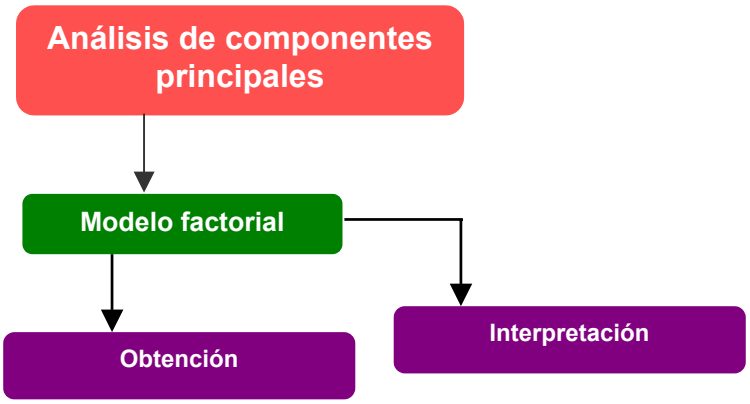
ANÁLISIS DE COMPONENTES PRINCIPALES

Autor: Manuel Terrádez Gurrea (mterradez@uoc.edu).

ESQUEMA DE CONTENIDOS



INTRODUCCIÓN

El Análisis de Componentes Principales (ACP) es una técnica estadística de síntesis de la información, o reducción de la dimensión (número de variables). Es decir, ante un banco de datos con muchas variables, el objetivo será reducirlas a un menor número perdiendo la menor cantidad de información posible.

Los nuevos componentes principales o factores serán una combinación lineal de las variables originales, y además serán independientes entre sí.

Un aspecto clave en ACP es la interpretación de los factores, ya que ésta no viene dada a priori, sino que será deducida tras observar la relación de los factores con las variables iniciales (habrá, pues, que estudiar tanto el signo como la magnitud de las correlaciones). Esto no siempre es fácil, y será de vital importancia el conocimiento que el experto tenga sobre la materia de investigación.

OBJETIVOS

- Entender por qué es importante reducir la dimensión en un problema estadístico.
- Saber aplicar el análisis de componentes principales, con ayuda de Minitab.
- Conocer pautas para elegir el modelo más adecuado para nuestro problema.
- Interpretar los factores del modelo obtenido.

CONOCIMIENTOS PREVIOS

Aparte de estar iniciado en el uso del paquete estadístico Minitab, resulta muy conveniente haber leído con profundidad los siguientes *math-blocks*:

- Estadística descriptiva.
- Correlación y regresión lineal múltiple.

CONCEPTOS FUNDAMENTALES

Fases de un análisis de componentes principales

Análisis de la matriz de correlaciones

Un análisis de componentes principales tiene sentido si existen altas correlaciones entre las variables, ya que esto es indicativo de que existe información redundante y, por tanto, pocos factores explicarán gran parte de la variabilidad total.

Selección de los factores

La elección de los factores se realiza de tal forma que el primero recoja la mayor proporción posible de la variabilidad original; el segundo factor debe recoger la máxima variabilidad posible no recogida por el primero, y así sucesivamente. Del total de factores se elegirán aquéllos que recojan el porcentaje de variabilidad que se considere suficiente. A éstos se les denominará componentes principales.

Análisis de la matriz factorial

Una vez seleccionados los componentes principales, se representan en forma de matriz. Cada elemento de ésta representa los coeficientes factoriales de las variables (las correlaciones entre las variables y los componentes principales). La matriz tendrá tantas columnas como componentes principales y tantas filas como variables.

Interpretación de los factores

Para que un factor sea fácilmente interpretable debe tener las siguientes características, que son difíciles de conseguir:

- Los coeficientes factoriales deben ser próximos a 1.
- Una variable debe tener coeficientes elevados sólo con un factor.
- No deben existir factores con coeficientes similares.

Cálculo de las puntuaciones factoriales

Son las puntuaciones que tienen los componentes principales para cada caso, que nos permitirán su representación gráfica.

Se calculan mediante la expresión: $X_{ij} = a_{i1} \cdot Z_{1j} + ... + a_{ik} \cdot Z_{kj} = \sum_{s=1}^{k} a_{is} \cdot Z_{sk}$

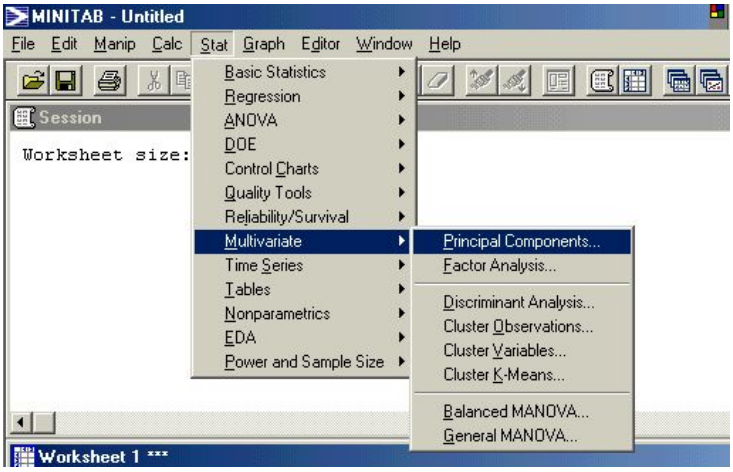
Los *a* son los coeficientes y los *Z* son los valores estandarizados que tienen las variables en cada uno de los sujetos de la muestra.

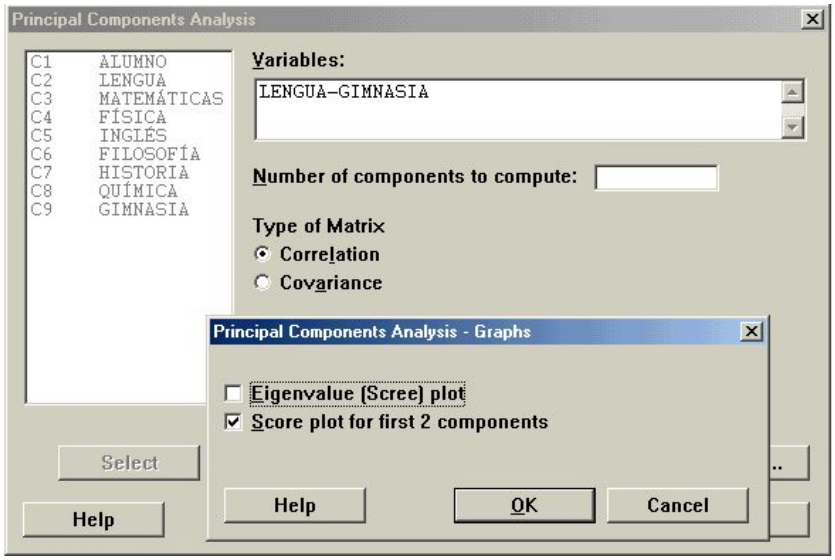
CASOS PRÁCTICOS CON SOFTWARE

Calificaciones escolares

Vamos a utilizar los datos del archivo **asignaturas.mtw**, que recogen las calificaciones de los 15 alumnos de una clase en diversas asignaturas.

Stat → Multivariate → Principal Components...





La salida que nos ofrece Minitab es la siguiente:

Principal Component Analysis

Eigenanalysis of the Correlation Matrix

Eigenvalue	3,7104	2,8608	0,9535	0,2156	0,1513	0,0628
Proportion	0,464	0,358	0,119	0,027	0,019	0,008
Cumulative	0,464	0,821	0,941	0,968	0,986	0,994

Eigenvalue	0,0317	0,0139
Proportion	0,004	0,002
Cumulative	0,998	1,000

Variable	PC1	PC2	PC3	PC4	PC5	PC6
LENGUA	0,500	0,085	-0,028	-0,235	0,434	0,112
MATEMÁTI	-0,113	0,555	0,133	-0,254	-0,245	-0,686
FÍSICA	-0,052	0,575	0,076	0,059	0,386	0,093
INGLÉS	0,499	0,037	-0,005	-0,550	0,102	0,001
FILOSOFÍ	0,450	0,122	-0,303	0,702	0,145	-0,340
HISTORIA	0,493	0,064	-0,011	0,027	-0,736	0,140
QUÍMICA	-0,073	0,574	-0,021	0,135	-0,163	0,611
GIMNASIA	0,187	-0,069	0,940	0,250	0,052	-0,002

Variable	PC7	PC8
LENGUA	-0,372	0,589
MATEMÁTI	-0,247	0,075
FÍSICA	0,696	0,126
INGLÉS	0,115	-0,651
FILOSOFÍ	-0,087	-0,232
HISTORIA	0,318	0,300
QUÍMICA	-0,436	-0,239
GIMNASIA	-0,066	-0,084

En primer lugar nos aparecen los valores propios (eigenvalue) de cada componente principal, y justo debajo la proporción de varianza explicada (proportion) por cada una de ellos y la varianza explicada acumulada (cumulative).

Los datos de varianza explicada son muy importantes para saber cuántos componentes principales vamos a utilizar en nuestro análisis. No hay una regla definida sobre el número que se debe utilizar, con lo cual deberemos decidir en función del número de variables iniciales (hay que recordar que se trata de reducirlas en la medida de lo posible) y de la proporción de varianza explicada acumulada.

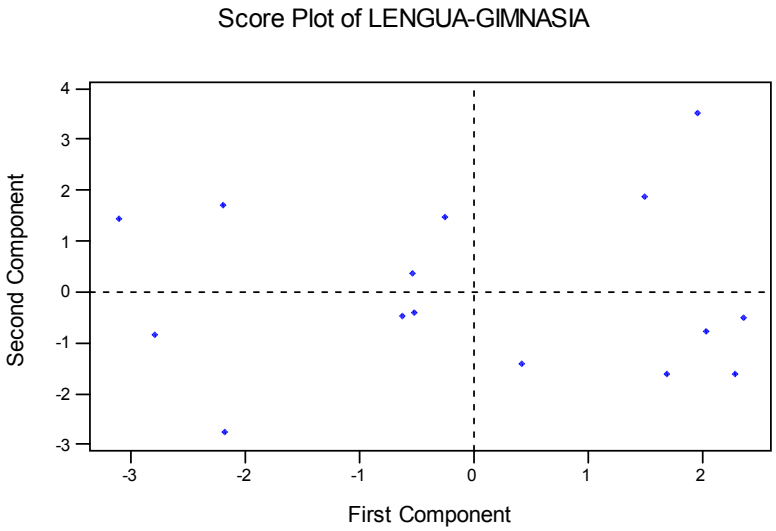
En este caso, parece razonable quedarse con los 3 primeros componentes principales, ya que con ellos se explica el 94,1% de la varianza, y teniendo en cuenta que añadiendo uno más sólo ganamos un 2,7%, y quitando uno perdemos un 12%.

Finalmente, nos aparecen las correlaciones de cada componente principal con cada variable: esto nos ayudará a interpretar las variables.

En este caso, vemos que PC1 tiene la mayor correlación positiva con las asignaturas LENGUA, INGLÉS, HISTORIA y FILOSOFÍA, mientras que tiene correlación negativa con MATEMÁTICAS y casi nula con el resto de asignaturas. Por tanto, es claro que estamos hablando de la facilidad para las asignaturas de Letras.
En cuanto a PC2, ocurre justo al contrario, ya que tiene correlación positiva con FÍSICA, QUÍMICA y MATEMÁTICAS, y cercana a 0 con el resto de asignaturas. Evidentemente, se está refiriendo a la facilidad en las asignaturas de Ciencias.

Por último, PC3 tiene una correlación positiva muy alta (casi 1) con GIMNASIA, con lo cual habría que interpretarla como la facilidad en dicha asignatura, bastante independiente del resto.

También obtenemos el gráfico en dos dimensiones de PC1 y PC2, donde podemos ver la variabilidad de las observaciones, y si existe alguna que ofrezca un valor extrañamente alto o bajo en cada eje.



Barómetro empresarial

Procedemos de forma análoga con el archivo **merco.mtw**, que contiene los datos del Barómetro Merco, publicados por CincoDías en marzo de 2001, y que consiste en una clasificación de las 50 empresas con más prestigio, en función de su puntuación en las siguientes variables:

- REF: Resultados económico-financieros.
- CPS: Calidad producto/servicio.
- CCCL: Cultura corporativa y calidad laboral.
- ERSC: Ética y responsabilidad social corporativa.
- DGPI: Dimensión global y presencia internacional.
- IDI: Investigación, desarrollo e innovación.

Stat → Multivariate → Principal components...

Principal Component Analysis

Eigenanalysis of the Correlation Matrix

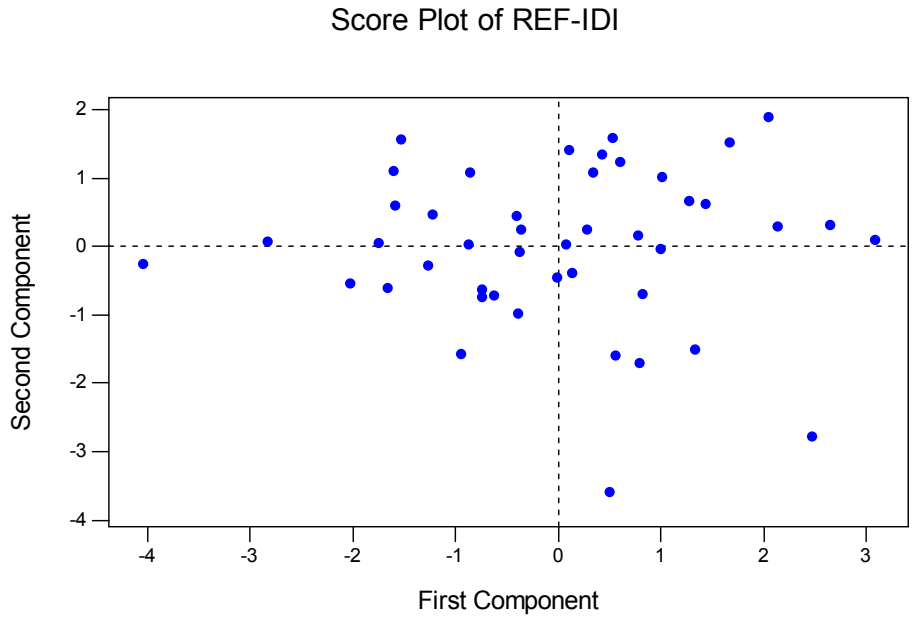
45 cases used 4 cases contain missing values

Eigenvalue	2,1520	1,3205	1,1834	0,7683	0,4062	0,1695
Proportion	0,359	0,220	0,197	0,128	0,068	0,028
Cumulative	0,359	0,579	0,776	0,904	0,972	1,000

Variable	PC1	PC2	PC3	PC4	PC5	PC6
REF	-0,015	0,743	0,296	0,250	0,538	0,093
CPS	0,053	0,066	0,855	-0,215	-0,464	0,021
CCCL	0,582	-0,187	0,038	0,439	-0,063	0,654
ERSC	0,626	-0,101	0,107	0,216	0,123	-0,725
DGPI	0,245	0,629	-0,405	-0,036	-0,614	-0,050
IDI	0,454	0,056	-0,068	-0,807	0,316	0,187

A la vista de la salida del Minitab, resultaría difícil decantarse por reducir la dimensión a la mitad, alcanzando un 77,6% de varianza explicada (con los 3 primeros Componentes Principales) o llegar al 90,4% de varianza con un componente principal más.

Veamos el gráfico en dos dimensiones de PC1 y PC2:



En cuanto a la interpretación de los componentes, observamos que PC1 tiene la mayor correlación positiva con las variables ERSC y CCCL, con lo cual sus valores positivos podrían asimilarse con aquellas compañías que destacan por sus valores intangibles (cultura de empresa, ética profesional, etc.)

PC2 tiene una correlación positiva muy alta con la variable REF y también destacable con DGPI, con lo cual, a diferencia del caso anterior, parece indicar que valores positivos de este componente los obtendrían las empresas con mejores datos objetivos (buena salud financiera, gran dimensión).

Por su parte, PC3 tiene una correlación positiva muy alta con CPS, y correlación negativa con DGPI, y por tanto lo podríamos asociar con aquellas compañías más "cercanas" al usuario, que destacan más por la calidad del servicio que por su presencia internacional.

Por último, PC4 tiene una correlación negativa muy alta con IDI, y de ahí que pudiéramos asociar sus valores positivos con las empresas de corte más tradicional, que no destacan principalmente por la investigación y la innovación.

En cuanto al gráfico en dos dimensiones de PC1 y PC2, observamos que hay gran dispersión en el primer componente, mientras que en el segundo la mayoría de observaciones se sitúan en los valores centrales, aunque hay algunos datos que destacan por sus valores distintos (especialmente los negativos), y que son los que cabría estudiar más a fondo.

Clientes bancarios

En el archivo **clientes bancarios.mtw** aparecen los datos de las oficinas de una entidad bancaria.

Las variables son las siguientes: TIPO (tipo de oficina: **S**ucursal o **D**elegación), PROMEDIO (promedio de transacciones por cliente), CLIENTES (incremento/decremento de clientes respecto al ejercicio anterior), TRANSACCIONES (incremento/decremento de transacciones respecto al ejercicio anterior) y VOLUMEN (volumen de clientes), y se desea estudiar con detalle la variable CLIENTES, para detectar su relación con el resto de variables.

Stat → Multivariate → Principal components...

Principal Component Analysis

Eigenanalysis of the Correlation Matrix

Eigenvalue	1,9481	1,0577	0,9826	0,0115
Proportion	0,487	0,264	0,246	0,003
Cumulative	0,487	0,751	0,997	1,000

Variable	PC1	PC2	PC3	PC4
PROMEDIO	0,021	-0,477	0,878	-0,013
CLIENTES	-0,714	-0,043	0,004	0,699
TRANSACC	-0,647	0,363	0,204	-0,639
VOLUMEN	0,268	0,799	0,432	0,321

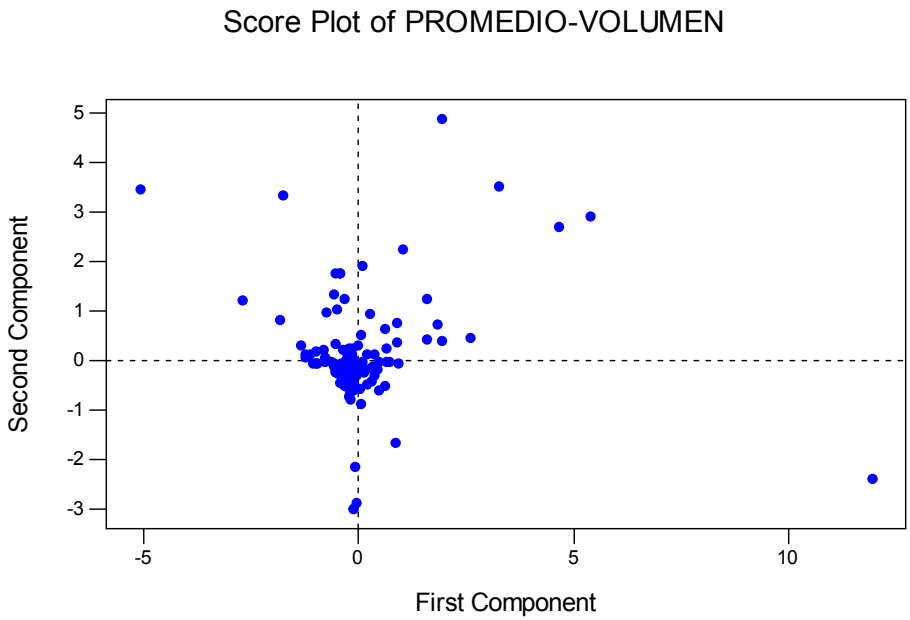
A la vista de la salida que ofrece Minitab, podemos afirmar que los resultados que ofrece el ACP en este caso no son demasiado buenos, ya que tan sólo logramos reducir la dimensión en una variable si optamos por el modelo con 3 componentes (99,7% de varianza explicada), que parece el más lógico ya que con 2 sólo explicaríamos el 75% de la varianza.

En cuanto a la interpretación de los componentes, observamos que PC1 tiene alta correlación negativa con las variables CLIENTES y TRANSACCIONES, con lo cual sus valores positivos podrían asimilarse a aquellas oficinas que durante el ejercicio han disminuido su número de clientes y transacciones respecto al ejercicio anterior.

PC2 tiene la mayor correlación (positiva) con la variable VOLUMEN, mientras que es también reseñable su correlación negativa con PROMEDIO. Esto parece indicar que valores positivos de este componente los obtendrían las oficinas con una gran cartera de clientes, no necesariamente muy activos.

Por último, PC3 tiene una correlación positiva y muy alta con la variable PROMEDIO, y por tanto lo podríamos asociar con oficinas cuyos cientes son muy activos, ya que realizan un gran número de transacciones.

En cuanto al gráfico en dos dimensiones de PC1 y PC2, observamos que la gran mayoría de las observaciones se acumulan en los valores centrales de ambos componentes, aunque hay algunos datos que destacan por sus valores distintos, y que son los que cabría estudiar más a fondo.


Score Plot of PROMEDIO-VOLUMEN
Second Component
First Component
5
4
3
2
1
0
-1
-2
-3
-5
0
5
10

BIBLIOGRAFÍA

[1] Baró, J. y Alemany, R. (2000): "Estadística II". Ed. Fundació per a la Universitat Oberta de Catalunya. Barcelona.

[2] Peña Sánchez de Rivera, D. (1987): "Estadística. Modelos y Métodos. Volumen 2". Alianza Editorial. Madrid. ISBN: 84-206-8110-5

[3] Johnson, R. R. (1996): "Elementary statistics". Belmont, etc. : Duxbury, cop

[4] Martín-Guzmán, P. (1991): "Curso básico de estadística económica". AC, DL. Madrid. ISBN: 84-7288-142-3

ENLACES

http://www.5campus.org/leccion/anamul

Lección sobre Análisis Multivariante (Universidad de Zaragoza)

http://www.uniovi.es/UniOvi/Apartados/Departamento/Psicologia/metodos/tutor.1/fac3.html

Artículo "ANÁLISIS FACTORIAL vs COMPONENTES PRINCIPALES" de la Universidad de Oviedo

http://www.inegi.gob.mx/difusion/espanol/niveles/jly/nivbien/componentes.html

Selección de variables a través de la técnica de Componentes Principales