-1)^2}\sum_{j\in\{u,v\}}\sigma(l,j)\,\sigma\left(m,\overline{j}\right) + \frac{1}{(n-1)^2}b_2(l,m)\,b_2(u,v) +$$

$$+\frac{r(n-r)}{n(n-1)^2}\sum_{\underline{j}\in\{\{l,m\},\{u,v\}\}} b_2\left(\underline{j}\right)\left[a\left(\underline{j}\right) + \tfrac{1}{n-r}\sigma\left(\underline{j}\right)\right] +$$

$$+\frac{r^2}{n^2(n-1)^2}\left[(n-r)^2 a(l,m,u,v)+(n-r)\sum_{\underline{j}\in\ C_2(l,m,u,v)} a\left(\underline{j}\right)\sigma\left(\overline{j}\right)+\right.$$

$$\left.+\sum_{j\in\{m,u,v\}}\sigma(l,j)\,\sigma\left(\overline{j}\right)\right]=$$

$$=\frac{1}{(n-1)^2}\left\{(n-r-1)\sum_{j\in\{u,v\}}\sigma(l,j)\,\sigma\left(m,\overline{j}\right)\right.$$

$$+c(l,m)\cdot c(u,v)+\frac{r^2}{n^2}\sum_{j\in\{u,v\}}\sigma(l,j)\,\sigma\left(m,\overline{j}\right)+$$

$$\left.+\frac{r^2(n-r)}{n^2}\left[\sum_{j_1\in\{l,m\}}\sum_{j_2\in\{u,v\}} a\left(j_1,j_2\right)\sigma\left(\overline{j}_1,\overline{j}_2\right)\right]\right\}.$$

Y agrupando convenientemente, se obtiene la expresión (3.8).

■

A partir del teorema 3.3.10, se deduce el sesgo condicionado del producto de dos factores del tipo $S_{lm}-\sigma_{lm}$, el cual se presenta en el siguiente corolario:

Corolario 3.3.11 *Sea* $\underline{X}=(X_1,...,X_p)'\sim\mathcal{N}_p\left(\underline{\mu},\mathbf{\Sigma}\right)$ *con* $\underline{\mu}=(\mu_1,...,\mu_p)'$ *y* $\mathbf{\Sigma}=(\sigma_{lm})$. *Sea* $\underline{X}_1,...,\underline{X}_n$ *una muestra aleatoria del vector* $\underline{X}$. *Sea* $I=\{i_1,...,i_r\}\subset\{1,...,n\}$, $\mathbf{X}_I=\left[\underline{X}_{i_1}...\underline{X}_{i_r}\right]$ *y* $\mathbf{x}_I\in\mathcal{M}_{p\times r}$. *Entonces, para* $l,m,u,v=1,...,p$,

$$S\left(\mathbf{x}_I;\left(S_{lm}-\sigma_{lm}\right)\left(S_{uv}-\sigma_{uv}\right)\right)=$$

$$=\frac{1}{(n-1)^2}\left\{\frac{r\left(r-n^2\right)}{n^2}\sum_{j\in\{u,v\}}\sigma(l,j)\,\sigma\left(m,\overline{j}\right)+\right.$$

$$\left.+c(l,m)\cdot c(u,v)+\frac{r^2(n-r)}{n^2}\sum_{j_1\in\{l,m\}}\sum_{j_2\in\{u,v\}} a\left(j_1,j_2\right)\sigma\left(\overline{j}_1,\overline{j}_2\right)\right\}.$$

El sesgo condicionado obtenido en el corolario anterior, será útil para el cálculo aproximado del sesgo condicionado de autovalores y autovectores de la matriz de covarianzas muestrales, en el que se tendrán en cuenta únicamente los términos de orden n^{-2} y superior. Por ello, es conveniente expresar el sesgo condicionado de $\left(S_{lm}-\sigma_{lm}\right)\left(S_{uv}-\sigma_{uv}\right)$ en tales términos:

$$S\left(\mathbf{x}_I;\left(S_{lm}-\sigma_{lm}\right)\left(S_{uv}-\sigma_{uv}\right)\right)=$$

$$= \frac{1}{(n-1)^2}\left\{-r\sum_{j\in\{u,v\}}\sigma(l,j)\,\sigma\left(m,\overline{j}\right) + \right.$$

$$\left. + \left[b_2(l,m) + ra(l,m)\right]\left[b_2(u,v) + ra(u,v)\right]\right\} + O\left(n^{-3}\right)$$

Utilizando el lema 3.3.8, se tiene que

$$b_2(l,m) + ra(l,m) = -r\sigma_{lm} + (r-1)\,s^I_{lm} + r\left(\overline{x}^I_l - \mu_l\right)\left(\overline{x}^I_m - \mu_m\right) =$$
$$= -r\sigma_{lm} + \sum_{i\in I}\left(x_{il} - \mu_l\right)\left(x_{im} - \mu_m\right).$$

Y por ello, se puede expresar, para $l,m,u,v = 1,...,p$.
$S(\mathbf{x}_I; (S_{lm} - \sigma_{lm})(S_{uv} - \sigma_{uv})) =$

$$= \frac{1}{(n-1)^2}\left\{ - r\left(\sigma_{lu}\sigma_{mv} + \sigma_{lv}\sigma_{mu}\right) + \right.$$
$$+ \left[-r\sigma_{lm} + \sum_{i\in I}\left(x_{il} - \mu_l\right)\left(x_{im} - \mu_m\right)\right]\cdot$$
$$\left.\cdot\left[-r\sigma_{uv} + \sum_{i\in I}\left(x_{iu} - \mu_u\right)\left(x_{iv} - \mu_v\right)\right]\right\} + O\left(n^{-3}\right). \quad (3.9)$$

El siguiente paso a dar es calcular el valor esperado de términos de la forma $(S_{lm} - \sigma_{lm})(S_{uv} - \sigma_{uv})(S_{wz} - \sigma_{wz})$, condicionado a $\mathbf{X}_I = \mathbf{x}_I$. Para ello, previamente, es necesario obtener las esperanzas condicionadas del producto de tres factores del tipo $B_j(l,m)$. Como para conseguir el objetivo marcado, únicamente se utilizarán términos de orden n^{-2} o superiores, estas esperanzas condicionadas se expresarán según dichos órdenes.

Lema 3.3.12 *Sea* $\underline{X} = (X_1,...,X_p)' \sim \mathcal{N}_p\left(\underline{\mu}, \mathbf{\Sigma}\right)$ *con* $\underline{\mu} = (\mu_1,...,\mu_p)'$ *y* $\mathbf{\Sigma} = (\sigma_{lm})$. *Sea* $\underline{X}_1,...,\underline{X}_n$ *una muestra aleatoria del vector* $\underline{X}$, $I = \{i_1,...,i_r\} \subset \{1,...,n\}$, $\mathbf{X}_I = \left[\underline{X}_{i_1}...\underline{X}_{i_r}\right]$ *y* $\mathbf{x}_I \in \mathcal{M}_{p\times r}$. *Entonces,*

1. $E\left[B_1(l,m)\,B_1(u,v)\,B_1(w,z) \mid \mathbf{X}_I = \mathbf{x}_I\right] =$

$$= \frac{1}{(n-1)^2}\sum_{\underline{j}\in\{\{u,v\},\{w,z\}\}}\sum_{j_1\in\underline{j}}\sum_{j_2\in\overline{\underline{j}}}\sigma(l,j_1)\,\sigma(m,j_2)\,\sigma\left(\overline{j}_1,\overline{j}_2\right) + O\left(n^{-3}\right). \quad (3.10)$$

2. $E\left[B_1\left(l,m\right)B_1\left(u,v\right)B_3\left(w,z\right)\ \middle|\ \mathbf{X}_I=\mathbf{x}_I\right]=$

$$=\frac{r}{\left(n-1\right)^2}\sum_{j\in\{u,v\}}\sigma\left(l,j\right)\sigma\left(m,\overline{j}\right)a\left(w,z\right)+O\left(n^{-3}\right). \tag{3.11}$$

3. $E\left[B_1\left(l,m\right)B_3\left(u,v\right)B_3\left(w,z\right)\ \middle|\ \mathbf{X}_I=\mathbf{x}_I\right]=0.$

4. $E\left[B_3\left(l,m\right)B_3\left(u,v\right)B_3\left(w,z\right)\ \middle|\ \mathbf{X}_I=\mathbf{x}_I\right]$ *es un término de orden* n^{-3}.

Demostración

1. A partir del lema 3.3.4,

$$E\left[B_1\left(l,m\right)B_1\left(u,v\right)B_1\left(w,z\right)\ \middle|\ \mathbf{X}_I=\mathbf{x}_I\right]=$$
$$=\frac{\left(n-r-1\right)^3}{\left(n-1\right)^3}E\left[\left(S_{lm}^{(I)}-\sigma_{lm}\right)\left(S_{uv}^{(I)}-\sigma_{uv}\right)\left(S_{wz}^{(I)}-\sigma_{wz}\right)\right]=$$
$$=\frac{n-r-1}{\left(n-1\right)^3}\sum_{\underline{j}\in\{\{u,v\},\{w,z\}\}}\sum_{j_1\in\underline{j}}\sum_{j_2\in\overline{\underline{j}}}\sigma\left(l,j_1\right)\sigma\left(m,j_2\right)\sigma\left(\overline{j}_1,\overline{j}_2\right)+O\left(n^{-3}\right),$$

de lo que se tiene la expresión (3.10).

2. Debido a la independencia entre el vector de medias muestrales y la matriz de covarianzas muestrales, y utilizando los lemas 3.3.3 y 3.3.4,

$$E\left[B_1\left(l,m\right)B_1\left(u,v\right)B_3\left(w,z\right)\ \middle|\ \mathbf{X}_I=\mathbf{x}_I\right]=$$
$$=\frac{1}{\left(n-1\right)^3}E\left[\left(S_{lm}^{(I)}-\sigma_{lm}\right)\left(S_{uv}^{(I)}-\sigma_{uv}\right)\left(\overline{x}_w^I-\overline{X}_w^{(I)}\right)\left(\overline{x}_z^I-\overline{X}_z^{(I)}\right)\right]=$$
$$=\frac{1}{\left(n-1\right)^3}E\left[\left(S_{lm}^{(I)}-\sigma_{lm}\right)\left(S_{uv}^{(I)}-\sigma_{uv}\right)\right]E\left[\left(\overline{x}_w^I-\overline{X}_w^{(I)}\right)\left(\overline{x}_z^I-\overline{X}_z^{(I)}\right)\right]=$$
$$=\frac{r\left(n-r\right)\left(n-r-1\right)}{n\left(n-1\right)^3}\sum_{j\in\{u,v\}}\sigma\left(l,j\right)\sigma\left(m,\overline{j}\right)\left[a\left(w,z\right)+\frac{1}{n-r}\sigma\left(w,z\right)\right]+$$
$$+O\left(n^{-3}\right),$$

de donde se deduce la expresión (3.11).

3. Análogamente,

$$E\left[B_1\left(l,m\right)B_3\left(u,v\right)B_3\left(w,z\right)\ \middle|\ \mathbf{X}_I=\mathbf{x}_I\right]=$$

$$= \frac{1}{(n-1)^3} E\left[\left(S_{lm}^{(I)} - \sigma_{lm}\right) \prod_{j\in\{u,v,w,z\}} \left(\overline{x}_j^I - \overline{X}_j^{(I)}\right)\right] =$$

$$= \frac{1}{(n-1)^3} E\left[S_{lm}^{(I)} - \sigma_{lm}\right] E\left[\prod_{j\in\{u,v,w,z\}} \left(\overline{x}_j^I - \overline{X}_j^{(I)}\right)\right] = 0.$$

4. Por último,

$$E\left[B_3(l,m)\, B_3(u,v)\, B_3(w,z) \mid \mathbf{X}_I = \mathbf{x}_I\right] =$$

$$= \left[\frac{r(n-r)}{n(n-1)}\right]^3 E\left[\prod_{j\in\{l,m,.u,v,w,z\}} \left(\overline{x}_j^I - \overline{X}_j^{(I)}\right)\right] \equiv O\left(n^{-3}\right)$$

■

A partir del lema 3.3.12, se puede calcular la esperanza del producto de tres factores del tipo $S_{lm} - \sigma_{lm}$.

Teorema 3.3.13 *Sea* $\underline{X} = (X_1, ..., X_p)' \sim \mathcal{N}_p\left(\underline{\mu}, \mathbf{\Sigma}\right)$ *con* $\underline{\mu} = (\mu_1, ..., \mu_p)'$ *y* $\mathbf{\Sigma} = (\sigma_{lm})$. *Sea* $\underline{X}_1, ..., \underline{X}_n$ *una muestra aleatoria del vector* $\underline{X}$. *Sea* $I = \{i_1, ..., i_r\} \subset \{1, ..., n\}$, $\mathbf{X}_I = \left[\underline{X}_{i_1} ... \underline{X}_{i_r}\right]$ *y* $\mathbf{x}_I \in \mathcal{M}_{p\times r}$. *Entonces,para* $l, m, u, v, w, z = 1, ..., p$,

$$E\left[\left(S_{lm} - \sigma_{lm}\right)\left(S_{uv} - \sigma_{uv}\right)\left(S_{wz} - \sigma_{wz}\right) \mid \mathbf{X}_I = \mathbf{x}_I\right] =$$

$$= \frac{1}{(n-1)^2}\left\{\sum_{\underline{j}\in\{\{u,v\},\{w,z\}\}} \sum_{j_1\in\underline{j}} \sum_{j_2\in\overline{\underline{j}}} \sigma(l, j_1)\,\sigma(m, j_2)\,\sigma\left(\overline{j}_1, \overline{j_2}\right) + \right.$$

$$+ \sum_{\underline{j}\in\{\{l,m\},\{u,v\},\{w,z\}\}} \sum_{i\in I} \left[\prod_{j_1\in\underline{j}} \left(x_{ij_1} - \mu_{j_1}\right) - \sigma\left(\underline{j}\right)\right] \cdot$$

$$\left. \cdot \sum_{\underline{j}_2\in\overline{\underline{j}}/\underline{j}_2=\left\{j_3,\overline{j}_3\right\}} \sum_{j_4\in\overline{\underline{j}}_2} \sigma(j_3, j_4)\,\sigma\left(\overline{j}_3, \overline{j}_4\right)\right\} + O\left(n^{-3}\right). \qquad (3.12)$$

Demostración

A partir de la expresión (3.6) se tiene la descomposición
$\left(S_{lm} - \sigma_{lm}\right)\left(S_{uv} - \sigma_{uv}\right)\left(S_{wz} - \sigma_{wz}\right) =$

$$= \sum_{j_1=1}^{3} B_{j_1}(l,m) \sum_{j_2=1}^{3} B_{j_2}(u,v) \sum_{j_3=1}^{3} B_{j_3}(w,z).$$

Teniendo en cuenta los distintos resultados obtenidos en los lemas 3.3.9 y 3.3.12, se obtiene que

$$E\left[(S_{lm}-\sigma_{lm})(S_{uv}-\sigma_{uv})(S_{wz}-\sigma_{wz}) \mid \mathbf{X}_I = \mathbf{x}_I\right] =$$

$$= \frac{1}{(n-1)^2}\left\{ \sum_{\underline{j}\in\{(u,v),(w,z)\}} \sum_{j_1\in\underline{j}} \sum_{j_2\in\overline{\underline{j}}} \sigma(l,j_1)\,\sigma(m,j_2)\,\sigma\left(\overline{j}_1,\overline{j}_2\right) + \right.$$

$$+ \sum_{\underline{j}\in\{(l,m),(u,v),(w,z)\}} b_2\left(\underline{j}\right) \sum_{\underline{j}_2\in\overline{\underline{j}}/\underline{j}_2=\{j_3,\overline{j}_3\}} \sum_{j_4\in\overline{\underline{j}}_2} \sigma(j_3,j_4)\,\sigma\left(\overline{j}_3,\overline{j}_4\right) +$$

$$\left. + r \sum_{\underline{j}\in\{(l,m),(u,v),(w,z)\}} a\left(\underline{j}\right) \sum_{\underline{j}_2\in\overline{\underline{j}}/\underline{j}_2=\{j_3,\overline{j}_3\}} \sum_{j_4\in\overline{\underline{j}}_2} \sigma(j_3,j_4)\,\sigma\left(\overline{j}_3,\overline{j}_4\right) \right\} + O\left(n^{-3}\right).$$

Como consecuencia del lema 3.3.8, se tiene que

$$\begin{aligned} b_2(l,m) + r a(l,m) &= -r\sigma(l,m) + (r-1)\, s^I_{lm} + r a(l,m) = \\ &= -r\sigma(l,m) + \sum_{i\in I} (x_{il}-\mu_l)(x_{im}-\mu_m), \end{aligned}$$

y por tanto, se obtiene el resultado (3.12).

■

Del teorema 3.3.13 se deduce una aproximación del sesgo condicionado del producto de tres factores del tipo $S_{lm}-\sigma_{lm}$.

Corolario 3.3.14 *Sea* $\underline{X} = (X_1,...,X_p)' \sim \mathcal{N}_p\left(\underline{\mu},\mathbf{\Sigma}\right)$ *con* $\underline{\mu} = (\mu_1,...,\mu_p)'$ *y* $\mathbf{\Sigma} = (\sigma_{lm})$. *Sea* $\underline{X}_1,...,\underline{X}_n$ *una muestra aleatoria del vector* $\underline{X}$. *Sea* $I = \{i_1,...,i_r\} \subset \{1,...,n\}$, $\mathbf{X}_I = \left[\underline{X}_{i_1}...\underline{X}_{i_r}\right]$ *y* $\mathbf{x}_I \in \mathcal{M}_{p\times r}$. *Entonces, para* $l,m,u,v,w,z = 1,...,p$,

$$S\left(\mathbf{x}_I;(S_{lm}-\sigma_{lm})(S_{uv}-\sigma_{uv})(S_{wz}-\sigma_{wz})\right) =$$

$$= \frac{1}{(n-1)^2} \sum_{\underline{j}\in\{\{l,m\},\{u,v\},\{w,z\}\}} \sum_{i\in I} \left[\prod_{j_1\in\underline{j}} (x_{ij_1}-\mu_{j_1}) - \sigma\left(\underline{j}\right)\right] \cdot$$

$$\cdot \sum_{\underline{j}_2\in\overline{\underline{j}}/\underline{j}_2=\{j_3,\overline{j}_3\}} \sum_{j_4\in\overline{\underline{j}}_2} \sigma(j_3,j_4)\,\sigma\left(\overline{j}_3,\overline{j}_4\right) + O\left(n^{-3}\right)$$

La esperanza condicionada del producto cuatro factores del tipo $B_j(l,m)$, $j=1,2,3$, es de orden n^{-3}, excepto para $B_1(a,b)\,B_1(l,m)\,B_1(u,v)\,B_1(w,z)$, que salvo término del mismo orden, coincide con el valor esperado del producto de cuatro factores del tipo $S_{lm}-\sigma_{lm}$. Por ello, con este conjunto de resultados, se tienen todas las herramientas necesarias para calcular una aproximación del sesgo condicionado de un autovalor de $\mathbf{S}$, con el orden anteriormente justificado, ya que el sesgo condicionado del producto de cuatro o más factores del tipo $S_{lm}-\sigma_{lm}$, es de orden n^{-3} o inferior.

3.4 Sesgo condicionado de un autovalor muestral

En esta sección se obtiene una aproximación del sesgo condicionado de los autovalores de $\mathbf{S}$ y $\widehat{\boldsymbol{\Sigma}}$. Para ello, se utilizan dos propiedades conocidas de éstos: el desarrollo en serie recogido en el teorema 3.3.1 y su invarianza ante transformaciones ortogonales.

Basándose en la primera propiedad, se considera como aproximación de $S\left(\mathbf{x}_I;\widetilde{\lambda}_k\right)$, $\mathcal{S}\left(\mathbf{x}_I;\widetilde{\lambda}_k\right)=S\left(\mathbf{x}_I;\widetilde{\lambda}_k-R_k\right)$. Y partiendo de la segunda propiedad, la determinación de dicha aproximación se realizará a través del vector aleatorio de componentes principales $\underline{Y}=\mathbf{A}\left(\underline{X}-\underline{\mu}\right)$.

Teorema 3.4.1 *Sea* $\underline{X}\sim\mathcal{N}_p\left(\underline{\mu},\boldsymbol{\Sigma}\right)$, *con* $\boldsymbol{\Sigma}$ *tal que sus autovalores,* $\lambda_1,...,\lambda_p$, *sean simples y no nulos. Sea* $\underline{\alpha}_j$ *un autovector de* $\boldsymbol{\Sigma}$ *asociado a* λ_j, $j=1,...,p$. *Sea* $\underline{X}_1,...,\underline{X}_n$ *una muestra aleatoria del vector* $\underline{X}$ *y* $\widetilde{\lambda}_1\geq\widetilde{\lambda}_2\geq\cdots\geq\widetilde{\lambda}_p$ *los autovalores de* $\mathbf{S}$. *Entonces, para tamaño muestral suficientemente grande, para* $k=1,...,p$,

$$\begin{aligned}\mathcal{S}\left(\mathbf{x}_I;\widetilde{\lambda}_k\right) &= \frac{1}{n-1}\left[\sum_{i\in I}y_{ik}^2-r\lambda_k\right]+\frac{1}{(n-1)^2}\left[r\lambda_k-\left(\sum_{i\in I}y_{ik}\right)^2\right]+\\ &+\frac{1}{(n-1)^2}\sum_{j\neq k}\frac{\lambda_k\lambda_j}{(\lambda_j-\lambda_k)^2}\sum_{i\in I}\left(y_{ij}^2-y_{ik}^2\right)-\\ &-\frac{1}{(n-1)^2}\sum_{j\neq k}\frac{1}{\lambda_j-\lambda_k}\left[\sum_{i\in I}y_{ik}y_{ij}\right]^2+O\left(n^{-3}\right),\end{aligned}\tag{3.13}$$

donde $y_{ij}=\underline{\alpha}_j'\left(\underline{x}_i-\underline{\mu}\right)$, $j=1,...,p$ *y* $\underline{x}_i\in\mathbf{x}_I$.

Demostración

Dado $\underline{X}$, el vector aleatorio asociado de componentes principales es $\underline{Y} = \mathbf{A}\left(\underline{X} - \underline{\mu}\right) \sim \mathcal{N}_p\left(\underline{0}, \mathbf{\Lambda}\right)$, donde $\mathbf{\Lambda} = (\phi_{lm}) \quad / \quad \phi_{lm} = \lambda_l \delta_{lm}$.

Sean $\underline{Y}_i = (Y_{i1}, ..., Y_{ip})' = \mathbf{A}\left(\underline{X}_i - \underline{\mu}\right)$, $i = 1, ..., n$. La matriz de covarianzas muestrales asociada viene dada por $\mathbf{T} = \mathbf{ASA}'$, la cual tiene idénticos autovalores que $\mathbf{S}$, por ser una transformación ortogonal de ésta.

Sea $\mathbf{Y}_I$ el conjunto de los r vectores $\underline{Y}_i = \mathbf{A}\left(\underline{X}_i - \underline{\mu}\right)$, con $i \in I$ e $\mathbf{y}_I$ el conjunto de las observaciones $\underline{y}_i = (y_{ij})$, con $i \in I$.

$\mathcal{S}\left(\mathbf{x}_I; \widetilde{\lambda}_k\right)$ se puede calcular a partir de $\mathcal{S}\left(\mathbf{y}_I; \widetilde{\lambda}_k\right)$, ya que

$$\begin{aligned} \mathcal{S}\left(\mathbf{x}_I; \widetilde{\lambda}_k\right) &= E\left[\widetilde{\lambda}_k - R_k \mid \mathbf{X}_I = \mathbf{x}_I\right] - E\left[\widetilde{\lambda}_k - R_k\right] = \\ &= E\left[\widetilde{\lambda}_k - R_k \mid \mathbf{Y}_I = \mathbf{y}_I\right] - E\left[\widetilde{\lambda}_k - R_k\right] = \mathcal{S}\left(\mathbf{y}_I; \widetilde{\lambda}_k\right). \end{aligned}$$

A partir del desarrollo (3.2), denotando por $\phi_{lm} = \lambda_l \delta_{lm}$, se obtiene que

$$\begin{aligned} \mathcal{S}\left(\mathbf{y}_I; \widetilde{\lambda}_k\right) &= S\left(\mathbf{y}_I; T_{kk}\right) - \sum_{j \neq k} \frac{1}{\lambda_j - \lambda_k} S\left(\mathbf{y}_I; T_{kj}^2\right) - \\ &\quad - \sum_{j \neq k} \frac{1}{\left(\lambda_j - \lambda_k\right)^2} S\left(\mathbf{y}_I; \left(T_{kk} - \lambda_k\right) T_{kj}^2\right) + \\ &\quad + \sum_{j \neq k} \sum_{l \neq k} \frac{S\left(\mathbf{y}_I; T_{kj}\left(T_{jl} - \phi_{jl}\right) T_{lk}\right)}{\left(\lambda_j - \lambda_k\right)\left(\lambda_l - \lambda_k\right)}. \end{aligned} \tag{3.14}$$

A continuación se determina cada sumando de (3.14):

- Usando el teorema 3.3.7, se tiene que

$$S\left(\mathbf{y}_I; T_{kk}\right) = \frac{1}{n-1}\left[-\frac{r\left(n-1\right)}{n}\lambda_k + \sum_{i \in I} y_{ik}^2 - \frac{r^2}{n}\left(\overline{y}_k^I\right)^2\right],$$

- A partir de la expresión (3.9), para $j \neq k$,

$$S\left(\mathbf{y}_I; T_{kj}^2\right) = \frac{1}{\left(n-1\right)^2}\left\{-r\lambda_k\lambda_j + \left(\sum_{i \in I} y_{ik} y_{ij}\right)^2\right\} + O(n^{-3}).$$

- Considerando el resultado recogido en el corolario 3.3.14, para $j \neq k$,

$$S\left(\mathbf{y}_I; (T_{kk} - \lambda_k) T_{kj}^2\right) = \frac{1}{(n-1)^2} \lambda_k \lambda_j \sum_{i \in I} \left(y_{ik}^2 - \lambda_k\right) + O\left(n^{-3}\right),$$

y para $j \neq k$ y $l \neq k$, se tiene que

$$S\left(\mathbf{y}_I; T_{kj} (T_{jl} - \phi_{jl}) T_{lk}\right) = \frac{\delta_{jl}}{(n-1)^2} \lambda_k \lambda_j \sum_{i \in I} \left(y_{ij}^2 - \lambda_j\right) + O\left(n^{-3}\right).$$

Sustituyendo en (3.14) las expresiones obtenidas, se tiene la igualdad (3.13).

■

Una expresión alternativa para $\mathcal{S}\left(\mathbf{x}_I; \widetilde{\lambda}_k\right)$ es la que se muestra a continuación, la cual se obtiene a partir de la expresión (3.13), teniendo en cuenta que $y_{ij}^2 = I\left(\underline{x}_i; \lambda_j, F\right) + \lambda_j$.

Corolario 3.4.2 *En las condiciones del teorema 3.4.1, si F es la función de distribución del vector $\underline{X}$, se verifica que*

$$\begin{aligned}
\mathcal{S}\left(\mathbf{x}_I; \widetilde{\lambda}_k\right) = {} & \frac{1}{n-1} I\left(\mathbf{x}_I; \lambda_k, F\right) + \frac{1}{(n-1)^2} \left[\left[r\lambda_k - \left(\sum_{i \in I} y_{ik} \right)^2 \right] - \right. \\
& - \sum_{j \neq k} \frac{1}{\lambda_j - \lambda_k} \left[\left(\sum_{i \in I} y_{ik} y_{ij} \right)^2 - \lambda_k \lambda_j \right] + \\
& \left. + \sum_{j \neq k} \frac{\lambda_k \lambda_j}{(\lambda_j - \lambda_k)^2} \sum_{i \in I} \left[I\left(\underline{x}_i; \lambda_j, F\right) - I\left(\underline{x}_i; \lambda_k, F\right) \right] \right] + O\left(n^{-3}\right).
\end{aligned}$$

El resultado del teorema 3.4.1 se puede particularizar al caso simple, $I = \{i\}$.

Corolario 3.4.3 *En las condiciones del teorema 3.4.1*

$$\mathcal{S}\left(\underline{x}_i;\widetilde{\lambda}_k\right) = \frac{1}{n-1}\left(y_{ik}^2-\lambda_k\right) - \frac{1}{(n-1)^2}\left\{\left(y_{ik}^2-\lambda_k\right) + \right.$$
$$\left. + \sum_{j\neq k}\left[\frac{\lambda_k\lambda_j}{\left(\lambda_j-\lambda_k\right)^2}\left(y_{ij}^2-y_{ik}^2\right) - \frac{1}{\lambda_j-\lambda_k}y_{ik}^2y_{ij}^2\right]\right\} + O\left(n^{-3}\right),$$

donde $y_{ij} = \underline{\alpha}_j'\left(\underline{x}_i - \underline{\mu}\right)$.

Una expresión alternativa a la anterior, la cual se utilizará posteriormente es

$$\mathcal{S}\left(\underline{x}_i;\widetilde{\lambda}_k\right) = \frac{1}{n}I\left(\underline{x}_i;\lambda_k,F\right) + \frac{1}{(n-1)^2}\sum_{j\neq k}\frac{1}{\left(\lambda_j-\lambda_k\right)^2}\cdot \tag{3.15}$$
$$\cdot\left[\lambda_k y_{ik}^2 I\left(\underline{x}_i;\lambda_j,F\right) - \lambda_j y_{ij}^2 I\left(\underline{x}_i;\lambda_k,F\right)\right] + O\left(n^{-3}\right).$$

Finalmente, debido a la relación existente entre los autovalores de $\widehat{\mathbf{\Sigma}}$ y $\mathbf{S}$, por lo que se puede considerar como aproximación de $S\left(\mathbf{x}_I;\widehat{\lambda}_k\right)$

$$\mathcal{S}\left(\mathbf{x}_I;\widehat{\lambda}_k\right) = \frac{n-1}{n}\mathcal{S}\left(\mathbf{x}_I;\widetilde{\lambda}_k\right).$$

Una expresión de dicha aproximación se recoge en el siguiente teorema.

Teorema 3.4.4 *Sea* $\underline{X} \sim \mathcal{N}_p\left(\underline{\mu},\mathbf{\Sigma}\right)$, *con* $\mathbf{\Sigma}$ *tal que sus autovalores,* $\lambda_1,...,\lambda_p$, *sean simples y no nulos. Sea* $\underline{\alpha}_j$ *un autovector de* $\mathbf{\Sigma}$ *asociado a* λ_j, $j=1,...,p$. *Sea* $\underline{X}_1,...,\underline{X}_n$ *una muestra aleatoria del vector* $\underline{X}$ *y* $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \cdots \geq \widehat{\lambda}_p$ *los autovalores de* $\widehat{\mathbf{\Sigma}}$. *Entonces, para tamaño muestral suficientemente grande, para* $k=1,...,p$,

$$\mathcal{S}\left(\mathbf{x}_I;\widehat{\lambda}_k\right) = \frac{1}{n}\left[\sum_{i\in I} y_{ik}^2 - r\lambda_k\right] + \frac{1}{n^2}\left\{\left[r\lambda_k - \left(\sum_{i\in I} y_{ik}\right)^2\right] - \right.$$
$$\left. - \sum_{j\neq k}\frac{1}{\lambda_j-\lambda_k}\left[\sum_{i\in I} y_{ik}y_{ij}\right]^2 + \sum_{j\neq k}\frac{\lambda_k\lambda_j}{(\lambda_j-\lambda_k)^2}\sum_{i\in I}\left(y_{ij}^2 - y_{ik}^2\right)\right\} +$$
$$+O\left(n^{-3}\right),$$

donde $y_{ij} = \underline{\alpha}_j'\left(\underline{x}_i - \underline{\mu}\right)$.

Otras expresiones para $\mathcal{S}\left(\mathbf{x}_I;\widehat{\lambda}_k\right)$ y $\mathcal{S}\left(\underline{x}_i;\widehat{\lambda}_k\right)$ se obtienen de forma análoga, utilizando las expresiones alternativas de $\mathcal{S}\left(\mathbf{x}_I;\widetilde{\lambda}_k\right)$ y $\mathcal{S}\left(\underline{x}_i;\widetilde{\lambda}_k\right)$.

Como consecuencia de los resultados anteriores, se pueden realizar los siguientes comentarios e interpretaciones:

Nota 3.4.1 1. *Para tamaño muestral suficientemente grande, el término dominante en* $\mathcal{S}\left(\mathbf{x}_I;\widetilde{\lambda}_k\right)$ *es*

$$\frac{1}{n-1}\sum_{i\in I}\left(y_{ik}^2 - \lambda_k\right) = \frac{r}{n-1}\left(\sum_{i\in I}\frac{y_{ik}^2}{r} - \lambda_k\right).$$

Teniendo en cuenta que $Y_k = \underline{\alpha}_k'\left(\underline{X} - \underline{\mu}\right)$, *y por lo tanto,* $E\left[Y_k\right] = 0$ *y* $\lambda_k = var\left(Y_k\right) = E\left[Y_k^2\right]$, *entonces, el momento muestral de orden dos, se puede considerar como un estimador de la varianza.*

Así, $r^{-1}\sum_{i\in I} y_{ik}^2$, *constituye una estimación de* λ_k *a partir de las observaciones de índices en* I. *Por ello, el término de orden* n^{-1} *de la aproximación del sesgo condicionado de* $\widetilde{\lambda}_k$, *compara la variación de las puntuaciones de la* k*-ésima componente principal de las observaciones consideradas, con la variación de la* k*-ésima componente principal poblacional.*

Luego, si la dispersión que presenta un grupo de observaciones varía notablemente respecto a la poblacional, dicho conjunto proporcionará un valor elevado de $\mathcal{S}\left(\mathbf{x}_I;\lambda_k\right)$ *y se podrá interpretar como un conjunto de*

observaciones altamente influyente sobre el estimador de la ***varianza*** *de la k-ésima componente principal,* λ_k.

Los términos de orden n^{-2} *están constituidos por sumandos en los que intervienen factores del tipo* $(\lambda_j - \lambda_k)^{-1}$, *los cuales reflejan pueden tener un valor considerablemente alto ante la existencia de autovalores de la matriz de covarianzas poblacional muy próximos.*

2. *En el caso en el que el Análisis de Influencia se lleve a cabo de forma individual para cada observación, el término de orden* n^{-1}, *se puede escribir como*

$$\frac{1}{n-1}\left(y_{ik}^2 - \lambda_k\right) = \frac{1}{n-1}\left(y_{ik}^2 - E\left[Y_k^2\right]\right),$$

que representa la desviación entre el cuadrado del valor de la k-ésima componente principal de la observación $\underline{x}_i$ *y su valor esperado.*

Cuando dicho término es pequeño, pueden detectarse valores fuertemente influyentes sobre el autovalor bajo estudio, a través del término de orden n^{-2}.

Para interpretar esta situación se puede suponer, en el caso de análisis de influencia individual, que $y_{ik}^2 \simeq \lambda_k$. *Así, según se deduce de (3.15), se verifica que*

$$S\left(\underline{x}_i; \widetilde{\lambda}_k\right) \simeq \frac{1}{(n-1)^2}\lambda_k^2 \sum_{j \neq k} \frac{y_{ij}^2 - \lambda_j}{(\lambda_j - \lambda_k)^2} \tag{3.16}$$

La expresión (3.16) representa la suma ponderada de las desviaciones de los cuadrados de los valores de cada componente principal para la observación $\underline{x}_i$, *con respecto a sus valores esperados, con pesos*

$$\frac{\lambda_k^2}{(n-1)^2(\lambda_j - \lambda_k)^2},$$

es decir, los sumandos de más peso son aquellos que corresponden a autovalores más próximos a λ_k. *En este sumatorio pueden existir compensaciones ya que* $y_{ij}^2 - \lambda_j$ *puede tomar tanto valores positivos como negativos. Pero, sí se verifica que*

$$\left| \lambda_k^2 \sum_{j \neq k} \frac{y_{ij}^2 - \lambda_j}{(\lambda_j - \lambda_k)^2} \right| \leq \lambda_k^2 \sum_{j \neq k} \frac{|y_{ij}^2 - \lambda_j|}{(\lambda_j - \lambda_k)^2}.$$

Si para autovalores próximos a λ_k, el valor de la componente asociada a $\underline{x}_i$ es muy próxima, en valor absoluto, a su desviación típica, y para autovalores lejanos a λ_k, el valor de la componente principal no es muy elevado, en valor absoluto, respecto a su desviación típica, el valor del sesgo condicionado de $\widetilde{\lambda}_k$ será pequeño.

Por otro lado, si para una observación, los valores de las componentes principales son similares en valor absoluto a las desviaciones típicas de las mismas, excepto una de ellas y_{ij_0}, que se separa marcadamente, además de detectarse un valor alto en $S\left(\underline{x}_i; \widetilde{\lambda}_{j_0}\right)$, en el sesgo condicionado del resto de los autovalores, se detectarán valores más altos, cuanto más próximo sea el autovalor a λ_{j_0}.

3. *El sesgo condicionado de un autovalor no está acotado, por lo que pueden existir observaciones altamente influyentes.*

Capítulo 4

SESGO CONDICIONADO DE UN AUTOVECTOR

4.1 Introducción

El objetivo de este capítulo es el cálculo bajo hipótesis de normalidad de una aproximación del sesgo condicionado de un autovector de la matriz de covarianzas muestrales asociado a un autovalor simple. Para un autovalor simple, en la bibliografía se encuentra un desarrollo que se basa en la diagonalización de la matriz de covarianzas, el cual se ha recogido en el teorema 3.3.1. Siguiendo esta idea, en el presente capítulo se muestra un desarrollo en serie de un autovector de la matriz de covarianzas muestrales a partir del caso diagonal, del cual se obtienen los primeros coeficientes.

En la sección 4.2 de este capítulo, se obtienen los primeros coeficientes del desarrollo en serie de un autovector de la matriz de covarianzas muestrales en función de parámetros poblacionales y de los elementos de dicha matriz, para posteriormente, en la siguiente sección, calcular una aproximación de su sesgo condicionado. Aunque el valor esperado no es objetivo de este trabajo, también se halla, ya que se utilizará en el capítulo siguiente para justificar la estimación realizada del sesgo condicionado de un autovector.

4.2 Desarrollo en serie de un autovector

El desarrollo en serie (3.2) de un autovalor simple de $\mathbf{S}$ se basa en la propiedad de invarianza de los autovalores ante transformaciones ortogonales. La transformación ortogonal utilizada es

$$\mathbf{T} = \mathbf{ASA}'. \tag{4.1}$$

La relación existente entre un autovector de $\mathbf{S}$, $\underset{\sim}{\widetilde{\alpha}}_k$, y un autovector de $\mathbf{T}$, $\underset{\sim}{\widetilde{\beta}}_k$, asociados a un mismo autovalor simple, $\widetilde{\lambda}_k$, es $\mathbf{A}'\underset{\sim}{\widetilde{\beta}}_k = \underset{\sim}{\widetilde{\alpha}}_k$.

De nuevo, para la obtención del desarrollo en serie de $\underset{\sim}{\widetilde{\alpha}}_k$, se hará a partir de la transformación ortogonal (4.1).

Es conocido que (Girshick [21]), fijada una matriz simétrica definida positiva, $\mathbf{\Sigma_0}$, existe un entorno de dicha matriz, $\mathcal{N}(\mathbf{\Sigma_0}) \subset \mathcal{D}_p$, con

$$\mathcal{D}_p = \{\mathbf{\Sigma} \in \mathcal{M}_{p\times p} \quad / \quad \mathbf{\Sigma} \text{ simétrica definida positiva}\},$$

donde las funciones

$$\lambda_k^{\mathbf{\Sigma_0}} : \mathcal{N}(\mathbf{\Sigma_0}) \longrightarrow \mathbb{R}^+; \qquad \underset{\sim}{\alpha}_k^{\mathbf{\Sigma_0}} : \mathcal{N}(\mathbf{\Sigma_0}) \longrightarrow \mathbb{R}^p, \qquad k = 1, ..., p,$$

tales que si $\mathbf{\Sigma} \in \mathcal{N}(\mathbf{\Sigma_0})$, $\lambda_1^{\mathbf{\Sigma_0}}(\mathbf{\Sigma}) > \lambda_2^{\mathbf{\Sigma_0}}(\mathbf{\Sigma}) > \cdots > \lambda_p^{\mathbf{\Sigma_0}}(\mathbf{\Sigma}) > 0$, son los autovalores de $\mathbf{\Sigma}$ y $\underset{\sim}{\alpha}_k^{\mathbf{\Sigma_0}}(\mathbf{\Sigma})$ es un autovector de $\mathbf{\Sigma}$ asociado a $\lambda_k^{\mathbf{\Sigma_0}}(\mathbf{\Sigma})$ y además dichas funciones son analíticas en $\mathcal{N}(\mathbf{\Sigma_0})$.

Para evitar posibles ambigüedades en el sentido del autovector, se puede fijar algún criterio como aquél en el que la componente de mayor valor absoluto de $\underset{\sim}{\alpha}_k^{\mathbf{\Sigma_0}}(\mathbf{\Sigma_0})$ tenga signo positivo. Al ser $\underset{\sim}{\alpha}_k^{\mathbf{\Sigma_0}}$ analítica en $\mathcal{N}(\mathbf{\Sigma_0})$, si es necesario, dicho entorno se puede reducir de forma que se verifique el criterio anterior para toda matriz perteneciente a él.

Para obtener el desarrollo en serie de $\underset{\sim}{\widetilde{\alpha}}_k$, a partir de la transformación ortogonal (4.1), se tomará como matriz $\mathbf{\Sigma_0}$ la matriz de autovalores de la matriz de covarianzas del vector aleatorio bajo estudio, $\mathbf{\Lambda}$.

Debido a la convergencia *casi seguro* de $\mathbf{S}$ a $\mathbf{\Sigma_0}$, se tiene directamente el mismo tipo de convergencia de $\mathbf{T}$ a $\mathbf{\Lambda}$. Entonces, para tamaño muestral suficientemente grande, para que $\mathbf{T}$ pertenezca con probabilidad uno al entorno $\mathcal{N}(\mathbf{\Lambda})$, en el cual $\underset{\sim}{\alpha}_k^{\mathbf{\Lambda}}$ es analítica, es válido el desarrollo en serie

$$\begin{aligned}
\underset{\sim}{\widetilde{\beta}}_k = {} & \underset{\sim}{\alpha}_k^{\mathbf{\Lambda}}(\mathbf{\Lambda}) + \sum_l \sum_{m\geq l} \left.\frac{\partial \underset{\sim}{\alpha}_k^{\mathbf{\Lambda}}(\mathbf{\Sigma})}{\partial \sigma_{lm}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}} (T_{lm} - \delta_{lm}\lambda_l) + \\
& + \frac{1}{2} \sum_l \sum_{m\geq l} \sum_u \sum_{v\geq u} \left.\frac{\partial^2 \underset{\sim}{\alpha}_k^{\mathbf{\Lambda}}(\mathbf{\Sigma})}{\partial \sigma_{lm}\partial \sigma_{uv}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}} (T_{lm} - \delta_{lm}\lambda_l)(T_{uv} - \delta_{uv}\lambda_u) + \\
& + \frac{1}{3!} \sum_l \sum_{m\geq l} \sum_u \sum_{v\geq u} \sum_w \sum_{z\geq w} \left.\frac{\partial^3 \underset{\sim}{\alpha}_k^{\mathbf{\Lambda}}(\mathbf{\Sigma})}{\partial \sigma_{lm}\partial \sigma_{uv}\partial \sigma_{wz}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}} \cdot \\
& \cdot (T_{lm} - \delta_{lm}\lambda_l)(T_{uv} - \delta_{uv}\lambda_u)(T_{wz} - \delta_{wz}\lambda_w) + R_k(\mathbf{T}), \qquad (4.2)
\end{aligned}$$

donde $\mathbf{T} = (T_{lm})$ y $R_k(\mathbf{T})$ incluye todos los términos formados por 7productos de cuatro o más factores del tipo $T_{lm} - \delta_{lm}\lambda_l$.

Este desarrollo se va a utilizar como base del cálculo del sesgo condicionado sobre un autovector muestral. Es evidente la necesidad de obtener las derivadas sucesivas del autovector respecto a las componentes de la matriz. A partir de los siguientes lemas, recogidos por Harville [27], se obtiene la expresión de la primera derivada de un autovector y las herramientas necesarias para el cálculo de las derivadas de orden superior.

Teorema 4.2.1 *Sea* $\mathbf{P}$ *una matriz* $p \times p$, *simétrica, de funciones definidas sobre un conjunto* $\mathcal{H} \subset \mathbb{R}^m$ *tal que* $\mathbf{P}(\underline{t}) = (p_{ij}(\underline{t}))$ *con* $p_{ij} : \mathcal{H} \longrightarrow \mathbb{R}$. *Sea* $\underline{t}_0$ *un punto interior de* $\mathcal{H}$ *para el cual* $\mathbf{P}$ *es continuamente diferenciable. Sea* λ^* *un autovalor simple de* $\mathbf{P}(\underline{t}_0)$ *y* $\underline{\alpha}^*$ *un autovector unitario asociado a* λ^*. *Entonces, para algún entorno* $\mathcal{N}(\underline{t}_0)$ *de* $\underline{t}_0$, *existe una función real* $\lambda : \mathcal{N}(\underline{t}_0) \longrightarrow \mathbb{R}$, *y una función vectorial* $\underline{\alpha} : \mathcal{N}(\underline{t}_0) \longrightarrow \mathbb{R}^p$ *tales que*

1. $\lambda(\underline{t}_0) = \lambda^*$ *y* $\underline{\alpha}(\underline{t}_0) = \underline{\alpha}^*$.
2. *Para todo* $\underline{t} \in \mathcal{N}(\underline{t}_0)$, $\lambda(\underline{t})$ *es un autovalor de* $\mathbf{P}(\underline{t})$ *y* $\underline{\alpha}(\underline{t})$ *un autovector unitario asociado al autovalor* $\lambda(\underline{t})$ *de forma que los signos de los elementos no nulos de* $\underline{\alpha}(\underline{t})$ *son los mismos que los de* $\underline{\alpha}^*$.
3. λ *y* $\underline{\alpha}$ *son continuamente diferenciables en* $\underline{t}_0$.

A partir del teorema 4.2.1 se obtiene que $\forall \underline{t} \in \mathcal{N}(\underline{t}_0)$, $\mathbf{P}(\underline{t})\underline{\alpha}(\underline{t}) = \lambda(\underline{t})\underline{\alpha}(\underline{t})$, y diferenciando ambos miembros de la igualdad $\forall \underline{t} \in \mathcal{N}(\underline{t}_0)$, $\forall j = 1, ..., m$

$$\mathbf{P}(\underline{t})\left[\frac{\partial}{\partial t_j}\underline{\alpha}(\underline{t})\right] + \left[\frac{\partial}{\partial t_j}\mathbf{P}(\underline{t})\right]\underline{\alpha}(\underline{t}) = \lambda(\underline{t})\left[\frac{\partial}{\partial t_j}\underline{\alpha}(\underline{t})\right] + \left[\frac{\partial}{\partial t_j}\lambda(\underline{t})\right]\underline{\alpha}(\underline{t}). \tag{4.3}$$

De la expresión (4.3), se obtiene, a través de propiedades básicas de matrices y diferenciación de éstas, las siguientes igualdades (Harville [27]):

$\forall\, \underline{t} \in \mathcal{N}(\underline{t}_0)$ y $\forall\, j = 1, ..., m$

$$\frac{\partial}{\partial t_j}\lambda(\underline{t}) = [\underline{\alpha}(\underline{t})]' \cdot \left[\frac{\partial}{\partial t_j}\mathbf{P}(\underline{t})\right] \cdot \underline{\alpha}(\underline{t}), \tag{4.4}$$

$$\frac{\partial}{\partial t_j}\underline{\alpha}(\underline{t}) = -\mathbf{H}^+(\underline{t}) \cdot \left[\frac{\partial}{\partial t_j}\mathbf{P}(\underline{t})\right] \cdot \underline{\alpha}(\underline{t}), \tag{4.5}$$

donde $\mathbf{H}(\underline{t}) = \mathbf{P}(\underline{t}) - \lambda(\underline{t})\mathbf{I}_p$ y $\mathbf{H}^+(\underline{t})$ es la inversa de Moore-Penrose de $\mathbf{H}(\underline{t})$.

Las igualdades recogidas en (4.4) y (4.5) se pueden aplicar al caso especial donde $\underline{t}$ sea un vector de dimensión $p(p+1)/2$, cuyas componentes

constituyan los elementos de una matriz simétrica definida positiva. Por ello, basta considerar la matriz de funciones $\boldsymbol{\Sigma}(\underline{\sigma})$ definida en

$$\mathcal{H}^* = \Big\{\underline{\sigma} = (\sigma_{11}, \sigma_{12}, ..., \sigma_{1p}, \sigma_{22}, ..., \sigma_{2p}, ..., \sigma_{p-1,p}, \sigma_{pp}) \in \mathbb{R}^{\frac{p(p+1)}{2}} / \\ / \quad (\sigma_{ij}) \in \mathcal{D}_p \text{ con } \sigma_{ij} = \sigma_{ji}\Big\},$$

de forma que la componente (i,j) de $\boldsymbol{\Sigma}(\underline{\sigma})$ viene determinada por la función

$$\begin{array}{ccc} \mathcal{H}^* & \longrightarrow & \mathbb{R} \\ \underline{\sigma} & \longmapsto & \begin{cases} \sigma_{ij} & si \quad i \leq j \\ \sigma_{ji} & si \quad i > j. \end{cases} \end{array}$$

Así pues, dado que $\mathcal{H}^*$ es un conjunto abierto de $\mathbb{R}^{\frac{p(p+1)}{2}}$, todos sus puntos son interiores, y como consecuencia, a partir de (4.4) y (4.5) se puede enunciar el siguiente teorema; para las funciones $\lambda_k^{\boldsymbol{\Sigma}_0}$ y $\underline{\alpha}_k^{\boldsymbol{\Sigma}_0}$, para cierto $\boldsymbol{\Sigma}_0 \in \mathcal{D}_p$, las cuales verifican las propiedades enunciadas en el teorema 4.2.1.

Teorema 4.2.2 *Sea $\boldsymbol{\Sigma}_0 \in \mathcal{D}_p$ y $\lambda_k^{\boldsymbol{\Sigma}_0}(\boldsymbol{\Sigma}_0)$ un autovalor simple de $\boldsymbol{\Sigma}_0$. Entonces, existe un entorno de $\boldsymbol{\Sigma}_0$, $\mathcal{N}(\boldsymbol{\Sigma}_0)$, tal que $\forall \boldsymbol{\Sigma} \in \mathcal{N}(\boldsymbol{\Sigma}_0)$,*

1. $\dfrac{\partial}{\partial \sigma_{lm}} \lambda_k^{\boldsymbol{\Sigma}_0}(\boldsymbol{\Sigma}) = (2 - \delta_{lm})\, \alpha_{kl}^{\boldsymbol{\Sigma}_0}(\boldsymbol{\Sigma})\, \alpha_{km}^{\boldsymbol{\Sigma}_0}(\boldsymbol{\Sigma}),$

2. $\dfrac{\partial}{\partial \sigma_{lm}} \underline{\alpha}_k^{\boldsymbol{\Sigma}_0}(\boldsymbol{\Sigma}) = -\Big[\alpha_{km}^{\boldsymbol{\Sigma}_0}(\boldsymbol{\Sigma})\, \underline{g}_l^k(\boldsymbol{\Sigma}) + (1 - \delta_{lm})\, \alpha_{kl}^{\boldsymbol{\Sigma}_0}(\boldsymbol{\Sigma})\, \underline{g}_m^k(\boldsymbol{\Sigma})\Big],$

3. $\dfrac{\partial}{\partial \sigma_{lm}} \alpha_{ki}^{\boldsymbol{\Sigma}_0}(\boldsymbol{\Sigma}) = -\left[\alpha_{km}^{\boldsymbol{\Sigma}_0}(\boldsymbol{\Sigma})\, g_{il}^k(\boldsymbol{\Sigma}) + (1 - \delta_{lm})\, \alpha_{kl}^{\boldsymbol{\Sigma}_0}(\boldsymbol{\Sigma})\, g_{im}^k(\boldsymbol{\Sigma})\right],$

donde $\alpha_{ki}^{\boldsymbol{\Sigma}_0}(\boldsymbol{\Sigma})$ es la i-ésima componente del vector $\underline{\alpha}_k^{\boldsymbol{\Sigma}_0}(\boldsymbol{\Sigma})$, $\underline{g}_l^k(\boldsymbol{\Sigma})$ es la l-ésima columna de $\left[\boldsymbol{\Sigma} - \lambda_k^{\boldsymbol{\Sigma}_0}(\boldsymbol{\Sigma})\, \mathbf{I}_p\right]^+$, cuyo $i-$ésimo elemento se denota por $g_{il}^k(\boldsymbol{\Sigma})$.

Otros resultados necesarios para posteriores desarrollos giran en torno a la diferenciación de la matriz de Moore-Penrose, también recogido por Harville [27], los cuales serán aplicados a la matriz $\mathbf{F}_k(\boldsymbol{\Sigma}) = \boldsymbol{\Sigma} - \lambda_k^{\boldsymbol{\Sigma}_0}(\boldsymbol{\Sigma})\, \mathbf{I}$.

Lema 4.2.3 *Sea $\mathbf{F} = (f_{ij})$ una matriz $p \times p$ de funciones definidas sobre el conjunto $\mathcal{H}^* \subset \mathbb{R}^{\frac{p(p+1)}{2}}$, $f_{ij} : \mathcal{H}^* \longrightarrow \mathbb{R}$. Sea $\underline{t}_0$ un punto interior de $\mathcal{H}^*$, en el cual $\mathbf{F}$ es continuamente diferenciable y tal que para un entorno de $\underline{t}_0$, el rango de $\mathbf{F}$ es constante. Entonces,*

1. $$\frac{\partial}{\partial t_j}\mathbf{F}^{+}(\underset{\sim}{t})\Big|_{\underset{\sim}{t}=\underset{\sim}{t}_0} = -\mathbf{F}^{+}(\underset{\sim}{t}_0)\left[\frac{\partial}{\partial t_j}\mathbf{F}(\underset{\sim}{t})\Big|_{\underset{\sim}{t}=\underset{\sim}{t}_0}\right]\mathbf{F}^{+}(\underset{\sim}{t}_0) +$$
$$+\mathbf{F}^{+}(\underset{\sim}{t}_0)\left(\mathbf{F}^{+}(\underset{\sim}{t}_0)\right)'\left[\frac{\partial}{\partial t_j}\mathbf{F}(\underset{\sim}{t})\Big|_{\underset{\sim}{t}=\underset{\sim}{t}_0}\right]'\left(\mathbf{I}_p - \mathbf{F}(\underset{\sim}{t}_0)\mathbf{F}^{+}(\underset{\sim}{t}_0)\right) +$$
$$+\left(\mathbf{I}_p - \mathbf{F}^{+}(\underset{\sim}{t}_0)\mathbf{F}(\underset{\sim}{t}_0)\right)\left[\frac{\partial}{\partial t_j}\mathbf{F}(\underset{\sim}{t})\Big|_{\underset{\sim}{t}=\underset{\sim}{t}_0}\right]'\left(\mathbf{F}^{+}(\underset{\sim}{t}_0)\right)'\mathbf{F}^{+}(\underset{\sim}{t}_0).$$

2. *Además,* $\mathbf{F}^{+}\mathbf{F}$ *y* $\mathbf{F}\mathbf{F}^{+}$ *son continuamente diferenciables en* $\underset{\sim}{t}_0$ *y*

(a) $$\frac{\partial\left[\mathbf{F}^{+}(\underset{\sim}{t})\mathbf{F}(\underset{\sim}{t})\right]}{\partial t_j}\Big|_{\underset{\sim}{t}=\underset{\sim}{t}_0} = \mathbf{F}^{+}(\underset{\sim}{t}_0)\left[\frac{\partial}{\partial t_j}\mathbf{F}(\underset{\sim}{t})\Big|_{\underset{\sim}{t}=\underset{\sim}{t}_0}\right]\left(\mathbf{I}_p - \mathbf{F}^{+}(\underset{\sim}{t}_0)\mathbf{F}(\underset{\sim}{t}_0)\right) +$$
$$+\left[\mathbf{F}^{+}(\underset{\sim}{t}_0)\left[\frac{\partial}{\partial t_j}\mathbf{F}(\underset{\sim}{t})\Big|_{\underset{\sim}{t}=\underset{\sim}{t}_0}\right]\left(\mathbf{I}_p - \mathbf{F}^{+}(\underset{\sim}{t}_0)\mathbf{F}(\underset{\sim}{t}_0)\right)\right]',$$

(b) $$\frac{\partial\left[\mathbf{F}(\underset{\sim}{t})\mathbf{F}^{+}(\underset{\sim}{t})\right]}{\partial t_j}\Big|_{\underset{\sim}{t}=\underset{\sim}{t}_0} = \left(\mathbf{I}_p - \mathbf{F}(\underset{\sim}{t}_0)\mathbf{F}^{+}(\underset{\sim}{t}_0)\right)\left[\frac{\partial}{\partial t_j}\mathbf{F}(\underset{\sim}{t})\Big|_{\underset{\sim}{t}=\underset{\sim}{t}_0}\right]\mathbf{F}^{+}(\underset{\sim}{t}_0) +$$
$$+\left[\left(\mathbf{I}_p - \mathbf{F}(\underset{\sim}{t}_0)\mathbf{F}^{+}(\underset{\sim}{t}_0)\right)\left[\frac{\partial}{\partial t_j}\mathbf{F}(\underset{\sim}{t})\Big|_{\underset{\sim}{t}=\underset{\sim}{t}_0}\right]\mathbf{F}^{+}(\underset{\sim}{t}_0)\right]'.$$

Dada la dificultad para obtener las expresiones recogidas en los teoremas 4.2.2 y 4.2.3, en primer lugar, se estudiará el caso diagonal, transformándose posteriormente al caso general, a través de la relación existente entre una matriz simétrica y su diagonalización.

Dada una matriz diagonal $\mathbf{\Lambda} = diag\{\lambda_1, \lambda_2, ..., \lambda_p\} \in \mathcal{D}_p$ y una matriz $\mathbf{\Sigma} \in \mathcal{D}_p$, en esta sección se utilizará la notación que se describe a continuación:

- $\mathbf{D}_{dif,k} = (d_{ij})$, con
$$d_{ij} = \delta_{ij}(1 - \delta_{ik})(\lambda_i - \lambda_k) = \begin{cases} \lambda_i - \lambda_k & si \quad i = j \neq k \\ 0 & en\ caso\ contrario. \end{cases}$$

- $\mathbf{D}_k^{*}(t) = (d_{ij})$, con $d_{ij} = t\delta_{ij}\delta_{ik} = \begin{cases} t & si \quad i = j = k \\ 0 & en\ caso\ contrario. \end{cases}$

- $\mathbf{D}_k(t) = (d_{ij})$, con $d_{ij} = t\delta_{ij}(1 - \delta_{ik}) = \begin{cases} t & si \quad i = j \neq k \\ 0 & en\ caso\ contrario. \end{cases}$

- $\mathbf{J}^{*}_{uv} = (j_{lm})$, con $j_{l,m} = \delta_{lu}\delta_{mv} = \begin{cases} 1 & si \quad l = u \; y \; m = v \\ 0 & en \; caso \; contrario. \end{cases}$

- $\mathbf{J}_{uv} = (j_{lm})$, con
  $j_{l,m} = \delta_{lu}\delta_{mv} + \delta_{lv}\delta_{mu}\left(1 - \delta_{uv}\right) =$
  $= \begin{cases} 1 & si \quad (l = u \; y \; m = v) \; \acute{o} \; (l = v \; y \; m = u) \\ 0 & en \; caso \; contrario. \end{cases}$

- $\lambda_k\left(\mathbf{\Sigma}\right) = \lambda_k^{\mathbf{\Lambda}}\left(\mathbf{\Sigma}\right)$ y $\underline{\alpha}_k\left(\mathbf{\Sigma}\right) = \underline{\alpha}_k^{\mathbf{\Lambda}}\left(\mathbf{\Sigma}\right)$.

- $\mathbf{F}_k\left(\mathbf{\Sigma}\right) = \mathbf{\Sigma} - \lambda_k\left(\mathbf{\Sigma}\right)\mathbf{I}_p$.

- $\underline{e}_l = \left(\delta_{li}\right)_{i=1,...,p}$.

- $\gamma_{lk} = \begin{cases} \dfrac{1}{\lambda_l - \lambda_k} & si \quad l \neq k \\ 0 & si \quad l = k. \end{cases}$

Esta sección se divide en distintos apartados, en los cuales se calculan las derivadas sucesivas de un autovector evaluado cuando la matriz de covarianzas poblacional es diagonal. El primer término del desarrollo, en el caso diagonal, trivialmente, es $\underline{\alpha}_k\left(\mathbf{\Lambda}\right) = \underline{e}_k$.

4.2.1 Derivada de primer orden de un autovector

En este apartado, se va a obtener la expresión de la derivada de un autovector de $\mathbf{\Sigma}$, cuando dicha matriz es diagonal. Para ello se debe tener en cuenta el siguiente resultado:

Teorema 4.2.4 *Sea* $\mathbf{\Lambda} = diag\left\{\lambda_1, \lambda_2, ..., \lambda_p\right\} \in \mathcal{D}_p$, *con* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. *Si* λ_k *es simple,*

1. $\mathbf{F}_k^{+}\left(\mathbf{\Lambda}\right) = diag\left\{\gamma_{1k}, \gamma_{2k}, ..., \gamma_{pk}\right\}$.
2. $\mathbf{F}_k\left(\mathbf{\Lambda}\right)\mathbf{F}_k^{+}\left(\mathbf{\Lambda}\right) = \mathbf{F}_k^{+}\left(\mathbf{\Lambda}\right)\mathbf{F}_k\left(\mathbf{\Lambda}\right) = \mathbf{D}_k\left(1\right) \quad y$
   $\mathbf{I}_p - \mathbf{F}_k\left(\mathbf{\Lambda}\right)\mathbf{F}_k^{+}\left(\mathbf{\Lambda}\right) = \mathbf{D}_k^{*}\left(1\right)$.
3. $\left(\mathbf{F}_k^{+}\left(\mathbf{\Lambda}\right)\right)'\mathbf{F}_k^{+}\left(\mathbf{\Lambda}\right) = \left(\mathbf{D}_{dif,k}^{+}\right)^2$.

Demostración

El apartado **1** se obtiene de forma directa a partir de la inversa de Moore-Penrose, de una matriz diagonal $\mathbf{D} = (d_{ij})$ $(d_{ij} = 0, \forall i \neq j)$:

$$\mathbf{D}^{+} = \left(d_{ij}^{+}\right) \quad / \quad d_{ij}^{+} = \begin{cases} \dfrac{1}{d_{ij}} & si \quad d_{ij} \neq 0 \\ 0 & si \quad d_{ij} = 0. \end{cases}$$

Los apartados **2** y **3** se obtienen por propiedades básicas del cálculo matricial.

■

Nota 4.2.1 *Puede observarse que las columnas de* $\mathbf{F}_k^{+}(\mathbf{\Lambda})$ *son*

$$\underline{g}_l^k = \gamma_{lk}\underline{e}_l, \qquad l = 1, ..., p. \tag{4.6}$$

En adelante, a la derivada de una función de la matriz $\mathbf{\Sigma}$, $f(\mathbf{\Sigma})$, evaluada sobre una matriz diagonal, se denotará por $\left.\dfrac{\partial f}{\partial \sigma_{lm}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}}$, que realmente representa $\left.\dfrac{\partial f(\mathbf{\Sigma}(\underline{\sigma}))}{\partial \sigma_{lm}}\right|_{\underline{\sigma}=\underline{\sigma}_0}$ donde si $\sigma_{ij}^0 = \lambda_i \delta_{ij}$,

$$\underline{\sigma}_0 = \left(\sigma_{11}^0, \sigma_{12}^0, ..., \sigma_{1p}^0, \sigma_{22}^0, ..., \sigma_{2p}^0, ..., \sigma_{p-1,p}^0, \sigma_{pp}^0\right).$$

Lema 4.2.5 *Sea* $\mathbf{\Lambda} = diag\{\lambda_1, \lambda_2, ..., \lambda_p\}$ *con* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$. *Si* λ_k *es simple, entonces,*

1. $\left.\dfrac{\partial \lambda_k(\mathbf{\Sigma})}{\partial \sigma_{lm}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}} = \delta_{lm}\delta_{lk}.$

2. $\left.\dfrac{\partial \underline{\alpha}_k(\mathbf{\Sigma})}{\partial \sigma_{lm}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}} = \begin{cases} -\dfrac{1}{\lambda_m - \lambda_k}\underline{e}_m & si \quad l = k \neq m \\ -\dfrac{1}{\lambda_l - \lambda_k}\underline{e}_l & si \quad l \neq m = k \\ \underline{0} & en\ caso\ contrario. \end{cases}$

3. $\left.\dfrac{\partial \alpha_{kj}(\mathbf{\Sigma})}{\partial \sigma_{lm}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}} = \begin{cases} -\dfrac{1}{\lambda_j - \lambda_k} & si \quad k = l \neq m = j \ \acute{o}\ l = j \neq m = k \\ 0 & en\ caso\ contrario. \end{cases}$

Demostración

1. Teniendo en cuenta que $\underline{\alpha}_k(\mathbf{\Lambda}) = \underline{e}_k$ y $\lambda_k(\mathbf{\Lambda}) = \lambda_k$, haciendo uso del apartado **1** del teorema 4.2.2,

$$\left.\frac{\partial \lambda_k}{\partial \sigma_{lm}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}} = (2 - \delta_{lm})\,\delta_{kl}\delta_{km} = \delta_{kl}\delta_{km}.$$

2. Del apartado **2** del teorema 4.2.2 y de la expresión (4.6), se tiene que

$$\left.\frac{\partial \underline{\alpha}_k}{\partial \sigma_{lm}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}} = -\left[\delta_{km}\gamma_{lk}\underline{e}_l + (1 - \delta_{lm})\,\delta_{kl}\gamma_{mk}\underline{e}_m\right],$$

de donde se obtiene el resultado dado en el apartado **2.**

3. Se obtiene de forma directa del apartado **2.**

■

4.2.2 Derivada de segundo orden de un autovector

En el apartado **2** del teorema 4.2.2 se observa que para calcular la derivada de segundo orden de un autovector, $\underline{\alpha}_k$, es preciso hallar $\left.\frac{\partial \mathbf{F}_k^+}{\partial \sigma_{uv}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}}$. A su vez, previamente es necesario el cálculo de $\left.\frac{\partial \mathbf{F}_k}{\partial \sigma_{uv}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}}$, como se observa en la expresión del apartado **1** del lema 4.2.3.

En los resultados enunciados en éste y apartados sucesivos, en general, existe una larga casuística para las distintas relaciones entre los subíndices de las componentes de la matriz $\mathbf{\Sigma}$, respecto de las que se está derivando, y el índice k considerado. Por ello, en dichos enunciados, se muestra únicamente cada una de las posibles situaciones que se pueden dar, omitiendo situaciones análogas obtenidas tras permutación de dichos índices. En tal caso, se señalará con $(*)$ la existencia de omisión de casos análogos. Aquellos casos que no aparezcan en los enunciados, ni sean análogos a alguno de ellos, corresponden a relaciones en las que la derivada es nula, lo cual queda justificado en la propia demostración de los resultados.

Lema 4.2.6 *Sea* $\mathbf{\Lambda} = diag\{\lambda_1, \lambda_2, ..., \lambda_p\}$ *con* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$. *Si* λ_k *es simple, entonces,*

$$\left.\frac{\partial \mathbf{F}_k(\mathbf{\Sigma})}{\partial \sigma_{uv}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}} = \mathbf{J}_{uv} - \delta_{uv}\delta_{uk}\mathbf{I}_p. \tag{4.7}$$

Demostración

Teniendo en cuenta que para toda matriz de $\mathcal{D}_p$, $\dfrac{\partial}{\partial\sigma_{uv}}\boldsymbol{\Sigma} = \mathbf{J}_{uv}$ y el apartado **1** del lema 4.2.5, se tiene de forma inmediata (4.7). ■

Utilizando el lema 4.2.6, se puede determinar $\left.\dfrac{\partial \mathbf{F}_k^+}{\partial\sigma_{uv}}\right|_{\boldsymbol{\Sigma}=\boldsymbol{\Lambda}}$.

Lema 4.2.7 *Sea* $\boldsymbol{\Lambda} = diag\{\lambda_1, \lambda_2, ..., \lambda_p\}$ *con* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$. *Si* λ_k *es simple, entonces,* (*)

$$\left.\frac{\partial \mathbf{F}_k^+(\boldsymbol{\Sigma})}{\partial\sigma_{uv}}\right|_{\boldsymbol{\Sigma}=\boldsymbol{\Lambda}} = \begin{cases} \left(\mathbf{D}_{dif,k}^+\right)^2 & si \quad k = u = v \\ \dfrac{1}{(\lambda_v - \lambda_k)^2}\mathbf{J}_{uv} & si \quad k = u \neq v \\ -\dfrac{1}{(\lambda_u - \lambda_k)(\lambda_v - \lambda_k)}\mathbf{J}_{uv} & si \quad u \neq k \neq v. \end{cases} \tag{4.8}$$

Demostración

A partir del apartado **1** del lema 4.2.3 para $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}$ y los apartados **2** y **3** del lema 4.2.4, se obtiene que

$$\left.\frac{\partial \mathbf{F}_k^+(\boldsymbol{\Sigma})}{\partial\sigma_{uv}}\right|_{\boldsymbol{\Sigma}=\boldsymbol{\Lambda}} = \mathbf{S}_1 + \mathbf{S}_2 + \mathbf{S}_3,$$

donde

$$\begin{aligned} \mathbf{S}_1 &= -\mathbf{F}_k^+(\boldsymbol{\Lambda}) \left.\frac{\partial \mathbf{F}_k(\boldsymbol{\Sigma})}{\partial\sigma_{uv}}\right|_{\boldsymbol{\Sigma}=\boldsymbol{\Lambda}} \mathbf{F}_k^+(\boldsymbol{\Lambda}), \\ \mathbf{S}_2 &= \left(\mathbf{D}_{dif,k}^+\right)^2 \left(\mathbf{J}_{uv} - \delta_{uv}\delta_{uk}\mathbf{I}_p\right) \mathbf{D}_k^*(1). \\ \mathbf{S}_3 &= \mathbf{D}_k^*(1) \left(\mathbf{J}_{uv} - \delta_{uv}\delta_{uk}\mathbf{I}_p\right) \left(\mathbf{D}_{dif,k}^+\right)^2 = \mathbf{S}_2'. \end{aligned}$$

A continuación, se calcula cada uno de los sumandos anteriores:

- Teniendo en cuenta que $\mathbf{F}_k^+(\boldsymbol{\Lambda}) = \mathbf{D}_{dif,k}^+$ y la expresión (4.7),

$$\mathbf{S}_1 = \begin{cases} -\mathbf{D}_{dif,k}^+ \mathbf{D}_k(-1)\mathbf{D}_{dif,k}^+ & si \quad k = u = v \\ -\mathbf{D}_{dif,k}^+ \mathbf{J}_{uv} \mathbf{D}_{dif,k}^+ & en \quad caso \quad contrario. \end{cases}$$

Dado que $-\mathbf{D}_{dif,k}^+ \mathbf{D}_k(-1) = \mathbf{D}_{dif,k}^+$ y que

$$\mathbf{D}_{dif,k}^+ \mathbf{J}_{uv} \mathbf{D}_{dif,k}^+ = \begin{cases} \dfrac{1}{(\lambda_u - \lambda_k)(\lambda_v - \lambda_k)}\mathbf{J}_{uv} & si \quad k \neq u \quad y \quad k \neq v \\ \boldsymbol{\Theta} & en \quad caso \quad contrario, \end{cases}$$

se tiene que

$$\mathbf{S}_1 = \begin{cases} \left(\mathbf{D}^+_{dif,k}\right)^2 & si \quad k = u = v \\ -\dfrac{1}{(\lambda_u - \lambda_k)(\lambda_v - \lambda_k)}\mathbf{J}_{uv} & si \quad k \neq u \quad y \quad k \neq v \\ \boldsymbol{\Theta} & en \quad caso \quad contrario. \end{cases}$$

- Para $\mathbf{S}_2$ se pueden distinguir los casos,

$$\mathbf{S}_2 = \begin{cases} \left(\mathbf{D}^+_{dif,k}\right)^2 \mathbf{D}_k(-1)\,\mathbf{D}^*_k(1) & si \quad k = u = v \\ \left(\mathbf{D}^+_{dif,k}\right)^2 \mathbf{J}_{uv}\mathbf{D}^*_k(1) & en \quad caso \quad contrario. \end{cases}$$

Teniendo en cuenta que $\mathbf{D}_k(s)\,\mathbf{D}^*_k(t) = \boldsymbol{\Theta}$, $\forall s,t$ y $\mathbf{J}_{uv}\mathbf{D}^*_k(t) = \boldsymbol{\Theta}$, si $k \neq u$ y $k \neq v$, entonces, se tiene que

$$\mathbf{S}_2 = \begin{cases} \dfrac{1}{(\lambda_u - \lambda_k)^2}\mathbf{J}^*_{uv} & si \quad u \neq v = k \\ \dfrac{1}{(\lambda_v - \lambda_k)^2}\mathbf{J}^*_{vu} & si \quad k = u \neq v \\ \boldsymbol{\Theta} & en \quad caso \quad contrario. \end{cases}$$

Dado que $\mathbf{J}^*_{uv} + \mathbf{J}^*_{vu} = \mathbf{J}_{uv}$, cuando $u \neq v$, sumando $\mathbf{S}_1, \mathbf{S}_2, \mathbf{S}'_2$, se obtiene (4.8).

■

Nota 4.2.2 *Se observa que las columnas de* $\left.\dfrac{\partial \mathbf{F}^+_k(\boldsymbol{\Sigma})}{\partial \sigma_{uv}}\right|_{\boldsymbol{\Sigma}=\boldsymbol{\Lambda}}$ *son* $(*)$

$$\left.\frac{\partial \underline{g}^k_l(\boldsymbol{\Sigma})}{\partial \sigma_{uv}}\right|_{\boldsymbol{\Sigma}=\boldsymbol{\Lambda}} = \begin{cases} \dfrac{1}{(\lambda_l - \lambda_k)^2}\underline{e}_l & si \quad k = u = v \neq l \\ \dfrac{1}{(\lambda_v - \lambda_k)^2}\underline{e}_v & si \quad l = u = k \neq v \\ \dfrac{1}{(\lambda_l - \lambda_k)^2}\underline{e}_k & si \quad k = u \neq v = l \\ -\dfrac{1}{(\lambda_u - \lambda_k)(\lambda_v - \lambda_k)}\underline{e}_v & si \quad v \neq k \neq u = l. \end{cases} \tag{4.9}$$

A continuación se obtiene la expresión de la derivada de segundo orden de un autovector de una matriz simétrica, evaluada sobre una matriz diagonal.

Lema 4.2.8 *Sea* $\mathbf{\Lambda} = diag\{\lambda_1, \lambda_2, ..., \lambda_p\}$ *con* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$. *Si* λ_k *es simple, entonces,* $(*)$

$$\left.\frac{\partial^2 \underline{\alpha}_k(\mathbf{\Sigma})}{\partial\sigma_{lm}\partial\sigma_{uv}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}} = \begin{cases} -\dfrac{1}{(\lambda_v-\lambda_k)^2}\underline{e}_v & si \quad l=m=k=u\neq v \\ \dfrac{1}{(\lambda_l-\lambda_k)(\lambda_m-\lambda_k)}\underline{e}_l & si \quad m=u\neq k=v\neq l \\ -\dfrac{1}{(\lambda_l-\lambda_k)^2}\underline{e}_k & si \quad k=m=u\neq v=l. \end{cases}$$

Demostración

A partir del apartado **2** del teorema 4.2.2 se obtiene que

$$\frac{\partial^2 \underline{\alpha}_k}{\partial\sigma_{lm}\partial\sigma_{uv}} = -\left[\frac{\partial\alpha_{km}}{\partial\sigma_{uv}}\underline{g}_l^k + \alpha_{km}\frac{\partial \underline{g}_l^k}{\partial\sigma_{uv}} + (1-\delta_{lm})\left(\frac{\partial\alpha_{kl}}{\partial\sigma_{uv}}\underline{g}_m^k + \alpha_{kl}\frac{\partial \underline{g}_m^k}{\partial\sigma_{uv}}\right)\right]. \tag{4.10}$$

Teniendo en cuenta el apartado **3** del lema 4.2.5, y las expresiones (4.6) y (4.9), y realizando una distinción de casos según los índices k, l, m, u, v, se obtiene

- Si $l = m$:
  - Si $l = m = u \neq k = v$:
$$\left.\frac{\partial^2 \underline{\alpha}_k}{\partial\sigma_{lm}\partial\sigma_{uv}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}} = \frac{1}{(\lambda_l-\lambda_k)^2}\underline{e}_l.$$
  - Si $l = m = k = u \neq v$:
$$\left.\frac{\partial^2 \underline{\alpha}_k}{\partial\sigma_{lm}\partial\sigma_{uv}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}} = -\frac{1}{(\lambda_v-\lambda_k)^2}\underline{e}_v.$$
- El caso $u = v$, es análogo a $l = m$, por la simetría respecto a los pares (l, m) y (u, v).
- Si $l \neq m$ y $u \neq v$:
  - Si $l \neq u = m \neq v = k \neq l$:
$$\left.\frac{\partial^2 \underline{\alpha}_k}{\partial\sigma_{lm}\partial\sigma_{uv}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}} = \frac{1}{(\lambda_l-\lambda_k)(\lambda_m-\lambda_k)}\underline{e}_l.$$

– Si $k = m = u \neq v = l$:

$$\left.\frac{\partial^2 \underline{\alpha}_k}{\partial\sigma_{lm}\partial\sigma_{uv}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}} = -\frac{1}{(\lambda_m - \lambda_k)^2}\underline{e}_k.$$

– Si $v \neq k = m \neq u = l \neq v$:

$$\left.\frac{\partial^2 \underline{\alpha}_k}{\partial\sigma_{lm}\partial\sigma_{uv}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}} = \frac{1}{(\lambda_u - \lambda_k)(\lambda_v - \lambda_k)}\underline{e}_v.$$

El resto de las posibles situaciones son análogas a éstas o nulas.

■

Nota 4.2.3 *En razonamientos posteriores son necesarias las derivadas de segundo orden de las componentes de los autovectores; por ello se detallan a continuación* (∗)

- *Si* $u = w = z = k \neq l = v$ *ó* $l = u = z = k \neq v = w$,

$$\left.\frac{\partial^2 \alpha_{kl}(\mathbf{\Sigma})}{\partial\sigma_{uv}\partial\sigma_{wz}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}} = -\frac{1}{(\lambda_v - \lambda_k)^2}. \tag{4.11}$$

- *Si* $z = u \neq k = v \neq w = l$,

$$\left.\frac{\partial^2 \alpha_{kl}(\mathbf{\Sigma})}{\partial\sigma_{uv}\partial\sigma_{wz}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}} = \frac{1}{(\lambda_w - \lambda_k)(\lambda_z - \lambda_k)}. \tag{4.12}$$

4.2.3 Derivada de tercer orden de un autovector

A partir de la expresión (4.10), se observa que para la obtención de la derivada de tercer orden de $\underline{\alpha}_k$, es necesario conocer la expresión de las columnas de la matriz $\dfrac{\partial_k^2 \mathbf{F}_k^+}{\partial\sigma_{uv}\partial\sigma_{wz}}$. Por ello, previamente, debido a la expresión dada en el apartado **1** del lema 4.2.3 y a la propia definición de $\mathbf{F}_k$, se hallan $\dfrac{\partial^2\lambda_k}{\partial\sigma_{wz}\partial\sigma_{uv}}$, $\dfrac{\partial \mathbf{F}_k\mathbf{F}_k^+}{\partial\sigma_{wz}}$, $\dfrac{\partial \mathbf{F}_k^+\mathbf{F}_k}{\partial\sigma_{wz}}$ y $\dfrac{\partial_k^2\mathbf{F}_k}{\partial\sigma_{uv}\partial\sigma_{wz}}$.

Lema 4.2.9 *Sea* $\mathbf{\Lambda} = diag\{\lambda_1, \lambda_2, ..., \lambda_p\}$ *con* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$. *Si* λ_k *es simple, entonces,* (∗)

$$\left.\frac{\partial^2 \lambda_k(\mathbf{\Sigma})}{\partial\sigma_{uv}\partial\sigma_{wz}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}} = -\frac{2}{\lambda_u - \lambda_k} \quad si \quad u = w \neq v = z = k\,. \tag{4.13}$$

Demostración

Derivando respecto a σ_{wz} la expresión del apartado **1** del teorema 4.2.2, se tiene que

$$\frac{\partial^2 \lambda_k(\boldsymbol{\Sigma})}{\partial \sigma_{uv} \partial \sigma_{wz}} = (2 - \delta_{uv}) \left[\frac{\partial \alpha_{ku}(\boldsymbol{\Sigma})}{\partial \sigma_{wz}} \alpha_{kv}(\boldsymbol{\Sigma}) + \alpha_{ku}(\boldsymbol{\Sigma}) \frac{\partial \alpha_{kv}(\boldsymbol{\Sigma})}{\partial \sigma_{wz}} \right].$$

En particular, cuando $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}$, utilizando el apartado **3** del lema 4.2.5, se obtiene que el único caso en el que dicha derivada es no nula, salvo análogos, es para $u = w \neq k = v = z$ en el que

$$\left. \frac{\partial^2 \lambda_k}{\partial \sigma_{uv} \partial \sigma_{wz}} \right|_{\boldsymbol{\Sigma}=\boldsymbol{\Lambda}} = -\frac{2}{\lambda_u - \lambda_k}.$$

■

Lema 4.2.10 *Sea* $\boldsymbol{\Lambda} = diag\{\lambda_1, \lambda_2, ..., \lambda_p\}$ *con* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$. *Si* λ_k *es simple, entonces,*

$$\left. \frac{\partial \mathbf{F}_k(\boldsymbol{\Sigma}) \mathbf{F}_k^+(\boldsymbol{\Sigma})}{\partial \sigma_{wz}} \right|_{\boldsymbol{\Sigma}=\boldsymbol{\Lambda}} = \left. \frac{\partial \mathbf{F}_k^+(\boldsymbol{\Sigma}) \mathbf{F}_k(\boldsymbol{\Sigma})}{\partial \sigma_{wz}} \right|_{\boldsymbol{\Sigma}=\boldsymbol{\Lambda}} = \delta_{wk} \gamma_{zk} \mathbf{J}_{kz} + \delta_{zk} \gamma_{wk} \mathbf{J}_{kw}. \tag{4.14}$$

Demostración

Particularizando la expresión obtenida en el apartado **2b** del lema 4.2.3 para $\mathbf{F} = \mathbf{F}_k$, en el caso en el que $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}$, teniendo en cuenta el apartado **2** en el lema 4.2.4 y la expresión (4.7), se tiene que

$$\left. \frac{\partial \mathbf{F}_k \mathbf{F}_k^+}{\partial \sigma_{wz}} \right|_{\boldsymbol{\Sigma}=\boldsymbol{\Lambda}} = \mathbf{S}_1 + \mathbf{S}_1',$$

donde $\mathbf{S}_1 = \mathbf{D}_k^*(1) \left(\mathbf{J}_{wz} - \delta_{wz} \delta_{wk} \mathbf{I}_p \right) \mathbf{D}_{dif,k}^+$

Operando, se tiene que

$$\begin{aligned} \mathbf{S}_1 &= \mathbf{D}_k^*(1) \mathbf{J}_{wz} \mathbf{D}_{dif,k}^+ = \mathbf{D}_k^*(1) \left[\delta_{wk} \gamma_{zk} \mathbf{J}_{kz}^* + (1 - \delta_{wz}) \delta_{zk} \gamma_{wk} \mathbf{J}_{kw}^* \right] = \\ &= \delta_{wk} \gamma_{zk} \mathbf{J}_{kz}^* + \delta_{zk} \gamma_{wk} \mathbf{J}_{kw}^*, \end{aligned}$$

de donde se obtiene (4.14).

Por otro lado, a partir de la expresión del apartado **2b** del lema 4.2.3, para $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}$ se tiene que

$$\left. \frac{\partial \mathbf{F}_k^+ \mathbf{F}_k}{\partial \sigma_{wz}} \right|_{\boldsymbol{\Sigma}=\boldsymbol{\Lambda}} = \mathbf{S}_1' + \mathbf{S}_1 = \left. \frac{\partial \mathbf{F}_k \mathbf{F}_k^+}{\partial \sigma_{wz}} \right|_{\boldsymbol{\Sigma}=\boldsymbol{\Lambda}}.$$

■

Lema 4.2.11 *Sea* $\mathbf{\Lambda} = diag\{\lambda_1, \lambda_2, ..., \lambda_p\}$ *con* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$. *Si* λ_k *es simple, entonces,* $(*)$

$$\left.\frac{\partial^2 \mathbf{F}_k(\mathbf{\Sigma})}{\partial\sigma_{uv}\partial\sigma_{wz}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}} = \frac{2}{\lambda_u - \lambda_k}\mathbf{I}_p \quad si \qquad u = w \neq k = v = z\,. \tag{4.15}$$

Demostración

La derivada de segundo orden de $\mathbf{F}_k$ viene dada por

$$\left.\frac{\partial^2 \mathbf{F}_k}{\partial\sigma_{uv}\partial\sigma_{wz}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}} = \left.\frac{\partial^2 \mathbf{\Sigma}}{\partial\sigma_{uv}\partial\sigma_{wz}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}} - \left.\frac{\partial^2 \lambda_k}{\partial\sigma_{uv}\partial\sigma_{wz}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}} \mathbf{I}_p.$$

Teniendo en cuenta que $\dfrac{\partial^2 \mathbf{\Sigma}}{\partial\sigma_{uv}\partial\sigma_{wz}} = \mathbf{\Theta}$, cualquiera que sea $\mathbf{\Sigma} \in \mathcal{D}_p$ y la expresión (4.13), se obtiene directamente (4.15).

■

Con los resultados anteriores, se dispone de todos los elementos necesarios para hallar $\left.\dfrac{\partial^2 \mathbf{F}_k^+}{\partial\sigma_{uv}\partial\sigma_{wz}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}}$, matriz cuyas columnas son necesarias en el cálculo de la derivada de tercer orden de un autovector de $\mathbf{\Sigma}$. Por ello, en el resultado que se recoge a continuación, se detalla la casuística completa relativa a los subíndices de las componentes de la matriz $\mathbf{\Sigma}$ respecto de las que se deriva, salvo casos análogos y nulos, pudiendo conocer así de forma directa la expresión de las columnas de $\left.\dfrac{\partial^2 \mathbf{F}_k^+}{\partial\sigma_{uv}\partial\sigma_{wz}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}}$.

Lema 4.2.12 *Sea* $\mathbf{\Lambda} = diag\{\lambda_1, \lambda_2, ..., \lambda_p\}$ *con* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$. *Si* λ_k *es simple, entonces,* $(*)$

1. *Si* $w = z = k = u = v$,

$$\left.\frac{\partial^2 \mathbf{F}_k^+(\mathbf{\Sigma})}{\partial\sigma_{uv}\partial\sigma_{wz}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}} = 2\left(\mathbf{D}_{dif,k}^+\right)^3.$$

2. *Si* $w = z = k \neq u = v$,

$$\left.\frac{\partial^2 \mathbf{F}_k^+(\mathbf{\Sigma})}{\partial\sigma_{uv}\partial\sigma_{wz}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}} = -\frac{2}{(\lambda_u - \lambda_k)^3}\mathbf{D}_u^*(1).$$

3. *Si* $w = z = k \neq u \neq v \neq k$,

$$\left.\frac{\partial^2 \mathbf{F}_k^+(\mathbf{\Sigma})}{\partial\sigma_{uv}\partial\sigma_{wz}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}} = -\left[\frac{1}{(\lambda_u - \lambda_k)^2(\lambda_v - \lambda_k)} + \frac{1}{(\lambda_u - \lambda_k)(\lambda_v - \lambda_k)^2}\right]\mathbf{J}_{uv}.$$

4. *Si* $w = z = k = v \neq u$,

$$\left.\frac{\partial^2 \mathbf{F}_k^+(\boldsymbol{\Sigma})}{\partial\sigma_{uv}\partial\sigma_{wz}}\right|_{\boldsymbol{\Sigma}=\boldsymbol{\Lambda}} = \frac{2}{(\lambda_u - \lambda_k)^3}\mathbf{J}_{uk}.$$

5. *Si* $u = k = w \neq z = v$,

$$\left.\frac{\partial^2 \mathbf{F}_k^+(\boldsymbol{\Sigma})}{\partial\sigma_{uv}\partial\sigma_{wz}}\right|_{\boldsymbol{\Sigma}=\boldsymbol{\Lambda}} = \frac{2}{(\lambda_v - \lambda_k)^3}\mathbf{D}_k^*(1) - \frac{4}{(\lambda_v - \lambda_k)^3}\mathbf{D}_v^*(1) - \\ -\frac{2}{\lambda_v - \lambda_k}\left(\mathbf{D}_{dif,k}^+\right)^2.$$

6. *Si* $v \neq u = k = w \neq z \neq v$,

$$\left.\frac{\partial^2 \mathbf{F}_k^+(\boldsymbol{\Sigma})}{\partial\sigma_{uv}\partial\sigma_{wz}}\right|_{\boldsymbol{\Sigma}=\boldsymbol{\Lambda}} = -\left[\frac{1}{(\lambda_v - \lambda_k)(\lambda_z - \lambda_k)^2} + \frac{1}{(\lambda_v - \lambda_k)^2(\lambda_z - \lambda_k)}\right]\mathbf{J}_{vz}.$$

7. *Si* $k = w \neq z = u = v$,

$$\left.\frac{\partial^2 \mathbf{F}_k^+(\boldsymbol{\Sigma})}{\partial\sigma_{uv}\partial\sigma_{wz}}\right|_{\boldsymbol{\Sigma}=\boldsymbol{\Lambda}} = -\frac{2}{(\lambda_u - \lambda_k)^3}\mathbf{J}_{uk}.$$

8. *Si* $v \neq k = w \neq z = u \neq v$,

$$\left.\frac{\partial^2 \mathbf{F}_k^+(\boldsymbol{\Sigma})}{\partial\sigma_{uv}\partial\sigma_{wz}}\right|_{\boldsymbol{\Sigma}=\boldsymbol{\Lambda}} = -\left[\frac{1}{(\lambda_u - \lambda_k)^2(\lambda_v - \lambda_k)} + \frac{1}{(\lambda_u - \lambda_k)(\lambda_v - \lambda_k)^2}\right]\mathbf{J}_{vk}.$$

9. *Si* $u = w = z = v \neq k$,

$$\left.\frac{\partial^2 \mathbf{F}_k^+(\boldsymbol{\Sigma})}{\partial\sigma_{uv}\partial\sigma_{wz}}\right|_{\boldsymbol{\Sigma}=\boldsymbol{\Lambda}} = \frac{2}{(\lambda_u - \lambda_k)^3}\mathbf{D}_u^*(1).$$

10. *Si* $k \neq u = w \neq z \neq k$ *y* $z \neq v \neq k$,

$$\left.\frac{\partial^2 \mathbf{F}_k^+(\boldsymbol{\Sigma})}{\partial\sigma_{uv}\partial\sigma_{wz}}\right|_{\boldsymbol{\Sigma}=\boldsymbol{\Lambda}} = \frac{1}{(\lambda_u - \lambda_k)(\lambda_v - \lambda_k)(\lambda_z - \lambda_k)}\mathbf{J}_{zv}.$$

11. *Si* $k \neq u = w \neq z = v \neq k$,

$$\left.\frac{\partial^2 \mathbf{F}_k^+(\boldsymbol{\Sigma})}{\partial\sigma_{uv}\partial\sigma_{wz}}\right|_{\boldsymbol{\Sigma}=\boldsymbol{\Lambda}} = \frac{2}{(\lambda_u - \lambda_k)(\lambda_v - \lambda_k)}\left[\frac{1}{\lambda_v - \lambda_k}\mathbf{D}_v^*(1) + \frac{1}{\lambda_u - \lambda_k}\mathbf{D}_u^*(1)\right].$$

Demostración

Considerando en el apartado **1** del lema 4.2.3, $\mathbf{F} = \mathbf{F}_k$, con derivada respecto a σ_{uv} y derivando de nuevo respecto a σ_{wz}, se tiene que

$$\left.\frac{\partial^2 \mathbf{F}_k^+(\mathbf{\Sigma})}{\partial\sigma_{uv}\partial\sigma_{wz}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}} = \mathbf{S}_1 + \mathbf{S}_2 + \mathbf{S}_3, \tag{4.16}$$

donde

$$\begin{aligned}
\mathbf{S}_1 &= \frac{\partial}{\partial\sigma_{wz}}\left[-\mathbf{F}_k^+\left(\frac{\partial\mathbf{F}_k}{\partial\sigma_{uv}}\right)\mathbf{F}_k^+\right]\Bigg|_{\mathbf{\Sigma}=\mathbf{\Lambda}},\\
\mathbf{S}_2 &= \frac{\partial}{\partial\sigma_{wz}}\left[\mathbf{F}_k^+\left(\mathbf{F}_k^+\right)'\left(\frac{\partial\mathbf{F}_k}{\partial\sigma_{uv}}\right)'\left(\mathbf{I}_p - \mathbf{F}_k\mathbf{F}_k^+\right)\right]\Bigg|_{\mathbf{\Sigma}=\mathbf{\Lambda}},\\
\mathbf{S}_3 &= \frac{\partial}{\partial\sigma_{wz}}\left[\left(\mathbf{I}_p - \mathbf{F}_k^+\mathbf{F}_k\right)\left(\frac{\partial\mathbf{F}_k}{\partial\sigma_{uv}}\right)'\left(\mathbf{F}_k^+\right)'\mathbf{F}_k^+\right]\Bigg|_{\mathbf{\Sigma}=\mathbf{\Lambda}}.
\end{aligned}$$

A su vez, teniendo en cuenta los apartados **2** y **3** del lema 4.2.4 y la expresión (4.7), dichos sumandos se pueden descomponer de la forma

$$\begin{aligned}
\mathbf{S}_1 &= \mathbf{S}_{11} + \mathbf{S}_{12} + \mathbf{S}_{13},\\
\mathbf{S}_2 &= \mathbf{S}_{21} + \mathbf{S}_{22} + \mathbf{S}_{23} + \mathbf{S}_{24},\\
\mathbf{S}_3 &= \mathbf{S}_{31} + \mathbf{S}_{32} + \mathbf{S}_{33} + \mathbf{S}_{34},
\end{aligned}$$

donde

$$\begin{aligned}
\mathbf{S}_{11} &= -\left.\frac{\partial\mathbf{F}_k^+}{\partial\sigma_{wz}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}}\left(\mathbf{J}_{uv} - \delta_{uv}\delta_{uk}\mathbf{I}_p\right)\mathbf{D}_{dif,k}^+ =\\
&= -\left.\frac{\partial\mathbf{F}_k^+}{\partial\sigma_{wz}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}}\left[\gamma_{vk}\mathbf{J}_{uv}^* + \left(1-\delta_{uv}\right)\gamma_{uk}\mathbf{J}_{vu}^* - \delta_{uv}\delta_{uk}\mathbf{D}_{dif,k}^+\right],\\
\mathbf{S}_{12} &= -\mathbf{D}_{dif,k}^+\left.\frac{\partial^2\mathbf{F}_k}{\partial\sigma_{uv}\partial\sigma_{wz}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}}\mathbf{D}_{dif,k}^+,\\
\mathbf{S}_{13} &= \mathbf{S}_{11}',\\
\mathbf{S}_{21} &= \left.\frac{\partial\mathbf{F}_k^+}{\partial\sigma_{wz}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}}\mathbf{D}_{dif,k}^+\left(\mathbf{J}_{uv} - \delta_{uv}\delta_{uk}\mathbf{I}_p\right)\mathbf{D}_k^*(1) =\\
&= \left.\frac{\partial\mathbf{F}_k^+}{\partial\sigma_{wz}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}}\left[\delta_{vk}\gamma_{uk}\mathbf{J}_{uv}^* + \left(1-\delta_{uv}\right)\delta_{uk}\gamma_{vk}\mathbf{J}_{vu}^*\right],\\
\mathbf{S}_{22} &= \mathbf{D}_{dif,k}^+\left.\frac{\partial\mathbf{F}_k^+}{\partial\sigma_{wz}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}}\left(\mathbf{J}_{uv} - \delta_{uv}\delta_{uk}\mathbf{I}_p\right)\mathbf{D}_k^*(1) =\\
&= \mathbf{D}_{dif,k}^+\left.\frac{\partial\mathbf{F}_k^+}{\partial\sigma_{wz}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}}\left[\delta_{vk}\mathbf{J}_{uv}^* + \left(1-\delta_{uv}\right)\delta_{uk}\mathbf{J}_{vu}^* - \delta_{uv}\delta_{uk}\mathbf{D}_k^*(1)\right],\\
\mathbf{S}_{23} &= \left(\mathbf{D}_{dif,k}^+\right)^2\left.\frac{\partial^2\mathbf{F}_k}{\partial\sigma_{uv}\partial\sigma_{wz}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}}\mathbf{D}_k^*(1),
\end{aligned}$$

$$\mathbf{S}_{24} = -\left(\mathbf{D}^{+}_{dif,k}\right)^{2}\left(\mathbf{J}_{uv} - \delta_{uv}\delta_{uk}\mathbf{I}_{p}\right)\left.\frac{\partial \mathbf{F}_k\mathbf{F}_k^{+}}{\partial\sigma_{wz}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}} =$$

$$= -\left[\gamma_{uk}^{2}\mathbf{J}^{*}_{uv} + (1-\delta_{uv})\,\gamma_{vk}^{2}\mathbf{J}^{*}_{vu} - \delta_{uv}\delta_{uk}\left(\mathbf{D}^{+}_{dif,k}\right)^{2}\right]\left.\frac{\partial \mathbf{F}_k\mathbf{F}_k^{+}}{\partial\sigma_{wz}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}},$$

y por último,

$$\mathbf{S}_{31} = \mathbf{S}'_{24}, \mathbf{S}_{32} = \mathbf{S}'_{23}, \mathbf{S}_{33} = \mathbf{S}'_{22}, \mathbf{S}_{34} = \mathbf{S}'_{21}.$$

Por lo tanto, $\mathbf{S}_3 = \mathbf{S}'_2$ y por lo tanto, para el cálculo de $\left.\dfrac{\partial^2 \mathbf{F}_k^{+}(\mathbf{\Sigma})}{\partial\sigma_{uv}\partial\sigma_{wz}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}}$, bastará hallar $\mathbf{S}_1$ y $\mathbf{S}_2$.

Para el cálculo de $\mathbf{S}_1$ y $\mathbf{S}_2$, es necesario la distinción de casos recogida en las expresiones obtenidas para $\left.\dfrac{\partial \mathbf{F}_k^{+}}{\partial\sigma_{wz}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}}$, $\left.\dfrac{\partial \mathbf{F}_k\mathbf{F}_k^{+}}{\partial\sigma_{wz}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}}$ y $\left.\dfrac{\partial^2 \mathbf{F}_k}{\partial\sigma_{uv}\partial\sigma_{wz}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}}$ donde se utiliza (4.8), (4.14) y (4.15). Se analizarán los casos $w = z = k$, $w = z \neq k$, $w \neq z \neq k$. El caso $w \neq z = k$ se omitirá, por ser análogo a aquél en el que $w = z \neq k$.

- Si $w = z = k$:

  $$\mathbf{S}_{11} = -\left(\mathbf{D}^{+}_{dif,k}\right)^{2}\left[\gamma_{vk}\mathbf{J}^{*}_{uv} + (1-\delta_{uv})\,\gamma_{uk} - \delta_{uv}\delta_{uk}\mathbf{D}^{+}_{dif,k}\right] =$$

  $$= \delta_{uv}\delta_{uk}\left(\mathbf{D}^{+}_{dif,k}\right)^{3} - \gamma_{uk}^{2}\gamma_{vk}\mathbf{J}^{*}_{uv} - (1-\delta_{uv})\,\gamma_{uk}\gamma_{vk}^{2}\mathbf{J}^{*}_{vu},$$

  $$\mathbf{S}_{12} = \mathbf{\Theta}.$$

  Por lo tanto,

  $$\mathbf{S}_1 = 2\delta_{uv}\delta_{uk}\left(\mathbf{D}^{+}_{dif,k}\right)^{3} - \left(\gamma_{uk}^{2}\gamma_{vk} + \gamma_{uk}\gamma_{vk}^{2}\right)\mathbf{J}_{uv}.$$

  Por otro lado,

  $$\mathbf{S}_{21} = \left(\mathbf{D}^{+}_{dif,k}\right)^{3}\left[\delta_{vk}\gamma_{uk}\mathbf{J}^{*}_{uv} + (1-\delta_{uv})\,\delta_{uk}\gamma_{vk}\mathbf{J}^{*}_{vu}\right] =$$

  $$= \delta_{kv}\gamma_{uk}^{3}\mathbf{J}^{*}_{uv} + (1-\delta_{uv})\,\delta_{ku}\gamma_{vk}^{3}\mathbf{J}^{*}_{vu},$$

  $$\mathbf{S}_{22} = \mathbf{S}_{21},$$

  $$\mathbf{S}_{23} = \mathbf{S}_{24} = \mathbf{\Theta}.$$

  Luego,

  $$\mathbf{S}_2 + \mathbf{S}'_2 = 2\left[\delta_{kv}\gamma_{uk}^{3} + \delta_{ku}\gamma_{vk}^{3}\right]\mathbf{J}_{uv},$$

  Y sumando $\mathbf{S}_1$, $\mathbf{S}_2$ y $\mathbf{S}'_2$, considerando los posibles casos en función de los índices k, u, v, w, z, se obtienen los resultados recogidos en los apartados **1**, **2**, **3** y **4** del enunciado.

- Si $k = w \neq z$ (caso análogo para $k = z \neq w$):

  Realizando cálculos para cada uno de los sumando que conforman $\mathbf{S}_1$, se obtiene,

$$\begin{aligned}\mathbf{S}_{11} &= -\gamma_{zk}^2 \mathbf{J}_{wz} \left[\gamma_{vk} \mathbf{J}_{uv}^* + (1-\delta_{uv})\gamma_{uk} - \delta_{uv}\delta_{uk}\mathbf{D}_{dif,k}^+\right] = \\ &= \gamma_{zk}^2 \{\delta_{uv}\delta_{uk}\gamma_{zk}\mathbf{J}_{kz}^* - \gamma_{vk}\left(\delta_{zu}\mathbf{J}_{kv}^* + \delta_{uk}\mathbf{J}_{zv}^*\right) - \\ &- (1-\delta_{uv})\gamma_{uk}\left[\delta_{zv}\mathbf{J}_{ku}^* + \delta_{vk}\mathbf{J}_{zu}^*\right]\}.\end{aligned}$$

  Para el segundo sumando,

$$\mathbf{S}_{12} = -2\gamma_{zk}\left[\delta_{uk}\delta_{vz} + \delta_{vk}\delta_{uz}\right]\left(\mathbf{D}_{dif,k}^+\right)^2.$$

  Entonces, teniendo en cuenta que

$$\mathbf{J}_{zv}^* + \mathbf{J}_{vz}^* = \begin{cases} 2\mathbf{D}_v^*(1) & si \quad z = v \\ \mathbf{J}_{vz} & si \quad z \neq v, \end{cases}$$

  es posible obtener la expresión de $\mathbf{S}_1$,

$$\begin{aligned}\mathbf{S}_1 &= \gamma_{zk}^2 \{\delta_{uv}\delta_{uk}\gamma_{zk}\mathbf{J}_{kz} - \gamma_{vk}\left[\delta_{zu}\mathbf{J}_{vk} + \delta_{uk}\left(\mathbf{J}_{zv}^* + \mathbf{J}_{vz}^*\right)\right] - \\ &- (1-\delta_{uv})\gamma_{uk}\left[\delta_{zv}\mathbf{J}_{uk} - \delta_{vk}\gamma_{uk}\left(\mathbf{J}_{zu}^* + \mathbf{J}_{uz}^*\right)\right]\} - \\ &-2\gamma_{zk}\left[\delta_{uk}\delta_{vz} + \delta_{vk}\delta_{uz}\right]\left(\mathbf{D}_{dif,k}^+\right)^2.\end{aligned}$$

  Por otra parte, los sumandos de $\mathbf{S}_2$ se pueden expresar como sigue,

$$\begin{aligned}\mathbf{S}_{21} &= \gamma_{zk}^2 \mathbf{J}_{wz}\left[\delta_{vk}\gamma_{uk}\mathbf{J}_{uv}^* + (1-\delta_{uv})\delta_{uk}\gamma_{vk}\mathbf{J}_{vu}^*\right] = \\ &= \gamma_{zk}^3\left[\delta_{kv}\delta_{zu} + \delta_{ku}\delta_{zv}\right]\mathbf{D}_k^*(1), \\ \mathbf{S}_{22} &= \mathbf{D}_{dif,k}^+\gamma_{zk}^2\mathbf{J}_{wz}\left[\delta_{vk}\mathbf{J}_{uv}^* + (1-\delta_{uv})\delta_{uk}\mathbf{J}_{vu}^* - \delta_{uv}\delta_{uk}\mathbf{D}_k^*(1)\right] = \\ &= \gamma_{zk}^3\left[\delta_{kv}\delta_{ku}\mathbf{J}_{zv}^* + \delta_{ku}(1-\delta_{uv})\delta_{kv}\mathbf{J}_{zu}^* - \delta_{uv}\delta_{uk}\mathbf{J}_{zk}^*\right] = \mathbf{\Theta}, \\ \mathbf{S}_{23} &= \left(\mathbf{D}_{dif,k}^+\right)^2 2\left[\gamma_{vk}\delta_{uk}\delta_{vz} + \gamma_{uk}\delta_{vk}\delta_{uz}\right]\mathbf{D}_k^*(1) = \mathbf{\Theta}, \\ \mathbf{S}_{24} &= -\left[\gamma_{uk}^2\mathbf{J}_{uv}^* + (1-\delta_{uv})\gamma_{vk}^2\mathbf{J}_{vu}^* - \delta_{uv}\delta_{uk}\left(\mathbf{D}_{dif,k}^+\right)^2\right]\gamma_{zk}\mathbf{J}_{kz} = \\ &= \gamma_{zk}\left[\delta_{uv}\delta_{uk}\gamma_{zk}^2\mathbf{J}_{zk}^* - \gamma_{uk}^2\left[\delta_{vk}\mathbf{J}_{uz}^* + \delta_{vz}\mathbf{J}_{uk}^*\right] - \right. \\ &- \left.(1-\delta_{uv})\gamma_{vk}^2\left[\delta_{uk}\mathbf{J}_{vz}^* + \delta_{uz}\mathbf{J}_{vk}^*\right]\right].\end{aligned}$$

  Simplificando adecuadamente, se obtiene que

$$\begin{aligned}\mathbf{S}_2 + \mathbf{S}_2' &= 2\gamma_{zk}^3\left[\delta_{kv}\delta_{zu} + \delta_{ku}\delta_{zv}\right]\mathbf{D}_k^*(1) + \\ &+\gamma_{zk}\left[\delta_{uv}\delta_{uk}\gamma_{zk}^2\mathbf{J}_{zk} - \gamma_{uk}^2\left[\delta_{vk}\left(\mathbf{J}_{uz}^* + \mathbf{J}_{zu}^*\right) + \delta_{vz}\mathbf{J}_{uk}\right] -\end{aligned}$$

$$- (1-\delta_{uv})\,\gamma_{vk}^2\,[\delta_{uk}\,(\mathbf{J}^*_{vz}+\mathbf{J}^*_{zv})+\delta_{uz}\mathbf{J}_{vk}]].$$

Al igual que en el caso anterior, esta expresión se puede simplificar en función de los subíndices, obteniéndose las expresiones recogidas en los apartados **5**, **6**, **7** y **8**.

- Si $w \neq k \neq z$, razonando de forma similar al caso anterior, se obtiene

$$\begin{aligned}\mathbf{S}_{11} &= \gamma_{wk}\gamma_{zk}\mathbf{J}_{wz}\left[\gamma_{vk}\mathbf{J}^*_{uv}+(1-\delta_{uv})\,\gamma_{uk}\mathbf{J}^*_{vu}-\delta_{uv}\delta_{uk}\mathbf{D}^+_{dif,k}\right]= \\ &= \gamma_{wk}\gamma_{zk}\left[\gamma_{vk}\left[\delta_{zu}\mathbf{J}^*_{wv}+(1-\delta_{wz})\,\delta_{uw}\mathbf{J}^*_{zv}\right]+\right. \\ &\quad +(1-\delta_{uv})\,\gamma_{uk}\left[\delta_{zv}\mathbf{J}^*_{wu}+(1-\delta_{wz})\,\delta_{vw}\mathbf{J}^*_{zu}\right]- \\ &\quad \left.-\delta_{uv}\delta_{uk}\left[\gamma_{zk}\mathbf{J}^*_{wz}+(1-\delta_{wz})\,\gamma_{wk}\mathbf{J}^*_{zw}\right]\right],\end{aligned}$$

$$\mathbf{S}_{12}=\boldsymbol{\Theta}.$$

Entonces,

$$\begin{aligned}\mathbf{S}_1 &= \gamma_{wk}\gamma_{zk}\left[\gamma_{vk}\left[\delta_{zu}\left(\mathbf{J}^*_{wv}+\mathbf{J}^*_{vw}\right)+(1-\delta_{wz})\,\delta_{uw}\left(\mathbf{J}^*_{zv}+\mathbf{J}^*_{vz}\right)\right]+\right. \\ &\quad +(1-\delta_{uv})\,\gamma_{uk}\left[\delta_{zv}\left(\mathbf{J}^*_{wu}+\mathbf{J}^*_{uw}\right)+(1-\delta_{wz})\,\delta_{vw}\left(\mathbf{J}^*_{zu}+\mathbf{J}^*_{uz}\right)\right]- \\ &\quad \left.-\delta_{uv}\delta_{uk}\left[\gamma_{zk}+(1-\delta_{wz})\,\gamma_{wk}\right]\left(\mathbf{J}^*_{wz}+\mathbf{J}^*_{zw}\right)\right].\end{aligned}$$

Por otro lado, se determinan los sumandos de $\mathbf{S}_2$:

$$\begin{aligned}\mathbf{S}_{21} &= \gamma_{wk}\gamma_{zk}\mathbf{J}_{wz}\left[\delta_{vk}\gamma_{uk}\mathbf{J}^*_{uv}+(1-\delta_{uv})\,\delta_{uk}\gamma_{vk}\mathbf{J}^*_{vu}\right]= \\ &= \gamma_{wk}\gamma_{zk}\left[\delta_{vk}\gamma_{uk}\left[\delta_{zu}\mathbf{J}^*_{wv}+(1-\delta_{wz})\,\delta_{wu}\mathbf{J}^*_{zv}\right]+\right. \\ &\quad \left.+(1-\delta_{uv})\,\delta_{uk}\gamma_{vk}\left[\delta_{zv}\mathbf{J}^*_{wu}+(1-\delta_{wz})\,\delta_{wv}\mathbf{J}^*_{zu}\right]\right], \\ \mathbf{S}_{22} &= \gamma_{wk}\gamma_{zk}\mathbf{D}^+_{dif,k}\mathbf{J}_{wz}\left[\delta_{vk}\mathbf{J}^*_{uv}+(1-\delta_{uv})\,\delta_{uk}\mathbf{J}^*_{vu}-\delta_{uv}\delta_{uk}\mathbf{D}^*_k(1)\right]= \\ &= \gamma_{wk}\gamma_{zk}\left[\delta_{vk}\delta_{zu}\gamma_{wk}\mathbf{J}^*_{wv}+\delta_{ku}\delta_{zv}\gamma_{wk}\mathbf{J}^*_{wu}+\right. \\ &\quad \left.+(1-\delta_{wz})\left[\delta_{kv}\delta_{wu}\gamma_{zk}\mathbf{J}^*_{zv}+\delta_{ku}\delta_{wv}\gamma_{zk}\mathbf{J}^*_{zu}\right]\right],\end{aligned}$$

$$\mathbf{S}_{23}=\mathbf{S}_{24}=\boldsymbol{\Theta}.$$

Entonces,

$$\begin{aligned}\mathbf{S}_2+\mathbf{S}'_2 &= \gamma_{wk}\gamma_{zk}\left[\delta_{vk}\delta_{zu}\left(\gamma_{uk}+\gamma_{wk}\right)\mathbf{J}_{wz}+(1-\delta_{wz})\,\delta_{uw}\delta_{vk}\left(\gamma_{uk}+\gamma_{zk}\right)\mathbf{J}_{zv}+\right. \\ &\quad \left.+\delta_{uk}\delta_{zv}\left(\gamma_{kv}+\gamma_{wk}\right)\mathbf{J}_{wu}+(1-\delta_{wz})\,\delta_{uk}\delta_{vw}\left(\gamma_{vk}+\gamma_{zk}\right)\mathbf{J}_{zu}\right].\end{aligned}$$

Para reducir el número de casos a distinguir, se omitirán aquellos en los que k coincide con u ó v, ya que sería un caso análogo al de que k coincida con w ó z, casos que ya han sido estudiados. Entonces,

$$\begin{aligned}\left.\frac{\partial^2\mathbf{F}^+_k}{\partial\sigma_{uv}\partial\sigma_{wz}}\right|_{\boldsymbol{\Sigma}=\boldsymbol{\Lambda}} &= \gamma_{wk}\gamma_{zk}\left[\delta_{uw}\gamma_{vk}\left(\mathbf{J}^*_{zv}+\mathbf{J}^*_{vz}\right)+\right. \\ &\quad +\delta_{vw}\left(1-\delta_{uv}\right)\gamma_{uk}\left(\mathbf{J}^*_{zu}+\mathbf{J}^*_{uz}\right)+\delta_{zu}\left(1-\delta_{wz}\right)\gamma_{vk}\left(\mathbf{J}^*_{vw}+\mathbf{J}^*_{wv}\right)+ \\ &\quad \left.+\,\delta_{zv}\left(1-\delta_{wz}\right)\left(1-\delta_{uv}\right)\gamma_{uk}\left(\mathbf{J}^*_{wu}+\mathbf{J}^*_{uw}\right)\right].\end{aligned}$$

cuya expresión se simplifica al considerar los casos recogidos en los apartados **9**, **10** y **11**.

■

Por último, se obtiene la derivada de tercer orden de un autovector:

Lema 4.2.13 *Sea* $\mathbf{\Lambda} = diag\{\lambda_1, \lambda_2, ..., \lambda_p\}$ *con* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$. *Si* λ_k *es simple, entonces,* $(*)$

1. Si $l = m = v \neq k = u = w = z$,

$$\left.\frac{\partial^3 \underline{\alpha}_k(\mathbf{\Sigma})}{\partial\sigma_{lm}\partial\sigma_{uv}\partial\sigma_{wz}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}} = \frac{2}{(\lambda_l - \lambda_k)^3}\underline{e}_l.$$

2. Si $k = u \neq l = m = v = w = z$,

$$\left.\frac{\partial^3 \underline{\alpha}_k(\mathbf{\Sigma})}{\partial\sigma_{lm}\partial\sigma_{uv}\partial\sigma_{wz}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}} = -\frac{2}{(\lambda_l - \lambda_k)^3}\underline{e}_l.$$

3. Si $l = m = u = w = z = k \neq v$,

$$\left.\frac{\partial^3 \underline{\alpha}_k(\mathbf{\Sigma})}{\partial\sigma_{lm}\partial\sigma_{uv}\partial\sigma_{wz}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}} = -\frac{2}{(\lambda_v - \lambda_k)^3}\underline{e}_v.$$

4. Si $l = m = w \neq k = v \neq z = u$,

$$\left.\frac{\partial^3 \underline{\alpha}_k(\mathbf{\Sigma})}{\partial\sigma_{lm}\partial\sigma_{uv}\partial\sigma_{wz}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}} = -\frac{1}{(\lambda_l - \lambda_k)^2(\lambda_u - \lambda_k)}\underline{e}_l.$$

5. Si $k = u = w \neq l = m = v = z$,

$$\left.\frac{\partial^3 \underline{\alpha}_k(\mathbf{\Sigma})}{\partial\sigma_{lm}\partial\sigma_{uv}\partial\sigma_{wz}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}} = \frac{2}{(\lambda_l - \lambda_k)^3}\underline{e}_k.$$

6. Si $z \neq k = u \neq l = m = v = w \neq z$,

$$\left.\frac{\partial^3 \underline{\alpha}_k(\mathbf{\Sigma})}{\partial\sigma_{lm}\partial\sigma_{uv}\partial\sigma_{wz}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}} = -\frac{1}{(\lambda_l - \lambda_k)^2(\lambda_z - \lambda_k)}\underline{e}_z.$$

7. Si $l = m = u = w = k \neq v = z$,

$$\left.\frac{\partial^3 \underline{\alpha}_k(\mathbf{\Sigma})}{\partial\sigma_{lm}\partial\sigma_{uv}\partial\sigma_{wz}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}} = -\frac{2}{(\lambda_v - \lambda_k)^3}\underline{e}_k.$$

8. Si $v \neq k = l = m = w \neq u = z \neq v$,

$$\left.\frac{\partial^3 \underline{\alpha}_k(\boldsymbol{\Sigma})}{\partial\sigma_{lm}\partial\sigma_{uv}\partial\sigma_{wz}}\right|_{\boldsymbol{\Sigma}=\boldsymbol{\Lambda}} = \left[\frac{1}{(\lambda_u-\lambda_k)^2(\lambda_v-\lambda_k)} + \frac{1}{(\lambda_u-\lambda_k)(\lambda_v-\lambda_k)^2}\right]\underline{e}_v.$$

9. Si $k = l = u = w \neq m = v = z$,

$$\left.\frac{\partial^3 \underline{\alpha}_k(\boldsymbol{\Sigma})}{\partial\sigma_{lm}\partial\sigma_{uv}\partial\sigma_{wz}}\right|_{\boldsymbol{\Sigma}=\boldsymbol{\Lambda}} = \frac{9}{(\lambda_m-\lambda_k)^3}\underline{e}_m.$$

10. Si $z \neq k = l = u = w \neq m = v \neq z$,

$$\left.\frac{\partial^3 \underline{\alpha}_k(\boldsymbol{\Sigma})}{\partial\sigma_{lm}\partial\sigma_{uv}\partial\sigma_{wz}}\right|_{\boldsymbol{\Sigma}=\boldsymbol{\Lambda}} = \left[\frac{2}{(\lambda_m-\lambda_k)(\lambda_z-\lambda_k)^2} + \frac{1}{(\lambda_m-\lambda_k)^2(\lambda_z-\lambda_k)}\right]\underline{e}_z.$$

11. Si $l = u = w \neq m = v \neq z = k \neq l$,

$$\left.\frac{\partial^3 \underline{\alpha}_k(\boldsymbol{\Sigma})}{\partial\sigma_{lm}\partial\sigma_{uv}\partial\sigma_{wz}}\right|_{\boldsymbol{\Sigma}=\boldsymbol{\Lambda}} = -\frac{2}{(\lambda_l-\lambda_k)^2(\lambda_m-\lambda_k)}\underline{e}_l.$$

12. Si $k = l = u \neq m = z \neq v = w \neq k$,

$$\left.\frac{\partial^3 \underline{\alpha}_k(\boldsymbol{\Sigma})}{\partial\sigma_{lm}\partial\sigma_{uv}\partial\sigma_{wz}}\right|_{\boldsymbol{\Sigma}=\boldsymbol{\Lambda}} = \left[\frac{1}{(\lambda_v-\lambda_k)(\lambda_m-\lambda_k)^2} + \frac{1}{(\lambda_v-\lambda_k)^2(\lambda_m-\lambda_k)}\right]\underline{e}_k.$$

13. Si $v, w \neq l = u \neq m = z \neq v, w$ y $k = v \neq w$,

$$\left.\frac{\partial^3 \underline{\alpha}_k(\boldsymbol{\Sigma})}{\partial\sigma_{lm}\partial\sigma_{uv}\partial\sigma_{wz}}\right|_{\boldsymbol{\Sigma}=\boldsymbol{\Lambda}} = -\frac{1}{(\lambda_l-\lambda_k)(\lambda_m-\lambda_k)(\lambda_w-\lambda_k)}\underline{e}_w.$$

Demostración

Para demostrar estas expresiones es necesario derivar la expresión (4.10), obteniéndose la derivada de tercer orden de un autovector de $\boldsymbol{\Sigma}$, que viene dada por

$$\frac{\partial^3 \underline{\alpha}_k}{\partial\sigma_{lm}\partial\sigma_{uv}\partial\sigma_{wz}} = -\left[\frac{\partial^2\alpha_{km}}{\partial\sigma_{uv}\partial\sigma_{wz}}\underline{g}_l^k + \frac{\partial\alpha_{km}}{\partial\sigma_{uv}}\frac{\partial\underline{g}_l^k}{\partial\sigma_{wz}} + \frac{\partial\alpha_{km}}{\partial\sigma_{wz}}\frac{\partial\underline{g}_l^k}{\partial\sigma_{uv}} + \alpha_{km}\frac{\partial^2\underline{g}_l^k}{\partial\sigma_{uv}\partial\sigma_{wz}}\right.$$
$$\left. + (1-\delta_{lm})\left[\frac{\partial^2\alpha_{kl}}{\partial\sigma_{uv}\partial\sigma_{wz}}\underline{g}_m^k + \frac{\partial\alpha_{kl}}{\partial\sigma_{uv}}\frac{\partial\underline{g}_m^k}{\partial\sigma_{wz}} + \frac{\partial\alpha_{kl}}{\partial\sigma_{wz}}\frac{\partial\underline{g}_m^k}{\partial\sigma_{uv}} + \alpha_{kl}\frac{\partial^2\underline{g}_m^k}{\partial\sigma_{uv}\partial\sigma_{wz}}\right]\right].$$

Teniendo en cuenta el apartado **3** del lema 4.2.5 y las expresiones (4.6), (4.9), (4.11), (4.12) y el lema 4.2.12, y distinguiendo los casos según los valores de los índices, se obtienen las expresiones recogidas en el enunciado. Para ello es conveniente observar que para que $\left.\frac{\partial^3 \underline{\alpha}_k(\boldsymbol{\Sigma})}{\partial\sigma_{lm}\partial\sigma_{uv}\partial\sigma_{wz}}\right|_{\boldsymbol{\Sigma}=\boldsymbol{\Lambda}}$ conste de algún sumando no nulo, el índice k debe coincidir con alguno de los subíndices l, m, u, v, w, z.

■

4.2.4 Desarrollo en serie de un autovector muestral

A partir de los resultados previos, en este apartado se deduce la expresión del desarrollo de un autovector muestral en función de los autovectores poblacionales, para posteriormente calcular su esperanza y el sesgo condicionado.

Teorema 4.2.14 *Sea $\underline{X}$ un vector aleatorio p-dimensional con matriz de covarianzas no singular, $\boldsymbol{\Sigma}$, cuyos autovalores son $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ y autovectores ortonormales asociados, $\underline{\alpha}_1, ..., \underline{\alpha}_p$, respectivamente. Sean $\boldsymbol{\Lambda} = diag\{\lambda_1, \lambda_2, ..., \lambda_p\}$ y $\mathbf{A} = [\underline{\alpha}_1, ..., \underline{\alpha}_p]'$. Sea $\underline{X}_1, ..., \underline{X}_n$, una muestra aleatoria de $\underline{X}$, $\mathbf{S}$ la matriz de covarianzas muestrales con autovalores $\widetilde{\lambda}_1 \geq \widetilde{\lambda}_2 \geq \cdots \geq \widetilde{\lambda}_p$ y autovectores ortonormales asociados $\widetilde{\underline{\alpha}}_j$, $j = 1, ..., p$. Si λ_k es simple y $\mathbf{T} = (T_{lm}) = \mathbf{ASA}'$, para tamaño muestral suficientemente grande, $\widetilde{\underline{\alpha}}_k$ se puede expresar según el desarrollo en serie convergente*

$$
\begin{aligned}
\widetilde{\underline{\alpha}}_k = \underline{\alpha}_k &- \sum_{l \neq k} \frac{1}{\lambda_l - \lambda_k} T_{lk} \underline{\alpha}_l - \frac{1}{2} \sum_{l \neq k} \frac{1}{(\lambda_l - \lambda_k)^2} T_{lk}^2 \underline{\alpha}_k + \\
&+ \sum_{l \neq k} \frac{1}{(\lambda_l - \lambda_k)^2} T_{lk} \left[(T_{ll} - \lambda_l) - (T_{kk} - \lambda_k) \right] \underline{\alpha}_l + \\
&+ \sum_{l \neq k} \sum_{m \neq k,l} \frac{1}{(\lambda_l - \lambda_k)(\lambda_m - \lambda_k)} T_{lm} T_{mk} \underline{\alpha}_l + \\
&+ \sum_{l \neq k} \frac{1}{(\lambda_l - \lambda_k)^3} T_{lk} \left[\frac{3}{2} T_{lk}^2 - \left[(T_{ll} - \lambda_l) - (T_{kk} - \lambda_k) \right]^2 \right] \underline{\alpha}_l + \\
&+ \sum_{l \neq k} \sum_{m \neq k,l} \frac{1}{(\lambda_l - \lambda_k)^2 (\lambda_m - \lambda_k)} \cdot \\
&\cdot \left[T_{lm} T_{mk} \left[(T_{kk} - \lambda_k) - (T_{ll} - \lambda_l) \right] + \left[T_{km}^2 - T_{lm}^2 \right] T_{kl} \right] \underline{\alpha}_l + \\
&+ \sum_{l \neq k} \sum_{m \neq k,l} \frac{1}{(\lambda_l - \lambda_k)(\lambda_m - \lambda_k)^2} T_{mk} \cdot \\
&\cdot \left[\left[(T_{kk} - \lambda_k) - (T_{mm} - \lambda_m) \right] T_{lm} + \frac{1}{2} T_{km} T_{kl} \right] \underline{\alpha}_l - \\
&- \sum_{l \neq k} \sum_{m \neq k,l} \sum_{u \neq l,m,k} \frac{1}{(\lambda_l - \lambda_k)(\lambda_m - \lambda_k)(\lambda_u - \lambda_k)} T_{um} T_{uk} T_{lm} \underline{\alpha}_l + \\
&+ \sum_{l \neq k} \sum_{m \neq k,l} \frac{1}{(\lambda_l - \lambda_k)(\lambda_m - \lambda_k)^2} T_{km} T_{lk} T_{lm} \underline{\alpha}_k + \\
&+ \sum_{l \neq k} \frac{1}{(\lambda_l - \lambda_k)^3} T_{kl}^2 \left[(T_{ll} - \lambda_l) - (T_{kk} - \lambda_k) \right] \underline{\alpha}_k + R_k(\mathbf{S}) .
\end{aligned}
$$

donde $k = 1, ..., p$ *y* $R_k(\mathbf{S})$ *representa los términos en los que intervienen productos de cuatro o más factores del tipo* $T_{lm} - \lambda_l \delta_{lm}$.

Demostración

La expresión del desarrollo en serie de $\underset{\sim}{\widetilde{\alpha}}_k$, se puede obtener a través del cálculo previo del desarrollo de $\underset{\sim}{\widetilde{\beta}}_k(\mathbf{T})$, aplicando posteriormente la relación existente entre ambos autovectores. Para ello, se sustituyen en la expresión general (4.2), las derivadas de primer, segundo y tercer órdenes para el caso en que la matriz correspondiente sea diagonal, ya que $\mathbf{A\Sigma A}' = \mathbf{\Lambda}$,.

Si se denota por $\mathbf{D}_j(\mathbf{\Sigma})$ cada uno de los sumandos de dicho desarrollo, se puede expresar

$$\underset{\sim}{\widetilde{\beta}}_k(\mathbf{T}) = \mathbf{D}_1(\mathbf{\Lambda}) + \mathbf{D}_2(\mathbf{\Lambda}) + \mathbf{D}_3(\mathbf{\Lambda}) + \mathbf{D}_4(\mathbf{\Lambda}) + R_k(\mathbf{\Lambda}).$$

A continuación se estudia cada uno de los sumandos del desarrollo de $\underset{\sim}{\widetilde{\beta}}_k$. Para ello se aplican los lemas 4.2.5, 4.2.8 y 4.2.13, considerando el número de resultados análogos, a los señalados en los lemas correspondientes.

- $\mathbf{D}_1(\mathbf{\Lambda}) = \underset{\sim}{\alpha}_k(\mathbf{\Lambda}) = \underset{\sim}{e}_k$.
- $$\mathbf{D}_2(\mathbf{\Lambda}) = \sum_l \left.\frac{\partial \underset{\sim}{\alpha}_k(\mathbf{\Sigma})}{\partial \sigma_{ll}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}} (T_{ll} - \lambda_l) + \frac{1}{2}\sum_l \sum_{m \neq l} \left.\frac{\partial \underset{\sim}{\alpha}_k(\mathbf{\Sigma})}{\partial \sigma_{lm}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}} T_{lm} =$$
$$= -\sum_l \gamma_{lk} T_{lk} \underset{\sim}{e}_l.$$
- $$\mathbf{D}_3(\mathbf{\Lambda}) = \frac{1}{2}\sum_l \sum_{m \geq l} \sum_u \sum_{v \geq u} \left.\frac{\partial^2 \underset{\sim}{\alpha}_k(\mathbf{\Sigma})}{\partial \sigma_{lm} \partial \sigma_{uv}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}} (T_{lm} - \delta_{lm}\lambda_l)(T_{uv} - \delta_{uv}\lambda_u)$$

se puede descomponer de la siguiente forma:

$$\mathbf{D}_3(\mathbf{\Lambda}) = \frac{1}{2}[\mathbf{D}_{31}(\mathbf{\Lambda}) + \mathbf{D}_{32}(\mathbf{\Lambda}) + \mathbf{D}_{33}(\mathbf{\Lambda}) + \mathbf{D}_{34}(\mathbf{\Lambda})],$$

donde

$$\mathbf{D}_{31}(\mathbf{\Lambda}) = \frac{1}{4}\sum_l \sum_{m \neq l} \sum_u \sum_{v \neq u} \left.\frac{\partial^2 \underset{\sim}{\alpha}_k(\mathbf{\Sigma})}{\partial \sigma_{lm} \partial \sigma_{uv}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}} T_{lm} T_{uv} =$$
$$= \frac{1}{4}\left[8\sum_l \sum_{m \neq l} \gamma_{lk}\gamma_{mk} T_{lm} T_{mk} \underset{\sim}{e}_l - 4\sum_l \gamma_{lk}^2 T_{lk}^2 \underset{\sim}{e}_k\right] =$$
$$= 2\sum_l \gamma_{lk} \sum_{m \neq l} \gamma_{mk} T_{lm} T_{mk} \underset{\sim}{e}_l - \sum_l \gamma_{lk}^2 T_{lk}^2 \underset{\sim}{e}_k,$$

$$\mathbf{D}_{32}(\mathbf{\Lambda}) = \frac{1}{2}\sum_l \sum_{m \neq l} \sum_u \left.\frac{\partial^2 \underset{\sim}{\alpha}_k(\mathbf{\Sigma})}{\partial \sigma_{lm} \partial \sigma_{uu}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}} T_{lm}(T_{uu} - \lambda_u) =$$

$$= \frac{1}{2}\left[-2\sum_l \gamma_{lk}^2 \left(T_{kk}-\lambda_k\right) T_{lk}\underline{e}_l + 2\sum_l \gamma_{lk}^2 \left(T_{ll}-\lambda_l\right) T_{lk}\underline{e}_l\right] =$$

$$= \sum_l \gamma_{lk}^2 T_{lk}\left[\left(T_{ll}-\lambda_l\right)-\left(T_{kk}-\lambda_k\right)\right]\underline{e}_l,$$

$$\mathbf{D}_{33}\left(\mathbf{\Lambda}\right) = \frac{1}{2}\sum_l \sum_u \sum_{v\neq u} \left.\frac{\partial^2 \underline{\alpha}_k\left(\mathbf{\Sigma}\right)}{\partial\sigma_{ll}\partial\sigma_{uv}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}} \left(T_{ll}-\lambda_l\right) T_{uv} = \mathbf{D}_{32}\left(\mathbf{\Lambda}\right),$$

$$\mathbf{D}_{34}\left(\mathbf{\Lambda}\right) = \sum_l \sum_u \left.\frac{\partial^2 \underline{\alpha}_k\left(\mathbf{\Sigma}\right)}{\partial\sigma_{ll}\partial\sigma_{uu}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}} \left(T_{ll}-\lambda_l\right)\left(T_{uu}-\lambda_u\right) = \underline{0}.$$

Por lo tanto,

$$\mathbf{D}_3\left(\mathbf{\Lambda}\right) = \sum_l \gamma_{lk} \sum_{m\neq l} \gamma_{mk} T_{lm} T_{mk}\underline{e}_l - \frac{1}{2}\sum_l \gamma_{lk}^2 T_{lk}^2 \underline{e}_k +$$

$$+ \sum_l \gamma_{lk}^2 T_{lk}\left[\left(T_{ll}-\lambda_l\right)-\left(T_{kk}-\lambda_k\right)\right]\underline{e}_l.$$

Teniendo en cuenta la simetría entre los pares de índices (l,m), (u,v), (w,z), se puede escribir

$$\mathbf{D}_4\left(\mathbf{\Lambda}\right) = \frac{1}{6}\left[\mathbf{D}_{41}\left(\mathbf{\Lambda}\right) + 3\mathbf{D}_{42}\left(\mathbf{\Lambda}\right) + 3\mathbf{D}_{43}\left(\mathbf{\Lambda}\right) + \mathbf{D}_{44}\left(\mathbf{\Lambda}\right)\right],$$

donde

$$\mathbf{D}_{41}\left(\mathbf{\Lambda}\right) = \frac{1}{8}\left[\sum_l \sum_{m\neq l} \sum_u \sum_{v\neq u} \sum_w \sum_{z\neq w} \left.\frac{\partial^3 \underline{\alpha}_k\left(\mathbf{\Sigma}\right)}{\partial\sigma_{lm}\partial\sigma_{uv}\partial\sigma_{wz}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}} T_{lm}T_{uv}T_{wz}\right] =$$

$$= \frac{1}{8}\left[8\sum_l 9\gamma_{lk}^3 T_{lk}^3 \underline{e}_l + 24\sum_l \sum_{m\neq l}\left[2\gamma_{lk}^2\gamma_{mk} + \gamma_{lk}\gamma_{mk}^2\right] T_{km}^2 T_{kl}\underline{e}_l -\right.$$

$$-24\sum_l \sum_{m\neq l} 2\gamma_{lk}^2\gamma_{mk} T_{lm}^2 T_{kl}\underline{e}_l +$$

$$+24\sum_l \sum_{m\neq l}\left[\gamma_{lk}\gamma_{mk}^2 + \gamma_{lk}^2\gamma_{mk}\right] T_{km}T_{lk}T_{lm}\underline{e}_k -$$

$$\left. - 48\sum_l \sum_{m\neq l} \sum_{u\neq l,m,k} \gamma_{lk}\gamma_{mk}\gamma_{uk} T_{um}T_{uk}T_{lm}\underline{e}_l\right],$$

$$\mathbf{D}_{42}\left(\mathbf{\Lambda}\right) = \frac{1}{4}\left[\sum_l \sum_{m\neq l} \sum_u \sum_{v\neq u} \sum_w \left.\frac{\partial^3 \underline{\alpha}_k\left(\mathbf{\Sigma}\right)}{\partial\sigma_{lm}\partial\sigma_{uv}\partial\sigma_{ww}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}} T_{lm}T_{uv}\left(T_{ww}-\lambda_w\right)\right] =$$

$$= \frac{1}{4}\left[-8\sum_l \sum_{m\neq l} \gamma_{lk}^2\gamma_{mk}\left(T_{ll}-\lambda_l\right) T_{ku}T_{lu}\underline{e}_l +\right.$$

$$+4\sum_l 2\gamma_{lk}^3\left(T_{ll}-\lambda_l\right) T_{kl}^2\underline{e}_k -$$

$$\left. -8\sum_l \sum_{m\neq l} \gamma_{lk}\gamma_{mk}^2\left(T_{mm}-\lambda_m\right) T_{km}T_{ml}\underline{e}_l\right.$$

$$-4\sum_{l}2\gamma_{lk}^{3}\left(T_{kk}-\lambda_{k}\right)T_{kl}^{2}\underline{e}_{k}+$$

$$+8\sum_{l}\sum_{m\neq l}\left[\gamma_{lk}\gamma_{mk}^{2}+\gamma_{lk}^{2}\gamma_{mk}\right]\left(T_{kk}-\lambda_{k}\right)T_{lm}T_{mk}\underline{e}_{l}\Bigg],$$

$$\mathbf{D}_{43}\left(\mathbf{\Lambda}\right)=\frac{1}{2}\left[\sum_{l}\sum_{m\neq l}\sum_{u}\sum_{w}\left.\frac{\partial^{3}\underline{\alpha}_{k}\left(\mathbf{\Sigma}\right)}{\partial\sigma_{lm}\partial\sigma_{uu}\partial\sigma_{ww}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}}T_{lm}\left(T_{uu}-\lambda_{u}\right)\left(T_{ww}-\lambda_{w}\right)\right]=$$

$$=\frac{1}{2}\Bigg[4\sum_{l}2\gamma_{lk}^{3}\left(T_{ll}-\lambda_{l}\right)T_{lk}\left(T_{kk}-\lambda_{k}\right)\underline{e}_{l}-$$

$$-2\sum_{l}2\gamma_{lk}^{3}\left(T_{ll}-\lambda_{l}\right)^{2}T_{lk}\underline{e}_{l}-2\sum_{l}2\gamma_{lk}^{3}\left(T_{kk}-\lambda_{k}\right)^{2}T_{kl}\underline{e}_{l}\Bigg],$$

$$\mathbf{D}_{44}\left(\mathbf{\Lambda}\right)=\sum_{l}\sum_{u}\sum_{w}\left.\frac{\partial^{3}\underline{\alpha}_{k}\left(\mathbf{\Sigma}\right)}{\partial\sigma_{ll}\partial\sigma_{uu}\partial\sigma_{ww}}\right|_{\mathbf{\Sigma}=\mathbf{\Lambda}}\left(T_{ll}-\lambda_{l}\right)\left(T_{uu}-\lambda_{u}\right)\left(T_{ww}-\lambda_{w}\right)=$$

$$=\underline{0}.$$

Así,

$$\mathbf{D}_{4}\left(\mathbf{\Lambda}\right)=\frac{3}{2}\sum_{l}\gamma_{lk}^{3}T_{lk}^{3}\underline{e}_{l}+\frac{1}{2}\sum_{l}\sum_{m\neq l}\left[2\gamma_{lk}^{2}\gamma_{mk}+\gamma_{lk}\gamma_{mk}^{2}\right]T_{km}^{2}T_{kl}\underline{e}_{l}-$$

$$-\sum_{l}\sum_{m\neq l}\gamma_{lk}^{2}\gamma_{mk}T_{lm}^{2}T_{kl}\underline{e}_{l}+$$

$$+\frac{1}{2}\sum_{l}\sum_{m\neq l}\left[\gamma_{lk}\gamma_{mk}^{2}+\gamma_{lk}^{2}\gamma_{mk}\right]T_{km}T_{lk}T_{lm}\underline{e}_{k}-$$

$$-\sum_{l}\sum_{m\neq l}\sum_{u\neq l,m}\gamma_{lk}\gamma_{mk}\gamma_{uk}T_{um}T_{uk}T_{lm}\underline{e}_{l}-$$

$$-\sum_{l}\sum_{m\neq l}\gamma_{lk}^{2}\gamma_{mk}\left(T_{ll}-\lambda_{l}\right)T_{km}T_{lm}\underline{e}_{l}+$$

$$+\sum_{l}\gamma_{lk}^{3}\left(T_{ll}-\lambda_{l}\right)T_{kl}^{2}\underline{e}_{k}-\sum_{l}\sum_{m\neq l}\gamma_{lk}\gamma_{mk}^{2}\left(T_{mm}-\lambda_{m}\right)T_{km}T_{ml}\underline{e}_{l}$$

$$-\sum_{l}\gamma_{lk}^{3}\left(T_{kk}-\lambda_{k}\right)T_{kl}^{2}\underline{e}_{k}+$$

$$+\sum_{l}\sum_{m\neq l}\left[\gamma_{lk}\gamma_{mk}^{2}+\gamma_{lk}^{2}\gamma_{mk}\right]\left(T_{kk}-\lambda_{k}\right)T_{lm}T_{mk}\underline{e}_{l}-$$

$$+2\sum_{l}\gamma_{lk}^{3}\left(T_{ll}-\lambda_{l}\right)T_{lk}\left(T_{kk}-\lambda_{k}\right)\underline{e}_{l}-$$

$$-\sum_{l}\gamma_{lk}^{3}\left(T_{ll}-\lambda_{l}\right)^{2}T_{lk}\underline{e}_{l}-\sum_{l}\gamma_{lk}^{3}\left(T_{kk}-\lambda_{k}\right)^{2}T_{kl}\underline{e}_{l}\Bigg].$$

Por lo tanto, sumando los términos $\mathbf{D}_{1}\left(\mathbf{\Lambda}\right)$, $\mathbf{D}_{2}\left(\mathbf{\Lambda}\right)$, $\mathbf{D}_{3}\left(\mathbf{\Lambda}\right)$, $\mathbf{D}_{4}\left(\mathbf{\Lambda}\right)$ y $R_{k}\left(\mathbf{T}\right)$ se obtiene el desarrollo de $\underline{\alpha}_{k}\left(\mathbf{T}\right)$, a partir del cual se obtiene el

desarrollo de $\underline{\widetilde{\alpha}}_k$.

■

Nota 4.2.4 *El desarrollo dado por el teorema 4.2.14 es suficiente para poder calcular el sesgo condicionado de un autovector, aunque es posible obtener la expresión del desarrollo en serie en función de los elementos de* $\mathbf{S}-\mathbf{\Sigma}$, $s_{lm}-\sigma_{lm}$, *ya que* $\mathbf{T}-\mathbf{\Lambda}=\mathbf{A}\left(\mathbf{S}-\mathbf{\Sigma}\right)\mathbf{A}'$ *y, por tanto,*

$$T_{lm}-\delta_{lm}\lambda_l=\underline{\alpha}_l'\left(\mathbf{S}-\mathbf{\Sigma}\right)\underline{\alpha}_m=\sum_{q_1,q_2}\alpha_{l,q_1}\left(S_{q_1,q_2}-\sigma_{q_1,q_2}\right)\alpha_{m,q_2}.$$

A continuación se obtiene la esperanza de un autovector de $\mathbf{S}$, asociado a un autovalor simple, extendiendo así la aproximación dada en la expresión 1.6.

Teorema 4.2.15 *En las condiciones del teorema 4.2.14, se verifica que*

$$E\left[\underline{\widetilde{\alpha}}_k\right]=\left\{1-\frac{1}{2\left(n-1\right)}\sum_{l\neq k}\frac{\lambda_l\lambda_k}{\left(\lambda_l-\lambda_k\right)^2}\right\}\underline{\alpha}_k+R\left(n^{-2}\right) \tag{4.17}$$

donde $R\left(n^{-2}\right)$ *representa una serie convergente de términos de órdenes* n^{-2} *e inferiores.*

Demostración

Basta utilizar el desarrollo en serie de un autovector muestral dado en el teorema 4.2.14 y el lema 3.3.4.

■

4.3 Sesgo condicionado de un autovector muestral

El objetivo del presente capítulo, como ya se indicó en la introducción del mismo, es el cálculo de una aproximación del sesgo condicionado de un autovector de la matriz de covarianzas muestrales, el cual, se obtiene en esta sección, tras los resultados obtenidos en las secciones previas. En particular, dicho cálculo se basa en el desarrollo en serie obtenido en el teorema 4.2.14. De nuevo, la aproximación vendrá dada por $\mathcal{S}\left(\mathbf{x}_I;\underline{\widetilde{\alpha}}_k\right)=S\left(\mathbf{x}_I;\underline{\widetilde{\alpha}}_k-R_k\left(\mathbf{S}\right)\right)$.

Teorema 4.3.1 *Sea $\underline{X}$ un vector aleatorio p-dimensional con matriz de covarianzas no singular, $\boldsymbol{\Sigma}$, cuyos autovalores son $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ y autovectores ortonormales asociados, $\underline{\alpha}_1, ..., \underline{\alpha}_p$, respectivamente. Sean $\boldsymbol{\Lambda} = diag\{\lambda_1, \lambda_2, ..., \lambda_p\}$ y $\mathbf{A} = [\underline{\alpha}_1, ..., \underline{\alpha}_p]'$. Sea $\underline{X}_1, ..., \underline{X}_n$, una muestra aleatoria de $\underline{X}$, $\mathbf{S}$ la matriz de covarianzas muestrales con autovalores $\widetilde{\lambda}_1 \geq \widetilde{\lambda}_2 \geq \cdots \geq \lambda_p$ y autovectores ortonormales asociados $\widetilde{\underline{\alpha}}_j$, $j = 1, ..., p$. Si λ_k es simple y $\mathbf{T} = (T_{lm}) = \mathbf{ASA}'$, para tamaño muestral suficientemente grande, $\widetilde{\underline{\alpha}}_k$ se puede expresar según el desarrollo en serie convergente*

$$
\begin{aligned}
\mathcal{S}\left(\mathbf{x}_I; \widetilde{\underline{\alpha}}_k\right) = & -\frac{1}{n-1} \sum_{l \neq k} \frac{1}{\lambda_l - \lambda_k} \sum_{i \in I} y_{il} y_{ik} \underline{\alpha}_l \\
& + \frac{1}{(n-1)^2} \left[\sum_{l \neq k} \frac{1}{\lambda_l - \lambda_k} \left(r^2 \overline{y}_k^I \overline{y}_l^I - 2 \sum_{i \in I} y_{il} y_{ik} \right) \underline{\alpha}_l + \sum_{l \neq k} \frac{1}{(\lambda_l - \lambda_k)^2} \cdot \right. \\
& \left. \cdot \sum_{i \in I} y_{il} y_{ik} \left[\left(-r\lambda_l + \sum_{i \in I} y_{il}^2 \right) - \left(-r\lambda_k + \sum_{i \in I} y_{ik}^2 \right) \right] \underline{\alpha}_l \right] + \\
& + \sum_{l \neq k} \sum_{m \neq k,l} \frac{1}{(\lambda_l - \lambda_k)(\lambda_m - \lambda_k)} \sum_{i \in I} y_{il} y_{im} \sum_{i \in I} y_{im} y_{ik} \underline{\alpha}_l - \\
& - \frac{1}{2} \sum_{l \neq k} \frac{1}{(\lambda_l - \lambda_k)^2} \left[-r\lambda_l \lambda_k + \left(\sum_{i \in I} y_{il} y_{ik} \right)^2 \right] \underline{\alpha}_k + \\
& + \frac{1}{2} \sum_{l \neq k} \frac{\lambda_l \lambda_k}{(\lambda_l - \lambda_k)^3} \sum_{i \in I} y_{il} y_{ik} \underline{\alpha}_l + \\
& + \sum_{l \neq k} \sum_{m \neq k,l} \frac{\lambda_m}{(\lambda_l - \lambda_k)(\lambda_m - \lambda_k)} \sum_{i \in I} y_{ik} y_{il} \underline{\alpha}_l + \\
& + \frac{1}{2} \sum_{l \neq k} \sum_{m \neq k,l} \frac{\lambda_k \lambda_m}{(\lambda_l - \lambda_k)(\lambda_m - \lambda_k)^2} \sum_{i \in I} y_{ik} y_{il} \underline{\alpha}_l \\
& + \sum_{l \neq k} \frac{\lambda_k \lambda_l}{(\lambda_l - \lambda_k)^3} \left[\left(\sum_{i \in I} y_{il}^2 - r\lambda_l \right) - \left(\sum_{i \in I} y_{ik}^2 - r\lambda_k \right) \right] \underline{\alpha}_k + \\
& + O\left(n^{-3}\right)
\end{aligned}
\tag{4.18}
$$

Demostración

A partir del desarrollo dado en el teorema 4.2.14, es posible calcular el sesgo condicionado de un autovector muestral, hallando el sesgo condicionado de cada uno de los sumandos. Para obtener los primeros coeficientes, basta

hacer uso de los lemas 3.3.7, 3.3.14 y de la expresión (3.9),

$$\begin{aligned}
\mathcal{S}\left(\mathbf{x}_I;\widetilde{\underline{\alpha}}_k\right) = & -\frac{1}{n-1}\sum_{l\neq k}\frac{1}{\lambda_l-\lambda_k}\sum_{i\in I}y_{il}y_{ik}\underline{\alpha}_l + \\
& +\frac{1}{(n-1)^2}\left[r^2\sum_{l\neq k}\frac{1}{\lambda_l-\lambda_k}\overline{y}_k^I\overline{y}_l^I\underline{\alpha}_l+\sum_{l\neq k}\frac{1}{(\lambda_l-\lambda_k)^2}\cdot\right. \\
& \left.\cdot\sum_{i\in I}y_{il}y_{ik}\left[\left(-r\lambda_l+\sum_{i\in I}y_{il}^2\right)-\left(-r\lambda_k+\sum_{i\in I}y_{ik}^2\right)\right]\underline{\alpha}_l+\right. \\
& \left.+\sum_{l\neq k}\sum_{m\neq k,l}\frac{1}{(\lambda_l-\lambda_k)(\lambda_m-\lambda_k)}\left[\sum_{i\in I}y_{il}y_{im}\sum_{i\in I}y_{im}y_{ik}-\lambda_m\sum_{i\in I}y_{ik}y_{il}\right]\underline{\alpha}_l-\right. \\
& \left.-\frac{1}{2}\sum_{l\neq k}\frac{1}{(\lambda_l-\lambda_k)^2}\left[-r\lambda_l\lambda_k+\left(\sum_{i\in I}y_{il}y_{ik}\right)^2\right]\underline{\alpha}_k+\right. \\
& \left.+\sum_{l\neq k}\frac{1}{(\lambda_l-\lambda_k)^3}\sum_{i\in I}y_{il}y_{ik}\left(-2\lambda_l^2-2\lambda_k^2+\frac{9}{2}\lambda_l\lambda_k\right)\underline{\alpha}_l+\right. \\
& \left.+\frac{1}{2}\sum_{l\neq k}\sum_{m\neq k,l}\frac{\lambda_k\lambda_m}{(\lambda_l-\lambda_k)(\lambda_m-\lambda_k)^2}\sum_{i\in I}y_{ik}y_{il}\underline{\alpha}_l+\right. \\
& \left.+\sum_{l\neq k}\frac{\lambda_k\lambda_l}{(\lambda_l-\lambda_k)^3}\left[\left(\sum_{i\in I}y_{il}^2-r\lambda_l\right)-\left(\sum_{i\in I}y_{ik}^2-r\lambda_k\right)\right]\underline{\alpha}_k\right\}+ \\
& +O\left(n^{-3}\right).
\end{aligned}$$

Y agrupando adecuadamente se obtiene (4.18). ■

Este resultado se puede particularizar para $I=\{i\}$, donde $r=1$.

Corolario 4.3.2 *En las condiciones del teorema 4.3.1,*

$$\begin{aligned}
\mathcal{S}\left(\underline{x}_i;\widetilde{\underline{\alpha}}_k\right) = & -\frac{1}{n-1}\sum_{l\neq k}\frac{1}{\lambda_l-\lambda_k}y_{il}y_{ik}\underline{\alpha}_l \\
& +\frac{1}{(n-1)^2}\left[-\sum_{l\neq k}\frac{1}{\lambda_l-\lambda_k}y_{il}y_{ik}\underline{\alpha}_l+\sum_{l\neq k}\frac{1}{(\lambda_l-\lambda_k)^2}\cdot\right. \\
& \left.\cdot\sum_{i\in I}y_{il}y_{ik}\left[\left(y_{il}^2-\lambda_l\right)-\left(y_{ik}^2-\lambda_k\right)\right]\underline{\alpha}_l+\right. \\
& \left.+\sum_{l\neq k}\sum_{m\neq k,l}\frac{1}{(\lambda_l-\lambda_k)(\lambda_m-\lambda_k)}y_{il}y_{im}^2y_{ik}\underline{\alpha}_l-\right.
\end{aligned}$$

$$
\begin{aligned}
&-\frac{1}{2}\sum_{l\neq k}\frac{1}{(\lambda_l-\lambda_k)^2}\left[y_{il}^2y_{ik}^2-\lambda_l\lambda_k\right]\underline{\alpha}_k+\\
&+\frac{1}{2}\sum_{l\neq k}\frac{\lambda_l\lambda_k}{(\lambda_l-\lambda_k)^3}y_{il}y_{ik}\underline{\alpha}_l+\\
&+\sum_{l\neq k}\sum_{m\neq k,l}\frac{\lambda_m}{(\lambda_l-\lambda_k)(\lambda_m-\lambda_k)}y_{ik}y_{il}\underline{\alpha}_l+\\
&+\frac{1}{2}\sum_{l\neq k}\sum_{m\neq k,l}\frac{\lambda_k\lambda_m}{(\lambda_l-\lambda_k)(\lambda_m-\lambda_k)^2}y_{ik}y_{il}\underline{\alpha}_l\\
&+\sum_{l\neq k}\frac{\lambda_k\lambda_l}{(\lambda_l-\lambda_k)^3}\left[\left(y_{il}^2-\lambda_l\right)-\left(y_{ik}^2-\lambda_k\right)\right]\underline{\alpha}_k\Bigg]+\\
&+O\left(n^{-3}\right).
\end{aligned}
$$

Nota 4.3.1 *Debido a la igualdad entre los autovectores de* $\mathbf{S}$ *y* $\boldsymbol{\Sigma}$, *se verifica que*

$$
\begin{aligned}
E\left[\widehat{\underline{\alpha}}_k\right] &= E\left[\widetilde{\underline{\alpha}}_k\right],\\
S\left(\underline{x}_I;\widehat{\underline{\alpha}}_k\right) &= S\left(\mathbf{x}_I;\widetilde{\underline{\alpha}}_k\right).
\end{aligned}
$$

Se puede establecer la siguiente relación con la función de influencia de un autovector de la matriz de covarianzas:

Corolario 4.3.3 *En las condiciones del teorema 4.3.1, si* F *es la función de distribución del vector* $\underline{X}$,

$$
\begin{aligned}
\mathcal{S}\left(\mathbf{x}_I;\widetilde{\underline{\alpha}}_k\right) &= \frac{1}{n-r}I\left(\mathbf{x}_I;\underline{\alpha}_k,F\right)+O\left(n^{-2}\right)=\\
&= \frac{1}{n-r}\sum_{i\in I}I\left(\underline{x}_i;\underline{\alpha}_k,F\right)+O\left(n^{-2}\right) \qquad y\\
\mathcal{S}\left(\underline{x}_i;\widetilde{\underline{\alpha}}_k\right) &= \frac{1}{n-1}I\left(\underline{x}_i;\underline{\alpha}_k,F\right)+O\left(n^{-2}\right).
\end{aligned}
$$

Nota 4.3.2 *Dado que el sesgo condicionado de un vector es otro vector de la misma dimensión, para evaluar la influencia que ejerce una observación sobre dicho vector es necesario el uso de normas vectoriales. No obstante, tiene sentido plantearse qué tipo de observaciones proporcionan un valor nulo del sesgo condicionado, cualquiera que sea el tamaño muestral. Esto ocurrirá, cuando cada uno de los términos de orden* n^{-j} *sean nulos,* $j=1,2,\ldots$.

En el caso de Análisis de Influencia individual, para una observación $\underline{x}_i$, *el primer término es nulo cuando la función de influencia es nula y esto ocurre cuando:*

1. *El valor de la k-ésima componente principal de dicha observación es nula. Esta situación tiene una doble interpretación:*

   (**a**) *El valor de la k-ésima componente principal coincide con su valor esperado.*

   (**b**) *La proyección de $\underline{x}_i$ sobre $\underline{\alpha}_k$ es la misma que la de $\underline{\mu}$ sobre $\underline{\alpha}_k$.*

   Además, en tal caso,

$$S\left(\underline{x}_i;\widetilde{\underline{\alpha}}_k\right) \simeq \frac{1}{(n-1)^2}\sum_{l\neq k}\frac{\lambda_l\lambda_k}{(\lambda_l-\lambda_k)^2}\left[\frac{1}{2}+\frac{1}{\lambda_l-\lambda_k}\left[y_{il}^2-\lambda_l+\lambda_k\right]\right]\underline{\alpha}_k =$$

$$= \frac{1}{(n-1)^2}\sum_{l\neq k}\frac{\lambda_l\lambda_k}{(\lambda_l-\lambda_k)^2}\left[\frac{y_{il}^2}{\lambda_l-\lambda_k}-\frac{1}{2}\right]\underline{\alpha}_k.$$

2. *Cuando el valor del resto de las componentes principales es nula. Además de coincidir el valor de dichas componentes principales con su valor esperado, se deduce que la dirección de $\underline{x}_i-\underline{\mu}$ coincide con la de $\underline{\alpha}_k$ y por lo tanto $|y_{ik}| = \left\|\underline{x}_i-\underline{\mu}\right\|$. Además,*

$$S\left(\underline{x}_i;\widetilde{\underline{\alpha}}_k\right) \simeq -\frac{1}{(n-1)^2}\sum_{l\neq k}\frac{\lambda_l\lambda_k}{(\lambda_l-\lambda_k)^2}\left[\frac{1}{2}+\frac{1}{\lambda_l-\lambda_k}\left[-\lambda_l-y_{ik}^2+\lambda_k\right]\right]\underline{\alpha}_k =$$

$$= -\frac{1}{(n-1)^2}\sum_{l\neq k}\frac{\lambda_l\lambda_k}{(\lambda_l-\lambda_k)^2}\left[\frac{1}{2}+\frac{y_{ik}^2}{\lambda_l-\lambda_k}\right]\underline{\alpha}_k.$$

*En ambos casos, **1** y **2**, la dirección del sesgo condicionado está determinada principalmente por el autovector estudiado, y puede ocurrir que no sea nulo. Por lo tanto, el sesgo condicionado y la función de influencia, en general, no se anulan para los mismos valores del vector $\underline{x}_i$.*

Capítulo 5

MEDIDAS DE INFLUENCIA

5.1 Introducción

La utilización práctica del sesgo condicionado como medida de influencia requiere la solución de algunos problemas. Entre éstos, se ha de citar que el sesgo condicionado es un parámetro poblacional, por lo que en la práctica, es necesario obtener un estimador.

Por otra parte, tiene la misma dimensión que el estadístico sobre el que se realiza el análisis. Por tanto, cuando éste es un vector o una matriz, se han de utilizar métricas adecuadas para cuantificar la influencia. En particular, en el Análisis de Componentes Principales, cuando el interés se centra en un autovector, será necesaria la utilización de métricas vectoriales. Además, al ser la reducción de la dimensionalidad uno de los objetivos fundamentales del Análisis de Componentes Principales, el objetivo se centrará en un conjunto de autovalores y autovectores. Para los primeros, será conveniente obtener el sesgo condicionado sobre la suma de los mismos, como diagnóstico de influencia sobre la variabilidad total explicada por las correspondientes componentes principales. Para un conjunto de autovectores, éstos pueden considerarse como una matriz, siendo necesario el uso de normas matriciales.

Este capítulo aborda la problemática anteriormente descrita, bajo la hipótesis de normalidad. En la sección 5.2, se propone una estimación del sesgo condicionado de los autovalores y autovectores de la matriz de covarianzas muestrales. Y en las sucesivas, se presentan distintas medidas de influencia para un autovalor, un autovector, un conjunto de autovalores y un conjunto de autovectores.

Es conveniente señalar que el Análisis de Influencia dentro del Análisis de Componentes Principales, llevado a cabo a través del sesgo condicionado, está desarrollado cuando los autovalores poblacionales correspondientes son

simples, por lo que es conveniente verificar dicha hipótesis previo al Análisis de Influencia.

Los distintos diagnósticos propuestos se hacen desde el punto de vista del análisis de influencia múltiple, para un conjunto de observaciones simultáneamente, siendo directa la traslación al estudio de la influencia individual, $I = \{i\}$.

5.2 Estimación del sesgo condicionado en el ACP

Las expresiones (3.13) y (4.18) para $\mathcal{S}\left(\mathbf{x}_I; \widetilde{\lambda}_k\right)$ y $\mathcal{S}\left(\mathbf{x}_I; \underset{\sim}{\widetilde{\alpha}}_k\right)$ dependen de parámetros desconocidos. Por ello, para estudiar la influencia de una observación o de un conjunto de observaciones en los estimadores de los parámetros de interés, es necesario obtener estimadores de ambos sesgos condicionados.

Muñoz Pichardo y otros [50], para un estadístico T tal que $E[T]$ no depende del tamaño muestral, proponen $T - T^{(I)}$ como estimador del sesgo condicionado de T, motivado por verificarse que

$$E\left[T - T^{(I)} \mid \mathbf{X}_I = \mathbf{x}_I\right] = S\left(\mathbf{x}_I; T\right).$$

En el caso en el que T represente un autovalor, $\widetilde{\lambda}_k$, o un autovector, $\underset{\sim}{\widetilde{\alpha}}_k$, la igualdad anterior no se verifica, al ser estimadores sesgados. No obstante, a partir de los momentos recogidos en el teorema 1.2.5, se observa que para ambos se cumple que

$$\begin{aligned} E\left[T - T^{(I)} \mid \mathbf{X}_I = \mathbf{x}_I\right] &= \left\{E\left[T \mid \mathbf{X}_I = \mathbf{x}_I\right] - E\left[T\right]\right\} + \left\{E\left[T\right] - E\left[T^{(I)}\right]\right\} = \\ &= S\left(\mathbf{x}_I; T\right) + A_n. \end{aligned}$$

donde $\lim_{n \to +\infty} A_n = 0$.

Así, siguiendo la línea marcada por Muñoz Pichardo y otros [50], podrían considerarse los estimadores del sesgo condicionado de un autovalor de $\mathbf{S}$, $\widetilde{\lambda}_k$, y de un autovector unitario asociado, $\underset{\sim}{\widetilde{\alpha}}_k$, dados por

$$\widetilde{S}\left(\mathbf{x}_I;\widetilde{\lambda}_k\right) = \widetilde{\lambda}_k - \widetilde{\lambda}_k^{(I)} \qquad \text{y}$$
$$\widetilde{S}\left(\mathbf{x}_I;\underline{\widetilde{\alpha}}_k\right) = \underline{\widetilde{\alpha}}_k - \underline{\widetilde{\alpha}}_k^{(I)},$$

los cuales representan el cambio real observado tras eliminar un conjunto de observaciones.

Se ha comentado en capítulos anteriores la problemática que surge con la comparación directa de los autovalores y autovectores de la matriz de covarianzas muestrales, utilizando la muestra completa y con la obtenida al omitir un conjunto de observaciones. Estos problemas son, por un lado, la dificultad práctica para asegurar la correspondencia en la ordenación de los autovalores, por otro lado, la determinación del sentido de los autovectores asociados y finalmente la complejidad computacional que surge en la resolución de los problemas de autovalores. Los dos primeros problemas quedan solucionados para tamaño muestral suficientemente grande: el orden de los autovalores se conserva y los signos de las componentes tras la omisión son los mismos del autovector calculado al considerar el total de las observaciones. Sin embargo, en la práctica es difícil saber cuándo se puede considerar que el tamaño muestral es suficientemente grande. Por ello, para la estimación del sesgo condicionado, es más conveniente el uso de aproximaciones basadas en los desarrollos recogidos en el teorema 2.4.7, lo que también soluciona el problema computacional citado. Así, los estimadores que se proponen del sesgo condicionado de un autovalor y de un autovector, se obtienen truncando los desarrollos en serie, es decir,

$$\begin{aligned}
\widetilde{S}_t\left(\mathbf{x}_I;\widetilde{\lambda}_k\right) &= \frac{1}{n-r-1}\widetilde{\nu}_{I,k} - \frac{1}{2\left(n-r-1\right)^2}\widetilde{\pi}_{I,k}, \\
\widetilde{S}_t\left(\mathbf{x}_I;\underline{\widetilde{\alpha}}_k\right) &= \frac{1}{n-r-1}\underline{\widetilde{\beta}}_{I,k} - \frac{1}{2\left(n-r-1\right)^2}\underline{\widetilde{\gamma}}_{I,k},
\end{aligned} \tag{5.1}$$

donde los coeficientes $\widetilde{\nu}_{I,k}$, $\widetilde{\pi}_{I,k}$, $\underline{\widetilde{\beta}}_{I,k}$ y $\underline{\widetilde{\gamma}}_{I,k}$ se describen en el teorema 2.4.7.

El cálculo del sesgo condicionado de un autovalor o un autovector, se basa en la hipótesis de simplicidad del autovalor poblacional correspondiente. En el cálculo del estimador se impone la condición de simplicidad del autovalor muestral para que tengan sentido los factores del tipo $\left(\widetilde{\lambda}_j - \widetilde{\lambda}_k\right)^{-1}$.

5.3 Medidas de influencia para un autovalor

El análisis de influencia sobre los autovalores de la matriz de covarianzas es de gran importancia en el Análisis de Componentes Principales, ya que en función de éstos se determina el número de componentes principales que se retienen cuando se realiza una reducción de la dimensión del vector aleatorio bajo estudio. Luego, una observación que influya marcadamente sobre un autovalor puede modificar la dimensión del espacio escogido.

En la sección anterior se ha propuesto un estimador del sesgo condicionado de un autovalor muestral. El único inconveniente que plantea su uso, como diagnóstico de influencia, es el signo, inconveniente que puede evitarse considerando su valor absoluto o su cuadrado.

Por otra parte, con el objeto de cuantificar adecuadamente la influencia en función de una escala de referencia, conviene considerar su relativización respecto a alguna magnitud de interés, de forma que la medida resultante pueda interpretarse en términos relativos. Dos magnitudes que pueden jugar este papel son: el propio autovalor y una estimación de la varianza o desviación típica del mismo.

Al ser $var\left(\widetilde{\lambda}_k\right)$ un parámetro desconocido, para utilizarla como elemento de relativización en las medidas de influencia, es necesario realizar una estimación. Para ello se propone, basándose en la aproximación que se deriva del teorema 1.2.5, el estimador

$$\widetilde{var}\left(\widetilde{\lambda}_k\right) = \frac{2\widetilde{\lambda}_k^2}{n-1}\left[1 - \frac{1}{n-1}\sum_{j\neq k}\frac{\widetilde{\lambda}_j^2}{\left(\widetilde{\lambda}_j - \widetilde{\lambda}_k\right)^2}\right]. \tag{5.2}$$

Siguiendo la línea marcada en el comienzo de esta sección, se proponen los siguientes diagnósticos de influencia para medir el efecto que ejerce el conjunto de observaciones, $\underline{\mathbf{x}}_I$, $I = \{i_1, ..., i_r\}$, sobre un autovalor simple de $\mathbf{S}$, $\widetilde{\lambda}_k$.

1. Medida M_1 asociada al k-ésimo autovalor debida al conjunto de observaciones subindicadas por $I = \{i_1, ..., i_r\}$,

$$M_1\left(\widetilde{\lambda}_k; I\right) = \frac{\left|\widetilde{S}_t\left(\mathbf{x}_I; \widetilde{\lambda}_k\right)\right|}{\widetilde{\lambda}_k}.$$

Representa la razón de cambio experimentado por la varianza de la k-ésima componente principal muestral, al omitir las observaciones de índices en I. En consecuencia, esta medida cuantifica la influencia

en términos de la variabilidad explicada por la k-ésima componente principal.

Pack y otros [53] utilizan una medida similar a M_1, en el caso de las componentes principales basadas en la matriz de correlación. En esta medida, el numerador es la diferencia entre $\widetilde{\lambda}_k$ y $\widetilde{\lambda}_k^{(I)}$, lo que es equivalente a $\widetilde{S}\left(\mathbf{x}_I; \widetilde{\lambda}_k\right)$. El uso de $\widetilde{S}_t\left(\mathbf{x}_I; \widetilde{\lambda}_k\right)$ evita los problemas derivados del uso del autovalor tras la omisión de las observaciones.

2. Medida M_2 asociada al k-ésimo autovalor debida al conjunto de observaciones subindicadas por $I = \{i_1, ..., i_r\}$,

$$M_2\left(\widetilde{\lambda}_k; I\right) = \frac{\left|\widetilde{S}_t\left(\mathbf{x}_I; \widetilde{\lambda}_k\right)\right|}{\sum_j \widetilde{\lambda}_j}.$$

Representa la razón del cambio experimentado por la variabilidad explicada de la k-ésima componente principal, al omitir las observaciones de índices en I, en términos de la variabilidad total explicada.

3. Medida M_3 asociada al k-ésimo autovalor debida al conjunto de observaciones subindicadas por $I = \{i_1, ..., i_r\}$,

$$M_3\left(\widetilde{\lambda}_k; I\right) = \frac{\left|\widetilde{S}_t\left(\mathbf{x}_I; \widetilde{\lambda}_k\right)\right|}{\left[\widetilde{var}\left(\widetilde{\lambda}_k\right)\right]^{\frac{1}{2}}}.$$

Representa la razón del cambio experimentado por la variabilidad explicada por la k-ésima componente principal, al omitir las observaciones de índices en I, en términos de la desviación típica de la misma.

Este patrón de estandarización, basado en medir la influencia relativizada con las unidades de la desviación típica del estadístico en cuestión, es utilizado con frecuencia en el Análisis de Influencia en gran número de técnicas estadísticas.

Nota 5.3.1 *Fijado el autovalor, las tres medidas propuestas, M_1, M_2 y M_3 son proporcionales, y, por ello, toman valores elevados para los mismos conjuntos de observaciones.*

Sin embargo, cada uno de los diagnósticos tiene su propia interpretación. La relativización realizada en M_2, es la misma para todos los autovalores. Teniendo en cuenta que, en términos absolutos, la perturbación que experimenta un autovalor tras omitir un conjunto de observaciones en su cálculo,

no se refleja de la misma forma para autovalores pequeños que para autovalores grandes, M_2 dará mayor relevancia a la influencia sobre autovalores asociados a las primeras componentes principales, que son las que generalmente sufren mayores cambios en términos absolutos. M_1 y M_3, desde dos criterios diferentes, realizan una relativización distinta de la variación de cada autovalor, tratando de equiparar la relevancia en todas las componentes principales. De esta forma es posible la comparación de la influencia que ejerce una misma observación sobre distintos autovalores, reflejando sobre cuál de ellos es mayor el efecto. La medida M_3 proporciona mayor relevancia a la influencia experimentada por autovalores muestrales con menor varianza.

Además, para tamaño muestral suficientemente grande, fijada una observación, el orden de las medidas M_1 y M_3 para los distintos autovalores es el mismo, ya que, en tal caso, $\left[\widetilde{var}\left(\widetilde{\lambda}_k\right)\right]^{\frac{1}{2}} \simeq \sqrt{2\left(n-1\right)^{-1}}\,\widetilde{\lambda}_k$, que es, salvo constante multiplicativa, el mismo término que se utiliza en la relativización dada en M_1.

Nota 5.3.2 *El numerador de M_1, M_2 y M_3, tiene como término dominante $\widetilde{\nu}_{I,k} = \sum_{i\in I}\left(\widetilde{y}_{ik}^2 - \widetilde{\lambda}_k\right)$, que será el que determine en mayor medida la influencia ejercida por cada observación. En el caso de análisis de influencia individual, existirán valores destacadamente altos para los diagnósticos de influencia si el valor de la componente principal correspondiente, $\widetilde{y}_{ik}$, se desvía marcadamente de la desviación típica de la misma, $\widetilde{\lambda}_k^{\frac{1}{2}}$. Aún no ocurriendo esto, puede que las medidas propuestas tomen valores elevados debido al término $\widetilde{\pi}_{i,k} = -2\widetilde{y}_{ik}^2\left[1+\sum_{j\neq k}\widetilde{y}_{ij}^2\left(\widetilde{\lambda}_j-\widetilde{\lambda}_k\right)^{-1}\right]$, o bien, por la presencia de valores de componentes principales elevados o bien, ante la existencia de autovalores próximos a $\widetilde{\lambda}_k$. Sin embargo, los términos del sumatorio pueden ser tanto positivos como negativos y hay posibilidad de compensaciones.*

5.4 Medidas de influencia para un autovector

La importancia del Análisis de Influencia sobre los autovectores de la matriz de covarianzas se debe al interés por conocer la variación que experimenta el espacio de las componentes principales que se retienen en el estudio.

En este análisis, tanto si se realiza a través de las funciones de influencia como a través del sesgo condicionado, las herramientas obtenidas tienen la

misma dimensión que el vector aleatorio bajo estudio. Para que sea posible la comparación de la influencia ejercida por distintas observaciones o por conjuntos de ellas, es necesario el uso de medidas unidimensionales. En la bibliografía, se han propuesto medidas basadas en la función de influencia muestral, sobre las que se aplica la norma euclídea o, equivalentemente, el coseno formado por el autovector obtenido mediante la muestra completa y el obtenido tras la omisión de una observación. Algunas trabajos realizados bajo dichos enfoques son los de Pack y otros [53], Brooks [7], Jolliffe [35].

El caso de análisis de influencia individual puede extenderse directamente al caso de omisión múltiple, calculando dichas medidas bajo la omisión de un conjunto de observaciones.

En esta sección se proponen distintos diagnósticos para evaluar la influencia de un conjunto de observaciones sobre un autovector. Hay que tener en cuenta que las medidas anteriormente citadas tienen el inconveniente de basarse en los autovectores obtenidos tras la omisión de observaciones, lo que plantea, además del problema computacional de tener que resolver un problema de autovalores para cada observación o conjunto de observaciones, la dificultad de la elección adecuada del autovector asociado tras la omisión (problema de ordenación de autovalores) y del sentido de dicho autovector. Por ello, en esta memoria se propone, el uso de medidas unidimensionales basadas en la estimación truncada del sesgo condicionado. Las propuestas se fundamentarán en un tipo de normas vectoriales, las $(\mathbf{Q}, c)$-normas, citadas en la sección 1.1.4 del capítulo 1.

Con el objetivo de utilizar una expresión abreviada de $\widetilde{S}_t\left(\mathbf{x}_I; \underline{\widetilde{\alpha}}_k\right)$, se denotarán por

$$\underline{\widetilde{\beta}}_{I,k} = \sum_j c_{jk}\left(I\right) \underline{\widetilde{\alpha}}_j \quad \text{y} \quad \underline{\widetilde{\gamma}}_{I,k} = \sum_j d_{jk}\left(I\right) \underline{\widetilde{\alpha}}_j,$$

donde

$$\begin{aligned} c_{kk}\left(I\right) &= 0, \\ d_{kk}\left(I\right) &= -\sum_{j \neq k} \frac{1}{\left(\widetilde{\lambda}_j - \widetilde{\lambda}_k\right)^2} \left[\sum_{i \in I} \widetilde{y}_{ij} \widetilde{y}_{ik}\right]^2. \end{aligned}$$

y para $j \neq k$,

$$\begin{aligned} c_{jk}\left(I\right) &= -\sum_{i \in I} \frac{\widetilde{y}_{ij} \widetilde{y}_{ik}}{\widetilde{\lambda}_j - \widetilde{\lambda}_k}, \\ d_{jk}\left(I\right) &= 2 \left[\frac{1}{\widetilde{\lambda}_j - \widetilde{\lambda}_k} \sum_{i_1 \in I} \widetilde{y}_{i_1 j} \sum_{i_2 \in I} \widetilde{y}_{i_2 k} - \frac{1}{\widetilde{\lambda}_j - \widetilde{\lambda}_k} \sum_{i \in I} \widetilde{y}_{ij} \widetilde{y}_{ik} + \right. \end{aligned}$$

$$+\sum_{l\neq k}\frac{1}{\left(\widetilde{\lambda}_j-\widetilde{\lambda}_k\right)\left(\widetilde{\lambda}_l-\widetilde{\lambda}_k\right)}\sum_{i_1\in I}\widetilde{y}_{i_1l}\widetilde{y}_{i_1k}\sum_{i_2\in I}\widetilde{y}_{i_2j}\widetilde{y}_{i_2l}-$$

$$-\frac{1}{\left(\widetilde{\lambda}_j-\widetilde{\lambda}_k\right)^2}\sum_{i_1\in I}\widetilde{y}_{i_1j}\widetilde{y}_{i_1k}\sum_{i_2\in I}\widetilde{y}_{i_2k}^2\Bigg].$$

En general, se denotará por

$$e_{jk}(I)=\frac{1}{n-r-1}\left[c_{jk}(I)-\frac{1}{2(n-r-1)}d_{jk}(I)\right].$$

Así, la estimación truncada propuesta del sesgo condicionado de un autovector de $\mathbf{S}$, dada en (5.1), se puede expresar de la forma,

$$\widetilde{S}_t(\mathbf{x}_I;\underline{\widetilde{\alpha}}_k)=\sum_j e_{jk}(I)\cdot\underline{\widetilde{\alpha}}_j. \tag{5.3}$$

Se proponen los siguientes diagnósticos para medir la influencia que ejerce un conjunto de observaciones, $\mathbf{x}_I$, $I=\{i_1,...,i_r\}$, sobre un autovector de $\mathbf{S}$, asociado a un autovalor simple:

1. Medida M_4 asociada al k-ésimo autovector debida al conjunto de observaciones subindicadas por $I=\{i_1,...,i_r\}$,

$$M_4(\underline{\widetilde{\alpha}}_k,I)=\left\|\widetilde{S}_t(\mathbf{x}_I;\underline{\widetilde{\alpha}}_k)\right\|_2=\left\|\widetilde{S}_t(\mathbf{x}_I;\underline{\widetilde{\alpha}}_k)\right\|_{(\mathbf{I}_p,1)}.$$

   Se observa que, debido a la ortogonalidad de los autovectores $\underline{\widetilde{\alpha}}_1,...,\underline{\widetilde{\alpha}}_p$, se verifica que

$$M_4^2(\underline{\widetilde{\alpha}}_k,I)=\sum_j e_{jk}^2(I),$$

   por lo que , salvo constante, cada sumando de $M_4^2(\underline{\widetilde{\alpha}}_k,I)$ mide la influencia sobre $\underline{\widetilde{\alpha}}_k$ en la dirección de cada componente principal, $\underline{\widetilde{\alpha}}_j$.

2. Medida M_5 asociada al k-ésimo autovector debida al conjunto de observaciones subindicadas por $I=\{i_1,...,i_r\}$,

$$M_5(\underline{\widetilde{\alpha}}_k,I)=\left\|\widetilde{S}_t(\mathbf{x}_I;\underline{\widetilde{\alpha}}_k)\right\|_{(\mathbf{S},\frac{1}{n-1})}.$$

   Este diagnóstico se puede expresar como

$$M_5^2(\underline{\widetilde{\alpha}}_k,I)=\sum_j\widetilde{\lambda}_j e_{jk}^2(I),$$

ponderando la influencia en la dirección de cada autovector, $\widetilde{\underline{\alpha}}_j$, mediante el autovalor asociado, $\widetilde{\lambda}_j$, dando así mayor relevancia a las direcciones correspondientes a componentes principales de mayor variabilidad.

La medida $M_5\left(\widetilde{\underline{\alpha}}_k, I\right)$ puede considerarse una aproximación de $\left\|\widetilde{S}\left(\mathbf{x}_I; \widetilde{\underline{\alpha}}_k\right)\right\|_{\left(\mathbf{S}, \frac{1}{n-1}\right)}$, medida de interés en el Análisis de Influencia en componentes principales, ya que

$$\begin{aligned}\left\|\widetilde{S}\left(\mathbf{x}_I; \widetilde{\underline{\alpha}}_k\right)\right\|^2_{\left(\mathbf{S}, \frac{1}{n-1}\right)} &= (n-1)\left(\widetilde{\underline{\alpha}}_k - \widetilde{\underline{\alpha}}_k^{(I)}\right)' \mathbf{S}\left(\widetilde{\underline{\alpha}}_k - \widetilde{\underline{\alpha}}_k^{(I)}\right) = \\ &= \left(\widetilde{\underline{\alpha}}_k - \widetilde{\underline{\alpha}}_k^{(I)}\right)' \widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}\left(\widetilde{\underline{\alpha}}_k - \widetilde{\underline{\alpha}}_k^{(I)}\right) = \\ &= \left\|\widetilde{\underline{y}}_k - \widetilde{\underline{y}}_k^{(I)}\right\|_2^2\end{aligned}$$

donde $\widetilde{\underline{y}}_k = \widetilde{\mathbf{X}}\widetilde{\underline{\alpha}}_k$ e $\widetilde{\underline{y}}_k^{(I)} = \widetilde{\mathbf{X}}\widetilde{\underline{\alpha}}_k^{(I)}$. Esta medida de influencia es un caso particular de la propuesta por Critchley [15] para Análisis de Influencia sobre un conjunto de componentes principales, la cual se recoge en el apartado 5.6 de este capítulo. Como alternativa de esta medida, debido de nuevo a los inconvenientes reiterados que provoca en la práctica el uso de $\widetilde{\underline{\alpha}}_k - \widetilde{\underline{\alpha}}_k^{(I)}$, se opta por $M_5\left(\widetilde{\underline{\alpha}}_k, I\right)$.

Nota 5.4.1 *Para tamaño muestral suficientemente grande, en el caso de análisis de influencia individual, el término dominante de* $M_4^2\left(\widetilde{\underline{\alpha}}_k, I\right)$ *es*

$$\frac{1}{(n-r-1)^2}\sum_j c_{jk}^2(I) = \frac{1}{(n-r-1)^2}\sum_{j\neq k}\frac{\widetilde{y}_{ij}^2\widetilde{y}_{ik}^2}{\left(\widetilde{\lambda}_j - \widetilde{\lambda}_k\right)^2}.$$

Los casos en los que una observación afecta de forma destacada en el cálculo de los coeficientes de la k-ésima componente principal, es decir, sobre $\widetilde{\underline{\alpha}}_k$, *se recoge en los siguientes puntos:*

- *El valor de alguna de las componentes principales para la i-ésima observación es elevado, en relación al resto de las componentes.*
- *Existen autovalores muy próximos, tanto que, a pesar de ser el tamaño muestral grande,* $\left[(n-r-1)\left(\widetilde{\lambda}_j - \widetilde{\lambda}_k\right)\right]^{-1}$ *sea considerable.*
- *No dándose ninguna de las condiciones anteriores, se produce una combinación de los efectos moderados de ambos, de forma que se potencia el impacto conjunto.*

Nota 5.4.2 *Las medidas M_4 y M_5 pueden tomar valores altos para distintos conjuntos de observaciones debido a la diferente definición de cada una de ellas. Cada uno de los diagnósticos debe interpretarse en términos de lo que representan y comparar la eficacia de alguno de ellos frente a los otros puede conducir a equívoco, en tanto que cada uno mide la influencia sobre un autovector muestral evaluado de distinta forma la influencia en cada dirección.*

5.5 Medidas de influencia para un conjunto de autovalores

El estudio de la influencia en el Análisis de Componentes Principales se centra, fundamentalmente, en las componentes principales seleccionadas, que son las que se usarán en estudios posteriores. Por lo tanto, la existencia de observaciones altamente influyentes en las componentes principales no seleccionadas, no suele ser objeto de interés del Análisis de Influencia.

Cuando el Análisis de Componentes Principales se lleva a cabo con el objetivo de reducir la dimensión de la variable estudiada, el interés del Análisis de Influencia se centra sobre los parámetros correspondientes al conjunto de variables seleccionadas. El criterio más habitual de selección del número de componentes principales utilizadas para reducir la dimensión se basa en la proporción de variabilidad explicada por un conjunto de componentes principales. Así, el estudio de la influencia ejercida sobre el conjunto de las k primeras componentes principales, se puede plantear mediante el sesgo condicionado de la suma de los autovalores, que como consecuencia de la linealidad de la esperanza, es la suma de los sesgos condicionados de dichos autovalores. De nuevo, se recurre a aproximaciones obtenidas a través de los truncamientos de los desarrollos implicados. Así,

$$\widetilde{S}_t\left(\mathbf{x}_I; \sum_{j=1}^{k} \widetilde{\lambda}_j\right) = \sum_{j=1}^{k} \widetilde{S}_t\left(\mathbf{x}_I; \widetilde{\lambda}_j\right)$$

Al igual que en la sección 5.3, es interesante considerar medidas adimensionales, relativizadas respecto a los autovalores o respecto a la desviación típica de los mismos.

Un estimador de $var\left(\sum_{l=1}^{k} \widetilde{\lambda}_l\right)$ viene dado mediante la estimación directa de cada uno de los parámetros que la componen, omitiendo los términos de

orden n^{-2}, cuya expresión se puede derivar del teorema 1.2.5.

$$\widetilde{var}\left(\sum_{l=1}^{k}\widetilde{\lambda}_l\right) = 2\left(n-1\right)^{-1}\sum_{l=1}^{k}\widetilde{\lambda}_l^2 \tag{5.4}$$

Se proponen como medidas de influencia sobre el conjunto de los k primeros autovalores:

1. Medida M_6 asociada a la suma de los k primeros autovalores debida al conjunto de observaciones subindicadas por $I = \{i_1, ..., i_r\}$,

$$M_6\left(1..k; I\right) = \frac{\left|\widetilde{S}_t\left(\mathbf{x}_I; \sum_{l=1}^{k}\widetilde{\lambda}_l\right)\right|}{\sum_{l=1}^{k}\widetilde{\lambda}_l} = \frac{\left|\sum_{l=1}^{k}\widetilde{S}_t\left(\mathbf{x}_I; \widetilde{\lambda}_l\right)\right|}{\sum_{l=1}^{k}\widetilde{\lambda}_l}.$$

   Representa la razón de cambio experimentado por la variabilidad explicada por las k primeras componentes principales muestrales.

   En consecuencia, esta medida cuantifica la influencia en términos de la variabilidad explicada por las k primeras componentes principales.

2. Medida M_7 asociada a la suma de los k primeros autovalores debida al conjunto de observaciones subindicadas por $I = \{i_1, ..., i_r\}$,

$$M_7\left(1..k; I\right) = \frac{\left|\widetilde{S}_t\left(\mathbf{x}_I; \sum_{l=1}^{k}\widetilde{\lambda}_l\right)\right|}{\sum_{l=1}^{p}\widetilde{\lambda}_l} = \frac{\left|\sum_{l=1}^{k}\widetilde{S}_t\left(\mathbf{x}_I; \widetilde{\lambda}_l\right)\right|}{\sum_{l=1}^{p}\widetilde{\lambda}_l}.$$

   Representa la razón del cambio experimentado por la variabilidad explicada de las k primeras componentes principales en términos de la variabilidad total.

3. Medida M_8 asociada a la suma de los k primeros autovalores debida al

conjunto de observaciones subindicadas por $I = \{i_1, ..., i_r\}$,

$$M_8(1..k; I) = \frac{\left|\widetilde{S}_t\left(\mathbf{x}_I; \sum_{j=1}^{k}\widetilde{\lambda}_j\right)\right|}{\left[\widetilde{var}\left(\sum_{j=1}^{k}\widetilde{\lambda}_j\right)\right]^{\frac{1}{2}}} = \frac{\left|\sum_{j=1}^{k}\widetilde{S}_t\left(\mathbf{x}_I; \widetilde{\lambda}_j\right)\right|}{\left[\widetilde{var}\left(\sum_{j=1}^{k}\widetilde{\lambda}_j\right)\right]^{\frac{1}{2}}}.$$

Representa la razón del cambio experimentado por la variabilidad explicada por las k primeras componentes principales en términos de la desviación típica de la misma.

Nota 5.5.1 *Las medidas M_6, M_7 y M_8, constituyen una generalización de M_1, M_2 y M_3, respectivamente, por lo que los comentarios realizados en las notas 5.3.1 y 5.3.2 se pueden extender a estas medidas.*

Por otro lado, hay que destacar que si la influencia ejercida sobre un autovalor entre los k primeros, es elevada, no tiene porqué ser alta en la suma debido a la posibilidad de compensaciones.

5.6 Medidas de influencia para un conjunto de autovectores

Cuando el interés del Análisis de Influencia se centra en un conjunto de k autovectores, generalmente los primeros, éstos se pueden presentar como la submatriz $\widetilde{\mathbf{A}}_k = [\underset{\sim}{\widetilde{\alpha}}_1, ..., \underset{\sim}{\widetilde{\alpha}}_k] \in \mathcal{M}_{p\times k}$, obteniéndose que

$$\widetilde{S}_t\left(\mathbf{x}_I; \widehat{\mathbf{A}}_k\right) = \left[\widetilde{S}_t\left(\mathbf{x}_I; \underset{\sim}{\widetilde{\alpha}}_1\right), ..., \widetilde{S}_t\left(\mathbf{x}_I; \underset{\sim}{\widetilde{\alpha}}_k\right)\right] \in \mathcal{M}_{p\times k},$$

por lo que se debe recurrir a normas matriciales para medir la influencia ejercida por una o varias observaciones.

En general, las normas que se han utilizado en el Análisis de Influencia son las $(\mathbf{Q}, \mathbf{C})$-normas, citadas en la sección 1.1.4 del capítulo 1. En el caso particular de $\mathbf{C} = c\mathbf{I}_k$, con c escalar positivo, entonces,

$$\|\mathbf{B}\|^2_{(\mathbf{Q},c\mathbf{I}_k)} = \frac{1}{c}tr\left(\mathbf{B}'\mathbf{Q}\mathbf{B}\right) = \frac{1}{c}\sum_{l=1}^{k}\underset{\sim}{b}_l'\mathbf{Q}\underset{\sim}{b}_l = \sum_{l=1}^{k}\|\underset{\sim}{b}_l\|^2_{(\mathbf{Q},c)},$$

siendo $\underset{\sim}{b}_l$ la l-ésima columna de $\mathbf{B}$.

A continuación, se proponen distintas medidas de influencia sobre un conjunto de autovectores:

1. Medida M_9 asociada a $\widetilde{\mathbf{A}}_k$ debida al conjunto de observaciones subindicadas por $I = \{i_1, ..., i_r\}$,

$$M_9\left(\widetilde{\mathbf{A}}_k; I\right) = \left\|\widetilde{\mathbf{A}}_k\right\|_{(\mathbf{I}_p, \mathbf{I}_k)}.$$

Esta medida se puede expresar de la forma,

$$M_9^2\left(\widetilde{\mathbf{A}}_k; I\right) = \sum_{j=1}^{k} \left\|\widetilde{S}_t\left(\mathbf{x}_I; \underline{\widetilde{\alpha}}_q\right)\right\|^2_{(\mathbf{I}_p, 1)} = \sum_{j=1}^{k} M_4^2\left(\underline{\widetilde{\alpha}}_j; I\right).$$

2. Medida M_{10} asociada a $\widetilde{\mathbf{A}}_k$ debida al conjunto de observaciones subindicadas por $I = \{i_1, ..., i_r\}$,

$$M_{10}\left(\widetilde{\mathbf{A}}_k; I\right) = \left\|\widetilde{\mathbf{A}}_k\right\|_{(\mathbf{S}, \frac{1}{n-1}\mathbf{I}_k)}.$$

Esta medida puede expresar también de la forma

$$M_{10}^2\left(\widetilde{\mathbf{A}}_k; I\right) = \sum_{j=1}^{k} \left\|\widetilde{S}_t\left(\mathbf{x}_I; \underline{\widetilde{\alpha}}_j\right)\right\|^2_{(\mathbf{S}, \frac{1}{n-1}\mathbf{I}_k)} = \sum_{j=1}^{k} M_5^2\left(\underline{\widetilde{\alpha}}_j; I\right).$$

M_{10} guarda relación con el diagnóstico propuesto por Critchley [15]

$$\Phi = \left[tr\left(\left(\mathbf{Y}_k - \mathbf{Y}_k^{(i)}\right)'\left(\mathbf{Y}_k - \mathbf{Y}_k^{(i)}\right)\right)\right]^{\frac{1}{2}}$$

donde $\mathbf{Y}_k$ es la matriz de los valores de las primeras k componentes principales e $\mathbf{Y}_k^{(i)}$, es la misma, bajo la omisión de la observación $\underline{x}_i$. Por tanto, Φ es un caso particular de una $(\mathbf{Q}, \mathbf{C})$-norma de $\widetilde{S}\left(\underline{x}_i; \widetilde{\mathbf{A}}_k\right)$, en la que $\mathbf{Q} = \mathbf{S}$ y $\mathbf{C} = \frac{1}{n-1}\mathbf{I}_k$, ya que

$$\begin{aligned}
\left\|\mathbf{Y}_k - \mathbf{Y}_k^{(i)}\right\|_{(\mathbf{I}_p, \mathbf{I}_k)} &= \left[tr\left(\left(\mathbf{Y}_k - \mathbf{Y}_k^{(i)}\right)'\left(\mathbf{Y}_k - \mathbf{Y}_k^{(i)}\right)\right)\right]^{\frac{1}{2}} = \\
&= \left[tr\left(\left(\widetilde{\mathbf{A}}_k - \widetilde{\mathbf{A}}_k^{(i)}\right)\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}\left(\widetilde{\mathbf{A}}_k - \widetilde{\mathbf{A}}_k^{(i)}\right)'\right)\right]^{\frac{1}{2}} = \\
&= \left[tr\left((n-1)\left(\widetilde{\mathbf{A}}_k - \widetilde{\mathbf{A}}_k^{(i)}\right)\mathbf{S}\left(\widetilde{\mathbf{A}}_k - \widetilde{\mathbf{A}}_k^{(i)}\right)'\right)\right]^{\frac{1}{2}} = \\
&= \left\|\widetilde{S}\left(\underline{x}_i; \widetilde{\mathbf{A}}_k\right)\right\|_{(\mathbf{S}, \frac{1}{n-1}\mathbf{I}_k)}
\end{aligned}$$

Para evitar los distintos problemas prácticos mencionados anteriormente, la propuesta de Critchley puede ser sustituida por la medida M_{10}.

Nota 5.6.1 *Las medidas M_9 y M_{10} son una generalización de M_4 y M_5. Además, al ser dichas medidas aditivas respecto a los autovectores considerados, a partir del análisis de influencia sobre cada autovector, los conjuntos de observaciones más influyentes para los k primeros autovectores serán aquéllos que influyan más en cada uno de los autovectores.*

5.7 Aplicaciones

Para ilustrar este trabajo se presentan tres ejemplos prácticos. El primero de ellos corresponde a un conjunto de datos ya utilizado en la bibliografía para el Análisis de Componentes Principales y en particular para el Análisis de Influencia en este campo. Este ejemplo permite la comparación del análisis de influencia realizado mediante el sesgo condicionado con las técnicas utilizadas por otros autores. El segundo ejemplo corresponde a un conjunto de datos generados aleatoriamente, en el cual se han introducido deliberadamente dos observaciones extremas procedentes de otras poblaciones. En él se pretende confirmar la validez de las medidas propuestas al detectar estas observaciones como altamente influyentes. En el tercer ejemplo, también generado aleatoriamente, se muestra la utilidad de la estimación truncada del sesgo condicionado, ante la existencia de autovalores cercanos que provocan la permutación en el orden de los autovalores correspondientes tras la omisión de una observación determinada.

El análisis de influencia llevado a cabo en los tres ejemplos, se realiza desde el punto de vista individual, mediante un programa realizado en MAPLE que se adjunta en el apéndice A. En éste se han obtenido distintos diagnósticos para evaluar la influencia ejercida por las distintas observaciones en los estadísticos de interés.

En dicho programa se establecen algunos criterios, que es conveniente fijar previamente:

- Los autovalores se ordenan en sentido decreciente.
- Los autovectores de la matriz $\mathbf{S}$ se escogen de forma que la coordenada de mayor valor absoluto tenga signo positivo.

Además, en el Análisis de Componentes Principales, el sentido de los autovectores no es de importancia, sino su dirección. En el Análisis de Influencia, se escoge uno de los dos autovectores unitarios posibles para el estudio

de la variabilidad en dicha dirección. Por ello, en la función de influencia muestral, si los autovectores no se comparan de forma adecuada, se pueden detectar observaciones altamente influyentes cuando realmente la dirección del autovector no ha variado sensiblemente. Por ello es necesario fijar algún criterio sobre el autovector calculado tras la omisión de observaciones. En las aplicaciones que se acompañan, con objeto de comparar con las medidas propuestas en el presente capítulo, se ha tomado un criterio sugerido por Brooks [7], que consiste en escoger el autovector tras la omisión con menor ángulo con el obtenido utilizando la muestra completa. Otro criterio lógico es fijar la mayor componente en valor absoluto con signo positivo, que debe coincidir, para tamaño muestral suficientemente grande, con el signo de la misma en el autovector calculado mediante toda la muestra.

En general, en el Análisis de Influencia, algunos autores proponen para las medidas de influencia puntos de corte a partir de los cuales se deben considerar las observaciones influyentes. Sin embargo, otros autores proponen medir la influencia de todas las observaciones y considerar como influyentes aquellas que obtengan valores considerablemente mayores en relación con el resto. Ello puede realizarse a través de una representación gráfica de los valores del diagnóstico considerado frente al índice de casos. Ambas estrategias están cargadas con cierto nivel de subjetividad. No obstante, posiblemente, la segunda esté más en la línea de la propia definición de observación influencia, por lo que en esta memoria se ha optado por ella.

5.7.1 Aplicación 1: Conjunto de datos de Kendall

Kendall [37] proporcionó los datos de la composición de veinte muestras de tierra. En cada una de ellas se mide el contenido de cieno (X_1), el contenido de arcilla (X_2), la materia orgánica (X_3) y la acidez dada por el pH (X_4). Estos datos se han utilizado en distintos artículos para ilustrar el Análisis de Influencia en Componentes Principales (Critchley [15], Tanaka [62], [64], Bénasséni [6], Wang y Nyquist [70] y Shi [59]).

El vector de medias y la matriz de covarianzas muestrales de las cuatro variables estudiadas son, respectivamente,

$$\overline{\underline{X}} = [22.735, 10.505, 2.535, 6.65]',$$

$$\mathbf{S} = \begin{bmatrix} 79.73818421 & 22.38455263 & 1.526605263 & .1107894737 \\ 22.38455263 & 13.81944737 & -.5843947369 & .02500000000 \\ 1.526605263 & -.5843947369 & .6434473685 & .03236842105 \\ .1107894737 & .02500000000 & .03236842105 & .2626315790 \end{bmatrix}.$$

Los autovalores de $\mathbf{S}$, ordenados en sentido decreciente, sus autovectores unitarios asociados y los porcentajes de variabilidad explicada por las componentes principales vienen recogidos en la tabla 5.1. Así, a la vista de los porcentajes de variabilidad acumulados, es suficiente considerar las dos primeras componentes principales.

	$k=1$	$k=2$	$k=3$	$k=4$
$\widetilde{\lambda}_k$	86.640282	7.0935865	.47140339	.25843819
% de variab.	91.718059	7.5093245	.49903120	.27358463
% acum. de variab.	91.718059	99.227384	99.726415	100
$\widetilde{\underline{\alpha}}_k$	.95578526	−.28801387	−.05901045	.00634830
	.29368118	.94519099	.14154620	−.01816655
	.01497175	−.15381233	.97857851	−.13602099
	.00131651	−.00194084	.13735550	.99051903

Tabla 5.1: Autovalores, autovectores y porcentajes de variabilidad.

A continuación se muestran los valores de las diferentes medidas propuestas y algunos resultados complementarios, útiles para analizar la influencia de las distintas observaciones en los autovalores y autovectores de la matriz de covarianzas muestrales.

En primer lugar, los valores de las componentes principales para cada una de las observaciones vienen recogidos en la tabla 5.2. Mediante las representaciones gráficas dadas en la figura 5.1, se pueden detectar las observaciones con valores de componentes principales, en valor absoluto, más elevados.

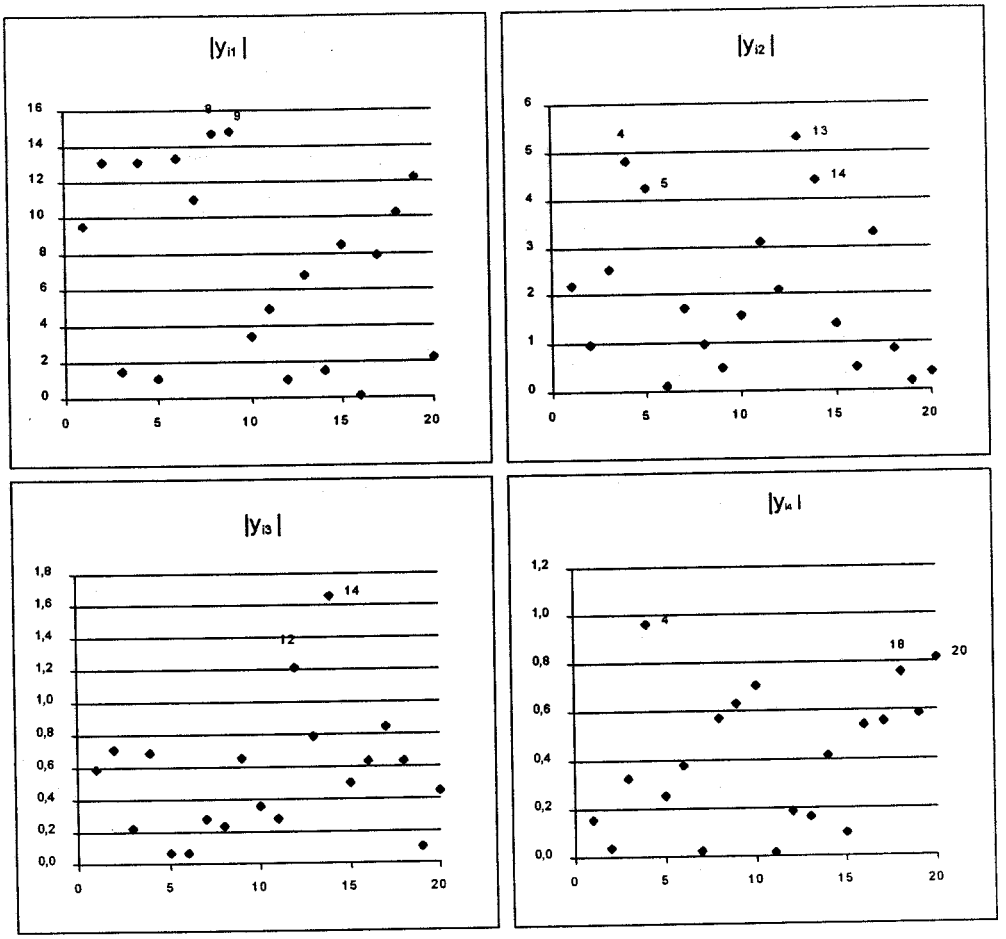

Figura 5.1: Componentes principales, en valor absoluto

	y_{i1}	y_{i2}	y_{i3}	y_{i4}
$i = 1$	−9.556807810	2.202617258	−.5866455116	−.1540246885
$i = 2$	−13.07013056	.9870446058	−.7072802353	−.0340512671
$i = 3$	−1.457765160	2.536032246	.2264804871	.3288506840
$i = 4$	13.07343411	4.803413436	.6920554090	−.9620676794
$i = 5$	−1.060206027	4.233377335	.06784312693	.2526891975
$i = 6$	−13.29393702	.1222527588	−.06683448842	.3805268037
$i = 7$	−10.99691271	−1.707662065	.2760893723	.02346337588
$i = 8$	14.67926366	.9808347053	−.2313668544	.5697595325
$i = 9$	14.83858681	−.4844785369	−.6606535195	.6262058951
$i = 10$	3.337461367	1.563783357	−.3562770329	.7042548569
$i = 11$	4.887304971	3.090409805	.2746808701	.01194812464
$i = 12$	−1.011571080	−2.090355366	1.209406733	.1828900024
$i = 13$	6.798669297	−5.282543718	−.7882937424	−.1624669901
$i = 14$	1.457003418	−4.401296484	1.658280129	.4157070123
$i = 15$	8.410084506	−1.381997489	−.4972418665	−.09636901280
$i = 16$	−.1415324027	−.4895149234	.6315417358	−.5426543608
$i = 17$	7.822063877	−3.271409051	−.8490120303	−.5552954146
$i = 18$	−10.26936646	−.8608192510	−.6303648880	−.7587606154
$i = 19$	−12.24230993	−.1949506673	−.1042748921	.5856137453
$i = 20$	−2.203332831	−.3547379588	.4418671972	−.8162192017

Tabla 5.2: Valores de las componentes principales

Para llevar a cabo el análisis de influencia a través del sesgo condicionado, es necesario contrastar previamente la hipótesis de normalidad multivariante. En el programa adjunto en el apéndice A se obtienen los valores experimentales de los coeficientes de sesgo y curtosis, $b_{1,4}^{exp} = 6.5798$ y $b_{2,4}^{exp} = 19.4892$, respectivamente, y aplicando el contraste de hipótesis propuesto por Mardia [45], basada en dichos coeficientes, se obtiene la región crítica, con un nivel de significación del 5%,

$$RC = \{b_{1,4} \geq 8.8\} \cup \{b_{2,4} \geq 27.1\} \cup \{b_{2,4} \leq 18.4\},$$

por lo que no existe evidencia significativa para rechazar la normalidad tetradimensional de los datos.

Aunque el sesgo condicionado se puede utilizar como herramienta para el análisis de influencia siempre que los autovalores muestrales sean simples, es conveniente estudiar la existencia de autovalores poblacionales múltiples, especialmente entre los primeros, en los que se centra fundamentalmente el Análisis de Influencia. En el ejemplo mostrado, los dos primeros autovalores son bastante diferentes por lo que no existe sospecha de multiplicidad. No obstante, en el programa que se adjunta en el apéndice A, existe un módulo en el que es posible realizar dicho contraste.

En las tablas 5.3 y 5.4 se muestran, para los dos primeros autovalores, los valores de la estimación del sesgo condicionado, $\widetilde{S}\left(\underline{x}_i; \widetilde{\lambda}_j\right)$, $j = 1, 2$. (calculados únicamente teniendo en cuenta el orden que ocupan los autovalores tras la omisión de una observación), la estimación truncada, $\widetilde{S}_t\left(\underline{x}_i; \widetilde{\lambda}_j\right)$, $j = 1, 2$, y las medidas M_1^2, M_2^2 y M_3^2, para cada una de las observaciones $\underline{x}_i$, $i = 1, ..., 20$.

En ambas tablas se refleja la validez de la aproximación de $\widetilde{S}\left(\underline{x}_i; \widetilde{\lambda}_k\right)$ mediante $\widetilde{S}_t\left(\underline{x}_i; \widetilde{\lambda}_k\right)$, $k = 1, 2$, donde para las observaciones con mayor valor de la estimación del sesgo condicionado, los errores relativos cometidos son inferiores al 1%. Análogamente se pueden comparar las estimaciones del sesgo condicionado del resto de los autovalores y de los cuatro autovectores.

i	$\widetilde{S}\left(\underset{\sim}{x}_i;\widetilde{\lambda}_1\right)$	$\widetilde{S}_t\left(\underset{\sim}{x}_i;\widetilde{\lambda}_1\right)$	$M_1^2\left(\widetilde{\lambda}_1,\{i\}\right)$	$M_2^2\left(\widetilde{\lambda}_1,\{i\}\right)$	$M_3^2\left(\widetilde{\lambda}_1,\{i\}\right)$
1	.50720699	.52417794	.00003660	.00003660	0.00034427
2	5.16531355	5.1948352	.00359504	.00302422	.03381384
3	−4.68964283	−4.6892725	.00292934	.00246422	0.0275525
4	4.99332687	5.0478338	.00339445	.00285548	0.03192721
5	−4.74842956	−4.7482176	.00300345	.00252656	0.02824959
6	5.52042784	5.5493295	.00410243	.00345104	.03858620
7	2.24268010	2.2643321	.00068303	.00057457	.00642438
8	7.77440142	7.8119138	.00812971	.00683887	.07646556
9	8.05243201	8.0901232	.00871907	.00733466	.08200895
10	−4.16334344	−4.1614615	.00230702	.00194071	.02169915
11	−3.42597346	−3.4215573	.00155958	.00131195	.01466893
12	−4.75374713	−4.7535704	.00301023	.00253226	.02831332
13	−2.16460520	−2.1539233	.00061804	.00051991	.00581314
14	−4.69108726	−4.6906781	.00293110	.00246570	.02756908
15	−.68361070	−.67152143	.00006007	.00005053	.00056502
16	−4.81217813	−4.8121750	.00308491	.00259509	.02901575
17	−1.26535085	−1.2530186	.00020915	.00017594	.00196728
18	1.34633819	1.3643246	.00024796	.00020859	.00233231
19	3.94876241	3.9734514	.00210327	.00176931	.01978275
20	−4.52963193	−4.5288344	.00273233	.00229849	.02569945

Tabla 5.3: Análisis de influencia sobre $\widetilde{\lambda}_1$.

i	$\widetilde{S}\left(\underline{x}_i;\widetilde{\lambda}_2\right)$	$\widetilde{S}_t\left(\underline{x}_i;\widetilde{\lambda}_2\right)$	$M_1^2\left(\widetilde{\lambda}_2,\{i\}\right)$	$M_2^2\left(\widetilde{\lambda}_2,\{i\}\right)$	$M_3^2\left(\widetilde{\lambda}_2,\{i\}\right)$
1	−.09220849	−.09322301	.00017270	$.973 \cdot 10^{-6}$	.00188256
2	−.32973588	−.33072591	.00217372	.00001225	.02369413
3	−.01794358	−.016872170	$.565 \cdot 10^{-5}$	$.319 \cdot 10^{-7}$	$6.166 \cdot 10^{-5}$
4	1.10906272	1.0971591	.02392255	.00013489	.26076159
5	.65408016	.65708950	.00858059	.00004838	.09353054
6	−.39309281	−.39311022	.00307112	.00001731	.03347596
7	−.20804308	−.20950312	.00087226	$.491 \cdot 10^{-5}$	.00950791
8	−.32814491	−.32979415	.00216149	.00001218	.02356081
9	−.37800622	−.37840779	.00284569	.00001604	.03101876
10	−.25071433	−.25031945	.00124525	$.702 \cdot 10^{-5}$	.01357355
11	.17341524	.17449459	.00060510	$.341 \cdot 10^{-5}$	.00659581
12	−.14165307	−.14071830	.00039352	$.221 \cdot 10^{-5}$	.00428948
13	1.27869938	1.2839630	.03276221	.00018474	.35711612
14	.70841538	.71714621	.01022077	.00005763	.11140888
15	−.27688725	−.27707352	.00152566	$.860 \cdot 10^{-5}$	.01663008
16	−.38015463	−.38011228	.00287138	.00001619	.03129883
17	.25305499	.25382610	.00128038	$.722 \cdot 10^{-5}$	.01395651
18	−.34772357	−.34793163	.00240578	.00001356	.02622359
19	−.39161446	−.39164448	.00304826	.00001718	.03322679
20	−.38675592	−.38673429	.00297230	.00001676	.03239886

Tabla 5.4: Análisis de influencia sobre $\widetilde{\lambda}_2$.

Debido a la gran cantidad de resultados numéricos y la ventaja que proporciona el uso de representaciones gráficas, en las figuras 5.2, 5.3, 5.4, 5.5, 5.6 y 5.7 se representan los valores del diagnóstico frente al índice de casos. Estas gráficas permiten identificar visualmente y de forma rápida observaciones altamente influyentes. En dichas gráficas, se señalan aquellas observaciones con valor del diagnóstico más elevado.

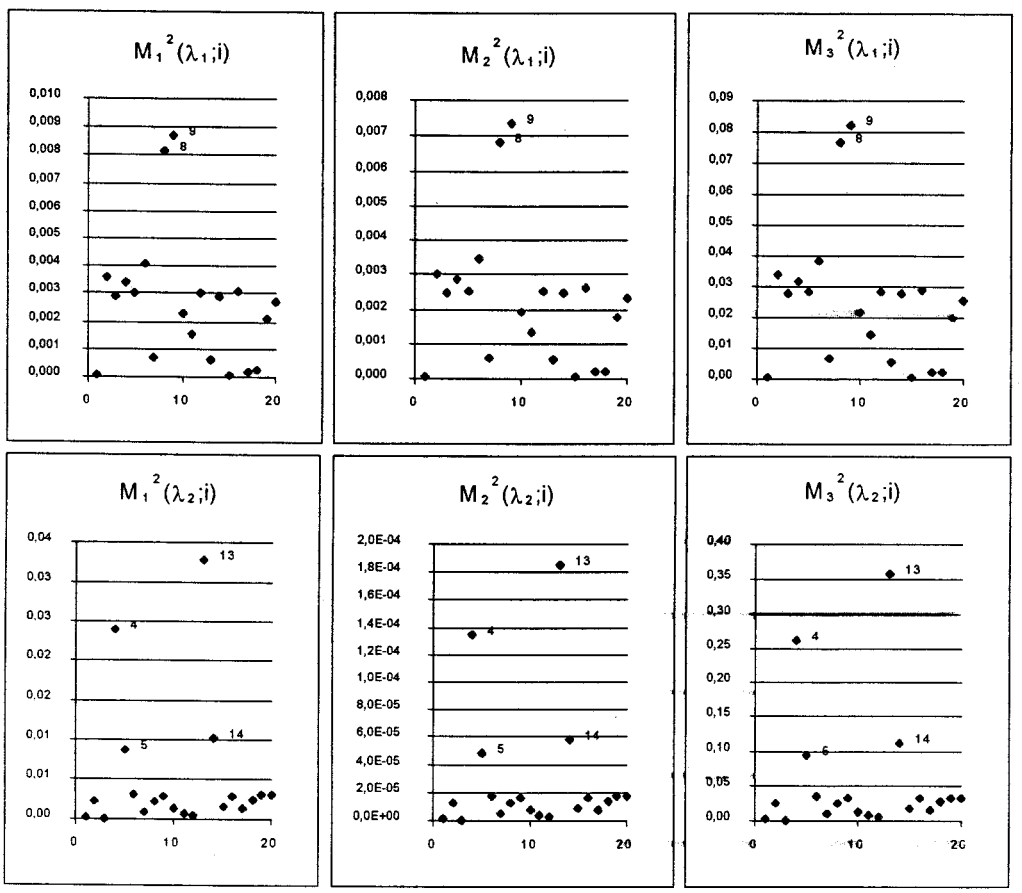

Figura 5.2: Análisis de influencia sobre $\widetilde{\lambda}_1$ y $\widetilde{\lambda}_2$.

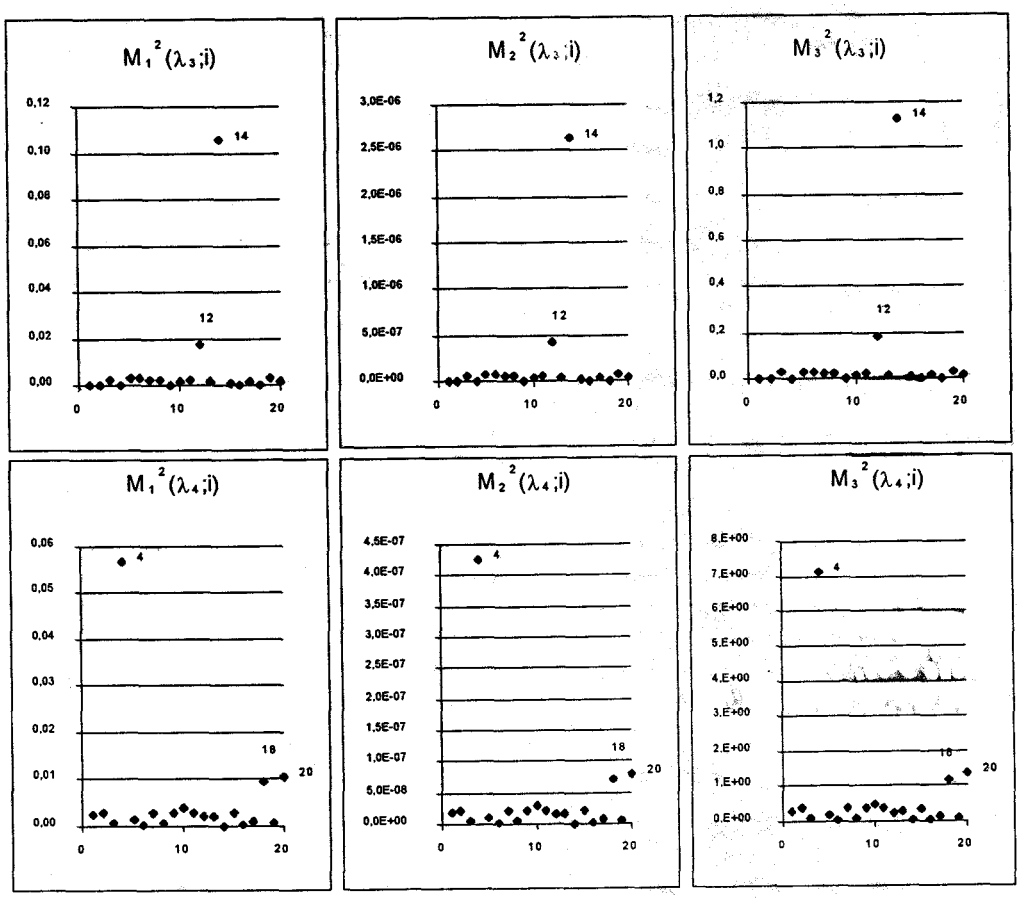

Figura 5.3: Análisis de influencia sobre $\widetilde{\lambda}_3$ y $\widetilde{\lambda}_4$.

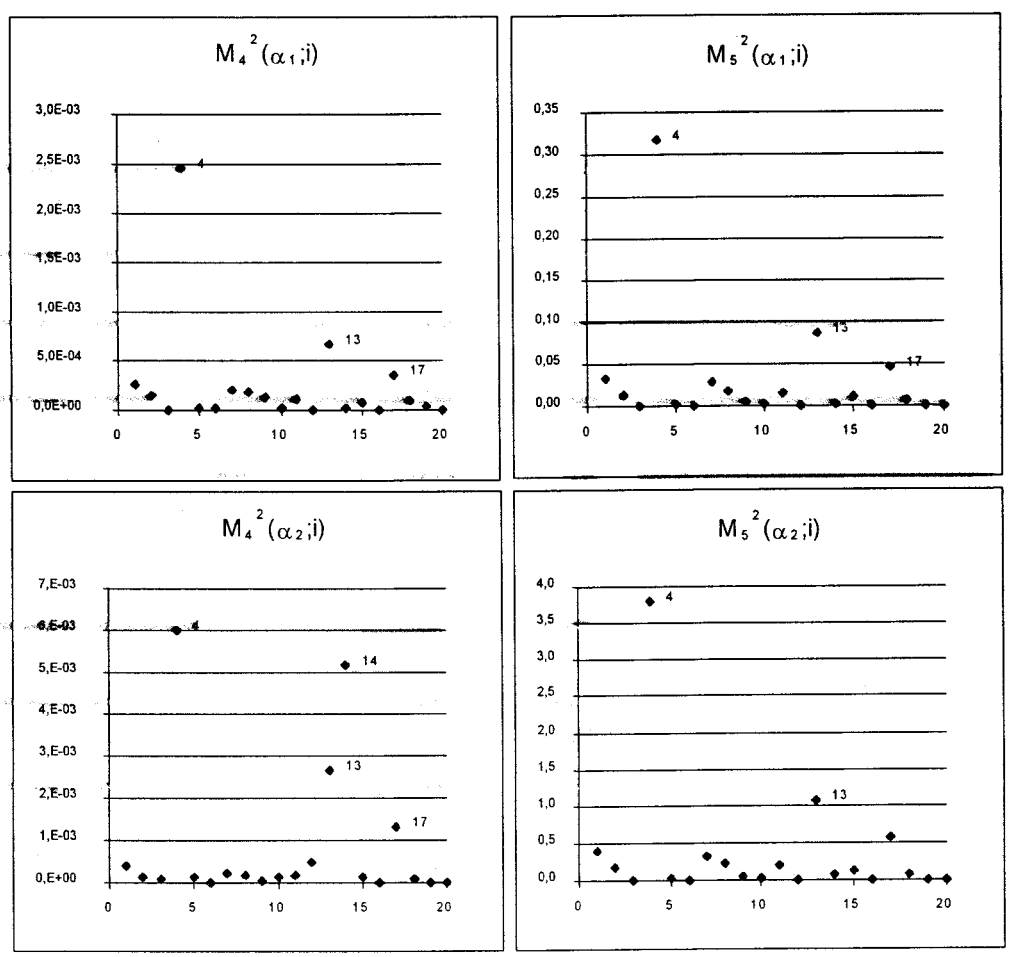

Figura 5.4: Análisis de influencia sobre $\underset{\sim}{\alpha}_1$ y $\underset{\sim}{\alpha}_2$.

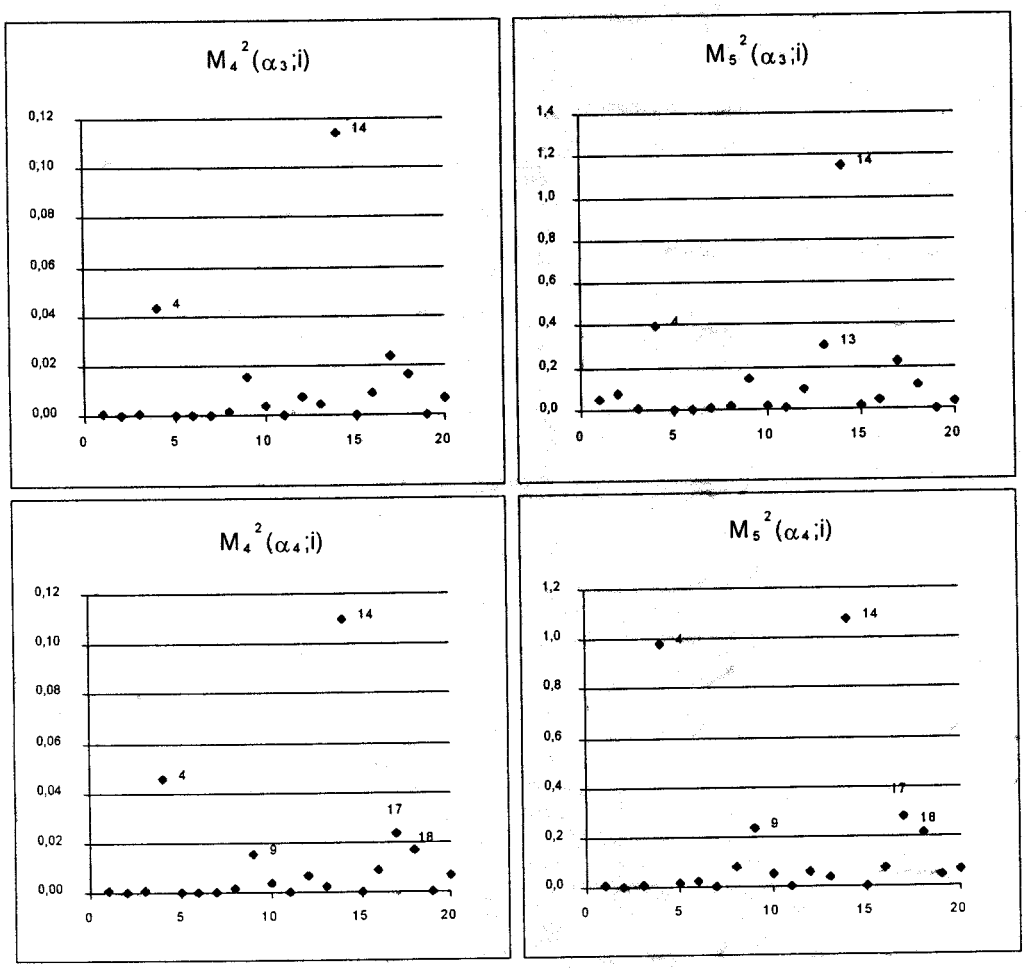

Figura 5.5: Análisis de influencia sobre $\underline{\widetilde{\alpha}}_3$ y $\underline{\widetilde{\alpha}}_4$.

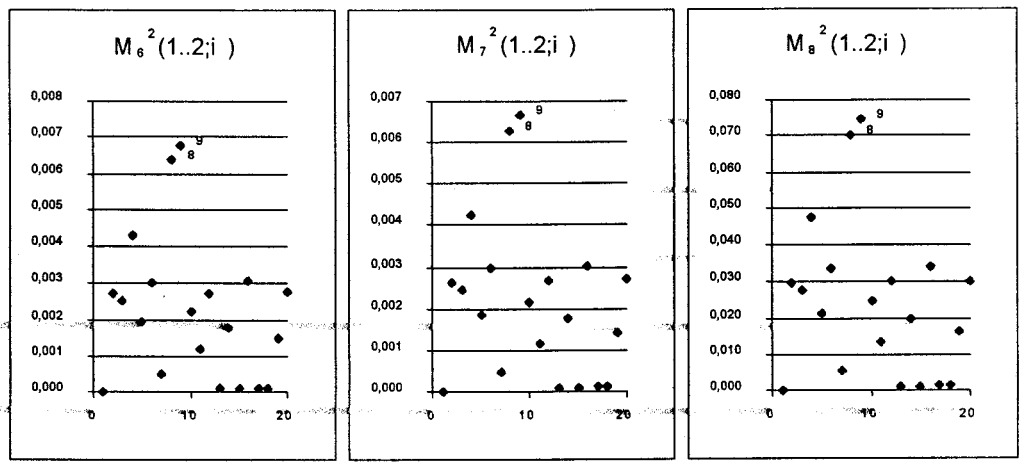

Figura 5.6: Análisis de influencia sobre $\widetilde{\lambda}_1 + \widetilde{\lambda}_2$.

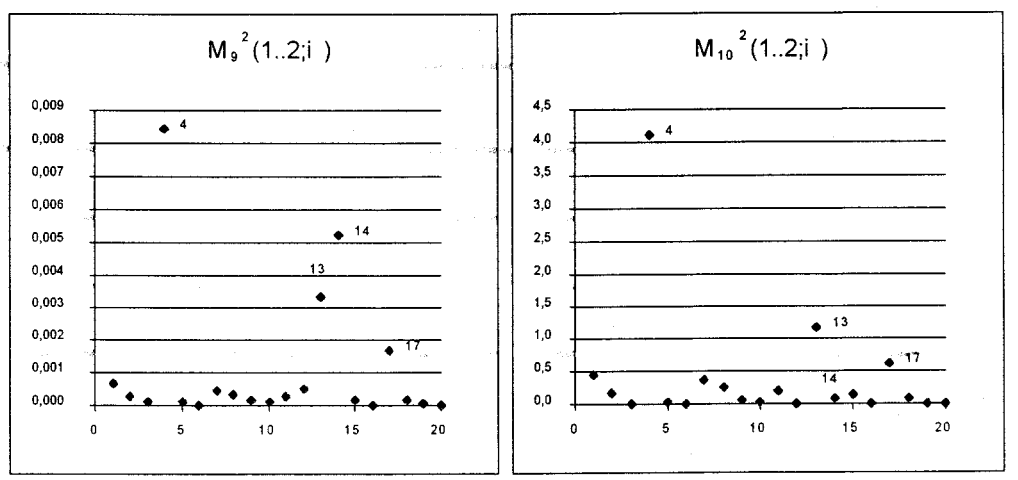

Figura 5.7: Análisis de influencia sobre $\underline{\widetilde{\alpha}}_k$ y $\underline{\widetilde{\alpha}}_k$, conjuntamente.

En la tabla 5.5 se recogen los valores más destacados en cada una de las medidas, en orden decreciente en el caso en el que existan varias, señalando entre paréntesis aquellas observaciones que se puedan considerar de influencia moderada, es decir, siendo destacables, existen otras más sobresalientes.

	M_1^2, M_2^2, M_3^2		M_4^2	M_5^2
$\widetilde{\lambda}_1$	$9, 8$	$\underline{\widetilde{\alpha}}_1$	$4, (13)$	$4, (13)$
$\widetilde{\lambda}_2$	$13, 4, (14, 5)$	$\underline{\widetilde{\alpha}}_2$	$4, 14, (13)$	$4, (13)$
$\widetilde{\lambda}_3$	$14, (12)$	$\underline{\widetilde{\alpha}}_3$	$14, (4)$	14
$\widetilde{\lambda}_4$	$4, (20, 18)$	$\underline{\widetilde{\alpha}}_4$	$14, (4)$	$14, 4$
	M_6^2, M_7^2, M_8^2		M_9^2	M_{10}^2
$\widetilde{\lambda}_1 + \widetilde{\lambda}_2$	$8, 9, (4)$	$\underline{\widetilde{\alpha}}_1, \underline{\widetilde{\alpha}}_2$	$4, (14, 13)$	$4, (13)$

Tabla 5.5: Observaciones influyentes en el ACP

A partir de la tabla 5.5 se pueden realizar los siguientes comentarios:

- Una observación que influye de forma destacada para un autovalor, no tiene porqué hacerlo para su autovector asociado o para otros autovalores: las observaciones 8 y 9 sólo presentan valores elevados de las medidas de influencia para el primer autovalor (salvo en la suma de los primeros autovalores). Del mismo modo, una observación de influencia elevada para un autovector, no tiene porqué influir de forma destacada en su autovalor asociado, como ocurre para las observaciones 4 y 14 sobre $\underline{\widetilde{\alpha}}_1$ y $\underline{\widetilde{\alpha}}_4$, respectivamente. En cambio, es menos frecuente que una observación sea altamente influyente sobre un autovector, y no lo sea para ninguno más, ya que se debe mantener la ortogonalidad de los autovectores.

- En general, las observaciones que ejercen mayor influencia sobre los autovalores son las de mayor valor absoluto en la puntuación de la componente principal correspondiente, como puede verse en la figura 5.1.

- Las medidas M_4^2 y M_5^2 no siempre detectan las mismas observaciones altamente influyentes: la observación 14 es altamente influyente sobre $\underline{\widetilde{\alpha}}_2$ de acuerdo a M_4^2 pero no para M_5^2; la observación 4 destaca para M_5^2 sobre $\underline{\widetilde{\alpha}}_4$, aunque sólo de forma moderada para M_4^2.

La posibilidad de detección de distintas observaciones altamente influyentes se debe a la diferente forma de llevar a medidas unidimensionales el sesgo condicionado de un autovector. Las medidas M_4^2 y M_5^2 se pueden escribir en función de los autovectores de $\mathbf{S}$, según la expresión (5.3). En ambas aparecen los términos $e_{jk}^2(I)$, los cuales se muestran en las tablas 5.6 y 5.7 para $\underset{\sim}{\widetilde{\alpha}}_2$ y $\underset{\sim}{\widetilde{\alpha}}_4$, respectivamente.

Se observa que el caso 14 tiene un valor elevado para M_4^2 en el segundo autovector, debido a los coeficientes tercero y cuarto de la estimación del sesgo condicionado. Esto se suaviza al ponderar por un autovalor pequeño y por ello, mediante la medida M_5^2, que da mayor importancia a los primeros coeficientes, $e_{12}^2(i)$ y $e_{12}^2(i)$, la observación 14 tiene un valor bajo.

En la cuarta observación, los primeros coeficientes de la estimación del sesgo condicionado de $\underset{\sim}{\widetilde{\alpha}}_4$, $e_{14}^2(i)$ y $e_{14}^2(i)$, son elevados respecto al resto de las observaciones, y los últimos no son altos, lo cual se acentúa más al ponderar con los autovalores. Por ello, mediante M_5^2 se detecta como observación más influyente que con M_4^2.

	$e_{12}^2(i)$	$e_{22}^2(i)$	$e_{32}^2(i)$	$e_{42}^2(i)$
$i=1$	0.000241616	$2.9112 \cdot 10^{-8}$	0.000142554	$9.20266 \cdot 10^{-6}$
$i=2$	0.000100819	$3.338 \cdot 10^{-9}$	$4.32793 \cdot 10^{-5}$	$9.4118 \cdot 10^{-8}$
$i=3$	$6.60903 \cdot 10^{-6}$	$1.437 \cdot 10^{-9}$	$2.57996 \cdot 10^{-5}$	$5.08942 \cdot 10^{-5}$
$i=4$	0.002292484	$4.22723 \cdot 10^{-6}$	0.001317194	0.002367289
$i=5$	$9.58573 \cdot 10^{-6}$	$2.08 \cdot 10^{-9}$	$7.68545 \cdot 10^{-6}$	$9.92644 \cdot 10^{-5}$
$i=6$	$1.62247 \cdot 10^{-6}$	0	$5.918 \cdot 10^{-9}$	$1.8009 \cdot 10^{-7}$
$i=7$	0.000201295	$8.813 \cdot 10^{-9}$	$1.92156 \cdot 10^{-5}$	$1.3009 \cdot 10^{-7}$
$i=8$	0.000132964	$3.929 \cdot 10^{-9}$	$4.83914 \cdot 10^{-6}$	$2.75338 \cdot 10^{-5}$
$i=9$	$3.31507 \cdot 10^{-5}$	$3.7 \cdot 10^{-10}$	$9.51725 \cdot 10^{-6}$	$8.02524 \cdot 10^{-6}$
$i=10$	$1.3312 \cdot 10^{-5}$	$3.321 \cdot 10^{-9}$	$2.28718 \cdot 10^{-5}$	$8.37817 \cdot 10^{-5}$
$i=11$	0.000113368	$6.567 \cdot 10^{-9}$	$6.09394 \cdot 10^{-5}$	$1.07737 \cdot 10^{-7}$
$i=12$	$2.11652 \cdot \cdot 10^{-6}$	$5.3281 \cdot 10^{-8}$	0.000472434	$1.01185 \cdot 10^{-5}$
$i=13$	0.000638437	$9.00821 \cdot 10^{-7}$	0.001940517	$7.64784 \cdot 10^{-5}$
$i=14$	$\mathbf{1.86199 \cdot 10^{-5}}$	$\mathbf{3.98069 \cdot 10^{-6}}$	**0.004868267**	**0.000284624**
$i=15$	$7.20805 \cdot 10^{-5}$	$2.514 \cdot 10^{-9}$	$3.75742 \cdot 10^{-5}$	$1.32351 \cdot 10^{-6}$
$i=16$	$2.313 \cdot 10^{-9}$	$3.2 \cdot 10^{-11}$	$6.6766 \cdot 10^{-6}$	$4.62647 \cdot 10^{-6}$
$i=17$	0.000336677	$2.91783 \cdot 10^{-7}$	0.000685887	0.000274041
$i=18$	$4.32436 \cdot 10^{-5}$	$1.893 \cdot 10^{-9}$	$2.38086 \cdot 10^{-5}$	$3.23675 \cdot 10^{-5}$
$i=19$	$3.37261 \cdot 10^{-6}$	$3 \cdot 10^{-12}$	$3.5327 \cdot 10^{-8}$	$1.04586 \cdot 10^{-6}$
$i=20$	$2.95748 \cdot 10^{-7}$	$1.4 \cdot 10^{-11}$	$1.72022 \cdot 10^{-6}$	$5.50925 \cdot 10^{-6}$

Tabla 5.6: Coeficientes $e_{j2}^2(i)$

	$e_{14}^2(i)$	$e_{24}^2(i)$	$e_{34}^2(i)$	$e_{44}^2(i)$
$i=1$	$1.26474 \cdot 10^{-6}$	$1.07269 \cdot 10^{-5}$	0.000775943	$7.9554 \cdot 10^{-8}$
$i=2$	$1.27662 \cdot 10^{-7}$	$1.16283 \cdot 10^{-7}$	$6.14756 \cdot 10^{-5}$	$3.92 \cdot 10^{-10}$
$i=3$	$1.08211 \cdot 10^{-7}$	$5.22272 \cdot 10^{-5}$	0.000407352	$4.4843 \cdot 10^{-8}$
$i=4$	**0.000132275**	**0.002824261**	**0.042071616**	**0.000250322**
$i=5$	$3.9093 \cdot 10^{-8}$	$9.94685 \cdot 10^{-5}$	$2.55797 \cdot 10^{-5}$	$2.286 \cdot 10^{-9}$
$i=6$	$1.3156 \cdot 10^{-5}$	$1.77353 \cdot 10^{-7}$	$5.10718 \cdot 10^{-5}$	$7.49 \cdot 10^{-10}$
$i=7$	$3.4627 \cdot 10^{-8}$	$1.33362 \cdot 10^{-7}$	$3.59004 \cdot 10^{-6}$	$2E-12$
$i=8$	$3.89437 \cdot 10^{-5}$	$2.76534 \cdot 10^{-5}$	0.001368335	$3.79533 \cdot 10^{-7}$
$i=9$	$5.64397 \cdot 10^{-5}$	$9.56462 \cdot 10^{-6}$	0.015539752	$3.41578 \cdot 10^{-5}$
$i=10$	$2.56671 \cdot 10^{-6}$	$8.93716 \cdot 10^{-5}$	0.003711565	$4.76689 \cdot 10^{-6}$
$i=11$	$1.746 \cdot 10^{-9}$	$1.11512 \cdot 10^{-7}$	$9.07399 \cdot 10^{-7}$	0
$i=12$	$2.8455 \cdot 10^{-8}$	$1.94 \cdot 10^{-5}$	0.006609823	$2.7873 \cdot 10^{-6}$
$i=13$	$1.0156 \cdot 10^{-6}$	$9.79018 \cdot 10^{-5}$	0.002224664	$3.39522 \cdot 10^{-7}$
$i=14$	$5.34071 \cdot 10^{-7}$	0.000777301	0.108429472	0.000265045
$i=15$	$3.44176 \cdot 10^{-7}$	$1.48419 \cdot 10^{-6}$	0.000197095	$6.217 \cdot 10^{-9}$
$i=16$	$2.983 \cdot 10^{-9}$	$5.67768 \cdot 10^{-6}$	0.008465984	$1.59891 \cdot 10^{-5}$
$i=17$	$1.34775 \cdot 10^{-5}$	0.000375195	0.023030245	$5.89167 \cdot 10^{-5}$
$i=18$	$3.47985 \cdot 10^{-5}$	$3.87668 \cdot 10^{-5}$	0.016430218	$6.10058 \cdot 10^{-5}$
$i=19$	$2.56923 \cdot 10^{-5}$	$1.03573 \cdot 10^{-6}$	0.000258896	$1.9027 \cdot 10^{-8}$
$i=20$	$1.48799 \cdot 10^{-6}$	$6.10223 \cdot 10^{-6}$	0.006874925	$1.96193 \cdot 10^{-5}$

Tabla 5.7: Coeficientes $e_{j4}^2(i)$

- Para la suma de los dos primeros autovalores, según las medidas M_6, M_7 y M_8, destacan moderadamente las observaciones 8 y 9. Estas son las mismas observaciones que para $\widetilde{\lambda}_1$, lo cual es lógico, pues un cambio en el primer autovalor, en términos absolutos, en general es mayor que en el segundo autovalor. La observación 4 tiene una influencia bastante elevada sobre el segundo autovalor e influencia moderada en la suma. En cambio, la observación 13, que influye más que 4 para $\widetilde{\lambda}_2$, no se destaca para $\widetilde{\lambda}_1 + \widetilde{\lambda}_2$. Esto se produce, en parte, por ser de signo opuesto la estimación del sesgo condicionado de ambos autovalores.

- Para el conjunto de los dos primeros autovectores, debido a la aditividad de las medidas M_9^2 y M_{10}^2 respecto a M_4^2 y M_5^2, destaca fundamentalmente la observación 4, lo mismo que ocurría con cada uno de ellos por separado. Situación análoga sucede con el caso 13 de forma moderada. Y siendo muy destacada la observación 14 para $\underline{\widetilde{\alpha}}_2$ con M_4^2, lo es de forma moderada para el conjunto de ambos autovectores, con M_9^2.

- Se puede comparar el análisis de influencia llevado a cabo basado en la estimación truncada del sesgo condicionado, con el realizado mediante otras técnicas por diversos autores.

  Por ejemplo, en el análisis realizado por Critchley [15], basado en las funciones de influencia muestrales, las observaciones que se detectan con influencia elevada para los autovalores son básicamente las mismas, salvo algunas excepciones. En el estudio realizado por dicho autor, las observaciones 8 y 9 toman valores de la función de influencia ligeramente más elevados que el resto para el primer autovalor, y mediante este criterio no es claro clasificarlas como observaciones influyentes. Por otro lado, la observación 14 se detecta como de influencia moderada para el último autovalor, mediante la función de influencia muestral, pero no mediante las medidas basadas en la estimación truncada del sesgo condicionado.

  Bénasséni [6] estudia la influencia que ejerce cada observación en el espacio generado por todos los posibles conjuntos de autovectores, a través de medidas de proximidad de subespacios vectoriales. En la tabla 5.8 se recogen aquellas observaciones más destacadas para cada uno de los autovectores y para el espacio generado por los dos primeros, $\langle \underline{\widetilde{\alpha}}_1, \underline{\widetilde{\alpha}}_2 \rangle$.

  Comparando la tabla 5.8 con la tabla 5.5, se observa que en general, las observaciones detectadas como de influencia elevada para los distintos diagnósticos de influencia son las mismas, excepto la observación 17

$\widetilde{\underline{\alpha}}_1$	$4,(13,17)$
$\widetilde{\underline{\alpha}}_2$	$4,14,(13,17)$
$\widetilde{\underline{\alpha}}_3$	$14,4$
$\widetilde{\underline{\alpha}}_4$	$14,4$
$\langle\widetilde{\underline{\alpha}}_1,\widetilde{\underline{\alpha}}_2\rangle$	$14,4,(13,17)$

Tabla 5.8: Benasseni: Observaciones influyentes.

que, aunque no destaca marcadamente mediante los diagnósticos propuestos, toma los valores más elevados dentro de las observaciones no claramente distinguidas. En el caso de dos autovectores, hay que tener en cuenta que el análisis de influencia que se realiza es diferente: en este trabajo se trata de evaluar la influencia ejercida por una o varias observaciones en un conjunto de autovectores; mientras que Bénasséni [6] evalúa la influencia sobre el espacio generado por dichos autovectores. A pesar de ello, las observaciones detectadas son básicamente las mismas.

Desde la perspectiva de la influencia local, mediante los diagnósticos propuestos por Shi [59], las observaciones que ejercen más influencia sobre los estadísticos de interés se recogen en la tabla 5.9, las cuales son fundamentalmente las mismas, aunque existe una alteración en la intensidad de la influencia de la observación 4, para el segundo autovector.

$\widetilde{\lambda}_1$		$\widetilde{\underline{\alpha}}_1$	$4,(13)$
$\widetilde{\lambda}_2$	$13,4$	$\widetilde{\underline{\alpha}}_2$	$14,13,4$
$\widetilde{\lambda}_3$	$14,12$	$\widetilde{\underline{\alpha}}_3$	$14,4$
$\widetilde{\lambda}_4$	4	$\widetilde{\underline{\alpha}}_4$	$14,4$

Tabla 5.9: Shi: observaciones influyentes.

Como conclusión, con este ejemplo se muestra la validez de las medidas de influencia propuestas en el trabajo, para detectar observaciones altamente influyentes en el Análisis de Componentes Principales. Los casos destacados son básicamente los mismos que en los estudios realizado mediante otras técnicas, aunque con algunos matices diferenciales propios.

5.7.2 Aplicación 2: Inclusión de observaciones extrañas

Los datos de este ejemplo, los cuales se adjuntan en el Apéndice B, se han generado de forma aleatoria. Representan una muestra de tamaño 50, en la que las 48 primeras observaciones corresponden a un mismo vector normal tetradimensional de vector de medias

$$[45, 50, -50, 13]'$$

y matriz de covarianzas

$$\begin{bmatrix} 110 & -75 & 65 & 2 \\ -75 & 365 & -160 & -3 \\ 65 & -160 & 205 & -2 \\ 2 & -3 & -2 & 1 \end{bmatrix}.$$

Las dos últimas observaciones, 49 y 50, pueden consideradas como atípicas, al generarse a partir de poblaciones distintas, con vectores de medias, respectivamente, $[35, 90, -70, 13]'$ y $[65, 14, -60, 16]'$, y matriz de covarianzas la misma que el resto de la muestra.

En las figuras 5.8 y 5.9, se observa que:

- La observación 2 destaca para $\widetilde{\lambda}_1$ y $\underline{\widetilde{\alpha}}_1$, según se refleja en los gráficos correspondientes a $M_1^2\left(\widetilde{\lambda}_1; i\right)$ y $M_4^2\left(\underline{\widetilde{\alpha}}_1; i\right)$.
- La observación 49 destaca únicamente para $\widetilde{\lambda}_1$.
- La observación 50 destaca para los dos últimos autovalores y para todos los autovectores, según se refleja en los gráficos correspondientes a $M_1^2\left(\widetilde{\lambda}_k; i\right)$, $k = 3, 4$ y $M_4^2\left(\underline{\widetilde{\alpha}}_k; i\right)$, $k = 1, 2, 3, 4$.

Es de resaltar, que las observaciones 2 y 49 toman valores ligeramente destacados para la primera componente principal, mientras que la observación 50 los tiene para las dos últimas, como puede verse en la figura 5.10. Además, se verifica que $\underline{x}_{49} - \overline{\underline{x}} \simeq 50.194 \cdot \underline{\widetilde{\alpha}}_1$, es decir, $\underline{x}_{49} - \overline{\underline{x}}$ es prácticamente paralelo al primer autovector y por tanto casi ortogonal al resto. Por ello, la influencia sobre los autovectores es muy pequeña.

Como conclusión, en este ejemplo se muestra la utilidad de los diagnósticos de influencia propuestos, al detectar como influyentes observaciones que proceden de poblaciones extrañas, y por lo tanto propensas a provocar cambios considerables en los resultados del Análisis de Componentes Principales. Además se detecta un caso, 2, que procediendo de la misma población, debido a la aleatoriedad de la distribución, destaca del resto.

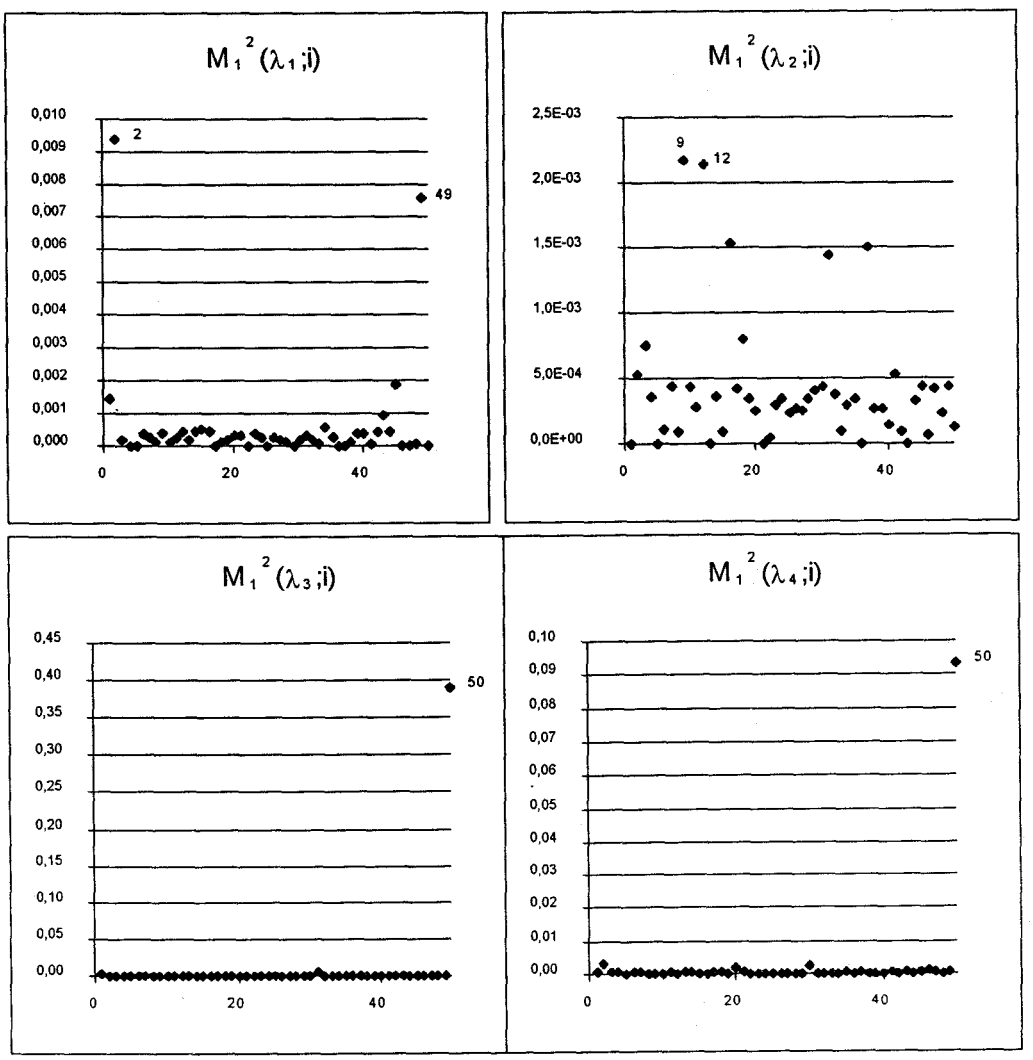

Figura 5.8: Análisis de influencia sobre $\widetilde{\lambda}_k$, $k = 1, ..., 4$.

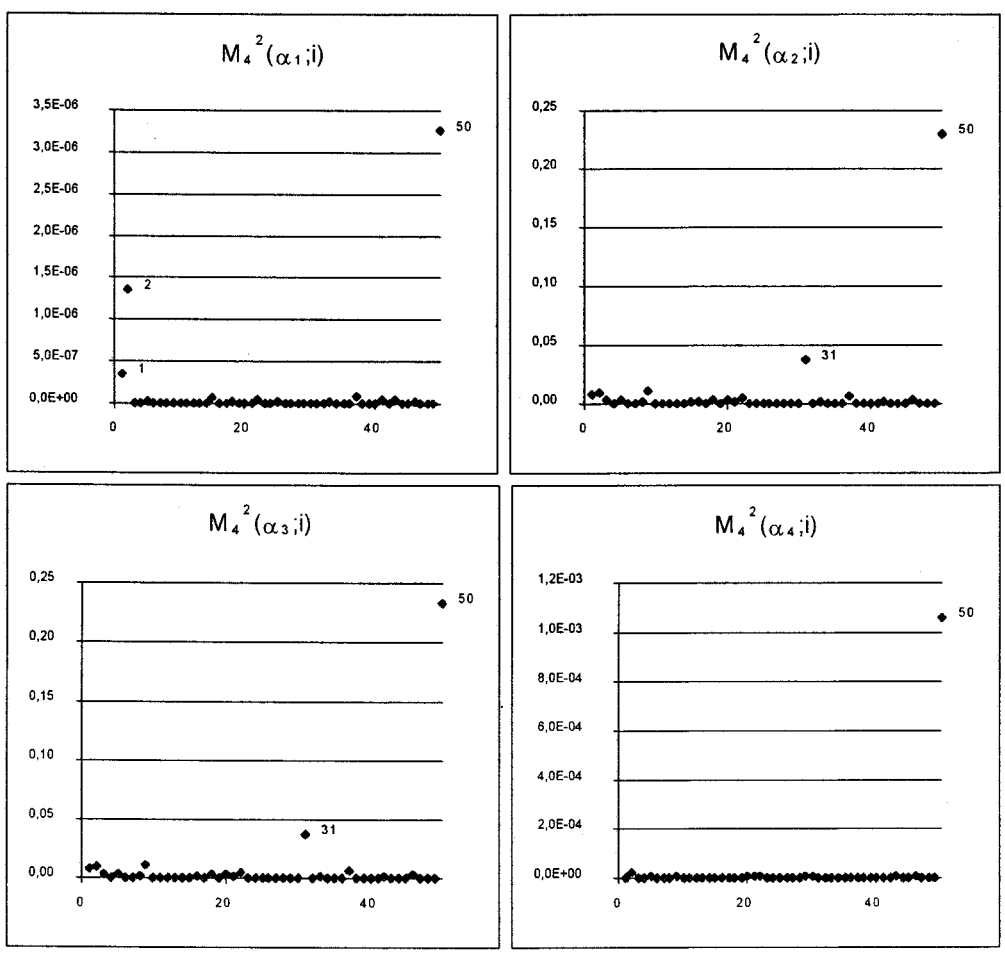

Figura 5.9: Análisis de influencia sobre $\underline{\widetilde{\alpha}}_k$ $k = 1, ..., 4$

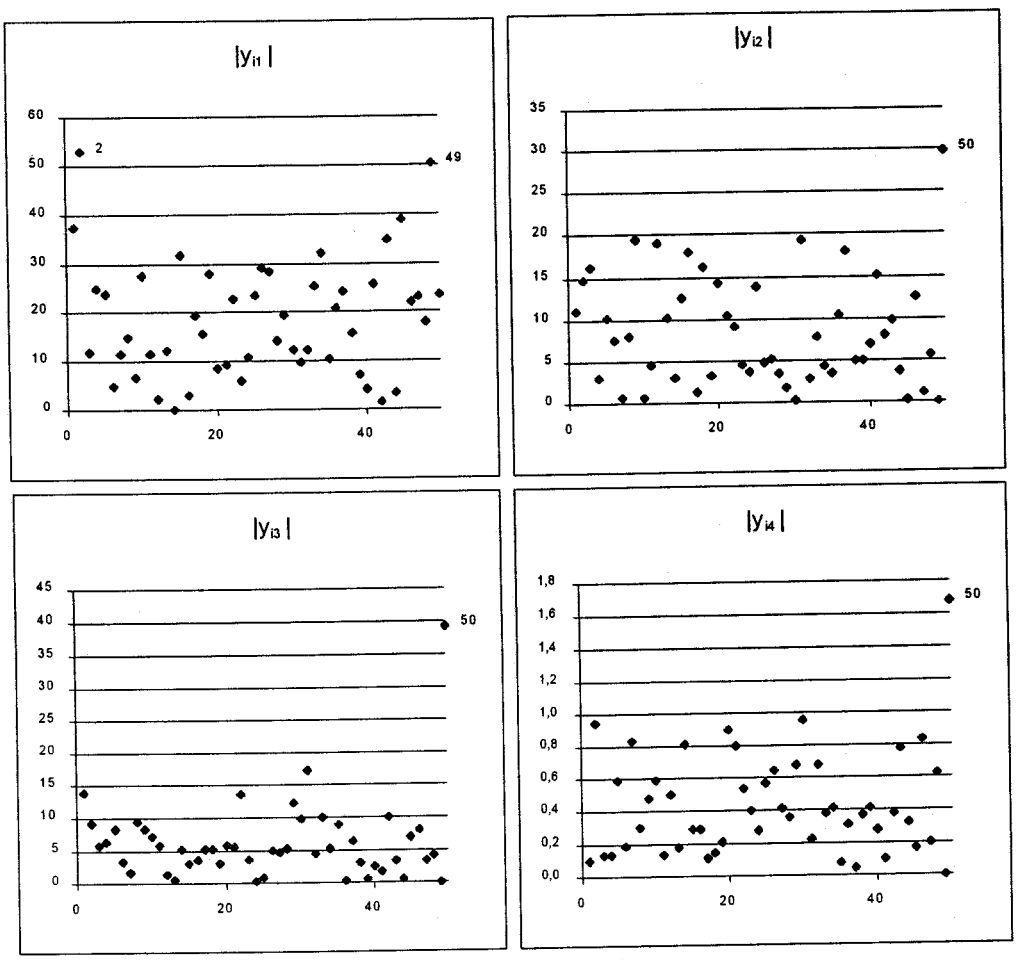

Figura 5.10: Componentes principales, en valor absoluto.

5.7.3 Aplicación 3: El problema del orden

Los datos de este ejemplo, los cuales se adjuntan en el Apéndice C, se han generado de forma aleatoria, según una ley normal tetradimensional de vector de medias

$$[20, 15, 10, 14]'$$

y matriz de covarianzas

$$\begin{bmatrix} 9.25 & 0 & 0 & 0 \\ 0 & 8.75 & 0 & 0 \\ 0 & 0 & 6.25 & 5.5 \\ 0 & 0 & 5.5 & 5 \end{bmatrix}.$$

Los autovalores de la matriz de covarianzas muestrales y sus autovectores unitarios asociados se recogen en la tabla 5.10.

	$k=1$	$k=2$	$k=3$	$k=4$
$\widetilde{\lambda}_k$	12.70302200	9.519176897	8.643221800	.06290442
$\widetilde{\underline{\alpha}}_k$	$-.5532811829$	.7138684063	.4292660818	.001568979698
	$-.1182354066$	$-.5773711641$	.8078105832	$-.01024637365$
	.6078011142	.2967026848	.2924501392	$-.6760312397$
	.5572056415	.2626822277	.2786496145	.7368000495

Tabla 5.10: Autovalores y autovectores

En este trabajo se ha comentado reiteradamente el inconveniente derivado del cálculo de la estimación del sesgo condicionado mediante la comparación directa de los autovectores, ya que ante la existencia de autovalores cercanos, o de tamaños muestrales pequeños, es posible que no se realice de forma adecuada. Así ocurre en el presente ejemplo. La detección de este tipo de situaciones no es sencilla ya que a veces es difícil distinguir, a primera vista, si el hecho de que los diagnósticos de influencia sobre un autovector sean elevados, se deba a que realmente se trate de una observación altamente influyente, o por lo contrario, a una comparación indecuada de los estadísticos.

En las tablas 5.11 y 5.12, se muestran los autovalores y autovectores de las matrices de covarianzas muestrales obtenidos tras la omisión de las observaciones 2 y 92, respectivamente.

	$k=1$	$k=2$	$k=3$	$k=4$
$\widetilde{\lambda}_k^{(2)}$	12.82054698	8.757904887	8.486883075	.06046609
$\widetilde{\underline{\alpha}}_k^{(2)}$	−.5742497674	.6486105334	−.4995352936	−.002464030390
	−.09871801964	.5508359123	.8287332506	−.005979067405
	.5986467731	.3826896710	−.1879452975	−.6781203589
	.5496489773	.3597676879	−.1683515812	.7349224164

Tabla 5.11: Autovalores y autovectores tras la omisión de la observación 2

	$k=1$	$k=2$	$k=3$	$k=4$
$\widetilde{\lambda}_k^{(92)}$	12.79832496	8.887882345	8.020129327	.06291679
$\widetilde{\underline{\alpha}}_k^{(92)}$	−.5860758139	.7480407062	−.3113523684	.003153591971
	−.07533123133	.3323121207	.9400652970	−.01308040227
	.5945319966	.4224183354	−.1110757441	−.6750974993
	.5453183024	.3893154326	−.08366073862	.7376058055

Tabla 5.12: Autovalores y autovectores tras la omisión de la observación 92

En primer lugar, se observa una mayor proximidad entre los autovalores segundo y tercero, especialmente para la omisión de la segunda observación. La omisión de la segunda observación, aparentemente, provoca el cambio de signo de la segunda componente tanto sobre el segundo autovector como sobre el tercero. Pero se puede observar que los signos de las componentes de $\widetilde{\underline{\alpha}}_2^{(2)}$ son los mismos que los de $\widetilde{\underline{\alpha}}_3$, y los de las componentes de $\widetilde{\underline{\alpha}}_3^{(2)}$ son los opuestos de los de $\widetilde{\underline{\alpha}}_2$. Esto revela la posibilidad de que exista una permutación en el orden de los autovalores. Un comentario similar se puede realizar para la observación 92.

A continuación, se comparan tres estudios, ilustrados gráficamente en la figura 5.11: El primero se realiza utilizando como diagnóstico de influencia $\|FIM\,(\underline{x}_i;\widetilde{\underline{\alpha}}_k)\|_2^2$ (simbolizado el la figura 5.11 por a_k, $k=2,3$). El segundo, se realiza corrigiendo el diagnóstico anterior, alterando el sentido del tercer

autovector tras la omisión y permutando los autovectores segundo y tercero en las observaciones 2 y 92 (simbolizado el la figura 5.11 por b_k, $k = 2, 3$). El tercer estudio utiliza M_4^2 (simbolizado el la figura 5.11 por c_k, $k = 2, 3$).

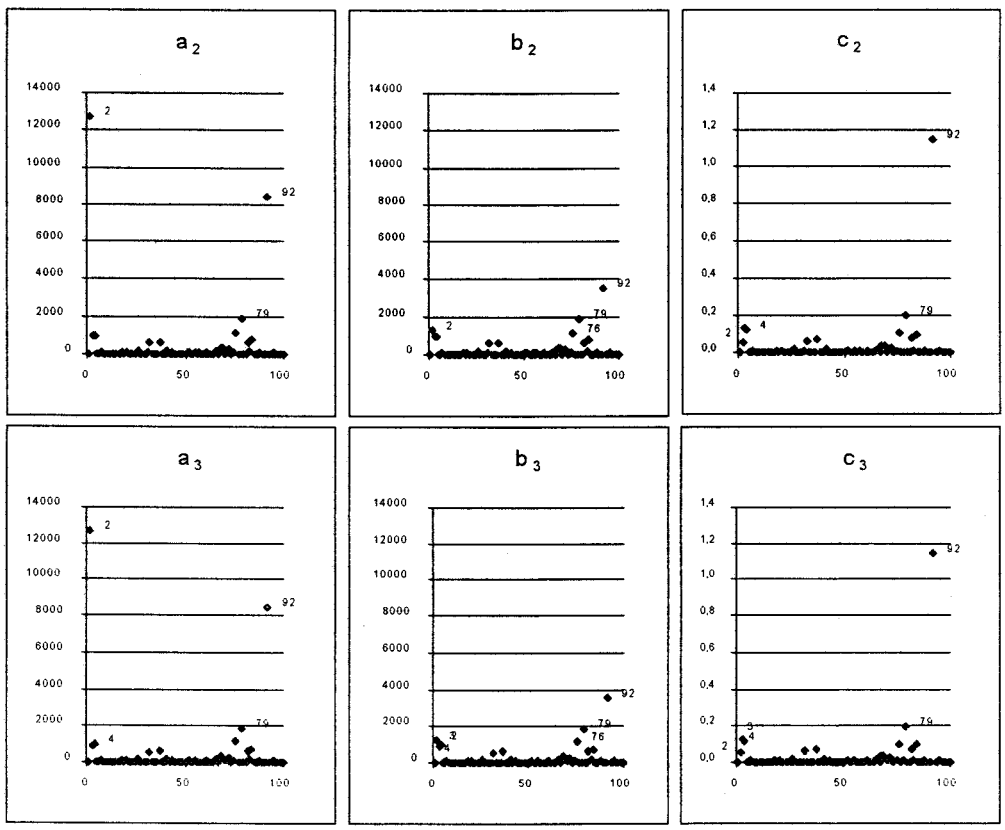

Figura 5.11: Análisis de influencia sobre $\underline{\widetilde{\alpha}}_2$ y $\underline{\widetilde{\alpha}}_3$.

$$
\begin{aligned}
a_k &: FIM\left(\underline{x}_i; \underline{\widetilde{\alpha}}_k\right), \quad k = 2, 3, \\
b_k &: FIM\left(\underline{x}_i; \underline{\widetilde{\alpha}}_k\right) \text{ corregida} \quad k = 2, 3, \\
c_k &: S\left(\underline{x}_i; \underline{\widetilde{\alpha}}_k\right), \quad k = 2, 3,
\end{aligned}
$$

A partir de las gráficas representadas en dicha figura, se advierte que para la observación 2, $FIM\,(\underline{x}_2;\widetilde{\underline{\alpha}}_k)$, $k=2,3$, se habían calculado de forma incorrecta, ya que al realizar el análisis de influencia sobre los autovectores $\widetilde{\underline{\alpha}}_2$ y $\widetilde{\underline{\alpha}}_3$, a partir de su norma euclídea, se detecta como la observación más influyente (véase a_2 y a_3); en cambio, al corregir el cálculo de $FIM\,(\underline{x}_2;\widetilde{\underline{\alpha}}_k)$, $k=2,3$, esta observación proporciona un valor del diagnóstico notablemente más reducido (véase b_2 y b_3). Mediante el uso de la estimación truncada de $S\,(\underline{x}_2;\widetilde{\underline{\alpha}}_k)$, $k=2,3$, se obtiene directamente esta conclusión (véase c_2 y c_3).

En cambio, al analizar la observación 92, tanto al comparar los autovectores directamente, como al permutarlos para el cálculo de $\widetilde{S}\,(\underline{x}_{92};\widetilde{\underline{\alpha}}_k)$, $k=2,3$, el valor del diagnóstico de influencia utilizado, es elevado respecto al resto. Esto se refleja de la misma forma al utilizar $\widetilde{S}_t\,(\underline{x}_{92};\widetilde{\underline{\alpha}}_k)$, $k=2,3$.

En conclusión, puede observarse que las medidas propuestas en esta memoria inciden en menor medida en el problema de autovalores muy próximos que las propuestas por otros autores basados en la comparación directa de los autovalores o autovectores asociados obtenidos utilizando sin y con la omisión del caso bajo estudio.

Apéndice A

Implementación de medidas de influencia

El programa básico utilizado para ilustrar los ejemplos anteriores, se ha realizado mediante MAPLE. Las salidas, en general, se realizan a ficheros para su posterior tratamiento mediante Microsoft Excel, principalmente para realizar las representaciones gráficas presentadas. Las instrucciones de dicho programa, se exponen a continuación.

```
> with(linalg):
  with(plots):
  Digits:=10:
```

p =dimensión del vector de variables bajo estudio.

```
> p:=4:
```

Fichero de datos:

```
> Matriz_datos:=fopen('DatosKendall',READ):
    X:=readdata(Matriz_datos,p):
  fclose(Matriz_datos):
```

Comprobación del rango de la matriz de datos y determinación del número de observaciones= n:

```
> rg:=rank(X);
  n:=rowdim(X):
```

Cálculo del vector de medias y la matriz de covarianzas muestrales:

```
> E:=n->vector(n,1):
  mhu:=vector(p):
  mhu:=scalarmul(&*(transpose(X),E(n)),(1./n));
  Xtilde:=proc(XX) local p,n,i,Xmedia;
    p:=coldim(XX);
    n:=rowdim(XX);
    Xmedia:=array(1..n,1..p):
    Xmedia:=scalarmul(&*(E(n),transpose(E(n)),XX),(1./n)):
    evalm(XX-Xmedia):
  end:
  S:=proc(XX)
    local n,XXtild;
    n:=rowdim(XX);
    XXtild:=Xtilde(XX):
    scalarmul(&*(transpose(XXtild),XXtild),(1./(n-1))):
  end:
  Xtild:=Xtilde(X):
  SX:=S(X);
```

Procedimiento para la ordenación de los autovalores de una matriz, en sentido decreciente, y de las filas de la matriz de autovalores / multiplicidad / autovector unitario asociado, de la misma forma:

```
> T:=proc(SXX) local T1,p,w,lambda,T2,k,j,t,u;
    p:=rowdim(SXX);
    lambda:=vector(p);
    T1:=eigenvects(SXX);
    w:=vector(p,k->k):
    for k from 1 to p do
      lambda[k]:=T1[k][1]:
    od:
    T2:=array(1..p,1..3):
    for k from 2 to p do
      for j from k-1 by -1 to 1 do
        if lambda[j+1]>lambda[j] then
          t:=lambda[j]; u:=w[j]; lambda[j]:=lambda[j+1];
          w[j]:=w[j+1];
          lambda[j+1]:=t;
          w[j+1]:=u;
        fi;
      od;
    od;
```

```
    for k from 1 to p do
      for j from 1 to 3 do
        T2[k,j]:=T1[w[k],j]:
      od:
    od:
    T2:
  end:
```

Construcción del vector de autovalores, lambda, y de la matriz de autovectores, A, asignando signo positivo a la mayor componente, en valor absoluto, de cada autovector.

```
> A:=array(1..p,1..p):
  TS:=T(SX):
  lambda:=vector(p,k->TS[k,1]):
  for k from 1 to p do
    z:=op(TS[k,3]):
    a:=op(convert(z,list)):
    if abs(min(a))>max(a) then
      z:=scalarmul(z,-1):
    fi:
    for j from 1 to p do
      A[k,j]:=z[j]:
    od:
  od:
  print(lambda):
 print(A):
```

Cálculo del porcentaje de variabilidad de las componentes principales:

```
> sumalambda:=sum('lambda[j]','j'=1..p):
  print(sumalambda);
  for j from 1 to p do
    print(j,100*lambda[j]/sumalambda):
    print(100*sum('lambda[l]','l'=1..j)/sumalambda):
  od:
```

Matriz de componentes principales:

```
> CP:=array(1..n,1..p):
  CP:=evalm(&*(Xtild,transpose(A))):
  cp:=fopen(CompPrinc,WRITE):
    writedata(cp,CP):
  fclose(cp):
```

Contraste de normalidad multivariante:

```
> G:=array(1..n,1..n):
  G:=evalm(&*(Xtild,inverse(SX),transpose(Xtild))):
  sesgo:=sum('sum('G[l,m]^3','l'=1..n)','m'=1..n)/(n^2);
  curtosis:=sum('G[l,l]^2','l'=1..n)/n;
```

Contraste de igualdad de los autovalores, desde el a+1 hasta el a+h:

```
> a:=2:
  h:=2:
  prod:=1:
  lambdamedia:=sum('lambda[j]','j'=a+1..a+h)/h:
  Test1:=-n*ln(product('lambda[l]','l'=a+1..a+h)/lambdamedia^h);
  f:=(h-1)*(h+2)/2;
```

Contraste de igualdad de los últimos autovalores, desde el a+1 hasta el p:

```
> a:=2:
  h:=p-a:
  prod:=1:
  Test2:=-(n-1-a-(2*h^2+h+2)/6*h+lambdamedia^2*sum('1./(lambda[j]
  -lambdamedia)^2','j'=1..a))*ln(product('lambda[l]','l'=a+1..p)/
  lambdamedia^(p-a));
```

Cálculo de autovalores (ordenados en sentido decreciente) y autovectores (signo de forma que el coseno con el autovector calculado con la muestra completa sea positivo), tras la omisión de una observación y cálculo de la estimación del sesgo condicionado de autovalores y autovectores mediante la omisión de observaciones.

```
> SClambda:=array(1..n,1..p):
  Eomis:=E(n-1):
  Aomis:=array(1..n,1..p,1..p):
  for i from 1 to n do
    Xomis:=delrows(X,i..i):
    T2:=T(S(Xomis)):
    lambdaomis:=vector(p,k->T2[k,1]):
    for k from 1 to p do
      v:=subvector(A,k,1..p):
      z:=op(T2[k,3]):
      if dotprod(v,z)<0 then
```

```
      z:=scalarmul(z,-1):
    fi:
    SClambda[i,k]:=lambda[k]-lambdaomis[k]:
    for j from 1 to p do
      Aomis[i,k,j]:=z[j]:
      SCalfa[i,k,j]:=A[k,j]-Aomis[i,k,j]:
    od:
  od:
od:
```

Coeficiente gamma (γ_{jk}):

```
> gam:=array(1..p,1..p):
  for k from 1 to p do
    gam[k,k]:=0.:
    for j from 1 to k-1 do
      gam[j,k]:=1./(lambda[j]-lambda[k]):gam[k,j]:=-gam[j,k]:
    :od:
  od:
```

Coeficientes de los desarrollos de los autovalores y autovectores tras la omisión y estimación truncada del sesgo condicionado de los autovalores y autovectores de S

```
> SClambda_trunc:=array(1..n,1..p):
  nhu:=array(1..n,1..p):
  pi:=array(1..n,1..p):
  beta:=array(1..n,1..p,1..p):
  vgamma:=array(1..n,1..p,1..p):
  for i from 1 to n do
    for k from 1 to p do
      nhu[i,k]:=CP[i,k]^2-lambda[k];
      b21[i,k]:=sum('CP[i,j]^2*gam[j,k]','j'=1..p);
      b22[i,k]:=sum('CP[i,j]^2*gam[j,k]^2','j'=1..p);
      pi[i,k]:=-2*CP[i,k]^2*(1+b21[i,k]);
      for l from 1 to p do
        beta[i,k,l]:=-CP[i,k]*sum('gam[j,k]*CP[i,j]*A[j,l]',
        'j'=1..p):
        vgamma[i,k,l]:=-CP[i,k]^2*b22[i,k]*A[k,l]-2*b21[i,k]*
        beta[i,k,l]-2*CP[i,k]^3*sum('CP[i,j]*gam[j,k]^2*A[j,l]',
        'j'=1..p);
      od;
```

```
      SClambda_trunc[i,k]:=nhu[i,k]/(n-2.)-pi[i,k]/(2.*(n-2.)^2):
      for l from 1 to p do
        SCalfa_trunc[i,k,l]:=beta[i,k,l]/(n-2)-vgamma[i,k,l]/(2*
        (n-2)^2):
      od:
    od;
    od;
```

Estimación de la varianza de los autovalores:

```
> VARlambda:=vector(p,k->2*lambda[k]^2*(1.-sum('(lambda[j]*gam[j,k])^2',
'j'=1..p)/(n-1))/(n-1):
```

Cálculo de las medidas de influencia de cada observación sobre cada autovalor:

```
> Autoval1:=fopen(autoval1,WRITE):
  for i from 1 to n do
    for k from 1 to p do
    M1_t[i,k]:=abs(SClambda_trunc[i,k])/lambda[k];
    M2_t[i,k]:=abs(SClambda_trunc[i,k])/sumalambda:
    M3_t[i,k]:=abs(SClambda_trunc[i,k])/sqrt(VARlambda[k]):
    M1_tcuadrado[i,k]:=M1_t[i,k]^2:
    M2_tcuadrado[i,k]:=M2_t[i,k]^2:
    M3_tcuadrado[i,k]:=M3_t[i,k]^2:
    fprintf(Autoval1,'%06.12f %06.12f %06.12f ',M1_tcuadrado[i,k],
    M2_tcuadrado[i,k],M3_tcuadrado[i,k]):
  od:
  fprintf(Autoval1,' \n'):
  od:
  fclose(Autoval1):
```

Cálculo de las medidas de influencia de cada observación sobre cada autovector y cálculo de los coeficientes de cada autovector en la medida M_4^2:

```
> Autovect1:=fopen(autovect1,WRITE):
  v:=vector(p):
  M4_tcuadrado:=array(1..n,1..p):
  M5_tcuadrado:=array(1..n,1..p):
  for i from 1 to n do
    for k from 1 to p do
      v:=vector(p,l->SCalfa_trunc[i,k,l]):
      for j from 1 to p do alfa:=subvector(A,j,1..p):
```

```
      fprintf(Autovect1,'%06.12f ',dotprod(v,alfa)^2):
    od:
    M4_tcuadrado[i,k]:=norm(v,2)^2;
    M5_tcuadrado[i,k]:=innerprod(v,SX,v)*(n-1);
    fprintf(Autovect1,'%06.12f %06.12f  ',M4_tcuadrado[i,k],
    M5_tcuadrado[i,k]):
  od:
  fprintf(Autovect1,' \n'):
od:
fclose(Autovect1):
```

Cálculo de las medidas de influencia de cada observación para el conjunto de los k primeros autovalores:

```
> k:=2:
  Autoval2:=fopen(autoval2,WRITE):
  VARsumlambda:=sum('2*lambda[j]^2/(n-1)','j'=1..k):
  M6_tcuadrado:=vector(n,i->(sum('SClambda_trunc[i,j]','j'=1..k)/
  sum('lambda[j]','j'=1..k))^2):
  M7_tcuadrado:=vector(n,i->(sum('SClambda_trunc[i,j]','j'=1..k)/
  sumalambda)^2):
  M8_tcuadrado:=vector(n,i->(sum('SClambda_trunc[i,j]','j'=1..k)^2/
  VARsumlambda)):
  for i from 1 to n do
    fprintf(Autoval2,'%06.12f %06.12f %06.12f \n ',M6_tcuadrado[i],
    M7_tcuadrado[i],M8_tcuadrado[i]):
  od:
  fclose(Autoval2):
```

Cálculo de las medidas de influencia de cada observación para un conjunto de k autovectores:

```
> k:=2:
  Autovect2:=fopen(autovect2,WRITE):
  for i from 1 to n do fprintf(Autoval2,'%06.12f %06.12f \n ',
    'j'=1..k)):
  od:
  fclose(Autoval2):
```

Apéndice B

Datos de la aplicación 2

Los datos generados aleatoriamente para la Aplicación 2, indicando en primer lugar el numero de caso, son:

1	35.42541070654989	11.20862456038594	-32.09772992134095	13.03353436663747
2	70.02301284670830	12.15843459963799	-17.02186489105225	12.63386866450310
3	51.65135371685030	66.52320116758350	-47.33985239267350	13.32905709743500
4	29.84964580833912	66.58596475422380	-61.88358056545260	12.78132553398609
5	25.54298692941666	61.85139065980910	-65.50106167793270	12.34371465444565
6	44.45352977514267	47.93619132041930	-40.56964820623398	13.19811666011811
7	47.51711143553260	38.30237002670765	-44.42642408609391	12.48836457729340
8	35.02833893895149	31.10606399178505	-43.48314368724823	13.29442694783211
9	42.04215953499079	64.48956654220820	-35.69651532173157	13.08020947128535
10	43.66012497246266	25.60587559640408	-32.02453055977821	13.59675765037537
11	38.86020787805319	35.76974340528250	-44.73294675350190	12.95555552095175
12	51.33037638664250	60.37483906745910	-38.32347917556763	13.50344324111939
13	44.51450380682946	63.42297118902210	-49.27575835585590	13.23495110869408
14	46.31646626256410	45.86719490773980	-54.92437442764640	12.66576457023621
15	42.90070939064026	80.83385658264160	-58.79978489875790	13.35422122478485
16	54.19745898246770	55.20984923839570	-37.24890065193176	13.41442883014679
17	42.50533736124635	31.01852900162339	-38.48585122823715	12.97703889384866
18	27.75385724566877	51.59681216441100	-67.17060287296770	13.02048101089895
19	32.04274947941303	68.98744936287400	-65.28971111774450	12.82911075651646
20	45.68834930658340	48.83952111005780	-32.71005356311798	13.72073328495026
21	45.96373665332790	33.94772630929947	-55.49317926168440	12.82333281636238
22	21.50929284095764	61.67337821424010	-61.64587305486200	13.24274027347565
23	39.28905642032623	40.42696511745453	-48.45947504043580	12.75231364369393
24	41.96445956826211	58.55376395583150	-53.04009753465650	12.86072143912316
25	30.73603349924088	59.30452936887740	-71.54076659679410	12.71917086839676

26	43.80859744548798	21.19551002979279	-36.24077248573303	13.82750546932221
27	42.20312248170376	73.86509881913660	-63.21894860267640	12.83843354880810
28	42.70775775611401	57.20679695904260	-62.20379844307900	13.78449052572251
29	48.04141428973530	62.23423895146700	-67.03695851564410	12.98839342501015
30	54.43972468376160	37.03937482833863	-48.46165871620180	12.69472682476044
31	17.55498432368040	45.34389483183620	-60.63200557231900	12.97235321253538
32	42.56267310678959	55.92879150807860	-60.33688363432880	14.06596374511719
33	42.32523596286774	63.77305799722670	-73.40978878736500	13.26890036463738
34	48.06636205315590	23.83684673905373	-28.15482783317566	13.46907451748848
35	51.86152501404290	36.73246596753598	-51.41029527783390	13.58208161592484
36	52.20092192292210	36.31714162230492	-31.82981091737747	13.45858088135719
37	36.03020392358303	17.95314966142178	-47.46925044059750	13.24049066007137
38	47.52081772685050	31.98419511318207	-47.12349072098730	13.09400960803032
39	43.63512390851975	56.43183663487440	-50.30715402960780	13.55967378616333
40	42.54122617840767	55.05803307890890	-46.00559836626050	12.72699490189553
41	56.90959130972620	35.14749085158110	-27.00090861320496	13.06869947165251
42	30.35844704508782	44.66225662827492	-51.71868377923970	13.29252085089684
43	41.18933725357056	81.77456068992620	-62.37550520896910	13.87882196903229
44	44.23386687040329	52.60390125215050	-49.45104698836800	13.50641530752182
45	39.59388347342611	79.43973028287290	-72.96842876821760	13.19581611454487
46	60.63117864727970	36.43041071295739	-33.80287319421768	12.61788275837898
47	45.26723757386210	28.15551546216011	-37.32484567165375	13.37059816718102
48	41.48187491297722	30.02649071812630	-42.94036251306534	12.56958347558975
49	30.49429394715336	89.00699014014860	-75.65999828737680	13.04332010161457
50	70.94655359722674	10.07357431389390	-78.00257229804990	16.97610423900187

Apéndice C

Datos de la aplicación 3

Los datos generados aleatoriamente en la Aplicación 3, indicando en primer lugar el numero de caso, son:

1	20.20679032802582	15.20120619982481	17.56443162262440	15.28342029452324
2	25.80899882316590	8.56989522731731	18.61912792921066	14.95742040872574
3	24.16654276847840	16.97434258460999	19.96766906976700	17.22319209575653
4	9.99885574562109	13.98795320391660	16.13303371146321	13.61470787972212
5	16.15612876415253	11.16228792667390	12.10647976398468	10.17942130565643
6	21.29769474267960	16.18869996070862	10.44272369146347	8.45662081241608
7	20.71611945331097	11.93018746376038	14.43241080641747	11.60769432783127
8	16.97736859321594	16.76656168699265	12.66112815961242	10.12536446005106
9	23.18186020851136	15.48125682771206	11.06102949380875	9.33346664905548
10	19.45681355893612	18.58054590225220	16.02318595349789	13.37823322415352
11	18.32117176055908	14.73333312571049	12.55139626562595	9.83025130629540
12	18.01699842082827	18.02065944671631	15.75425866246223	13.15348535776138
13	22.05237412452698	16.40970665216446	15.50975425541401	12.62986750900746
14	20.06427325494588	12.99458742141724	13.36661662906408	10.78984262049198
15	22.53031408786774	17.12532734870911	13.88267779350281	11.85102415084839
16	24.56806445121765	17.48657298088074	8.917380928993230	7.41190004348755
17	25.29024842381478	14.86223336309195	10.53631669282913	8.37657177448273
18	15.09694230556488	15.12288606539369	14.30513855814934	12.08702385425568
19	18.25510054826736	13.97466453909874	15.39294098317623	12.64962637424469
20	18.01150281666429	19.32439970970154	16.57164575159550	13.46059508621693
21	18.54578021168709	10.50701648896534	12.66586878895760	9.81139582395554
22	15.36865222454071	16.46644164085390	15.17922848463059	12.98674619197846
23	18.58762910962105	13.52388186216350	12.94653204083443	10.72597515583038
24	20.65115791559220	14.16432863473892	13.30285210907459	10.60227367281914
25	16.23092412948608	13.31502521038055	17.74691173434258	15.14251786470413
26	18.45437955856323	19.96503281593323	15.94380670785904	13.12853959202767

27	21.28189939260483	14.03060129284859	15.78034099936486	13.70628577470779
28	19.34056027233601	19.70694315433502	17.67420297861099	14.80630075931549
29	20.81276986002922	14.93036055006087	15.07059053331614	12.34133691340685
30	21.59699833393097	13.16836094856262	17.55110046267510	14.85256391763687
31	12.87831091880799	14.83411927521229	19.29883199930191	16.41465675830841
32	19.27992920577526	21.39578247070313	11.97804754227400	9.59331126511097
33	18.91189333796501	16.61340218782425	14.88816049695015	12.64609569311142
34	20.71194660663605	17.81444710493088	13.60550412535667	10.90778106451035
35	20.36655287444592	18.49248969554901	12.35383436083794	10.44217333197594
36	22.61260354518890	17.75148528814316	9.916378766298300	7.82027167081833
37	15.31364834308624	16.44294396042824	11.01143294514623	8.51333598257042
38	19.51336789131165	15.56405764818192	14.30486610531807	11.52528136968613
39	20.85732322931290	18.35804271697998	15.55983524024487	12.96352919936180
40	21.21710026264191	13.36196941137314	15.40770930796862	12.92902229726315
41	25.58365642190021	15.41219682991505	14.10465681552887	12.05014783143997
42	16.53143382072449	16.75512510538101	15.21060500014574	12.71724460832775
43	21.68236267566681	20.27293181419373	14.55881321430207	11.61541461944580
44	20.67792689800263	18.03849184513092	14.71181324124336	12.42705780267716
45	20.11906890198589	16.17489668726921	16.25203701853752	13.82878309488297
46	20.26308573459176	12.70729655027390	14.87716226279736	12.53761371970177
47	19.58370560407639	7.223589003086090	17.77274239063263	14.51863372325897
48	18.59345862269402	12.41750085353851	14.28567035496235	11.57845345139504
49	15.50372469425201	15.99379509687424	16.40909698605537	13.45757001638413
50	15.89700371855639	14.85662543401122	15.40390487760305	12.83170779049397
51	13.06911897659302	15.73705530166626	18.93846826255322	16.24593642354012
52	15.23628091812134	13.22288614511490	13.07793736457825	11.73341608047485
53	21.22528162598610	18.51054060459137	14.32662242650986	12.32666015625000
54	21.45961719751358	17.44882732629776	14.74358745664358	12.36749567091465
55	21.70362865924835	11.79978275299072	15.36739063262940	12.66019776463509
56	17.70386064052582	16.95715320110321	16.19537815451622	13.38756543397904
57	19.33965717256069	16.36064359545708	10.95837275683880	8.24622085690498
58	16.50163519382477	12.05463355779648	15.59262507408857	12.76370056718588
59	16.18052613735199	13.28201705217362	14.53753951191902	11.67223042249680
60	14.68538630008698	15.02720938658084	17.34525620937348	14.00587666034699
61	17.70375764369965	19.77803492546082	11.00353395566344	8.54271737486124
62	18.47444677352905	17.75268226861954	16.04697746038437	13.66751992702484
63	20.43251653015614	13.91494667530060	12.84444340318442	10.29499174654484
64	22.18455857038498	15.21289238942904	15.32695207931102	12.81400271877647
65	23.10145580768586	15.62123094499111	15.42821399867535	12.59447358548641
66	18.62733438611031	20.53852808475495	15.92861090600491	13.22481891512871

67	14.51853740215300	14.85338123515248	21.39051979780197	18.50902903079987
68	22.27572864294052	10.27049374580380	13.72432039678097	11.08601084351540
69	22.85722422599793	10.60257482528687	12.86004085838795	10.44302853941918
70	15.01389124393460	10.05287893429051	17.02910411357880	13.99551123380661
71	22.05088138580322	15.67652821540833	16.33677176758647	13.87276978045702
72	19.68904263526201	10.12421854032198	17.06086209416390	13.73550018668175
73	18.95649829506874	13.87484246492386	18.23826691508293	15.54222041368485
74	21.02526366710663	10.67239093780518	10.03216671943665	7.80402016639710
75	24.24939799308777	16.05164998769760	15.07235226035118	12.32531318068504
76	14.78755247592930	17.42763447761536	10.35562855005264	8.42468535900116
77	24.21267282962799	13.89068806171417	14.64339715242386	12.55340766906738
78	23.55071365833283	17.11308395862579	14.63923192396760	12.17655002325773
79	25.85100966789289	13.61264669895172	18.58072549104691	15.21777415275574
80	21.68979704380036	14.36433796584606	12.53457864373923	10.65975125879049
81	18.45874297618866	12.48527759313583	15.40307151898742	12.60738169774413
82	22.91756087541580	15.89646565100937	19.48855324089527	16.76984593272209
83	21.75566695450011	23.51990675926209	15.78686350584030	12.98625528812409
84	13.39283347129820	14.11269378662110	11.90660911798477	9.78184592723846
85	21.86998802423477	14.06755596399307	11.99570354819298	9.63606935739517
86	16.40218496322632	14.08158105611801	14.76565623283386	13.02141904830933
87	17.60627937316895	13.31066519021988	14.67316880822182	12.57812243700028
88	19.87807361781597	19.72634768486023	13.90565791726112	11.46447992324829
89	20.38206221163273	13.10503667593002	13.99609772861004	11.31642881035805
90	19.10637924075127	13.89449653029442	14.22863513231278	10.92405641078949
91	15.95573413372040	14.73802236467600	19.11062583327294	16.25604945421219
92	14.20520913600922	24.89087677534511	11.20931529998779	8.79941582679749
93	23.71158325672150	12.31711202859879	14.95998043939471	12.51460272818804
94	20.19612086564303	12.23517680168152	15.98092269897461	13.70419239997864
95	20.00529731068943	21.38073348999023	16.55069521069527	14.40007776021958
96	18.67679378390312	11.81236660480499	11.10884124040604	9.27211642265320
97	16.56950545310974	11.33105957508087	14.55278646945953	11.46788620948792
98	17.65797126293183	11.54370510578156	15.90589186549187	12.82303813099861
99	22.93581634759903	16.30595755577087	14.46206609904766	11.72082719206810
100	24.64693570137024	15.82187139987946	8.38490530848503	5.98423892259598

Referencias bibliográficas

[1] ANDREWS, D.F.; GNANADESIKAN, R.; WARNER, J.L. (1972). Methods for Assessing Multivariate Normality. *Bell Laboratories Memorandum.*

[2] ATKINSON, T.W. (1963). Asymptotic Theory for Principal Component Analysis. *Ann. Math. Statist.*, 34, 122-148.

[3] BECKMAN, R.J.; NACHTSHEIM, C.J.; COOK, D. (1987). Diagnostics for Mixed-Models Analysis of Variance. *Technometrics,* 29, 413-426.

[4] BELSLEY, D.A.; KUH, E.; WELSH, R.E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* New York: John Wiley & Sons.

[5] BÉNASSÉNI, J. (1988). Sensitivity of Principal Component Analysis to Data Perturbation. *Data Analysis and Informatics, V.* Editado por Diday. Elsevier Science Publishers, B.V. North Holand. 303-310.

[6] BÉNASSÉNI, J. (1990). Sensitivity Coefficients for the Subspaces Spanned by Principal Components. *Commun. Statist.-Theory Meth.*, 19, 2021-2034.

[7] BROOKS, S.P. (1994). Diagnostics for Principal Components: Influence Functions as Diagnostic Tools. *The Statistician* 43, 483-494.

[8] CALDER, P. (1984). Influence Functions for Principal Components and their Variances. Trabajo no publicado. Mathematical Institute, University of Kent at Canterbury.

[9] CALDER, P. (1986). Influence Functions in Multivariate Analysis. Tesis Doctoral. University of Kent at Canterbury.

[10] CAMPBELL, N. A. (1978). The Influence Function as an Aid in Outlier Detection in Discriminant Analysis. *J. Appl. Statist.*, 27 , 251-258.

[11] CASTAÑO TOSTADO, E.; TANAKA, Y. (1990). Some Comments on Escoufier's RV- Coefficient as a Sensitivity Measure in Principal Component Analysis. *Commun. Statist.-Theory Meth.*, 19, 4619-4626.

[12] CHATTERJEE, S.; HADI, A.S. (1988). *Sensitivity Analysis in Linear Resgression.* New York: John Wiley.

[13] COOK, R.D.; WEISBERG, S. (1982). *Residual and Influence in Regression.* Chapman & Hall.

[14] COOK, R.D. (1986). Assessment of Local Influence. *Journal of Statistical Society,* Series B, 48, 133-169 (con discusión).

[15] CRITCHLEY, F. (1985). Influence in Principal Component Analysis. *Biometrika*, 72, 627-636.

[16] DAUDIN, J.J.; DUBY, C.; TRECOURT, P. (1988). Stability of Principal Component Analysis Studied by the Bootstrap Method. *Statistics*, 19, 241-258.

[17] DEVLIN, S.J.; GNANADESIKAN, R.; KETTENRING, J.R. (1975). Robust Estimation and Outlier Detection with Correlation Coefficients. *Biometrika*, 62, 531-545.

[18] ESCOUFIER, Y. (1973). Le Traitement des Variables Vectorielles. *Biometrics*, 29, 751-760.

[19] FLURY, B. (1988). *Common Principal Components & Related Multivariate Models.* New York: John Wiley and Sons.

[20] FUNG,W.-K. (1992). Joint Influence Function and Multicollinearity. *Computational Statistics,* 7, 81-89.

[21] GIRSHICK, M.A. (1939). On the Sampling Theory of Roots of Determinantal Equations. *Ann. Math. Statist.,* 10, 203-224.

[22] GNANADESIKAN, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations.* New York: John Wiley & Sons.

[23] HAMPEL, F.R. (1968). Contributions to the Theory of Robust Estimation. Tesis Doctoral. University of California at Berkeley.

[24] HAMPEL, F.R. (1973). Robust Estimation: A Condensed Partial Survey. *Z. Wahr. verw. Geb,* 27, 87-104.

[25] HAMPEL, F.R. (1974). The Influence Curve and its Role in Robust Estimation. *J. Amer. Statist. Assoc.*, 69, 383-393.

[26] HAMPEL, F.R.; RONCHETTI, E.M.; ROUSSEEUW, P.J.; STAHEL, A.W. (1986). *Robust Statistics: The Approach Based on Influence Functions.* New York: John Wiley & Sons.

[27] HARVILLE, D.A. (1997). *Matrix Algebra from a Statistician's Perspective.* Springer.

[28] HOTELLING, H. (1933). Analysis of a Complex of Statistical Variables into Principal Components. *J. Educ. Psychol.*, 24, 471-441.

[29] HUBER, P. (1981). *Robusts Statistics.* New York: John Wiley & Sons.

[30] JACKSON, J.E. (1991). *A User's Guide to Princpal Components.* New York: John Wiley & Sons.

[31] JAMES, A.T. (1969). Test of Equality of Latent Rootsof the Covariance Matrix. En P.R. Krishnaiah (Ed.), *Multivariate Analysis*, Vol II, 205-218. Academic Press: New York.

[32] JOHNSON N.L.; KOTZ, S. (1972). *Distributions in Statistics: Continuous Multivariate Distribution.* John Wiley and Sons.

[33] JIMÉNEZ GAMERO, M.D. (1994). Análisis de Muestras Generadas en el Proceso de Simulación Bootstrap. Tesis Doctoral. Universidad de Sevilla.

[34] JIMÉNEZ GAMERO, M.D.; MUÑOZ PICHARDO, J.M.; MUÑOZ REYES, A. (1995). Medidas de Influencia en las Estimaciones Bootstrap. *Actas del XXII Congreso Nacional de Estadística e Investigación Operativa.*

[35] JOLLIFFE, I.T. (1986). *Principal Component Analysis.* New York: Springer-Verlag.

[36] KATO, T. (1980). *Perturbation Theory for Linear Operators.* Springer.

[37] KENDALL, M. G. (1980). *Multivariate Analysis.* (2^a edición). Charles Griffin & Company LTD.

[38] KOTZ, S.; JOHNSON, N.L. (1982). *Enciclopedy of Statistical Sciences.* New York: John Wiley and Sons.

[39] KRZANOWSKI, W.J. (1984). Sensitivity of Principal Components. *J. Roy. Stat. Soc.*, Series B, 46, 558-563.

[40] LAWLEY, D.N. (1956). Test of Significance for the Latent Roots of Covariance and Correlation Matrices. *Biometrika* 43, 128-36.

[41] LAWRENCE, A.J. (1988). Regression Transformations Diagnostics Using Local Influence. *J. Am. Statist. Assoc.*, 83, 1067-1072.

[42] MALLOWS, C.L. (1973). Influence Functions. Conferencia no publicada presentada en la Working Conference on Robust Regression en el National Bureau of Economics Research en Cambridge, Mass.

[43] MALLOWS, C.L. (1975). On some Topics in Robustness. Trabajo no publicado. *Bell Telephone Laboratories report.* Murray Hill, N.J.

[44] MALLOWS, C.L. (1976). On some Topics in Robustness. *Bell Laboratories Memorandum.*

[45] MARDIA, K.V. (1970). Measures of Multivariate Skewness and Kurtosis with Applications. *Biometrika* 57, 519-530.

[46] MERTENS, B.J.A. (1998). Exact Principal Component Influence Measures Applied to the Analysis of Spectroscopy Data on Rice. *J. Appl. Statist.*, 47, 527-542.

[47] MILLER, K.S. (1964). *Multidimensional Gaussian Distributions.* New York: John Wiley & Sons.

[48] MORENO REBOLLO, J.L.; MUÑOZ REYES, A.; MUÑOZ PICHARDO, J.M. (1999). Influence Diagnostic in Survey Sampling: Conditional Bias. *Biometrika*, 86, 923-928

[49] MUIRHEAD, R.J. (1982). *Aspects of Multivariate Statistical Theory.* New York: John Wiley & Sons.

[50] MUÑOZ PICHARDO, J.M; MUÑOZ GARCÍA, J.; MORENO REBOLLO, J.L.; PINO MEJÍAS, R. (1995). A New Approach to Influence Analysis in Linear Models. *Sankhyā: The Indian Journal of Statistics*, Series A, 57, 393-409.

[51] MUÑOZ PICHARDO, J.M; MUÑOZ GARCÍA, J.; FERNÁNDEZ PONCE, J.M.; LÓPEZ BLÁZQUEZ, F. (1998). Local Influence on the General Linear Model. *Sankhyā: The Indian Journal of Statistics*, Series B, 60, 269-292

[52] MUÑOZ PICHARDO, J.M; MUÑOZ GARCÍA, J.; FERNÁNDEZ PONCE, J.M.; JIMÉNEZ GAMERO, M.D. (2000). Influence Analysis on Multivariate Linear General Models. *Commun. Statist.-Theory Meth.*, 29, 529-547.

[53] PACK, P.; JOLLIFFE, I.T; MORGAN, B.J.T. (1988). Influential Observations in Principal Component Analysis: a Case Study. *J. Appl. Statist.*, 15, 39-52.

[54] PEARSON, K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space. *Philos. Mag.*, 2, 559-572.

[55] RADHAKRISHNAN, R.; KSHIRSAGAR, A.M. (1981). Influence Functions for Certain Parameters in Multivariate Analysis. *Commun. Statist.-Theory Meth.*, Series A, 10, 515-529.

[56] RELLICH, F. (1969). *Perturbation Theory of Eigenvalue Problems.* Gordon and Breach.

[57] ROBERT, P.; ESCOUFIER, Y. (1976). A Unifying Tool for Linear Multivariate Statistical Methods: the RV Coefficient. *J. Appl. Statist.*, 25, 257-265.

[58] SEBER, G.A.F. (1984). *Multivariate Observations.* New York: John Wiley & Sons.

[59] SHI, L. (1997). Local Influence in Principal Components Analysis. *Biometrika*, 84, 175-178.

[60] SIBSON, R. (1979). Studies in the Robustness of Multidimensional Scaling: Perturbational Analysis of Classical Scaling. *J.R. Statist. Soc.*, Series B, 41, 217-229.

[61] TANAKA, Y. (1984). Sensitivity Analysis in Hayashi's Third Method of Quantification. *Behaviormetrika*, 16, 31-44.

[62] TANAKA, Y. (1988). Sensitivity Analysis in Principal Component Analysis: Influence on the Subspace Spanned by Principal Components. *Commun. Statist.-Theory Meth.*, 17, 3157-3175.

[63] TANAKA, Y. (1989). Sensitivity Analysis in Principal Component Analysis: Influence on the Subspace Spanned by Principal Components. *Commun. Statist.-Theory Meth.*, 18, 4305. (Corrección de [62]).

[64] TANAKA, Y.; CASTAÑO TOSTADO, E. (1990). Quadratic Perturbation Expansions of Certain Functions of Eigenvalues and Eigenvectors and their Applications to Sensitivity Analysis in Multivariate Methods. *Commun. Statist.-Theory Meth.*, 19, 2943-2965.

[65] TANAKA, Y.; TARUMI, T. (1986). Sensitivity Analysis in Hayashi's Secon Method of Quantification. *J.Japan Statist. Soc.* 16, 37-52.

[66] TANAKA, Y.; TARUMI, T. (1988). A Numerical Investigation on Sensitivity. Analysis in Multivariate Methods. *Data Analysis and Informatics, V.* Editado por Diday. Elsevier Science Publishers, B.V. North Holand. 291-301.

[67] TARUMI, T. (1986). Sensitivity Analysis of Descriptive Multivariate Methods Formulated by Generalized Singular Value Descomposition. *Math. Japonica,* 31, 957-977.

[68] TARUMI, T.; TANAKA, Y. (1986). Statistical Software SAM- Sensitivity Analysis in Multivariate Methods. *COMPSTAT*, 351-356.

[69] THOMAS, W.; COOK, R.D. (1990). Assessing Influence on Predictions from Generalized Linear Models. *Technometrics*, 32, 59-65.

[70] WANG, S-G.; NYQUIST H. (1991). Effects on the Eigenstructure of a Data Matrix when Deleting an Observation. *Computational Statistics & Data Analysis* 11. 179-188.

[71] WILKINSON, J.H. (1965). *The Algebraic Eigenvalue Problem.* Oxford University Press.

[72] WISHART, J. (1928). The Generalised Product Moment Distribution in Samples from a Normal Multivariate Population. *Biometrika* 20A, 32-52.

UNIVERSIDAD DE SEVILLA

Reunido el Tribunal integrado por los abajo firmantes en el día de la fecha, para juzgar la Tesis Doctoral de

D. Alicia Enguix González

titulada Análisis de influencia en componentes principales

y habiendo otorgado la calificación de Sobresaliente Cum laude

por unanimidad

Sevilla, 16 de Julio de 2001

El Vocal,

El Vocal,

El Vocal,

El Presidente

El Secretario,

El Doctorando,

Alicia Enguix