QÜESTIIÓ, vol. 22, 1, p. 39-68, 1998

UN PROCEDIMIENTO PARA OBTENER CLUSTERS UTILIZANDO LA D.V.S. DE UNA MATRIZ. COMPARACIONES CON EL BIPLOT Y CON EL MODELO Q-FACTORIAL

JUAN L. GONZÁLEZ CABALLERO*
MARIANO J. VALDERRAMA BONNET**

Durante las últimas décadas, el análisis de un conjunto de n individuos medidos en p variables, proporcionando una matriz de datos $X_{n,p}$, mediante técnicas de representación que utilizan la Descomposición en Valores Singulares (DVS) de la matriz $X_{n,p}$ (o alguna derivada), han permitido resumir la información que aportan los datos en alguna forma óptima, siendo muy útil para indicar la presencia de clusters entre los n individuos y/o para prevenir ante posibles clasificaciones erróneas producidas por técnicas de agrupamiento más complejas. En este artículo estudiaremos un procedimiento que puede utilizarse en ocasiones para obtener clasificaciones naturales de un conjunto de datos, basado en la representación biplot y en el modelo Q-factorial que puede obtenerse a partir de la DVS.

A procedure of clustering using the SVD of a matrix. Comparisons with the biplot and with the Q-factor model.

Palabras clave: Descomposición en valores singulares; modelos factoriales; representación biplot; varianza generalizada; procedimientos de cluster no jerárquicos.

Clasificación AMS (1991): 62H25, 62H30.

* Juan L. González Caballero. Dpto. de Matemáticas. Facultad de Medicina. Universidad de Cádiz.

** Mariano J. Valderrama Bonnet. Dpto. de Estadística e I.O. Facultad de Farmacia. Universidad de Granada.

– Recibido en mayo de 1996.
– Aceptado en septiembre de 1997.

1. INTRODUCCIÓN

La Descomposición en Valores Singulares (DVS) de una matriz $X_{n,p}$ de rango r ($\leq p < n$) es un resultado descrito por primera vez por el matemático inglés Sylvester [20], que permite descomponer X como

$$X = V\Lambda U', \tag{1}$$

donde $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_r)$ es una matriz diagonal donde $\lambda_1^2, \ldots, \lambda_r^2$ son los autovalores positivos de $X'X$, $U = [u_1, \ldots, u_r]$ es la matriz de autovectores ortonormales de $X'X$ y $V = [v_1, \ldots, v_r]$ es la matriz de autovectores ortonormales de XX', correspondientes a los autovalores $\lambda_1^2, \ldots, \lambda_r^2$.

Su importancia estadística se debe a Eckart y Young [3] y Householder y Young [13], que mostraron que si $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r$, el mejor ajuste en el sentido de mínimos cuadrados de la matriz X por una de rango $q \leq r$ viene dado por la matriz

$$X_{(q)} = V_{(q)}\Lambda_{(q)}U'_{(q)}, \tag{2}$$

con $\Lambda_{(q)} = \text{diag}(\lambda_1, \ldots, \lambda_q)$, $U_{(q)} = [u_1, \ldots, u_q]$ y $V_{(q)} = [v_1, \ldots, v_q]$, es decir, $X_{(q)}$ minimiza

$$||X - M||^2 = \sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ij} - m_{ij})^2.$$

Además, una medida absoluta de la bondad de este ajuste puede definirse por la proximidad a 1 de

$$||X_{(q)}||^2/||X||^2 = \sum_{\alpha=1}^{q}\lambda_\alpha^2 / \sum_{\alpha=1}^{r}\lambda_\alpha^2. \tag{3}$$

La DVS, o la descomposición espectral de una matriz simétrica que es un caso particular de DVS, es el fundamento de muchas de las técnicas de reducción y representación de datos, como el análisis de componentes principales, que pueden verse, por ejemplo, en Jolliffe [17] o Jackson [14], y el de coordenadas principales en Cuadras [2], algunos modelos factoriales descritos, por ejemplo, en Reyment y Jöreskog [19] o en Jambu [15], el análisis de correspondencias, descrito por Greenacre [11], ó la técnica de representación biplot, iniciada por Gabriel [7], entre otras.

En los últimos años pueden encontrarse numerosas referencias en la literatura sobre Análisis Cluster (Gnanadesikan [9], Gordon [10], Everitt [4]), 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, se utilizan para este propósito conjuntamente con otros procedimientos de obtención de grupos homogéneos, no sólo para indicar su presencia sino también para prevenir ante pretensiones excesivas de grupos producidas por procedimientos más complejos.

En este artículo pretendemos utilizar la DVS para proponer un procedimiento geométrico que permita obtener clusters naturales dentro de un conjunto de datos. Este procedimiento está inspirado en las representaciones geométricas que el biplot y, sobre todo, el modelo factorial, permiten hacer de cualquier conjunto de individuos.

2. EL MODELO DE REPRESENTACIÓN BIPLOT

El biplot es un modelo de representación basado en la descomposición (1), que fue introducido por Gabriel [7] y relacionado también posteriormente por Greenacre [11] con el análisis de correspondencias. Dada una matriz de datos $X_{n,p}$ (n individuos y p variables), el biplot proporciona una descripción conjunta, exacta o aproximada según el rango de X, de los n individuos y las p variables en dos dimensiones.

A partir de la descomposición (1), puede definirse una matriz diagonal Λ^{α} con $0 \leq \alpha \leq 1$, cuyos elementos sean $\lambda_1^{\alpha}, \ldots, \lambda_r^{\alpha}$, y otra similar $\Lambda^{1-\alpha}$ para formar las matrices

$$G = V\Lambda^{\alpha}, \qquad H' = \Lambda^{1-\alpha}U'. \tag{4}$$

Estas dos matrices permiten expresar X como

$$X = G \cdot H', \tag{5}$$

lo cual es equivalente a que cada elemento de X se pueda escribir como

$$x_{ij} = g_i' \cdot h_j, \qquad i = 1, \ldots, n; j = 1, \ldots, p \tag{6}$$

siendo g_i y h_j los vectores formados por las filas de G y H respectivamente.

Si, como hemos supuesto en la introducción, el rango de X es r, los vectores g_i y h_j tienen r componentes, permitiéndonos obtener una representación conjunta de las n filas y las p columnas de la matriz X en un espacio r-dimensional. Gabriel [7] denomina a los vectores g_i 'efectos fila' y a los h_j 'efectos columna' que intervienen de forma multiplicativa en cada elemento x_{ij} de X.

En el caso de que X tenga rango 2, esta representación puede hacerse en el plano, con las consiguientes ventajas de interpretación, denominándose a este tipo de representación *biplot* de X. En general, si rg(X)$>$ 2, las propiedades mínimo cuadráticas de la descomposición (1) nos permiten obtener una representación aproximada de X en el

plano, tomando las dos primeras componentes de g_i y h_j. Este tipo de aproximación puede ser utilizada también para la representación en espacios de dimensión mayor que 2, aunque éstas sean menos claras desde el punto de vista gráfico. Gabriel [8] se refiere a ellas como *bimodels*, reservando el término biplot para la representación en dimensión 2.

La no unicidad en la obtención del biplot de X dada por el escalar α en la definición de G y H en (4), puede evitarse dando valores particulares de α o imponiendo condiciones al modelo gráfico obtenido con los vectores g_i y h_j.

El biplot que más información puede darnos en cuanto a las relaciones entre las filas (individuos) de X, con vistas al descubrimiento de posibles grupos entre ellos, es el que se obtiene para $\alpha = 1$. Con este valor se tiene

$$G = V_{(2)}\Lambda_{(2)}, \qquad H' = U'_{(2)}, \tag{7}$$

que verifica

$$H'H = I_2 \tag{8}$$

Este biplot, además de (6), verifica que (Gabriel, [7]):

$$XX' = GG', \tag{9}$$

es decir, las relaciones entre las filas de X respecto de la métrica euclídea, pueden ser representadas por las de los vectores g con la misma métrica. Esta relación (9) permite obtener, para dos filas cualesquiera x_i y x_k de X, que:

$$x'_i x_k \sim g'_i g_k, \tag{10}$$

$$||x_i|| \sim ||g_i||, \tag{11}$$

$$\cos(x_i, x_k) \sim \cos(g_i, g_k), \tag{12}$$

$$||x_i - x_k|| \sim ||g_i - g_k|| \tag{13}$$

indicando (x,y) el ángulo entre los vectores x e y.

Las propiedades anteriores pueden ser utilizadas al inspeccionar el biplot definido por (7) cuando se pretenden encontrar clusters entre las filas de X. En concreto, vamos a analizar el ejemplo que nos proporcionan los conocidos datos de los tres tipos de Iris que Fisher [5] utilizó por primera vez en problemas de discriminación. Como se sabe, estos datos consisten en las medidas de las longitudes y anchuras (en mm.) 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 nosotros sólo utilizaremos aquí las 10 primeras plantas (Tabla 1) para analizar un conjunto más manejable. En ella se han ordenado los datos de forma que las 3 primeras flores

son del Tipo Setosa (A), las 3 siguientes del Tipo Versicolor (B) y las 4 últimas del Tipo Virgínica (C), como se aprecia en la última columna. Esta clasificación será ignorada en principio, pero nos ayudará posteriormente para ir analizando los resultados obtenidos.

Tabla 1. *Las 10 primeras flores Iris utilizadas por Fisher (1936)*

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

La Figura 1 muestra el biplot definido por (7) para la matriz $X_{10,4}$ de la Tabla 1 cuando se utilizan los datos centrados, es decir, medidos respecto de la media de cada columna. Es evidente que esta transformación no cambia las distancias relativas entre las flores. Se han representado las filas mediante el correspondiente número y las columnas mediante las correspondientes direcciones.

Figura 1. Biplot de los datos Iris centrados por columnas.

La inspección del biplot de la Figura 1 sólo nos permite descubrir una diferencia clara entre las flores del tipo A y las de los tipos B y C. Entre las flores de estos dos últimos tipos se aprecian ligeras diferencias, pero éstas no permiten obtener resultados claros en cuanto a posibles agrupaciones entre ellas, ni en cuanto a que coincidan estas agrupaciones con la clasificación de tipos.

Debe notarse que cualquier procedimiento que se utilice para formar clusters entre los individuos de un conjunto, debe partir de una medida de proximidad entre los individuos (distancia, disimilaridad, similaridad, etc.), y de una técnica clasificatoria (jerárquica, no jerárquica, etc.) que mediante un algoritmo apropiado obtenga los grupos.

En este caso, la medida de proximidad que estamos utilizando es la distancia euclídea entre los individuos, cuya representación aproximada nos proporciona el biplot, y cuyo cuadrado puede expresarse, en virtud de la regla del paralelogramo, como

$$\begin{aligned} d_{ik}^2 &= d_i^2 + d_k^2 - 2 \cdot d_i \cdot d_k \cdot \cos\theta_{ik} \\ &= (d_i - d_k)^2 + 2 \cdot d_i \cdot d_k \cdot (1 - \cos\theta_{ik}) \end{aligned} \tag{14}$$

donde hemos llamado a $||x_i - x_k|| = d_{ik}$, $||x_i|| = d_i$ y $||x_k|| = d_k$, siendo $||\cdot||$ la norma euclídea y θ_{ik} el ángulo que forman los vectores x_i y x_k. La igualdad (14) significa que la distancia entre dos individuos cualesquiera depende de la diferencia entre sus distancias al origen y del ángulo que forman. Dos individuos estarán próximos cuando sus distancias respectivas al origen sean muy parecidas y, además, el ángulo que formen sea próximo a 0°. En el biplot de la Figura 1, las flores 1 y 2 están próximas porque distan del origen aproximadamente lo mismo y forman un ángulo próximo a 0°, en cambio, las flores 5 y 8 no estan próximas porque aunque distan del origen aproximadamente lo mismo, el ángulo que forman es próximo a 120°, y tampoco están próximas las flores 6 y 8 aunque formen un ángulo próximo a 0° ya que tienen distancias al origen distintas.

La técnica clasificatoria debe actuar en base a las distancias entre los individuos y, por tanto, teniendo en cuenta las diferencias entre sus normas y los ángulos que forman. Es evidente que aunque el biplot es capaz de representar bien estas relaciones, esto no es suficiente para proporcionar una regla clara de clasificación entre los individuos. La decisión sobre el número de grupos y sobre la asignación de los elementos a cada grupo necesita algo más que la simple inspección de una representación biplot.

3. LOS MODELOS FACTORIALES EN MODO R Y Q

Reyment y Jöreskog [19] y Jambu [15], describen que la descomposición (1) permite también, dada la matriz de datos $X_{n,p}$, obtener un conjunto de direcciones principales

que recojan en orden decreciente la máxima variabilidad, tanto en el espacio de los individuos ($\mathbb{R}^p$) como en el de las variables ($\mathbb{R}^n$), considerados aquéllos y éstas como puntos o vectores dentro de dichos espacios, con coordenadas las filas o las columnas de X.

3.1. El modelo factorial para variables

En concreto, para la matriz $X_{n,p}$ previamente centrada por columnas, puede obtenerse un modelo factorial para las variables (o en modo R) de la forma

$$X_{n,p} = F_{n,q} \cdot (A_{p,q})' + E_{n,p} \tag{15}$$

tomando $F = V_q$ y $A = U_q \cdot \Lambda_q$. La matriz F representa por columnas un conjunto de variables ideales o factores, siendo las filas de F las puntuaciones que alcanzan los individuos en ellos. La matriz A se denomina matriz de pesos o cargas factoriales, y sus elementos permiten cuantificar la relacion entre las variables originales y los factores. Por último, la matriz E o de términos error, representa la información contenida en X que no es capaz de explicar el conjunto de factores de F.

Desde el punto de vista geométrico, el modelo (15) escrito sin la matriz E

$$X_{n,p} \approx F_{n,q} \cdot (A_{p,q})' \tag{16}$$

representa que cada variable puede escribirse como combinación lineal de los factores extraídos, es decir, el conjunto de variables podría considerarse que, salvo errores, está contenido en un subespacio de dimensión q cuya base ortonormal la forman los q factores. Nótese que la elección del número de factores no tiene por qué ser dos o tres.

Aunque la formulación (15) del modelo factorial no cumple con las hipótesis del modelo clásico en lo que se refiere a los factores específicos ([12]), sí lo hace en cuanto a los factores comunes, ya que la matriz de covarianzas S verifica que:

$$(n-1) \cdot S = X'X \approx AA'$$

La solución (15) es una de las técnicas más utilizadas para extraer una primera solución factorial ortogonal, de la que, posteriormente, pueden derivarse otras realizando rotaciones ortogonales u oblicuas con ella.

Conviene notar que, básicamente, los criterios analíticos que se utilizan para realizar las rotaciones están inspirados en las condiciones dadas por Thurstone [21] para obtener una estructura factorial más simple. Tanto si los factores rotados siguen siendo ortogonales como si se obtienen factores oblicuos, los *criterios de estructura simple de Thurstone* pretenden convertir los factores originales en otros que pasen,

aproximadamente por los posibles clusters de variables, de forma que cada una tenga pesos muy altos (positivos o negativos) en uno de los factores y prácticamente nulos en el resto (Cuadras, [2], pp. 169-173). Esta idea de agrupación de las variables a través de los factores es la que trataremos de utilizar para hacer lo mismo con los individuos de un conjunto, en el que se sospeche la existencia de grupos naturales.

3.2. El modelo factorial para individuos

Al igual que se hizo con las variables, puede pensarse en obtener de (1) un modelo factorial para los individuos (o en modo Q) de la forma

$$X_{n,p} = B_{n,q} \cdot (G_{p,q})' + E_{n,p} \tag{17}$$

tomando $B = V_q \cdot \Lambda_q$ y $G = U_q$, donde B, G y E representan lo mismo que en el modelo para variables, cambiando éstas por los individuos.

También en este caso, el modelo (17) escrito sin el término error E

$$X_{n,p} \approx B_{n,q} \cdot (G_{p,q})' \tag{18}$$

representa geométricamente que pueden encontrarse q factores (individuos ideales) respecto de los cuales los n individuos iniciales pueden expresarse como combinación lineal de ellos.

Debemos notar que al igual que en el modelo R-factorial la matriz de pesos A se obtuvo a partir de la descomposición espectral de $X'X$, matriz que salvo una constante representa las covarianzas entre las variables, en el modelo Q-factorial la matriz de pesos B se obtiene de la descomposición de XX', que representa los productos escalares de los vectores fila de X.

Además, de la forma de extraer la matriz de pesos B en (17) y la relación (2) se deduce que

$$B = X \cdot U_q,$$

es decir, inicialmente los elementos de B representan las coordenadas de los vectores fila de X respecto del sistema de autovectores U_q y, por tanto, esta matriz B representa los cosenos de los ángulos que forman cada individuo con cada factor (cosenos directores), multiplicados por las longitudes de tales vectores, es decir, si es x_i el vector de $\mathbb{R}^p$ que representa al individuo i,

$$\begin{aligned} b_{i,j} &= \text{coordenada de } x_i \text{ en la dirección } u_j \\ &= \cos\theta_{i,j} \cdot ||x_i|| \end{aligned}$$

con $\theta_{i,j}$ el ángulo que forma el vector x_i con el vector u_j.

Por otra parte, el modelo Q-factorial nos va a permitir rotar los factores, ortogonal u oblicuamente, con los mismos criterios utilizados con las variables (Cuadras, [2]), para obtener factores nuevos que determinen, aproximadamente, los posibles clusters existentes dentro del conjunto de individuos. Cuando se transforman los factores mediante rotaciones, en general oblicuas, la nueva matriz B^r (matriz de pesos rotada) sigue teniendo el mismo significado en cuanto a las coordenadas respecto de los nuevos factores, sin embargo no sirve para obtener los cosenos directores. Es necesario para estos casos definir una nueva matriz asociada al modelo factorial, denominada matriz de estructura factorial $S_{n,q}$, en la que sus elementos representan tales cosenos multiplicados por las longitudes de los vectores fila de X. La relación entre ambas matrices viene dada por

$$S_{n,q} = B^r_{n,q} \cdot \Theta_{q,q} \tag{19}$$

con $\Theta_{q,q}$ matriz de cosenos de los ángulos que forman los factores. Claramente, si $\Theta_{q,q} = I$ en el caso ortogonal, $S_{n,q} = B^r_{n,q}$.

Todo lo comentado anteriormente para las matrices de pesos y estructuras tiene la misma validez si partimos de una matriz de datos W en la que los vectores fila w_i son los de la matriz X normalizados. Los elementos de la matriz WW' son ahora los cosenos de los ángulos que forman cada pareja de filas de W o de X. Con la descomposición espectral de WW' se obtiene una nueva matriz de pesos, que seguiremos llamando B, y si aplicamos una rotación oblicua obtendremos las matrices B^r y S, siendo ahora los elementos de S los cosenos directores de los vectores fila de X respecto de los factores respectivos. Estos cosenos directores oscilarán entre -1 y 1, y tienen un significado claro en cuanto al grado de asociación de cada individuo con cada factor, de forma que podrían utilizarse para obtener una clasificación de los individuos según el factor al que se asocien con valores próximos a los extremos.

Hay que aclarar que el modelo Q-factorial obtenido al normalizar las filas de X no tiene una relación fácil de expresar respecto al que se obtiene con el obtenido sin normalizarlas. Lo que ocurre es algo parecido al problema de obtener las componentes principales de un conjunto de variables antes y después de su tipificación, es decir, con la matriz de covarianzas o con la de correlaciones. Los argumentos sobre la homogenización de la dispersión de las variables para usar la tipificación o la recomendación de que no se tipifiquen cuando tengan dispersiones parecidas, pueden considerarse con los individuos en cuanto a las normas euclídeas que presentan las filas de X.

A pesar de lo expuesto hasta ahora, el procedimiento que pasa por obtener el modelo Q-factorial de la matriz W, no es tampoco lo suficientemente adecuado para detectar las diferencias entre algunos clusters. Concretamente, pueden encontrarse ejemplos de conjuntos de datos (Figura 2) en los que, existiendo grupos, un análisis factorial en modo Q por el procedimiento expuesto antes no detectaría algunos de ellos. Sim-

plemente, el hecho de que exista más de un grupo (clusters 1 y 2) en alguna de las direcciones de máxima variabilidad, hará que el factor que determine tal dirección aglutine a todos los grupos que se encuentran en dicha dirección. También puede ocurrir que existan grupos de elementos (cluster 3) en los que todas las direcciones principales «pesen» por igual.

Figura 2. Ejemplo de conjunto de datos con grupos naturales entre los que algunos no pueden ser detectados con las asociaciones que establece el modelo Q-factorial con la matriz normalizada por filas.

Lo que ocurre en estos casos puede ser debido, por una parte, a que los individuos forman ángulos muy parecidos entre sí aunque tienen normas muy diferentes (en los casos 1 y 2) y, por tanto, los elementos de la matriz de estructuras asocian al mismo factor a individuos de clusters diferentes y, por otra, a que los individuos tienen normas muy parecidas y pequeñas en relación al resto, pero forman ángulos muy distintos (en el caso 3) y, por tanto, los elementos de la matriz de estructuras asocian a los individuos de un cluster a varios factores a la vez.

Figura 3. Ejemplos de datos que clasificaría mal el modelo Q-factorial con la matriz normalizada por filas.

La figura 3 aclara lo anterior. Los individuos representados por el vector 1 y 2 forman un ángulo próximo a 0° y, por tanto, tendrán un coeficiente de asociación próximo a 1 que hará que queden asociados al mismo factor, a pesar de que sus diferencias de longitud debería clasificarlos en factores distintos. Los individuos 1, 3 y 4 forman ángulos muy diferentes, pero sus longitudes son parecidas y pequeñas en relación a otras, lo cual aconsejaría que estuvieran asociados al mismo factor.

En el caso de las flores Iris, centradas por columnas, analizadas en la sección anterior, la Tabla 2 representa los valores de la matriz B obtenidos para el modelo (17) cuando se toman los 4 factores posibles y los valores singulares correspondientes. Nótese que si representamos las flores respecto de los 2 primeros factores obtendremos el biplot de la Figura 1. Además, la penúltima fila de la tabla indica el tanto por ciento de la variabilidad total que es capaz de explicar cada factor y la última el que explican los q primeros según (3). Por ejemplo, el biplot de la Figura 1 representa el 98.46% de la variabilidad presente en los datos.

Tabla 2. *Matriz de pesos iniciales B del modelo Q-factorial para los datos Iris centrados por columnas*

Iris	Factor 1	Factor 2	Factor 3	Factor 4
I1	−29.27	−0.99	1.87	−0.69
I2	−30.50	0.61	−1.58	0.71
I3	−34.46	2.16	0.05	0.19
I4	8.67	−2.33	2.29	−1.71
I5	7.39	−7.34	0.27	2.18
I6	8.78	2.84	−3.70	−0.18
I7	19.33	0.36	−2.32	1.09
I8	20.88	3.74	0.47	0.45
I9	12.11	−2.53	−1.74	−2.66
I10	17.06	3.47	4.38	0.62
Val. sing.	66.56	10.30	7.27	4.20
% Var. expl.	96.16	2.30	1.15	0.39
% Var. acum.	96.16	98.46	99.61	100.00

Cualquier clasificación de las flores que queramos obtener analizando las asociaciones de la Tabla 2, deben tener en cuenta los dos parámetros (14) de los que dependen las distancias relativas entre las flores: las diferencias de distancias al origen y el ángulo que forman. En su lugar queremos que tal clasificación dependa sólo de un parámetro: el ángulo que forman.

Con el mismo ejemplo, si después de centrar por columnas se normalizan los datos por filas, la Tabla 3 nos muestra la matriz de pesos inicial obtenida por (17) para los tres primeros factores, así como los valores singulares y los porcentajes de variabilidad total que representan.

Tabla 3. *Matriz de pesos inicial con los tres primeros factores para el modelo Q-factorial obtenido con los datos Iris, centrados por columnas y normalizados por filas*

Iris	Factor 1	Factor 2	Factor 3
I1	−0.99	−0.08	0.04
I2	−0.99	−0.003	−0.06
I3	−1.00	0.02	−0.001
I4	0.93	−0.25	0.21
I5	0.71	−0.66	−0.15
I6	0.87	0.39	−0.27
I7	0.98	0.07	−0.11
I8	0.97	0.19	0.07
I9	0.96	−0.11	−0.13
I10	0.93	0.16	0.29
Val. sing.	2.98	0.87	0.52
% Var. expl.	88.6	7.5	2.7
% Var. acum.	88.6	96.1	98.8

Podemos ahora utilizar los 2 o 3 primeros factores, ya que en ambos casos el porcentaje de variabilidad explicada es alto, y aplicar un criterio de rotación oblicua para obtener direcciones que clarifiquen más la agrupación de los individuos.

En este caso hemos utilizado el criterio de rotación oblicua oblimín directo ([16]) y hemos representado en la Tabla 4 las matrices de estructuras obtenidas con los 2 primeros factores (izquierda) y con los 3 primeros factores (derecha).

Una simple inspección de ambas matrices de la Tabla 4 nos permite ver que los posibles grupos que pueden determinar los factores, buscando los elementos con valor absoluto más alto en cada individuo son, por un lado, el formado por las flores del grupo Setosa (I1, I2 y I3) que tienen una asociación alta con la dirección negativa del Factor 1 y, por tanto, cosenos directores próximos a -1. Por otro lado, las flores de los grupos Versicolor y Virgínica (I4, I6, I7, I8, I9 y I10) que se asocian al mismo factor 1 en su dirección positiva sin posibilidad clara de distinción entre los dos grupos, mientras que la flor I5 del grupo Versicolor quedaría asociada a la dirección negativa del Factor 2.

Tabla 4. *Matrices de estructuras con dos y tres factores para el modelo Q-factorial con los datos Iris, centrados por columnas y normalizados por filas, con la clasificación obtenida según el mayor grado de asociación con los factores*

	Mat. de estr. (2 fact.)		Mat. de estr. (3 fact.)		
Iris	Factor 1	Factor 2	Factor 1	Factor 2	Factor 3
I1	-0.99	0.53	-0.99	0.58	0.01
I2	-0.98	0.59	-0.98	0.62	-0.11
I3	-0.98	0.62	-0.98	0.65	-0.06
I4	0.87	-0.75	0.87	-0.74	0.34
I5	0.58	-0.94	0.58	-0.96	0.15
I6	0.93	-0.21	0.93	-0.31	-0.35
I7	0.99	-0.53	0.98	-0.59	-0.06
I8	0.99	-0.43	0.99	-0.46	0.05
I9	0.92	-0.66	0.92	-0.71	-0.03
I10	0.95	-0.44	0.96	-0.43	0.26

Aparece en este ejemplo el problema que comentábamos anteriormente, debido a que los dos grupos Versicolor y Virgínica están asociados al mismo factor porque están muy próximos respecto a los ángulos que forman entre ellos y, a la vez, con la dirección positiva del primer factor principal. Veremos si una transformación previa de los datos puede resolver este problema.

4. UN PROCEDIMIENTO QUE PERMITE OBTENER CLUSTERS

Los problemas comentados en la sección anterior pueden tener solución. En esta sección se propone un procedimiento de obtención de clusters a partir de los factores obtenidos en el modelo Q-factorial propuesto en (17), sobre una matriz $W_{n,p+1}$ obtenida al realizar ciertas transformaciones sobre la inicial $X_{n,p}$. Dichas transformaciones están orientadas a evitar los problemas de carácter geométrico que impedían establecer una asociación uno a uno entre las distancias originales entre individuos y el coeficiente de asociación obtenido entre ellos después de centrar por columnas y normalizar por filas la matriz de datos.

4.1. Preparación de los datos. Obtención de W

Partiremos de un conjunto de datos cualquiera $X_{n,p}$, donde las variables analizadas sean de tipo cuantitativo y en el que estudios previos realizados sobre ellos hayan determinado la existencia de posibles agrupaciones naturales.

Para tal matriz X, nuestro objetivo va a ser encontrar un conjunto de factores o individuos ideales tales que podamos generar con ellos, aproximadamente, el conjunto original de datos y, al mismo tiempo, nos permitan encontrar las posibles agrupaciones que existan entre ellos.

La idea sobre la que trabajaremos es la siguiente: si en los datos existen grupos homogéneos, y utilizamos la métrica euclídea para determinar la proximidad entre los puntos que representan a los individuos, una buena forma de detectarlos sería que pudiéramos «verlos» desde una perspectiva apropiada (fuera de la nube de puntos que determinan), que nos permita descubrirlos con las direcciones que desde este punto exterior pasen por los centros de tales grupos. Esta idea intuitiva vamos a concretarla realizando algunas transformaciones sobre X que, aunque en un primer momento distorsionan las distancias originales entre ellos, veremos que permiten clasificarlos de acuerdo a las distancias originales que presentan.

En primer lugar, seguiremos suponiendo que la matriz de datos X está centrada por columnas, lo cual permitirá situar a la nube de puntos en torno al origen de coordenadas de $\mathbb{R}^p$.

En segundo lugar, incluiremos la nube de puntos de $\mathbb{R}^p$ en un espacio de una dimensión más, $\mathbb{R}^{p+1}$, añadiendo una columna de valores constante a X. Obtendremos una matriz

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

donde la última columna representa que se ha efectuado una traslación en $\mathbb{R}^{p+1}$ de cada vector que representa a los individuos mediante el vector $t = (0, \cdots, 0, \lambda)$ del espacio $\mathbb{R}^{p+1}$, es decir, se desplazarán a través de la dimensión *p+1* una longitud λ, de forma que este desplazamiento permita «ver» desde el origen los grupos homogeneos con mayor claridad.

En el tercer paso, normalizamos los vectores fila que representan a los individuos, para que las longitudes originales de cada vector fila no influyan en los elementos de la matriz de estructura. En este paso se obtienen las proyecciones de los vectores fila de T en la bola de radio unidad del espacio $\mathbb{R}^{p+1}$. La matriz que tiene por filas estas

proyecciones será

$$W = \begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1p} & w_{1,p+1} \\ w_{21} & w_{22} & \cdots & w_{2p} & w_{1,p+1} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ w_{n1} & w_{n2} & \cdots & w_{np} & w_{1,p+1} \end{pmatrix}$$

a la que le aplicaremos el modelo (17):

$$W_{n,p+1} \approx B_{n,q} \cdot (G_{p+1,q})' \tag{20}$$

La matriz B se obtiene ahora descomponiendo espectralmente WW', cuyos elementos son los productos escalares entre cada dos filas y, por tanto, los cosenos entre los ángulos que desde el origen forman los individuos desplazados en $\mathbb{R}^{p+1}$. Veremos a continuación que estos cosenos van a estar cerca de 1 (ó -1) para individuos próximos respecto a la métrica euclídea y cerca de 0 para individuos alejados entre sí.

Proposición 1:

Sean dos puntos cualesquiera P_i y P_k de $\mathbb{R}^p$, llamemos θ_{ik} el ángulo que forman los vectores OP_i, OP_k y d_i, d_k sus normas. Supongamos que incluimos estos puntos en un espacio de dimensión $p+1$, desplazándolos a través de la dimensión $p+1$ una distancia $\lambda > 0$, siendo OP_i' y OP_k' los nuevos vectores desde el origen, α_{ik} el ángulo que forman y t_i, t_k sus normas. Entonces

1. *Los valores del $cos\alpha_{ik}$ pueden expresarse como*

$$cos\alpha_{ik} = \frac{\lambda^2 + h d_i^2 cos\theta_{ik}}{\sqrt{d_i^2 + \lambda^2}\sqrt{h^2 d_i^2 + \lambda^2}} \tag{21}$$

   siendo $d_k = h \cdot d_i$, para $h > 0$

2. *Los valores del $cos\alpha_{ik}$, que oscilan están -1 y 1, están próximos a 1 para puntos P_i y P_k, situados fuera de un entorno del origen, próximos entre sí.*

3. *Los valores del $cos\alpha_{ik}$ son menores que $cos\theta_{ik}$ cuando los puntos P_i y P_k, situados fuera de un entorno del origen, tienden a estar alejados.*

4. *Los valores del $cos\alpha_{ik}$ están próximos a 1 para puntos P_i y P_k situados en un entorno del origen, independientemente del ángulo θ_{ik} que formen.*

Demostración:

1. Ayudándonos de la Figura 4,

Figura 4. Representación de elementos desplazados en $\mathbb{R}^{p+1}$.

se tiene que la expresión (14) puede escribirse indistintamente como

$$\begin{aligned} d_{ik}^2 &= d_i^2 + d_k^2 - 2d_i d_k \cos\theta_{ik} \\ &= t_i^2 + t_k^2 - 2t_i t_k \cos\alpha_{ik}. \end{aligned} \tag{22}$$

Pero como $t_i^2 = d_i^2 + \lambda^2$ y $t_k^2 = d_k^2 + \lambda^2$, sustituyendo en (22) y simplificando, se obtiene

$$-2d_i d_k \cos\theta_{ik} = 2\lambda^2 - 2\sqrt{d_i^2 + \lambda^2}\sqrt{h^2 d_i^2 + \lambda^2}\cos\alpha_{ik}$$

de donde, sustituyendo d_k por $h \cdot d_i$ y despejando $\cos\alpha_{ik}$ se llega a la expresión (21).

2. Para cualquier desplazamiento $\lambda > 0$ de los puntos en la dirección $p+1$, sea un punto P_i no contenido en un entorno del origen de $\mathbb{R}^p$, es decir con distancia al origen $d_i > M$.

   Los puntos próximos a P_i serán aquellos que estén en una cierta bola de centro P_i y radio ε : $B^p(P_i, \varepsilon)$. Para estos puntos debe verificarse que $d_{ik} = d(P_i, P_k) < \varepsilon$, y utilizando (14):

$$d_{ik}^2 = (d_i - d_k)^2 + 2 \cdot d_i \cdot d_k \cdot (1 - \cos(\theta_{ik})) = (h-1)^2 d_i^2 + 2h d_i^2 (1 - \cos(\theta_{ik})) < \varepsilon^2,$$

   lo cual implica que

$$(h-1)^2 + 2h(1 - \cos\theta_{ik}) < \frac{\varepsilon^2}{d_i^2} < \frac{\varepsilon^2}{M^2}, \tag{23}$$

donde $\frac{\varepsilon^2}{M^2}$ debe ser una cantidad próxima a 0 si ε es suficientemente pequeño en relación a M.

Por tanto, la desigualdad (23) nos indica que los puntos próximos a P_i son aquellos para los que los valores de h y del $\cos\theta_{ik}$ estan en un entorno de 1. Pero estos valores son transformados según la función (21), que es continua, en valores de $\cos(\alpha_{ik})$ próximos a 1, ya que se verifica que

$$\lim_{\substack{h \to 1 \\ \cos\theta_{ik} \to 1}} \frac{\lambda^2 + h d_i^2 \cos\theta_{ik}}{\sqrt{d_i^2+\lambda^2}\sqrt{h^2 d_i^2+\lambda^2}} = \frac{\lambda^2+d_i^2}{d_i^2+\lambda^2} = 1.$$

Esta última expresión indica que puntos inicialmente próximos entre sí van a tener un coseno próximo a 1 cuando son desplazados en $\mathbb{R}^{p+1}$.

3. Por otra parte, si seguimos considerando un punto P_i no contenido en un entorno del origen de $\mathbb{R}^p$, para cualquier valor de $\cos\theta_{ik}$ se verifica que

$$\lim_{h\to\infty} \frac{\lambda^2 + h d_i^2 \cos\theta_{ik}}{\sqrt{d_i^2+\lambda^2}\sqrt{h^2 d_i^2+\lambda^2}} = \frac{d_i \cos\theta_{ik}}{\sqrt{d_i^2+\lambda^2}} < \cos\theta_{ik},$$

lo cual indica que, independientemente del ángulo inicial que formen los puntos P_i y P_k, a medida que los puntos se alejen más, el valor de $\cos\alpha_{ik}$ será menor que el de $\cos\theta_{ik}$.

4. Por último, el límite

$$\lim_{d_i\to 0} \frac{\lambda^2 + h d_i^2 \cos\theta_{ik}}{\sqrt{d_i^2+\lambda^2}\sqrt{h^2 d_i^2+\lambda^2}} = 1$$

indica que con los puntos que se encuentran en un entorno del origen de coordenadas, vamos a obtener al desplazarlos cosenos próximos a 1, independientemente del ángulo inicial que formen.

■

La Figura 5 es una representación de las curvas de superficie obtenidas con el Software Mathematica [23] para el $\cos\alpha_{ik}$, cuando se toma $\lambda = 1$ y $d_i = 1$ y se hace variar h en [0,5] y θ_{ik} en $[-\pi, \pi]$. El sombreado de las superficies que encierran indican la evolución del valor del $\cos\alpha_{ik}$, desde 1 (color blanco) hasta 0 (color negro). Puede observarse cómo a medida que h se va haciendo más grande y θ_{ik} se aleja de 0°, las curvas se van oscureciendo.

Figura 5. Representación mediante curvas de superficies de $\cos\alpha_{ik}$.

Por otra parte, también puede comprobarse que estas curvas de superficie se alargan mucho más hacia la derecha cuando se aumenta el valor de d_i en relación al de $\lambda = 1$. Esto nos indica que cuando las distancias d_i son todas más grandes, las diferencias de distancias entre puntos próximos serán tambien más grandes y los valores del $\cos\alpha_{ik}$ seguirán siendo próximos a 1 incluso para valores de h mayores.

Estos resultados nos permiten utilizar los elementos de WW' en el modelo Q-factorial como medidas de similaridad que van a conservar la configuración inicial de los puntos. Las características del modelo factorial obtenido permiten establecer una regla clara de clasificación entre los elementos, obteniendo al mismo tiempo el número de posibles clusters y la asignación de los elementos a ellos.

El modelo (17), para la matriz W, obtiene un conjunto de q factores, cuyo número puede venir determinado por la proximidad a 1 del coeficiente (3). Cada uno de estos factores tiene posibilidad de determinar dos clusters, uno en su dirección positiva y otro en la negativa. Denotemos estas direcciones por $g_1^+, g_2^+, \cdots, g_q^+, g_1^-, g_2^-, \cdots, g_q^-$.

Una vez obtenidos los factores, la matriz de estructuras $\mathbf{S}= (s_{ij})$ que se obtiene después de realizar una rotación oblicua, será la que determine el número de clusters y asigne cada elemento a su grupo correspondiente. En concreto, para cada elemento, la estructura máxima en valor absoluto determinará la asignación a la dirección correspondiente, es decir:

$$I_i \in \begin{cases} g_j^+ & \text{si} \quad \max_{p=1,\cdots,q} |s_{ip}| = s_{ij} \\ g_j^- & \text{si} \quad \max_{p=1,\cdots,q} |s_{ip}| = -s_{ij} \end{cases}$$

4.2. Obtención del desplazamiento de la nube de puntos

En las transformaciones efectuadas a la matriz $X_{n,p}$ hasta llegar a la $W_{n,p+1}$ se ha dejado un punto sin resolver: ¿qué distancia desplazaremos los puntos en la dimensión $p+1$?

Si recordamos que la traslación a través de esta dimensión la hemos realizado con el objetivo de 'tener una perspectiva mejor' desde el origen para detectar los posibles grupos homogéneos en los datos, parece lógico que a la hora de determinar cuánto debemos desplazarnos, es decir el valor de λ, lo hagamos de forma que tal valor haga máxima la dispersión que presentan los puntos desde el origen.

Pueden tomarse varias formas de medir la dispersión que presentan los datos. La variación total, obtenida con la traza de la matriz de varianzas-covarianzas de W podría ser un buen coeficiente si no fuera porque va a ser siempre n el número de filas de X, independientemente del valor de λ. En efecto,

$$\mathrm{tr}(W'W) = \sum_{j=1}^{p+1}\sum_{i=1}^{n} w_{ij}^2 = \sum_{i=1}^{n}\sum_{j=1}^{p+1} w_{ij}^2 = \sum_{i=1}^{n} 1 = n$$

En su lugar, Mardia et al. [18] recogen el concepto de varianza generalizada que Wilks [22], y anteriormente también Frisch [6], definieron y que utilizaremos para este propósito.

Definicion (Mardia et al., [18], p.13)

Dado un vector aleatorio con p variables observadas sobre n individuos y dada la matriz de varianzas-covarianzas muestrales $S_{pp} = (s_{jk})$, siendo

$$s_{jk} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)(x_{ik} - \overline{x}_k) \qquad j,k = 1,\cdots,p,$$

se define la varianza generalizada como el determinante de S

$$VG = \det(S) = \mid S \mid$$

La varianza generalizada es una medida que resume la dispersión general de las p variables en el conjunto de los n individuos. Puede ser utilizado para determinar el desplazamiento imponiendo que el valor λ lo haga máximo para la nube de puntos determinada por W. Se probará a continuación que este máximo existe para una función que es múltiplo de la varianza generalizada de W: $f(\lambda) = \det(W'W)$

Proposición 2:

Dada la matriz $X_{n,p}$ centrada por columnas, $T_{n,p+1}$ la obtenida al trasladar los datos por una dimensión más, cuya última columna tiene por elementos el parámetro λ, y la matriz normalizada por las filas $W_{n,p+1}$, la dispersión que presentan los datos en $\mathbb{R}^{p+1}$ desde el origen, es directamente proporcional a $det(W'W)$. Además, la función $f(\lambda) = det(W'W)$ alcanza un máximo finito en $[0,\infty]$.

Demostración:

Con los datos centrados, considerados como puntos de $\mathbb{R}^p$, la varianza generalizada es el determinante de la matriz de covarianzas $S = (1/n)X'X$.

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

Cuando λ tiende a ∞, tal volumen dependerá de

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

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

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

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

Tenemos, por tanto, que la función $f(\lambda) = \det(W'W)$ toma el valor $f(0) = 0$, y para un cierto valor k en adelante puede hacerse tan pequeña como se quiera.

Con esto, la función $f(\lambda)$, que es continua en el intervalo [0,k] por ser sumas y productos de funciones continuas, está acotada, con lo que se sigue por continuidad que debe alcanzarse un máximo absoluto en tal intervalo, es decir que tendremos un valor de λ para el cual la nube de puntos desde el origen tendrá mayor dispersión.

■

4.3. Análisis del ejemplo de los datos Iris mediante el procedimiento propuesto

Los resultados obtenidos con las 10 primeras flores Iris, para la matriz T y W son

$$T = \begin{pmatrix} -9.1 & 2.7 & -25.1 & -11.9 & 6.7 \\ -13.1 & 3.7 & -25.1 & -10.9 & 6.7 \\ -13.1 & 5.7 & -29.1 & -11.9 & 6.7 \\ 5.9 & -2.3 & 6.9 & 1.1 & 6.7 \\ 2.9 & -8.3 & 5.9 & 1.1 & 6.7 \\ -0.1 & 1.7 & 8.9 & 4.1 & 6.7 \\ 4.9 & -2.3 & 16.9 & 8.1 & 6.7 \\ 7.9 & 0.7 & 16.9 & 10.1 & 6.7 \\ 3.9 & -2.3 & 11.9 & 1.1 & 6.7 \\ 9.9 & 0.7 & 11.9 & 9.1 & 6.7 \end{pmatrix} \quad W = \begin{pmatrix} -0.30 & 0.09 & -0.83 & -0.39 & 0.22 \\ -0.42 & 0.12 & -0.80 & -0.35 & 0.21 \\ -0.37 & 0.16 & -0.82 & -0.34 & 0.19 \\ 0.51 & -0.20 & 0.59 & 0.09 & 0.58 \\ 0.23 & -0.66 & 0.47 & 0.08 & 0.53 \\ 0.00 & 0.14 & 0.74 & 0.34 & 0.55 \\ 0.24 & -0.11 & 0.82 & 0.39 & 0.32 \\ 0.35 & 0.03 & 0.76 & 0.45 & 0.30 \\ 0.27 & -0.16 & 0.82 & 0.07 & 0.46 \\ 0.51 & 0.03 & 0.62 & 0.47 & 0.35 \end{pmatrix}$$

El cálculo de la función a maximizar se realiza a través de la obtención de los autovalores de $W' \cdot W$ y la maximización de su producto. Ambas tareas se realizan con subrutinas de FORTRAN de la librería IMSL.

La Tabla 5 contiene las distancias cuadradas originales de las 10 flores (triangular inferior) y los coeficientes de asociación dados en (21). En ella pueden apreciarse cómo estos coeficientes se aproximan a 1 tanto más cuanto que las distancias son menores.

Tabla 5. *Matriz de distancias cuadradas originales (triangular inferior) y de similaridades (triangular superior) entre las 10 flores transformadas en W*

Iris	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10
I1		.991	.993	−.558	−.425	−.599	−.836	−.833	−.689	−.759
I2	18		.997	−.615	−.465	−.570	−.836	−.841	−.709	−.791
I3	41	21		−.614	−.494	−.581	−.842	−.834	−.721	.773
I4	1443	1565	1890		.846	.769	.854	.842	.936	.874
I5	1395	1505	1846	46		.583	.717	.618	.810	.619
I6	1494	1554	1885	65	127		.909	.895	.875	.828
I7	2385	2485	2904	150	210	121		.980	.937	.930
I8	2541	2655	3066	194	308	165	22		.892	.976
I9	1732	1838	2203	29	73	50	75	131		.847
I10	2175	2307	2676	114	230	135	60	30	109	

Al obtener con W la solución dada por (17) con los tres primeros factores, se obtiene una matriz de estructuras dada en la Tabla 6, donde vemos cómo los Iris que tienen estructura más alta con el factor primero son del 6° al 10°, los que se asocian con el segundo factor son del 1° al 3°, y con el tercer factor el 4° y 5°. Puede establecerse así una clasificación originada por los factores, que como puede comprobarse agrupa

a los Iris del grupo Setosa en el segundo factor, a dos Iris del grupo Versicolor en el tercero y a los cuatro del grupo Virginica y uno del Versicolor en el primero.

Tabla 6. *Matriz de estructuras después de rotar con tres factores en el modelo Q-factorial para los datos Iris transformados en* **W**, *con la clasificación obtenida*

Iris	Factor 1	Factor 2	Factor 3
I1	−0.62	0.95	0.37
I2	−0.59	0.94	0.41
I3	−0.60	0.94	0.44
I4	0.80	−0.54	−0.87
I5	0.51	−0.36	−0.94
I6	1.00	−0.59	−0.60
I7	0.91	−0.80	−0.69
I8	0.93	−0.82	−0.61
I9	0.85	−0.66	−0.81
I10	0.90	−0.77	0.62

Combinando la clasificación original de los datos originales con la obtenida se tiene la Tabla 7, la cual pone en evidencia que, a excepción del Iris 6°, hecho que era de esperar si se observan sus distancias respectivas en la Tabla 5, los demás han sido bien clasificados en un 90% de los casos.

Tabla 7. *Correspondencia entre clasificación original y clasificación obtenida con el procedimiento propuesto para los datos Iris*

Iris	Grupo	Clasif.
I1	A	g_2^+
I2	A	g_2^+
I3	A	g_2^+
I4	B	g_3^-
I5	B	g_3^-
I6	B	g_1^+
I7	C	g_1^+
I8	C	g_1^+
I9	C	g_1^+
I10	C	g_1^+

Cuando se ha utilizado el procedimiento con las 150 flores *Iris* (Fisher, 1936), 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. Es evidente que el procedimiento de clasificación que se ha propuesto no sólo depende del número de factores que se tomen en el modelo Q-factorial, sino de todos aquellos parámetros que influyan al detectar separaciones entre los grupos: dimensión de los datos, número de grupos, número de elementos en cada grupo, etc.

5. ESTUDIO DE SIMULACIÓN DEL PROCEDIMIENTO Y COMPARACIÓN CON EL MÉTODO KM

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 existan 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 sin quitar generalidad y aleatoriedad a la muestra elegida. Para ello, utilizando la generación de números aleatorios entre 0 y 1 se ha obtenido para cada uno de los 100 conjuntos:

1. Un número aleatorio entre 1 y 5 que determine la dimensión del espacio inicial de los datos.
2. Un número aleatorio entre 2 y 6 que determine el número de grupos definido en cada conjunto.
3. Para cada grupo, un punto aleatorio del espacio determinado que será considerado como centro del grupo, exigiendo que la distancia mínima entre dos centros cualesquiera dentro de un conjunto sea al menos de 3 unidades.
4. Por último, en cada grupo se elije un número de elementos, con la restricción de que el conjunto en total tenga 50, y se han obtenido las coordenadas de los elementos sumando o restando a las coordenadas del centro un número aleatorio entre 1 y 3, calculado mediante un coeficiente de dispersión que se determina para cada grupo y coordenada.

Cuando se ha utilizado el procedimiento descrito en la sección 4 con los 100 conjuntos de elementos generados, hemos realizado los cálculos para distintas cotas impuestas a los autovalores de $W'W$ que, como se sabe, limitará en mayor o menor medida el número de factores a extraer en el modelo Q-factorial. Así, la Tabla 8 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.

Tabla 8. *Porcentaje de elementos bien clasificados de la muestra generada, según el porcentaje mínimo de variabilidad exigido a cada factor elegido en el modelo*

% EXIGIDO	1	5	10	15	20	25	30
% BIEN CLASIF.	83.42	85.24	85.86	85.12	84.36	82.88	82.88

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 total. 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 y, por tanto, más clusters. 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, como hemos dicho anteriormente, este procedimiento puede también utilizarse para determinar el número de grupos en un conjunto. Si analizamos las direcciones (positivas o negativas) que han intervenido en cada uno de los 100 conjuntos para determinar las clasificaciones, podemos enfrentar en la Tabla 9 los parámetros número de grupos inicial con el número de direcciones significativas:

Tabla 9. *Número de grupos inicial frente a número de direcciones significativas obtenidas en los 100 conjuntos generados*

	N° de grupos inicial				
N° dir.	2	3	4	5	6
2	21	25	4	5	
3		15	12	2	2
4			5	4	
5				1	

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 ha obtenido el mismo número de grupos iniciales o uno menos.

Por otra parte, se ha confeccionado un algoritmo para obtener soluciones que clasifiquen a cada uno de los 100 conjuntos generados, utilizando el procedimiento no jerárquico de minimizar la traza de la matriz de dispersión dentro de un número de grupos dado (Everitt, [4]). Este criterio es equivalente al de colocar cada individuo en el cluster cuyo centro esté más próximo a él (algoritmo de K-medias), cuya partición resultante depende del número de clusters del que partamos y, en ocasiones (Blashfield, [1]), de la partición inicial de la que se parta.

Los resultados obtenidos con este algoritmo, teniendo en cuenta que se ha iniciado con la partición correcta en cada conjunto, muestran que no ha sido capaz de clasificar bien al 100 % de los elementos como debería 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 estos resultados con los obtenidos con el procedimiento propuesto, vemos que el de minimizar la traza se comporta de forma parecida cuando en el primero se toman los autovalores capaces de explicar al menos el 10 % de la dispersión de los datos, donde se obtenían el 85.86 % de datos bien clasificados. Pero además, hay que tener en cuenta que el primero tiene, por un lado, la ventaja sobre el segundo de que no necesita el conocimiento previo del número de grupos para obtener este porcentaje de clasificaciones y, por otro lado, que los resultados obtenidos no dependen de ninguna partición u ordenación inicial de los elementos.

6. CONCLUSIONES

Se ha propuesto un procedimiento de clasificación en el que, utilizando algunas transformaciones geométricas y el modelo Q-factorial, se ha conseguido dejar la decisión del número de clusters y la asignación de elementos al propio procedimiento. Los resultados, sin ser excelentes, no son peores que los de uno de los métodos más utilizados dentro del análisis cluster, con la ventaja de que no tenemos que partir de ningún conocimiento sobre el número de clusters.

No obstante, es evidente que el procedimiento depende del criterio de elección de los factores en el modelo Q-factorial. Los estudios de simulación parecen aconsejar que sean elegidos aquéllos que expliquen al menos entre un 5% y un 15% de la variabilidad total.

Además, el número de clusters que puede determinar está limitado por el doble de las direcciones factoriales (positivas y negativas) obtenidas, que dependen a su vez del número de variables. Esto desaconseja que el procedimiento pueda utilizarse en

conjuntos de datos donde el número de variables utilizadas sea pequeño en relación con el de posibles clusters o se sospeche la existencia de un número de clusters muy superior al de variables. De todas formas, futuros estudios pueden perfilar mejor todas estas características.

AGRADECIMIENTOS

Los autores agradecen al Profesor C.M. Cuadras las numerosas sugerencias realizadas a lo largo de su elaboración definitiva.

Así mismo, este trabajo ha sido parcialmente financiado por el proyecto de investigación PS96-1436 de la DGICYT, Ministerior de Educación y Cultura.

REFERENCIAS

[1] **Blashfield, R.K.** (1976). «Mixture model tests of cluster analysis. Accuracy of four agglomerative hierarchical methods». *Psychol. Bulletin*, **83**, 377-385.

[2] **Cuadras, C.M.** (1991). *Métodos de Análisis Multivariante*. 2ª ed. PPU, Barcelona.

[3] **Eckart, C.** and **Young, G.** (1936). «Approximation of one matrix by another of lower rank». *Psychometrika*, **1**, 211-218.

[4] **Everitt, B.** (1993). *Cluster Analysis*, 3ª ed., Edward Arnold.

[5] **Fisher, R.A.** (1936). «The use of multiple measurements in taxonomic problems». *Ann. Eugen.*, **7**, 179-188.

[6] **Frisch, R.** (1929). «Correlation an scatter in statistical variables». *Nordic Statistical Journal*, **8**, 36-102.

[7] **Gabriel, K.R.** (1971). «The biplot-graphic display of matrices with applications to principal component analysis». *Biometrika*, **58**, 453-467.

[8] **Gabriel, K.R.** (1981). «Biplot display of multivariate matrices for inspection of data and diagnosis». In *Interpreting Multivariate Data* (ed. V. Barnett), 147-173. Wiley, Chichester.

[9] **Gnanadesikan, R.** (1977). *Methods for Statistical Data Analysis of Multivariate Observations*. Wiley, New York.

[10] **Gordon, A.D.** (1981).*Classification*. Chapman and Hall, London.

[11] **Greenacre, M.J.** (1984). *Theory and Applications of Correspondence Analysis*. Academic Press, London.

[12] **Harman, H.H.** (1976). *Modern Factor Analysis*. The University of Chicago Press.

[13] **Householder, A.S.** and **Young, G.** (1938). «Matrix approximation and latent roots». *Am. Math. Monthly*, **45**, 165-71.

[14] **Jackson, J.E.** (1990). *A User's Guide to Principal Components*. Wiley, New York.

[15] **Jambu, M.** (1991). *Exploratory and Multivariate Data Analysis*. Academic Press.

[16] **Jennrich, R.I.** and **Sampson, P.F.** (1966). «Rotation for simple loadings». *Psychometrica*, **31**, 313-323.

[17] **Jolliffe, I.T.** (1986). *Principal Component Analysis*. Springer-Verlag New York Inc.

[18] **Mardia, K.V., Kent, J.T.** and **Bibby, J.M.** (1979). *Multivariate Analysis*. Academic Press, London.

[19] **Reyment, R.** and **Jöreskog, K.G.** (1993). *Applied Factor Analysis in the Natural Sciences*. 2ª ed. Cambridge U.P.

[20] **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 orthogonal substitution». *Messenger of Mathematics*, **19**, 42-46.

[21] **Thurstone, L.L.** (1947). *Multiple factor Analysis*. The University of Chicago Press.

[22] **Wilks, S.S.** (1932). «Certain generalizations in the analysis of variance». *Biometrika*, **24**, 471-494.

[23] **Wolfram, S.** (1991). *Mathematica: A System for Doing Mathematics by Computer*. 2ª ed. Addison-Wesley.

ENGLISH SUMMARY

A PROCEDURE OF CLUSTERING USING THE SVD OF A MATRIX. COMPARISONS WITH THE BIPLOT AND WITH THE Q-FACTOR MODEL

JUAN L. GONZÁLEZ CABALLERO*
MARIANO J. VALDERRAMA BONNET**

In this paper we intend to use the SVD to propose a geometric procedure that allows us to obtain natural clusters within a data set. It's inspired by the geometric representations that the biplot and, especially, the Q-factor model allow us to make of any individual or object set. It transforms first the $\mathbf{X}_{n,p}$ *data matrix in* $\mathbf{W}_{n,p+1}$*, so that on obtaining the Q-factor model of* **W***, natural sets with greater clarity appear. Besides centering the* **X** *data in columns, they move through another dimension and they are normalized in rows (***W** *matrix). The distance of displacement in the* $(p+1)$ *dimension is looked for to maximize det(***W***'***W***), in order to maximize the generalized variance. The results that the paper include show that this distance exists and is finite and that the elements of* **WW***' matrix, that represent the cosins between the new row vectors, are next to 1 for individuals of a same cluster and far from 1 for individuals of different clusters. The paper finishes with analyzing the Iris data (Fisher[5]) with the proposed procedure, and with seeing its behaviour with a random sample of data sets with clusters, and subsequently its efficacy compared with the nonhierarchical clustering procedure of minimizing the trace of the within-group dispersion matrix.*

Keywords: Singular value decomposition; factor models; biplot representation; generalized variance; nonhierarchical clustering methods.

AMS Classification (1991): 62H25, 62H30.

* Juan L. González Caballero. Dpto. de Matemáticas. Facultad de Medicina. Universidad de Cádiz.

** Mariano J. Valderrama Bonnet. Dpto. de Estadística e I.O. Facultad de Farmacia. Universidad de Granada.

– Received May 1996.
– Accepted September 1997.

The Singular Value Decomposition (SVD) of a matrix $X_{n,p}$ is a result described for the first time by the English mathematician Sylvester (1889), which allows us to write X in (1), using the principal directions obtained in the $\mathbb{R}^p$ and $\mathbb{R}^n$ spaces, where the row and column vectors of the X matrix can be represented, respectively. Its statistic importance is owed to Eckart & Young (1936) and Householder & Young (1938), who showed its use in obtaining the best least-squares fit of the X matrix for one matrix of less rank.

The SVD, and the spectral decomposition of a square matrix that is a special case of the SVD, are the foundation of many reduction techniques and data representation, as principal components, principal coordinates, some factor models, the correspondence analysis or the technique of biplot representation.

In the last few years, numerous references can be found in the literature about Cluster Analysis (Gnanadesikan (1977), Gordon (1981), Everitt (1993)), in wich the use of geometric procedures of data representation are suggested that, in most cases, refer to some of the previous techniques. Although none of tem are specifically designed for to indicate the presence of clusters within data, they are used for this purpose together with other procedures for obtaining homogenous groups, not only to indicate its presence but also for preventing excesive claims for cluster structure produced by more complex techniques clustering.

In this paper we intend to use the SVD to propose a geometric procedure that allows us to obtain natural clusters within a data set. This procedure is inspired by the geometric representations that the biplot and, especially, the Q-factor model allow us to make of any individual or objet set.

In section 2, we present, in outline, the biplot representation model with the main results (4) to (13), that allow us to use it to obtain representations in the plan of element set that, sometimes, let us discover homogeneous. With the first ten Iris flowers (Table 1), used by Fisher (1936) in discrimination problems (3 of type Setosa, 3 of type Versicolor and 4 of type Virginica), the biplot representations are obtained when the data for columns are centred (Figure 1). This representation show that some natural sets can be discovered because the 3 Setosa flowers are separated from the rest.

In section 3, the R and Q factor models are introduced by means of SVD. Through the structure matrix obtained after rotating the solution (17) with the oblique rotation criterion proposed by Jennrich and Sampson (14), the Q-factor model of centred by columns and normalized in rows matrix, allows us to obtain (Table 4) practically the same natural sets as with the biplot representation.

In section 4, we propose a procedure that transforms first the $X_{n,p}$ data matrix in $W_{n,p+1}$, so that on obtaining the Q-factor model of W, natural sets with greater cla-

rity appear. Besides centering the X data in columns, they move through another dimension (T matrix), and they are normalized in rows (W matrix). The distance of displacement in the $(p+1)$ dimension is looked for to maximize det($W'W$), in order to maximize the generalized variance. The results that includes the section show that this distance exists and is finite and that the elements of WW' matrix, that represent the cosins between the new row vectors, are next to 1 for individuals of a same cluster and far from 1 for individuals of different clusters. The section finish on analyzing the Iris data with the proposed procedure again, and we obtain a structure matrix (Table 6) for the Q-factor model of W which is capable of clasifyng well 90 % of the elements.

In section 5, in order to evaluate the proposed procedure, we have seen its behavior with a random sample of data sets with clusters, and subsequently its efficacy compared with the nonhierarchical clustering procedure of minimizing the trace of the within-group dispersion matrix. Here, we can see that the percentages of well classified elements with boths procedures are very similar.