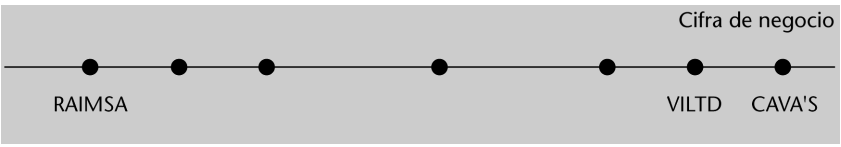
2. Análisis de componentes principales

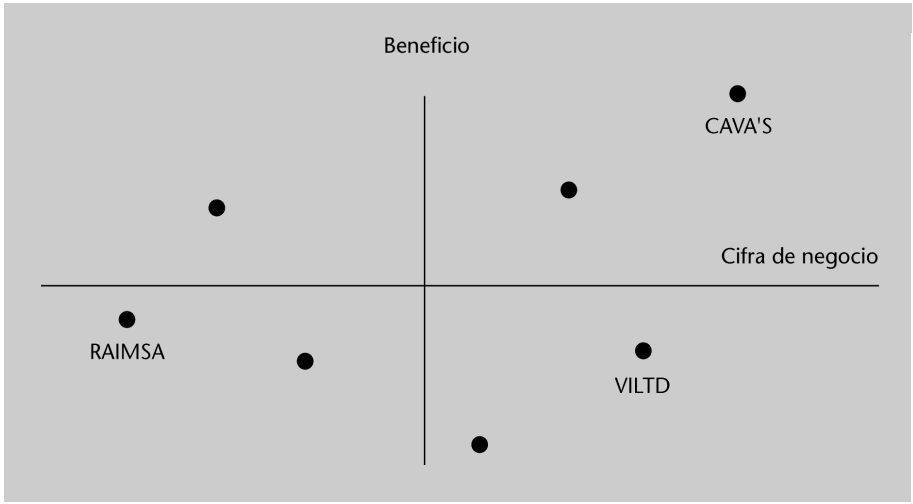
2.1. Introducción

Cuando obsevamos pocas variables en un colectivo de elementos, es relativamente cómodo ordenar los datos y hacer grupos de comportamiento homogéneo. La cuestión se complica cuando el número de variables observadas es tan grande que no permite una lectura fácil a partir de las simples representaciones gráficas o de las medidas clásicas de descripción. Es entonces cuando hay que utilizar métodos de síntesis de la gran cantidad de información disponible, reducir el número de variables y poner la información más al alcance del analista.

Si sobre todo el censo o sobre una muestra de empresas viticultoras del Alto Penedés calculásemos la cifra de negocio de cada una, podríamos representar los resultados en un diagrama de puntos:



Si hubiésemos observado dos variables: cifra de negocio y beneficios, también sería fácil su representación gráfica como una nube de puntos en un plano:



Incluso con tres variables (cifra de negocio, beneficios y gastos en publicidad) podríamos intentar realizar una clasificación de las empresas que las agrupase de acuerdo con estos tres criterios, sin embargo, no olvidemos que, a medida que aumentase el número de indicadores, también aumentaría la dificultad de describir el comportamiento de las empresas.

Es lógico pensar que este tipo de estudios es multidimensional y un análisis exhaustivo exigiría la observación de muchas variables: costes, plantilla, inversiones, márgenes comerciales, gastos de promoción, existencias, etc. La lista de indicadores puede ser tan larga como se quiera; cuantas más variables haya, más información y, a la vez, más complicación tendrá el analista. Ahora, las técnicas clásicas de descripción ya aprendidas son insuficientes; hay que recurrir a métodos que disminuyan la dimensionalidad del estudio, que lo hagan más fácil y que también retengan la mayor parte de la información contenida en las variables observadas inicialmente. Uno de estos métodos es el análisis de componentes principales (ACP).

El análisis de componentes principales...

... presenta numerosas aplicaciones en el marketing, como pueden ser la segmentación de mercados, las tipologías de productos y de empresas y las preferencias de los consumidores. La metodología del ACP es uno de los instrumentos más valiosos en los estudios de mercado.

Si después de estudiar una veintena de variables sobre las empresas viticultoras del Alto Penedés fuésemos capaces de reducirlas, por ejemplo, sólo a tres indicadores de síntesis de todas las variables observadas: dimensión de la empresa, productividad y análisis financiero, habríamos conseguido hacer comprensibles los resultados.

Las cuestiones que surgen ahora son:

1) Al pasar de las ocho variables iniciales a tres indicadores nuevos, se pierde una parte de la información que tenemos (lógicamente, se perderá más información cuanto más ejes queramos utilizar).

2) La etiqueta o concepto que asociamos a los indicadores nuevos no viene dada a priori, sino que se les atribuye un significado después de observar la relación funcional entre componentes nuevos (indicadores) y las variables iniciales, lo cual no siempre es fácil.

3) Una ventaja que se deriva de esta reducción de ejes es que ahora los nuevos componentes son independientes entre sí y este hecho es importante porque anula la posibilidad de que se sobrepongan conceptos.

Actividad

2.1. Imaginaos que queréis describir a los estudiantes matriculados en la UOC en la diplomatura de Empresariales. ¿Qué variables podríais utilizar? Al final saldría una lista larguísima: edad, altura, número de calzado, ..., asignaturas elegidas, horas de estudio, conexiones realizadas, .., nivel de renta, gastos de ocio, ..., inteligencia, agresividad, ...

Suponiendo que nos limitemos únicamente a cuestiones académicas, indicad una docena de variables que sean objetivamente evaluables, pasad la encuesta a un grupo de compañeros y haced una lista de los resultados. Comprobad cómo una información tan amplia sobrepasa el ojo clínico de cualquier analista, aunque sea experimentado.

¿Se da duplicidad en la información por una cierta redundancia en las preguntas? Las correlaciones entre las variables observadas os pueden dar la respuesta. Ahora es necesario que argumentéis la necesidad de hacer más fácil el estudio disminuyendo el número de variables que hay que utilizar y evitando duplicidades en las cuestiones.

En este apartado del análisis de componentes principales aprenderéis:

- Cuál es el objetivo del análisis de los componentes principales: la reducción de la dimensionalidad de los datos.

- Cuál es el procedimiento para la obtención de los componentes principales.
- Cómo se interpretan los resultados obtenidos en el análisis.

2.2. Matriz de datos y objetivos del análisis

Se tiene una muestra (o población) de I elementos en los cuales se han medido J variables con el objetivo de explicar un comportamiento determinado o de agruparlos en categorías y se ha llegado a la matriz de información (X_{ij}), con las variables dispuestas por columnas y los elementos por filas:

	X_1	X_2	...	X_j	...	X_J
I_1	X_{11}	X_{12}	...	X_{1j}	...	X_{1J}
I_2	X_{21}	X_{22}	...	X_{2j}	...	X_{2J}
...	...	...	...	...	...	...
I_i	X_{i1}	X_{i2}	...	X_{ij}	...	X_{iJ}
...	...	...	...	...	...	...
I_I	X_{I1}	X_{I2}	...	X_{Ij}	...	X_{IJ}

El análisis de componentes principales pretende reducir la dimensionalidad de la matriz de datos hasta conseguir un número inferior de variables nuevas (Z_j) o componentes principales con las características siguientes:

– Los componentes principales son combinaciones lineales de las variables originales.

– Los componentes principales no están en correlación entre sí.

– El número de componentes principales debe ser, a la vez, pequeño (para que el análisis sea eficaz) y suficiente (para absorber la mayor parte de la información de las variables iniciales).

Se trata, pues, de una técnica de condensación de datos en la que:

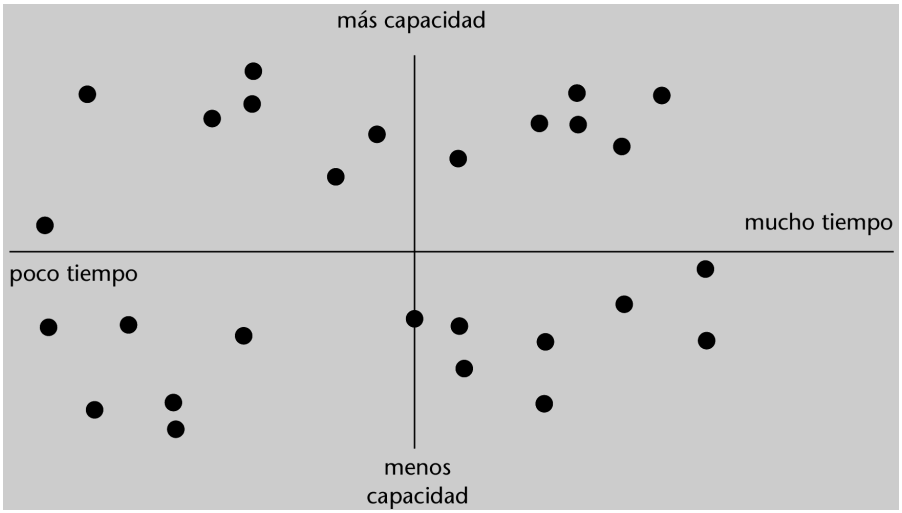
$$\left.\begin{array}{l} Z_1 = f_1(X_1, X_2, \ldots, X_J) \\ Z_2 = f_2(X_1, X_2, \ldots, X_J) \\ \quad\ldots\ldots\ldots\ldots \\ Z_J = f_J(X_1, X_2, \ldots, X_J) \end{array}\right\} \begin{array}{l} f_j \text{ es combinación lineal} \\ \left.\begin{array}{l} f_j \text{ y } f_{j'} (Z_j \text{ y } Z_{j'}) \text{ son independientes} \end{array}\right\} \end{array}$$

Nos planteamos encontrar estos componentes en una cantidad suficiente para hacer viable el estudio, cómoda la lectura y alta la capacidad explicativa.

Después de haber realizado la actividad 2.1, imaginaos qué fácil sería si hubiese resultado que la mayor parte de la información contenida en aquellas doce variables iniciales la pudiésemos sintetizar en dos componentes:

Z_1 = tiempo dedicado a la UOC,
Z_2 = capacidad del estudiante,

y que, además, éstas fuesen cuestiones independientes. Tendríamos resuelto el problema; los estudiantes se encontrarían localizados en un sistema bivariante de coordenadas y podrían ser clasificados con facilidad.



Sin entrar en la formalización del procedimiento, veamos cuáles serían los pasos que se deben seguir para determinar los componentes principales y para facilitar su lectura.

2.3. Procedimiento para la obtención de los componentes principales

Los pasos que se deben seguir son:

1) Obtención de la matriz de coeficientes de correlación entre todas las variables:

$$R = \begin{pmatrix} 1 & r_{12} & \dots & r_{1J} \\ r_{21} & 1 & \dots & r_{2J} \\ \dots & \dots & 1 & \dots \\ r_{J1} & r_{J2} & \dots & 1 \end{pmatrix} \qquad r_{jj'} = \frac{S_{jj'}}{S_j S_{j'}} \quad \text{y} \quad r_{jj'} = r_{j'j}$$

Nota

De hecho, podríamos haber trabajado con la matriz de varianzas y de covarianzas; no obstante, al fin y al cabo, una correlación no es más que una covarianza con variables estandarizadas.

2) Se calculan los valores propios a partir de los resultados de λ en la ecuación:

$$|R - \lambda I_J| = 0$$

En los valores propios

Notad que la suma de todas las soluciones coincide con el número de variables observadas:

$\lambda_1 + \lambda_2 + \ldots + \lambda_J = J$

Los valores propios están vinculados a los componentes principales que buscamos. El valor propio mayor λ_1 se asocia al primer componente Z_1, el siguiente λ_2 a Z_2, etc.

3) La dispersión total de las J variables observadas constituye la información de que disponemos inicialmente. Puesto que trabajamos con variables tipificadas, la suma de las varianzas será J, cifra que hemos repartido entre los nuevos factores principales.

Así pues, el primer componente absorbe una proporción de λ_1/J de la información inicial, los dos primeros componentes absorben una proporción de $(\lambda_1 + \lambda_2)/J$ del total y, si tomásemos I componentes, esta proporción retenida sería:

$$\frac{\lambda_1 + \lambda_2 + \ldots + \lambda_I}{J}$$

Esta expresión se tiene que entender como la capacidad explicativa de los componentes Z_1, Z_2, ..., Z_I, que permite determinar el número de componentes principales que necesitamos para conseguir una determinada bondad en el estudio.

Lógicamente, las J variables admiten hasta J componentes, y retienen el 100% de la información, pero no habríamos ganado nada si hubiésemos pasado a un nuevo sistema J-dimensional. Se trata de quedarnos con pocos componentes y, a la vez, retener la máxima información posible.

4) Para calcular las funciones que determinan cada uno de los componentes:

$$Z_I = u_{I1}X_1 + u_{I2}X_2 + \ldots + u_{IJ}X_J$$

hay que obtener los vectores característicos que contienen los coeficientes de las ecuaciones:

$$u_I = \begin{pmatrix} u_{I1} \\ u_{I2} \\ \ldots \\ u_{IJ} \end{pmatrix}$$

de manera que estén normalizados $\sum u_{IJ}^2 = 1$ y que sean independientes de los otros vectores $\sum u_{IJ} u_{I',J} = 0$:

$$1.^{er} \text{ componente } (R - \lambda_1 I)u_1 = 0$$
$$u_1' u_1 = 1$$

$$2.^{o} \text{ componente } (R - \lambda_2 I)u_2 = 0$$
$$u_2' u_2 = 1$$
$$u_1' u_2 = 0$$

y así sucesivamente hasta encontrar los vectores característicos de todos los componentes principales que hayamos fijado.

Ahora, los nuevos factores resultantes son independientes:

$$r_{Z_1 Z_2} = 0 \qquad r_{Z_1 Z_3} = 0 \qquad \ldots$$

5) Podemos proyectar las observaciones en un nuevo sistema de ejes sustituyendo simplemente los datos iniciales –estandarizados convenientemente– en las ecuaciones respectivas; se comprobará con facilidad que los nuevos datos presentan un valor medio igual a cero:

$$\bar{Z}_1 = \bar{Z}_2 = \ldots = \bar{Z}_J = 0$$

Actividad

2.2. Suponemos tres ratios financieras calculadas sobre cinco cajas de ahorros:

Caja	Ratio 1 (X_1)	Ratio 2 (X_2)	Ratio 3 (X_3)
A	23	22	45
B	45	38	74
C	34	24	47
D	19	7	15
E	52	44	83

Calculad las correlaciones que se dan entre las tres variables y obtened los valores propios λ_1, λ_2 y λ_3.

Comprobad cómo el primer factor es capaz de absorber prácticamente el 98% de la información total, lo cual justifica que calculemos sólo un componente principal Z_1.

Obtened el vector característico asociado a Z_1; veréis que resulta la ecuación:

$$Z_1 = 0{,}571X_1 - 0{,}581X_2 - 0{,}579X_3$$

Proyectad las cinco cajas de ahorros sobre el nuevo eje y veréis la ordenación conseguida (recordad que es necesario sustituir X_1, X_2 y X_3 por los valores estandarizados).

Si hubieseis calculado los tres componentes Z_2 y Z_3, veríais que salen las ecuaciones siguientes:

$$Z_2 = -0{,}817X_1 + 0{,}332X_2 + 0{,}472X_3$$
$$Z_1 = 0{,}082X_1 - 0{,}734X_2 - 0{,}664X_3$$

Podríais proyectar las cinco cajas en cada uno de los nuevos factores; comprobad que ahora los resultados tienen una media de cero y que no están en correlación.

$$\bar{Z}_1 = \bar{Z}_2 = \bar{Z}_3 = 0 \quad \text{y} \quad r_{Z_1Z_2} = r_{Z_1Z_3} = r_{Z_2Z_3} = 0$$

También podríamos demostrar el cumplimiento de las condiciones exigidas a los vectores:

$$u_1' u_1 = (-0{,}571)^2 + (-0{,}581)^2 + (-0{,}579)^2 = 1$$

$$u_2' u_2 = 1$$

$$u_3' u_3 = 1$$

$$u_1' u_2 = (-0{,}571)(-0{,}581) + (-0{,}581)(0{,}332) + (-0{,}579)(0{,}472) = 0$$

$$u_1' u_3 = 0$$

$$u_2' u_3 = 0$$

2.4. Interpretación de los resultados

Si trabajamos con menos ejes, será más fácil agrupar los resultados y clasificarlos en categorías. Sin embargo, esto será eficaz en la medida en que sepamos qué quieren decir estos nuevos componentes principales y, por tanto, las tipologías de los diferentes grupos que salen. Resulta poco útil formar categorías de elementos sin saber a qué criterio responden.

La interpretación de los componentes es fácil de conseguir en teoría, pero normalmente es bastante difícil en la práctica. Se puede hacer una primera aproximación a partir de las proyecciones conseguidas de los elementos; según si somos más o menos conocedores de la realidad que analizamos, puede ser bastante esclarecedora la posición que ocupan los elementos en cada nuevo eje.

Suponed que, estudiando las estadísticas de los municipios españoles y aplicando un análisis de componentes principales, Z_1 tiene valores altos para: Santander, Barcelona, Alicante, Cádiz... y muy bajos para León, Madrid, Jaén, Albacete... Empezaréis a pensar que seguramente Z_1 se identifica con algún concepto que mide la distancia del municipio al mar.

De forma análoga, las proyecciones sobre Z_2, Z_3, ... pueden ayudar a interpretar el concepto que traducen.

De todos modos, será más definitivo el estudio de las correlaciones entre las variables X_1, X_2, ..., X_J y los componentes encontrados Z_1, Z_2, ..., Z_J.

Como ya sabemos...

... cada componente es una combinación lineal de todas las variables, pero siempre hay algunas de mayor peso que pueden ser relevantes para etiquetar el componente.

Definimos la correlación entre X_j y Z_l a partir de la relación:

$$r_{X_j Z_l} = \sqrt{\lambda_l} u_{lj}$$

Así, para las J variables y para los componentes seleccionados, tendríamos:

	Z_1	Z_2	...
X_1	$\sqrt{\lambda_1} u_{11}$	$\sqrt{\lambda_2} u_{21}$	...
X_1	$\sqrt{\lambda_1} u_{12}$	$\sqrt{\lambda_2} u_{22}$	...
...	...	...	...
X_J	$\sqrt{\lambda_1} u_{1J}$	$\sqrt{\lambda_2} u_{2J}$	...

El signo y la magnitud de las correlaciones son fundamentales para dar significado a los componentes; las correlaciones extremas son aquellas que marcan la etiqueta de cada nuevo factor.

Z_1 debe tener un significado estrechamente vinculado a las variables con las que esté más relacionada: directamente cuando la correlación sea positiva e inversamente cuando sea negativa; lógicamente, Z_1 es un factor que no tiene nada que ver con las variables que presenten correlación muy baja.

Actividad

2.3. Suponemos que sobre veinticinco modelos de automóviles hemos observado quince características: velocidad máxima, capacidad del maletero, consumo de gasolina por ciudad, etc. Esto nos ha permitido hacer un ACP a partir del cual hemos seleccionado dos componentes Z_1 y Z_2 que retienen el 82,3% de la información inicial.

No disponemos de la proyección de los diferentes modelos en los nuevos ejes, pero sí de las correlaciones entre las quince características observadas y los componentes, de las cuales hemos seleccionado las más relevantes.

Z_1 está muy correlacionada con: la cilindrada, la aceleración, el consumo de carburante a 90 km/h y el consumo de carburante a 20 km/h.

Z_2 está muy correlacionada con: la longitud del coche, la distancia entre los ejes de las ruedas y la medida de las ruedas.

¿Qué interpretación tendríais que hacer de Z_1 y Z_2 que permitiese obtener una descripción fácil de los veinticinco coches observados?

Llegaréis con facilidad a la conclusión de que Z_1 es un identificador de la potencia y de las prestaciones mecánicas, y que Z_2 se asocia con las dimensiones del coche.

Ejemplo

Realizamos ahora un ejemplo simulado, paso a paso, de fácil solución sin tener que utilizar el soporte informático. Tenemos veintiséis municipios para los cuales hemos calculado la distribución porcentual del voto en las últimas elecciones al Parlamento de Cataluña:

	Partidos politicos					
Municipio	CiU	PSC	PP	ERC	ICV	Otros
1	32	37	11	6	9	5
2	42	20	8	13	12	5
3	27	41	12	3	7	10
4	48	32	6	8	6	1
5	33	25	20	4	12	6
...	...	...	...	...	...	...
26	53	21	4	12	9	1

Queremos hacer un ACP que ofrezca una lectura más cómoda de los resultados de la votación.

La matriz de coeficientes de correlación entre las variables ha dado los resultados siguientes:

	CiU	PSC	PP	ERC	ICV	Otros
CiU	1					
PSC	– 0,654	1				
PP	– 0,755	0,185	1			
ERC	0,808	– 0,760	– 0,731	1		
ICV	– 0,097	– 0,671	0,452	0,264	1	
Otros	– 0,918	0,521	0,628	– 0,653	0,127	1

Ahora ya podemos obtener los valores propios de cada componente:

$$0 = |R - \lambda I| = \begin{vmatrix} 1-\lambda & -0{,}654 & \dots & -0{,}918 \\ -0{,}654 & 1-\lambda & \dots & -0{,}097 \\ \dots & \dots & \dots & \dots \\ -0{,}918 & -0{,}097 & \dots & 1-\lambda \end{vmatrix}$$

La solución de este determinante nos lleva a una ecuación de sexto grado del tipo:

$$\lambda^6 + b\,\lambda^5 + c\,\lambda^4 + d\,\lambda^3 + e\,\lambda^2 + f\,\lambda + g = 0$$

que admite seis raíces:

$\lambda_1 = 3{,}69$ $\lambda_2 = 1{,}776$ $\lambda_3 = 0{,}438$

$\lambda_4 = 0{,}078$ $\lambda_5 = 0{,}012$ $\lambda_6 = 0{,}006$

$\sum \lambda_j = 6$

Si sólo retenemos un componente principal, podríamos absorber el 3,69/6 = 61,5% de toda la información; y si tomamos dos, el (3,69 + 1,776)/6 = 91,1%. Resulta lógico que cuantos más componentes haya, se tendrá más bondad en el análisis, pero, en cambio, la interpretación será más difícil.

Para calcular el primer componente:

$$Z_1 = u_{11}\text{CiU} + u_{12}\text{PSC} + u_{13}\text{PP} + u_{14}\text{ERC} + u_{15}\text{ICV} + u_{16}\text{Otros}$$

tenemos que obtener el vector característico:

$$(R - \lambda_1 I)\, u_1 = 0 \text{ , con } u_1' u_1 = 1$$

$$\begin{vmatrix} -2{,}69 & -0{,}654 & \ldots & -0{,}918 \\ -0{,}654 & -2{,}69 & \ldots & -0{,}097 \\ \ldots & \ldots & \ldots & \ldots \\ -0{,}918 & -0{,}097 & \ldots & -2{,}69 \end{vmatrix} \times \begin{vmatrix} u_{11} \\ u_{12} \\ \ldots \\ u_{16} \end{vmatrix} = 0$$

$$u_{11}^2 + u_{12}^2 + \ldots + u_{16}^2 = 1$$

Si resolvemos el sistema, obtenemos:

$$u_1 = \begin{pmatrix} 0{,}704 \\ -0{,}283 \\ \ldots \\ -0{,}155 \end{pmatrix}$$

Operaríamos de manera análoga para conseguir el segundo componente:

$$Z_2 = u_{12}\text{CiU} + u_{22}\text{PSC} + u_{13}\text{PP} + u_{14}\text{ERC} + u_{15}\text{ICV} + u_{16}\text{Otros}$$

$$(R - \lambda_2 I) = 0, \text{ con } u_2' u_2 = 1 \text{ y } u_1' u_2 = 0$$

y obtendríamos:

$$u_2 = \begin{pmatrix} 0{,}602 \\ -0{,}791 \\ \ldots \\ -0{,}136 \end{pmatrix}$$

Los dos componentes encontrados son:

Z_1 = 0,704 CiU – 0,283 PSC + ... – 0,155 Otros (61,5% de bondad)
Z_2 = 0,602 CiU – 0,791 PSC + ... – 0,136 Otros (29,6% de bondad)

sobre los cuales se proyectan las estandarizaciones de los resultados de los seis municipios.

Así, en los nuevos ejes, las proyecciones del primer municipio son:

$$Z_{11} = 0{,}704\left(\frac{32-\overline{CiU}}{S_{CiU}}\right) - 0{,}283\left(\frac{37-\overline{PSC}}{S_{PSC}}\right) + \ldots - 0{,}155\left(\frac{5-\overline{Otros}}{S_{Otros}}\right) = 0{,}11$$

$$Z_{21} = 0{,}602\left(\frac{32-\overline{CiU}}{S_{CiU}}\right) - 0{,}791\left(\frac{37-\overline{PSC}}{S_{PSC}}\right) + \ldots - 0{,}136\left(\frac{5-\overline{Otros}}{S_{Otros}}\right) =$$

$$= -0{,}23$$

y para todos los municipios observados tendríamos:

Municipio	Z_1	Z_2
1	0,11	– 0,23
2	1,25	0,17
3	– 1,31	– 1,35
4	1,46	0,95
5	– 0,85	1,21
...	...	...
26	2,17	– 0,46

Podríamos comprobar que:

$$\overline{Z}_1 = 0 \ , \ \overline{Z}_2 = 0 \text{ y } r_{Z_1 Z_2} = 0$$

Caben interpretaciones de los componentes según la correlación que presentan con las variables observadas; entre otros cálculos, obtendríamos:

$$r(CiU, Z_1) = \sqrt{\lambda_1}u_{11} = \sqrt{0{,}625} \times 0{,}704 = 0{,}557$$

$$r(Otros, Z_2) = \sqrt{\lambda_2}u_{26} = \sqrt{0{,}296}(-0{,}136) = -0{,}074$$

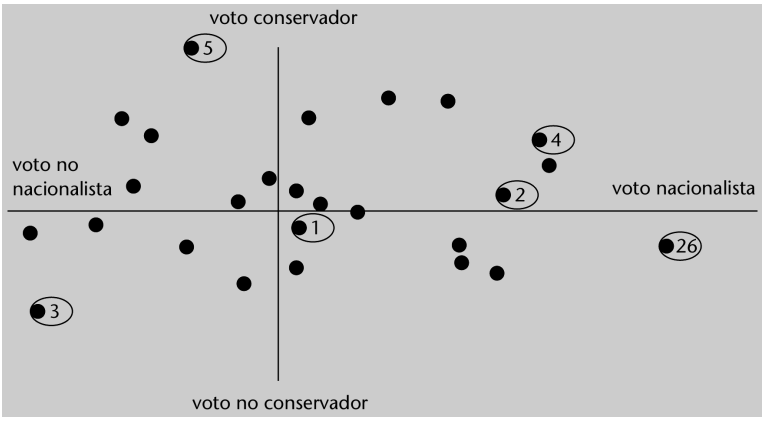
y, para todos los casos, lo que se muestra en el siguiente cuadro de correlaciones:

	Z_1	Z_2
CiU	0,557	0,383
PSC	– 0,223	– 0,430
PP	– 0,806	0,512
ERC	0,792	– 0,427
ICV	0,211	– 0,816
Otros	– 0,123	– 0,074

Se podría hacer la siguiente interpretación del mismo:

Z_1 = voto nacionalista,
Z_2 = voto conservador,

que vuelve a situar los veintiséis municipios de acuerdo con estos dos criterios y que permite un análisis más cómodo que el que se conseguiría con la información inicial:



Actividades

2.4. Hemos realizado un estudio acerca de diferentes aspectos relativos a diez grandes superficies de venta y hemos valorado de 0 a 10 las cuestiones siguientes:

– Facilidad de acceso al complejo comercial (A).
– Comodidad de aparcamiento (B).
– Bondad de los precios (C).
– Calidad de los productos (D).
– Servicio de guardería (E).
– Ofertas y promociones de artículos (F).
– Variedad de productos (G).
– Área de descanso y de ocio (H).

Hemos llegado a los siguientes resultados valorativos:

Cuadro de puntuaciones:

A	B	C	D	E	F	G	H	
1	6	9	8	7	6	7	9	6
2	7	9	5	3	7	5	4	7
3	6	8	5	3	7	4	4	6
4	5	5	7	2	4	4	3	4
5	6	7	9	9	7	8	9	5
6	8	9	9	7	7	7	8	8
7	2	1	5	6	2	4	5	3
8	7	8	3	2	6	2	3	6
9	4	3	8	6	2	9	7	4
10	5	6	7	7	4	8	8	6

Si calculáis la correlación entre todas estas características, comprobaréis que hay variables muy correlacionadas etre sí y que, por tanto, hay mucha información redundante.

Correlaciones entre variables:

	A	B	C	D	E	F	G	H
A	1							
B	0,941	1						
C	0,077	0,040	1					
D	– 0,136	– 0,097	0,751	1				
E	0,882	0,926	0,022	– 0,052	1			
F	– 0,051	– 0,053	0,846	0,805	– 0,132	1		
G	0,053	0,115	0,812	0,949	0,067	0,868	1	
H	0,903	0,894	0,073	0,000	0,792	0,064	18	1

Para simplificar los resultados de este estudio, tenéis que efectuar un análisis de los componentes principales. Veréis que podéis llegar a obtener hasta ocho valores propios (λ_j).

Valor propio	3,7031	3,5150	0,3285	0,2535	0,0958	0,0608	0,0422	0,0013
Proporción	0,463	0,439	0,041	0,032	0,012	0,008	0,005	0,000
Acumulativa	0,463	0,902	0,943	0,975	0,987	0,995	1,000	1,000

Ahora tendríais que justificar que os decidís sólo por dos factores principales (Z_1 y Z_2) y después tendríais que encontrar sus vectores característicos:

Variable	CP1	CP2
Acceso	– 0,489	– 0,132
Aparcamiento	– 0,496	– 0,124
Precios	– 0,136	0,463
Calidad	– 0,072	0,496
Guardería	– 0,471	– 0,134
Promoción	– 0,088	0,494
Variedad	– 0,169	0,490
Descanso	– 0,483	– 0,073

Esto os tiene que permitir proyectar las diez grandes superficies del estudio en un sistema de dos dimensiones y discutir las posiciones que ocupan según los nuevos indicadores Z_1 y Z_2. Os pueden ayudar las correlaciones, que calcularéis entre las ocho variables iniciales y los componentes nuevos.

	A	B	C	D	E	F	G	H
Z_1	– 0,942	– 0,955	– 0,262	– 0,138	– 0,906	– 0,169	– 0,325	– 0,929
Z_2	– 0,247	– 0,233	0,868	0,930	– 0,252	0,926	0,918	– 0,136

A continuación únicamente falta identificar los conceptos que engloban tanto Z_1 como Z_2.
Veréis que Z_1 se asocia a aspectos complementarios y de servicios, mientras que Z_2 es un indicador de mercado.

2.5. El cuadro que se muestra a continuación recopila los resultados conseguidos por un grupo de quince adolescentes en las pruebas atléticas siguientes:

P1: 100 metros lisos (en segundos).
P2: 200 metros lisos (en segundos).
P3: lanzamiento de peso (en metros).
P4: lanzamiento de disco (en metros).
P5: salto de longitud (en metros).
P6: 3.000 metros lisos (en minutos).
P7: salto de altura (en metros).
P8: 5.000 metros lisos (en minutos).
P9: triple salto (en metros).
P10: jabalina (en metros).
P11: 50 metros lisos (en segundos).

	P1	P2	P3	P4	P5	P6
1	13,4	28,8	7,32	37,40	3,93	14,518
2	13,9	29,7	7,09	34,15	4,11	14,658
3	14,8	31,4	8,71	43,55	4,45	16,870
4	12,9	27,9	5,78	30,05	3,70	19,502
5	13,3	28,5	5,99	29,95	3,87	16,770
6	15,1	32,5	6,14	32,10	5,02	15,778
7	13,8	29,4	6,67	33,35	4,05	15,302
8	12,9	27,7	8,73	44,10	3,91	18,347
9	15,0	31,9	6,45	32,25	4,87	19,418
10	14,4	30,3	9,11	45,60	4,23	14,826
11	14,1	30,1	6,34	31,70	4,71	18,606
12	12,8	28,0	6,80	34,25	3,77	15,812

	P1	P2	P3	P4	P5	P6
13	15,0	31,9	9,14	45,70	4,55	17,962
14	13,6	29,1	9,08	45,95	3,99	18,298
15	13,1	28,3	7,32	36,60	3,86	14,733

	P7	P8	P9	P10	P11
1	1,43	21,2484	9,04	67,32	7,80
2	1,61	21,5004	9,32	61,47	8,05
3	1,60	25,4820	10,00	78,39	8,35
4	1,35	30,2196	8,50	55,18	7,55
5	1,37	25,3020	9,07	56,12	7,75
6	1,70	23,5164	10,35	56,89	8,65
7	1,55	22,6596	9,20	60,03	8,00
8	1,41	28,1406	8,92	77,23	7,55
9	1,78	30,0684	10,38	58,05	8,60
10	1,73	21,8028	9,71	79,15	8,25
11	1,76	28,6068	10,52	57,06	8,15
12	1,27	23,5776	8,70	61,65	7,50
13	1,69	27,4476	10,20	80,69	8,55
14	1,49	28,0524	9,17	82,71	7,90
15	1,36	21,6354	8,76	65,88	7,65

Para elaborar una clasificación más cómoda de los participantes, hemos decidido efectuar un análisis de los componentes principales. Veréis que salen los valores propios siguientes:

Valor propio	Proporción	Acumulativa
0,7580	0,523	0,523
2,9796	0,271	0,794
1,9462	0,177	0,971
0,1850	0,017	0,988
0,0976	0,009	0,997
0,0242	0,002	0,999
0,0042	0,000	1,000
0,0024	0,000	1,000
0,0015	0,000	1,000
0,0012	0,000	1,000
0,0000	0,000	1,000

Teniendo en cuenta estos resultados, tenéis que justificar que os quedáis sólo con tres factores principales, a los cuales corresponderán los coeficientes que vemos aquí (vectores característicos):

Variable	CP1	CP2	CP3
100 m (P1)	– 0,407	0,000	0,094
200 m (P2)	– 0,403	0,034	0,098
Lanz. de peso (P3)	– 0,085	– 0,564	– 0,057
Lanz. de disco (P4)	– 0,083	– 0,563	– 0,076
Salto de longitud (P5)	– 0,395	0,135	0,014
3.000 m (P6)	– 0,087	0,098	– 0,690
Salto de altura (P7)	– 0,390	0,028	0,043
5.000 m (P8)	– 0,087	0,098	– 0,690
Triple salto (P9)	– 0,402	0,072	– 0,005
Jabalina (P10)	– 0,071	0,565	– 0,091
50 m (P11)	– 0,408	0,028	0,095

Si partimos de las ecuaciones que caracterizan a los componentes, encontraremos la proyección de cada atleta en los nuevos ejes (después de haber tipificado los resultados en cada prueba):

Atleta	Z1	Z2	Z3
1	1,65511	– 0,47131	1,51021
2	0,25803	0,42526	1,72854
3	– 2,19657	– 1,63623	– 0,02101
4	2,92188	1,93690	– 2,22897
5	2,06382	1,64760	– 0,01841
6	– 3,32226	1,93981	1,48237
7	0,68917	0,78782	1,21968
8	1,95017	– 1,92668	– 1,84459
9	– 3,51133	2,06703	– 1,37375
10	– 1,40265	– 2,38132	1,38170
11	– 1,93254	2,02394	– 1,06570
12	3,17963	0,35715	0,40038
13	– 3,22884	– 1,94573	– 0,79644
14	0,39772	– 2,44456	– 1,62322
15	2,47865	– 0,37968	1,24922

La interpretación de los resultados exige calcular previamente la correlación entre las variables originales y los nuevos componentes principales; se hará una mención especial de las correlaciones más altas.

	Z_1	Z_2	Z_3
P1	– 0,978	0,000	0,131
P2	– 0,968	0,059	0,137
P3	– 0,205	– 0,973	– 0,079
P4	– 0,200	– 0,973	– 0,106
P5	– 0,949	0,232	0,020
P6	– 0,208	0,169	– 0,963
P7	– 0,937	0,048	0,059
P8	– 0,208	0,169	– 0,963
P9	– 0,965	0,124	– 0,007
P10	– 0,170	– 0,976	– 0,127
P11	– 0,979	0,049	0,133

Ahora podemos acabar el problema fácilmente, otorgando significado a los tres componentes principales que vuelven a situar a los participantes de las pruebas atléticas.