UNIVERSIDAD DE SEVILLA

FACULTAD DE MATEMÁTICAS

DEPARTAMENTO DE ESTADÍSTICA E
INVESTIGACION OPERATIVA

# ALGUNAS APORTACIONES A LOS MÉTODOS DE OPTIMIZACIÓN DEL ANÁLISIS CLUSTER MEDIANTE LA D.V.S.

Juan Luis González Caballero

Cádiz, Abril 1996

**UNIVERSIDAD DE SEVILLA**

FACULTAD DE MATEMÁTICAS

# ALGUNAS APORTACIONES A LOS METODOS DE OPTIMIZACION DEL ANALISIS CLUSTER MEDIANTE LA D.V.S.

Memoria presentada por
**Juan Luis González Caballero**
para optar al grado de
Doctor en Ciencias Matemáticas.

Doctorando,

**D. Juan L. González Caballero**

V°.B°. Ponente,

**Dr. D. Rafael Infante Macías**
Catedrático de Estadística e I.O.
Universidad de Sevilla.

V°.B°. Director,

**Dr. D. Mariano J. Valderrama Bonnet**
Catedrático de Estadística e I.O.
Universidad de Granada.

V°.B°. Director,

**Dr. D. José Ramírez Labrador**
Profesor Titular de Análisis Matemático
Universidad de Cádiz.

Cádiz, Abril 1996

# Introducción

La Descomposición en Valores Singulares (DVS) de una matriz $X_{n,p}$ es un resultado descrito por primera vez en 1889 por el matemático inglés Sylvester, que permite descomponer $X$ utilizando las direcciones principales obtenidas en los espacios $\mathbb{R}^p$ y $\mathbb{R}^n$ donde pueden representarse los vectores fila y columna, respectivamente, de la matriz $X_{n,p}$.

Su importancia estadística se debe a Eckart y Young (1936) y Householder y Young (1938), que mostraron la utilidad de la DVS para obtener el mejor ajuste, en el sentido de mínimos cuadrados, de la matriz $X$ por una de rango menor.

La DVS, ó la descomposición espectral de una matriz cuadrada que es un caso particular de ella, es el fundamento de muchas de las técnicas de reducción y representación de datos, como el Análisis de Componentes Principales, véase por ejemplo Jolliffe (1986) ó Jackson (1990), algunos Modelos Factoriales descritos, por ejemplo, en Reyment y Jöreskog (1993) ó en Jambu (1991), el Análisis de Correspondencias, descrito por Greenacre (1984), ó la técnica de Representación Biplot, iniciada por Gabriel (1971), entre otras.

En los últimos años, pueden encontrarse numerosas referencias en la literatura sobre Análisis Cluster (Gnanadesikan (1977), Gordon (1981), Everitt (1993)), en las que se sugieren el empleo de procedimientos geométricos de representación de los datos que, en la mayoría de los casos, se refieren a alguna de las técnicas mencionadas anteriomente. Aunque ninguna de ellas está específicamente diseñada para indicar la presencia de clusters entre los datos, se utilizan para este propósito conjuntamente con otros procedimien-

tos de obtención de grupos homogéneos, no sólo para indicar su presencia sino también para prevenir ante posibles clasificaciones erróneas producidas por técnicas de cluster más complejas.

Entre las técnicas mencionadas anteriormente, el Modelo Factorial es un término genérico que se utiliza para describir una variedad de modelos, diseñados para analizar las interrelaciones que existen dentro de un conjunto de variables o de elementos - individuos u objetos -. Su origen hay que buscarlo en los estudios que Pearson y Spearman realizaron a principios de este siglo, sobre el intento de medir la capacidad y el comportamiento humanos. Posteriormente, autores como Thurstone (1947), Cattell (1952,1965), Harmann (1976), Lawley y Maxwell (1963,1971) y Reyment y Jöreskog (1993) , entre otros, han ido aportando resultados nuevos al modelo, depurándolo en sus planteamientos en algunos casos, proponiendo o mejorando soluciones en otros, y, en general, construyendo todo aquello que en la actualidad sustenta la teoría en la que se basa.

Aunque el primer Modelo Factorial que surgió se utilizó para analizar variables (denominado en modo R), Burt (1917, 1937) y Stephenson (1936) introdujeron el Modelo Factorial para analizar los elementos (denominado en modo Q), y posteriormente aparecen otros tipos de Modelos Factoriales que pueden verse descritos en Cattell (1965).

El Modelo Q-Factorial, que en la mayoría de los casos sólo ha sido utilizado en Psicología para obtener tipos de individuos con determinados patrones de comportamiento (Fleiss y otros, 1971), se ha considerado también a veces como una técnica de clasificación (Overall y Klett, 1972) en la que los tipos de individuos obtenidos clasificaban a los originales. En cualquiera de los dos casos, su utilización ha sido muy criticada por numerosos autores como Baggaley (1964), Lorr (1966), Jones (1968), Fleiss y Zubin (1969) ó Fleiss y otros (1971).

La DVS permite la formulación conjunta de los Modelos Factoriales en modo R y Q, como lo introducen Reyment y Jöreskog (1993) en algunos casos particulares para analizar objetos geológicos. Esta formulación conjunta, que se encuentra también en Jambu (1991), puede evitar algunas de las críticas cuando se utiliza como una técnica clasificatoria en general, pero sigue teniendo algunos problemas.

En el capítulo primero de esta memoria se introducen los resultados más importantes de los dos tipos de Modelo Factorial mencionados y su formulación conjunta mediante la DVS. Además, en la sección correspondiente al Modelo Q-Factorial se recogen los problemas más importantes de este Modelo que han sido objeto de críticas.

En el capítulo segundo se comienza introduciendo el Análisis Cluster, haciendo un repaso breve de los tipos y procedimientos de clasificación más importantes, y deteniéndonos en los procedimientos de optimización del Análisis Cluster.

Las aportaciones principales de esta memoria se encuentran en las siguientes secciones del capítulo segundo y en las del capítulo tercero.

En las correspondientes al capítulo segundo, se propone un procedimiento de clasificación general (PROCED) que permite obtener grupos naturales en conjuntos donde se sospecha que existen más de uno, con la restricción de que tal número de grupos no sea superior a la dimensión de los datos más uno.

El procedimiento se basa en obtener el Modelo Q-Factorial derivado de la DVS de una matriz $W_{n,p+1}$, obtenida al transformar la matriz de datos original $X_{n,p}$. La nube de puntos de $I\!R^p$ formada por las filas de $X_{n,p}$, se centra por columnas para que el centro de gravedad de la nube se sitúe en el origen de coordenadas de $I\!R^p$, se introduce en un espacio de una dimensión más $I\!R^{p+1}$, se traslada en la dimensión $p+1$ a una cierta distancia del origen y se normaliza por filas hasta llegar a $W_{n,p+1}$. Estas transformaciones efectuadas sobre $X$ facilitan que los factores obtenidos con el Modelo Q-Factorial para $W_{n,p+1}$ y, sobre todo, los que se obtienen realizando rotaciones oblicuas con ellos, sean capaces de descubrir las posibles agrupaciones naturales que pueden existir en el conjunto de elementos.

El procedimiento se prueba analizando un conjunto de datos bien conocido: las 150 flores Iris de tres tipos (Setosa, Virgínica y Versicolor) que Fisher (1936) utilizó en problemas de discriminación, del que se llega a clasificar bien un 82.66 % del total.

La última sección del capítulo se dedica a un estudio de simulación del

procedimiento, en el que se genera de forma aleatoria una muestra de 100 conjuntos con 50 elementos cada uno entre los que existen grupos definidos, llegándose a obtener un porcentaje medio del 85.86 % de elementos bien clasificados, y finaliza con una discusión al compararlo con uno de los procedimientos de optimización del Análisis Cluster, el de minimizar la traza de la matriz de dispersión dentro de los grupos (MINTRAZ), que obtiene la mejor agrupación con este criterio conociendo el número de grupos y partiendo de una ordenación dada. En este análisis comparativo se llega a la conclusión de que, en general, los dos procedimientos se comportan de forma similar en cuanto a la agrupación de los elementos, si bien PROCED tiene la ventaja de que no necesita conocer el número de grupos previamente. Además, se comprueba con muestras generadas aleatoriamente que PROCED es capaz de obtener mejores resultados que MINTRAZ en determinados tipos de conjuntos.

En el capítulo tercero, se propone una modificación del procedimiento estudiado en el anterior, para obtener uno más general (PROCGEN) que no tenga la restricción de que el número de grupos no exceda la dimensión de los datos más uno. Para ello, se van extrayendo de forma iterada los elementos mejor agrupados en torno a uno de los factores obtenidos con los mismos pasos que en PROCED, terminándose el proceso cuando no queden elementos por clasificar.

El procedimiento se prueba con un ejemplo simulado donde los grupos están muy claramente separados, y en el que PROCGEN clasifica bien al 100 % de los elementos. El capítulo finaliza también realizando un estudio de simulación, donde PROCGEN es capaz de mejorar los resultados de PROCED con un 87.40 % de elementos bien clasificados, y una discusión al compararlo con MINTRAZ, donde se observa cómo a medida que el número de grupos es mayor PROCGEN es capaz de obtener mejores resultados que MINTRAZ, a pesar de que éste último procedimiento parte de la clasificación original.

La memoria concluye con un apartado donde se apuntan algunos de los problemas que han quedado abiertos al finalizarla y con dos Apéndices y una relación de referencias bibliográficas. El Apéndice A recoge el fichero de los datos de las flores Iris utilizado al introducir PROCED en el capítulo segundo. El Apéndice B recoge los programas elaborados para la realización

de los estudios de simulación y comparativos desarrollados en los capítulos segundo y tercero.

Para finalizar, quisiera manifestar mi profundo agradecimiento a todas aquellas personas que me han ayudado a la realización de esta memoria.

A todos mis compañeros del Departamento de Matemáticas, por su colaboración y estímulo durante la realización de esta memoria.

A mi familia, por la comprensión y paciencia que han tenido conmigo durante todo este tiempo.

Al Profesor D. Rafael Infante Macías, sin cuyo apoyo desinteresado este trabajo no hubiera visto la luz.

De manera especial, a los Profesores D. Mariano J. Valderrama Bonnet y D. José Ramírez Labrador, directores de esta memoria, por el constante apoyo y asesoramiento que siempre me han dado a lo largo de su elaboración.

Por último, y aunque ya no esté entre nosotros, a mi amigo y maestro el Profesor D. Antonino García Rendón, por la confianza que siempre depositó en mí.

Cádiz, Abril 1996

# Contenido

# Capítulo 1

# El Modelo Factorial

## 1.1 Introducción

El Modelo de Análisis Factorial (A.F.)[109] es un término genérico que se utiliza para describir una variedad de técnicas, diseñadas para analizar las interrelaciones que existen dentro de un conjunto de variables o de elementos - individuos u objetos -.

Aunque tales técnicas pueden diferir bastante en sus objetivos y en el modelo matemático subyacente en ellas, todas tienen un objetivo común : la obtención de un número pequeño de factores - variables o elementos teóricos no observables -, que se construyen de forma que contengan la información esencial existente en el conjunto de datos de partida.

Su origen hay que buscarlo en los estudios que Pearson y Spearman realizaron a principios de este siglo, sobre el intento de medir la capacidad y el comportamiento humanos. Los métodos que utilizaron intentaban analizar los resultados obtenidos en un grupo de individuos al pasarles una batería de tests psicológicos. Algunos de estos tests fueron diseñados para medir varios aspectos de la capacidad mental que se suponía debían están relacionados. La técnica que propusieron intentó "explicar" estas relaciones mediante la búsqueda de un número pequeño de factores que contuvieran la información

esencial sobre las relaciones existentes entre los tests.

A partir de aquí, han sido muchos los autores que a lo largo de la primera mitad de este siglo han ido aportando resultados nuevos al modelo, depurándolo en sus planteamientos en algunos casos, proponiendo o mejorando soluciones en otros, y, en general, construyendo todo aquello que en la actualidad sustenta la teoría en la que se basa. Pueden encontrarse la mayoría de las referencias que han ido recogiendo todas estas aportaciones en los manuales de Thurstone [126], Cattell [22], Harmann [61], Lawley y Maxwell [89] y Reyment et al. [109], entre otros.

En el modelo de A.F., los factores se construyen según un modelo lineal, de forma que permitan expresar a las variables o los elementos de partida mediante combinaciones lineales de tales factores - salvo términos de error -. De esta forma, si esto se consigue con un número pequeño en relación al número de variables o de elementos, se habrá conseguido reducir la complejidad total que presentan. En este sentido, puede decirse que la técnica permite reducir la dimensión de los datos.

Sin embargo, es evidente que un planteamiento tan sencillo del modelo no puede tener una solución fácil. Inicialmente, la técnica tiene infinitas soluciones, que dependerán de muchos aspectos que quedan abiertos en su planteamiento. Sólo añadiendo algunas hipótesis adicionales y/o algunos criterios, más o menos objetivos, pueden conseguirse obtener soluciones únicas, las cuales generalmente se obtienen con criterios de convergencia. En Harmann [61], se hace una extensa exposición de diferentes hipótesis y criterios que pueden utilizarse.

La diversidad de soluciones posibles, unido a que los factores que se obtienen nunca son directamente observables, y por lo tanto estarán sujetos a interpretaciones, hacen que esta técnica tenga detractores. Algunos investigadores no son partidarios de su utilización al carecer de una solución única, obtenida con criterios objetivos. Sin embargo, son muchos los que piensan que a partir de la introducción del concepto de *estructura simple* debido a Thurstone [125], se reduce bastante el campo de la subjetividad, y aunque este concepto no proporciona soluciones únicas, si permite obtener soluciones objetivas, en cierto sentido, para todos los investigadores.

En la introducción del capítulo hemos dicho que el A.F. puede utilizarse para analizar tanto variables como elementos, y conviene que distingamos ambos casos. El primero que surgió fué el análisis de las variables, al que se le denominó **Análisis Factorial en modo R**, al ser la matriz de correlaciones R la que se utilizaba para extraer los factores. El análisis de los elementos no apareció hasta que Burt [15] y [16] y Stephenson [120] lo introdujeron con la denominación **Análisis Factorial Invertido** o **Análisis Factorial en modo Q**, al denominar Q a la matriz de correlaciones entre los individuos. Posteriormente aparecen otros tipos de A.F. que pueden verse descritos en Cattell [23].

A pesar del tiempo transcurrido desde su aparición, son muy pocas las referencias existentes sobre el A.F. en modo Q. Es probable que se deba a dos hechos. Por un lado, a que su desarrollo está concebido, en principio, como una traslación al caso de los individuos de la mayoría de las técnicas desarroladas para las variables. Por otro lado, a los problemas de objetividad de las soluciones de este tipo de técnica se añaden otros, derivados de la dificultad de interpretación de conceptos que se definen para variables, pero no para individuos, o del intento de asimilar esta técnica con un procedimiento de Análisis Cluster. No obstante, algunos resultados obtenidos con posterioridad a su aparición, y la aplicación que se ha hecho de ella para analizar determinados problemas, aconsejan no abandonar por completo su utilización, si bien deben tenerse presente sus limitaciones.

En las dos secciones siguientes **1.2** y **1.3**, se hará una breve descripción de las dos técnicas, mostrando, sobre todo en el modo Q, algunos problemas que distintos autores han apuntado hasta el momento, y en la sección **1.4** se introduce una forma de analizar un conjunto de datos mediante las dos técnicas al mismo tiempo, utilizando la Descomposición en Valores Singulares de una matriz rectangular.

## 1.2 El modelo factorial para variables

### 1.2.1 El modelo, hipótesis, consecuencias

La idea básica que subyace en el modelo de A.F. es que las $p$ componentes de un vector aleatorio $\mathbf{x} = (x_1, ..., x_p)$ puedan expresarse, excepto un término de error, como funciones lineales de $m(< p)$ variables aleatorias hipotéticas o *factores comunes* $f_1, ..., f_m$, es decir

$$\begin{aligned} x_1 &= a_{11}f_1 + \cdots + a_{1m}f_m + e_1 \\ x_2 &= a_{21}f_1 + \cdots + a_{2m}f_m + e_2 \\ &\vdots \\ x_p &= a_{p1}f_1 + \cdots + a_{pm}f_m + e_p \end{aligned} \tag{1.1}$$

donde $a_{jk}, j = 1, 2, \ldots, p; k = 1, 2, \ldots, m$ son constantes a determinar denominadas *pesos factoriales*, y $e_j, j = 1, 2, \ldots, p$ son los términos error, llamados también *factores únicos* por ser para cada $j$, $e_j$ 'específico' de $x_j$, mientras que los factores $f_k$ son 'comunes' a todas las variables $x_j$.

Además de que $m < p$ (incluso mucho menor que $p$ para que las variables sean explicadas por un número reducido de factores), se impondrá en el modelo inicialmente que la totalidad de los $(m + p)$ factores sean variables incorreladas, para que así la parte de variabilidad de una cualquiera de las variables explicada por un factor, no tenga relación (en sentido lineal) con la que explican los demás factores.

Las ecuaciones (1.1) pueden escribirse en forma matricial mediante

$$\mathbf{x} = \mathbf{A}\mathbf{f} + \mathbf{e} \tag{1.2}$$

con lo que el modelo se construye, determinando la llamada matriz factorial (o *patrón factorial*)

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ a_{p1} & a_{p2} & \cdots & a_{pm} \end{pmatrix}$$

cuyos elementos, los pesos factoriales, representan la relación existente entre los factores comunes y las variables.

Sólo estos elementos van a tener interés en el modelo propuesto inicialmente, ya que los errores o factores únicos se incluyen en el modelo (1.1) dada la imposibilidad , en general, de expresar de forma exacta $p$ variables en función de un número reducido m de factores. De este modo, en estos factores únicos puede recogerse tanto la dispersión de cada variable que no son capaces de explicar los factores, como la variabilidad presente en la muestra de datos de la que partimos para construir el modelo, en el caso de que ésta provenga de una población. En los casos en los que el modelo no quiera extenderse a una población más amplia, los factores únicos sólo recogerán la primera variabilidad.

El hecho de que en la Estadística en general, las relaciones lineales entre variables se midan a través de las covarianzas o las correlaciones entre ellas, explica que para determinar el modelo (1.1) se utilicen unas u otras, contando, además, con aquellas restricciones que sea necesario imponer al problema superdeterminado de obtener la matriz patrón A. Esto hace además que los factores comunes deban entenderse como la dimensionalidad subyacente en la población o la muestra de datos, que relaciona y explica las relaciones y asociaciones existentes entre las variables que se miden en ellas.

Si a la simplificación de que los factores comunes y específicos sean incorrelados le añadimos que

$$E[\mathbf{e}] = \mathbf{0} \qquad E[\mathbf{f}] = \mathbf{0} \qquad E[\mathbf{x}] = \mathbf{0} \tag{1.3}$$

estaremos imponiendo, en primer lugar, una hipótesis estándar para los términos error en la mayoría de los modelos estadísticos. La segunda es conveniente para simplificar, sin suponer una pérdida de generalidad. La tercera puede no ser cierta, en cuyo caso, la expresión (1.2) puede cambiarse a

$$\mathbf{x} = \mu + \mathbf{A}\mathbf{f} + \mathbf{e} \tag{1.4}$$

donde $E[\mathbf{x}] = \mu$. La ecuación (1.4) introduce una ligera complicación algebraica comparada con la (1.2), pero como no hay pérdida de generalidad, suele utilizarse esta última.

Por su parte, la incorrelación entre los factores comunes y específicos puede expresarse como

$$\begin{aligned} E[\mathbf{ee}'] &= \mathbf{\Psi} \text{ (diagonal)} \\ E[\mathbf{fe}'] &= \mathbf{0} \text{ (una matriz de ceros)} \\ E[\mathbf{ff}'] &= \mathbf{I_m} \text{ (la matriz identidad).} \end{aligned} \tag{1.5}$$

Ésta última, puede relajarse y permitir que los factores sean correlados (oblicuos) en lugar de incorrelados (ortogonales). Como veremos más adelante, esto suele hacerse aplicando una rotación oblicua a alguna solución ortogonal.

Con todo esto, al modelo (1.2) descrito con las hipótesis (1.3) y (1.5) se le denominará *modelo factorial ortogonal*, mientras que si suponemos $E[\mathbf{ff}'] \neq \mathbf{I_m}$, se denominará *modelo factorial oblicuo.*

Para el modelo ortogonal, la matriz factorial $\mathbf{A}$ puede caracterizarse a través de la matriz $\mathbf{\Sigma}$ mediante el siguiente resultado debido a Thurstone [126]:

**Teorema** *El modelo (1.2) con las hipótesis (1.3) y (1.5) verifica que*

$$\mathbf{\Sigma} = \mathbf{AA}' + \mathbf{\Psi} \tag{1.6}$$

Este resultado es conocido como la identidad fundamental que debe verificar toda matriz factorial ortogonal.

Además de lo anterior, el modelo ortogonal verifica que

$$COV(\mathbf{x}, \mathbf{f}) = E[\mathbf{xf}'] = \mathbf{A}E[\mathbf{ff}'] + E[\mathbf{ef}'] = \mathbf{A}. \tag{1.7}$$

es decir, la estructura de covarianzas vendrá dada por

$$Var(x_j) = a_{j1}^2 + \cdots + a_{jm}^2 + \psi_j \qquad j = 1, \ldots, p \tag{1.8}$$

$$\begin{aligned} Cov(x_j, x_h) &= a_{j1}a_{h1} + \cdots + a_{jm}a_{hm} \qquad j, h = 1, \ldots, p \\ Cov(x_j, f_k) &= a_{jk} \qquad j = 1, \ldots, p; \quad k = 1, \ldots, m \end{aligned} \tag{1.9}$$

La expresión (1.8) indica que la varianza de la variable i-ésima puede dividirse en dos partes, como decíamos en la introducción. Por un lado tendremos la parte de varianza que son capaces de explicar los factores comunes

$$h_j^2 = a_{j1}^2 + \cdots + a_{jm}^2 \qquad j = 1, \ldots, p \tag{1.10}$$

denominada *comunalidad*. Por otro, se tendrá la varianza debida al factor específico, $\psi_j$, denominada *unicidad* o *varianza específica*. De este modo

$$Var(x_j) = \sigma_{jj} = h_j^2 + \psi_j \qquad j = 1, \ldots, p \tag{1.11}$$

### Modelo con la matriz de correlaciones

Todo el planteamiento anterior puede simplificarse aún más si se utiliza el hecho de que la correlación entre dos variables es invariante por transformaciones del tipo $(x_j - a)/b$ y suponemos que las variables observadas se han estandarizado mediante la transformación

$$z_j = \frac{x_j - \mu_j}{\sigma_j} \qquad j = 1, \ldots, p$$

con $\mu_j$ y $\sigma_j^2$ la media y varianza de cada variable $x_j$.

En este supuesto, si seguimos llamando a las variables $x_j$, se deduce de (1.6) y (1.7) que,

$$\mathbf{R} = \mathbf{A}\mathbf{A}' + \mathbf{\Psi} \tag{1.12}$$

$$COV(\mathbf{x}, \mathbf{f}) = COR(\mathbf{x}, \mathbf{f}) = \mathbf{A}. \tag{1.13}$$

y, por tanto

$$1 = h_j^2 + \psi_j \qquad j = 1, \ldots, p \tag{1.14}$$

### Interpretación geométrica

Si consideramos cada una de las $p$ variables como un vector de un espacio vectorial y asociamos su desviación típica a la norma del vector:

$$\| x_j \| = \sigma_j = \sqrt{Var(x_j)} \qquad j = 1, \ldots, p$$

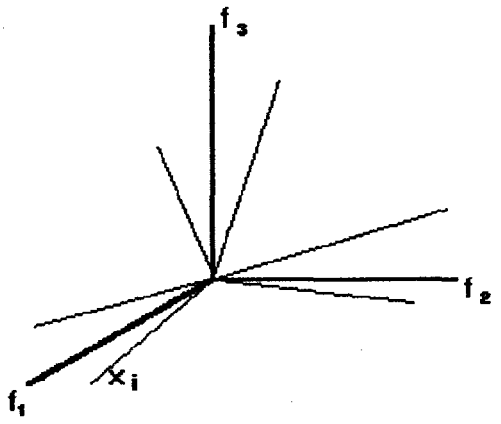


Figura 1.1: Expresión de las variables reducidas $x_j^*$ en combinación lineal de los factores.

y la covarianza entre dos variables al producto escalar asociado a la norma, se tendrá que es lícito asociar el coeficiente de correlación entre ellas al coseno del ángulo que forman: $r = \cos\theta$.

Si las variables son estandarizadas, entonces estarán representadas por vectores de norma unidad. En el caso de que fueran linealmente independientes (rango $p$), estos vectores estarán en un espacio $p$-dimensional, y sus extremos en una esfera de radio unidad. En este contexto, los factores comunes serán vectores unitarios ortogonales (o no), de forma que las $p$ variables puedan expresarse, una vez restadas sus unicidades, en función de ellos como

$$x_j^* = x_j - e_j = a_{j1}f_1 + \ldots + a_{jm}f_m \qquad j = 1, \ldots, p$$

En la Figura 1.1, puede verse cómo los vectores tendrán normas inferiores a uno en general (serán $h_j$), y que los pesos $a_{jk}$ son las proyecciones ortogonales (en este caso los factores se suponen incorrelados) de cada variable reducida en cada factor.

Esta interpretación vectorial, que puede encontrarse en Dempster [32], puede justificarse teóricamente por el hecho de que el conjunto de variables

aleatorias sobre una población, con las operaciones suma y producto por un escalar, tiene estructura de espacio vectorial. Además, las variables con varianza finita y esperanza nula forman un subespacio vectorial del anterior, en el cual la covarianza entre dos variables: $\text{cov}(x_j, x_k)$ es un producto escalar que define una norma, que es la varianza, y un ángulo entre los vectores $x_j$ y $x_k$ cuyo coseno es el coeficiente de correlación.

### 1.2.2 Obtención de soluciones primarias

El modelo factorial (1.1) supone que los $p+p(p-1)/2 = p(p+1)/2$ elementos de la matriz de covarianzas $\mathbf{\Sigma}$ o los $p(p-1)/2$ de la matriz de correlaciones $\mathbf{R}$ del vector aleatorio $\mathbf{x}$, deben ser reproducidos por los $p \cdot m$ pesos de $\mathbf{A}$ y los $p$ específicos $e_j$, según

$$\mathbf{\Sigma} = \mathbf{A}\mathbf{A}' + \mathbf{\Psi}$$

Cuando $m = p$, cualquier matriz de covarianzas puede reproducirse exactamente como un producto de $\mathbf{A}\mathbf{A}'$, siendo $\mathbf{\Psi} = \mathbf{0}$. Sin embargo, el problema surge cuando $m$ es mucho menor que $p$, con objeto de que el modelo proporcione una explicación más "simple" de la matriz $\mathbf{\Sigma}$ con menos parámetros.

No siempre es esto posible desde el punto de vista algebraico, e incluso puede que en ocasiones exista solución numérica pero no sea consistente con el problema. En general, la solución de (1.6) depende de $\mathbf{\Sigma}$, del número de variables $p$ y del número de factores $m$ que se pretendan determinar.

Históricamente, son muchos los procedimientos que se han desarrollado para extraer los factores, la mayoría de los cuales están recogidos en Harmann [61]. Unos parten de la estimación de las comunalidades, para trabajar después con la matriz reducida $\mathbf{\Sigma}^* = \mathbf{\Sigma} - \mathbf{\Psi}$. Otros parten de un número de factores y tratan de obtener los pesos minimizando los residuales $e_j$. Los hay también que utilizan métodos basados en la extracción de las componentes principales, otros que utilizan razonamientos de tipo geométrico o de agrupamiento de las variables y algunos, también, que tratan el Análisis Factorial como un método estadístico y obtienen las soluciones utilizando técnicas de estimación.

Por otra parte, una vez encontrada una solución $\mathbf{A}$, correspondiente a un patron factorial de un conjunto de factores ortogonales, ésta no es única. En efecto, si es $\mathbf{T}$ una matriz ortogonal $m \cdot m$ ($\mathbf{TT'} = \mathbf{T'T} = \mathbf{I}$), que represente una rotación rígida de los factores, entonces la expresión (1.2) puede escribirse

$$\mathbf{x} = \mathbf{Af} + \mathbf{e} = \mathbf{ATT'f} + \mathbf{e} = \mathbf{A^*f^*} + \mathbf{e}$$

con $\mathbf{A^*} = \mathbf{AT}, \mathbf{f^*} = \mathbf{T'f}$ y siendo

$$E[\mathbf{f^*}] = \mathbf{T'}E[\mathbf{f}] = \mathbf{0}$$

$$COV[\mathbf{f^*}] = \mathbf{T'}COV[\mathbf{f}]\mathbf{T} = \mathbf{T'T} = \mathbf{I}_m$$

es decir, $\mathbf{A^*} = \mathbf{AT}$ también será una solución del problema, siendo $\mathbf{T}$ cualquier matriz ortogonal. Además, es imposible distinguir - sólo con el conocimiento de las variables iniciales $\mathbf{x}$ - entre las matrices $\mathbf{A}$ y $\mathbf{A^*}$, ya que las dos verifican las condiciones del modelo.

Lo anterior da lugar a plantear el problema de buscar una solución para el modelo factorial, imponiendo condiciones que busquen primero una solución cualquiera, y después realizar rotaciones rígidas de ejes -mediante matrices ortogonales- que permitan obtener patrones factoriales más fáciles de interpretar, ó bien, que atiendan a determinados criterios que se establezcan.

Esta forma de extraer soluciones rotando los factores, puede generalizarse no imponiendo la condición de ortogonalidad entre ellos y buscando un conjunto de factores oblicuos, también bajo ciertos criterios.

### 1.2.3 Rotación de soluciones

Uno de los propósitos del Análisis Factorial es describir la configuración de las variables de la forma más simple posible.

El planteamiento realizado desde el punto de vista geométrico, supone que lo que se busca en el modelo, salvo los términos error, es un conjunto pequeño de vectores que permitan expresar a las variables como combinaciones lineales de ellos. Esto lleva inmediatamente a la conclusión de que una vez encontrado este conjunto, no puede ser único salvo que contenga un sólo vector. En

realidad, una solución al problema no es más que una base del subespacio vectorial de dimensión $m$ en la que, salvo errores, podrían 'incluirse' los vectores $x_j$. Por lo tanto, cualquier rotación, ortogonal o no, de estos vectores encontrados, proporcionaría una nueva base y, en consecuencia, un nuevo conjunto de factores.

Es por ello que, una vez encontrada una solución, el modelo factorial puede obtener soluciones derivadas de ésta realizando rotaciones que atiendan a algunos criterios, generalmente de simplicidad.

Thurstone [125], formuló el concepto de *estructura simple* atendiendo a la necesidad de unificar tales criterios de simplicidad. Este modelo de estructura simple, trata de dar unos criterios sobre las infinitas soluciones factoriales que pueden extraerse de un mismo conjunto de datos, con el fin de que distintos investigadores obtengan las mismas soluciones para el mismo problema. Los criterios pueden resumirse en los siguientes puntos (Cuadras, [30]):

- Cada fila de la matriz de pesos deberá tener un cero por lo menos.
- Si hay $m$ factores comunes, cada columna de la matriz de pesos deberá tener $m$ ceros por lo menos.
- Para todo par de columnas de la matriz de pesos deberá haber varias variables cuyas entradas se anulen en una columna pero no en la otra.
- Para todo par de columnas de la matriz de pesos, una gran proporción de las variables deberán tener entradas nulas en ambas columnas cuando hay cuatro o más factores.
- Para todo par de columnas de la matriz de pesos, deberá haber sólo un número pequeño de variables con entradas no nulas en ambas columnas.

La traslación de estos criterios tan generales en términos matemáticos precisos, fué una tarea que llevó varios años y el esfuerzo de muchos investigadores. Como consecuencia de ello, se han desarrollado numerosas soluciones, que pueden encontrarse desarrolladas con detalle en Harmann [61]. Éstas pueden dividirse básicamente en dos tipos de soluciones: las que mantienen la ortogonalidad de los factores utilizando rotaciones rígidas de soluciones

primarias, obedeciendo a criterios de optimización que consigan acercar la solución a la estructura simple de Thurstone; y las que relajan la hipótesis inicial de ortogonalidad, buscando rotaciones entre los factores que puedan transformar las soluciones primarias en soluciones oblicuas que obedezcan también a criterios de optimización.

Hay que resaltar también, que los procedimientos de rotación oblicuos introducen en el modelo (1.1) los conceptos de correlaciones entre los factores y de estructura factorial.

### Patrones y estructuras factoriales

Hasta ahora, en el patrón considerado, los factores comunes $f_k, k = 1, \ldots, m$ se suponían incorrelados, al igual que los factores únicos y éstos con aquéllos, pero, en general, en el modelo factorial los factores comunes pueden no imponerse incorrelados pues en ocasiones esta hipótesis no se sostiene desde el punto de vista de la interpretación del resultado. Así, los valores $r_{f_k f_l}$ pueden ser en general distintos de cero, o también,

$$E[\mathbf{f}\mathbf{f}'] \neq \mathbf{I}_m$$

En cualquiera de las dos hipótesis sobre si los factores son incorrelados o correlados, el A.F. proporciona no sólamente la matriz A del patrón factorial, sino también una matriz con las correlaciones entre las variables originales y los factores. Se denomina *estructura factorial* a la matriz **S** con tales correlaciones

$$\mathbf{S} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1m} \\ s_{21} & s_{22} & \cdots & s_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ s_{p1} & s_{p2} & \cdots & s_{pm} \end{pmatrix}$$

es decir, $s_{jk} = r_{x_j f_k}$, de forma que la solución factorial completa necesita del patrón y la estructura factoriales. Es importante tener presente la relación funcional entre ambas matrices, expresada mediante las ecuaciones (1.15), donde queda claro que en el caso de que supongamos incorrelados los factores, el patrón coincide con la estructura.

$$
\begin{aligned}
r_{x_j f_1} &= a_{j1} + a_{j2} r_{f_1 f_2} + \cdots + a_{jk} r_{f_1 f_k} + \cdots + a_{jm} r_{f_1 f_m} \qquad (1.15)\\
&\vdots \\
r_{x_j f_k} &= a_{j1} r_{f_k f_1} + a_{j2} r_{f_k f_2} + \cdots + a_{jk} + \cdots + a_{jm} r_{f_k f_m} \\
&\vdots \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (1.16)\\
r_{x_j f_m} &= a_{j1} r_{f_m f_1} + a_{j2} r_{f_m f_2} + \cdots + a_{jk} r_{f_m f_k} + \cdots + a_{jm}
\end{aligned}
$$

La relación (1.15) puede también expresarse en notación matricial como en (1.17)

$$\mathbf{S} = \mathbf{A}\mathbf{\Phi} \qquad (1.17)$$

con

$$\mathbf{\Phi} = \begin{pmatrix} 1 & r_{f_1 f_2} & \cdots & r_{f_1 f_m} \\ r_{f_2 f_1} & 1 & \cdots & r_{f_2 f_m} \\ \cdots & \cdots & \cdots & \cdots \\ r_{f_p f_1} & r_{f_p f_2} & \cdots & r_{f_p f_m} \end{pmatrix}$$

**Procedimientos de rotación**

En el Análisis Factorial, la rotación ortogonal se basa en la búsqueda de una matriz $\mathbf{T_{m,m}}$, donde $m$ es el número de factores, que transforme la matriz $\mathbf{A_{n,m}}$ en otra $\mathbf{B_{n,m}} = \mathbf{A_{n,m}} \cdot \mathbf{T_{m,m}}$, de forma que $\mathbf{B}$ optimize un criterio determinado.

Entre la variedad de criterios que se han propuesto, algunos de los más utilizados son los que se basan en el hecho de que la comunalidad de cualquier elemento permanece invariable ante una transformación ortogonal, es decir,

$$\sum_{j=1}^{m} b_{ij}^2 = \sum_{j=1}^{m} a_{ij}^2 = h_i^2 \qquad i = 1, \ldots, n,$$

de donde se deduce que la suma de los cuadrados de las comunalidades (cua-

drados por filas de la matriz de pesos) es constante:

$$\sum_{i=1}^{n}\sum_{j=1}^{m} b_{ij}^4 + 2 \cdot \sum_{i=1}^{n}( \sum_{j<k=1}^{m} b_{ij}^2 b_{ik}^2) = \text{cte.}$$

De esta expresión surge entre otros, el criterio *cuartimax* (Carroll [18]), obteniendo la rotación que maximiza

$$Q = \sum_{i=1}^{n}\sum_{j=1}^{m} b_{ij}^4$$

o que minimiza

$$N = \sum_{i=1}^{n}( \sum_{j<k=1}^{m} b_{ij}^2 b_{ik}^2)$$

Por otro lado, otro criterio muy utilizado es el de Kaiser [84], denominado *varimax*, cuyo objetivo es obtener una matriz **B** más simplificada en el sentido de que sus columnas contengan pesos lo más cercanos posible a 0, 1 ó -1. Para ello el criterio que adopta es maximizar la varianza de los cuadrados de los pesos normalizados por las comunalidades $h_i^2$. Tal varianza es para una columna $j$:

$$s_j^2 = \frac{1}{n}\sum_{i=1}^{n}(\frac{b_{ij}}{h_i})^4 - \frac{1}{n^2}(\sum_{i=1}^{n}\frac{b_{ij}^2}{h_i^2})^2,$$

y para todas las columnas es

$$s^2 = \sum_{j=1}^{m} s_j^2 = \frac{1}{n}\sum_{j=1}^{m}\sum_{i=1}^{n}(\frac{b_{ij}}{h_i})^4 - \frac{1}{n^2}\sum_{j=1}^{m}(\sum_{i=1}^{n}\frac{b_{ij}^2}{h_i^2})^2.$$

En realidad la función que propone maximizar es $n^2 \cdot s^2$:

$$V = n \cdot \sum_{j=1}^{m}\sum_{i=1}^{n}(\frac{b_{ij}}{h_i})^4 - \sum_{j=1}^{m}(\sum_{i=1}^{n}\frac{b_{ij}^2}{h_i^2})^2$$

y el procedimiento que utiliza es iterado, formando ciclos de transformaciones ortogonales entre todas las parejas posibles de factores, hasta que los cambios en el valor de V no sean significativos.

En el caso de realizar una rotación oblicua, hay que tener en cuenta que además de la matriz de pesos (o patrón factorial) como solución final, surge también la matriz de estructura factorial, la cual contiene las relaciones entre los factores y los individuos. Ambas matrices coinciden cuando los factores son ortogonales.

Así, si llamamos $\mathbf{B_{n,m}}$ a la matriz de pesos con los factores oblicuos y $\mathbf{\Theta_{m,m}}$ a la matriz de correlaciones entre cada pareja de factores, la estructura factorial oblicua viene dada por

$$\mathbf{S_{n,m}} = \mathbf{B_{n,m}} \cdot \mathbf{\Theta_{m,m}}$$

donde, claramente, si $\mathbf{\Theta_{m,m}} = \mathbf{I}$ (en el caso ortogonal), $\mathbf{S_{n,m}} = \mathbf{B_{n,m}}$.

Una forma de obtener una solución oblicua parte de una solución ortogonal $\mathbf{A_{n,m}}$, y mediante un criterio apropiado busca una transformación $\mathbf{T_{m,m}}$ que pase de la matriz $\mathbf{A_{n,m}}$ a la estructura factorial oblicua $\mathbf{S_{n,m}}$:

$$\mathbf{S_{n,m}} = \mathbf{A_{n,m}} \cdot \mathbf{T_{m,m}}$$

Posteriormente, se calcularía

$$\mathbf{\Theta_{m,m}} = \mathbf{T'_{m,m}} \cdot \mathbf{T_{m,m}}$$

y la matriz de pesos

$$\mathbf{B_{n,m}} = \mathbf{S_{n,m}} \cdot \mathbf{\Theta^{-1}_{m,m}}$$

Nótese también que podría ponerse

$$\mathbf{B} = \mathbf{S\Theta^{-1}} = \mathbf{S(T'T)^{-1}} = \mathbf{ST^{-1}(T')^{-1}} = \mathbf{A(T')^{-1}}$$

Entre los criterios que más se utilizan para obtener la matriz $\mathbf{T}$ anterior está el debido a Kaiser [84] y Carroll [19], denominados métodos *oblimin*, en los cuales se busca minimizar la función

$$C = \sum_{j \leq k=1}^{m} (n \sum_{i=1}^{n} \frac{s_{ij}^2}{h_i^2} \frac{s_{ik}^2}{h_i^2} - \sum_{i=1}^{n} \frac{s_{ij}^2}{h_i^2} \sum_{i=1}^{n} \frac{s_{ik}^2}{h_i^2})$$

es decir, minimizar las covarianzas de los cuadrados de los elementos de la estructura final **S**. Puede comprobarse fácilmente que este criterio es equivalente al propuesto por Kaiser en el caso ortogonal, y es considerado por este autor como el método más obvio de relajar la restricción de ortogonalidad. De todas formas, atendiendo a cuestiones de oblicuidad de los factores, la solución definitiva al criterio la propuso Carroll [20] añadiendo un parámetro $\gamma$ con lo que quedaría

$$C^* = \sum_{j \leq k=1}^{m} (n \sum_{i=1}^{n} \frac{s_{ij}^2}{h_i^2} \frac{s_{ik}^2}{h_i^2} - \gamma \sum_{i=1}^{n} \frac{s_{ij}^2}{h_i^2} \sum_{i=1}^{n} \frac{s_{ik}^2}{h_i^2})$$

Nótese que para $\gamma = 0$ este criterio minimizaría la expresión N dada anteriormente para rotaciones ortogonales, sólo que aquí sería para obtener factores oblicuos. A este caso particular de los métodos oblimin se le denomina criterio *cuartimin*.

Una segunda forma de obtener soluciones oblicuas parte también de la solución ortogonal **A** y, mediante un criterio apropiado, encuentra directamente el patrón factorial oblicuo **B**. Jennrich y Sampson [77] proponen que el criterio a optimizar sea el mismo que el de Carroll, pero con la matriz de pesos **B**, es decir, minimizar:

$$F(B) = \sum_{j \leq k=1}^{m} (\sum_{i=1}^{n} \frac{b_{ij}^2}{h_i^2} \frac{b_{ik}^2}{h_i^2} - \frac{\delta}{n} \sum_{i=1}^{n} \frac{b_{ij}^2}{h_i^2} \sum_{i=1}^{n} \frac{b_{ik}^2}{h_i^2})$$

Como $\mathbf{B} = \mathbf{A}(\mathbf{T}')^{-1}$, el problema equivale a encontrar una matriz de transformación **T** que minimice $F(\mathbf{A}(\mathbf{T}')^{-1})$, sujeta a la condición $\text{diag}(\mathbf{T}'\mathbf{T}) = \mathbf{I}$, y se realiza de forma iterada mediante ciclos completos de transformaciones oblicuas entre todas las posibles parejas de factores, hasta que los cambios de $F(\mathbf{B})$ no sean significativos.

El parámetro $\delta$ tiene relación con la mayor o menor oblicuidad que se le quiera dar a los factores. Cuando varía entre 1 y valores negativos, pasando por 0, la oblicuidad de los factores varía de menos a más, siendo la solución más adoptada la de $\gamma = 0$.

### 1.2.4 Medición de los factores

En el Análisis Factorial, los factores pueden ser considerados como funciones de las variables originales. Por ello, puede tener interés en muchas ocasiones, que determinemos el valor de esta nueva variable en cada objeto. A estos valores se les denomina *puntuaciones factoriales.*

La utilidad de las puntuaciones factoriales puede ser doble. Por un lado, al haber menos factores que variables originales, la representación que puede hacerse de los objetos con tales puntuaciones será sensiblemente más sencilla, manteniendo casi la misma cantidad de información. Por otro, en el caso de que los factores sean ortogonales, las nuevas variables serán incorreladas, y esto supone una ventaja importante al utilizar otras técnicas estadísticas.

Se han propuesto, también, varios métodos para encontrar estas puntuaciones factoriales [61], pero quizás el que más se utiliza es el que se basa en la estimación por mínimos cuadrados.

Básicamente, esta técnica utiliza el hecho de que el modelo (1.1) es lineal, donde las filas $x_j$ de la matriz $\mathbf{X}$ y la matriz $\mathbf{A}$ son conocidas a estas alturas, y los $e_j$ representan los errores del modelo. Con esto, pueden obtenerse estimaciones de los $f_k$ a partir de minimizar

$$\sum_{j=1}^{p}(x_j - a_{j1}f_1 - \cdots - a_{im}f_m)^2$$

cuya solución [30], es:

$$\mathbf{f} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{X}$$

## 1.3 El modelo factorial para los elementos de un conjunto de datos

Como se ha comentado en la introducción, el modelo de A.F. para individuos, o en modo Q, es una técnica multivariante iniciada por Burt [15],[16] y Stephenson [120] como una alternativa al modelo para variables, o en modo R, introducido pocos años antes. Su objetivo era el análisis de conductas y comportamientos psicológicos mediante la realización de una batería de tests en un conjunto de individuos. Para ello, en lugar de tratar de buscar unos pocos factores que permitieran explicar el comportamiento desde el punto de vista de la información que proporcionan los tests, se pensó que ante un grupo de tests, las puntuaciones obtenidas podrían utilizarse para buscar algunos factores que representaran diferentes tipos de individuos subyacentes en la población.

El modelo factorial en modo Q se basa en los mismos postulados que los del modo R, trasladándolos a los individuos. No obstante, tiene interés que comentemos las diferencias y similitudes que hay entre ambos, así como algunos problemas que se han apuntado de él por el hecho de trasladar conceptos definidos para variables a individuos.

### 1.3.1 Planteamiento del modelo

Como en la sección anterior, partiremos también de un conjunto de $p$ variables $x_1, \ldots, x_p$ que se han medido en un conjunto de $n$ individuos $I_1, \ldots, I_n$, encontrándonos con una matriz de datos como

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

donde $x_{ij}$ representa el valor que toma la variable $x_j$ en el individuo $I_i$. Una fila representa las puntuaciones de las $p$ variables en un individuo, mientras que una columna contiene todos los valores alcanzados por una variable en los individuos.

En ocasiones, se considerarán las variables $x_j$ con media nula - variables centradas -, e incluso con varianza unidad - variables estandarizadas -. Tanto una como otra situación no hacen perder generalidad al desarrollo, aunque sí pueden originar resultados diferentes, como se comentará más adelante.

El modelo matemático trata de encontrar un conjunto de $m$ individuos, denominados factores comunes que llamaremos $g_1, ..., g_m$, donde necesariamente debe ser $m \leq n$, y $n$ factores únicos ó de error $e_1, ..., e_n$, tales que cada individuo original $I_i$ pueda ser expresado de forma lineal como en (1.19).

$$\begin{aligned} I_1 &= a_{11}g_1 + \cdots + a_{1m}g_m + e_1 \\ I_2 &= a_{21}g_1 + \cdots + a_{2m}g_m + e_2 \\ &\vdots \\ I_n &= a_{n1}g_1 + \cdots + a_{nm}g_m + e_n \end{aligned} \tag{1.19}$$

Los coeficientes $a_{ik}$, $i = 1, 2, \ldots, n$, $k = 1, 2, \ldots, m$ son constantes a determinar denominadas *pesos factoriales*, y $e_i$, $i = 1, 2, \ldots, p$ son los términos error, llamados también *factores únicos* por ser para cada $e_i$ 'específico' para cada individuo $I_i$, mientras que los factores $g_k$ son 'comunes' a todos los individuos.

Desde el punto de vista geométrico (Figura 1.2), al igual que en el modo R, los individuos representados por las filas, menos los factores únicos, pueden expresarse linealmente respecto de la base formada por los factores comunes.

El modelo (1.19), cuando se formula para las variables, utiliza conceptos estadísticos como la media o la varianza para una variable, o la covarianza o correlación para cada pareja de variables. Como ya se dijo, estos conceptos gozan de una clara interpretación geométrica en $I\!R^n$ ya que cada una de las $p$ columnas de la matriz X pueden considerarse como las coordenadas de un vector desde el origen en dicho espacio. Estos vectores - a los que por comodidad les suprimiremos la flecha - $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_p}$, originan a su vez otros $p$ vectores desviación $\mathbf{d_1}, \mathbf{d_2}, \ldots, \mathbf{d_p}$ cuando se resta a cada $\mathbf{x_j}$ el vector con módulo el valor medio de $\mathbf{x_j}$ y cuya dirección forma ángulos iguales con cada eje de $I\!R^n$

$$\mathbf{d_j} = \mathbf{x_j} - \overline{x}_j \mathbf{1} \qquad j = 1, \ldots, p \tag{1.20}$$

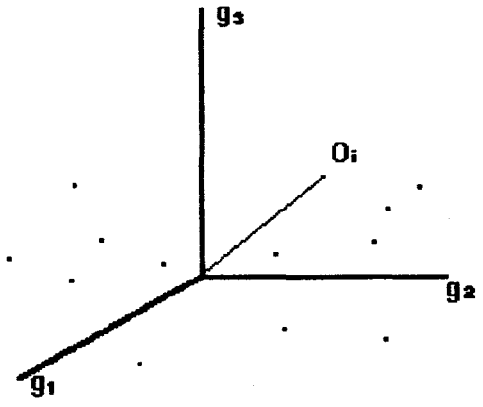


Figura 1.2: Representación de los individuos reducidos respecto de los factores

siendo $\mathbf{1}$ el vector de $I\!R^n$ cuyas componentes son todas iguales a 1.

Si utilizamos la métrica euclídea, claramente la norma cuadrada de los vectores desviación $\mathbf{d_j}$ será

$$\|\mathbf{d_j}\|^2 = \mathbf{d_j}\mathbf{d_j} = \sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2 \qquad j = 1, \ldots, p$$

lo que significa que la longitud de cada vector desviación es proporcional a la desviación típica de la variable $\mathbf{x_j}$.

Además, para cada dos vectores desviación $\mathbf{d_j}, \mathbf{d_h}$:

$$\mathbf{d_j}\mathbf{d_h} = \sum_{i=1}^{n}(x_{ij} - \overline{x}_j)(x_{ih} - \overline{x}_h) = (n-1)s_{jh} \qquad j, h = 1, \ldots, p$$

de forma que, si $\theta_{jh}$ es el ángulo formado por los vectores $\mathbf{d_j}$ y $\mathbf{d_h}$, puede verse fácilmente que el coeficiente de correlación entre las variables $\mathbf{x_j}$ y $\mathbf{x_h}$ es

$$r_{jh} = \frac{\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)(x_{ih} - \overline{x}_h)}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2}\sqrt{\sum_{i=1}^{n}(x_{ih} - \overline{x}_h)^2}}$$

$$= \frac{s_{jh}}{\sqrt{s_{jj}}\sqrt{s_{hh}}}$$
$$= cos(\theta_{jh}) \qquad j,h = 1,\ldots,p$$

Volviendo al modelo para individuos (1.19), podemos trasladar estos conceptos geométricos que nos ayuden a su interpretación.

Si consideramos las filas de $\mathbf{X}$, $\mathbf{I_1},...,\mathbf{I_n}$, como puntos del espacio $I\!R^p$, el centro de gravedad estará en el punto $\mathbf{C}(\overline{x}_1,\ldots,\overline{x}_p)$, ó en el $\mathbf{C}(0,\ldots,0)$ según sean los datos originales o centrados, y la dispersión total que presenta esta nube de puntos puede medirse relacionándola con el volúmen de un hiperelipsoide de la forma

$$(\mathbf{x}-\overline{x})S^{-1}(\mathbf{x}-\overline{x}) = c^2$$

En este contexto, si suponemos - sin pérdida de generalidad -, que las variables están centradas, los vectores desde el origen asociados a cada punto $\mathbf{I_i}$ pueden ser tratados igual que en el espacio de las variables, definiéndose los vectores desviación como

$$\mathbf{d}_i^* = \mathbf{I}_i - \overline{I}_i\mathbf{1} \qquad i = 1,\ldots,n \tag{1.21}$$

con

$$\overline{I}_i = \frac{1}{p}\sum_{j=1}^{p} x_{ij} \qquad i = 1,\ldots,n$$

y todos los demás conceptos de desviación típica de un individuo, y coeficiente de correlación entre dos individuos - todas ellas para el conjunto de variables utilizadas -, pueden introducirse utilizando la métrica euclídea definida en $I\!R^p$.

No se les puede objetar nada, desde el punto de vista geométrico, a tales conceptos, aunque sí desde el punto de vista de la interpretación estadística. Es evidente que si bien en determinados casos en los que la naturaleza de las variables es muy homogéneo, puede tener sentido los conceptos de media, dispersión ó correlación entre individuos, cuando esto no ocurra serán de difícil justificación.

Una alternativa que plantea menos problemas de interpretación es el hecho de considerar los vectores $\mathbf{I}_i$ sin restar los valores medios $\overline{I}_i$. En ese caso, la norma euclídea de cada $\mathbf{I}_i$,

$$\|\mathbf{I}_i\| = \sqrt{\sum_{j=1}^{p} x_{ij}^2} \qquad i = 1, \ldots, n$$

representará la norma del vector $\vec{\mathbf{OI}}_i$, y el producto escalar será el habitual

$$< \mathbf{I}_i, \mathbf{I}_h >= \sum_{j=1}^{p} x_{ij} x_{hj} = \|\mathbf{I}_i\| \|\mathbf{I}_h\| \cos(\theta_{ih}) \qquad i, h = 1, \ldots, n$$

con lo que el coeficiente de correlación o de similaridad entre dos individuos sería

$$q_{ih} = \frac{\sum_{j=1}^{p} x_{ij} x_{hj}}{\sqrt{\sum_{j=1}^{p} x_{ij}^2} \sqrt{\sum_{j=1}^{p} x_{hj}^2}} =< \frac{\mathbf{I}_i}{\|\mathbf{I}_i\|}, \frac{\mathbf{I}_h}{\|\mathbf{I}_h\|} >= \cos(\theta_{ih}) \tag{1.22}$$

con $i, h = 1, \ldots, n$, representando el coseno del ángulo que forman los vectores $\vec{\mathbf{OI}}_i$ y $\vec{\mathbf{OI}}_h$

### 1.3.2 Obtención de soluciones factoriales

Independientemente del tipo de coeficiente de correlación o de asociación entre los individuos que se adopte, el Análisis Factorial en modo Q utiliza la matriz de coeficientes

$$\mathbf{Q} = \begin{pmatrix} 1 & q_{12} & \cdots & q_{1n} \\ q_{21} & 1 & \cdots & q_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ q_{n1} & q_{n2} & \cdots & 1 \end{pmatrix}$$

para extraer los factores formulados en el modelo (1.19)

El procedimiento comienza por suponer - sin pérdida de generalidad - que los vectores $\mathbf{I}_i$ se han normalizado, e imponer, inicialmente, hipótesis de

ortogonalidad y normalidad entre los factores comunes, que pueden resumirse en

$$< g_k, g_l >= \begin{cases} 1 & \text{si} \quad k = l \\ 0 & \text{si} \quad k \neq l \end{cases} \qquad k, l = 1, \ldots, m$$

y también entre los factores comunes con los específicos

$$< g_k, e_i >= 0, \qquad k = 1, \ldots, m \quad i = 1, \ldots, n$$

mientras que entre los específicos se supone sólo la ortogonalidad

$$< e_i, e_h >= \begin{cases} \psi_i & \text{si} \quad i = h \\ 0 & \text{si} \quad i \neq h \end{cases} \qquad i, h = 1, \ldots, n$$

Con estas hipótesis es fácil deducir, al igual que se hizo para las variables, la identidad fundamental del modelo factorial ([126]), que permite caracterizar los coeficientes $a_{ik}$: **Teorema** *Se verifica que*

$$q_{ih} = \sum_{k=1}^{m} a_{ik} a_{hk}, \quad i \neq h, \quad i,h = 1, \ldots, n \tag{1.23}$$

$$q_{ii} = \sum_{k=1}^{m} a_{ik}^2 + \psi_i \tag{1.24}$$

*o en notación matricial, siendo* $\mathbf{\Psi}$ *la matriz diagonal de las* $\psi_i$

$$\mathbf{Q} = \mathbf{A}\mathbf{A}' + \mathbf{\Psi} \tag{1.25}$$

Conviene destacar, al igual que en el modelo para variables, cómo las similaridades entre cada pareja de individuos $\mathbf{I}_i$ e $\mathbf{I}_h$ pueden reproducirse (1.23) a través de las filas $i$ y $h$ de la matriz de pesos $\mathbf{A}$, mientras que las similaridades de un individuo consigo mismo (1.24) vienen dadas por dos términos

$$q_{ii} = \sum_{k=1}^{m} a_{ik}^2 + \psi_i^2 = h_i^2 + \psi_i$$

el primero llamado *comunalidad*, representando la parte de longitud del vector $\mathbf{I}_i$ que son capaces de explicar los factores comunes, y el segundo $\psi_i$, que será la que explique el factor específico.

Además, utilizando las hipótesis de partida del modelo, cada peso factorial $a_{ik}$ tiene un significado muy claro: es el coseno del ángulo que forman el vector $\mathbf{I}_i$ con el factor $g_k$

$$< \frac{\mathbf{I}_i}{\|\mathbf{I}_i\|}, g_k >= a_{ik} \tag{1.26}$$

Históricamente, los procedimientos que se han desarrollado para extraer los factores en el modo R, se han utilizado también en su mayoría para el caso de analizar individuos. Exceptuando los procedimientos que tratan al A.F. como un método estadístico y obtienen las soluciones utilizando técnicas de estimación partiendo de hipótesis sobre las distribuciones de las variables, el modo Q puede utilizar cualquiera de los apuntados en la sección anterior, sobre todo aquellos que tienen en cuenta procedimientos algebraicos y geométricos, en los cuales, la mayoría de las veces lo que se pretende es encontrar agrupaciones entre los individuos analizados.

Respecto a la rotación de los factores, el modelo factorial en modo Q, en analogía con el modo R, una vez obtenida una solución puede obtener soluciones derivadas de ésta, realizando rotaciones que atiendan a algunos criterios, generalmente de simplificidad. Los criterios de estructura simple de Thurstone [125], y su traslación a procedimientos matemáticos precisos, permiten realizar rotaciones ortogonales u oblicuas que obedezcan a criterios de optimización.

Hay que resaltar también aquí, que los procedimientos de rotación oblicuos introducirán en el modelo (1.19) un elemento nuevo: las asociaciones existentes entre los factores.

En efecto, cuando se suponen ortogonales los factores, el modelo queda completamente determinado por la matriz A de pesos, también llamada *patrón factorial*, que expresa linealmente a los individuos originales respecto de los factores. Pero cuando el objetivo del análisis, como será nuestro caso, no aconseja imponer esta hipótesis por ser muy restrictiva, los ángulos entre los factores serán, en general, oblícuos y por tanto

$$< g_k, g_l > \neq \begin{cases} 1 & \text{si} \quad k = l \\ 0 & \text{si} \quad k \neq l \end{cases} \qquad k, l = 1, \ldots, m$$

A la matriz con los productos escalares anteriores se le suele denominar $\mathbf{\Phi}$ y será del tipo

$$\mathbf{\Phi} = \begin{pmatrix} 1 & < g_1, g_2 > & \cdots & < g_1, g_m > \\ < g_2, g_1 > & 1 & \cdots & < g_2, g_m > \\ \cdots & \cdots & \cdots & \cdots \\ < g_m, g_1 > & < g_m, g_2 > & \cdots & 1 \end{pmatrix}$$

Esta nueva situación hace que la relación (1.26) no sea válida, sino que se verifiquen en este caso, para $i = 1, \ldots, n$, las ecuaciones

$$\begin{aligned} < \frac{\mathbf{I}_i}{\|\mathbf{I}_i\|}, g_1 > &= a_{i1} + a_{i2} < g_1, g_2 > + \cdots + a_{ik} < g_1, g_k > + \cdots + a_{im} < g_1, g_m > \\ &\vdots \\ < \frac{\mathbf{I}_i}{\|\mathbf{I}_i\|}, g_k > &= a_{i1} < g_k, g_1 > + a_{i2} < g_k, g_2 > + \cdots + a_{ik} + \cdots + a_{im} < g_k, g_m > \\ &\vdots \\ < \frac{\mathbf{I}_i}{\|\mathbf{I}_i\|}, g_m > &= a_{i1} < g_m, g_1 > + a_{i2} < g_m, g_2 > + \cdots + a_{ik} < g_m, g_k > + \cdots + a_{im} \end{aligned}$$

en las que las asociaciones entre los individuos originales y los factores obtenidos vienen dados por los pesos y también por las asociaciones que existan entre los factores. Es evidente de las ecuaciones anteriores que

$$\mathbf{S} = \mathbf{A}\mathbf{\Phi} \tag{1.27}$$

con $s_{ik} =< \frac{\mathbf{I}_i}{\|\mathbf{I}_i\|}, g_k >$, denominándose a $\mathbf{S}$ la matriz de *estructura factorial.* Nótese además, que en el caso de que los factores sean ortogonales, el patrón factorial $\mathbf{A}$ coincide con la estructura factorial $\mathbf{S}$.

### 1.3.3 Medición de los factores

Por último, al igual que el A.F. en modo R, la técnica para individuos permite un paso más. Una vez obtenidos el patrón y la estructura, los cuales determinan totalmente el espacio de los factores y la relación lineal de los individuos respecto de ellos, es interesante si queremos utilizar los factores en

otros análisis o buscar asociaciones dominantes entre las variables, conocer cuál sería la puntuación teórica que, en caso de que existieran tales individuos patrones, alcanzarían en cada una de las variables utilizadas en el análisis.

Los procedimientos que se han utilizado para este fín han sido también varios, aunque no han proliferado con tanta diversidad como los que se utilizan para obtener los patrones factoriales. Como en el modo R, el más utilizado cuando se trata del modo Q es el que está basado nuevamente en el procedimiento de mínimos cuadrados, expuesto en la sección **1.2**

### 1.3.4 Algunos problemas que plantea el Análisis Factorial para los elementos de un conjunto

Como ya se ha comentado, desde sus inicios los investigadores que han empleado más frecuentemente este tipo de análisis son los psicólogos y psiquiatras que se dedican a estudiar tipologías de desórdenes mentales. Más recientemente, se ha utilizado también en investigaciones de mercado en las que no sólo interesa estudiar las relaciones entre items o variables, sino también las existentes entre personas u objetos, y en el análisis de variables y objetos de tipo geológico. Incluso existen algunas referencias que lo proponen como técnica que puede emplearse en el análisis cluster como en Overall y Klett [106].

Sin embargo, su utilización ha sido muy criticada en la mayoría de los casos, sobre todo cuando se ha intentado utilizar como un procedimiento de clasificación sin tomar algunas precauciones.

En efecto, independientemente de la forma de extraer o rotar los factores, Fleiss et al. [49] comentan que a pesar de ser la técnica en modo Q correcta desde el punto de vista teórico, como escribe Cattell [23], su utilización para identificar 'tipos' puede tener problemas.

Entre otros, Baggaley [5] cita el problema de generalizar los resultados obtenidos con el Q-análisis, ya que los diferentes factores encontrados entre los individuos, sólo pueden ser válidos para el conjunto de tests o variables utilizadas. Este conjunto no puede considerarse como una muestra de variables

extraida de una población compuesta por las infinitas variables que pueden utilizarse en un estudio. Esta objección nos parece obvia si el objetivo final de la obtención de estos 'tipos' es buscar algún tipo de clasificación entre los diferentes comportamientos psicológicos. Es bien conocido que cualquier técnica de Análisis clasificatorio - o Análisis Cluster -, busca grupos entre un conjunto de individuos atendiendo a las propiedades que estos presentan. Es tarea del investigador seleccionar las variables que midan estas propiedades, pero en ocasiones puede haber aspectos que escapen al control y que pueden cambiar el sentido de la clasificación y, por tanto, el resultado final.

Por otra parte, respecto al modelo, Zubin Y Fleiss [133] cuestionan el significado del modelo lineal utilizado en el Análisis Factorial cuando se aplica a personas en lugar de variables y Lorr [93] cuestiona los criterios de rotación de los factores tipo. Tanto uno como otro son problemas si se pretende utilizar el modelo en modo Q como si los individuos fueran variables desde el punto de vista estadístico, pero no desde el punto de vista geométrico, que será para nosotros el que nos interese de la técnica.

Jones [82] apunta que cuando los individuos tienen pesos apreciables en más de un factor, no pueden ser asignados a un tipo único. En este caso, pensamos que estos problemas pueden resolverse mediante un adecuado tratamiento del modelo como expondremos en la próximas secciones, de forma que los factores obtenidos sean congruentes en todo momento con lo que se pretende de ellos.

Por último, Fleiss y Zubin [48] comentan, entre otras cuestiones, que el coeficiente de correlación entre los individuos no es una buena medida de similaridad. Aunque ya hemos comentado este problema en la sección primera al introducir el coeficiente (1.22), nos detendremos a analizar este último problema por la importancia que tendrá en el desarrollo posterior de esta memoria.

### 1.3.5 El coeficiente de correlación como medida de similaridad

Fleiss y Zubin [48], en un trabajo en el que tratan algunos aspectos sobre las técnicas de clasificación, comentan que algunos de los métodos que se utilizan para este fin no son satisfactorios. En concreto, analizan aquellos procedimientos que emplean el coeficiente de correlación entre los individuos como medida de similaridad, entre los que se encuentran el Análisis Factorial en modo Q.

Por un lado, se refieren al problema de la estandarización previa de los datos. El hecho de que en muchas aplicaciones, las variables que describen a los individuos que se pretende clasificar, sean de diferentes tipos - categóricas, ordinales, continuas - o, incluso, siendo de tipo continuo, no estén medidas en las mismas unidades, aconseja a estandarizar los datos antes de calcular un coeficiente de similaridad entre los individuos o una distancia que provengan de un producto escalar. De esta forma, las aportaciones que proporcionan cada variable a tales coeficientes son comparables al tener todas varianzas unitarias. Sin embargo, estos autores, y posteriormente también Duda Y Hart [35] muestran ejemplos en los que la estandarización no ayuda a la distinción entre los diferentes grupos que puedan existir entre los individuos, sino que lo dificulta.

Desde el punto de vista matemático, la razón está en que se utiliza para estandarizar la desviación típica de todos los datos, donde se supone que existen grupos diferentes, cada uno con una dispersión y un tamaño propios. Así, si suponemos la existencia de dos grupos, cada uno con medias $\mu_1$ y $\mu_2$ y varianzas iguales $\sigma$, y cada grupo representa proporciones p y q, con $p+q=1$, respecto del total, la desviación típica total no tenderá a ser $\sigma$ sino

$$\sigma' = \sqrt{\sigma^2 + pq(\mu_1 - \mu_2)^2}$$

con lo cual, mientras mayor sea la diferencia entre los valores medios $\mu_1$ y $\mu_2$, mayor será el valor por el que se dividen las diferencias entre los individuos y se reducirán las contribuciones de cada variable en la distancia entre ellos. En estos casos, existe la posibilidad de que la estandarización debilite el poder de discriminación que tiene cada variable. Es evidente que si se pudieran estimar

las varianzas de cada grupo y utilizarlas en la estandarización, se evitaría el problema anterior. En Saunders y Schucman [113] se analizaron datos en los que se conocían estas varianzas, y se obtuvieron buenos resultados.

Un segundo problema que afecta a la estandarización, antes de calcular el coeficiente de correlación o la distancia entre individuos, es ignorar las posibles correlaciones entre las variables. Al igual que con las varianzas, las correlaciones existentes dentro de cada grupo pueden diferir mucho de la existente en la muestra entera. De esta forma, al ignorarlas en la estandarización podemos estar realizando transformaciones de los datos en cada grupo que no tendrán las mismas consecuencias para todos ellos. Esto puede hacer que las diferencias entre dos de ellos se enmascaren más. Una solución poco utilizada es la que parte de las primeras coordenadas principales para realizar el análisis.

Otros problemas que plantea el coeficiente de correlación como medida de similaridad pueden resumirse en las siguientes cuestiones. Para empezar, ¿qué significa?. No parece que pueda utilizarse, como ocurre con las variables, para realizar predicciones en un individuo de la puntuacion que alcanzará un nuevo test, en base a la puntuación que alcanza en otro individuo, porque en realidad no estamos trabajando con una muestra aleatoria de tests escogida de un universo de tests. ¿Cómo interpretar una correlación nula?, ¿y una correlación de $-1$?.

Fleiss y Zubin [48] proponen un ejemplo hipotético, en el que consideran tres individuos (Figura 1.3) con puntuaciones

$$I(-1, -\frac{1}{2}, 0, \frac{1}{2}, 1), \quad II(-1, -\frac{1}{2}, 0, \frac{1}{2}, \frac{3}{2}) \quad \text{y} \quad III(-1, 0, 1, 2, 3)$$

Como se aprecia, las puntuaciones del individuo III son el doble que las del I más 1, mientras que las del individuo II son las mismas que las del I en todas las variables excepto en la última, que pasa de ser 1 a $\frac{3}{2}$. Es difícil aceptar que los individuos I y III sean más similares - el coeficiente de correlación es 1 - que los individuos I y II - con coeficiente 0.986 -, que sólo se diferencian en una puntuación.

Si, como en la sección anterior, seguimos llamando $\mathbf{I}_1, \ldots, \mathbf{I}_n$ a las filas

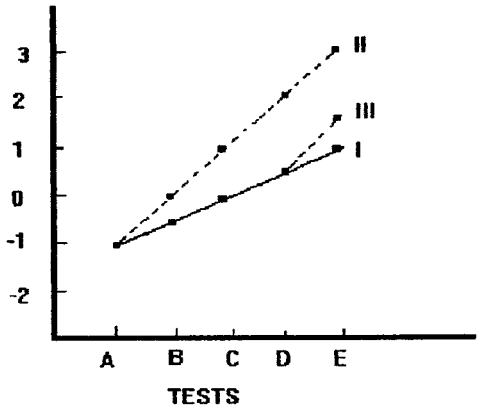


Figura 1.3: Representación de tres individuos con cinco tests

de la matriz $\mathbf{X}$, y llamamos $\overline{I}_i$ a la media de la fila $i$, $\mathbf{d}_i^* = \mathbf{I}_i - \overline{I}_i\mathbf{1}$ como se definió en (1.21), es inmediato comprobar que

$$d^2(\mathbf{I}_i, \mathbf{I}_h) = p(\overline{I}_i - \overline{I}_h)^2 + \|\mathbf{d}_i^*\|^2 + \|\mathbf{d}_h^*\|^2 - 2\|\mathbf{d}_i^*\|\|\mathbf{d}_h^*\|q_{ih}^* \qquad (1.28)$$

$i, h = 1, \ldots, n$, con $q_{ih}^*$ el coeficiente de correlación entre los individuos $\mathbf{I}_i$ e $\mathbf{I}_h$

$$q_{ih}^* = \frac{\sum_{j=1}^p (x_{ij} - \overline{I}_i)(x_{hj} - \overline{I}_h)}{\sqrt{\sum_{j=1}^p (x_{ij} - \overline{I}_i)^2}\sqrt{\sum_{j=1}^p (x_{hj} - \overline{I}_h)^2}}$$

La igualdad (1.28) muestra que la distancia euclídea entre dos individuos, considerados como puntos o vectores del espacio $\mathbb{R}^p$, puede descomponerse en dos componentes. La primera refleja las diferencias de nivel, es decir, las diferencias entre las puntuaciones medias de cada individuo (Figura 1.4), mientras que el resto mide las diferencias de forma que pueden encontrarse entre los dos individuos, tanto en la dispersión respecto del nivel medio, como en el grado de relación lineal existente entre los dos.

Lo que ocurre en el ejemplo propuesto por Fleiss y Zubin [48] está claro. La distancia al cuadrado entre los individuos I y II, intuitivamente más similares, se descompondría en

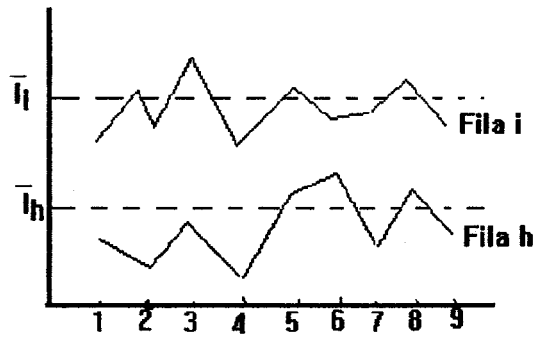


Figura 1.4: Diferencias entre el nivel medio y la dispersión entre dos individuos

$$0.25 = 0.05 + 0.20$$

mientras que entre el individuo I y el III, daría

$$7.5 = 5 + 2.5$$

lo que indica que la mayor o menor proximidad entre dos individuos no va a estar directamente relacionado sólo con el coeficiente $q^*_{ih}$, sino que depende mucho de las diferencias de niveles. Es evidente que si queremos corregir este problema, la solución apunta a comparar igualando primero los niveles medios.

Algo parecido ocurre también con el coeficiente de correlación sin centrar los datos como el $q_{ih}$ introducido en (1.3.4). Es fácil ver que

$$d^2(\mathbf{I}_i, \mathbf{I}_h) = \sum_{j=1}^{p} x_{ij}^2 + \sum_{j=1}^{p} x_{hj}^2 - 2 \cdot \sqrt{\sum_{j=1}^{p} x_{ij}^2} \cdot \sqrt{\sum_{j=1}^{p} x_{hj}^2} \cdot q_{ih}$$

lo que indica nuevamente que la proximidad entre dos individuos no sólo dependerá del coeficiente $q_{ih}$, sino también del módulo de cada vector $\mathbf{I}_i$ e

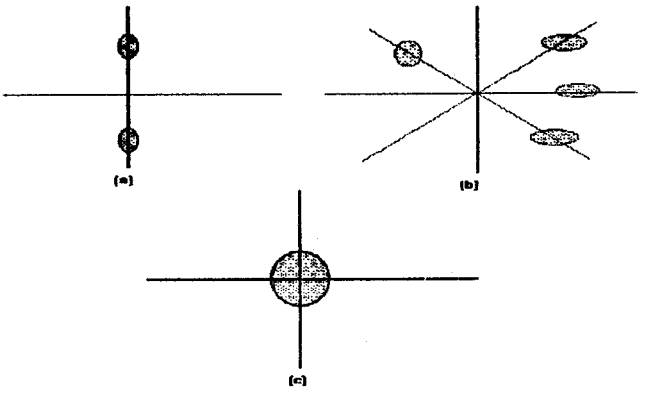

Figura 1.5: Algunos casos (Fleiss y Zubin, 1969) en los que la correspondencia entre clusters y factores no es unívoca

$\mathbf{I}_h$. De ahí que, como veremos en las secciones siguientes, este coeficiente de similaridad se utilice entre los individuos previamente normalizados.

Un último aspecto que apuntan Fleiss y Zubin [48] y Stewart [121] es que los factores no son clusters. Los clusters se definen como individuos que se encuentran relativamente cerca entre sí en el espacio, mientras que un factor hay que entenderlo como un eje que desde el punto de vista lineal, se obtiene haciendolo pasar por grupos más o menos homogeneos de individuos (Figura 1.5 a),b),c)). No hay reglas bien establecidas para definir un cluster y sí están muy claras para definir un factor. Cattell [24] proporciona un tratamiento más completo de esta distinción, apuntando que en algunos conjuntos de datos, puede haber más clusters que factores, aunque un cluster siempre pueda acercarse a un factor particular. El hecho de que el análisis factorial proporcione un conjunto de ejes coordenados, permite definir numerosos clusters por la posición de los individuos en cualquier dimensión o en combinaciones lineales de ellas.

El problema anterior está también relacionado con el hecho de que el número máximo de factores extraibles de una matriz de correlaciones no puede ser mayor que el rango de la matriz, es decir, en el caso en el que

utilicemos las Q correlaciones, no pueden extraerse más de $p - 1$ factores. Esto limita el número de tipos, patrones o clusters que pueden obtenerse con este procedimiento, y la pregunta es obvia: ¿qué hace un investigador que elija trabajar con tres o cuatro variables y que sospeche que existen al menos cuatro o cinco tipos diferentes? ¿debe añadir nuevas variables aunque estas no sean necesarias, sólo para aumentar la posibilidad de extraer más factores?.

Es evidente que algunos de los problemas planteados no tienen solución tal y como está concebida la técnica en modo Q. Sin embargo, veremos en la siguiente sección cómo se han resuelto algunos de ellos en casos concretos, e introduciremos algunas modificaciones que nos van a permitir utilizar la técnica como un procedimiento para descubrir clusters dentro de un conjunto de individuos.

## 1.4 Formulación conjunta de los modelos mediante la D.V.S.

Como se ha comentado, en su origen el modelo de Análisis Factorial se introdujo para tratar de expresar un conjunto de variables aleatorias $x_1, \ldots, x_p$ observadas y correladas, mediante un conjunto menor de variables aleatorias independientes $f_1, \ldots, f_q, q < p$, llamadas factores, que dividen cada variable $x_i$ en dos partes no observadas, una "sistemática" formada por una combinación lineal de los factores $f_j$ y otra de "error", $e_i$

$$x_i = (\sum_{j=1}^{q} a_{ij} f_j) + e_i \qquad i = 1, \ldots, p \tag{1.29}$$

de tal forma que las componentes de error $e_i$ sean tan pequeñas como sea posible y tengan entre sí correlaciones próximas a cero (idealmente nulas).

Una de las técnicas más utilizadas para obtener una primera determinación de los factores es el método de las componentes principales, basado en las ideas geométricas y algebraicas del primitivo trabajo de Pearson [107]. En él se obtienen, a partir de la descomposición espectral de la matriz de covarianzas o correlaciones, un nuevo conjunto de variables independientes que son capaces de explicar, por turno, la máxima cantidad de la varianza de las variables. Cada variable nueva obtenida puede considerarse como un factor, de manera que si un pequeño grupo de ellas es capaz de explicar un porcentaje alto de la variabilidad total, el resto puede englobarse en los errores $e_i$ del modelo (1.29). Con posterioridad a la determinación del número y composición de los factores, pueden realizarse rotaciones ortogonales u oblicuas de tales factores, que faciliten su interpretación.

La Descomposición espectral de una matriz cuadrada puede considerarse como un caso particular de la denominada Descomposición en Valores Singulares (**DVS**), establecida por vez primera para matrices rectangulares por Eckart y Young [37]. Su introducción nos va a permitir analizar cualquier matriz de datos $\mathbf{X}_{n,p}$ con $n$ el número de individuos u objetos y $p$ el número de variables, bajo las hipótesis del modelo factorial, no sólo en el espacio de las variables (ó en modo R) sino también en el de los objetos (ó en modo Q), tratando de buscar en cada caso factores que expliquen la dispersión exis-

tente en los elementos de cada espacio. Además, la **DVS** nos va a permitir establecer una estrecha relación entre los modelos factoriales obtenidos para ambos tipos de elementos de una misma matriz $\mathbf{X}$, de la que extraeremos importantes consecuencias.

### 1.4.1 El Análisis Factorial Moderno

Los distintos modos de utilizar el modelo de A.F. dentro de un conjunto de datos $\mathbf{X}$ han presentado desde sus comienzos caminos a veces paralelos y a veces distintos. En esta sección pretendemos utilizar la formulación introducida por Lebart y al. [90] y posteriormente Jambu [70], para darle un tratamiento unificado y, a partir de él, hacer algunas propuestas en las que el Análisis Factorial pueda utilizarse como una técnica de clasificación.

En cualquiera de las dos formas de utilizarlo, subyace la idea de que el modelo factorial pretende dar el mejor resúmen del conjunto de datos $\mathbf{X}_{n,p}$. Esto significa, desde el punto de vista algebraico, tratar de reconstruir los $n \cdot p$ elementos de la matriz $\mathbf{X}_{n,p}$ con el menor número de datos posibles y tal que los elementos $x_{ij}$ se aproximen lo más posible a esta reconstrucción.

El caso más simple es aquél en el que $\mathbf{X}$ puede expresarse como

$$\mathbf{X}_{n,p} \approx \mathbf{v}_1 \cdot \mathbf{u}_1' \tag{1.30}$$

donde $\mathbf{v}_1$ y $\mathbf{u}_1$ son vectores columna de $n$ y $p$ componentes respectivamente, con lo que la matriz de $n \cdot p$ elementos se habrá reconstruido con $n+p$ valores.

Hay que notar que en el caso de que la expresión anterior fuera una igualdad, significaría que todos los vectores fila de la matriz $\mathbf{X}$ pueden ponerse como combinación lineal de uno ($\mathbf{u}_1'$), y analogamente ocurriría con los vectores columna. Como se sabe, en ese caso la matriz sería de rango 1, lo cual desecharemos al ser un caso trivial. Es por ello que, si como es usual, el rango de $\mathbf{X}$ es $p$ ($p < n$), lo más probable sea que la matriz $n \cdot p$ originada con el producto $\mathbf{v}_1 \cdot \mathbf{u}_1'$ tenga elementos no muy cercanos a los de $\mathbf{X}$.

Si en lugar de expresar $\mathbf{X}$ como un producto de dos vectores, lo hacemos

como una suma de dos productos

$$\mathbf{X}_{n,p} \approx \mathbf{v}_1 \cdot \mathbf{u}_1' + \mathbf{v}_2 \cdot \mathbf{u}_2'$$

estaremos aproximándonos más a los elementos de $\mathbf{X}$ al añadir un sumando más a los de (1.30). Los errores que se cometan al aproximar los $n \cdot p$ elementos por los $2p + 2n$ serán por tanto menores que antes.

Así podría seguirse, añadiendo más productos de vectores hasta que los errores que se cometieran al aproximar los elementos de $\mathbf{X}$ fueran despreciables. ¿Cuando sería ese momento?.

Hay varias formas de responder a esta pregunta, aunque todas llevarían al concepto común del rango de la matriz $\mathbf{X}$. Sin embargo, hay una definición introducida por Horst [65] para el rango de una matriz a partir de los posibles productos en que puede descomponerse, que conecta muy bien con lo que pretendemos.

### Descomposición de una matriz en factores. Rango de una matriz

En el álgebra escalar, se sabe que la descomposición de un número en producto de factores no es, en general, única. Análogamente, cualquier matriz puede expresarse como producto de dos factores de infinitas formas posibles. Aceptando la primera afirmación, que es evidente, puede demostrarse fácilmente que son infinitas las formas de descomponer $\mathbf{X}$, ya que si

$$\mathbf{X} = \mathbf{Y} \cdot \mathbf{W}$$

donde $\mathbf{Y}$ y $\mathbf{W}$ son dos matrices cumpliendo la propiedad de multiplicidad para matrices -número de columnas de $\mathbf{Y}$ coincide con número de filas de $\mathbf{W}$-, buscando una matriz $\mathbf{A}$, cuadrada y ortonormal, de orden apropiado se tendrá que

$$\mathbf{Y}^* = \mathbf{Y} \cdot \mathbf{A}$$

y

$$\mathbf{W}^* = \mathbf{A}' \cdot \mathbf{W}$$

verifican también que

$$\mathbf{X} = \mathbf{Y} \cdot \mathbf{A} \cdot \mathbf{A}' \cdot \mathbf{W} = \mathbf{Y}^* \cdot \mathbf{W}^*$$

al ser $\mathbf{AA'} = \mathbf{I}$.

Como se ha visto, la propiedad anterior se utiliza con frecuencia en el análisis factorial en modo R, ya que en él, el objetivo principal es descomponer la matriz de covarianzas o correlaciones, $\mathbf{R} = \mathbf{X'X}$, en un producto de dos matrices que verifiquen ciertas restricciones.

Es evidente que si encontramos una solución $\mathbf{R} = \mathbf{S} \cdot \mathbf{T}$, también lo será $\mathbf{R} = \mathbf{S^*} \cdot \mathbf{T^*}$ con $\mathbf{S^*} = \mathbf{A} \cdot \mathbf{S}$ ,$\mathbf{T^*} = \mathbf{A'} \cdot \mathbf{T}$ y $\mathbf{A}$ una matriz ortonormal elegida de acuerdo con las restricciones impuestas. Este es el fundamento de las rotaciones de factores que pueden efectuarse para conseguir soluciones con estructura más simple.

De la misma forma podría utilizarse en el modo Q, para la descomposición de una matriz del tipo $\mathbf{Q} = \mathbf{XX'}$, donde sus elementos serán ahora correlaciones o asociaciones entre los individuos de la matriz $\mathbf{X}$.

Un concepto que va a tener una gran importancia en la técnica de Análisis Factorial, cuando se trata de determinar la dimensión mínima posible de los factores de una matriz, es el de rango. Aunque tradicionalmente pueda definirse en términos de los menores ó de la dependencia o independencia de las filas y columnas de la matriz, aquí lo veremos desde otro punto de vista.

**Definicion**(Horst, [65]) *El rango de una matriz* $\mathbf{X}$ *es el menor orden común entre todos los pares de matrices cuyo producto es la matriz* $\mathbf{X}$.

Es evidente que esta definición es equivalente a la dada mediante los menores de la matriz, y que de forma inmediata pueden extraerse las siguientes propiedades:

**Lema**

1. *El rango de una matriz no puede exceder su dimensión menor*
2. *El rango de una matriz obtenida mediante* $\mathbf{R} = \mathbf{X'X}$ *ó* $\mathbf{Q} = \mathbf{XX'}$, *no puede exceder de la dimensión menor de* $\mathbf{X}$. *De hecho, los rangos de* $\mathbf{R}$ *y* $\mathbf{Q}$ *serán iguales al de* $\mathbf{X}$.

De la definición de rango anterior, Horst extrae el concepto de *matriz*

*básica*:

**Definicion**

*Una matriz se denomina básica si su rango es igual a su orden menor.*

Una consecuencia inmediata se recoge en el siguiente resultado:

**Lema**

*Una matriz básica no puede ser expresada como el producto de dos matrices cuyo orden común sea menor que el menor orden de la matriz.*

y, por supuesto, estos conceptos de rango y de matriz básica estarán relacionados con los de dependencia e independencia lineal de los vectores fila o columna de la matriz $\mathbf{X}$.

**Lema**

*Dada la matriz $\mathbf{X}_{n,p}$ de rango $r \leq p < n$, sólo $r$ de los $p$ vectores columna o de los $n$ vectores fila son linealmente independientes, de forma que los subespacios generados por los vectores fila o columna son de dimensión $r$. Además pueden encontrarse para estos subespacios sendas bases con $r$ elementos de entre las columnas o las filas, no de forma única.*

Así pues, en el caso general en que la matriz $\mathbf{X}_{n,p}$ sea de rango $r \leq p < n$, la descomposición de (1.30) sería exacta con $r$ vectores fila y $r$ vectores columna de la forma

$$\mathbf{X}_{n,p} = \mathbf{v}_1 \cdot \mathbf{u}_1' + \mathbf{v}_2 \cdot \mathbf{u}_2' + \ldots + \mathbf{v}_r \cdot \mathbf{u}_r' \tag{1.31}$$

pudiendo utilizar sólo $q$ términos ($q < r$) para obtener una representación aproximada

$$\mathbf{X}_{n,p} \approx \mathbf{v}_1 \cdot \mathbf{u}_1' + \ldots + \mathbf{v}_q \cdot \mathbf{u}_q' + \varepsilon_{n,p} \tag{1.32}$$

**Interpretación geométrica**

La representación (1.31) puede entenderse mejor si asociamos a la matriz $\mathbf{X}$ dos representaciones geométricas, una de las $n$ filas consideradas como las

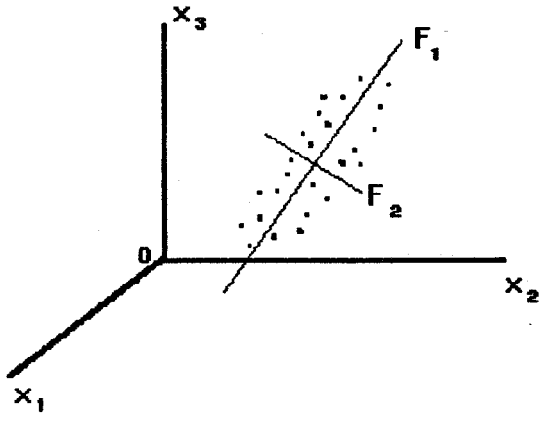


Figura 1.6: Dirección que mejor ajusta a la nube de puntos $I_i$

coordenadas de $n$ puntos o $n$ vectores en el espacio euclídeo $I\!R^p$, y otra de las $p$ columnas consideradas como vectores en el espacio $I\!R^n$.

**1. Análisis de los puntos de $I\!R^p$**

El conjunto de individuos o filas $\{I_i, i = 1, \ldots, n\}$ de $\mathbf{X}$ puede representarse como $n$ puntos de $I\!R^p$, cuyas coordenadas son los valores obtenidos por cada individuo en el conjunto $\{x_j, j = 1, \ldots, p\}$ de las $p$ variables o direcciones, que forman una base de $I\!R^p$.

En un principio, nuestra intención es buscar en este espacio una dirección (representada por un vector) respecto de la cual las coordenadas de los n puntos se aproximen lo más posible a las originales (Figura 1.6). En otros términos, y suponiendo definida la métrica euclídea en $I\!R^p$, se busca la recta que mejor ajuste a la nube de puntos formada por el conjunto $\{I_i, i = 1, \cdots, n\}$

Si es $\mathbf{u}_1$ el vector unitario en la dirección $F_1$ buscada, será un vector columna $p \cdot 1$, donde las $p$ coordenadas de $\mathbf{u}_1$ serán las determinadas por su expresión lineal respecto de la base de $I\!R^p$ formada por $\{\mathbf{x}_j, j = 1, \ldots, p\}$.

El producto $\mathbf{X} \cdot \mathbf{u}_1$ será un vector cuyas componentes son los productos escalares de los vectores $\vec{\mathbf{OI}}_i$ con $\mathbf{u}_1$, y por tanto serán las longitudes de las

proyecciones de los $n$ puntos en $F_1$. Si lo que se pretende es determinar $F_1$ con la condición de reproducir lo mejor posible los $I_i$ con sus proyecciones en esa dirección, el criterio para determinarla puede ser que la suma de cuadrados de las proyecciones sea máxima. Con esto, el problema a resolver será encontrar un vector $\mathbf{u}$, que supondremos unitario $\mathbf{u}'\mathbf{u} = 1$, y que verifique

$$\max_{\mathbf{u}} (\mathbf{Xu})'(\mathbf{Xu}) = \max_{\mathbf{u}} \mathbf{u}'\mathbf{X}'\mathbf{Xu}$$

Si ahora queremos ajustar la nube de puntos al mejor subespacio de dimensión 2 buscando dos direcciones $F_1$ y $F_2$ respecto de la cuales las coordenadas de los n puntos aproximen lo más posible a las originales, podemos razonar de la misma forma y tener que

- Este subespacio deberá contener al vector $\mathbf{u}_1$ obtenido anteriormente
- El vector $\mathbf{u}_2$, ortogonal a $\mathbf{u}_1$, se obtendrá de forma que $\mathbf{u}_2'\mathbf{X}'\mathbf{X}\mathbf{u}_2$ sea máxima y siendo $\mathbf{u}_1$ y $\mathbf{u}_2$ ortonormales, es decir, $\mathbf{u}_1'\mathbf{u}_1 = 1$, $\mathbf{u}_2'\mathbf{u}_2 = 1$ y $\mathbf{u}_1'\mathbf{u}_2 = 0$.

Los dos vectores $\mathbf{u}_1$ y $\mathbf{u}_2$ obtenidos generarán el subespacio buscado.

Así podríamos continuar tratando de ajustar la nube de puntos a un subespacio de dimensión $q$ buscando $q$ direcciones, para lo cual tendremos que

- Este subespacio deberá contener a los vectores $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{q-1}$ obtenidos anteriormente.
- El vector $\mathbf{u}_q$ se obtendrá de forma que $\mathbf{u}_q'\mathbf{X}'\mathbf{X}\mathbf{u}_q$ sea máxima, siendo $\mathbf{u}_q'\mathbf{u}_q = 1$ y $\mathbf{u}_\alpha'\mathbf{u}_\beta = 0$ para $\alpha \neq \beta = 1, 2, \ldots, q$.

Al igual que antes, los $q$ vectores $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{q-1}, \mathbf{u}_q$ generarán el subespacio buscado.

Conviene que nos paremos un poco para analizar lo que se está haciendo. Cada individuo $I_i$ de $\mathbb{R}^p$, que en un principio tenía p componentes respecto de

las $p$ direcciones $x_1, \ldots, x_p$, se va a poder expresar ahora de forma aproximada mediante $q$ componentes, las $q$ proyecciones que el punto $I_i$ determinará en las $q$ direcciones buscadas anteriormente. Las proyecciones de cada $I_i$ en cada $\mathbf{u}_\alpha$, $\alpha = 1, \ldots, q$ vendrán dadas por las $n$ componentes del vector columna $\mathbf{X}\mathbf{u}_\alpha$. Por último, aunque todavía no hemos hallado los vectores $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_q$, es evidente que éstos deberán estar referidos a la base $\mathbf{x}_1, \ldots, \mathbf{x}_p$ de $I\!R^p$, es decir, cada uno tendrá $p$ coordenadas respecto de esta base y, por lo tanto, podrían ser considerados como "nuevos individuos" o "factores" que tratan de explicar la variabilidad existente entre los $n$ de partida.

## 2. Análisis de los puntos de $I\!R^n$

En $I\!R^n$, el conjunto de las columnas de $\mathbf{X}$ formará también una nube de puntos o de vectores $\{\mathbf{x}_j, j = 1, \ldots, p\}$, cuyas coordenadas vendrán dadas respecto de la base formada por $\{I_i, i = 1, \ldots, n\}$.

Si pretendemos buscar también aquí el subespacio de dimensión $1, 2, \ldots, s$ que mejor ajuste a la nube de puntos, utilizando el mismo razonamiento, buscaremos primero la dirección $G_1$ que mejor la ajuste.

Si es $\mathbf{v}_1$ el vector unitario que la determina, las proyecciones de los $\mathbf{x}_j$ sobre él serán ahora $\mathbf{X}'\mathbf{v}_1$. Por tanto, habrá que maximizar

$$(\mathbf{X}'\mathbf{v}_1)'(\mathbf{X}'\mathbf{v}_1) = \mathbf{v}_1'\mathbf{X}\mathbf{X}'\mathbf{v}_1 \qquad \text{con} \qquad \mathbf{v}_1'\mathbf{v}_1 = 1$$

Análogamente al caso en $I\!R^p$, el subespacio de dimensión $s$ que mejor ajuste a la nube vendrá determinado por los vectores $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_s$ que maximizen $\mathbf{v}'\mathbf{X}\mathbf{X}'\mathbf{v}$ con las condiciones de que $\mathbf{v}_\alpha'\mathbf{v}_\alpha = 1$ y $\mathbf{v}_\alpha'\mathbf{v}_\beta = 0$ para $\alpha \neq \beta = 1, 2, \ldots, s$

También aquí conviene que analizemos lo que se está haciendo. Cada variable $\mathbf{x}_j$ de $I\!R^n$, que en un principio tiene $n$ componentes respecto de las $n$ direcciones $I_1, \ldots, I_n$, se va a poder expresar ahora de forma aproximada mediante $s$ componentes, las $s$ proyecciones que el vector $\mathbf{x}_j$ determinará en las $s$ direcciones buscadas anteriormente. Las proyecciones de cada $\mathbf{x}_j$ en cada $\mathbf{v}_\alpha$, $\alpha = 1, \ldots, s$ vendrán dadas por las $p$ componentes del vector columna $\mathbf{X}'\mathbf{v}_\alpha$. Y al igual que en $I\!R^p$, aunque todavía no hemos hallado los vectores $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_s$, es evidente que éstos deberán estar referidos a la

base $I_1, \ldots, I_n$ de $\mathbb{R}^n$, es decir, cada uno tendrá $n$ coordenadas respecto de esta base y, por lo tanto, podrían ser considerados como "nuevas variables" o "factores" que resumirían la información contenida en los $p$ de partida.

### La Descomposición en Valores Singulares de una matriz rectangular

La solución que ha quedado pendiente en los análisis de los puntos de $\mathbb{R}^p$ y $\mathbb{R}^n$ puede obtenerse a partir de la denominada Descomposición en Valores Singulares (**DVS**) de una matriz.

La **DVS** de una matriz $\mathbf{X}$ cualquiera, que puede considerarse como una generalización de la Descomposición Espectral de una matriz cuadrada, fue abordada por primera vez para este tipo de matrices por el matemático inglés Sylvester [122] en un artículo publicado en 1889 y, posteriormente, Autonne (1913,1915) y definitivamente Eckart y Young [37] completaron la extensión a matrices rectangulares. Su introducción nos va a permitir analizar cualquier matriz de datos $\mathbf{X}_{n,p}$, bajo las hipótesis del Modelo Factorial, en el espacio de las variables y en el de los individuos, tratando de buscar en cada caso factores que traten de explicar la dispersión existente en los elementos de cada espacio. Además, la **DVS** nos va a permitir establecer una estrecha relación entre los modelos factoriales obtenidos para ambos tipos de elementos de una misma matriz $\mathbf{X}$.

La versión más general de la **DVS** para el caso de matrices rectangulares reales está recogida en el siguiente resultado ([78]):

**Teorema**

*Dada una matriz* $\mathbf{X}$ *de orden nxp, con rango*$(\mathbf{X})$ *igual a* $r(\leq p < n)$, *existen una matriz* $\mathbf{U}_{p,r}$ *real, otra matriz* $\mathbf{V}_{n,r}$ *real y una matriz diagonal* $\Gamma_r$ *con elementos* $\gamma_1, \ldots, \gamma_r$ *positivos, denominados valores singulares de* $\mathbf{X}$, *tales que*

$$\mathbf{X} = \mathbf{V}\Gamma\mathbf{U}'$$

*donde,*

1. $\gamma_1^2, \gamma_2^2, \ldots, \gamma_r^2$ *son los autovalores y las columnas de* $\mathbf{U}$ *los correspon-*

*dientes autovectores en la descomposición espectral de la matriz cuadrada semidefinida positiva $X'X$ de orden $p \cdot p$ y de rango $r$.*

2. *Análogamente, $\gamma_1^2, \gamma_2^2, \ldots, \gamma_r^2$ y las columnas de $\mathbf{V}$ son los autovalores y autovectores de la matriz cuadrada semidefinida positiva $XX'$ de orden $n \cdot n$ y de rango $r$.*

3. *Tanto las columnas de $\mathbf{U}$ como las de $\mathbf{V}$ son ortonormales respectivamente.*

4. *Los autovalores positivos de $\mathbf{X'X}$ y $\mathbf{XX'}$ son los mismos. Además, si $\mathbf{u}_\alpha$ es el autovector de $\mathbf{X'X}$ y $\mathbf{v}_\alpha$ el autovector de $\mathbf{XX'}$, que corresponden ambos al autovalor ${\gamma_\alpha}^2$, se verifican las siguientes relaciones, para $\alpha = 1, \ldots, r$:*

$$\mathbf{u}_\alpha = \gamma_\alpha^{-1}\mathbf{X'v}_\alpha$$

*y*

$$\mathbf{v}_\alpha = \gamma_\alpha^{-1}\mathbf{Xu}_\alpha$$

*En forma matricial, las relaciones anteriores son*

$$\mathbf{V} = \mathbf{XU\Gamma}^{-1} \quad y \quad \mathbf{U} = \mathbf{X'V\Gamma}^{-1}$$

En la práctica, los pasos para realizar la descomposición en valores singulares son los siguientes:

1. Se calculan $\mathbf{XX'}$ ó $\mathbf{X'X}$, y se toma aquélla que sea de menor orden. Por lo general, será $\mathbf{X'X}$

2. Se calculan los autovalores positivos $\lambda_1, \lambda_2, \ldots, \lambda_r$ de $X'X$ y los correspondientes autovectores $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r$.

3. Se calculan los valores singulares $\gamma_1 = \sqrt{\lambda_1}, \gamma_2 = \sqrt{\lambda_2}, \ldots, \gamma_r = \sqrt{\lambda_r}$.

4. Se calculan $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r$ mediante

$$\mathbf{v}_\alpha = \gamma_\alpha^{-1}\mathbf{Xu}_\alpha \qquad \alpha = 1, 2, \ldots, r$$

Mediante la **DVS**, la matriz $\mathbf{X}$ puede expresarse de la forma

$$\mathbf{X} = \gamma_1 \mathbf{v}_1 \mathbf{u}'_1 + \gamma_2 \mathbf{v}_2 \mathbf{u}'_2 + \cdots + \gamma_r \mathbf{v}_r \mathbf{u}'_r \tag{1.33}$$

donde a $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r$ se les denomina vectores singulares a izquierda de $\mathbf{X}$, a $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r$ vectores singulares a derecha de $\mathbf{X}$ y a las matrices $\mathbf{v}_\alpha \mathbf{u}'_\alpha$ de dimensión $n \cdot p$, se les denomina planos singulares. Se dice que la matriz $\mathbf{X}$ de rango $r$ se ha descompuesto como combinación lineal de $r$ matrices de rango 1.

### Obtención de las direcciones en $\mathbb{R}^p$ y $\mathbb{R}^n$ y relaciones entre ellas

La obtención de las direcciones $F_1, F_2, \ldots, F_q$ en $\mathbb{R}^p$ y $G_1, G_2, \ldots, G_s$ en $\mathbb{R}^n$ está resuelta en el siguiente resultado ([70]):

**Teorema**

*Dada la matriz* $\mathbf{X}_{n,p}$ *de rango* $r \leq p < q$,

1. *Pueden encontrarse* $r$ *vectores* $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r$ *en* $\mathbb{R}^p$ *con la condición de que que* $\mathbf{u}'_\alpha \mathbf{X}' \mathbf{X} \mathbf{u}_\alpha$, $\alpha = 1, \ldots, r$ *sea máxima, siendo* $\mathbf{u}'_\alpha \mathbf{u}_\alpha = 1$ *y* $\mathbf{u}'_\alpha \mathbf{u}_\beta = 0$ *para* $\alpha \neq \beta = 1, 2, \ldots, q$.

2. *Los vectores* $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r$ *son los autovectores correspondientes a los autovalores positivos ordenados en orden decreciente de la matriz* $\mathbf{X}'\mathbf{X}$.

3. *Análogamente, pueden encontrarse* $r$ *vectores* $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r$ *en* $\mathbb{R}^n$ *con la condición de que* $\mathbf{v}'_\alpha \mathbf{X} \mathbf{X}' \mathbf{v}_\alpha$, $\alpha = 1, \ldots, r$ *sea máxima, siendo* $\mathbf{v}'_\alpha \mathbf{v}_\alpha = 1$ *y* $\mathbf{v}'_\alpha \mathbf{v}_\beta = 0$ *para* $\alpha \neq \beta = 1, 2, \ldots, q$.

4. *Los vectores* $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r$ *son los autovectores correspondientes a los autovalores positivos ordenados en orden decreciente de la matriz* $\mathbf{X}\mathbf{X}'$.

Los resultados anteriores no sólo indican qué son las direcciones que buscábamos en los dos espacios, sino que además, utilizando la **DVS** de $\mathbf{X}$ podemos deducir que:

- Los autovalores positivos $\lambda_1, \lambda_2, \ldots, \lambda_r$ de $\mathbf{X}'\mathbf{X}$ y $\mathbf{X}\mathbf{X}'$ son los mismos y, además, verifican que

$$\mathbf{u}_\alpha' \mathbf{X}'\mathbf{X}\mathbf{u}_\alpha = \lambda_\alpha \quad \alpha = 1, \ldots, r$$

y

$$\mathbf{v}_\alpha' \mathbf{X}\mathbf{X}'\mathbf{v}_\alpha = \lambda_\alpha \quad \alpha = 1, \ldots, r$$

es decir, son las sumas de los cuadrados de la proyecciones de los $I_i$ sobre las direcciones $F_\alpha$, lo cual es una medida de la dispersión que es capaz de explicar cada dirección. Análogamente, son las sumas de las proyecciones de las $\mathbf{x}_j$ sobre las direcciones $G_\alpha$.

- Los vectores $\mathbf{u}_\alpha$ y $\mathbf{v}_\alpha$ correspondientes al autovalor $\lambda_\alpha$ verifican para $\alpha = 1, \ldots, r$:

$$\mathbf{u}_\alpha = \lambda_\alpha^{-1/2} \mathbf{X}'\mathbf{v}_\alpha$$

y

$$\mathbf{v}_\alpha = \lambda_\alpha^{-1/2} \mathbf{X}\mathbf{u}_\alpha$$

es decir, las coordenadas de los puntos $I_i$ de $\mathbb{R}^p$ en cada eje $F_\alpha$ son proporcionales a las componentes del vector unitario $\mathbf{v}_\alpha$. Por su parte, las coordenadas de los puntos $x_j$ de $\mathbb{R}^n$ en cada eje $G_\alpha$ son proporcionales a las del vector unitario $\mathbf{u}_\alpha$. En forma matricial, las relaciones anteriores pueden escribirse como:

$$\mathbf{V} = \mathbf{\Lambda}^{-1/2}\mathbf{X}\mathbf{U} \qquad \text{y} \qquad \mathbf{U} = \mathbf{\Lambda}^{-1/2}\mathbf{X}'\mathbf{V} \tag{1.34}$$

- Las nubes de puntos de $\mathbb{R}^p$ y $\mathbb{R}^n$ pueden reconstruirse mediante el nuevo conjunto de coordenadas dado por las direcciones $F_\alpha$ y $G_\alpha$.

Para la reconstrucción exacta podemos utilizar la descomposición (1.33) que nos da la **DVS**

$$\mathbf{X}_{(n,p)} = \sum_{\alpha=1}^{r} \lambda_\alpha^{1/2} \mathbf{v}_\alpha \cdot \mathbf{u}_\alpha'$$

Si utilizáramos sólo la primera dirección $F_1$ o $G_1$, que sería la que recoge la máxima variabilidad dada por el autovalor $\lambda_1$, la matriz $\mathbf{X}$ puede aproximarse

en $I\!R^p$ por las coordenadas de los puntos $I_i$ a lo largo de $F_1$ y en $I\!R^n$ por las coordenadas de los puntos $x_j$ a lo largo de $G_1$. Se tendría entonces que:

$$\mathbf{X} \approx \lambda_1^{\frac{1}{2}} \mathbf{v}_1 \cdot \mathbf{u}_1'$$

Si tomáramos los $q$ primeros autovalores $\lambda_1, \ldots, \lambda_q$, de forma que $\lambda_1 + \ldots + \lambda_r$ represente una parte importante de la variabilidad de la nube de puntos en $I\!R^p$ o $I\!R^n$, se tendrá que:

$$\mathbf{X} \approx \sum_{\alpha=1}^{q} \lambda_\alpha^{\frac{1}{2}} \mathbf{v}_\alpha \cdot \mathbf{u}_\alpha'$$

y la reconstrucción será mejor cuanto que $\sum_{\alpha=1}^{q} \lambda_\alpha$ se acerque a la variabilidad total $\sum_{\alpha=1}^{r} \lambda_\alpha$. Por ello, se suele tomar como coeficiente de la "calidad" de la reconstrucción al cociente:

$$\tau_q = \frac{\sum_{\alpha=1}^{q} \lambda_\alpha}{\sum_{\alpha=1}^{r} \lambda_\alpha}$$

La expresión anterior de $\mathbf{X}$ puede escribirse también en forma matricial como:

$$\mathbf{X} \approx \mathbf{V}_q \mathbf{\Lambda}_q^{\frac{1}{2}} \mathbf{U}_q'$$

donde $\mathbf{U}_q$ y $\mathbf{V}_q$ son las matrices $p \cdot q$ y $n \cdot q$ cuyas columnas son los $q$ autovectores de $\mathbf{X}'\mathbf{X}$ y $\mathbf{X}\mathbf{X}'$ correspondientes a los $q$ primeros autovalores, y $\Lambda_q$ es la matriz diagonal cuyos elementos son tales autovalores.

### 1.4.2 Diferentes modelos de Análisis Factorial

El procedimiento de obtención de los ejes factoriales en $I\!R^p$ y $I\!R^n$ descritos en la sección anterior es general. En él se ha partido de una matriz de datos $\mathbf{X}$ que en principio puede ser la de los datos originales, pero que también podría ser una matriz de los datos transformada previamente.

Si tenemos en cuenta que las direcciones obtenidas en cada espacio son las determinadas por los autovectores de $\mathbf{X}'\mathbf{X}$ y de $\mathbf{X}\mathbf{X}'$, no es de extrañar

que nos planteemos ahora transformar la matriz $\mathbf{X}$ para que los productos anteriores cobren un significado especial. Pretendemos que los elementos

$$a_{jk} = \sum_{i=1}^{n} x_{ij} x_{ik} \qquad j, k = 1, \ldots, p$$

de la matriz $\mathbf{A} = \mathbf{X'X}$, y

$$b_{ik} = \sum_{j=1}^{p} x_{ij} x_{kj} \qquad i, k = 1, \ldots, n$$

de la matriz $\mathbf{B} = \mathbf{XX'}$ tengan un significado no sólo geométrico, sino también estadístico, que permitan llegar a conclusiones más claras.

Dependiendo de si estamos interesados en obtener factores entre las variables o entre los elementos, la matriz $\mathbf{X}$ puede transformarse para que $\mathbf{A}$ y $\mathbf{B}$ representen coeficientes o parámetros familiares. Veremos en las secciones siguientes algunos casos que permitirán mostrarnos cómo los modelos de Análisis Factorial para variables y para elementos, introducidos en las secciónes anteriores, están presentes en este más general.

Aunque el Análisis Factorial Moderno también contempla el caso en el que las variables tratadas sean cualitativas o mezcla de cualitativas y cuantitativas, el caso más frecuente y al que nos limitaremos aquí, es aquel en que las variables sean cuantitativas.

### El modelo para variables con la matriz de covarianzas

Si estamos interesados en descubrir factores entre las variables, representadas por vectores en $I\!R^n$, hemos visto que estas direcciones se obtienen descomponiendo espectralmente $\mathbf{X'X}$, matriz de dispersión de los puntos de $I\!R^p$.

Si el conjunto de las $p$ variables es heterogéneo en cuanto a sus valores medios y sus varianzas, a causa de que cada una mida magnitudes distintas, una forma de evitar tal heterogeneidad entre los valores medios es sustituir los elementos $x_{ij}$ de $\mathbf{X}$ por sus desviaciones respecto de la media $y_{ij} = x_{ij} - \overline{x}_j$. Gráficamente, en $I\!R^p$, esta transformación (Figura 1.7) tiene el efecto de

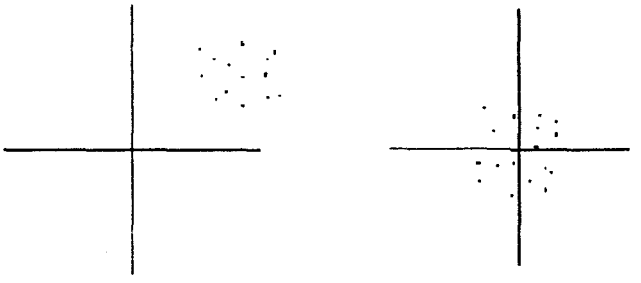

Figura 1.7: Centralización de una nube de puntos.

trasladar la nube de puntos al centro de coordenadas: mientras que en $I\!R^n$, los vectores $\mathbf{x}_j$ que representaban a las variables se habrán desplazado (Figura 1.8) por el vector $\mathbf{1}\overline{x}_j$ cuyas coordenas serán todas iguales a $\overline{x}_j$

Lo anterior no sufre cambios significativos si además utilizamos un coeficiente de normalización ($\sqrt{n}$) por el que dividiremos las desviaciones con el fin de obtener una matriz conocida al calcular $\mathbf{X}'\mathbf{X}$. En efecto, si la transformación que efectuamos es

$$y_{ij} = \frac{x_{ij} - \overline{x}_j}{\sqrt{n}}$$

entonces los elementos de la matriz $\mathbf{Y}'\mathbf{Y}$ serán

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

es decir, $\mathbf{Y}'\mathbf{Y}$ será la matriz de varianzas y covarianzas entre las $p$ variables.

Por otra parte, en $I\!R^n$, las variables $y_j$ han sido desplazadas respecto de las $x_j$ de forma que las longitudes de los vectores que forman cada variable serán

$$L(y_j) = \sqrt{\frac{\sum_{i=1}^{n} (x_{ij} - x_j)^2}{n}} = \sigma_j$$

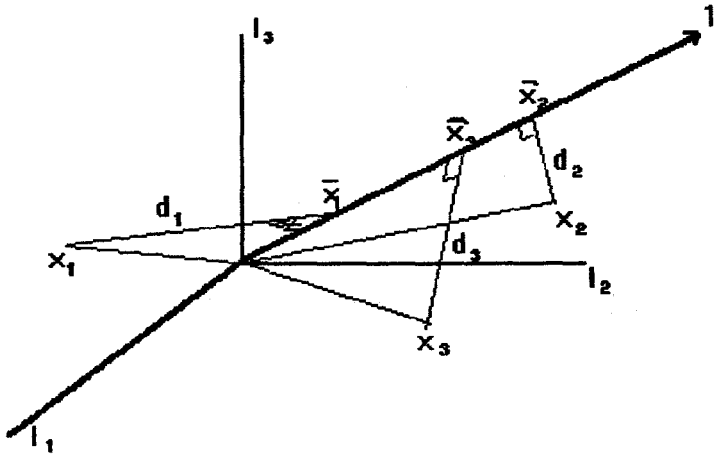


Figura 1.8: Vectores variables desplazados por $1\bar{x}_j$

es decir, las desviaciones típicas de cada variable.

Si son $\lambda_1, \lambda_2, \ldots, \lambda_q$, $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_q$ y $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_q$ los $q$ primeros autovalores y autovectores de $\mathbf{Y}'\mathbf{Y}$ y $\mathbf{Y}\mathbf{Y}'$, podremos poner

$$\mathbf{X} \approx \mathbf{V}_q \mathbf{\Lambda}_q^{\frac{1}{2}} \mathbf{U}_q' = \mathbf{F}\mathbf{A}'$$

con $\mathbf{F}$ la matriz $n \cdot q$ cuyas columnas van a ser las coordenadas que cada nueva dirección tiene respecto de los $I_i$ en $\mathbb{R}^n$, y $\mathbf{A}$ la matriz $p \cdot q$ cuyas columnas van a ser las coordenadas que tienen las $q$ nuevas direcciones respecto de las variables originales $y_j$ en $\mathbb{R}^p$.

Hay dos formas de elegir $\mathbf{F}$ y $\mathbf{A}$, originadas por la **DVS**, que aunque son esencialmente las mismas, difieren en la longitud que tendrán los factores:

- **Forma 1**

  Tomamos $\mathbf{F} = \mathbf{V}_q$ y $\mathbf{A} = \mathbf{U}_q \mathbf{\Lambda}_q^{1/2}$

  Es fácil probar que como una consecuencia de que las columnas de $\mathbf{Y}$ suman cero, las de $\mathbf{V}_q$ también. Por tanto, como $\mathbf{F}'\mathbf{F} = \mathbf{V}_q'\mathbf{V}_q = \mathbf{I}$ los factores serán incorrelados y estandarizados, es decir, representados

en $\mathbb{R}^n$ serán $q$ vectores perpendiculares y unitarios. Por otra parte, la matriz $\mathbf{A}$ será según (1.34)

$$\mathbf{A} = \mathbf{U}_q \mathbf{\Lambda}_q^{1/2} = \mathbf{X}'\mathbf{V}_q$$

es decir, tendrá por elementos $a_{j,\alpha}$ las proyecciones de cada variable $x_j$ en cada factor $F_\alpha$. Al ser en esta solución todos los factores unitarios, las columnas de $\mathbf{A}$ son directamente comparables para ver la importancia que representa cada factor $F_\alpha$ en la variable $x_j$. A esta matriz es a la que en la sección **1.2** se le ha denominado matriz de pesos, o patrón factorial.

Pero además, en esta forma la matriz $\mathbf{A}$ también representa las covarianzas entre las variables y los factores, es decir la matriz de estructura factorial. Si, en particular, deseáramos obtener la matriz de correlaciones entre las variables y los factores, sólo tendríamos que dividir cada fila de $\mathbf{A}$ por la desviación típica de las variables.

- **Forma 2**

  Tomamos $\mathbf{A} = \mathbf{U}_q$ y $\mathbf{F} = \mathbf{V}_q \mathbf{\Lambda}_q^{1/2}$

  Con esta elección, los factores siguen siendo incorrelados, pero $\mathbf{F}'\mathbf{F} = \mathbf{\Lambda}_q$, es decir, las desviaciónes típicas, y por tanto las longitudes de los factores no son iguales. Con esto, las columnas de $\mathbf{A}$ no van a ser directamente comparables.

### El modelo para variables con la matriz de correlaciones

Cuando la heterogeneidad entre las variables se debe no sólo a los valores medios, sino también a la dispersión de cada una, la transformación que puede hacerse es:

$$z_{ij} = \frac{x_{ij} - \overline{x}_j}{\sigma_j \sqrt{n}}$$

Con esto, los elementos de la matriz $\mathbf{Z}'\mathbf{Z}$ serán:

$$r_{jk} = \sum_{i=1}^{n} \frac{x_{ij} - \overline{x}_j}{\sigma_j \sqrt{n}} \cdot \frac{x_{ik} - \overline{x}_k}{\sigma_k \sqrt{n}} = \frac{\sigma_{jk}}{\sigma_j \sigma_k}$$

es decir, que $\mathbf{Z}'\mathbf{Z}$ será la matriz de correlaciones entre las $p$ variables.

Gráficamente, en $I\!R^p$ esta transformación tendrá el efecto de trasladar la nube de puntos al centro de coordenadas y homogeneizar la dispersión. Si ésta se representara por el volúmen de la nube de puntos, sería como pasar de un volúmen en forma de hiperelipsoide a otro con forma de hiperesfera.

Por otra parte, en $I\!R^n$, las variables $z_j$ han sido también desplazadas respecto de las $x_j$ y las longitudes de los vectores que forman las nuevas variables serán ahora la unidad :

$$L(z_j) = \sqrt{\frac{\sum_{i=1}^{n}(x_{ij} - x_j)^2}{\sigma_j \cdot n}} = 1$$

Al igual que antes, si son $\lambda_1, \lambda_2, \ldots, \lambda_q$, $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_q$ y $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_q$ los $q$ primeros autovalores y autovectores de $\mathbf{Z}'\mathbf{Z}$ y $\mathbf{Z}\mathbf{Z}'$, podremos poner

$$\mathbf{X} \approx \mathbf{V}_q \mathbf{\Lambda}_q^{1/2} \mathbf{U}_q' = \mathbf{F}\mathbf{A}'$$

y tendremos nuevamente las dos soluciones anteriores.

Nótese que, en este caso, la matriz patrón factorial coincide con la de estructura factorial.

### Modelos para los elementos de un conjunto

En las mismas hipótesis que se han utilizado para los modelos entre variables, la **DVS** permite utilizar la expresión:

$$\mathbf{X} \approx \mathbf{V}_q \mathbf{\Lambda}_q^{1/2} \mathbf{U}_q' = \mathbf{A}\mathbf{G}' \tag{1.35}$$

para formular un modelo en el que encontrar un conjunto de direcciones entre los elementos que permitan reconstruir la matriz $\mathbf{X}$. La matriz $\mathbf{G}$, de dimensión $p \cdot q$, representa por columnas las coordenadas que los vectores directores tendrán respecto de las variables originales $x_j$ en $I\!R^p$, mientras que la matriz $\mathbf{A}$, de dimensión $n \cdot q$, representa las coordenadas que tiene cada dirección respecto de los elementos $I_i$ en $I\!R^n$.

También aquí podremos elegir dos opcines para $\mathbf{G}$ y $\mathbf{A}$:

- **Opción 1**

  Tomamos $\mathbf{G} = \mathbf{U}_q$ y $\mathbf{A} = \mathbf{V}_q\mathbf{\Lambda}_q^{1/2}$

- **Opción 2**

  Tomamos $\mathbf{A} = \mathbf{V}_q$ y $\mathbf{G} = \mathbf{U}_q\mathbf{\Lambda}_q^{1/2}$

Los mismos argumentos de homogeneidad que se utilizaron para variables, hacen que se suela adoptar la opción 1, donde los factores, ademas de ser ortogonales, están normalizados y permiten ser comparados fácilmente mediante las columnas de $\mathbf{A}$.

En cuanto a las transformaciones que pueden hacerse sobre la matriz inicial $\mathbf{X}$ para obtener parámetros interpretables, como ya se ha comentado en la sección anterior, en un principio simplemente se intercambió el papel de las variables por los elementos (Burt [15],[16] y Stephenson [120]), y así como en los modelos para variables se estandarizaron los datos por las columnas, se hizo lo propio por las filas. El resultado es que si consideramos

$$\overline{I}_i = \frac{1}{p}\sum_{j=1}^{p} x_{ij} \quad \text{y} \quad SI_i^2 = \frac{1}{p}\sum_{j=1}^{p}(x_{ij} - \overline{I}_i)^2$$

y transformamos los datos $x_{ij}$ a

$$\frac{x_{ij} - \overline{I}_i}{SI_i \cdot \sqrt{p}}$$

los elementos de $\mathbf{XX}'$ serían una especie de coeficientes de correlación entre individuos

$$q_{ih}^* = \frac{\sum_{j=1}^{p}(x_{ij} - \overline{I}_i)(x_{hj} - \overline{I}_h)}{\sqrt{\sum_{j=1}^{p}(x_{ij} - \overline{I}_i)^2}\sqrt{\sum_{j=1}^{p}(x_{hj} - \overline{I}_h)^2}}$$

representando en $I\!R^p$ los cosenos de los ángulos que forman los vectores $I_i - \overline{I}_i \cdot \mathbf{1}$ y $I_k - \overline{I}_k \cdot \mathbf{1}$

Otra alternativa que plantea menos problemas de interpretación es la introducida también en la anterior sección. Veamos cómo se interpretaría desde el punto de vista de la **DVS**.

Si llamamos $n_i$ a la norma euclidea de cada $\mathbf{I}_i$, considerado como un vector en $I\!R^p$

$$n_i = \sqrt{\sum_{j=1}^{p} x_{ij}^2}$$

$n_i$ representará la norma del vector $\vec{OI}_i$.

Efectuando la transformación

$$w_{ij} = \frac{x_{ij}}{n_i}$$

se tendrá que la matriz $\mathbf{WW}'$ tiene ahora por elementos los coeficientes

$$q_{ih} = \frac{\sum_{j=1}^{p} x_{ij} x_{hj}}{\sqrt{\sum_{j=1}^{p} x_{ij}^2} \sqrt{\sum_{j=1}^{p} x_{hj}^2}}$$

que representan el coseno del ángulo que forman en $I\!R^p$ los vectores $\vec{\mathbf{OI}}_i$ y $\vec{\mathbf{OI}}_k$. Los valores entre $-1$ y 1, pasando por 0, que puede tomar este coeficiente representan la linealidad exacta entre ambos vectores en la misma y en distinta dirección, o la perpendicularidad en el caso de 0.

La opción 1 anterior no dará en este caso que

$$\mathbf{W} \approx \mathbf{AG}'$$

con $\mathbf{G} = \mathbf{U}_q$ y $\mathbf{A} = \mathbf{V}_q \mathbf{\Lambda}_q^{1/2}$. Las columnas de $\mathbf{G}$ serán los autovectores de $\mathbf{W}'\mathbf{W}$ y por tanto las direcciones en $I\!R^p$ que resumen a los individuos. Al ser ortonormales, facilitan la interpretación de las columnas de la matriz $\mathbf{A}$, donde se tendrá por columnas las coordenadas de los factores respecto de los individuos en $I\!R^n$.

Como también se ha mencionado en la sección anterior, este modelo de análisis es muy utilizado en algunos tipos de estudios geológicos para analizar rocas o sedimentos que se encuentran en un determinado lugar, desde el punto de vista de la composición mineralógica que presentan. Imbrie y Purdy [66], Imbrie [67], Imbrie y Van Andel [68] y posteriormente Jöreskog et al. [83] lo recogen como una de las técnicas de Análisis Factorial posibles a utilizar en Geología. En este contexto, la técnica pretende tres objetivos principales:

1. Encontrar el número mínimo, $k$, de factores o "composiciones tipo" diferentes que han servido como base a la formación de los objetos estudiados.

2. Especificar las composiciones de tales factores respecto de las especies minerales estudiadas.

3. Describir cada objeto en términos lineales de los factores, es decir, obtener para cada objeto la cantidad que aporta cada uno de los factores encontrados.

Generalmente, el carácter de la composición que presentan los objetos es porcentual, lo cual representa una cierta homogeneidad entre los elementos. En $I\!R^p$, la nube de puntos estará situada en el cuadrante de la esfera $p$-dimensional que corresponde al subconjunto

$$S = \{(x_1, x_2, \ldots, x_p)/x_i \geq 0 \quad \forall i\}$$

lo cual permite tener valores para el coeficiente $q_{ik}$ entre 0 y 1, asociando los valores próximos a 1 a aquellos objetos que presenten composiciones parecidas y próximos a 0 para los de composiciones dispares, no teniendo los problemas comentados en la sección anterior que apuntaban Fleiss y Zubin [48]. Además, este tipo de datos permite realizar el objetivo 2 anterior (Miesch [100]), es decir, especificar las composiciones porcentuales de los factores obtenidos respecto de las especies minerales estudiadas.

En cierto modo, el Análisis Factorial introducido por Imbrie y Purdy realiza una clasificación de los objetos analizados. En él, podemos utilizar la matriz de pesos **A** ó la matriz de estructura factorial si se realizan rotaciones oblicuas con los factores originales, para asignar cada objeto a un factor, de forma que todos los objetos asignados al mismo factor serán muy similares en cuanto a su composición. Además, de la matriz **G** se puede tener por columnas las composiciones porcentuales de los factores, que pueden considerarse como objetos ideales o patrones que representan a cada grupo.

En este punto, nos preguntamos si este procedimiento no podría ser utilizado como técnica de clasificación para un conjunto de datos cualquiera, quizás con las únicas hipótesis de que las variables utilizadas sean cuantitativas. Pensamos que el planteamiento expuesto en esta sección sobre el

modelo del Análisis Factorial Moderno, permite introducir algunas modificaciones que harán posible este propósito. Abordaremos estas ideas en los siguientes capítulos.

# Capítulo 2

# Un procedimiento basado en la D.V.S. que permite clasificar bajo ciertas hipótesis

## 2.1 Introducción

En este capítulo abordaremos el problema de encontrar agrupaciones naturales dentro de un conjunto de elementos del que se dispone de alguna información sobre la existencia de tales grupos homogéneos. No se va a restringir la naturaleza de los datos, en principio, más que la exigencia de que sean cuantitativos para que tenga sentido representarlos como puntos de un espacio métrico. En cuanto a número de grupos, forma o distribución de los grupos y proporción de cada grupo dentro del conjunto total, son aspectos que en algúnos casos serán considerados más adelante y en otros pueden ser objeto de futuras investigaciones.

Partiendo de este conjunto, representado por la matriz de datos $X_{n,p}$, y utilizando la terminología del Análisis Cluster donde se engloba este problema, nuestro objetivo será encontrar una partición del conjunto de elementos que obtenga los grupos existentes en el conjunto. Para ello, comenzaremos primero introduciendo en la sección **2.2** aquellos conceptos y métodos más

importantes que utiliza la técnica multivariante del Análisis Cluster.

En la sección **2.3**, queremos proponer un procedimiento de clasificación basado en el modelo factorial para elementos de un conjunto, expuesto al final del capítulo anterior. Este procedimiento será propuesto, en principio, con la restricción de que el número de grupos existentes en el conjunto no exceda al número de variables más uno, y será ilustrado con un ejemplo.

Por último, en la sección **2.4**, se hará un estudio de simulación generando una muestra grande de conjuntos de elementos en los que haya grupos naturales, que nos permitirá precisar hasta qué punto es capaz de clasificar el procedimiento, y al mismo tiempo se comparan los resultados obtenidos con uno de los procedimientos de clasificación dados en la sección **2.2**.

## 2.2 Los procedimientos del Análisis Cluster

### 2.2.1 El objetivo del Análisis Cluster

Podría definirse el Análisis Cluster como aquella técnica multivariante cuyo objetivo es realizar una clasificación de un grupo de individuos u objetos en grupos más pequeños, también llamados clusters, de forma que los objetos del mismo cluster sean lo más homogeneos posibles en algún sentido, mientras que los de diferentes clusters sean lo más heterogeneos posibles. Esta declaración de intenciones sobre lo que es el Análisis Cluster es muy sencilla en sus pretensiones, pero encierra una gran cantidad de problemas abiertos que no se han resuelto todavía. A pesar de que este tipo de procedimientos no son recientes dentro del Análisis Multivariante, hay algunos aspectos de la definición anterior que parecen tener difícil solución, en la mayoría de los casos debido a la subjetividad que encierran algunos problemas de clasificación.

Para empezar conviene que aclararemos algunos términos. En muchas aplicaciones del Análisis Cluster, se supone que los objetos pertenecen a grupos naturales mientras que, en otros casos, la cuestión es simplemente encontrar la mejor forma de agruparlos. En el primer caso se dice que estamos tratando con la **clasificación** de los objetos y en el segundo con la **disección**. Otros términos utilizados también cuando se trata de descubrir o confirmar que existen grupos naturales es el **reconocimiento de patrones** y la **taxonomía numérica**.

Durante los últimos cincuenta años se han desarrollado una gran variedad de técnicas de Análisis Cluster. Éstas han tratado siempre de dar solución a los dos problemas principales que tiene este tipo de Análisis. Por un lado, la construcción de diversos tipos de medidas de semejanza entre los objetos, dependiendo de lo que se entienda por homogeneidad en el conjunto de individuos u objetos que tratemos y del tipo de variables que se observen en ellos. Por otro, los problemas originados al buscar técnicas algorítmicas adecuadas que permitan ir formando los grupos con las características expuestas al principio.

Fruto de esta investigacin son las numerosas referencias que han aparecido desarrollando las cuestiones anteriores, sobre todo en los años 70 y 80. Entre las más destacables por contener una recopilación de todo lo aparecido sobre el tema figuran los manuales de Jardine y Sibson [75], Sneath y Sokal [116], Anderberg [1], Everitt [42] y [44], Hartigan [63] y Gordon [56].

De especial importancia para conocer el estado de la cuestión son los artículos de Cormack [28] y Everitt [43]. En el primero, Corkmack hace un repaso de la bibliografía que ha aparecido hasta ese momento sobre el tema, destacando aquellas que en su opinión merecen relevancia sobre temas como medidas de similaridad y de clusters, los principios en los que se basan las técnicas de clasificación empíricas y las limitaciones y fallos en su utilizacin. En este sentido, subraya la importancia de aquellos métodos basados en formulaciones matemáticas bien definidas, y sugiere otras formas de resumir datos alternativas a la clasificación. Por último, censura la tendencia creciente a ver la taxonomía numérica como una alternativa satisfactoria para clasificar ideas.

En el artículo de Everitt, se hace primero un repaso de los problemas aún no resueltos por el Análisis Cluster, a pesar del gran número de técnicas que hasta ese momento han aparecido. Problemas como el de elegir el mejor número de clusters, el mejor método según el tipo de datos o el problema de las interacciones "peligrosas" que pueden existir en algunos problemas entre métodos de cluster y algunas técnicas gráficas, son algunos de los que se comentan en el artículo con vistas a futuros desarrollos que intenten resolverlos.

### 2.2.2 Medidas de similaridad y disimilaridad

La mayoría de los métodos de análisis cluster se han desarrollado apoyándose en algún parámetro cuantitativo que fuera capaz de medir cuánto de parecidos o de distintos son dos individuos u objetos entre sí, o por utilizar un término que no hay que entender sólo en su acepción geométrica, cuánto de próximos están entre sí.

Hay muchas formas de medir la proximidad, dependiendo principalmente del tipo de datos utilizados, en particular del tipo de variables. Comentare-

mos, antes de enumerar algunas, ciertos aspectos de interés.

La elección de las variables utilizadas para describir a cada individuo que será clasificado, constituye un marco de referencia que tendrá una importancia capital en los futuros clusters. Es evidente que esta elección es la primera decisión en la que el juicio del investigador refleja los aspectos que para él serán relevantes en la clasificación. En consecuencia hay varias cuestiones que surgen al realizar esta elección :

1. ¿Son relevantes las variables elegidas para el tipo de clasificación que se desea realizar?

2. ¿Cuántas variables deben medirse en cada individuo?

No existe una base teórica que tenga una respuesta objetiva a estas preguntas. Mas bien deben ser argumentos de tipo empíricos y, a veces subjetivos, los que orienten sobre cuántas y cuáles deben ser las variables utilizadas. De Sarbo et al. [31] ha estudiado que es mejor no utilizar variables que no distingan a los clusters, que dejar fuera de la elección alguna que podría ser esclarecedora de la situación. Por último, también influirán en esta elección los tradicionalmente llamados "missing values". Este tipo de Análisis no permite reemplazarlos por valores estimados como en otras técnicas multivariantes, debido al desconocimiento que se tiene a priori de los valores medios de los posibles grupos a formar. Por ello habrá que ir seleccionando en un principio aquellos objetos y aquellas variables que no presenten muchos valores de este tipo. De todas formas, en determinados casos, Dixon [33] y Little and Rubin [92] han desarrollado algunos métodos de estimación que pueden utilizarse.

Una vez que las variables han sido elegidas, podemos encontrarnos con el problema de que no utilicen las mismas unidades, e incluso, que sean de tipos diferentes: categóricas, ordinales o cuantitativas. Para este último tipo de variables, la solución que suele adoptarse es la de estandarizar los datos, utilizando los valores medios y las desviaciones típicas de las variables calculadas sobre el conjunto completo de datos. A pesar de los problemas que Fleiss y Zubin [48] o Duda y Hart [35] apuntan que pueden existir con esta práctica, los intentos de resolver el problema utilizando las desviaciones típi-

cas de las variables dentro de los grupos, que en principio son desconocidos, plantean todavía más problemas.

Cuando las variables son de otros tipos o de tipos diferentes, se han dado varias sugerencias para utilizarlas. La solución más simple es convertirlas todas a variables binarias, pero esto conlleva una gran pérdida de información. Una alternativa más razonable es utilizar un coeficiente de similaridad (Gower [58]), que sea capaz de incorporar la información de diferentes tipos de variables. Otra posibilidad es analizar los objetos separadamente con grupos de variables del mismo tipo, y después intentar sintetizar los resultados de los diferentes estudios.

El siguiente paso, una vez resuelta la elección de las variables y su estandarización e importancia en el conjunto de objetos a clasificar, es la definición de una medida que represente la fuerte o débil relación que exista entre cada pareja de objetos, dados los valores en ambos de las p variables elegidas.

Se define una similaridad entre dos objetos $I_i$ e $I_k$ a alguna función de sus valores observados

$$s_{ik} = f(\mathbf{x_i}, \mathbf{x_k})$$

donde $\mathbf{x}_i' = [x_{i1}, x_{i2}, \ldots, x_{ip}]$ y $\mathbf{x}_k' = [x_{k1}, x_{k2}, \ldots, x_{kp}]$ son los valores que toman las variables elegidas para los dos objetos. Se han propuesto muchas funciones dependiendo, en parte, del tipo de las variables y, en parte del tipo de objetos. Listas completas de similaridades pueden encontrarse en Sneath y Sokal [116], Anderberg [1], Clifford y Stephenson [25], Cormack [28] y Gower [59].

Las similaridades son medidad a las que en general se les exige la simetría ($s_{ik} = s_{ki}$), aunque hay medidas asimétricas consideradas por Constantine y Gower [27]. La mayoría de estas medidas son no negativas y suelen estar definidas para que su valor más alto sea la unidad

$$0 \leq s_{ik} \leq 1$$

aunque algunas similaridades son correlaciones entre objetos, o cosenos entre vectores que representen a los objetos en un espacio euclídeo, y por tanto estarán entre

$$-1 \leq s_{ik} \leq 1$$

Asociada con cada similaridad que esté entre 0 y 1 hay una disimilaridad

$$d_{ik} = 1 - s_{ik}$$

que será simétrica y no negativa. El grado de similaridad entre dos objetos crecerá con $s_{ik}$ y decrecerá cuando crezca $d_{ik}$. Además es natural exigirle a esta medida que cada objeto tenga similaridad máxima consigo mismo, es decir, $s_{ii} = 1$ y $d_{ii} = 0$.

Por su parte, alguna medidas de disimilaridad tienen la propiedad de que

$$d_{ik} \leq d_{ij} + d_{jk} \; \forall i, j, k$$

en cuyo caso se les conoce como distancias métricas. Gower [59] da una lista bastante completa de ellas. Entre las más utilizadas para variables cuantitativas está la distancia Euclídea

$$d_{ik}^2 = \sum_{j=1}^{p} (x_{ij} - x_{kj})^2$$

que suele utilizarse con los datos estandarizados o, incluso teniendo en cuenta las covarianzas entre las variables en la llamada distancia de Mahalanobis

$$D_{ik}^2 = (\mathbf{x}_i - \mathbf{x}_k)' S^{-1} (\mathbf{x}_i - \mathbf{x}_k)$$

donde S es la matriz de covarianza del conjunto de datos.

### 2.2.3 Tipos y procedimientos de clasificación

Antes de utilizar alguno de los muchos procedimientos que se han dado hasta ahora para clasificar objetos, el analista debe decidir el tipo de clasificacion apropiada para sus datos. En la mayoría de los casos, la elección del procedimiento a emplear dependerá del tipo.

Generalmente, pueden distinguirse entre cuatro tipos de clasificación: **particiones, particiones difusas, clasificaciones jerárquicas, y clasificaciones cuasi-jerárquicas.**

1. **Clasificaciones disjuntas o particiones**, en las que el conjunto de n objetos $\{I_1, I_2, \ldots, I_n\}$ se divide en un número de subconjuntos disjuntos y todo elemento del conjunto está dentro de algún subconjunto. A veces puede ocurrir que algunos objetos permanezcan aislados sin estar en ningún grupo y formando cada uno de ellos uno, llamándose en ese caso una **clasificación no exhaustiva.** Obviamente. los diferentes grupos son más homogéneos cuanto más alto es su número y la separación entre ellos será mayor.

   Los procedimientos que se utilizan para este tipo de clasificaciones se basan en la optimización de criterios que detecten tales grupos, aunque generalmente necesitan conocer el número de clusters. Ésta es una de las principales dificultades al aplicarlos, de forma que en los análisis cluster de tipo exploratorio el investigador tendrá raramente bastante información para determinar el número apropiado de grupos.

2. **Clasificaciones difusas o no disjuntas**, que se dan en aquellos problemas en los que la naturaleza de los objetos permite clasificarlos en más de un grupo a la vez. En ese caso, los clusters se solapan y forman un recubrimiento no disjunto del conjunto de objetos.

   En este tipo de clasificaciones suele utilizarse la teoría de los conjuntos difusos como indica Bezdek [10] y [11].

   Dentro de este tipo de clasificaciones pueden enmarcarse los métodos recomendados por Anderberg [1] para encontrar clusters que se solapan en muestras de datos continuos, basadas en el análisis de la densidad puntual y en el análisis discriminante. En ellos se estiman los parámetros y el número de densidades diferentes que están mezcladas dentro del conjunto de datos.

   Por último, Jardine y Sibson [72],[73] y [74] utilizan teoría de grafos para construir este tipo de grupos.

3. **Clasificaciones jerárquicas**, en las que los objetos y los grupos pueden representarse gráficamente en forma de árbol genético o **dendograma**. Todos los grupos van surgiendo por la fusión de subgrupos que están por debajo en el árbol, o por la división de grupos que están por encima. En el extremo inferior de esta jerarquía están los diferentes objetos, cada uno en una clase, mientras que en el extremo superior

hay un grupo con todos los objetos. Si introducimos este dendograma en un sistema de coordenadas donde el eje OY represente una medida apropiada de homogeneidad dentro de los grupos, tendremos una jerarquía en la que en cada nivel se tendrá una colección de grupos disjuntos de la misma homogeneidad. Un desarrollo más completo puede encontrarse en múltiples referencias, entre las que destacan las de Sokal y Rohlf [118], Ward [129], Hartigan [62], Johnson [79], Lance y Willians [88], Jardine y Sibson [72], Wishart [131] y Calinsky y Harabasz [17].

Se hablará de un procedimiento de cluster **aglomerativo** si los grupos en la jerarquía se van obteniendo por la unión de subgrupos. Por el contrario, los procedimientos que van dividiendo paso a paso los grupos (comenzando por el grupo completo), se denominan **divisivos**

Los dendogramas pueden también construirse utilizando métodos para generar particiones, eligiendo un criterio de clasificación por el que, por ejemplo, cada uno de los grupos formados tenga un cierto grado de homogeneidad. Disminuyendo este criterio paso a paso se va formando la jerarquía.

Murtagh [104],[105] comenta los últimos resultados en algoritmos de clasificación jerárquicos y dendogramas y,además, obtiene distribuciones de probabilidad de **dendogramas aleatorios** que pueden ser utilizados como modelos de probabilidad para contrastar la aleatoriedad de las estructuras jerárquicas.

4. **Clasificaciones cuasi-jerárquicas**, que tienen las propiedades de las clasificaciones jerárquicas con la excepción de que los clusters de cada nivel pueden solaparse.

En general, tanto los procedimientos que obtienen grupos disjuntos como solapados de denominan **horizontales**, mientras que los procedimientos jerárquicos o cuasi-jerárquicos son **verticales**. En realidad, como un dendograma está formado por diferentes niveles, cada uno de los cuales es una clasificación disjunta o no, pueden valer muchas de las propiedades de los procedimientos horizontales para cada nivel.

Por último, es muy importante que el analista tenga una herramienta que le permita evaluar la clasificación, un **criterio de clasificación** que sea

tan natural y orientado al problema como sea posible. En una partición, los grupos deben ser **homogeneos** y estar **bien separados** mutuamente. Una medida de homogeneidad que a menudo se utiliza para los clusters $C_k$ que tengan al menos dos objetos es la media de todas las proximidades o distancias en $C_k$

$$hom(C_k) = \frac{1}{2n_k(n_k-1)} \sum_{I_i \in C_k} \sum_{I_j \in C_k} d_{ij}$$

donde $n_k$ es el número de objetos en el cluster k. Cuando se utiliza la distancia Euclídea como medida, otra forma de medir la homogeneidad es la varianza dentro de los grupos, obtenida con sólo sustituir las distancias $d_{ij}$ en la expresión anterior por $d_{ij}^2$. En ocasiones, también se puede medir la homogeneidad de un grupo con al menos dos objetos por los índices

$$hom(C_k) = \min_{I_i, I_j \in C_k} d_{ij}$$

$$hom(C_k) = \max_{I_i, I_j \in C_k} d_{Cj}$$

El valor de un índice de homogeneidad debe ser pequeño cuando la homogeneidad sea alta. Por ello, el índice de homogeneidad de los clusters con un sólo elemento es 0.

Por otra parte, la separación entre dos clusters $C_k$ y $C_l$ puede medirse por la media de las distancias de todas las parejas de objetos, uno de $C_k$ y otro de $C_l$

$$sep(C_k, C_l) = \frac{1}{n_k n_l} \sum_{I_i \in C_k} \sum_{I_j \in C_l} d_{ij}$$

Otras veces la separación se calcula utilizando alguno de los índices

$$sep(C_k, C_l) = \min_{I_i \in C_k, I_j \in C_l} d_{ij}$$

$$sep(C_k, C_l) = \max_{I_i \in C_k, I_j \in C_l} d_{ij}$$

Un índice de separación alto dará buenas separaciones. La separación de una clase $C_k$ del resto de la muestra puede definirse simplemente como

$$sep(C_k) = sep(C_k, S - C_k)$$

Es fácil ver que de las definiciones expuestas, hay algunas más fuertes para los grupos grandes, mientras que otras son más débiles y pequeños valores pueden indicar que no hay homogeneidad en los clusters.

Por último, una medida de la calidad de un cluster es la razón

$$b(C_k) = \frac{hom(C_k)}{sep(C_k)}$$

denominado **coeficiente B**, que se utiliza mucho en los métodos de optimización de análisis cluster tratados más extensamente en 2.4.4

El Análisis Cluster está constituido por una colección de muchas técnicas diversas cuyo fin es descubrir la estructura de los datos, algunas de ellas deducidas de hipótesis matemáticas realizadas sobre la muestra y otras basadas sólo en razonamientos heurísticos.

Un problema importante que está presente en todos los tipos de clasificación, y que en la mayoría de los casos hay que resolverlo de forma subjetiva, es el de la determinación del número correcto de grupos. No hay más que inspeccionar la cantidad de algoritmos diferentes que en la actualidad se ofrecen como solución, para darse cuenta de que no hay un procedimiento que nos proporcione una solución óptima en todos los casos y, por tanto, con cada procedimiento siempre se podrán encontrar ejemplos de datos que no queden bien clasificados.

Por otra parte, los algoritmos que están especialmente diseñados para analizar muestras donde existen grupos esféricos y compactos no suelen dar buenos resultados cuando los clusters son alargados y curvados. Otros procedimientos detectan bien la existencia de los grupos pequeños frente a los grandes, o la existencia de grupos más o menos esféricos y compactos, pero en el momento en que estas configuraciones fallan en alguna/s de sus características, no funcionan.

En general, todo investigador que vaya a utilizar técnicas de Análisis Cluster, necesita un cierto aprendizaje profesional en el que se familiarice con las propiedades y posibilidades de los algoritmos que puede aplicar a sus datos. Él debe conocer con detalle las posibilidades de preparación de sus datos (datos faltantes, estandarización, etc), y debe elegir la medida de

proximidad más apropiada para sus datos. La mayoría de los programas que necesita para realizar los análisis no los va a encontrar normalmente en los principales paquetes estadísticos al uso (BMDP, SPSS, SAS), por lo que tiene que acudir a pequeñas subrutinas que están disponibles o a confeccionarse sus propios programas.

Independientemente del tipo de clasificación que se pretenda, los procedimientos del Análisis Cluster pueden clasificarse en 4 grandes categorías:

1. **Métodos de Optimización**, que utilizan un criterio de clasificación que es optimizado sobre todas las posibles clasificaciones que pueden hacerse. Generalmente llevan aparejado un algoritmo que evalúa el criterio y obtiene la solución óptima o cuasi-óptima.

2. **Métodos recursivos de construcción de grupos alrededor de centros**, que comienzan con un elemento arbitrario $I_i$ del conjunto (llamado también núcleo) para ir construyendo un grupo alrededor de él de forma recursiva. El proceso se para cuando el grupo comienza a ser heterogeneo y se comienza de nuevo el proceso con otro elemento de los que se encuentran alejados del primero. Este procedimiento se repite hasta que todos los objetos son clasificados.

3. **Métodos basados en el análisis de la densidad puntual**, en los que los clusters se relacionan con áreas de puntos de alta densidad, llamados también modas. Además, bajo condiciones ideales, los clusters son separados por áreas de baja densidad, llamados "valles" de la mezcla de distribuciones.

4. **Métodos de unión/división**, diseñados principalmente para obtener los **clusters jerárquicos**. En ellos, el proceso se realiza secuencialmente de forma que en cada paso sólo un objeto o grupo de objetos cambia de grupo, y los grupos en cada paso se anidan con respecto a grupos previos. Así, una vez que un objeto se asigna a un grupo, nunca deja de pertenecer a él en todo el proceso del análisis. Pueden utilizarse procedimientos **aglomerativos** frente a otros **divisivos**, según que el procedimiento de formación de los grupos comience con n grupos, uno por objeto, y termine en un sólo grupo con todos los objetos, o que el procedimiento siga el orden inverso.

Para los procedimientos de la tercera clase, el punto de partida es la matriz de datos X de los datos originales o la matriz Z de los datos estandarizados. Para los demás, el punto de partida es una matriz de proximidades o similaridades definida entre todos los pares de objetos.

Puede añadirse una última clase de procedimientos denominada "**otros métodos**". El análisis de correspondencia, por ejemplo, proporciona la determinación de clusters para datos cualitativos (Benzécri [9]). O como se ha mencionado en el capítulo primero, el Análisis Factorial en modo Q puede utilizarse también en determinadas situaciones (Overall y Klett [106]). Podemos también citar aquí los métodos derivados de la teoría de los conjuntos difusos (Bezdek [10],[11]; Gitman y Levine [53]; Ruspini [112]) y a los métodos que están basados en la teoría de la decisión.

En la sección siguiente nos centraremos en las caracteristicas de los procedimientos de optimización, ya que algunos de sus resultados nos serán de utilidad en el procedimiento que propondremos en la sección **2.3**.

### 2.2.4 Métodos de optimización de Análisis Cluster

Las técnicas de optimización para conseguir clusters están basadas en minimizar o maximizar algún criterio numérico relacionado con las posibles particiones que puedan llevarse a cabo en un conjunto de objetos. La principal diferencia respecto de las técnicas jerárquicas es que con ellas no necesariamente se forman clasificaciones jerárquicas.

Son muchos los criterios numéricos de clasificación y los algoritmos de optimización que pueden utilizarse en este tipo de métodos, pero todos tratan de responder a las dos preguntas básicas que, una vez que se han elegido las variables, quiere responder cualquier analista que busca agrupaciones naturales en un conjunto de datos:

1. Si existen tales agrupaciones naturales, ¿cuántos grupos hay? ó ¿cuál es el mejor número de grupos $g$?

2. ¿dónde deben hacerse las divisiones? ó dado un número de grupos fijo

$g$, ¿cuál es la mejor subdivisión del conjunto de datos en ese número de grupos?

### Criterios de clasificación

La idea básica en la que se apoyan estos criterios es que asociada con cada posible partición de los $n$ objetos en $g$ grupos, puede definirse un coeficiente f$(n, g)$, que indica la "calidad" de tal partición. Definiendo tales índices de forma que el máximo (ó el mínimo) lo alcancen cuando la partición refleje la situación más clara de agrupamiento, se puede buscar tal partición y por tanto los grupos.

Se han sugerido muchos criterios para responder a las preguntas anteriores, pero los más utilizados son los que surgen al considerar las tres matrices de dispersión que pueden calcularse al realizar cualquier partición de un conjunto de $n$ datos en $g$ grupos:

$$\mathbf{T} = \frac{1}{n}\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \overline{\mathbf{x}})(\mathbf{x}_{ij} - \overline{\mathbf{x}})'$$

$$\mathbf{W} = \frac{1}{n-g}\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)(\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)' \tag{2.1}$$

$$\mathbf{B} = \sum_{i=1}^{g} n_i(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})'$$

donde las matrices $p \cdot p$ (con $p$ el número de variables) representan, respectivamente, la dispersión total, dentro de los grupos y entre los grupos, satisfaciendo la ecuación

$$\mathbf{T} = \mathbf{W} + \mathbf{B} \tag{2.2}$$

Para $p = 1$, esta ecuación representa una relación entre escalares, utilizada en el Análisis de la Varianza. En este caso, un criterio natural de agrupamiento podría ser elegir la partición correspondiente al valor mínimo de la suma de cuadrados dentro de los grupos o, lo que es equivalente, al valor máximo de la suma entre grupos.

Para $p > 1$, el criterio de clasificación para la ecuación anterior no está clara y se han sugerido varias alternativas.

1. **Minimización de la traza de W**

   Una extensión natural de la minimización de la suma de cuadrados dentro de los grupos es minimizar la suma de las sumas de cuadrados dentro de los grupos para todas las variables, es decir, minimizar tr(**W**).

   Puede demostrarse que este criterio es equivalente a minimizar la suma de distancias euclídeas cuadradas entre los individuos de cada grupo y su media, es decir

   $$E = \sum_{i=1}^{n} d_{i,c(i)}^2 \tag{2.3}$$

   donde $d_{i,c(i)}$ es la distancia euclídea entre el objeto $I_i$ y la media del grupo donde ha sido asignado.

   El primero que lo sugirió fue Singleton y Kautz [115], y está implícito en los métodos descritos por Forgey [50], Jancey [71], MacQueen [94] y Ball y Hall [6].

2. **Minimización del determinante de W**

   Uno de los tests para contrastar las diferencias entre grupos dentro del Análisis Multivariante de la Varianza, está basado en la razón de los determinantes entre las matrices de dispersión dentro de los grupos y total (Krzanowski [87]). Valores altos del cociente

   $$\frac{det(\mathbf{T})}{det(\mathbf{W})}$$

   indican que los vectores medios de cada grupo son diferentes. Con esta base, Friedman y Rubin [51] sugieren como criterio de clasificación maximizar este cociente. Si se tiene además en cuenta que para todas las posibles particiones de $n$ objetos en $g$ grupos, la matriz **T** permanece constante, maximizar el cociente anterior es equivalente a minimizar det(**W**). El criterio ha sido estudiado con más detalle por Marriott [98],[99].

3. **Maximización de la traza de $\mathbf{BW}^{-1}$**

   Otro criterio sugerido por Friedman y Rubin [51] es la maximización de la traza de la matriz obtenida multiplicando la matriz **B** y la inversa de **W**. Esta función es utilizada también en el contexto del Análisis Multivariante de la Varianza, y es equivalente a la que Rao [108] llama la generalización de la distancia de Mahalanobis [96] para más de dos grupos.

   Tanto $\mathrm{tr}(\mathbf{BW}^{-1})$ como $det(\mathbf{T})/det(\mathbf{W})$ pueden expresarse en términos de los autovalores $\lambda_i$ de $\mathbf{BW}^{-1}$, siendo

$$tr(\mathbf{BW}^{-1}) = \sum_{i=1}^{p} \lambda_i$$

$$\frac{det(\mathbf{T})}{det(\mathbf{W})} = \prod_{i=1} p(1+\lambda_i)$$

Otros criterios de optimización, no basados en funciones de las matrices de dispersión, se describen en Rubin [110] y Wallace y Boulton [128], esencialmente para datos con variables continuas. Spath [119] describe otros criterios para variables binarias y ordinales.

### Optimización del criterio de clasificación

Una vez que el criterio de clasificación se selecciona, hay que tratar de dar solución a las dos cuestiones planteadas en la introducción sobre el número de grupos y la mejor partición.

Suponiendo que se tuviera por algún medio el número de grupos $g$ que pueden formarse de forma natural en un conjunto de datos, se plantea un problema al optimizar el criterio elegido dado el elevado número de posibles particiones que pueden formarse. Hay que tener en cuenta que este número depende de $n$ y $g$, y es muy grande incluso para valores pequeños de $n$ y $g$. Viene dado por

$$\frac{1}{g!}\sum_{i=1}^{g}\binom{g}{i}(-1)^{g-i}i^n$$

que es del orden de $g^n/g!$ cuando $n$ es grande, de forma que evaluar todas las posibles particiones es prohibitivo en cuanto $n$ sea un poco grande.

En la práctica, la imposibilidad de poder examinar todas las posibles particiones ha traido el desarrollo de algoritmos que, partiendo de una dada, investigan el valor óptimo del criterio reordenando particiones a partir de la inicial y quedándose con las nuevas sólo en el caso de que suponga una mejora en la optimización. A este tipo de algoritmos se les denomina "hill-climbing". No obstante, es necesario afinar bien en las soluciones iniciales, y en general utilizar varias configuraciones iniciales para asegurar la validez de la solución final, ya que el resultado final del método puede depender en algunos casos de la partición inicial. Para más detalles sobre estos algoritmos puede verse MacQueen [94], Beale [7] y [8], Thorndike [124], McRae [95], Friedman y Rubin [51] y Blashfield [12]. Hartigan [63] da algunos resultados empíricos y analíticos sobre la conexión entre los posibles óptimos locales y el óptimo global y Marriott [99] sugiere que una convergencia débil y la obtención de agrupamientos diferentes cuando se parte de soluciones iniciales diferentes, indica que $g$ está elegido erróneamente y, en particular, que no hay evidencia de clusters naturales.

Por otra parte, Marriott [99] también proporciona algunos resultados muy interesantes sobre los cambios que producen en algunos de los criterios de clasificación el hecho de que un individuo se añada a un grupo ya formado. Muestra que, por ejemplo, añadir un individuo, I, al grupo $s$ con media $\overline{x}_s$ y de tamaño $n_s$, cambia $\mathbf{W}_s$ a $\mathbf{W}_s^* = \mathbf{W}_s + \mathbf{d}_s\mathbf{d}_s'$, donde

$$\mathbf{d}_s = (\mathbf{I} - \overline{x}_s)\left(\frac{n_s}{n_s+1}\right)^{\frac{1}{2}}$$

y en consecuencia

$$tr(\mathbf{W}_s^*) = tr(\mathbf{W}_s) + \mathbf{d}_s'\mathbf{d}_s$$
$$det(\mathbf{W}_s^*) = det(\mathbf{W}_s)(1 + \mathbf{d}_s'\mathbf{W}_s^{-1}\mathbf{d}_s)$$

Además de las referencias citadas que tratan sobre los algoritmos "hill-climbing", existen otras que introducen otros algoritmos o formas de abordar el problema. entre ellas, Jensen y al. [76], da detalles de un algoritmo de programación dinámica, Gordon y Henderson [55] dan un algoritmo basado en la técnica de paso descendente y Koontz y al. [85] utilizan un algoritmo

que investiga el óptimo global a traves de cotas. Por último, una alternativa interesante al problema de investigar la partición que minimiza la tr(W) la proporciona Calinsky y Harabasz [17], que utilizan el árbol de mínima expansión partiendo de la matriz de distancias euclídeas entre cada pareja de objetos.

No terminaremos esta sección sin apuntar algunas de las propiedades y los problemas que plantean los criterios expuestos.

Uno de los principales problemas del criterio tr(**W**) es que depende de las unidades de medida. Pueden obtenerse diferentes soluciones según se utilicen los datos originales o se estandaricen. Este fue quizás el principal motivo para que se introdujera el criterio det(**W**), que no depende de la escala en que se midan las variables.

Otro problema de la tr(**W**) es que impone una estructura esférica para los grupos que forma, incluso aunque éstos se presenten en los datos con otras formas. Sin embargo, nuevamente el criterio det(**W**) no restringe los grupos a que sean esféricos, aunque sí tiene el problema de suponer que todos los grupos tienen la misma forma, causando problemas cuando no ocurre esto.

En un intento de solucionar el problema de las formas, Scott y Symons [114] han sugerido cambiar det(**W**) por

$$\prod_{i=1}^{g} det(\mathbf{W}_i)^{n_i}$$

aunque aqui es necesaria la restricción de que cada cluster tenga al menos $p+1$ elementos para que los determinantes sean todos no nulos. Por su parte, Maronna y Jacovkis [97] han sugerido minimizar

$$\sum_{i=1}^{g}(n_i - 1)det(\mathbf{W}_i)^{\frac{1}{p}}$$

y Symons [123] sugiere dos nuevos criterios

$$n \ln det(\mathbf{W}) - 2\sum n_i \ln n_i$$

$$\sum(n_i \ln det(\mathbf{W}_i) - 2n_i \ln n_i)$$

### La selección del número de grupos g

Además de criterios subjetivos que puedan ser ayudados por otras técnicas de Análisis Multivariante, se han sugerido una gran variedad de métodos que pueden ayudar para determinarlo, la mayoría útiles en situaciones particulares.

El más general y razonable, aunque prohibitivo en los cálculos, es calcular el valor que optimiza el criterio de clasificación elegido para cada número de grupos $g$, representando este valor frente a $g$. Está claro que grandes cambios de nivel en la gráfica deben tomarse, generalmente, como sugerencias sobre el número de grupos. Tiene el defecto de ser un procedimiento subjetivo, pues por lo general no nos encontraremos con situaciones de una claridad objetiva en cuanto a estos cambios de nivel.

Beale [8], describe un contraste utilizando la F de Fisher, que puede utilizarse para analizar si una subdivisión en $g_2$ clusters es significativamente mejor que una subdivisión en un número menor de clusters $g_1$. El estadístico que utiliza es

$$F(g_1, g_2) = \frac{R_{g_1} - R_{g_2}}{R_{g_2}} / [\frac{n - g_1}{n - g_2}(\frac{g_2}{g_1})^{\frac{2}{p}} - 1]$$

donde $R_g = (n-g)S_g^2$ y $S_g^2$ es la desviación media cuadrática desde los centros de los clusters en la muestra. Esta cantidad se compara con la F de Fisher, con $p(g_2 - g_1)$ y $p(n - g_2)$ grados de libertad, de forma que un resultado significativo indica que una subdivisión en $g_2$ clusters es mejor que en $g_1$. Sin embargo, la experiencia con este procedimiento revela que sólo es útil cuando los clusters están bien separados y aproximadamente con forma esférica.

Otro criterio sugerido por Calinsky y Harabasz [17] es tomar el valor de $g$ que corresponde al máximo valor de C, donde

$$C = \frac{tr(\mathbf{B})}{g-1} / \frac{tr(\mathbf{W})}{n-g}$$

que es capaz de obtener razonablemente bien el número de grupos según estudiaron Milligan y Cooper [101].

Por último, Marriott [98] y Krzanowski y Lai [87] utilizan argumentos heurísticos para determinar el número de grupos $g$. El primero toma el valor de $g$ para el que $g^2 det(\mathbf{W})$ es mínimo, y los segundos obtienen el valor que hace máximo $C_g$, con

$$C_g = |DIFF(g)/DIFF(g+1)|$$

siendo

$$DIFF(g) = (g-1)^{\frac{2}{p}} \tilde{S}_{(g-1)} - g^{\frac{2}{p}} \tilde{S}_{(g)}$$

con $\tilde{S}_{(g)}$ el valor óptimo de la función $\mathrm{tr}(\mathbf{W})$ para un valor de $g$ dado.

## 2.3 El procedimiento PROCED para clasificar elementos de un conjunto. Ejemplo

Como se ha visto en la sección anterior, son muchos los procedimientos que pueden utilizarse para obtener clasificaciones de un conjunto de elementos en el que se sospeche que existen grupos homogéneos y, si bien es cierto que unos funcionan mejor que otros dependiendo de la naturaleza de los datos, no es menos cierto que no existe el método ideal de clasificación, pues es fácil buscar para cualquiera de ellos un ejemplo que no quede bien clasificado.

Con esta salvedad presente, en esta sección queremos proponer un procedimiento de clasificación basado en el modelo factorial para los elementos de un conjunto, expuesto al final del capítulo anterior. En concreto, allí se dijo que la **DVS** permite utilizar la expresión (1.35)

$$\mathbf{X} \approx \mathbf{V}_q \mathbf{\Lambda}_q^{\frac{1}{2}} \mathbf{U}_q' = \mathbf{A}\mathbf{G}'$$

para formular un modelo en el que encontrar un conjunto de direcciones, dadas por las columnas de **G**, entre los individuos que permitan reconstruir la matriz **X** y emplear, posteriormente, procedimientos de rotación clásicos dentro del Análisis Factorial, que permitan orientar tales direcciones hacia los centros de los distintos grupos y así poder determinar los posibles clusters naturales que hubiera en él.

En general, este procedimiento se puede enmarcar dentro de los procedimientos de optimización del Análisis Cluster, pues si bien en él no se va a utilizar la descomposición (2.2) de la matriz de dispersión **T** vista en la sección anterior para optimizar algunos de los criterios dados sobre las matrices **W** y/ó **B** de dispersión dentro de los grupos y entre los grupos, sin embargo el objetivo va a ser minimizar las distancias que existan entre los elementos de cada grupo determinado por cada factor y el propio factor que lo determina, de forma que el papel que juega el vector de valores medios dentro de cada grupo es el que tendría aquí el vector-factor que determina al grupo. Así, si la matriz de distancias dentro de los grupos, referidas a los

vectores-factores que los determinan es, salvo constantes,

$$\mathbf{W}^* = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \mathbf{g}_i)(\mathbf{x}_{ij} - \mathbf{g}_i)'$$

donde $\mathbf{x}_{ij}$ representa el elemento $j$ del grupo $i$ y $\mathbf{g}_i$ el factor que determina al grupo $i$, el criterio de minimizar la traza de $\mathbf{W}^*$ sería equivalente a minimizar

$$E^* = \sum_{i=1}^{n} d^2_{i,\mathbf{g}(i)}$$

donde $d_{i,\mathbf{g}(i)}$ es la distancia euclídea entre el elemento $I_i$ y el factor $g$ que determina el grupo donde está $I_i$.

Con el procedimiento que se propone, pretendemos tener una forma de determinar, al mismo tiempo, tanto el número de grupos (suponiendo que existan más de uno) como la mejor asignación de cada elemento a uno de los grupos, tratando así de resolver en un sólo procedimiento las dos cuestiones principales con las que se iniciaba la sección (**2.2.4**). Debemos, no obstante, imponer una condición restrictiva al número de grupos: dado que el número máximo de direcciones que obtendremos no puede exceder de $p+1$, limitaremos a esta cantidad también el número de grupos. Posteriormente, en el Capítulo 3, trataremos de abordar el problema sin que exista tal restricción.

Sokal y Michener [117], Overall y Klett [106] y Miyazaki y Seki [103] son las referencias más cercanas encontradas sobre algo similar, en ciertos aspectos, respecto de lo que proponemos. En las dos primeras, se utiliza el Análisis Factorial sobre la matriz de correlaciones entre individuos a los que se han aplicado una serie de tests psicológicos, buscando que los factores obtenidos se puedan identificar con patrones o individuos tipo. En la sección **1.3** se expusieron algunos de los problemas que podía acarrear el uso del coeficiente de correlación entre individuos o de la generalización de los resultados que puedan obtenerse con una muestra a una población en general. Por su parte, Miyazaki y Seki [103], aunque obtienen en su artículo un procedimiento de clasificación mediante el Análisis de Componentes Principales (al que denominan Clusters Principales), sólo utilizan éstas para ir obteniendo un grupo por cada Componente Principal del conjunto de datos que sea representativa en cuanto a la variabilidad que es capaz de explicar. Cada grupo se forma

tomando aquellos elementos más alejados a cada Componente, hasta que las que quedan son poco representativas y los elementos más alejados a ellas forman un último grupo. Este procedimiento no asegura que la partición obtenida sea óptima para algunos de los criterios objetivos que suelen utilizarse dentro del Análisis Cluster.

En nuestro caso, aunque proponemos un procedimiento que se basa en algunas de las técnicas utilizadas en las referencias citadas, sin embargo pensamos que las propiedades geométricas de las transformaciones que utilizamos nos permiten evitar gran parte de los problemas encontrados en ellas.

### 2.3.1 Preparación de los datos. Coeficiente de similaridad

Las transformaciones a las que nos referimos comienzan sobre la matriz $\mathbf{X}$ inicial, antes de obtener las direcciones de la descomposición (1.35).

Si, como hemos venido haciendo hasta ahora, consideramos cada fila de $\mathbf{X}$ como un punto de $I\!R^p$, lo primero que haremos con tal nube de puntos será centrarla respecto a las columnas. Esta operación va a permitirnos situar a la nube de puntos en torno al origen de coordenadas de $I\!R^p$, como se vió en la Figura 1.7, lo cual será de especial importancia en lo que sigue.

Una vez centrados los datos en $\mathbf{C}$, pueden adoptarse varias alternativas sobre el coeficiente de similaridad entre los elementos que se va a emplear o, lo que es lo mismo, sobre cómo transformar $\mathbf{C}$ en otra matriz $\mathbf{W}$ que sirva como base para entraer las direcciones buscadas con (1.35). Está claro que utilizar los valores medios y desviaciones típicas de las filas de $\mathbf{C}$ para definir coeficientes de correlación entre los individuos no es una solución aceptable desde el punto de vista estadístico. En su lugar, nos parece que la solución que adoptan Imbrie y Purdi [66] al dividir los elementos de cada fila por su módulo es, desde el punto de vista geométrico aceptable y, al menos, no rechazable desde el punto de vista estadístico. Existe una tercera solución que sería utilizar los vectores fila con las longitudes originales, en cuyo caso los coeficientes de similaridad dados por los elementos de $\mathbf{C}\mathbf{C}'$ dependerían no sólo del ángulo que formen los vectores fila, sino también del producto de

sus longitudes como se vió en la sección **(1.3)**.

Se tiene aquí, por tanto, la posibilidad de realizar dos descomposiciones en valores singulares de la matriz resultante, una con los vectores fila de **C** sin normalizar y otra normalizándolos, al igual que para el caso de las variables puede realizarse una doble descomposición con las variables centradas -que conlleva la descomposición de la matriz de varianzas y covarianzas- o con las variables tipificadas -que conlleva la descomposición de la matriz de correlaciones-.

Es conocido que en el caso de las variables (Jackson [69]), no hay una correspondencia uno a uno entre las direcciones obtenidas mediante la matriz de correlaciones y las obtenidas a partir de la matriz de covarianzas. En el dilema de utilizar la matriz de correlaciones en lugar de la de covarianzas, se apuntan razones sobre si las variables originales están medidas en diferentes unidades e, incluso, aunque sean las mismas unidades, el hecho de que tengan diferentes varianzas da un peso excesivo a algunas de ellas respecto de las otras. Al trasladar estos razonamientos a las direcciones que pretendemos encontrar entre los individuos, parece aconsejable que normalizemos la matriz **C** por filas, para evitar los mismos problemas de que pesen excesivamente en la descomposición los individuos con mayor longitud. Es por ello que adoptaremos la normalización como paso previo a la descomposición, sin descartar que puede haber casos en los que también pueda ir bien para nuestros propósitos la descomposición sin normalizar.

Además, antes de normalizar por filas, realizaremos una transformación geométrica después de obtener la matriz **C**. La idea sobre la que trabajaremos es la siguiente: si en los datos existen grupos homogeneos, y utilizamos la métrica euclidea para determinar la proximidad entre los puntos, una buena forma de detectarlos sería que pudiéramos "verlos" desde una perspectiva apropiada que nos permitiera descubrir tales grupos. Pueden encontrarse ejemplos de datos en los que, existiendo grupos entre ellos, un análisis factorial en modo Q clásico ó por el procedimiento expuesto antes, no detectaría tales grupos. Simplemente el hecho de que exista más de un grupo en alguna de las direcciones de máxima variabilidad como ocurría en la Figura 1.5, hará que el factor que determine tal dirección aglutine a todos los grupos que se encuentran en dicha dirección.

Lo que ocurre en estos casos es que los coeficientes de similaridad de objetos que estuvieran en distintos grupos podrían tener valores próximos a 1 al estar en una misma dirección, lo cual facilitaría que los objetos de estos grupos se asocien al mismo factor, y esto no es deseable si queremos obtener una clasificación que refleje correctamente la composición del conjunto de elementos.

Tenemos, pues, que evitar estos casos, buscando que los vectores unitarios que representan a los objetos formen ángulos próximos a 0º sólo cuando éstos pertenezcan al mismo grupo. Esta idea intuitiva puede concretarse realizando las siguientes operaciones:

Primero, incluimos la nube de puntos $\mathbf{C}$ de $I\!R^p$ en un espacio de una dimensión más, $I\!R^{p+1}$, sin más que añadir una columna de ceros a la matriz $\mathbf{C}_{n,p}$

$$\mathbf{C}^* = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1p} & 0 \\ c_{21} & c_{22} & \cdots & c_{2p} & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ c_{n1} & c_{n2} & \cdots & c_{np} & 0 \end{pmatrix},$$

que pasará a ser de dimensión $n \cdot (p+1)$. Tendremos con esto que en $I\!R^{p+1}$, cada individuo tendrá unas coordenadas en las $p$ primeras direcciones dadas por las filas de $\mathbf{C_{n,p}}$, y en la dirección $p+1$ tendrá coordenada 0.

Podemos ahora, en segundo lugar, hacer una traslación (Figura 2.1) en $I\!R^{p+1}$ de cada punto mediante el vector $\mathbf{t} = (0, \ldots, 0, \lambda)$ del espacio $I\!R^{p+1}$, es decir, se desplazarán a través de la dimensión $p+1$ una longitud $\lambda$, de forma que este desplazamiento permita "ver" desde el origen los grupos homogeneos con mayor claridad.

$$\mathbf{T} = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1p} & \lambda \\ c_{21} & c_{22} & \cdots & c_{2p} & \lambda \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ c_{n1} & c_{n2} & \cdots & c_{np} & \lambda \end{pmatrix}$$

El siguiente paso sería aplicar el modelo Q-factorial a la matriz $\mathbf{T}$, es decir, se pretende encontrar un grupo de $q$ direcciones que nos permitan expresar

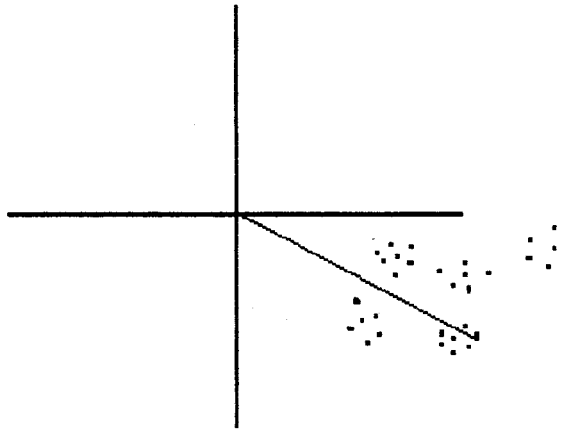

Figura 2.1: Traslación en $I\!R^{p+1}$ de cada punto mediante el vector $\mathrm{t} = (0, \ldots, 0, \lambda)$.

a los $n$ vectores fila en combinación lineal de ellas, de forma que si exiten grupos naturales, éstos puedan ser determinados mediante asociaciones altas con las direcciones encontradas (en sentido positivo o negativo).

Para ello, el método dado en la sección **1.4** obtiene primero la matriz diagonal

$$\mathbf{D} = \mathrm{diag}(\mathbf{T} \cdot \mathbf{T}')$$

cuyos elementos serán las normas euclideas al cuadrado de cada vector de $I\!R^{p+1}$ que, desde el origen, representa a cada individuo trasladado en las filas de $\mathbf{T}$. De esta forma, la matriz

$$\mathbf{W} = \mathbf{D}^{-1/2} \cdot \mathbf{T}$$

tendrá por filas vectores unitarios de $I\!R^{p+1}$ cuyas coordenadas dependerán de $\lambda$, distancia desde el origen a la que se han desplazado los datos.

En este punto, el modelo en modo Q define los coeficientes de asociación o similaridades que permiten expresar de forma numérica la cercanía o lejanía entre dos puntos de la nube.

Para ello definimos la matriz

$$\mathbf{Q} = \mathbf{W} \cdot \mathbf{W}'$$

cuyos elementos

$$q_{ik} = \cos\theta_{jk}$$

son, precisamente, los cosenos de los ángulos que forman cada pareja de vectores fila. **Q** será una matriz que dependerá de $\lambda$ al ser

$$\begin{aligned} q_{ik} &= \frac{\sum_{j=1}^{p+1} t_{ij}t_{kj}}{\sqrt{\sum_{j=1}^{p+1} t_{ij}^2}\sqrt{\sum_{j=1}^{p+1} t_{kj}^2}} \\ &= \frac{\sum_{j=1}^{p} c_{ij}c_{kj} + \lambda^2}{\sqrt{\sum_{j=1}^{p} c_{ij}^2 + \lambda^2}\sqrt{\sum_{j=1}^{p} c_{kj}^2 + \lambda^2}} \end{aligned}$$

Será esta matriz **Q** la que tomaremos como punto de partida para extraer las direcciones, pero veamos antes cómo obtener la coordenada $\lambda$ del desplazamiento en la dimensión $p+1$.

### Obtención del desplazamiento de la nube de puntos

La traslación a través de la dimensión $p+1$ de los puntos la hemos realizado con el objetivo de "ver" o "tener una perspectiva mejor" desde el origen para detectar los posibles grupos homogéneos en los datos. Es lógico que a la hora de determinar cuánto debemos desplazarnos, es decir el valor de $\lambda$, lo hagamos de forma que tal valor haga máxima la dispersión que presentan los puntos desde el origen. Anderson [2] define un concepto que vamos a utilizar para este propósito.

Consideremos p variables observadas en n unidades. Como se sabe, la matriz de varianzas y covarianzas es la que describe tanto la variación de cada variable como la conjunta de cada pareja de ellas:

$$\mathbf{S} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ s_{n1} & s_{n2} & \cdots & s_{np} \end{pmatrix}$$

con

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

Si deseamos obtener un parámetro que resuma la dispersión general de estas variables en el conjunto de estas n unidades, podemos dar la siguiente

**Definición**

*Se define el término varianza generalizada en las n unidades como el determinante de S*

$$VG = det(\mathbf{S}) = \mid \mathbf{S} \mid$$

La varianza generalizada proporciona una forma resumida de dar la información que encierran todas las varianzas y covarianzas en un sólo valor. Es evidente que cuando p = 1, este valor coincide con la varianza de la variable, y que cuando $p > 1$, en esta definición debe haber alguna pérdida de información sobre la estructura de varianzas y covarianzas de las variables. No obstante, presenta algunas propiedades interesantes. Una interpretación geométrica de $\mid \mathbf{S} \mid$ nos ayudará a introducirlas.

Si llamamos VOL al volúmen generado por los vectores formados por las columnas de la matriz de datos, previamente centrados por sus valores medios, $\mathbf{e}_1 = \mathbf{x}_1 - \overline{x}_1, \ldots, \mathbf{e}_p = \mathbf{x}_p - \overline{x}_p$ representados en la Figura 1.8, puede probarse con facilidad que

$$\mid \mathbf{S} \mid = \frac{VOL^2}{n^p}$$

lo cual indica que el volúmen depende de la varianza generalizada.

En el caso de datos estandarizados, la varianza generalizada se definirá como

$$\mid \mathbf{R} \mid = \frac{VOL^2}{n^p}$$

y el volumen dependerá también de ella.

Pero la varianza generalizada puede interpretarse también geométricamente en el espacio $I\!R^p$ donde puede representarse la nube de puntos de las unidades. La representación más intuitiva se refiere a la dispersión de éstos puntos respecto al punto medio $\overline{\mathbf{x}} = [\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_p]$. Si consideramos la distancia estadística (Mahalanobis [96]) entre dos puntos de $I\!R^p$, $\mathbf{x}'$ e $\mathbf{y}'$ como

$$d^2(\mathbf{x}, \mathbf{y}) = \mathbf{x}'\mathbf{S}^{-1}\mathbf{y}$$

puede verse fácilmente que la ecuación

$$(\mathbf{x} - \overline{x})\mathbf{S}^{-1}(\mathbf{y} - \overline{y}) = c^2$$

define un hiperelipsoide centrado en $\overline{\mathbf{x}}$ cuyo volumen está relacionado con $|\ \mathbf{S}\ |$. Concretamente

$$\text{Volumen de}\{\mathbf{x} : (\mathbf{x} - \overline{x})\mathbf{S}^{-1}(\mathbf{y} - \overline{y}) \leq c^2\} = k_p \mid S \mid c^p,$$

donde $k_p$ y $c$ son constantes, con lo que el volúmen será directamente proporcional a la varianza generalizada.

Volviendo a nuestro problema, estamos interesados en obtener la mejor forma de "ver" los datos desde el origen, proyectados en la hiperesfera de radio unidad, de forma que los grupos naturales queden al descubierto.

Un procedimiento para obtener la mejor forma de "ver" los datos puede ser maximizar el volúmen que éstos determinen en $I\!R^{p+1}$ desde el origen. Tal volúmen vendrá dado por un múltiplo de

$$\det(\mathbf{W}'\mathbf{W})$$

y, por lo tanto, dependerá de $\lambda$. En concreto, el volúmen será máximo cuando lo sea la función $f(\lambda) = \det(\mathbf{W}'\mathbf{W})$. Veremos en el siguiente resultado que este máximo se alcanza.

**Teorema 2.3.1** *Dada la matriz* $\mathbf{X}_{n,p}$, *sea la matriz centrada* $\mathbf{C}_{n,p}$, *sea* $\mathbf{T}_{n,p+1}$ *la obtenida al trasladar los datos por una dimensión más, cuya última columna tiene por elementos el parámetro* $\lambda$, *e indicamos por* $\mathbf{W}_{n,p+1}$ *la matriz normalizada por las filas. Entonces la dispersión que presentan los datos en* $I\!R^{p+1}$ *desde el origen, es directamente proporcional al* $det(\mathbf{W}'\mathbf{W})$. *Además, la función* $f(\lambda) = det(\mathbf{W}'\mathbf{W})$ *alcanza un máximo finito en* $[0, \infty]$.

DEMOSTRACIÓN

Con los datos centrados por columnas, considerados como puntos de $I\!R^p$, la varianza generalizada será la obtenida por el determinante de la matriz de covarianzas $\mathbf{S} = (1/n)\mathbf{C}'\mathbf{C}$.

Cuando estos datos se incluyen en un espacio de dimensión $p+1$, se desplazan y se normalizan por filas, el volúmen del hipercono determinado por $\det(\mathbf{W}'\mathbf{W})$ será 0 cuando $\lambda = 0$.

Cuando $\lambda$ tiende a $\infty$, tal volúmen será una función que dependerá de

$$f(\lambda) = \sum_{j_1 \ldots j_{p+1}} (-1)^{j_1 + \ldots + j_{p+1}} h_{1j_1} \cdots h_{(p+1)j_{p+1}}$$

con $j_1 \ldots j_{p+1}$ cualquier permutación del orden natural, y siendo los términos $h_{jk}$ los elementos de $\mathbf{H} = \mathbf{W}'\mathbf{W}$ de la forma

$$\begin{aligned} h_{jk} &= \sum_{i=1}^{n} \frac{c_{ij} c_{ik}}{\| t_i \|^2} \quad j,k = 1, \ldots, p \\ h_{j(p+1)} &= \sum_{i=1}^{n} \frac{c_{ij} \lambda}{\| t_i \|^2} \quad j = 1, \ldots, p \\ h_{(p+1)(p+1)} &= \sum_{i=1}^{n} \frac{\lambda^2}{\| t_i \|^2} \end{aligned}$$

con $t_i$ las filas de la matriz T. Estos términos tienden todos a 0 cuando $\lambda$ tiende a $\infty$, excepto el $h_{(p+1)(p+1)}$ que tiende a 1.

Tenemos por tanto que la función $f(\lambda) = \det(\mathbf{W}'\mathbf{W})$ toma el valor $f(0) = 0$, y para un cierto valor $k$ en adelante puede hacerse tan pequeña como se quiera, con lo cual $f$ estará acotada en un entorno de $\infty$.

Con esto, la función $f(\lambda)$, que es continua en $\mathbb{R}^+$ por ser sumas y productos de funciones continuas, está acotada en un compacto con lo que se sigue por continuidad que debe alcanzar su máximo absoluto en el intervalo $[0, k]$, es decir que tendremos un valor de $\lambda$ para el que "veremos" la nube de puntos desde el origen con mayor dispersión. ■

### 2.3.2 Obtención de los factores

Demostrada la existencia de un valor $\lambda_0$ que hace máxima la dispersión, el procedimiento extrae los factores siguiendo los pasos de la técnica propuesta en la sección **1.4**.

Aplicando tales resultados a la matriz $\mathbf{W}$, factorizaremos la matriz de coeficientes $\mathbf{Q} = \mathbf{W}\mathbf{W}'$, cuyos autovectores nos proporcionarán un conjunto de vectores independientes con los que podremos extraer los factores.

De forma análoga a (1.35), queremos expresar $\mathbf{W}$, aproximadamente, como el producto de una matriz de pesos factoriales y otra de puntuaciones de los factores, de la forma

$$\mathbf{W} \approx \mathbf{A}\mathbf{G}'$$

donde $\mathbf{A}$ es una matriz $n \cdot q$ y $\mathbf{G}$ una matriz $(p+1) \cdot q$, siendo $q < r = \text{rg}(\mathbf{W})$. El desarrollo para el vector $\mathbf{I}_i$ de $I\!R^{p+1}$, cuyas coordenadas son las de la fila $i$ de $\mathbf{W}$, será

$$\begin{aligned} \mathbf{I}_i &\approx (a_{i1}, a_{i2}, \ldots, a_{iq}) \cdot (\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_q)' \\ &\approx a_{i1}\mathbf{g}_1 + a_{i2}\mathbf{g}_2 + \cdots + a_{iq}\mathbf{g}_q \end{aligned}$$

con $\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_q$ las columnas de $\mathbf{G}$.

La solución más utilizada es la que se obtiene con $\mathbf{G} = \mathbf{U}_q$ y $\mathbf{A} = \mathbf{V}_q\mathbf{\Lambda}_q^{\frac{1}{2}}$, siendo las columnas de $\mathbf{U}_q$ los autovectores de $\mathbf{W}'\mathbf{W}$, y por tanto las direcciones en $I\!R^{p+1}$ que resumen a los individuos, las columnas de $\mathbf{V}_q$ los autovectores de $\mathbf{W}\mathbf{W}'$ y $\mathbf{\Lambda}_q^{\frac{1}{2}}$ los autovalores. Por su parte, $\mathbf{A}$ puede escribirse también como

$$\mathbf{A} = \mathbf{V}_q\mathbf{\Lambda}_q^{\frac{1}{2}} = \mathbf{W} \cdot \mathbf{U}_q$$

que facilita su obtención sin tener que determinar $\mathbf{W}\mathbf{W}'$

### El número de factores

El primer objetivo del análisis es determinar $q$, el número de factores independientes que generen los datos. Un valor máximo para $q$ es el rango de las matrices $\mathbf{W}$, $\mathbf{W}\mathbf{W}'$ ó $\mathbf{W}'\mathbf{W}$, que en la mayoría de los casos será $p+1$. Este rango es el que determinará el número de autovalores y autovectores en la descomposición espectral.

Sin embargo, en la mayoría de los casos es posible llegar a una aproximación buena de $\mathbf{W}$ con pocos términos de la descomposición espectral. En

estos casos, $q$ estará determinado por el rango aproximado de $\mathbf{W}$, y será el número de autovectores que utilicemos en la descomposición de $\mathbf{W}$.

Si se tuviera conocimiento por algún procedimiento previo del número $g$ de grupos que existen entre los datos, y éste valor de $g$ fuera menor que $p+1$, la decisión sobre el número de factores a elegir sería obviamente el valor de $g$. Pero en el Análisis Cluster, la determinación del número de clusters es un problema que en la mayoría de los casos viene asociado al de determinar los clusters, como se vió en la sección anterior. Es por ello que, en general, el valor de $g$ será desconocido, y tendremos que incluir en el procedimiento alguna forma de determinarlo. Vamos a asociar su extracción con el problema de determinar el número $q$ de factores de la descomposición que es capaz de explicar, salvo un error pequeño, la variabilidad existente en los datos.

Una forma de detectar tal valor de $q$ es utilizar el hecho de que la suma de los autovalores de $\mathbf{Q}$ ó $\mathbf{H}$ debe ser igual a su traza, es decir, $n$.

Como se vió en la sección **1.4**, un autovalor representa la suma de las proyecciones cuadradas de los vectores objetos en el autovector asociado. Por lo tanto, la razón de un autovalor respecto de la traza de $\mathbf{H}$, es una medida apropiada de la información contenida en el factor asociado. Si esta razón indica que un factor contribuye con una cantidad pequeña de información a la solución, puede decirse que este factor es insignificante. El número de autovalores útiles servirá, entonces, como una primera aproximación a $q$.

### Rotación de los factores

Nuestra intención, como se ha dicho anteriormente, es utilizar los medios necesarios para que los diferentes grupos que pudieran existir en un conjunto de datos, salgan a la luz con la ayuda de los factores que se obtienen en este tipo de análisis. Ya hemos mencionado la limitación que existe en cuanto a su número y la posibilidad de despreciar aquellos que aporten muy poca información al conjunto de datos. En nuestro caso, esta información debe ser poco relevante al utilizarla para descubrir grupos entre ellos.

La solución obtenida hasta ahora permite que los objetos originales pue-

dan representarse respecto de los factores obtenidos, los cuales serán ortogonales entre sí al proceder del sistema de autovectores de la matriz **H**. Pero esta representación no es única. Como ya se ha mencionado en el capítulo primero, al descomponer una matriz en factores, existen infinidad de transformaciones ortogonales **T** que permitan expresar **W** como producto

$$\mathbf{W} = \mathbf{A}^*(\mathbf{F}^*)'$$

siendo $\mathbf{A}^* = \mathbf{AT}$ y $\mathbf{F}^* = \mathbf{FT}$. Pero además, suprimiendo la hipótesis de ortogonalidad entre los factores, pueden obtenerse, a veces, soluciones factoriales que simplifiquen la matriz de pesos factoriales en cuanto a que cada factor pese más en un objeto determinado que en el resto.

En nuestro caso, una vez obtenida la solución dada por los primeros autovectores de **H**, es necesario buscar nuevas soluciones que definan los grupos con más claridad. En principio, los puntos que representan a los objetos estarán situados en direcciones que recogen la máxima variabilidad en orden decreciente del conjunto de datos, pero éstas no tienen porqué pasar por los centros de gravedad de los posibles grupos que existan entre los datos. Nuestro próximo objetivo al rotar los factores va a ser precisamente buscar nuevos factores, ortogonales o no, que pasen por estos centros de gravedad. Es en este punto donde el procedimiento que proponemos cobra su verdadera dimensión como procedimiento de optimización. Como se verá a continuación, las rotaciones que pueden efectuarse con los factores son transformaciones que permiten encontrar factores derivados más próximos a los grupos de elementos que existan en el conjunto de datos. La amplia variedad de criterios que los analístas factoriales han propuesto para obtener estas rotaciones, no responden más que al deseo de simplificar las matrices de pesos y estructura en orden a obtener valores extremos en ellas que clarifiquen las asociaciones existentes entre los factores y los elementos. Ya que tanto unos como otros se han normalizado, las distancias entre ellos dependen únicamente del coseno del ángulo que forman, según

$$\begin{aligned} d^2(\mathbf{I}_i, \mathbf{g}_j) &= \sum_{k=1}^{p} x_{ik}^2 + \sum_{k=1}^{p} g_{jk}^2 - 2 \cdot \sqrt{\sum_{k=1}^{p} x_{ik}^2} \cdot \sqrt{\sum_{k=1}^{p} g_{jk}^2} \cdot q_{ij} \\ &= 1 + 1 - 2 \cdot \cos\theta_{ij} \\ &= 2 \cdot (1 - \cos\theta_{ij}) \end{aligned}$$

De esta forma, minimizar expresiones del tipo

$$E^* = \sum_{i=1}^{n} d^2_{i,\mathbf{g}(i)},$$

con $\mathbf{g}(i)$ el factor que más pesa en el objeto $i$, es equivalente a maximizar los cosenos de los ángulos entre los factores y los individuos, es decir, maximizar los elementos extradiagonales de la matriz de estructura factorial, que es uno de los objetivos de las rotaciones factoriales.

El hecho de que los nuevos factores que buscamos tengan que pasar lo más cerca posible de los centros de gravedad exige que no tengan que ser, en general, necesariamente ortogonales. Es por ello que nosotros utilizaremos en nuestro procedimiento rotaciones oblicuas que sean capaces de obtener lo esperado.

De los procedimientos apuntados en al sección **1.2**, utilizaremos aqui el criterio propuesto por Jennrich y Sampson [77].

En él, se encuentra la matriz transformación $\mathbf{T}$ y el patrón oblicuo $\mathbf{B} = \mathbf{A}(\mathbf{T})^{-1}$, sujeta a la condición $\text{diag}(\mathbf{T}'\mathbf{T}) = \mathbf{I}$, minimizando

$$F(\mathbf{B}) = F(\mathbf{A}(\mathbf{T}')^{-1}) = \sum_{j\leq k=1}^{q} \left(\sum_{i=1}^{n} \frac{b^2_{ij}}{h^2_i}\frac{b^2_{ik}}{h^2_i} - \frac{\beta}{n}\sum_{i=1}^{n}\frac{b^2_{ij}}{h^2_i}\sum_{i=1}^{n}\frac{b^2_{ik}}{h^2_i}\right) \tag{2.4}$$

con $\beta = 0$.

El método actúa mediante secuencias de rotaciones simples, en las que se van tomando parejas de factores, por ejemplo $\mathbf{g}_1$ y $\mathbf{g}_2$, y calculando la rotación simple

$$\tilde{\mathbf{g}}_1 = t_1\mathbf{g}_1 + t_2\mathbf{g}_2 \tag{2.5}$$

con $t_1^2 + 2t_1t_2c_{12} + t_2^2 = 1$ y $c_{12}$ la correlación entre $\mathbf{g}_1$ y $\mathbf{g}_2$, de tal forma que la nueva matriz de pesos $\tilde{\mathbf{B}}$ haga mínimo $\mathrm{F}(\tilde{\mathbf{B}})$.

Las ecuaciones

$$\tilde{b}_1 = \frac{1}{t_1}a_1, \quad \tilde{b}_2 = \frac{-t_2}{t_1}a_1 + a_2, \quad \text{y} \quad \tilde{b}_i = a_1, \quad \text{para} \quad i \neq 1,2 \tag{2.6}$$

y el cambio de coordenadas

$$\gamma = \frac{1}{t_1}, \quad \delta = \frac{t_2}{t_1} \tag{2.7}$$

permiten poner $F(\tilde{\mathbf{B}})$ como un polinomio de cuarto grado, que es en realidad el que se minimiza eligiendo la raíz de su derivada que cumple ese propósito.

Desecho el cambio (2.7), se obtiene la nueva matriz de pesos $\tilde{\mathbf{B}}$ mediante (2.6) y la de correlaciones $\tilde{\mathbf{C}}$ mediante las ecuaciones (2.8)

$$\tilde{c}_{1j} = t_1 c_{1j} + t_2 c_{2j}, \quad \text{para} \quad j \neq 1; \quad \text{y} \quad \tilde{c}_{ij} = c_{ij}, \quad \text{para} \quad i, j \neq 1. \tag{2.8}$$

Las rotaciones se van tomando con todas las parejas posibles de factores hasta que $F(\mathbf{B})$ converge, siendo los valores finales de $\mathbf{B}$ y $\mathbf{C}$ las matrices de pesos y de correlaciones de la solución factorial rotada. Por último, la expresión (1.17) nos permite obtener la matriz $\mathbf{S}$ de estructura factorial.

### 2.3.3 Ejemplo

Un ejemplo real que vamos a utilizar para comprobar si el procedimiento funciona nos lo proporcionan los conocidos datos de los tres tipos de Iris que Fisher [46] utilizaró por primera vez en problemas de discriminación. Estos datos (Apéndice A) consisten en las medidas de las longitudes y anchuras de los sépalos y pétalos de 150 especímenes de plantas de Iris. Fueron utilizadas 50 plantas de cada uno de los tres tipos de Iris Setosa, Versicolor y Virgínica, aunque aquí sólo tomaremos las 10 primeras flores, entre las que hay 3 del Tipo Setosa (A), 3 del Tipo Versicolor (B) y 4 del Tipo Virgínica (C), según la Tabla 2.1.

| IRIS | Long. Sepa. | Anch. Sepa. | Long. Peta. | Anch. Peta. | Tipo Iris |
|---|---|---|---|---|---|
| 1 | 50 | 33 | 14 | 02 | A |
| 2 | 64 | 28 | 56 | 22 | C |
| 3 | 65 | 28 | 46 | 15 | B |
| 4 | 67 | 31 | 56 | 24 | C |
| 5 | 63 | 28 | 51 | 15 | C |
| 6 | 46 | 34 | 14 | 03 | A |
| 7 | 69 | 31 | 51 | 23 | C |
| 8 | 62 | 22 | 45 | 15 | B |
| 9 | 59 | 32 | 48 | 18 | B |
| 10 | 46 | 36 | 10 | 02 | A |

Tabla 2.1: Las 10 primeras flores Iris utilizadas por Fisher.

En la obtención de los resultados que siguen se ha utilizado el procedimiento PROCED, que será comentado en la sección siguiente más detenidamente y cuyo listado aparece en el Apéndice B. Primero se ha obtenido el modelo Q-factorial para estos datos sin desplazarlos en una dimensión más, y después haciendo el desplazamiento como se ha indicado anteriormente.

Con estos datos, después de centrarlos por columnas y normalizarlos por filas, se obtiene una matriz de pesos inicial (Tabla 2.2) dada por el modelo Q-factorial (1.35),

| Iris | Factor 1 | Factor 2 | Factor 3 |
|---|---|---|---|
| I1 | -0.99 | -0.08 | 0.04 |
| I2 | 0.98 | 0.07 | -0.11 |
| I3 | 0.93 | -0.25 | 0.21 |
| I4 | 0.97 | 0.19 | 0.07 |
| I5 | 0.96 | -0.11 | -0.13 |
| I6 | -0.99 | 0.00 | -0.06 |
| I7 | 0.93 | 0.16 | 0.29 |
| I8 | 0.71 | -0.66 | -0.15 |
| I9 | 0.87 | 0.39 | -0.27 |
| I10 | -1.00 | 0.02 | 0.00 |
| V.E. | 88.6% | 7.5% | 2.7% |

Tabla 2.2: Matriz de pesos iniciales del modelo Q-factorial para los datos de la Tabla 2.1.

y si este patrón factorial se rota oblicuamente utilizando el criterio de Jennrich y Sampson [77] con los dos primeros factores, se obtiene una matriz de pesos rotada (izquierda) y de estructuras (derecha) dadas por la Tabla 2.3.

| Iris | Factor 1 | Factor 2 | | Factor 1 | Factor 2 |
|---|---|---|---|---|---|
| I1 | -0.96 | 0.07 | | -0.99 | 0.53 |
| I2 | 0.95 | -0.07 | | 0.99 | -0.53 |
| I3 | 0.66 | -0.43 | | 0.87 | -0.75 |
| I4 | 1.02 | 0.06 | | 0.99 | -0.43 |
| I5 | 0.79 | -0.28 | | 0.92 | -0.66 |
| I6 | -0.91 | 0.16 | | -0.98 | 0.59 |
| I7 | 0.96 | 0.02 | | 0.95 | -0.44 |
| I8 | 0.16 | -0.86 | | 0.58 | -0.94 |
| I9 | 1.08 | 0.30 | | 0.93 | -0.21 |
| I10 | -0.88 | 0.19 | | -0.98 | 0.62 |

Tabla 2.3: Matriz de pesos rotada y de estructura con dos factores para el modelo Q-factorial con los datos de la Tabla 2.1.

Una simple inspección de la matriz de estructuras de la Tabla 2.3 nos permite ver que los posibles grupos que pueden determinar los factores son, por un lado, el formado por las flores del grupo Setosa (I1, I6 y I10) que tienen una asociación alta con la dirección negativa del Factor 1, con cosenos directores próximos a $-1$. Por otro lado, las flores de los grupos Versicolor y Virgínica (I2, I3, I4, I5, I7 y I9) se asocian al mismo factor 1 en su dirección positiva sin posibilidad clara de distinción entre los dos grupos, mientras que la flor I8 del grupo Versicolor quedaría asociada a la dirección negativa del Factor 2.

Si en lugar de tomar los dos primeros factores tomáramos los tres primeros, se tienen las matrices de pesos rotada y de estructuras dadas en la Tabla 2.4, que mantienen la misma asociación anterior.

| Iris | Fac. 1 | Fac. 2 | Fac. 3 | | Fac. 1 | Fac. 2 | Fac. 3 |
|---|---|---|---|---|---|---|---|
| I1 | -0.94 | 0.09 | 0.04 | | -0.99 | 0.58 | 0.01 |
| I2 | 0.92 | -0.13 | -0.10 | | 0.98 | -0.59 | -0.06 |
| I3 | 0.70 | -0.33 | 0.28 | | 0.87 | -0.74 | 0.34 |
| I4 | 1.04 | 0.08 | 0.04 | | 0.99 | -0.46 | 0.05 |
| I5 | 0.75 | -0.34 | -0.10 | | 0.92 | -0.71 | -0.03 |
| I6 | -0.91 | 0.13 | -0.07 | | -0.98 | 0.62 | -0.11 |
| I7 | 1.03 | 0.14 | 0.27 | | 0.96 | -0.43 | 0.26 |
| I8 | 0.11 | -0.90 | 0.003 | | 0.58 | -0.96 | 0.15 |
| I9 | 1.02 | 0.16 | -0.34 | | 0.93 | -0.31 | -0.35 |
| I10 | -0.88 | 0.19 | -0.02 | | -0.98 | 0.65 | -0.06 |

Tabla 2.4: Matriz de pesos rotada y de estructura con tres factores para el modelo Q-factorial con los datos de la Tabla 2.1.

Si ahora utilizamos el desplazamiento descrito en el procedimiento PROCED antes de extraer los factores, las primeras operaciones de centralización y la búsqueda de la distancia de desplazamiento que maximice la dispersión nos dan una matriz **T**,

$$\mathbf{T} = \begin{pmatrix} -9.1 & 2.7 & -25.1 & -11.9 & 6.7 \\ 4.9 & -2.3 & 16.9 & 8.1 & 6.7 \\ 5.9 & -2.3 & 6.9 & 1.1 & 6.7 \\ 7.9 & 0.7 & 16.9 & 10.1 & 6.7 \\ 3.9 & -2.3 & 11.9 & 1.1 & 6.7 \\ -13.1 & 3.7 & -25.1 & -10.9 & 6.7 \\ 9.9 & 0.7 & 11.9 & 9.1 & 6.7 \\ 2.9 & -8.3 & 5.9 & 1.1 & 6.7 \\ -0.1 & 1.7 & 8.9 & 4.1 & 6.7 \\ -13.1 & 5.7 & -29.1 & -11.9 & 6.7 \end{pmatrix}$$

Matriz T obtenida al desplazar una dimensión más los elementos de la Tabla 2.1 centrados por columnas.

y la matriz que se obtiene al normalizar los vectores fila será la **W**.

$$\mathbf{W} = \begin{pmatrix} -0.30 & 0.09 & -0.83 & -0.39 & 0.22 \\ 0.24 & -0.11 & 0.82 & 0.39 & 0.32 \\ 0.51 & -0.20 & 0.59 & 0.09 & 0.58 \\ 0.35 & 0.03 & 0.76 & 0.45 & 0.30 \\ 0.27 & -0.16 & 0.82 & 0.07 & 0.46 \\ -0.42 & 0.12 & -0.80 & -0.35 & 0.21 \\ 0.51 & 0.03 & 0.62 & 0.47 & 0.35 \\ 0.23 & -0.66 & 0.47 & 0.08 & 0.53 \\ 0.00 & 0.14 & 0.74 & 0.34 & 0.55 \\ -0.37 & 0.16 & -0.82 & -0.34 & 0.19 \end{pmatrix}$$

**Matriz W obtenida al normalizar por filas las matriz T.**

Los tres primeros autovalores de la matriz $\mathbf{W}'\mathbf{W}$ aparecen en la Tabla 2.5.

| Nº orden | Autovalor | Porc. inform. | Porc. inf. acumulada |
|---|---|---|---|
| $\lambda_1$ | 8.08 | 80.8 | 80.8 |
| $\lambda_2$ | 1.14 | 11.4 | 92.2 |
| $\lambda_3$ | 0.49 | 4.9 | 97.1 |

Tabla 2.5: Autovalores de $\mathbf{W}'\mathbf{W}$ y porcentaje de información que representan.

Dependiendo del número de factores que queramos tomar (según el porcentaje de la variabilidad explicada), la matriz de pesos factoriales $\mathbf{A}$ vendrá dada por el mismo número de columnas de la Tabla 2.6, obtenida mediante $\mathbf{A} = \mathbf{WG}$, donde en $\mathbf{G}$ se tienen por columnas los autovectores correspondientes a los autovalores de la Tabla 2.5.

| Iris | Factor 1 | Factor 2 | Factor 3 |
|---|---|---|---|
| I1 | -0.87 | 0.47 | 0.06 |
| I2 | 0.99 | 0.03 | 0.08 |
| I3 | 0.88 | 0.39 | -0.10 |
| I4 | 0.98 | -0.02 | 0.18 |
| I5 | 0.94 | 0.25 | -0.02 |
| I6 | -0.88 | 0.45 | 0.14 |
| I7 | 0.94 | 0.02 | 0.15 |
| I8 | 0.72 | 0.50 | -0.44 |
| I9 | 0.85 | 0.25 | 0.41 |
| I10 | -0.89 | 0.42 | 0.16 |
| V.E. | 80.8% | 11.4% | 4.9% |

Tabla 2.6: Matriz de pesos inicial en el modelo Q-factorial para los datos de la Tabla 2.1 transformados en $W$.

Podemos ahora tomar también dos o tres factores en el modelo inicial, dependiendo del porcentaje de variabilidad que queramos explicar. Esto puede ser relevante si se tiene en cuenta que a la hora de rotarlos no es lo mismo minimizar el criterio (2.4) con 2 factores que con 3. La información que se pierde al tomar menos factores puede hacer que oblige a los elementos a formar menos grupos de los que en realidad existen, y por otra parte, si se toman demasiados factores, el procedimiento de rotación puede hacer aparecer más grupos de los que hay. En la sección siguiente se tratará este tema con más profundidad mediante un estudio de simulación.

Los resultados para la nueva matriz de pesos rotada oblicuamente cuando se toman tres factores son los de la Tabla 2.7.

| Iris | Factor 1 | Factor 2 | Factor 3 |
|---|---|---|---|
| I1 | -0.06 | 0.94 | -0.07 |
| I2 | 0.56 | -0.35 | -0.19 |
| I3 | 0.42 | -0.01 | -0.61 |
| I4 | 0.67 | -0.38 | -0.04 |
| I5 | 0.48 | -0.15 | -0.45 |
| I6 | 0.02 | 0.95 | 0.01 |
| I7 | 0.64 | -0.32 | -0.09 |
| I8 | -0.08 | 0.01 | -0.99 |
| I9 | 1.07 | 0.09 | 0.01 |
| I10 | 0.03 | 0.93 | 0.06 |

Tabla 2.7: Matriz de pesos rotada con tres factores en el modelo Q-factorial para los datos de la Tabla 2.1 transformados en $W$.

La obtención de los factores oblicuos anteriores nos va a permitir llegar a las conclusiones que nos proponíamos respecto de los posibles grupos existentes entre los datos. Tanto la Tabla 2.7 anterior como, en especial, la que se obtendría con la matriz de estructura factorial (Tabla 2.8), pueden servirnos para llegar a una clasificación de acuerdo a las asociaciones que la matriz de estructuras establece entre los factores y los Iris, evidenciando las diferencias respecto del modelo extraído sin realizar el desplazamiento.

| Iris | Factor 1 | Factor 2 | Factor 3 |
|---|---|---|---|
| IRIS1 | -0.62 | 0.95 | 0.37 |
| IRIS2 | 0.91 | -0.80 | -0.69 |
| IRIS3 | 0.80 | -0.54 | -0.87 |
| IRIS4 | 0.93 | -0.82 | -0.61 |
| IRIS5 | 0.85 | -0.66 | -0.81 |
| IRIS6 | -0.59 | 0.94 | 0.41 |
| IRIS7 | 0.90 | -0.77 | 0.62 |
| IRIS8 | 0.51 | -0.36 | -0.94 |
| IRIS9 | 1.00 | -0.59 | -0.60 |
| IRIS10 | -0.60 | 0.94 | 0.44 |

Tabla 2.8: Matriz de estructuras después de rotar con tres factores en el modelo Q-factorial para los datos de la Tabla 2.1 transformados en $W$.

En la Tabla 2.8 vemos cómo los Iris que tienen coeficiente de asociación máximo con el primer factor oblicuo son el 2, 4, 5, 7 y 9, los que mejor se asocian con el segundo son el 1, 6 y 10, y con el tercero el 3 y 8. Puede establecerse así una clasificación originada por los factores, que como puede comprobarse agrupa a los Iris del grupo Virgínica y uno Versicolor en el primer factor, a los del grupo Setosa en el segundo, y a dos del grupo Versicolor en el tercero. Combinando la tabla de datos originales con esta clasificación se obtiene la Tabla 2.9,

| Iris | Grupo | Factor |
|---|---|---|
| I1 | A | 2 |
| I2 | C | 1 |
| I3 | B | 3 |
| I4 | C | 1 |
| I5 | C | 1 |
| I6 | A | 2 |
| I7 | C | 1 |
| I8 | B | 3 |
| I9 | B | 1 |
| I10 | A | 2 |

Tabla 2.9: Correspondencia entre clasificación original y clasificación obtenida con PROCED para los datos de la Tabla 2.1.

la cual pone en evidencia que, a excepción del Iris 9º, los demás han sido bien clasificados.

Si hubiéramos tomado los dos primeros factores, los resultados cambian algo. En efecto, la matriz de pesos rotada sería la de la Tabla 2.10,

| Iris | Factor 1 | Factor 2 |
|---|---|---|
| I1 | 0.02 | 0.94 |
| I2 | 0.64 | -0.43 |
| I3 | 0.98 | 0.00 |
| I4 | 0.55 | -0.50 |
| I5 | 0.86 | -0.17 |
| I6 | 0.00 | 0.92 |
| I7 | 0.59 | -0.43 |
| I8 | 1.01 | 0.20 |
| I9 | 0.80 | -0.13 |
| I10 | -0.05 | 0.89 |

Tabla 2.10: Matriz de pesos rotada con dos factores en el modelo Q-factorial para los datos de la Tabla 2.1 transformados en $W$.

y la de estructuras la de la Tabla 2.11,

| Iris | Factor 1 | Factor 2 |
|---|---|---|
| IRIS1 | -0.51 | 0.93 |
| IRIS2 | 0.88 | -0.80 |
| IRIS3 | 0.98 | -0.55 |
| IRIS4 | 0.84 | -0.82 |
| IRIS5 | 0.96 | -0.66 |
| IRIS6 | -0.53 | 0.92 |
| IRIS7 | 0.84 | -0.77 |
| IRIS8 | 0.90 | -0.38 |
| IRIS9 | 0.87 | -0.59 |
| IRIS10 | -0.56 | 0.92 |

Tabla 2.11: Matriz de estructuras después de rotar con dos factores en el modelo Q-factorial para los datos de la Tabla 2.1 transformados en $W$.

poniendo de manifiesto que ahora el Factor 1 aglutina al conjunto de los Iris Virgínica y Versicolor, mientras que el Factor 2 agruparía a los Iris Setosa.

Por último, no terminaremos la sección sin comentar que el procedimiento de clasificación que estamos proponiendo depende no sólo del número de

factores que se elijan en el modelo Q-factorial, sino también de la dimensión de los datos, número de grupos, número de elementos en cada grupo y, en general, todos aquellos factores que puedan influir a la hora de detectar las separaciones entre los grupos, cuando se realiza el desplazamiento de los elementos en una dimensión más.

Concretamente, cuando se ha utilizado PROCED con las 150 flores Iris (Apéndice A), los resultados obtenidos son algo peores que los anteriores, ya que se han conseguido clasificar bien a 124 de ellas (82.66 %), cuando se tomaron 2 factores, y a 111 (74 %) cuando se tomaron 3 factores.

## 2.4 Estudio de simulación utilizando PROCED. Comparacion con otros métodos de optimización del Análisis Cluster

Para evaluar el funcionamiento del procedimiento que se propone, se ha realizado una simulación generando una muestra de 100 conjuntos de datos, cada uno de ellos con 50 elementos entre los que existen grupos definidos.

Como el número de factores a tener en cuenta para la formación de esta muestra es muy elevado, se han acotado las posibilidades intentando no quitar generalidad y aleatoriedad a la muestra elegida.

En el Apéndice B se adjuntan una serie de Programas realizados con el Lenguaje FORTRAN 77 y la ayuda de algunas subrutinas de IMSL. Entre ellos, puede verse cómo el algoritmo GENALTO utiliza la generación de números aleatorios entre 0 y 1 para obtener:

1. Un número entre 1 y 5 que determine la dimensión del espacio inicial de los datos : $ndim$.

2. Un número entre 2 y 6 que determine el número de grupos definido en cada conjunto : $ngrup$.

3. Para cada grupo, se determina un punto aleatorio que será considerado como centro del grupo. Para ello se elige:

   (a) Un número aleatorio entre 0 y 50, que será el radio de la esfera con centro el origen donde esté situado el punto : $r(i)$, $i = 1, \ldots, ngrup$.

   (b) Cada coordenada del punto $c(i) = (x_1(i), \ldots, x_{ndim}(i)$ se calcula de forma aleatoria con la condición de que

   $$x_1(i)^2 + \cdots + x_{ndim}(i)^2 = r(i)^2$$

   (c) Además, se exige que la distancia mínima entre dos centros cualesquiera dentro de un conjunto sea al menos de 3.

(d) A continuación, en cada grupo ($i$) y coordenada ($j$), se determina aleatoriamente un coeficiente de dispersión $disp(i,j)$ entre 1 y 3 que determina en cada grupo la distancia máxima a la que estarán las coordenadas de los elementos de las del centro. Esta dispersión se determina coordenada a coordenada independientemente, de forma que los elementos de cada grupo $i$ pueden encontrarse distribuidos uniformemente dentro del hiperparalelepípedo cuyo centro sea $c(i)$ y cuyos lados midan $2 \cdot disp(i,j)$ para $j = 1, \cdots, ndim$.

(e) Por último, se elije un número de elementos ($kelem(i)$) para cada grupo con la restricción de que el conjunto en total tenga 50, y se van obteniendo las coordenadas de los elementos de cada grupo sumando o restando a las coordenadas del centro un número aleatorio entre 1 y 3 determinado por el coeficiente $disp(i,j)$.

Un Diagrama del Algoritmo GENALTO se presenta a continuación:

**Diagrama de flujo del algoritmo GENALTO**

**DIMENS.** Se dimensionan el vector de distancias de centros al origenr: $r(10)$, la matriz de centros por cada grupo y dimensión: $c(10,10)$, la matriz de dispersión por grupo y dimensión: $disp(10,10)$, el vector del número de elementos por grupo: $kelem(10)$ y el vector de coordenadas de los elementos: $x(10)$.

**DECLAR.** Subrutinas externas de IMSL para generar aleatorios: $rnset$, $rnunf$.

**INICIO** Valores aleatorios $rnset(0)$.

**SALIDA** Muestra de 100 conjuntos con 50 elementos cada uno y grupos entre los elementos. Se escribe en un fichero para su análisis posterior, teniendo la precaución de reflejar al comienzo de cada conjunto: un número de acuerdo al lugar que ocupa, la dimensión de los datos y el número de grupos, así como añadir al final de cada elemento el grupo al que pertenece.

**PARA** $l$ recorriendo de 1 a 100:

**Paso 1** Se obtiene la dimensión de los datos del conjunto: $ndim$,

**Paso 2** Se obtiene el número de grupos del conjunto: $ngrup$,

**Paso 3** Se escriben $l$, $ndim$, $ngrup$ en SALIDA,

**Paso 4** Se pone $cond$ a 1,

**Paso 5** MIENTRAS ($cond = 1$), HACER Pasos 5.1 a 5.4:

**Paso 5.1** Se obtienen las distancias de cada centro de grupo al origen: $r(i)$,

**Paso 5.2** Se obtiene los centros de cada grupo: $c(i,j)$, y se pone $cond$ a 0,

**Paso 5.3** Se calculan las distancias entre todos los centros,

**Paso 5.4** SI alguna distancia es menor que 3, ENTONCES se pone $cond = 1$.

**Paso 6** Se obtiene la dispersión en cada grupo y dimensión: $disp(i,j)$,

**Paso 7** Se obtiene el número de elementos en cada grupo: $kelem(i)$,

**Paso 8** Se generan las coordenadas de los elementos en cada grupo y dimensión con $c(i,j)$ y $disp(i,j)$, y se escriben en SALIDA.

**FIN**

El programa está preparado para obtener al mismo tiempo los 100 conjuntos en un fichero de salida, que ha servido de fichero de datos para el Algoritmo PROCED dado también en el Apéndice B.

Como puede apreciarse, PROCED va tomando cada uno de los 100 conjuntos generados por GENALTO y le aplica el procedimiento expuesto a lo largo de la sección **2.3**.

Para aplicar este procedimiento, el programa PROCED consta de un programa principal y varias subrutinas y funciones definidas, utilizadas a lo largo de él, que enumeramos a continuación:

1. La Subrutina ROT, que obtiene la matriz de pesos factoriales rotada por el criterio cuartimín directo de Jennrich y Sampson [77]. Esta Subrutina tiene asociadas las funciones CRIT y POL4, que se utilizan como funciones definidas dentro de ella.

2. La Función DET, donde se calcula el determinante de la matriz cuadrada **W'W** y con la ayuda de la subrutina UVMIF de IMSL se obtiene el desplazamiento que maximiza la dispersión.

3. La Subrutina TRAT, donde se parte de la matriz de datos inicial centrada por columnas, desplazada en una dimensión más y normalizada por filas y se obtiene el patrón factorial inicial.

**Diagrama de flujo del algoritmo PROCED**

**DIMENS.** Se dimensionan generosamente la matriz total de elementos: $t(10000, 10)$, la matriz temporal con los elementos de cada conjunto: $a(100, 10)$, la matriz de estructuras factoriales: $s(100, 10)$, el vector de normas de los elementos: $xnor(100)$.

**DECLAR.** Se declaran variables enteras para ir eligiendo el conjunto de elementos temporal: $comi$, $conj$, $ncolum$, $nfilas$, $ngrup$.

**DIMENS.** Se dimensionan vector de agrupación inicial: $grupin(100)$, agrupación final: $in(100)$ y matriz para contar el número de elementos bien clasificados: $matbu(20, -20 : 20)$.

**DECLAR.** Se declaran variables enteras para contar el número de elementos bien clasificados en cada conjunto: $buenos$ y el total: $buentotal$.

COMÚN Se dimensionan y declaran comunes en todo el algoritmo la matriz de elementos centrados por columnas: $c(100, 10)$, la matriz de elementos centrada por columnas, desplazada una dimensión más y normalizada por filas: $w(100, 10)$, la matriz producto W'W : $wpw(10, 10)$, el número de filas: $nfilas$, el número inicial de columnas: $ncolum$, el número de autovalores de W'W que se eligen: $kautv$, la distancia que se desplazan los datos en una dimensión más: $d$ y la cota inferior para elegir los autovalores de W'W : $cotaut$.

**COMÚN** Se dimensionan y declaran comunes con la subrutina ROT la matriz de pesos: $b(100, 10)$, la matriz de correlaciones entre los factores: $cor(10, 10)$, la matriz de autovectores de W'W : $avec(10, 10)$ y el número de columnas después de desplazar los elementos: $nncolum$.

**DECLAR.** Se declaran subrutinas y funciones externas definidas para efectuar rotaciones: TRAT, para calcular det(W'W ): DET, para el criterio a minimizar en las rotaciones: CRIT, para el polinomio a minimizar según el criterio de rotación: POL4 y las de IMSL para minimizar una función: UVMIF, para calcular los autovalores de una matriz cuadrada: EVLSF y para calcular los autovalores y autovectores de una matriz cuadrada: EVCSF

**ENTRADA** Se introduce la matriz total de datos con los 100 conjuntos generados por GENALTO y la cota mínima para elegir los autovalores.

**SALIDA** Se obtienen los resultados que se describen a continuación para cada conjunto de datos.

**INICIO** Se inicia el contador de elementos totales bien clasificados a 0.

**PARA** $l$ recorriendo de 1 a 100:

**Paso 1** Se obtiene la matriz y parámetros temporales: $l$, $ncolum$, $ngrup$, $a(i,j)$ y $grupin(i)$,

**Paso 2** Se centran los datos de $a(i,j)$ por columnas,

**Paso 3** Se obtiene la distancia $d$ con ayuda de DET y UVMIF,

**Paso 4** Se calcula W'W una vez desplazados los datos y el valor de $det(W'W)$,

**Paso 5** Se llama a la subrutina TRAT que nos dará el número de factores, los factores, la matriz de pesos inicial y la de correlaciones entre factores inicial,

**Paso 6** SI $(kautv > 1)$, HACER Pasos 6.1 y 6.2:

**Paso 6.1** Se normaliza por filas la matriz de pesos inicial,

**Paso 6.2** Con la ayuda de CRIT y ROT, se realizan rotaciones con los factores iniciales hasta que el valor del criterio a minimizar converge,

**Paso 7** Se obtiene la matriz de pesos rotada,

**Paso 8** Se obtiene la matriz de correlaciones entre los factores rotados,

**Paso 9** Se obtiene la matriz de estructura factorial,

**Paso 10** Utilizando la matriz de estructura factorial, se obtiene el vector $in(i)$ de agrupación final para cada elemento, asociándole a cada uno la dirección del factor con el que tiene mayor valor absoluto en la matriz de estructuras,

**Paso 11** Se escriben en SALIDA los datos, la agrupación inicial y la dirección del factor asignada,

**Paso 12** Se obtiene la matriz $matbu(i, j)$ que cuenta el número de elementos en cada número de grupo inicial y dirección,

**Paso 13** Se cuenta el número de elementos correctamente clasificados en el conjunto, asignando a cada grupo la dirección donde encuentra más elementos, escribiendo en SALIDA el valor de buenos.

**TERMINA** Finaliza el bucle de todos los conjuntos

**CUENTA** Cuenta y escribe elementos totales bien clasificados, realizando los Pasos A al E:

**Paso A** Con los elementos de la matriz *matbu*, donde sus elementos representan por filas el número de datos de cada grupo inicial que hay en cada uno de los grupos finales, se toma en primer lugar el máximo de sus elementos y se introduce en *buenos*.

**Paso B** Se ponen a 0 la fila y columna del elemento elegido, y con la matriz que queda se vuelve a tomar el máximo de sus elementos, incrementando *buenos* con este valor.

**Paso C** El Paso B se repite hasta que todos los elementos de *matbu* sean 0.

**Paso D** El número de elementos bien clasificados del conjunto de elementos temporal será el valor de *buenos*, y se escribe.

**Paso E** El número de elementos bien clasificados para los 100 conjuntos *buentotal*, se va incrementando con los valores de *buenos*, escribiéndolo al final.

**FIN**

**Diagrama de flujo de la Función CRIT**

**INICIO** Se define la función CRIT($nfilas$,$kautv$)

**COMÚN** Se dimensionan y declaran en común con la subrutina ROT la matriz de pesos: $b(100, 10)$, la matriz de correlaciones entre los factores: $cor(10, 10)$, la matriz de factores inicial: $avec(10, 10)$ y el número de columnas después de desplazar: $nncolum$.

**CÁLCULO** Se obtiene el valor del criterio a minimizar en las rotaciones.

**RETURN**

### Diagrama de flujo de la Función POL4

**INICIO** Se define la función POL4(x).

**COMÚN** Se declaran en común con la subrutina ROT los coeficientes del polinomio.

**CÁLCULO** Se obtiene el valor del polinomio.

**RETURN**

### Diagrama de flujo de la Subrutina ROT

**INICIO** Se define la subrutina ROT($nfilas, kautv$).

**DECLAR.** Se declara la función POL4 y la subrutina de IMSL UVMIF.

**COMÚN** Se dimensionan y declaran en común con la Función CRIT la matriz de pesos: $b(100, 10)$, la matriz de correlaciones entre los factores: $cor(10, 10)$, la matriz de factores inicial: $avec(10, 10)$ y el número de columnas después de desplazar: $nncolum$.

**COMÚN** Se declaran en común con la Función POL4 los coeficientes del polinomio.

**PARA** Eligiendo todas las parejas de factores posibles:

**Paso 1** Se obtienen los coeficientes del polinomio de $4^o$ grado,

**Paso 2** Con ayuda de POL4, se obtiene el valor que minimiza el polinomio: $xgam$,

**Paso 3** Se obtiene la nueva matriz de pesos con los dos factores rotados,

**Paso 4** Se obtienen los nuevos factores rotados,

**Paso 5** Se obtiene la nueva matriz de correlaciones entre factores.

**TERMINA** Finalizan los bucles que eligen cada pareja de factores.

**RETURN**

**Diagrama de flujo de la Función DET**

**INICIO** Se define la función DET($d$).

**DIMENS.** Se dimensionan el vector de autovalores temporales: $aval(10)$ y $avalord(10)$.

**COMÚN** Se dimensionan y declaran en común con el Programa Principal la matriz centrada: $c(100, 10)$, la matriz centrada, desplazada y normalizada: $w(100, 10)$, la matriz producto W'W : $wpw(10, 10)$ y los parámetros $nfilas$ y $ncolum$.

**Paso 1** Se inicializa $nncolum$ a $ncolum + 1$.

**Paso 2** Se añade la columna constante a $c(i, j)$ y se calculan las normas de cada fila.

**Paso 3** SI ($norma > 0$) ENTONCES se obtiene matriz $w(i, j)$ normalizando por filas la $c(i, j)$ desplazada.

**Paso 4** Se obtiene la matriz producto W'W .

**Paso 5** Mediante la subrutina de IMSL EVLSF se calculan los autovalores de W'W y, posteriormente, $det(W'W)$ en DET.

**RETURN**

**Diagrama de flujo de la Subrutina TRAT**

**INICIO** Se define la subrutina TRAT().

**DIMENS.** Se dimensionan los vectores de autovalores temporales: $aval(10)$, $oval(10)$ y la matriz de autovectores: $ovec(10, 10)$.

**COMÚN** Se dimensionan y declaran en común con el Programa Principal la matriz centrada: $c(100, 10)$, la matriz centrada, desplazada y normalizada: $w(100, 10)$, la matriz producto W'W : $wpw(10, 10)$ y los parámetros nfilas y $ncolum$, $kautv$, $d$ y $cotaut$.

**COMÚN** Se dimensionan y declaran en común con la Subrutina ROT la matriz de pesos: $b(100, 10)$, la matriz de correlaciones entre los factores: $cor(10, 10)$, la matriz de factores inicial: $avec(10, 10)$ y el número de columnas después de desplazar: $nncolum$.

**DECLAR.** Se declaran las funciones DET, CRIT, POL4 y las subrutinas de IMSL UVMIF, EVLSF Y EVCSF.

**Paso 1** Se inicializa $nncolum$ a $ncolum + 1$.

**Paso 2** Se obtiene la matriz producto W'W .

**Paso 3** Mediante la subrutina de IMSL EVCSF se calculan los autovalores autovectores de W'W .

**Paso 4** Mediante el contador $kautv$, se eligen los autovalores que superan la cota mínima $cotaut \cdot nfilas$.

**Paso 5** Con los autovalores en orden creciente, se obtiene los autovectores correspondiente, y con ello, los factores.

**Paso 6** Se obtiene la matriz de pesos inicial.

**Paso 7** Se obtiene la matriz de correlaciones entre factores inicial.

**RETURN**

Cuando se ha utilizado en PROCED el fichero de datos obtenido con GENALTO con los 100 conjuntos de elementos, hemos realizado los cálculos para distintas cotas impuestas a los autovalores de $\mathbf{W}'\mathbf{W}$ que, como se sabe, limitará en mayor o menor medida el número de factores a extraer en el modelo Q-factorial. Así, la Tabla 2.12 nos proporciona el porcentaje de elementos bien clasificados obtenido según el tanto por ciento de variabilidad mínimo que se ha exigido a cada factor para ser elegido en el modelo Q-factorial

| % EXIGIDO | 1 | 5 | 10 | 15 | 20 | 25 | 30 |
|---|---|---|---|---|---|---|---|
| % BIEN CLASIF. | 83.42 | 85.24 | 85.86 | 85.12 | 84.36 | 82.88 | 82.88 |

Tabla 2.12: Porcentaje de elementos bien clasificados de la muestra generada con GENALTO, según el porcentaje mínimo de variabilidad exigido a cada factor elegido en el modelo.

Como puede verse, el porcentaje mayor de elementos bien clasificados se obtiene cuando se toman los factores de forma que sean capaces de explicar al menos el 10% de la variabilidad, representada por los autovalores de la matriz $\mathbf{H}$. Este resultado es el esperado ya que, por un lado, si se toma un valor más bajo el número de factores elegidos aumenta y ello hace que, al rotarlos, aparezcan más direcciones de las que debería haber y, por tanto, más dispersión. Por otro, si se toma un valor más alto, pueden despreciarse direcciones antes de la rotación que sean significativas a la hora de formar grupos.

Por otra parte, este procedimiento puede también utilizarse para determinar el número de grupos en un conjunto, si éste es desconocido. Si analizamos las direcciones (positivas o negativas) significativas que han intervenido en cada uno de los 100 conjuntos para determinar las clasificaciones, podemos enfrentar en la Tabla 2.13 los parámetros número de grupos inicial con el número de direcciones significativas:

| | Nº de grupos inicial | | | | |
|---|---|---|---|---|---|
| Nº dir. | 2 | 3 | 4 | 5 | 6 |
| 2 | 21 | 25 | 4 | 5 | |
| 3 | | 15 | 12 | 2 | 2 |
| 4 | | | 5 | 4 | |
| 5 | | | | 1 | |

Tabla 2.13: Número de grupos inicial frente a número de direcciones significativas obtenidas en los 100 conjuntos generados por GENALTO.

Como puede verse, el procedimiento determina, en general, menos grupos de los que hay originalmente. La diagonal principal de la matriz formada con las 4 primeras columnas tiene por traza el número de conjuntos en los que el número de grupos inicial y el de direcciones obtenidas coinciden. En el 46% de los casos se ha obtenido el mismo número. Por otra parte, los elementos de la diagonal secundaria por encima de la principal representan aquellos conjuntos en los que el número de direcciones obtenidas ha sido una menos que el número de grupos inicial, obteniendo el 41% de estos casos. Resumiendo, en el 87% de los casos, se han obtenido el mismo número de grupos iniciales ó uno menos.

### 2.4.1 Comparación con otros procedimientos de optimización de Análisis Cluster

Además del estudio de simulación anterior, hemos comparado los resultados anteriores con los que se obtienen con uno de los métodos de optimización del Análisis Cluster mencionados en la sección **2.2**. Pensamos que, aunque la validez del método PROCED puede determinarse dado que la muestra que analizamos con él ya está clasificada, sin embargo es aconsejable que sus resultados sean comparados con los obtenidos con los mismos datos por otros procedimientos de clasificación existentes.

El programa MINTRAZ del Apéndice B es capaz de obtener la mejor clasificación en grupos de un conjunto, dado el número de grupos, con el

criterio de minimizar la traza de la matriz **W** de dispersión dentro de los grupos dada en (2.1).

**Diagrama de Flujo del algoritmo MINTRAZ**

**DIMENS.** Se dimensionan generosamente la matriz total de elementos: $t(10000, 10)$ y la matriz temporal con los elementos de cada conjunto: $a(200, 10)$.

**DIMENS.** Se dimensionan vector de agrupación inicial: $grupin(100)$, agrupación temporal: $in(100)$ y agrupación final: $lug(100)$.

**DIMENS.** Se dimensionan las matrices suma: $suma(20, 50)$ y suma de cuadrados: $suma2(20, 50)$ para cada grupo y dimensión.

**DIMENS.** Se dimensionan el vector de trazas de cada grupo: $trwv(20)$.

**DIMENS.** Se dimensionan los vectores, para cada dimensión, de las sumas y sumas de cuadrados viejas y nuevas de los posibles cambios: $sv(50)$, $sv2(50)$, $sn(50)$ y $sn2(50)$.

**DIMENS.** Se dimensionan el vector temporal del número de elementos de cada grupo: $kgrupo(20)$ y la matriz para contar el número de elementos bien clasificados: $matbu(100, 100)$.

**ENTRADA** Se introduce la matriz total de datos con los 100 conjuntos generados por GENALTO.

**SALIDA** Se escriben los resultados que se describen a continuación para cada conjunto de datos.

**INICIO** Se inicia el contador de elementos totales bien clasificados y el de reordenaciones a 0.

**PARA** $l$ recorriendo de 1 a 100:

**Paso 1** Se obtiene la matriz y parámetros temporales: $l, nfilas, ncolum, ngrup, a(i, j)$ y $in(i)$,

**Paso 2** Se inicializan a 0 la suma, suma de cuadrados y número de elementos de cada grupo: $suma(i)$, $suma2(i)$, $kgrupo(i)$,

**Paso 3** Se calculan la suma, suma de cuadrados y número de elementos de cada grupo iniciales: $suma(i)$, $suma2(i)$, $kgrupo(i)$,

**Paso 4** Se calculan las trazas iniciales de la matriz de dispersión dentro de cada grupo y la total: $trwv(i)$, $trw$,

**Paso 5** Se inicializan a cero sumas y sumas de cuadrados temporales: $sv(i)$, $sv2(i)$, $sn(i)$, $sn2(i)$,

**Paso 6** Se iniciliza a 0 un contador que indica cuantos elementos del conjunto se han intentado mover sin éxito desde el último cambio: $iban$,

**Paso 7** Se inicializa a 1 un contador de elementos del conjunto que los recorrerá de forma cíclica: $icon$,

**Paso 8** MIENTRAS $(iban < nfilas)$, HACER Pasos 8.1 a 8.6:

**Paso 8.1** Se inicializa a 1 un contador de incremento del número del grupo donde está el elemento, que recorrerá los demás grupos de forma cíclica: $ig$,

**Paso 8.2** Se obtiene el número del grupo inicial del elemento: $igv = in(icon)$,

**Paso 8.3** SI $(kgrup(igv) > 1)$ ENTONCES HACER Pasos 8.3.1 y 8.3.2:

**Paso 8.3.1** Se calculan suma, suma de cuadrados y traza de la matriz de dispersión del grupo donde está el elemento al quitarlo de él: $sv(i)$, $sv2(i)$ y $tngv$,

**Paso 8.3.2** MIENTRAS $(ig < ngrup$ y $iban > 0)$, HACER Pasos 8.3.2.1 a 8.3.2.3:

**Paso 8.3.2.1** Se calcula el número de grupo donde se prueba a colocar, obtenido de forma cíclica: $ign = igv + ig$ y SI $(ign > ngrup)$ ENTONCES $ign = ign - ngrup$,

**Paso 8.3.2.2** Se calculan suma, suma de cuadrados y traza de la matriz de dispersión del grupo donde se intenta cambiar, añadiéndole el elemento: $sn(i)$, $sn2(i)$ y $tngn$,

**Paso 8.3.2.3** SI (la traza total es menor) ENTONCES HACER Pasos 8.3.2.3.1 a 8.3.2.3.4, SI NO HACER Paso 8.3.2.3.5:

**Paso 8.3.2.3.1** Se pone $iban$ a 0,

**Paso 8.3.2.3.2** Se redefinen trazas, sumas y sumas de cuadrados con el elemento cambiado,

**Paso 8.3.2.3.3** Se cambia el índice del grupo por el del nuevo grupo,

**Paso 8.3.2.3.4** Se actualiza el número de elementos de cada grupo.

**Paso 8.3.2.3.5** Se incrementa en 1 el contador de incremento de grupo: $ig = ig + 1$.

**Paso 8.4** Grupo final donde está el elemento: $lug(icon) = in(icon)$,

**Paso 8.5** Incrementamos en 1 el contador de elementos para recorrerlos de forma cíclica: $icon = icon+1$ y SI $(icon > nfilas)$ ENTONCES $icon = icon - nfilas$,

**Paso 8.6** Incrementamos en 1 el contador de elementos que llevamos desde el último cambio. Si se recorren todos los elementos sin ningún cambio, termina.

**Paso 9** Se escriben en SALIDA elementos, agrupación inicial y agrupación final.

**Paso 10** Se calcula y escribe la traza de la matriz de dispersión dentro de los grupos de la ordenación final.

**Paso 11** Se obtiene la matriz $matbu(i, j)$ que cuenta el número de elementos en cada número de grupo inicial y final,

**Paso 12** Se cuenta el número de elementos bien clasificados en el conjunto, asignando a cada grupo el grupo final donde encuentra más elementos, escribiendo en SALIDA el valor de buenos.

**TERMINA** Finaliza el bucle de todos los conjuntos

**CUENTA** Cuenta y escribe en SALIDA los elementos totales bien clasificados y reordenaciones.

**FIN**

Los resultados obtenidos cuando se ha utilizado MINTRAZ con la misma muestra de conjuntos de datos que se utilizó en PROCED, teniendo en cuenta

que en el procedimiento de minimizar la traz(**W**) se parte del conocimiento del número correcto de grupos en cada conjunto, y de la clasificación original que traen los conjuntos al ser generados por GENALTO, muestra que MINTRAZ no ha sido capaz de clasificar bien al 100 % de los elementos como podría esperarse, sino que ha logrado 4285 elementos bien clasificados del total de los 5000, es decir, un 87.7 %. Además, hemos comprobado que el algoritmo ha cambiado la agrupación original, al menos en un elemento, en 64 de los 100 conjuntos, lo cual nos da una idea de lo poco eficiente que resulta este algoritmo.

Si comparamos éstos resultados con los obtenidos con PROCED, vemos que MINTRAZ se comporta de forma parecida cuando en aquél se tomaban los autovalores capaces de explicar al menos el 10 % de la dispersión de los datos, donde se obtenían el 85.83 % de datos bien clasificados. Pero además, hay que tener en cuenta, por un lado, que PROCED tiene la ventaja sobre MINTRAZ de que no necesita el conocimiento previo del número de grupos para obtener este porcentaje de clasificaciones y, por otro lado, que PROCED puede partir de cualquier ordenación inicial de los elementos para obtener los resultados que ha obtenido.

### 2.4.2 Estudio de simulación controlando el número de grupos

Por último, hemos realizado con PROCED un estudio de simulación con varias muestras en las que se ha controlado el número de grupos original de cada conjunto. Queremos ver con él en qué medida puede influir este parámetro en cuanto al número de elementos bien clasificados que pueden obtenerse con él. En la Tabla 2.13 veíamos que mientras aumentaba el número de grupos original, el número de direcciones que clasifican obtenidas con PROCED no aumenta en la misma medida, y pretendemos ver si éste es un factor a tener en cuenta cuando se vaya a utilizar el procedimiento.

Haciendo los cambios apropiados en GENALTO, se han generado 5 muestras de 100 conjuntos, de forma que cada una de ellas contiene sólo conjuntos de 50 elementos con 2, 3, 4, 5 y 6 grupos, respectivamente para cada muestra,

y por otra parte se ha tenido en cuenta que en cada caso el número de grupos no excediera a la dimensión de los datos más uno. A cada una de estas muestras se les ha aplicado PROCED, variando también el porcentaje mínimo de variabilidad explicada por cada dirección obtenida, y al mismo tiempo el procedimiento MINTRAZ, de forma que los porcentajes de elementos bien clasificados figuran en la Tabla 2.14.

| | PROCED % VARIABILIDAD EXIGIDA | | | | | | | | MINTRAZ |
|---|---|---|---|---|---|---|---|---|---|
| % BIEN CLASIF. | 1 | 5 | 10 | 15 | 20 | 25 | 30 | 35 | |
| Con 2 grupos | 78.72 | 81.50 | 86.66 | 87.86 | 89.98 | 91.30 | 93.24 | 92.72 | 94.84 |
| Con 3 grupos | 84.70 | 88.02 | 88.72 | 88.08 | 87.60 | 87.68 | 85.00 | 66.84 | 85.94 |
| Con 4 grupos | 88.16 | 88.70 | 87.48 | 85.32 | 82.92 | 80.00 | 76.60 | 76.38 | 80.14 |
| Con 5 grupos | 84.89 | 84.56 | 83.34 | 81.34 | 77.98 | 73.12 | 70.80 | 69.02 | 72.26 |
| Con 6 grupos | 87.04 | 84.52 | 81.56 | 78.60 | 72.44 | 66.16 | 64.36 | 55.34 | 65.24 |

Tabla 2.14: Porcentaje de elementos bien clasificados de las muestras generadas con GENALTO con un número constante de grupos, obtenidos con PROCED, para distintos valores del tanto por ciento de variabilidad exigida a los autovalores, y MINTRAZ.

El análisis de la Tabla 2.14 nos proporciona algunos resultados interesantes. Por un lado, se aprecia que en general se obtienen mejores clasificaciones cuando los conjuntos tienen menos grupos (2, 3 ó 4), alcanzándose para el caso de 2 grupos hasta un 93.24 % de elementos bien clasificados. Parece claro que PROCED es capaz de obtener mejor las direcciones que clasifican a los elementos cuando el número de grupos no es muy grande, lo cual parece lógico.

Por otro lado, el porcentaje de variabilidad mínima explicada por las direcciones puede tomarse alto cuando el número de grupos es menor que cuando éste es más de 3. Parece también lógico que sea necesario elegir más direcciones al principio, que después serán rotadas, cuando el número de grupos es mayor (4, 5 ó 6).

Por otra parte, pensamos que es también evidente en los dos resultados comentados anteriormente, la influencia del método de rotación oblicua empleado. Quizás la utilización de otros métodos como el mismo Oblimín Directo para otros valores de $\delta$ que no sean 0, puede obtener direcciones

más oblicuas entre sí, lo cual puede suponer un cambio importante en los resultados obtenidos anteriormente.

Finalmente, y para tener aquí también una referencia respecto a otro método de clasificación, los resultados obtenidos con MINTRAZ muestran un claro paralelismo con los de PROCED cuando cambia el número de grupos, siendo en todos los casos, excepto con 2 grupos, peores. Este hecho reafirma la idea de que MINTRAZ no es un buen procedimiento de clasificación por sí mismo, y que en la mayoría de los casos PROCED obtiene mejores resultados.

# Capítulo 3

# Una generalización de PROCED

## 3.1 Introducción

En el capítulo anterior se ha introducido un procedimiento que permite descubrir utilizando la DVS, la mejor agrupación existente en un conjunto de datos, donde se sospecha la existencia de más de un grupo y este número no supera al número de variables más uno.

En éste, se pretende analizar algunos aspectos del procedimiento para ver si podemos obtener generalizaciónes de él. Se trata de ver si puede quitarse la restricción sobre el número de grupos en cuanto a su cota superior, que hasta ahora era el número de variables más uno ($g \leq p + 1$).

En la sección **3.2**, introduciremos un algoritmo que generaliza al propuesto cuando se sabe que existen al menos dos grupos en el conjunto a analizar. El procedimiento se ilustrará con un ejemplo y, posteriormente, en la sección **3.3** se hará una simulación para estudiar su comportamiento.

## 3.2 El procedimiento PROCGEN

El hecho de que el número de grupos $g$ pueda ser superior a $p+1$, impide que en algunos casos obtengamos en la descomposición de la matriz de datos, previamente centrada y desplazada en una dimensión más, un número suficiente de factores que representen a la totalidad de los grupos, aunque esto no quiere decir que puedieran aparecer en el análisis factores que determinen dos grupos, uno en la dirección positiva del autovector que lo determina y el otro en la negativa. Cuando esto no sea así, habrá que recurrir a un procedimiento iterativo en el que en cada paso, aquellos individuos que aparezcan fuertemente asociados a un factor sean eliminados del conjunto de datos para realizar nuevamente el análisis con los elementos que queden sin clasificar.

El procedimiento iterativo tendría, por tanto, una serie de pasos en los que se irían obteniendo en cada uno un factor que sea capaz de agrupar a una serie de individuos, hasta que todos quedaran agrupados en alguno.

Con todo, los pasos del procedimiento para cuando $g \geq p+1$ están descritos en el programa PROCGEN del Apéndice B, realizado también como los demás en FORTRAN 77 y con la ayuda de algunas subrutinas de IMSL. Las operaciones que realiza pueden resumirse en las siguientes:

1. En el primer paso, se realizan las operaciones siguientes:

   (a) Las transformaciones descritas en la sección anterior sobre la matriz $\mathbf{X}_{n,p}$ hasta llegar a la matriz $\mathbf{W}_{n,p+1}$, cuyas filas representan las coordenadas de las proyecciones de cada individuo en la hiperesfera de radio unidad de $R^{p+1}$.

   (b) Posteriormente, se realiza la **DVS** de $\mathbf{W}$, calculando la matriz de pesos para los $p+1$ factores, cuyas coordenadas son las de los autovalores de $\mathbf{W}'\mathbf{W}$ y se eligen los $q$ primeros que verifiquen algún criterio.

   (c) A continuación se rotan oblicuamente los factores elegidos utilizando el método de Jennrich y Sampson [77], el cual permitirá que al menos uno de los factores se acerque al centroide de los elementos que se agrupen con él, separando a éstos del resto.

2. Con los elementos que quedan en el conjunto de datos, se vuelven a realizar los pasos b) y c), de forma que al final se deberá tener este conjunto de datos mermado por otro grupo de elementos que se haya asociado fuertemente a un factor

3. Así se procede hasta que todos los elementos queden asociados a alguno de los factores que habrán ido apareciendo en alguno de los pasos anteriores.

**Diagrama de Flujo del algoritmo PROCGEN**

**DIMENS.** Se dimensionan generosamente la matriz total de elementos: $t(10000, 10)$, la matriz temporal con los elementos de cada conjunto: $a(100, 10)$, la matriz de estructuras factoriales: $s(100, 10)$, el vector de normas de los elementos: $xnor(100)$.

**DECLAR.** Se declaran variables enteras para ir eligiendo el conjunto de elementos temporal: $comi, conj, ncolum, nfilas, ngrup$.

**DIMENS.** Se dimensionan los vectores de número de elementos: $kmed(-20 : 20)$ y medias: $valmed(-20 : 20)$ de los índices obtenidos por las direcciones que clasifican.

**DIMENS.** Se dimensionan los vectores de agrupación inicial: $grupin(100)$, indice de las direcciones: $in(100)$, agrupación temporal: $ordin(100)$, agrupación final: $lug(100)$, matriz para contar el número de elementos bien clasificados: $matbu(100, 100)$ y matriz de número de grupos inicial y final: $grupos(10, 10)$.

**DECLAR.** Se declaran variables enteras para contar el número de elementos bien clasificados en cada conjunto: $buenos$ y el total: $buentotal$.

**COMÚN** Se dimensionan y declaran comunes en todo el algoritmo la matriz de elementos centrados por columnas: $c(100, 10)$, la matriz de elementos centrada por columnas, desplazada una dimensión más y normalizada por filas: $w(100, 10)$, la matriz producto W'W: $wpw(10, 10)$, el número de filas: $nfilas$, el número inicial de columnas: $ncolum$, el número de autovalores de W'W que se eligen: $kautv$, la distancia que

se desplazan los datos en una dimensión más: d y la cota inferior para elegir los autovalores de W'W: *cotaut*.

**COMÚN** Se dimensionan y declaran comunes con la subrutina ROT la matriz de pesos: $b(100, 10)$, la matriz de correlaciones entre los factores: $cor(10, 10)$, la matriz de autovectores de W'W: $avec(10, 10)$ y el número de columnas después de desplazar los elementos: *nncolum*.

**DECLAR.** Se declaran subrutinas y funciones externas definidas: para efectuar rotaciones: TRAT, para calcular det(W'W): DET, para el criterio a minimizar en las rotaciones: CRIT, para el polinomio a minimizar según el criterio de rotación: POL4 y las de IMSL para minimizar una función: UVMIF, para calcular los autovalores de una matriz cuadrada: EVLSF y para calcular los autovalores y autovectores de una matriz cuadrada: EVCSF

**ENTRADA** Se introduce la matriz total de datos con los 100 conjuntos generados por GENALTO y la cota mínima para elegir los autovalores.

**SALIDA** Se obtienen los resultados que se describen a continuación para cada conjunto de datos.

**INICIO** Se inicia el contador de elementos totales bien clasificados a 0.

**PARA** $l$ recorriendo de 1 a 100:

**Paso 1** Se obtiene la matriz y parámetros temporales: $l$, *ncolum*, *ngrup*, $a(i, j)$ y $grupin(i)$,

**Paso 2** Se centran los datos de $a(i, j)$ por columnas,

**Paso 3** Se obtiene la distancia $d$ con ayuda de DET y UVMIF,

**Paso 4** Se calcula W'W una vez desplazados los datos y el valor de $det(W'W)$,

**Paso 5** Se inicializa a 0 el número del grupo de los elementos que en cada paso se van a ir separando del total: *ki*.

**Paso 6** MIENTRAS ($nfilas > 1$) HACER Pasos 6.1 a 6.10:

**Paso 6.1** Se incrementa *ki* en 1,

**Paso 6.2** Se llama a la subrutina TRAT que nos dará el número de factores, los factores, la matriz de pesos inicial y la de correlaciones entre factores inicial,

**Paso 6.3** SI ($kautv > 1$), HACER Pasos 6.3.1 y 6.3.2:

**Paso 6.3.1** Se normaliza por filas la matriz de pesos inicial,

**Paso 6.3.2** Con la ayuda de CRIT y ROT, se realizan rotaciones con los factores iniciales hasta que el valor del criterio a minimizar converge,

**Paso 6.4** Se obtiene la matriz de pesos rotada,

**Paso 6.5** Se obtiene la matriz de correlaciones entre los factores rotados,

**Paso 6.6** Se obtiene la matriz de estructura factorial,

**Paso 6.7** Utilizando la matriz de estructura factorial, se obtiene el vector $in(i)$ de agrupación final para cada elemento, asociándole a cada uno la dirección del factor con el que tiene mayor valor absoluto en la matriz de estructuras,

**Paso 6.8** Recorriendo todas las direcciones, se busca aquélla cuyos elementos asociados tienen la mayor asociación media en valor absoluto,

**Paso 6.9** A los elementos asociados a la dirección anterior le asignamos el grupo $ki$ y los separamos del conjunto total,

**Paso 6.10** Los elementos que quedan se meten en la nueva matriz W.

**Paso 7** Se incrementan en 1 el elemento correspondiente de la matriz $grupos(ngrup, ki)$,

**Paso 8** Se escriben en SALIDA los datos, la agrupación inicial y la agrupación final,

**Paso 9** Se obtiene la matriz $matbu(i, j)$ que cuenta el número de elementos en cada número de grupo inicial y final,

**Paso 10** Se cuenta el número de elementos bien clasificados en el conjunto, asignando a cada grupo inicial el grupo final donde encuentra más elementos, escribiendo en SALIDA el valor de *buenos*.

**TERMINA** Finaliza el bucle de todos los conjuntos

**CUENTA** Cuenta y escribe elementos totales bien clasificados y la matriz de grupos iniciales y finales con la matriz *matbu*.

**FIN**

Es evidente que en el procedimiento anterior hay algunas decisiones que tomar en algunos momentos que habrá que comentar con más detalle.

La primera de ellas es qué criterio adoptar para elegir el número de factores. En el estudio de simulación realizado en el capítulo anterior, se vió cómo el mayor porcentaje de elementos bien clasificados lo proporcionaba elegir aquellos autovalores de H que superaran al menos el 10% del número de datos. Aquí seguiremos utilizando esta cota en principio, sin perjuicio de que más adelante podamos usar otras.

La segunda está en el apartado C) del primer paso, en el que después de rotar los factores, hay que separar del conjunto inicial aquellos elementos que se asocien "fuertemente" a un factor. Aunque pueden darse varias formas de medir el grado de asociación entre los individuos y los factores, seguiremos utilizando el mismo criterio que en PROCED. Dado que tanto los individuos como los factores están normalizados por el procedimiento, utilizaremos la matriz de estructura factorial que como se sabe determina la asociación entre cada individuo y cada factor mediante el coseno del ángulo que forman al representarlos en $I\!R^{p+1}$

Conviene que nos paremos en este punto a hacer algunos comentarios. Como acabamos de decir, el procedimiento se ha estructurado de forma que los elementos que surgen del análisis Q estén referidos a la hiperesfera de radio unidad. Este hecho está basado en la normalización por filas que se realiza antes de obtener la DVS de W. Si esta normalización no se hubiera llevado a efecto, el análisis Q de la matriz T no hubiera dado a los elementos proyectados en al hiperesfera unidad, sino que cada uno tendría su propia longitud dentro de $R^{p+1}$ y esto llevaría a que además de la estructura factorial se tendrían las longitudes de cada individuo para determinar las posibles agrupaciones. Pensamos que, como hemos venido haciendo hasta ahora, conviene que simplifiquemos lo más posible los elementos en los que nos basaremos para determinar la proximidad entre los individuos. De todas

formas, podría investigarse si la cuestión de considerar también la longitud puede ampliar el abanico de casos que el procedimiento puede resolver.

Así pues, entre los factores que se obtengan con un porcentaje de al menos un 10% de información, se elegirá aquél que más fuertemente asociados tenga a los individuos. Una forma de determinar esta fortaleza puede ser determinar para cada factor un parámetro determinado por el valor medio de las asociaciones que tienen con el factor todos los individuos asociados a él. Estas asociaciones, dadas por los elementos de la estructura, se toman en valor absoluto. Así, si la matriz de estructura factorial para los q primeros factores ($q \leq p+1$) una vez rotados es

$$\mathbf{S} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1q} \\ s_{21} & s_{22} & \cdots & s_{2q} \\ \cdots & \cdots & \cdots & \cdots \\ s_{n1} & s_{n2} & \cdots & s_{nq} \end{pmatrix}$$

en cada fila $i$ el valor absoluto máximo de los $s_{ik}$, $; k = 1, \cdots, q$ determinará el factor al que el individuo se asocia. Si al factor $f_k$ se asocian los individuos $I_1, \cdots, I_{n_k}$, una medida de la fortaleza de tal asociación conjunta puede ser la media:

$$\mid \overline{s}_k \mid = \frac{1}{n_k} \sum_{l=1}^{n_k} \mid s_{lk} \mid$$

Otras medidas de tal fortaleza pueden ser también el máximo o el mínimo de todos los valores $\mid s_{lk} \mid$, $; l = 1, \cdots, n_k$:

$$\max_l \mid s_{lk} \mid$$

$$\min_l \mid s_{lk} \mid$$

### 3.2.1 Ejemplo

Analizaremos con el procedimiento iterativo PROCGEN un conjunto de datos simulado, difícil de clasificar si no se desplazan los datos en una dimensión más, en el que se conoce de antemano el número de grupos y éste es mayor

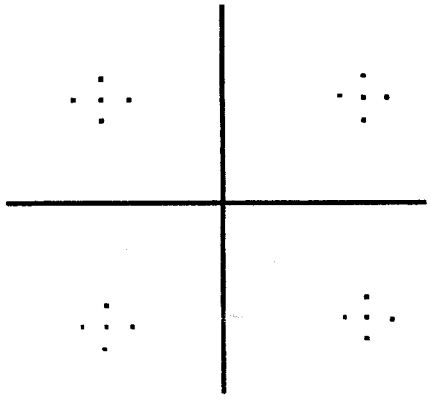

Figura 3.1: Datos con cuatro grupos en dimensión 2

que el número de variables más uno. Tomaremos los datos representados en la Figura 3.1 dados en la Tabla 3.1

| $x_1$ | $x_2$ |
|---|---|
| 2.0 | 2.0 |
| 2.0 | 1.5 |
| 2.0 | 2.5 |
| 1.5 | 2.0 |
| 2.5 | 2.0 |
| 2.0 | -2.0 |
| 2.0 | -1.5 |
| 2.0 | -2.5 |
| 1.5 | -2.0 |
| 2.5 | -2.0 |
| -2.0 | 2.0 |
| -2.0 | 1.5 |
| -2.0 | 2.5 |
| -1.5 | 2.0 |
| -2.5 | 2.0 |
| -2.0 | -2.0 |
| -2.0 | -1.5 |
| -2.0 | -2.5 |
| -1.5 | -2.0 |
| -2.5 | -2.0 |

Tabla 3.1: Matriz **A** de datos de la Figura 3.1.

En este caso, las primeras operaciones de centralizacion y la búsqueda de la distancia de desplazamiento que maximice la dispersión nos proporcionan

una matriz $\mathbf{T}_1$, y una matriz $\mathbf{W}_1$ obtenida al normalizar los vectores fila.

$$T_1 = \begin{pmatrix} 2.00 & 2.00 & 1.99 \\ 2.00 & 1.50 & 1.99 \\ 2.00 & 2.50 & 1.99 \\ 1.50 & 2.00 & 1.99 \\ 2.50 & 2.00 & 1.99 \\ 2.00 & -2.00 & 1.99 \\ 2.00 & -1.50 & 1.99 \\ 2.00 & -2.50 & 1.99 \\ 1.50 & -2.00 & 1.99 \\ 2.50 & -2.00 & 1.99 \\ -2.00 & 2.00 & 1.99 \\ -2.00 & 1.50 & 1.99 \\ -2.00 & 2.50 & 1.99 \\ -1.50 & 2.00 & 1.99 \\ -2.50 & 2.00 & 1.99 \\ -2.00 & -2.00 & 1.99 \\ -2.00 & -1.50 & 1.99 \\ -2.00 & -2.50 & 1.99 \\ -1.50 & -2.00 & 1.99 \\ -2.50 & -2.00 & 1.99 \end{pmatrix} \quad \mathbf{W}_1 = \begin{pmatrix} .58 & .58 & .58 \\ .63 & .47 & .62 \\ .53 & .66 & .53 \\ .47 & .63 & .62 \\ .66 & .53 & .53 \\ .58 & -.58 & .57 \\ .63 & -.47 & .62 \\ .53 & -.66 & .53 \\ .47 & -.63 & .62 \\ .66 & -.53 & .53 \\ -.58 & .58 & .57 \\ -.63 & .47 & .62 \\ -.53 & .66 & .53 \\ -.47 & .63 & .62 \\ -.66 & .53 & .53 \\ -.58 & -.58 & .57 \\ -.63 & -.47 & .62 \\ -.53 & -.66 & .53 \\ -.47 & -.63 & .62 \\ -.66 & -.53 & .53 \end{pmatrix}$$

**Matrices $\mathbf{T}_1$ y $\mathbf{W}_1$ obtenidas al desplazar una dimensión más y normalizar por filas, respectivamente, los elementos de la Tabla 3.1 centrados por columnas.**

Los autovalores de la matriz $\mathbf{W}_1'\mathbf{W}_1$ vienen dados en la Tabla 3.2,

| Nº orden | Autovalor | Porc. inform. | Porc. inf. acumulada |
|---|---|---|---|
| $\lambda_1$ | 6.667 | 33.34 | 33.34 |
| $\lambda_2$ | 6.666 | 33.33 | 67.67 |
| $\lambda_3$ | 6.666 | 32.33 | 100.00 |

Tabla 3.2: Autovalores de $\mathbf{W}_1'\mathbf{W}_1$ y porcentaje de información que representan para los datos de la Tabla 3.1.

y los autovectores en la Tabla 3.3.

Por tanto, las puntuaciones de los factores $\mathbf{G}$ vendrán dadas por los autovectores anteriores, mientras que la matriz de pesos factoriales $\mathbf{A}$, que de momento coincide con la de estructura factorial, vendrá dada por $\mathbf{A} = \mathbf{W}_1 \cdot \mathbf{G}$ en la Tabla 3.4.

| $u_1$ | $u_2$ | $u_3$ |
|---|---|---|
| 0.000 | 0.981 | -0.196 |
| 0.000 | 0.196 | 0.981 |
| 1.000 | 0.000 | 0.000 |

Tabla 3.3: Autovectores de $\mathbf{W}_1'\mathbf{W}_1$ correspondientes a los autovalores de la Tabla 3.2.

| Iris | Factor 1 | Factor 2 | Factor 3 |
|---|---|---|---|
| I1 | 0.57 | 0.68 | 0.45 |
| I2 | 0.62 | 0.70 | 0.34 |
| I3 | 0.53 | 0.65 | 0.55 |
| I4 | 0.62 | 0.58 | 0.52 |
| I5 | 0.53 | 0.75 | 0.39 |
| I6 | 0.57 | 0.45 | -0.68 |
| I7 | 0.62 | 0.52 | -0.58 |
| I8 | 0.53 | 0.39 | -0.75 |
| I9 | 0.62 | 0.34 | -0.70 |
| I10 | 0.53 | 0.54 | -0.65 |
| I11 | 0.57 | -0.45 | 0.68 |
| I12 | 0.62 | -0.52 | 0.58 |
| I13 | 0.53 | -0.39 | 0.75 |
| I14 | 0.62 | -0.34 | 0.70 |
| I15 | 0.53 | -0.54 | 0.65 |
| I16 | 0.57 | -0.68 | -0.45 |
| I17 | 0.62 | -0.70 | -0.34 |
| I18 | 0.53 | -0.65 | -0.54 |
| I19 | 0.62 | -0.58 | -0.52 |
| I20 | 0.53 | -0.75 | -0.39 |

Tabla 3.4: Matriz de pesos inicial en el modelo Q-factorial para los datos de la Tabla 3.1 transformados en $W_1$.

Pero como hemos dicho, las asociaciones dadas por matriz de estructura factorial suelen clarificarse más cuando se somete a los factores a una rotación oblicua. En este caso, después de tal rotación, la matriz de estructura factorial viene dada en la Tabla 3.5.

| Iris | Factor 1 | Factor 2 | Factor 3 |
|---|---|---|---|
| I1 | 0.58 | 0.82 | -0.00 |
| I2 | 0.62 | 0.77 | -0.11 |
| I3 | 0.53 | 0.84 | 0.08 |
| I4 | 0.62 | 0.77 | 0.11 |
| I5 | 0.53 | 0.84 | -0.09 |
| I6 | 0.58 | -0.00 | -0.82 |
| I7 | 0.62 | 0.11 | -0.77 |
| I8 | 0.53 | -0.09 | -0.84 |
| I9 | 0.62 | -0.11 | -0.77 |
| I10 | 0.53 | 0.09 | -0.84 |
| I11 | 0.58 | 0.00 | 0.82 |
| I12 | 0.62 | -0.11 | 0.77 |
| I13 | 0.53 | 0.09 | 0.84 |
| I14 | 0.62 | 0.11 | 0.77 |
| I15 | 0.53 | -0.09 | 0.84 |
| I16 | 0.58 | -0.82 | 0.00 |
| I17 | 0.62 | -0.77 | 0.11 |
| I18 | 0.53 | -0.84 | -0.08 |
| I19 | 0.62 | -0.77 | -0.11 |
| I20 | 0.53 | -0.84 | 0.09 |

Tabla 3.5: Matriz de estructuras después de rotar con tres factores en el modelo Q-factorial para los datos de la Tabla 3.1 transformados en $\mathbf{W_1}$.

Como puede verse, la simetría que presenta el conjunto de datos inicial hace que la elección del primer grupo de datos que se separe de total pueda ser cualquiera de los cuatro existentes. El valor medio de las asociaciones con los factores 2 ó 3 (en sentido positivo o negativo) es el mismo para los cuatro grupos : 0.808. El programa separa en primer lugar los elementos I6 a I10, asociados a la dirección negativa del factor 3.

Por su parte, con los elementos que quedan sin clasificar, puede volver a realizarse todo lo anteriormente expuesto. Obtenemos primero la nueva matriz $\mathbf{W}_2$.

$$\mathbf{W_2} = \begin{pmatrix} 0.58 & 0.58 & 0.58 \\ 0.63 & 0.47 & 0.62 \\ 0.53 & 0.66 & 0.53 \\ 0.47 & 0.62 & 0.62 \\ 0.66 & 0.53 & 0.53 \\ -0.58 & 0.58 & 0.58 \\ -0.63 & 0.47 & 0.62 \\ -0.53 & 0.66 & 0.53 \\ -0.47 & 0.62 & 0.62 \\ -0.66 & 0.53 & 0.53 \\ -0.58 & -0.58 & 0.58 \\ -0.62 & -0.47 & 0.62 \\ -0.53 & -0.66 & 0.53 \\ -0.47 & -0.62 & 0.62 \\ -0.66 & -0.53 & 0.53 \end{pmatrix}$$

Matriz $\mathbf{W_2}$ obtenida al desplazar una dimensión más y normalizar por filas los 15 elementos que quedan de la Tabla 3.1 centrados por columnas.

y los autovalores y autovectores de $\mathbf{W_2'W_2}$ están en las Tablas 3.6 y 3.7, respectivamente.

| Nº orden | Autovalor | Porc. inform. | Porc. inf. acumulada |
|---|---|---|---|
| $\lambda_1$ | 6.653 | 44.35 | 44.35 |
| $\lambda_2$ | 6.624 | 44.16 | 88.51 |
| $\lambda_3$ | 1.723 | 11.49 | 100.00 |

Tabla 3.6: Autovalores de $\mathbf{W_2'W_2}$ y porcentaje de información que representan.

| $u_1$ | $u_2$ | $u_3$ |
|---|---|---|
| -0.41 | 0.71 | 0.57 |
| 0.41 | 0.70 | -0.57 |
| 0.81 | 0.00 | 0.58 |

Tabla 3.7: Autovectores de $\mathbf{W_2'W_2}$ correspondientes a los autovalores de la Tabla 3.6.

La matriz de estructuras que nos interesa para clasificar, obtenida de rotar oblicuamente los pesos iniciales está en la Tabla 3.8.

| Iris | Factor 1 | Factor 2 | Factor 3 |
|---|---|---|---|
| I1 | 0.34 | 0.82 | 0.54 |
| I2 | 0.27 | 0.77 | 0.62 |
| I3 | 0.38 | 0.84 | 0.47 |
| I4 | 0.45 | 0.77 | 0.55 |
| I5 | 0.23 | 0.84 | 0.52 |
| I11 | 1.00 | -0.00 | 0.31 |
| I12 | 1.00 | -0.11 | 0.37 |
| I13 | 1.00 | 0.09 | 0.26 |
| I14 | 1.00 | 0.11 | 0.37 |
| I15 | 1.00 | -0.09 | 0.26 |
| I16 | 0.34 | -0.81 | 0.54 |
| I17 | 0.45 | -0.77 | 0.55 |
| I18 | 0.23 | -0.84 | 0.52 |
| I19 | 0.27 | -0.77 | 0.61 |
| I20 | 0.38 | -0.84 | 0.47 |

Tabla 3.8: Matriz de estructuras obtenida al rotar los pesos iniciales con los 15 elementos que quedan.

Es evidente que el grupo que surge ahora de este segundo paso es el de los elementos 10 al 15, que están fuertemente asociados al factor segundo.

Con los restantes elementos que aún permanecen sin clasificar puede volver a realizarse el análisis completo. Obtenemos la matriz $\mathbf{W}_3$,

$$\mathbf{W}_3 = \begin{pmatrix} 0.58 & 0.58 & 0.57 \\ 0.63 & 0.47 & 0.62 \\ 0.53 & 0.66 & 0.53 \\ 0.47 & 0.63 & 0.62 \\ 0.66 & 0.53 & 0.53 \\ -0.58 & -0.57 & 0.57 \\ -0.63 & -0.47 & 0.62 \\ -0.53 & -0.66 & 0.53 \\ -0.47 & -0.63 & 0.62 \\ -0.66 & -0.53 & 0.53 \end{pmatrix}$$

Matriz $\mathbf{W}_3$ obtenida al desplazar una dimensión más y normalizar por filas los 10 elementos que quedan de la Tabla 3.1 centrados por columnas.

los autovalores y dos primeros autovectores de $\mathbf{W}_3^t\mathbf{W}_3$ se dan en las Tablas 3.9 y 3.10,

| Nº orden | Autovalor | Porc. inform. | Porc. inf. acumulada |
|---|---|---|---|
| $\lambda_1$ | 6.582 | 65.82 | 65.82 |
| $\lambda_2$ | 3.334 | 33.34 | 99.16 |
| $\lambda_3$ | 0.084 | 0.84 | 100.00 |

Tabla 3.9: Autovalores de $\mathbf{W}_3'\mathbf{W}_3$ y porcentaje de información que representan.

| $u_1$ | $u_2$ |
|---|---|
| 0.707 | 0.000 |
| 0.707 | 0.000 |
| 0.000 | 1.000 |

Tabla 3.10: Autovectores de $\mathbf{W}_3'\mathbf{W}_3$ correspondientes a los dos primeros autovalores de la Tabla 3.9.

y la matriz de estructuras una vez rotados están en la Tabla 3.11,

| Iris | Factor 1 | Factor 2 |
|---|---|---|
| I1 | 0.82 | 0.57 |
| I2 | 0.77 | 0.62 |
| I3 | 0.84 | 0.53 |
| I4 | 0.77 | 0.62 |
| I5 | 0.84 | 0.53 |
| I16 | -0.82 | 0.58 |
| I17 | -0.77 | 0.62 |
| I18 | -0.84 | 0.53 |
| I19 | -0.77 | 0.62 |
| I20 | -0.84 | 0.53 |

Tabla 3.11: Matriz de estructuras obtenida al rotar los pesos iniciales con los 10 elementos que quedan.

que permite clasificar a cualquiera de los dos grupos, por ejemplo del I16 al I20.

Por último, los elementos que quedan del I1 al I5 pueden ser analizados de la misma forma. La matriz que forman es $\mathbf{W}_4$, los autovalores y autovectores significativos de $\mathbf{W}_4'\mathbf{W}_4$ están en las Tablas 3.12 y 3.13, y la matriz de estructuras una vez rotados en la Tabla 3.14.

$$\mathbf{W_4} = \begin{pmatrix} 0.58 & 0.58 & 0.57 \\ 0.62 & 0.47 & 0.62 \\ 0.53 & 0.66 & 0.53 \\ 0.47 & 0.63 & 0.62 \\ 0.66 & 0.53 & 0.53 \end{pmatrix}$$

**Matriz $\mathbf{W_4}$ obtenida al desplazar una dimensión más y normalizar por filas los 5 elementos que quedan de la Tabla 3.1 centrados por columnas.**

| Nº orden | Autovalor | Porc. inform. | Porc. inf. acumulada |
|---|---|---|---|
| $\lambda_1$ | 4.944 | 98.88 | 98.88 |

Tabla 3.12: Autovalores de $\mathbf{W_4'W_4}$ y porcentaje de información que representan.

| $u_1$ |
|---|
| 0.57 |
| 0.57 |
| 0.58 |

Tabla 3.13: Autovector de $\mathbf{W_4'W_4}$ correspondiente al autovalor de la Tabla 3.12.

| Iris | Factor 1 |
|---|---|
| I1 | 0.99 |
| I2 | 0.99 |
| I3 | 0.99 |
| I4 | 0.99 |
| I5 | 0.99 |

Tabla 3.14: Matriz de estructuras obtenida al rotar los pesos iniciales obtenidos con los 5 elementos últimos.

## 3.3 Estudio de simulación para PROCGEN. Comparación con otros métodos

Nuevamente, para evaluar el funcionamiento del procedimiento PROCGEN, se ha realizado un estudio de simulación, en principio con la misma muestra, a la que denominaremos MUESTRA1, que se utilizó para evaluar PROCED en la sección **2.4**.

Los resultados obtenidos en esta ocasión están en la Tabla 3.15, que nos proporciona el porcentaje de elementos bien clasificados según el tanto por ciento de variabilidad que exigimos a cada factor para ser elegido en el modelo Q-factorial.

| % EXIGIDO | 1 | 5 | 10 | 15 | 20 | 25 | 30 |
|---|---|---|---|---|---|---|---|
| % BIEN CLASIF. | 83.36 | 85.64 | 87.30 | 87.40 | 86.54 | 86.44 | 85.92 |

Tabla 3.15: Porcentaje de elementos bien clasificados de la MUESTRA1 generada con GENALTO, según el porcentaje mínimo de variabilidad exigido a cada factor elegido en el modelo.

Como puede verse comparando con la Tabla 2.12, el porcentaje mayor de elementos bien clasificados (87.30 %) ha aumentado respecto de los obtenidos con PROCED, obteniéndose cuando se toman los factores de forma que sean capaces de explicar al menos el 15% de la variabilidad, representada por los autovalores de la matriz H. Además, en este caso, este porcentaje es más estable en la banda entre el 5 y el 30 por ciento, lo cual hace que PROCGEN dependa en menor medida del porcentaje de variabilidad tomado. Este resultado es también el esperado si se tiene en cuenta que en PROCGEN se obtiene una optimización del método en cierto sentido, al ir eligiendo en cada paso la dirección más fuertemente asociada a alguno de los grupos.

Por otra parte, si utilizamos también PROCGEN para determinar el número de grupos en un conjunto, al enfrentar el número de direcciones finales que se han obtenido en cada uno de los 100 conjuntos con el número

| | Nº de grupos inicial | | | | |
|---|---|---|---|---|---|
| Nº dir. | 2 | 3 | 4 | 5 | 6 |
| 2 | 15 | 18 | 2 | 4 | |
| 3 | 2 | 16 | 10 | 3 | |
| 4 | 1 | 4 | 9 | 4 | 5 |
| 5 | 3 | | | 1 | 1 |
| 6 | | 1 | | | |
| 7 | | | | | |
| 8 | | 1 | | | |

Tabla 3.16: Número de grupos inicial frente a número de direcciones significativas obtenidas en los 100 conjuntos generados por GENALTO.

de grupos inicial, se obtiene la Tabla 3.16 para el caso de tomar los factores que expliquen al menos el 15% de la variabilidad:

Como puede verse, el procedimiento también determina en general menos grupos de los que hay originalmente, al igual que ocurría con PROCED. Analizando más detenidamente, hay un 41% de conjuntos en los que se determina el mismo número original, y un 39% en los que se determinan uno más ó uno menos que el número original, en total, un 80%.

Pero el procedimiento de clasificación PROCGEN se ha construido para que el número de grupos del conjunto no tenga la limitación de la dimensión de los datos más 1. Por ello, con una ligera modificación del programa GENALTO, puede generarse una segunda muestra en la que no exista esta restricción. De todas formas tampoco se permite que el número de grupos sea ilimitado, ya que como cada conjunto tiene 50 elementos, éste número podría ser una cota, aunque se ha puesto un tope máximo de 10 grupos, cualquiera que sea la dimensión del espacio donde están los datos, que sigue siendo entre 1 y 5.

Los resultados obtenidos con esta nueva muestra son ligeramente peores a los obtenidos con la primera, y se presentan en la Tabla 3.17, y el número de grupos obtenidos en cada conjunto enfrentados a los originales aparecen

| % EXIGIDO | 1 | 5 | 10 | 15 | 20 |
|---|---|---|---|---|---|
| % BIEN CLASIF. | 79.88 | 81.22 | 80.34 | 78.90 | 76.84 |

Tabla 3.17: Porcentaje de elementos bien clasificados de la MUESTRA2 generada con GENALTO, según el porcentaje mínimo de variabilidad exigido a cada factor elegido en el modelo.

| | Nº de grupos inicial | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Nº dir. | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 2 | 5 | 6 | | | | | | | |
| 3 | 1 | 6 | 3 | 6 | 2 | 2 | 1 | 3 | 3 |
| 4 | 1 | | 5 | 4 | 3 | 4 | 2 | 1 | 2 |
| 5 | 2 | | | 3 | 3 | 5 | 2 | 2 | 1 |
| 6 | 1 | | | | | 2 | 2 | | 2 |
| 7 | | | 1 | | | 1 | 1 | 3 | 1 |
| 8 | | 2 | | | | | | 1 | 1 |
| 9 | 1 | | | | | | | | |
| 10 | 2 | | | | | | | | |

Tabla 3.18: Número de grupos inicial frente a número de direcciones significativas obtenidas con los 100 conjuntos de la MUESTRA2 generados por GENALTO.

en la Tabla 3.18.

Al igual que se hizo en la sección **2.4**, compararemos ahora los resultados anteriores con la clasificación que de la MUESTRA2 ha sido capaz de realizar el procedimiento MINTRAZ. El porcentaje de elementos bien clasificados para esta nueva muestra es de 4011, lo que supone un 80.22 % que está en consonancia con los obtenidos en la Tabla 3.17.

Por último, y siguiendo también los pasos de la sección **2.4**, se han vuelto a generar, modificando ligeramente GENALTO, varias muestras en las que se ha controlado el número de grupos original de cada conjunto. En este caso, las posibilidades de PROCGEN nos permiten generar muestras de conjuntos con 2, 3, 4, 5 y 6 grupos, sin que éste número limite la dimensión de los datos. A cada una de las muestras generadas se les ha aplicado PROCGEN, variando también el porcentaje mínimo de variabilidad explicada por cada

| | PROCGEN % VARIABILIDAD EXIGIDA | | | | | | | MINTRAZ |
|---|---|---|---|---|---|---|---|---|
| % BIEN CLASIF. | 1 | 5 | 10 | 15 | 20 | 25 | 30 | |
| Con 2 grupos | 77.86 | 80.72 | 85.42 | 89.36 | 90.24 | 92.56 | 93.50 | 94.56 |
| Con 3 grupos | 87.26 | 88.68 | 89.42 | 88.94 | 88.26 | 88.04 | 86.90 | 90.90 |
| Con 4 grupos | 84.54 | 85.70 | 86.00 | 84.92 | 83.92 | 82.78 | 81.08 | 85.50 |
| Con 5 grupos | 84.78 | 86.30 | 84.76 | 82.36 | 80.16 | 76.78 | 73.94 | 82.54 |
| Con 6 grupos | 85.66 | 85.54 | 83.74 | 81.32 | 77.68 | 73.98 | 71.68 | 75.38 |

Tabla 3.19: Porcentaje de elementos bien clasificados de las muestras generadas con GENALTO con un número constante de grupos, obtenidos con PROCGEN y MINTRAZ.

dirección obtenida, y MINTRAZ, de forma que los porcentajes de elementos bien clasificados figuran en la Tabla 3.19, donde se pone nuevamente de manifiesto que también PROCGEN clasifica mejor a conjuntos con menos grupos, que a medida que el número de grupos aumenta es necesario exigir menos porcentaje de variabilidad explicada por los autovalores de H, y que el procedimiento MINTRAZ sólo mejora al PROCGEN cuando se tiene 2 ó 3 grupos en los conjuntos de datos.

# Algunos problemas abiertos

Son varias las problemas abiertos que quedan pendientes al terminar esta Memoria. Sólo queremos en este apartado enumerarlas, abriendo la posibilidad de que puedan ser objeto de futuros análisis.

En primer lugar, pensamos que pueden depurarse aún más los dos procedimientos propuestos, tratando de analizar cómo se comportan con otras muestras de conjuntos con grupos en los que se controlen más factores como:

- La forma de los grupos, generándolos mediante otros modelos aleatorios distinto del uniforme, por ejemplo el multinormal.
- La distancia entre los centros de los grupos.
- El número de elementos total en el conjunto y el de los elementos en cada grupo.
- La dimensión del espacio inicial de los elementos.

En segundo lugar, pueden estudiarse para los dos procedimientos algúnos algoritmos de parada que sean capaces de detectar la presencia de conjuntos sin grupos.

En tercer lugar, pueden utilizarse en los procedimientos otras técnicas de rotación oblicua, algunas de las cuales están descritas en la sección **1.2**. Incluso podría intentarse construir una técnica de rotación oblicua propia para los procedimientos.

En cuarto lugar, puede abordarse el estudio de los dos procedimientos sin normalizar por filas la matriz de datos desplazada en una dimensión más. En ese caso, las asociaciones de la matriz $\mathbf{TT}'$ no estarían comprendidas entre -1 y 1, al depender éstas de las normas de los vectores fila de $\mathbf{T}$, pero podrían adoptarse criterios para buscar los grupos que tuvieran en cuenta este hecho.

Por último, la diversidad de procedimientos de optimización y obtención del número de grupos de un conjunto de elementos, mencionados en la sección **2.2**, permiten realizar discusiones al compararlos con los propuestos en la Memoria, al igual que se ha hecho con MINTRAZ.

# Apéndice A

# Datos utilizados por R.A.Fisher

Datos recogidos por Fisher [46] sobre 150 especímenes de plantas de Iris, de las que se midieron las longitudes y anchuras de los sépalos (LS y AS) y de los pétalos (LP y AP). Fueron utilizadas 50 plantas de cada uno de los tres tipos de Iris Setosa (1), Versicolor (2) y Virgínica (3).

| LS | AS | LP | AP | TIPO |
|---|---|---|---|---|
| 50 | 33 | 14 | 02 | 1 |
| 64 | 28 | 56 | 22 | 3 |
| 65 | 28 | 46 | 15 | 2 |
| 67 | 31 | 56 | 24 | 3 |
| 63 | 28 | 51 | 15 | 3 |
| 46 | 34 | 14 | 03 | 1 |
| 69 | 31 | 51 | 23 | 3 |
| 62 | 22 | 45 | 15 | 2 |
| 59 | 32 | 48 | 18 | 2 |
| 46 | 36 | 10 | 02 | 1 |
| 61 | 30 | 46 | 14 | 2 |
| 60 | 27 | 51 | 16 | 2 |

65 30 52 20 3
56 25 39 11 2
65 30 55 18 3
58 27 51 19 3
68 32 59 23 3
51 33 17 05 1
57 28 45 13 2
62 34 54 23 3
77 38 67 22 3
63 33 47 16 2
67 33 57 25 3
76 30 66 21 3
49 25 45 17 3
55 35 13 02 1
67 30 52 23 3
70 32 47 14 2
64 32 45 15 2
61 28 40 13 2
48 31 16 02 1
59 30 51 18 3
55 24 38 11 2
63 25 50 19 3
64 32 53 23 3
52 34 14 02 1
49 36 14 01 1
54 30 45 15 2
79 38 64 20 3
44 32 13 02 1
67 33 57 21 3
50 35 16 06 1
58 26 40 12 2
44 30 13 02 1
77 28 67 20 3
63 27 49 18 3
47 32 16 02 1
55 26 44 12 2
50 23 33 10 2

| | | | | |
|---|---|---|---|---|
| 72 | 32 | 60 | 18 | 3 |
| 48 | 30 | 14 | 03 | 1 |
| 51 | 38 | 16 | 02 | 1 |
| 61 | 30 | 49 | 18 | 3 |
| 48 | 34 | 19 | 02 | 1 |
| 50 | 30 | 16 | 02 | 1 |
| 50 | 32 | 12 | 02 | 1 |
| 61 | 26 | 56 | 14 | 3 |
| 64 | 28 | 56 | 21 | 3 |
| 43 | 30 | 11 | 01 | 1 |
| 58 | 40 | 12 | 02 | 1 |
| 51 | 38 | 19 | 04 | 1 |
| 67 | 31 | 44 | 14 | 2 |
| 62 | 28 | 48 | 18 | 3 |
| 49 | 30 | 14 | 02 | 1 |
| 51 | 35 | 14 | 02 | 1 |
| 56 | 30 | 45 | 15 | 2 |
| 58 | 27 | 41 | 10 | 2 |
| 50 | 34 | 16 | 04 | 1 |
| 46 | 32 | 14 | 02 | 1 |
| 60 | 29 | 45 | 15 | 2 |
| 57 | 26 | 35 | 10 | 2 |
| 57 | 44 | 15 | 04 | 1 |
| 50 | 36 | 14 | 02 | 1 |
| 77 | 30 | 61 | 23 | 3 |
| 63 | 34 | 56 | 24 | 3 |
| 58 | 27 | 51 | 19 | 3 |
| 57 | 29 | 42 | 13 | 2 |
| 72 | 30 | 58 | 16 | 3 |
| 54 | 34 | 15 | 04 | 1 |
| 52 | 41 | 15 | 01 | 1 |
| 71 | 30 | 59 | 21 | 3 |
| 64 | 31 | 55 | 18 | 3 |
| 60 | 30 | 48 | 18 | 3 |
| 63 | 29 | 56 | 18 | 3 |
| 49 | 24 | 33 | 10 | 2 |
| 56 | 27 | 42 | 13 | 2 |

57 30 42 12 2
55 42 14 02 1
49 31 15 02 1
77 26 69 23 3
60 22 50 15 3
54 39 17 04 1
66 29 46 13 2
52 27 39 14 2
60 34 45 16 2
50 34 15 02 1
44 29 14 02 1
50 20 35 10 2
55 24 37 10 2
58 27 39 12 2
47 32 13 02 1
46 31 15 02 1
69 32 57 23 3
62 29 43 13 2
74 28 61 19 3
59 30 42 15 2
51 34 15 02 1
50 35 13 03 1
56 28 49 20 3
60 22 40 10 2
73 29 63 18 3
67 25 58 18 3
49 31 15 01 1
67 31 47 15 2
63 23 44 13 2
54 37 15 02 1
56 30 41 13 2
63 25 49 15 2
61 28 47 12 2
64 29 43 13 2
51 25 30 11 2
57 28 41 13 2
65 30 58 22 3

| | | | | |
|---|---|---|---|---|
| 69 | 31 | 54 | 21 | 3 |
| 54 | 39 | 13 | 04 | 1 |
| 51 | 35 | 14 | 03 | 1 |
| 72 | 36 | 61 | 25 | 3 |
| 65 | 32 | 51 | 20 | 3 |
| 61 | 29 | 47 | 14 | 2 |
| 56 | 29 | 36 | 13 | 2 |
| 69 | 31 | 49 | 15 | 2 |
| 64 | 27 | 53 | 19 | 3 |
| 68 | 30 | 55 | 21 | 3 |
| 55 | 25 | 40 | 13 | 2 |
| 48 | 34 | 16 | 02 | 1 |
| 48 | 30 | 14 | 01 | 1 |
| 45 | 23 | 13 | 03 | 1 |
| 57 | 25 | 50 | 20 | 3 |
| 57 | 38 | 17 | 03 | 1 |
| 51 | 38 | 15 | 03 | 1 |
| 55 | 23 | 40 | 13 | 2 |
| 66 | 30 | 44 | 14 | 2 |
| 68 | 28 | 48 | 14 | 2 |
| 54 | 34 | 17 | 02 | 1 |
| 51 | 37 | 15 | 04 | 1 |
| 52 | 35 | 15 | 02 | 1 |
| 58 | 28 | 51 | 24 | 3 |
| 67 | 30 | 50 | 17 | 2 |
| 63 | 33 | 60 | 25 | 3 |
| 53 | 37 | 15 | 02 | 1 |

# Apéndice B

# Programas

Algoritmos GENALTO, PROCED, MINTRAZ y PROCGEN de los Capítulos 2 y 3, realizados en FORTRAN 77 y ejecutados en el ordenador VAX del Centro de Cálculo de la Universidad de Cádiz.

```
C
C                              GENALTO
C
C  Generacion de una muestra aleatoria de 100 conjuntos con 50 elementos
C  cada uno, eligiendo de forma aleatoria los parametros: dimension del
C  espacio, numero de grupos, número de elementos en cada grupo, centros
C  de cada grupo con una distancia entre ellos de al menos 3 y dispersión
C  dentro de cada grupo y dimensión de al menos 3
C

C
C  Declaración de dimensiones de distancia de cada centro de grupo al
C  origen, matriz de centros por grupos, matriz de dispersión por grupos,
C  vector de elementos y número de elementos de cada grupo (hasta 10
C  grupos)
C
```

```
      Character *50 nombre
      Real r(10),c(10,10),disp(10,10),x(10)
      Integer kelem(10)

C
C  Declaración de funciones de IMSL utilizadas e inicio de contador
C  aleatorio
C

      External RNSET,RNUNF
      Call RNSET(0)

C
C  Apertura del fichero donde se grabarán resultados
C

      Type*,'nombre.tipo del fichero donde estan los datos'
      Accept 12,nombre
 12   Format(a50)
      Open(20,name=nombre,status='new')

C
C  Inicio del programa
C  Se generan los 100 conjuntos de datos en el mismo fichero
C

      do 1=1,100

C
C  Dimension del espacio de entre 1 y 5, número de grupos entre 2 y ndim+1
C  y escritura
C

      ndim=1+int(5*RNUNF())
      ngrup=2+int(ndim*RNUNF())
      write(20,*),1,ndim,ngrup
```

```
C
C  Formacion de los centros de los grupos
C

      cond=1
      do while (cond.eq.1)
      do i=1,ngrup
      ccua=0
      r(i)=int(50*RNUNF())!distancia de cada centro al origen
      do j=1,ndim-1
      cota=sqrt(r(i)*r(i)-ccua)
      c(i,j)=cota*RNUNF()
      ccua = ccua + c(i,j)*c(i,j)
      enddo
      c(i,ndim)=sqrt(r(i)*r(i)-ccua)
      enddo
      cond=0
      do i=1,ngrup-1
      do j=i+1,ngrup
      dist2=0
      do k=1,ndim
      dist2=dist2+(c(i,k)-c(j,k))*(c(i,k)-c(j,k))
      enddo
      if (dist2.lt.9) then
      cond=1
      endif
      enddo
      enddo
      enddo

C
C  Dispersion en cada grupo y dimension
C

      do i=1,ngrup
      do j=1,ndim
      disp(i,j) = 1 + int(3*RNUNF())
```

```
      enddo
      enddo

C
C  Número de elementos en cada grupo
C

      keletot=0
      do k=1,ngrup-1
      kelem(k)=1+int(50*RNUNF()/ngrup)
      keletot = keletot + kelem(k)
      enddo
      kelem(ngrup)=50-keletot

C
C  Generación de elementos en cada grupo y escritura
C

      do k=1,ngrup
      do i=1,kelem(k)
      do j=1,ndim
      x(j)=c(k,j) + disp(k,j)*(2*RNUNF()-1)
      enddo
      write(20,*),(x(j),j=1,ndim),k
      enddo
      enddo

      enddo
      CLOSE(20)
      end
```

```
C
C                                PROCED
C
C  Procedimiento de clasificacion utilizando el modelo factorial
C  en modo Q cuando se tienen un numero de grupos en el conjunto
C  de datos mayor que uno y menor o igual que la dimension de los
C  datos más uno
C

C
C  Declaración de dimensiones y de subrutinas y funciones externas de IMSL
C  utilizadas
C

      Character *50 nombre
      Real t(10000,10),a(100,10),s(100,10),xnor(100)
      Integer comi,conj,ncolum,nfilas,ngrup
      Integer grupin(100),in(100),matbu(20,-20:20)
      Integer buenos,buentotal
      Common c(100,10),w(100,10),wpw(10,10),nfilas,ncolum,kautv,d,cotaut
      Common /rot/ b(100,10),cor(10,10),avec(10,10),nncolum
      External TRAT,DET,CRIT,POL4,UVMIF,EVLSF,EVCSF

C
C  Introducción del fichero de datos y la cota mínima para elegir los
C  autovalores
C
      Type*,'Los datos del fichero deben contener para cada uno de los
      100 conjuntos el número de conjunto, la dimensión y el número de
      grupos en la primera línea, y en las siguientes lineas los valores
      de cada elemento y grupo inicial al que pertenece'
      Type*,'nombre.tipo del fichero donde estan los datos'
      Accept 12,nombre
 12   Format(a50)
      Open(10,name=nombre,status='old')
      i=1
      do while (i.le.10000)
```

```
      do k=1,100
      if (i.eq.50*(k-1)+k) then
      read(10,*),(t(i,j),j=1,3)
      ncolum=int(t(i,2))
      i=i+1
      endif
      enddo
      read(10,*,end=2),(t(i,j),j=1,ncolum+1)
      i=i+1
      enddo
 2    CLOSE(10) !LEIDOS
      Type*,'cota mínima para los autovalores'
      Accept *,cotaut

C
C
C     PROGRAMA PRINCIPAL
C
C

C
C  Apertura del fichero donde se graban los resultados
C

      Type*,'nombre.tipo del fichero donde se graban resultados'
      Accept 15,nombre
 15   Format(a50)
      Open(20,name=nombre,status='new') !se abre el fichero

C
C  Inicio de contador de elementos bien clasificados y asignación de conj,
C  nfilas, ncolum, ngrup, matriz de datos inicial en A, agrupación inicial
C  de sus elementos en grupin y escritura
C

      buentotal=0
      do l=1,100
```

```
      comi=50*(l-1)+1
      conj=t(comi,1)
      nfilas=50
      ncolum=t(comi,2)
      ngrup=t(comi,3)
      do i=1,nfilas
      do j=1,ncolum
      a(i,j)=t(comi+i,j)
      enddo
      grupin(i)=t(comi+i,ncolum+1)
      enddo
      write(20,*),'conjunto',l,'dimesion',ncolum,'grupos',ngrup
      write(20,*),'matriz de datos y agrupación inicial'
      do i=1,nfilas
      write(20,*),(a(i,j),j=1,ncolum),grupin(i)
      enddo

C
C  Cálculo de la matriz de datos centrada
C

      do j=1,ncolum
      suma=0
      do i=1,nfilas
      suma=suma+a(i,j)
      enddo
      xmedia=suma/nfilas
      do i=1,nfilas
      c(i,j)=a(i,j)-xmedia
      enddo
      enddo

C
C  Busqueda de la distancia de desplazamiento d óptima, con ayuda de la
C  función DET y la subrutina de IMSL UVMIF
C
```

```
      Call UVMIF(DET,.5,.4,10.,.01,200,d)
      write(20,*),'valor de d ',d

C
C  Cálculo de det(W'W) con ayuda de DET, una vez desplazados los datos
C

      vdet=-DET(d)
      write(20,*),'valor del det(WPW) desplazado ',vdet

C
C  Cálculo de los autovalores y autovectores de W'W eligiendo aquéllos
C  autovalores que cumplan la condición de ser mayores que cotaut*nfilas,
C  la matriz de pesos inicial y la de correlaciones entre los autovectores
C  con ayuda de la subrutina TRAT
C

      Call TRAT()

C
C  Comienzo de las rotaciones con la ayuda de la subrutina ROT y la función
C  CRIT y la función POL4, obteniendo y escribiendo las matrices de pesos
C  rotada, la de correlaciones de los nuevos factores y la de estructura
C  factorial
C

      if (kautv.gt.1) then ! kautv es el número de direcciones elegidas
      do i=1,nfilas ! tipificacion por filas de la matriz de pesos
      xnorma=0
      do j=1,kautv
      xnorma=xnorma+b(i,j)*b(i,j)
      enddo
      xnor(i)=sqrt(xnorma)
      if (xnor(i).gt.0.) then
      do j=1,kautv
      b(i,j)=b(i,j)/xnor(i)
      enddo
```

```
endif
enddo
vali=CRIT(nfilas,kautv)
valv=vali
Call ROT(nfilas,kautv)
valn=CRIT(nfilas,kautv)
write(20,*),'criterio',valn
do while ((valv-valn)/vali.gt.1.e-5)
Call ROT(nfilas,kautv)
valv=valn
valn= CRIT(nfilas,kautv)
write(20,*),'criterio',valn
enddo
endif
write(20,*),'matriz de pesos rotada'
do i=1,nfilas
do j=1,kautv
b(i,j)=b(i,j)*xnor(i)! devolvemos b(i,j) a su valor inicial
enddo
write(20,*),(b(i,j),j=1,kautv)
enddo
write(20,*),'matriz de correlaciones final'
do i=1,kautv
write(20,*),(cor(i,j),j=1,kautv)
enddo
write(20,*),'matriz de estructura factorial'
do i=1,nfilas
do j=1,kautv
s(i,j)=0
do k=1,kautv
s(i,j)=s(i,j)+b(i,k)*cor(k,j)
enddo
enddo
write(20,*),(s(i,j),j=1,kautv)
enddo

C
```

```
C  Agrupamientos de los elementos por la matriz de estructura factorial
C  y escritura de los datos con su agrupación inicial y final
C

      do i=1,nfilas
      xmax=abs(s(i,1))
      in(i)=nint(sign(1.,s(i,1))) ! 1*signo de s(i,1)
      do j=2,kautv
      if (abs(s(i,j)).gt.xmax) then
      xmax=abs(s(i,j))
      in(i)=nint(sign(float(j),s(i,j)))
      endif
      enddo
      enddo
      write(20,*),'datos, agrupacion inicial y agrupacion final'
      do i=1,nfilas
      write(20,*),(a(i,j),j=1,ncolum),grupin(i),in(i)
      enddo

C
C  Recuento de elementos bien clasificados
C

      do i=1,20
      do j=-20,20
      matbu(i,j)=0
      enddo
      enddo
      do k=1,nfilas
      matbu(grupin(k),in(k))=matbu(grupin(k),in(k))+1
      enddo
      buenos=0
      bmax=1
      do while (bmax.ne.0)
      bmax=0
      do i=1,20
      do j=-20,20
```

```
      if (matbu(i,j).gt.bmax) then
      bmax=matbu(i,j)
      ibm=i
      jbm=j
      endif
      enddo
      enddo
      buenos=buenos+bmax
      do k=-20,20
      matbu(ibm,k)=0
      enddo
      do k=1,20
      matbu(k,jbm)=0
      enddo
      enddo
      write(20,*),'numero de elementos bien clasificados',buenos
      buentotal=buentotal+buenos

      enddo
      write(20,*),'elementos totales bien clasificados ',buentotal
      CLOSE(20)

      END ! FIN DEL PROGRAMA PRINCIPAL

C
C
C     FUNCION CRITERIO A MINIMIZAR EN LAS ROTACIONES
C
C


      Real Function CRIT(nfilas,kautv)
      Common /ROT/ b(100,10),cor(10,10),avec(10,10),nncolum
      tem=0
      do i=1,kautv
      do j=i+1,kautv
      do k=1,nfilas
      tem=tem+b(k,i)*b(k,i)*b(k,j)*b(k,j)
```

```
      enddo
      enddo
      enddo
      CRIT=tem
      return
      end

C
C
C     FUNCION POLINOMICA UTILIZADA EN LA SUBRUTINA ROTACION
C
C

      Real Function POL4(x)
      Common /coef/ aa,bb,cc,dd,ee
      POL4=aa+x*(bb+x*(cc+x*(dd+x*ee)))
      return
      end

C
C
C     SUBRUTINA ROTACION
C
C

      Subroutine ROT(nfilas,kautv)
      External POL4,UVMIF
      Common /ROT/ b(100,10),cor(10,10),avec(10,10),nncolum
      common /coef/ aa,bb,cc,dd,ee

      do nf1=1,kautv ! se eligen dos factores
      do nf2=1,kautv
      if (nf1.ne.nf2) then
      aa=0
      bb=0
      cc=0
      dd=0
```

```
ee=0

do i=1,nfilas ! obtención de los coeficientes de POL4
v1=b(i,nf1)*b(i,nf1)
v2=b(i,nf2)*b(i,nf2)
v3=b(i,nf1)*b(i,nf2)
v4=-v1-v2
do j=1,kautv
v4=v4+b(i,j)*b(i,j) !suma de cuadrados menos los dos factores
enddo
aa=aa+v1*v4+v2*v4+v1*v2
bb=bb+2*cor(nf1,nf2)*v1*(v4+v2)-2*v3*(v1+v4)
cc=cc+v1*(v1+v2+2*v4)-4*cor(nf1,nf2)*v1*v3
dd=dd+2*cor(nf1,nf2)*v1*v1-2*v1*v3
ee=ee+v1*v1
enddo ! obtenidos los coeficientes de POL4

Call UVMIF(pol4,0.,2.,20.,.01,200,xdel) ! llamada para min. POL4
xgam=sqrt(1+2*cor(nf1,nf2)*xdel+xdel*xdel)

do i=1,nfilas ! nueva matriz de pesos
b(i,nf1)=b(i,nf1)*xgam
b(i,nf2)=b(i,nf2)-b(i,nf1)*xdel
enddo

do i=1,nncolum ! nuevas direcciones rotadas
avec(i,nf1)=(avec(i,nf1)+xdel*avec(i,nf2))/xgam
enddo

do i=1,kautv !nueva matriz de correlaciones
if (i.ne.nf1) then
cor(nf1,i)=(cor(nf1,i)+xdel*cor(nf2,i))/xgam
cor(i,nf1)=cor(nf1,i)
endif
enddo

endif
```

```
      enddo
      enddo

      return
      end

C
C
C     FUNCION DETERMINANTE UTILIZADA PARA EL CALCULO DEL DESPLAZAMIENTO Y
C     DE LA MATRIZ W'W
C
C

      Real Function DET(d)
      Real aval(10),avalord(10)
      Common c(100,10),w(100,10),wpw(10,10),nfilas,ncolum,kautv

      nncolum=ncolum+1

      do i=1,nfilas !añadimos la columna constante y norm. por filas
      xnorma=d*d
      do j=1,ncolum
      xnorma=xnorma+c(i,j)*c(i,j)
      enddo
      xnorma=sqrt(xnorma)
      if (xnorma.gt.0.) then
      do j=1,ncolum
      w(i,j)=c(i,j)/xnorma
      enddo
      w(i,nncolum)=d/xnorma
      endif
      enddo

      do i=1,nncolum! calculo de W'W
      do j=1,nncolum
      wpw(i,j)=0
      do k=1,nfilas
```

```
      wpw(i,j)=wpw(i,j)+w(k,i)*w(k,j)
      enddo
      enddo
      enddo

      Call EVLSF(nncolum,wpw,10,aval) ! autovalores de W'W

      vdet=1 ! calculo del determinante det(W'W)
      do i=1,nncolum
      vdet=vdet*aval(i)
      enddo
      DET=-vdet !la rutina de IMSL llamada minimiza
      return
      end

C
C
C     SUBRUTINA TRATAMIENTO DONDE SE OBTIENE LA DESCOMPOSICION ESPECTRAL
C     DE LA MATRIZ W'W, EL NUMERO DE DIRECCIONES PRINCIPALES ELEGIDAS Y
C     LA MATRIZ DE PESOS Y DE CORRELACIONES INICIAL
C
C

      Subroutine TRAT()
      Real aval(10),oval(10),ovec(10,10)
      Common c(100,10),w(100,10),wpw(10,10),nfilas,ncolum,kautv,d,cotaut
      Common /ROT/ b(100,10),cor(10,10),avec(10,10),nncolum
      External DET,CRIT,POL4,UVMIF,EVLSF,EVCSF
      nncolum=ncolum+1

      do i=1,nncolum! cálculo de W'W
      do j=1,nncolum
      wpw(i,j)=0
      do k=1,nfilas
      wpw(i,j)=wpw(i,j)+w(k,i)*w(k,j)
      enddo
      enddo
```

```
enddo

Call EVCSF(nncolum,wpw,10,oval,ovec,10) !autovalores y autovectores
kautv=0 !contador autovectores mayores que cotaut*nfilas
do i=1,nncolum
if (oval(i).gt.nfilas*cotaut) kautv=kautv+1

aval(i)=oval(nncolum-i+1) ! autovalores en orden decreciente
do j=1,nncolum
avec(i,j)=ovec(i,nncolum-j+1)
enddo
enddo

do j=1,kautv ! autovectores que interesan a norma euclidea 1
xnorma=0
do i=1,nncolum
xnorma=xnorma+avec(i,j)*avec(i,j)
enddo
xnorma=sqrt(xnorma)
if (xnorma.gt.0.) then
do i=1,nncolum
avec(i,j)=avec(i,j)/xnorma
enddo
endif
enddo

write(20,*),'autovalores y autovectores de W'W que interesan'
do j=1,kautv
write(20,*),aval(j),(avec(i,j),i=1,nncolum)
enddo

write(20,*),'matriz de pesos inicial' ! obtención y escritura
do i=1,nfilas
do j=1,kautv
b(i,j)=0
do k=1,nncolum
b(i,j)=b(i,j)+w(i,k)*avec(k,j)
```

```
enddo
enddo
write(20,*),(b(i,j),j=1,kautv)
enddo

write(20,*),'matriz de correlaciones inicial'!obtención y escritura
do i=1,kautv
do j=1,kautv
cor(i,j)=0
do k=1,nncolum
cor(i,j)=cor(i,j)+avec(k,i)*avec(k,j)
enddo
enddo
write(20,*),(cor(i,j),j=1,kautv)
enddo
return

end
```

```
C
C                              MINTRAZ
C
C  Cálculo de la ordenación que minimiza la traza de la matriz W de
C  dispersión dentro de los grupos dada en (2.1), cuando se
C  conoce el número de grupos y se parte de una ordenación cualquiera.
C  El programa realiza las ordenaciones para 100 conjuntos de datos y
C  cuenta los elementos bien clasificados
C

C
C  Declaración de dimensiones
C

      Character *50 nombre
      Real t(10000,50),a(100,50)! hasta 100 elementos y 50 valores
      Integer in(100),grupin(100),lug(100)! indices de grupos ini. y fin.
      Real suma(20,50), suma2(20,50) ! hasta 20 grupos, suma y suma de
      cuadrados
      Real trwv(20) ! para guardar las trazas de cada grupo
      Real sv(50),sv2(50),sn(50),sn2(50) ! sumas del posible cambio
      Integer kgrupo(20),matbu(100,100) ! hasta 20 grupos

C
C  Introducción y lectura de los datos
C

      Type*,'Los datos del fichero deben contener para cada uno de los
      100 conjuntos el número de conjunto, la dimensión y el número de
      grupos en la primera línea, y en las siguientes lineas los valores
      de cada elemento y grupo inicial al que pertenece'
      Type*,'nombre.tipo del fichero donde estan los datos'
      accept 12,nombre
 12   Format(a50)
      Open(10,name=nombre,status='old') !se abre el fichero
      i=1
```

```
      do while (i.le.10000)
      do k=1,100
      if (i.eq.50*(k-1)+k) then
      read(10,*),(t(i,j),j=1,3)
      ncolum=int(t(i,2))
      i=i+1
      endif
      enddo
      read(10,*,end=2),(t(i,j),j=1,ncolum+1)
      i=i+1
      enddo
 2    CLOSE(10)!LEIDOS

C
C  Apertura del fichero donde se graban los resultados
C

      Type*,'nombre.tipo del fichero de salida de los datos'
      Accept 12,nombre
      Open(20,name=nombre,status='new') !se abre el fichero

C
C  Inicio de contador de elementos bien clasificados y reordenaciones, y
C  asignación de conj, nfilas, ncolum, ngrup, matriz de datos inicial en
C  A, agrupación inicial de sus elementos en in y grupin y escritura
C

      buentotal=0 ! iniciamos a cero los bien clasificados
      reorden=0 ! iniciamos numero de reordenaciones a cero
      do l=1,100 ! contador de conjuntos
      comi=50*(l-1)+1 ! para leer los comienzos de cada conjunto
      conj=int(t(comi,1))
      nfilas=50
      ncolum=int(t(comi,2))
      ngrup=int(t(comi,3))
      do i=1,nfilas
      do j=1,ncolum
```

```
        a(i,j)=t(comi+i,j)
        enddo
        in(i)=int(t(comi+i,ncolum+1))
        grupin(i)=in(i)
        lug(i)=0
        enddo
        write(20,*),'conjunto',conj,'dimension',ncolum,'grupos',ngrup
        write(20,*),'matriz inicial'
        do i=1,nfilas
        write(20,*),(a(i,j),j=1,ncolum),in(i)
        enddo

C
C  Cálculos de las trazas iniciales de cada grupo y la total
C

        do i=1,ngrup !iniciamos a cero suma, suma2 y kgrupo
        do j=1,ncolum
        suma(i,j)=0
        suma2(i,j)=0
        enddo
        kgrupo(i)=0
        enddo

        do i=1,nfilas ! calculo de las medias de los grupos
        k=in(i)
        kgrupo(k)=kgrupo(k)+1
        do j=1,ncolum
        tem=a(i,j)
        suma(k,j)=suma(k,j)+tem
        suma2(k,j)=suma2(k,j)+tem*tem
        enddo
        enddo

        trw=0 ! calculo de las  trazas iniciales de cada grupo y la total
        do i=1,ngrup
        if(kgrupo(i).gt.0) then
```

```
        do j=1,ncolum
        trwv(i)=suma2(i,j)-suma(i,j)*suma(i,j)/kgrupo(i)
        enddo
        endif
        trw=trw+trwv(i)
        enddo
        write(20,*),'traza de w de la ordenacion inicial',trw

C
C COMIENZA EL CÁLCULO DE OTRAS ORDENACIONES MEJORES
C

        do i=1,ncolum ! inicializamos a cero vectores temporales
        sv(i)=0
        sv2(i)=0
        sn(i)=0
        sn2(i)=0
        enddo

        iban=0 ! bandera de cambio de grupo
        icon=1 ! contador individuo
        do while (iban.le.nfilas) ! se da una vuelta completa sin cambios
        ig=1 ! contador incremento grupo al que se pretende cambiar
        igv=in(icon) ! numero de grupo inicial del individuo
        if (kgrupo(igv).gt.1) then
        do i=1,ncolum
        tem=a(icon,i)
        sv(i)=suma(igv,i)-tem
        sv2(i)=suma2(igv,i)-tem*tem ! lo quitamos del grupo donde estaba
        tngv=sv2(i)-sv(i)*sv(i)/(kgrupo(igv)-1)
        enddo
        do while (ig.lt.ngrup.and.iban.gt.0) ! no hay cambios y no se han
        ! probado todos los grupos para este individuo
        ign=igv+ig
        if (ign.gt.ngrup) ign=ign-ngrup ! grupo al que se apunta
        do i=1,ncolum
        tem=a(icon,i)
```

```
        sn(i)=suma(ign,i)+tem
        sn2(i)=suma2(ign,i)+tem*tem ! lo añadimos en el grupo nuevo
        tngn=sn2(i)-sn(i)*sn(i)/(kgrupo(ign)+1)
        enddo
        if(tngn+tngv.lt.trwv(igv)+trwv(ign)) then ! se debe cambiar
        iban=0 ! cambiamos bandera, trazas y sumas
        trwv(igv)=tngv
        trwv(ign)=tngn
        do i=1,nfilas
        suma(ign,i)=sn(i)
        suma2(ign,i)=sn2(i)
        suma(igv,i)=sv(i)
        suma2(igv,i)=sv2(i)
        enddo
        in(icon)=ign ! lo añadimos al nuevo grupo
        kgrupo(igv)=kgrupo(igv)-1
        kgrupo(ign)=kgrupo(ign)+1
        else ! miramos al siguiente grupo
        ig=ig+1
        endif
        enddo ! fin de do while de grupos
        endif ! fin de condicional mas de un elemento en grupo de partida
        lug(icon)=in(icon)
        icon=icon+1
        if (icon.gt.nfilas) icon=icon-nfilas
        iban=iban+1
        enddo ! fin de do while de elementos
        do I=1,nfilas
        write(20,*),(a(i,j),j=1,ncolum),grupin(i),lug(i)
        enddo

C
C  Cálculo de la nueva traza de W
C

        trwf=0
        do i=1,ngrup
```

```
        if(kgrupo(i).gt.0) then
        do j=1,ncolum
        trwv(i)=suma2(i,j)-suma(i,j)*suma(i,j)/kgrupo(i)
        enddo
        endif
        trwf=trwf+trwv(i)
        enddo
        write(20,*),'traza de w de la ordenacion final',trwf

C
C  Recuento de elementos bien clasificados
C

        do i=1,100
        do j=1,100
        matbu(i,j)=0
        enddo
        enddo
        do k=1,nfilas
        matbu(grupin(k),lug(k))=matbu(grupin(k),lug(k))+1
        enddo
        buenos=0
        bmax=1
        do while (bmax.ne.0)
        bmax=0
        do i=1,100
        do j=1,100
        if (matbu(i,j).gt.bmax) then
        bmax=matbu(i,j)
        ibm=i
        jbm=j
        endif
        enddo
        enddo
        buenos=buenos+bmax
        do k=1,100
        matbu(ibm,k)=0
```

```
matbu(k,jbm)=0
enddo
enddo
if (buenos.ne.50) reorden=reorden+1
write(20,*),'numero de elementos bien clasificados', buenos
buentotal=buentotal+buenos

enddo ! fin contador de conjuntos

write(20,*),'elementos totales bien clasificados', buentotal
write(20,*),'numero de reordenaciones',reorden
CLOSE(20)

end ! FIN DEL PROGRAMA
```

```
C
C                          PROCGEN
C
C  Procedimiento de clasificacion utilizando el modelo factorial en modo Q
C  cuando se tiene más de un grupo en el conjunto de datos.
C
C


C
C  Declaración de dimensiones y de subrutinas y funciones externas de IMSL
C  utilizadas
C

      Character *50 nombre
      Real t(10000,10), a(100,10),s(100,10),xnor(100)
      Integer comi,conj,ncolum,nfilas,ngrup
      Integer kmed(-20:20),valmed(-20:20),grupos(10,10)
      Integer buenos, buentotal
      Integer grupin(100),in(100),lug(100),ordin(100),matbu(100,100)
      Common c(100,10),w(100,10),wpw(10,10),nfilas,ncolum,kautv,d,cotaut
      Common /ROT/ b(100,10),cor(10,10),avec(10,10),nncolum
      External TRAT,DET,CRIT,POL4,UVMIF,EVLSF,EVCSF

C
C  Introducción del fichero de datos y la cota mínima para elegir los
C  autovalores
C

      Type*,'Los datos del fichero deben contener para cada uno de los
      100 conjuntos el número de conjunto, la dimensión y el número de
      grupos en la primera línea, y en las siguientes lineas los valores
      de cada elemento y grupo inicial al que pertenece'
      Type*,'nombre.tipo del fichero donde estan los datos'
      Accept 12,nombre
      Type*,'cota mínima para los autovalores'
      Accept *,cotaut
 12   Format(a50)
```

```
        Open(10,name=nombre,status='old') ! se abre el fichero
        i=1
        do while (i.le.10000)
        do k=1,100
        if (i.eq.50*(k-1)+k) then
        read(10,*),(t(i,j),j=1,3)
        ncolum=int(t(i,2))
        i=i+1
        endif
        enddo
        read(10,*,end=2),(t(i,j),j=1,ncolum+1)
        i=i+1
        enddo
2       CLOSE(10) ! LEIDOS


C
C
C PROGRAMA PRINCIPAL
C
C

C
C  Apertura del fichero donde se graban los resultados
C

        Type*,'nombre.tipo del fichero donde se graban resultados'
        Accept 15,nombre
 15     Format(a50)
        Open(20,name=nombre,status='new') ! se abre el fichero

C
C  Inicio de contador de elementos bien clasificados y asignación de conj,
C  nfilas, ncolum, ngrup, matriz de datos inicial en A, agrupación inicial
C  de sus elementos en grupin y escritura
C
```

```
      buentotal=0
      do l=1,100
      comi=50*(l-1)+1
      conj=t(comi,1)
      nfilas=50
      ncolum=t(comi,2)
      ngrup=t(comi,3)
      do i=1,nfilas
      do j=1,ncolum
      a(i,j)=t(comi+i,j)
      enddo
      grupin(i)=t(comi+i,ncolum+1)
      ordin(i)=i
      lug(i)=0
      enddo
      write(20,*),'conjunto',conj,'dimesion',ncolum,'grupos',ngrup
      write(20,*),'matriz de datos y agrupación inicial'
      do i=1,nfilas
      write(20,*),(a(i,j),j=1,ncolum),grupin(i)
      enddo

C
C  Cálculo de la matriz de datos centrada
C

      do j=1,ncolum
      suma=0
      do i=1,nfilas
      suma=suma+a(i,j)
      enddo
      xmedia=suma/nfilas
      do i=1,nfilas
      c(i,j)=a(i,j)-xmedia
      enddo
      enddo

C
```

```
C  Busqueda de la distancia de desplazamiento d óptima, con ayuda de la
C  función DET y la subrutina de IMSL UVMIF
C

      Call UVMIF(DET,.5,.4,50.,.01,200,d)
      write(20,*),'valor de d ',d

C
C  Cálculo de det(W'W) con ayuda de DET, una vez desplazados los datos
C

      vdet=-DET(d)
      write(20,*),'valor del det(WPW) desplazado ',vdet

C
C  Mientras queden elementos por clasificar, cálculo de los autovalores y
C  autovectores de W'W, eligiendo aquéllos autovalores que cumplan la
C  condición de ser mayores que cotaut*nfilas, y obteniendo la matriz de
C  pesos inicial y la de correlaciones entre los autovectores con ayuda de
C  la subrutina TRAT
C
      ki=0
      do while (nfilas.gt.1)
      ki=ki+1
      call trat()

C
C  Comienzo de las rotaciones con la ayuda de la subrutina ROT y la función
C  CRIT y la función POL4, obteniendo y escribiendo las matrices de pesos
C  rotada, la de correlaciones de los nuevos factores y la de estructura
C  factorial
C

      do i=1,nfilas ! tipificacion por filas de la matriz de pesos
      xnorma=0
      do j=1,kautv
      xnorma=xnorma+b(i,j)*b(i,j)
```

```
enddo
xnor(i)=sqrt(xnorma)
if (xnor(i).gt.0.) then
do j=1,kautv
b(i,j)=b(i,j)/xnor(i)
enddo
endif
enddo
if (kautv.gt.1) then ! kautv es el número de direcciones elegidas
vali=CRIT(nfilas,kautv)
valv=vali
Call ROT(nfilas,kautv)
valn=CRIT(nfilas,kautv)
write(20,*),'criterio',valn
do while ((valv-valn)/vali.gt.1.e-5)
Call ROT(nfilas,kautv)
valv=valn
valn= CRIT(nfilas,kautv)
write(20,*),'criterio',valn
enddo
endif
write(20,*),'matriz de pesos rotada'
do i=1,nfilas
do j=1,kautv
b(i,j)=b(i,j)*xnor(i) !devolvemos b(i,j) a su valor inicial
enddo
write(20,*),(b(i,j),j=1,kautv)
enddo
write(20,*),'matriz de correlaciones final'
do i=1,kautv
write(20,*),(cor(i,j),j=1,kautv)
enddo
write(20,*),'matriz de estructura factorial'
do i=1,nfilas
do j=1,kautv
s(i,j)=0
do k=1,kautv
```

```
      s(i,j)=s(i,j)+b(i,k)*cor(k,j)
      enddo
      enddo
      write(20,*),(s(i,j),j=1,kautv)
      enddo

C
C  Agrupamientos de los elementos por la matriz de estructura factorial
C

      do i=1,nfilas
      xmax=abs(s(i,1))
      in(i)=nint(sign(1.,s(i,1))) ! 1*signo de s(i,1)
      do j=2,kautv
      if (abs(s(i,j)).gt.xmax) then
      xmax=abs(s(i,j))
      in(i)=nint(sign(float(j),s(i,j)))
      endif
      enddo
      enddo

C
C  Búsqueda de la mayor media en valor absoluto, clasificación final y
C  escritura de los datos con su agrupación inicial y final
C

      do i=-kautv,kautv
      valmed(i)=0
      kmed(i)=0
      enddo
      do i=1,nfilas
      valmed(in(i))=valmed(in(i))+abs(s(i,abs(in(i))))
      kmed(in(i))=kmed(in(i))+1
      enddo
      xmedma=0
      imedma=0
      do i=-kautv,kautv
```

```
        if (kmed(i).gt.0.) then
        if (valmed(i)/kmed(i).gt.xmedma) then
        xmedma=valmed(i)/kmed(i)
        imedma=i
        endif
        endif
        enddo
        kon=1
        do i=1,nfilas
        if (in(i).eq.imedma) then
        lug(ordin(i))=ki ! damos a los elementos clasificados el lugar ki
        write(20,*),ordin(i),grupin(ordin(i)),'grupclas',ki
        else
        do j=1,ncolum+1 ! los metemos en la nueva matriz para analizar
        w(kon,j)=w(i,j)
        enddo
        ordin(kon)=ordin(i)
        kon=kon+1
        endif
        enddo
        nfilas=kon-1
        write(20,*),'quedan por clasificar'
        do i=1,nfilas
        write(20,*),ordin(i),grupin(ordin(i))
        enddo
        enddo

        nfilas=50
        grupos(ngrup,ki)=grupos(ngrup,ki)+1
        write(20,*),'datos, agrupacion inicial y agrupacion final'
        do i=1,nfilas
        write(20,*),(a(i,j),j=1,ncolum),grupin(i),lug(i)
        enddo

C
C  Recuento de elementos bien clasificados
C
```

```
do i=1,100
do j=1,100
matbu(i,j)=0
enddo
enddo
do k=1,nfilas
matbu(grupin(k),lug(k))=matbu(grupin(k),lug(k))+1
enddo
buenos=0
bmax=1
do while (bmax.ne.0)
bmax=0
do i=1,100
do j=1,100
if (matbu(i,j).gt.bmax) then
bmax=matbu(i,j)
ibm=i
jbm=j
endif
enddo
enddo
buenos=buenos+bmax
do k=1,100
matbu(ibm,k)=0
matbu(k,jbm)=0
enddo
enddo
write(20,*),'numero de elementos bien clasificados ',buenos
buentotal=buentotal+buenos

enddo
write(20,*),'número de elementos en agrup. inicial y final'
do i=1,10
write(20,*)(grupos(i,j),j=1,10)
enddo
write(20,*),'elementos totales bien clasificados ',buentotal
```

```
      CLOSE(20)

      END ! FIN DEL PROGRAMA PRINCIPAL

C
C
C FUNCION CRITERIO A MINIMIZAR EN LAS ROTACIONES
C
C

      Real Function CRIT(nfilas,kautv)
      Common /ROT/ b(100,10),cor(10,10),avec(10,10),nncolum
      tem=0
      do i=1,kautv
      do j=i+1,kautv
      do k=1,nfilas
      tem=tem+b(k,i)*b(k,i)*b(k,j)*b(k,j)
      enddo
      enddo
      enddo
      CRIT=tem
      return
      end

C
C
C FUNCION POLINOMICA UTILIZADA EN LA SUBRUTINA ROTACION
C
C

      Real Function POL4(x)
      Common /coef/ aa,bb,cc,dd,ee
      POL4=aa+x*(bb+x*(cc+x*(dd+x*ee)))
      return
      end

C
```

```
C
C SUBRUTINA ROTACION
C
C
      Subroutine ROT(nfilas,kautv)
      External POL4,UVMIF
      Common /ROT/ b(100,10),cor(10,10),avec(10,10),nncolum
      Common /coef/ aa,bb,cc,dd,ee

      do nf1=1,kautv ! se eligen dos factores
      do nf2=1,kautv
      if (nf1.ne.nf2) then
      aa=0
      bb=0
      cc=0
      dd=0
      ee=0

      do i=1,nfilas ! obtención de los coeficientes de POL4
      v1=b(i,nf1)*b(i,nf1)
      v2=b(i,nf2)*b(i,nf2)
      v3=b(i,nf1)*b(i,nf2)
      v4=-v1-v2
      do j=1,kautv
      v4=v4+b(i,j)*b(i,j) !suma de cuadrados menos los dos factores
      enddo
      aa=aa+v1*v4+v2*v4+v1*v2
      bb=bb+2*cor(nf1,nf2)*v1*(v4+v2)-2*v3*(v1+v4)
      cc=cc+v1*(v1+v2+2*v4)-4*cor(nf1,nf2)*v1*v3
      dd=dd+2*cor(nf1,nf2)*v1*v1-2*v1*v3
      ee=ee+v1*v1
      enddo ! obtenidos los coeficientes de POL4

      Call UVMIF(pol4,0.,2.,20.,.01,200,xdel) ! llamada para min. POL4
      xgam=sqrt(1+2*cor(nf1,nf2)*xdel+xdel*xdel)
```

```
        do i=1,nfilas ! nueva matriz de pesos
        b(i,nf1)=b(i,nf1)*xgam
        b(i,nf2)=b(i,nf2)-b(i,nf1)*xdel
        enddo

        do i=1,nncolum ! nuevas direcciones rotadas
        avec(i,nf1)=(avec(i,nf1)+xdel*avec(i,nf2))/xgam
        enddo

        do i=1,kautv ! nueva matriz de correlaciones
        if (i.ne.nf1) then
        cor(nf1,i)=(cor(nf1,i)+xdel*cor(nf2,i))/xgam
        cor(i,nf1)=cor(nf1,i)
        endif
        enddo

        endif
        enddo
        enddo

        return
        end

C
C
C FUNCION DETERMINANTE UTILIZADA PARA EL CALCULO DEL DESPLAZAMIENTO
C Y DE LA MATRIZ W'W
C
C

        Real Function DET(d)
        Real aval(10),avalord(10)
        Common c(100,10),w(100,10),wpw(10,10),nfilas,ncolum,kautv

        nncolum=ncolum+1

        do i=1,nfilas !añadimos la columna constante y norm. por filas
```

```
        xnorma=d*d
        do j=1,ncolum
        xnorma=xnorma+c(i,j)*c(i,j)
        enddo
        xnorma=sqrt(xnorma)
        if (xnorma.gt.0.) then
        do j=1,ncolum
        w(i,j)=c(i,j)/xnorma
        enddo
        w(i,nncolum)=d/xnorma
        endif
        enddo

        do i=1,nncolum! calculo de W'W
        do j=1,nncolum
        wpw(i,j)=0 !iniciar a cero wpw
        do k=1,nfilas
        wpw(i,j)=wpw(i,j)+w(k,i)*w(k,j)
        enddo
        enddo
        enddo

        Call EVLSF(nncolum,wpw,10,aval) ! autovalores de W'W

        vdet=1 ! calculo del determinante
        do i=1,nncolum
        vdet=vdet*aval(i)
        enddo
        DET=-vdet !la rutina de IMSL llamada minimiza
        return
        end

C
C
C SUBRUTINA TRATAMIENTO DONDE SE OBTIENE LA DESCOMPOSICION ESPECTRAL
C DE LA MATRIZ W'W, EL NUMERO DE DIRECCIONES PRINCIPALES ELEGIDAS Y
C LA MATRIZ DE PESOS INICIAL
```

```
C
C

      Subroutine TRAT()
      Real aval(10),oval(10),ovec(10,10)
      Common c(100,10),w(100,10),wpw(10,10),nfilas,ncolum,kautv,d,cotaut
      Common /ROT/ b(100,10),cor(10,10),avec(10,10),nncolum
      External DET,CRIT,POL4,UVMIF,EVLSF,EVCSF

      nncolum=ncolum+1

      do i=1,nncolum! calculo de W'W
      do j=1,nncolum
      wpw(i,j)=0 !iniciar a cero wpw
      do k=1,nfilas
      wpw(i,j)=wpw(i,j)+w(k,i)*w(k,j)
      enddo
      enddo
      enddo

      Call EVCSF(nncolum,wpw,10,oval,ovec,10) !autovalores y autovectores
      kautv=0 !contador autovectores mayores que cotaut*nfilas
      do i=1,nncolum
      if (oval(i).gt.nfilas*cotaut) kautv=kautv+1

      aval(i)=oval(nncolum-i+1) ! autovalores en orden decreciente
      do j=1,nncolum
      avec(i,j)=ovec(i,nncolum-j+1)
      enddo
      enddo

      do j=1,kautv ! autovectores que interesan a norma euclidea 1
      xnorma=0
      do i=1,nncolum
      xnorma=xnorma+avec(i,j)*avec(i,j)
      enddo
      xnorma=sqrt(xnorma)
```

```
if (xnorma.gt.0.) then
do i=1,nncolum
avec(i,j)=avec(i,j)/xnorma
enddo
endif
enddo

write(20,*),'autovalores y autovectores de W'W que interesan'
do j=1,kautv
write(20,*),aval(j),(avec(i,j),i=1,nncolum)
enddo

write(20,*),'matriz de pesos inicial' ! obtención y escritura
do i=1,nfilas
do j=1,kautv
b(i,j)=0
do k=1,nncolum
b(i,j)=b(i,j)+w(i,k)*avec(k,j)
enddo
enddo
write(20,*),(b(i,j),j=1,kautv)
enddo

write(20,*),'matriz de correlaciones inicial'!obtención y escritura
do i=1,kautv !matriz de correlaciones
do j=1,kautv
cor(i,j)=0
do k=1,nncolum
cor(i,j)=cor(i,j)+avec(k,i)*avec(k,j)
enddo
enddo
write(20,*),(cor(i,j),j=1,kautv)
enddo

return
end
```

# Bibliografía


[1] Anderberg, M.R. (1973): *Cluster Analysis for Applications.* New York: Academic Press.

[2] Anderson, T.W. (1984): *An Introduction to Multivariate Analysis*, 2ª Ed., New York: Wiley.

[3] Autonne,L (1913): Sur les matrices hypohermitiennes et sur les matrices unitaries. *Comptes Rendus Acad. Sci. Paris*, **156**:1037-1055.

[4] Banfield, C.F. and Bassill, L.C. (1977): Algorithm AS113. A transfer algorithm for non-hierarchical classification. *Applied Statistics*, **26**:206-210.

[5] Baggaley, A.R. (1964):*Intermediate correlational methods.* New York: Wiley.

[6] Ball, G.H. y Hall, D.J. (1967): A clustering technique for summarizing multivariate data. *Behaviour Sci.*, **12**:153-155.

[7] Beale, E.M.L. (1969): *Cluster Analysis.* London:Scientific Control Systems.

[8] Beale, E.M.L. (1969): Euclidean cluster analysis. *Bull. I.S.I.*,**43**, Book 2, 92-94.

[9] Benzécri, J.P. (1976):*L'Analyse des Données. (Tome 1: La Taxonomie; Tome 2: L'Analyse des Correspondances).* Paris: Dunod.

[10] Bezdek,J.C. (1974) : Numerical Taxonomy with Fuzzy Sets. *Journal of Mathematical Biology*, 1: 57-71.

[11] Bezdek,J.C. (1974) : Cluster validity with Fuzzy Sets. *Journal of Cybernetics*, **3**: 58-73.

[12] Blashfield, R.K. (1976): Mixture model tests of cluster analysis: Accuracy of four agglomerative hierarchical methods. *Psychological Bulletin*, **49**: 499-520.

[13] Bock, H.H. (1985): On Some Significance Tests in Cluster Analysis. *Journal of Classification.* **2**:77-108.

[14] Burden, R.L. y Faires, J.D.: *Análisis Numérico.* Grupo Editorial Iberoamérica.

[15] Burt, C. (1917): *The Distribution and Relations of Educational Abilities.* London: P.S. King.

[16] Burt, C. (1937) : Correlations between persons. *British Journal of Psychology*, **28**, 59-96.

[17] Calinsky, T. y Harabasz, J. (1974): A Dendrite Method for Cluster Analysis. *Communications in Statistics*, **3.1**: 1-24.

[18] Carroll, J.B. (1953): Approximating simple structure in factor analysis. *Psychometrica*, **18**, 23-38.

[19] Carroll, J.B. (1957): Biquartimin criterion for rotation to oblique simple strucutre in factor analysis. *Science*, **126**, 1114-15.

[20] Carroll, J.B. (1960): IBM 704 program for generalizec analytic rotation solution in factor analysis. Harvard University, sin publicar.

[21] Carroll, J.B. (1961): The nature of the data, or how to choose a correlation coefficient. *Psychometrika*, **26**, nº 4: 347-372.

[22] Cattell,R.B. (1952): *Factor analysis.* New York: Harper & Bros

[23] Cattell,R.B. (1965): Factor analysis: An introduction to essentials. *Biometrics*,**21**, 190-215, 405-435.

[24] Cattell,R.B. (1978): *The Scientific Use of Factor analysis in the Behavioral and Life Sciences.* New York: Plenum Press.

[25] Clifford,H.T. y Stephenson,W (1975):*An Introduction to Numerical Classification.* New York: Academic Press.

[26] Cohen,J. (1969): $r_c$: A profile similarity coefficient invariant over variable reflection. *Psychological Bulletin*, **71**, nº 4: 281-284.

[27] Constantine,A.G. y Gower,J.C. (1978): Graphical representation of asymetric matrices. *Applied Statistics*, **27**:297-304.

[28] Cormack, R.M. (1971): A Review of Classification. *Journal of the Royal Stat. Soc. Ser. A.* **134**: 321-367.

[29] Cox, D.R. (1957): Note on grouping. *Journal of the American Statistical Association*, **52**: 543-547

[30] Cuadras,C.M. (1991) : *Métodos de Análisis Multivariante.* 2ª edic. Barcelona : PPU.

[31] De Sarbo,W.S., Carroll,J.D., Clark,L.A. and Green,P.E.(1984): Synthesized clustering: a method for amalgamating alternative clustering bases with differential weighting of variables. *Psychometrika*, **49**:57-58.

[32] Dempster,A.P. (1969): *Elements of Continuous Multivariate Analysis.* Addison-Wesley, Reading, Mass.

[33] Dixon,J.K. (1979): Pattern recognition with missing data. *IEEE Transactions on Systems, Man and Cybernetics*, SMC9:617-621.

[34] Dixon et al. (1990):*BMDP Statistical Software Manual. Vol. I y II.* University of California Press.

[35] Duda, R.O. and Hart, P.E. (1973): *Pattern Classification and Scene Analysis.* John Wiley and Sons.

[36] Eades, D.C. (1965): The inappropriateness of the correlation coefficient as a mesure of taxonomic resemblance. *Syst. Zool.*, **14**:98-100.

[37] Eckart, C. and Young, G.(1936): Approximation of one matrix by another of lower rank. *Psychometrika*,**1**, 211-218.

[38] Edwards, A.W.F. (1971): Distances between populations on the basis of gene frequencies. *Biometrics*, **27**:873-881.

[39] Eherenberg, A.S.C. (1968): On Methods: The Factor Analytic Search for Program Types. *Journal of Advertising Research*, **8**:55-63.

[40] Eherenberg, A.S.C. and Goodhart, G.J. (1976): Factor Analysis: Limitations and Alternatives. *Marketing Science Institute Working Paper No. 76-116.* Cambridge, Massachusetts: Marketing Science Institute.

[41] Engelman, L. y Hartigan, J.A. (1969): Percentage points of a test for cluster. *Journal of the American Statistical Association*, **64**: 1647-1648.

[42] Everitt, B.S. (1974): *Cluster Analysis*,London: Heinemann.

[43] Everitt, B.S. (1979): Unresolved Problems in Cluster Analysis. *Biometrics*, **35**: 169-181

[44] Everitt, B.S. (1993): *Cluster Analysis*, 3ª ed.,Edward Arnold.

[45] Fisher, LL. and Van Ness, J.W. (1971): Admissible clustering procedures. *Biometrika*, **58**:91-104.

[46] Fisher, R.A. (1936): The use of multiple measurements in taxonomic problems. *Ann. Eugen.*,**7**, 179-188.

[47] Fisher, W.D. (1958): On grouping for maximum homogeneity. *Journal of the American Statistical Association.* **53**: 789-798.

[48] Fleiss,J.L. and Zubin, J. (1969). On the metods and theory of clustering. *Multivariate Behaviour Res.*, **4**, 235-250.

[49] Fleiss et al. (1971). On the use of inverted factor analysis for generating typologies. *Journal of Abnormal Psychology*, **77**, 127-132.

[50] Forgy, E.W. (1965): Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, **21**: 768-769.

[51] Friedman, H.P. y Rubin, J. (1967): On Some Invariant Criteria for Grouping Data. *Jour. of Amer. Stat. Assoc.*, **62**: 1159-1178.

[52] Frisch, R (1929): Correlation an scatter in statistical variables. *Nordic Statistical Journal*, **8**, 36-102.

[53] Gitman, I. y Levine, M.D.(1970): An Algorithm for Detecting Unimodal Fuzzy Sets and Its Application as a Clustering Technique. *IEEE Transactions on Computers*, **C-19.7**:583-593.

[54] Godehardt, E. (1990): *Graphs as Structural Models: The Application of Graphs and Multigraphs in Cluster Analysis.* 2ª ed. Vieweg and Sohn.

[55] Gordon, A.D. y Henderson, J.J.( 1977): An Algorithm for Euclidean sum of squares classification. *Biometrics*, **33**:355-362.

[56] Gordon, A.D. (1981):*Classification.* London: Chapman and Hall.

[57] Gower, J.C. (1966): Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, **53**:325-338.

[58] Gower,J.C. (1971): A General Coefficient of Similarity and Some of Its Properties. *Biometrics*, **27**: 857-871.

[59] Gower, J.C. (1985):Measures of similarity, dissimilarity and distance. In *Encyclopedia of Statistical Sciences, Volume 5* (S. Kotz, N.L. Johson and C.B. Read, eds). New York: Wiley.

[60] Gutiérrez, R. y González, A. (1991): *Análisis Multivariable.* Los autores.

[61] Harmann, H.H. (1976):*Modern Factor Analysis.* The University of Chicago Press.

[62] Hartigan, J.A. (1967): Representation of Similarity Matrices by Trees. *Journal of the Amer. Stat. Assoc.*, **62**:1140-1158.

[63] Hartigan, J.A. (1975): *Clustering Algorithms.* New York: Wiley-Interscience.

[64] Hartigan, J.A. (1985): Statistical Theory in Clustering. *Journal of Classification*, **2**:63-76.

[65] Horst,P. (1965): *Factor analysis of data matrices.* New York: Holt, Rinehart & Winston.

[66] Imbrie,J. y Purdy, E. (1962): Classification of modern Bahamian carbonate sediments. In: Classification of Carbonate Rocks. *Am. Assoc. Pet. Geol. Mem.*, **7**: 253-272.

[67] Imbrie, J. (1963): Factor and vector analysis programs for analyzing geologic data. *Office Naval Res., Geogr. Branch. Tech. Rep.*, **6**, 83 pp.

[68] Imbrie, J. and Van Andel, T.H. (1964): Vector analysis of heavy mineral data. *Bull. Geol. Soc. Am.*, **75**: 1131-1156.

[69] Jackson, J.E.(1991): *A user's guide to Principal Components.* New York: Wiley.

[70] Jambu, M. (1991):*Exploratory and Multivariate Data Analysis.* Academic Press.

[71] Jancey, R.C. (1966): Multidimensional group analysis. *Aust. J. Bot.*, **14**: 127-130.

[72] Jardine, N. y Sibson, R. (1968): The Construction of Hierarchic and Non-hierarchic Classifications. *The Computer Journal*, **11**: 117-184.

[73] Jardine, N. y Sibson, R. (1968): A Model for Taxonomy. *Mathematical Biosciences*, **2**: 465-482.

[74] Jardine, N. y Sibson, R. (1971): Choice of Methods for Automatic Classification. *The Computer Journal*, **14**: 404-406.

[75] Jardine, N. y Sibson, R. (1971): *Mathematical Taxonomy.* New York: Wiley.

[76] Jensen,R.E. (1969):A dynamic programming algorithm for cluster analysis. *Op. Res.*, 1034-1056.

[77] Jennrich, R.I. y Sampson, P.F. (1966): Rotation for simple loadings. *Psychometrica*, **31**, 313-323.

[78] Jobson, J.D. (1992): *Applied Multivariate Data Analysis. Vol II: Categorical and Multivariate Methods.* New York: Springer-Verlag.

[79] Johnson, S.C. (1967): Hierarchical Clustering Schemes. Psychometrika, **32**: 241-254

[80] Johnson, R.A. y Wichern, D.W. (1992): *Applied Multivariate Statistical Analysis.* 3ª ed. Prentice-Hall.

[81] Jolliffe, I.T. (1986): *Principal Component Analysis.* New York: Springer-Verlag.

[82] Jones, K.J. (1968): Problems of grouping individuals and the method of modality. *Behavioral Science*, **13**, 496-511.

[83] Jöreskog, K.G., Klovan, J.E. and Reyment, R.A. (1976): *Geological Factor Analysis.* Elsevier Sc.

[84] Kaiser, H.F. (1958): The varimax criterion for analytic rotation in factor analysis. *Psychometrica*, **23**, 187-200.

[85] Koontz,W.L.G., Narendra, P.M. y Fukunaga,K. (1975):A branch and bound clustering algorithm. *IEEE Transactions on Computers*, **C-24**:908-915.

[86] Krzanowski, W.J. (1988): *Principles of Multivariate Analysis: A User's Perspective.* Oxford U.P.

[87] Krzanowski, W.J. y Lai, Y.T. (1988): A Criterion for Determining the Number of Groups in a Data Set Using Sum-of-Squares Clustering. *Biometrics*, **44**: 23-34.

[88] Lance, G.N. y Williams, W.T. (1967): A General Theory of Classificatory Sorting Strategies. I. Hierarchical Systems. *The Computer Journal*, **9**: 373-380.

[89] Lawley,D.N. y Maxwell,A.E.(1963,1971): *Factor Analysis as a statistical method.* London : Butterworth.

[90] Lebart,L., Morineau, A. y Warwick, K.M. (1984): *Multivariate descriptive statistical analysis.* Wiley.

[91] Lee, K.L. (1979): Multivariate Tests for Clusters. *Journal of the American Statistical Association*, **74**:708-714.

[92] Little,R.A. y Rubin,D.B. (1987):*Statistical Analysis with Missing Data.* New York: Wiley.

[93] Lorr, M. (Ed.) (1966): *Explorations in typing psychotics.* Oxford: Pergamon Press.

[94] MacQueen, J. (1967): Some Methods for Classification and Analysis of Multivariate Observations. *Proc. 5th. Berkeley Symp.*, **1**:281-297.

[95] McRae, D.J. (1971): MICKA, a Fortran IV iterative K-means cluster analysis program. *Behavioural Science*, **16**:423-424.

[96] Mahalanobis, P.C. (1936): On the Generalized Distance in Statistics. *Proc. Nat. Inst. Sci. India*, **2(1)**:49-55.

[97] Maronna, R y Jacovkis, P.M. (1974): Multivariate clustering procedures with variable metrics. *Biometrics*, **30**: 499-505.

[98] Marriott, F.H.C. (1971): Practical problems in a method of cluster analysis. *Biometrics*, **27**: 501-514.

[99] Marriott, F.H.C. (1982): Optimization methods of cluster analysis. *Biometrika*, **69**: 417-421.

[100] Miesch, A.T. (1976): Q-mode factor analysis of compositional data. *Computers and Geosciences.* **1**: 147-159.

[101] Milligan, G.W. y Cooper, M.C. (1985): An Examination of Procedures for Determining the Number of Clusters in a Data Set. *Psychometrika*, **50**:159-179.

[102] Mirkin, B.G. (1987): Additive Clustering and Qualitative Factor Analysis Methods for Similarity Matrices. *Journal of Classification.* **4**:7-31.

[103] Miyazaki, H. y Seki, Y. (1987): Principal Components and Principal Clusters. *Journal of Information and Optimization Sciences*, **8**: 189-199.

[104] Murtagh, F. (1983): Expected-Time Complexity Results for Hierarchic Clustering Algoritmhs which use Cluster Centres. *Information Technology: Research and Development*, **1**: 275-283.

[105] Murtagh, F. (1983): A Survey of Recent Advances in Hierarchical Clustering Algoritms. *The Computer Journal*, **26.4**: 354-359.

[106] Overall,J.E. y Klett,C.J. (1972):*Applied Multivariate Analysis.* New York: McGraw-Hill.

[107] Pearson, K. (1901): On lines and planes of closest fit to systems of points in space. *Phil. Mag., Ser. B*, **2**, 559-572.

[108] Rao,C.R. (1952): *Advanced Statistical Methods in Biometric Research.* New York: Wiley.

[109] Reyment, R and Jöreskog, K.G. (1993): *Applied Factor Analysis in the Natural Sciences.* 2ª ed. Cambridge U.P.

[110] Rubin,J. (1967): Optimal Classification into Groups: An Approach for Solving the Taxonomy Problem. *Journal of Theoretical Biology*, **15**: 103-144.

[111] Rubinstein, R.Y. (1981): *Simulation and the Monte Carlo method.* New York, Wiley.

[112] Ruspini, E.(1970): Numerical Methods for Fuzzy Clustering. *Information Science*, **2**: 319-350.

[113] Saunders, D.R. and Schucman, H. (1962): Syndrome analysis: an efficient procedure for isolating meaningful subgroups in a non-randon sample of a population. Paper read at Annual Meeting of Psychonomic Society, St. Louis, September.

[114] Scott, A.J. y Symon, M.J. (1971): Clustering methods based on likelihood ratio criteria. *Biometrics*, **27**:387-398.

[115] Singleton,R.C. y Kautz,W. (1965):*Minimum squared error clustering algorithm.* Stanford Research Institute.

[116] Sneath, P.H.A. y Sokal, R.R. (1973): *Numericla Taxonomy.* San Francisco: Freeman and Co.

[117] Sokal, R.R. y Michener, C.D. (1958): A statistical method for evaluating systematic relationships. *Univ. Kansas Sci. Bull.*, **38**:1409-1438.

[118] Sokal, R.R. y Rohlf, F.J.(1962): The Comparison of Dendograms by Objetive Methods. *Taxonomy*, **11**: 33-40.

[119] Spath,H. (1985):*Cluster Disecction and Analysis.* Chichester: Ellis Horwood Ltd.

[120] Stephenson, W. (1936): The inverted factor technique. *British Journal of Psychology*, **26**, 344-61.

[121] Stewart, D.W. (1981): The Application and Misapplication of Factor Analysis in Marketing Research. *Journal of Maketing Research*, **18**, 51-62.

[122] Sylvester, J.J. (1889): On the reduction of a bilinear quantic of the n.th order to the form of a sum of n products by a double orthpgonal substitution. *Messenger of Mathematics*, **19**, 42-46.

[123] Symons, M.J.(1981): Clustering criteria and multivariate normal mixtures. *Biometrics*, **37**:35-43.

[124] Thorndike, R.L. (1953): Who belongs in the family?. *Psychometrika*, 18: 267-276.

[125] Thurstone, L.L. (1935): *The vectors of mind.* The University of Chicago Press.

[126] Thurstone, L.L. (1947): *Multiple factor Analysis.* The University of Chicago Press.

[127] Walden,J. Smith,J.P. and Dackombe,R.V. (1992): The use of Simultaneous R- and Q-Mode Factor Analysis as a Tool for Assisting Interpretation of Mineral Magnetic Data. *Mathematical Geology*, **24**:227-247.

[128] Wallace,C.S. y Boulton,D.M. (1968):An information measure for classification. *Computer J.*, 11: 185-194.

[129] Ward, J.H. (1963): Hierarchical Grouping to Optimize an Objetive Function. *Journal of the American Statistical Association*, **58**:236-244.

[130] Wilks, S.S. (1932): Certain generalizations in the analysis of variance. *Biometrika*, **24**, 471-494.

[131] Wishart, D. (1969): An Algorithm for Hierarchical Classification. *Biometrics*, **25**: 165-170.

[132] Wong, M.A. (1982): A Hybrid Clustering Method for Identifying Hig-Density Clusters. *Journal of the American Statistical Association.* **77**:841-847.

[133] Zubin, J. and Fleiss, J.L. (1965): Taxonomy in the mental disorders- a historical perspective. *Symposium on Explorations in Typology with Special Reference to Psychotics.* New York: Human Ecology Fund.

Juan Luis González Caballero

Algunas aportaciones a los métodos de optimización del Análisis Cluster mediante la D.V.S.

Apto Cum Laude por unanimidad

21 Junio 96