COMPONENTES PRINCIPALES

ACP

Santiago de la Fuente Fernández

Fac. Ciencias Económicas y Empresariales

UAM – 2011

ANÁLISIS DE COMPONENTES PRINCIPALES

Cuando se recoge la información de una muestra de datos, lo más frecuente es tomar el mayor número posible de variables. Sin embargo, si tomamos demasiadas variables sobre un conjunto de objetos, por ejemplo 20 variables, tendremos que considerar $\binom{20}{2} = 180$ posibles coeficientes de correlación; si son 40 variables dicho número aumenta hasta 780.

Evidentemente, en este caso es difícil visualizar relaciones entre las variables. Otro problema que se presenta es la fuerte correlación que muchas veces se presenta entre las variables: si tomamos demasiadas variables (cosa que en general sucede cuando no se sabe demasiado sobre los datos o sólo se tiene ánimo exploratorio), lo normal es que estén relacionadas o que midan lo mismo bajo distintos puntos de vista. Por ejemplo, en estudios médicos, la presión sanguínea a la salida del corazón y a la salida de los pulmones están fuertemente relacionadas.

Se hace necesario, pues, reducir el número de variables. Es importante resaltar el hecho de que el concepto de mayor información se relaciona con el de mayor variabilidad o varianza. Cuanto mayor sea la variabilidad de los datos (varianza) se considera que existe mayor información, lo cual está relacionado con el concepto de entropía.

COMPONENTES PRINCIPALES

Estas técnicas fueron inicialmente desarrolladas por Pearson a finales del siglo XIX y posteriormente fueron estudiadas por Hotelling en los años 30 del siglo XX. Sin embargo, hasta la aparición de los ordenadores no se empezaron a popularizar.
Para estudiar las relaciones que se presentan entre p variables correlacionadas (que miden información común) se puede transformar el conjunto original de variables en otro conjunto de nuevas variables incorreladas entre sí (que no tenga repetición o redundancia en la información) llamado conjunto de componentes principales.

Las nuevas variables son combinaciones lineales de las anteriores y se van construyendo según el orden de importancia en cuanto a la variabilidad total que recogen de la muestra.

De modo ideal, se buscan m < p variables que sean combinaciones lineales de las p originales y que estén incorreladas, recogiendo la mayor parte de la información o variabilidad de los datos.

Si las variables originales están incorreladas de partida, entonces no tiene sentido realizar un análisis de componentes principales.

El análisis de componentes principales es una técnica matemática que no requiere la suposición de normalidad multivariante de los datos, aunque si esto último se cumple se puede dar una interpretación más profunda de dichos componentes.

CÁLCULO DE LOS COMPONENTES PRINCIPALES

Se considera una serie de variables $(x_1, x_2, \dots, x_p)$ sobre un grupo de objetos o individuos y se trata de calcular, a partir de ellas, un nuevo conjunto de variables $(y_1, y_2, \dots, y_p)$, incorreladas entre sí, cuyas varianzas vayan decreciendo progresivamente.

Cada y_j (j = 1, . . . , p) es una combinación lineal de las (x_1, x_2, ..., x_p) originales, es decir:

$$y_j = a_{j1}\,x_1 + a_{j2}\,x_2 + \ldots + a_{jp}\,x_p = a_j^{\circ}\,x$$

siendo $a_j^{\circ} = (a_{1j}, a_{2j}, \ldots, a_{pj})$ un vector de constantes, y $x = \begin{pmatrix} x_1 \\ \ldots \\ x_p \end{pmatrix}$

Obviamente, si lo que queremos es maximizar la varianza, como veremos luego, una forma simple podría ser aumentar los coeficientes a_{ij}. Por ello, para mantener la ortogonalidad de la transformación se impone que el módulo del vector $a_j^{\circ} = (a_{1j}, a_{2j}, \ldots, a_{pj})$ sea 1.

Es decir, $a_j^{\circ}.a_j = \sum_{k=1}^{p} a_{kj}^2 = 1$

El primer componente se calcula eligiendo a_1 de modo que y_1 tenga la mayor varianza posible, sujeta a la restricción $a_j^{\circ}.a_j = 1$. El segundo componente principal se calcula obteniendo a_2 de modo que la variable obtenida, y_2 esté incorrelada con y_1.

Del mismo modo se eligen (y_1, y_2, ... , y_p), incorrelados entre sí, de manera que las variables aleatorias obtenidas vayan teniendo cada vez menor varianza.

PROCESO DE EXTRACCIÓN DE FACTORES

Se elige a_1 de modo que se maximice la varianza de y_1 sujeta a la restricción de que $a_j^{\circ}.a_j = 1$

$$\mathrm{Var}(y_1) = \mathrm{Var}(a_1^{\circ}.x) = a_1^{\circ}.\Sigma a_1$$

El método habitual para maximizar una función de varias variables sujeta a restricciones el método de los *multiplicadores de Lagrange*.

El problema consiste en maximizar la función $a_1^{\circ}.\Sigma a_1$ sujeta a la restricción $a_j^{\circ}.a_j = 1$.

Se puede observar que la incógnita es precisamente a_1 (el vector desconocido que da la combinación lineal óptima).

Así, se construye la función L: $L(a_1) = a_1^{\circ}.\Sigma a_1 - \lambda(a_1^{\circ}.a_1 - 1)$

Para maximizar la función: $\frac{\vartheta L}{\vartheta a_1} = 2\Sigma a_1 - 2\lambda I a_1 = 0 \;\Rightarrow\; (\Sigma - \lambda I)a_1 = 0$

Esto es, en realidad, un sistema lineal de ecuaciones. Por el teorema de Roché-Frobenius, para que el sistema tenga una solución distinta de 0 la matriz (Σ − λI) tiene que ser singular. Esto implica que el determinante debe ser igual a cero:

$|\Sigma - \lambda I| = 0$ de este modo, λ es un autovalor de Σ.

La matriz de covarianzas Σ es de orden p y si además es definida positiva, tendrá p autovalores distintos, ($\lambda_1, \lambda_2, \ldots, \lambda_p$) tales que, por ejemplo, $\lambda_1 > \lambda_2 > \cdots > \lambda_p$.

Se tiene que, desarrollando la expresión anterior: $\begin{cases} (\sum - \lambda I)a_1 = 0 \\ \sum a_1 - \lambda I a_1 = 0 \\ \sum a_1 = \lambda I a_1 \end{cases}$

entonces, $\text{Var}(y_1) = \text{Var}(a_1^{\circ} x) = a_1^{\circ} \sum a_1 = a_1^{\circ} \lambda a_1 = \lambda a_1^{\circ} a_1 = \lambda 1 = \lambda$

Luego, para maximizar la varianza de y_1 se tiene que tomar el mayor autovalor, sea λ_1, y el correspondiente autovector a_1.

En realidad, a_1 es un vector que da la combinación de las variables originales que tiene mayor varianza, esto es, sí $a_1^{\circ} = (a_{11}, a_{12}, \ldots, a_{1p})$, entonces $y_1 = a_1^{\circ} x = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p$

El segundo componente principal, sea $y_2 = a_2^{\circ} x$, se obtiene mediante un argumento parecido. Además, se quiere que y_2 esté incorrelado con el anterior componente y_1, es decir, Cov(y_2, y_1) = 0.

Por lo tanto: $\text{Cov}(y_2, y_1) = \text{Cov}(a_2^{\circ} x, a_1^{\circ} x) == a_2^{\circ} \cdot E[(x - \mu)(x - \mu)^{\circ}] \cdot a_1 = a_2^{\circ} \sum a_1$

es decir, se requiere que $a_2^{\circ} \sum a_1 = 0$

Como se tenía que $\Sigma a_1 = \lambda a_1$, lo anterior es equivalente a $a_2^{\circ} \sum a_1 = a_2^{\circ} \lambda a_1 = \lambda a_2^{\circ} a_1 = 0$

Por tanto, $a_2^{\circ} a_1 = 0 \Rightarrow$ los vectores sean ortogonales.

De este modo, tendremos que maximizar la varianza de y_2, es decir, ($a_2 \Sigma a_2$), donde $\begin{cases} a_2^{\circ} a_2 = 1 \\ a_2^{\circ} a_1 = 0 \end{cases}$

Se toma la función: $L(a_2) = a_2^{\circ} \sum a_2 - \lambda(a_2^{\circ} a_2 - 1) - \delta a_2^{\circ} a_1$

con lo cual, $\frac{\vartheta L(a_2)}{\vartheta a_2} = 2\sum a_2 - 2\lambda a_2 - \delta a_1 = 0$

multiplicando por (a_1°) la expresión anterior, queda: $2a_1^{\circ} \sum a_2 - \delta = 0$, adviértase que $\begin{cases} a_1^{\circ} a_2 = a_2^{\circ} a_1 = 0 \\ a_1^{\circ} a_1 = 1 \end{cases}$

Luego, $\delta = 2a_1^{\circ} \sum a_2 = 2a_2^{\circ} \sum a_1 = 0$ ya que Cov(y_2, y_1) = 0.

De este modo,

$$\frac{\vartheta L(a_2)}{\vartheta a_2} = 2\sum a_2 - 2\lambda a_2 - \delta a_1 = 2\sum a_2 - 2\lambda a_2 = (\sum - \lambda I)a_2 = 0$$

Análogamente al caso anterior, elegimos λ como el segundo mayor autovalor de la matriz Σ con su autovector asociado a_2.

Los razonamientos anteriores se pueden extender, de modo que al j-ésimo componente le correspondería el j-ésimo autovalor.

Entonces todos los componentes y (en total p) se pueden expresar como el producto de una matriz formada por los autovectores, multiplicada por el vector x que contiene las variables originales ($x_1, \ldots, x_p$):

$$y = A\,x$$

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix} \qquad A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{pmatrix} \qquad x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} \qquad \text{siendo} \begin{cases} Var(y_1) = \lambda_1 \\ Var(y_2) = \lambda_2 \\ \vdots \\ Var(y_p) = \lambda_p \end{cases}$$

La matriz de covarianzas de y será: $\Delta = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix}$

porque $y_1, \ldots, y_p$ se han construido como variables incorreladas.

Se tiene que,

$$\Delta = Var(Y) = A' Var(X) A = A' \sum A \qquad \text{o bien,} \qquad \sum = A \Delta A'$$

ya que A es una matriz ortogonal ($a_i \overset{\circ}{} a_i = 1$ para todas sus columnas) por lo que $A A' = I$

PORCENTAJES DE VARIABLILIDAD

Sabemos que cada autovalor correspondía a la varianza del componente y_i, que se definía por medio del autovector a_i, es decir, $Var(y_i) = \lambda_i$.

Si sumamos todos los autovalores, tendremos la varianza total de los componentes, es decir:

$$\sum_{i=1}^{p} Var(y_i) = \sum_{i=1}^{p} \lambda_i = traza(\Delta) \text{ puesto que } \Delta \equiv \text{ matriz diagonal}$$

Por las propiedades del operador traza,

$$traza(\Delta) = traza(A' \sum A) = traza(\sum A' A) = traza(\sum) \qquad \text{pues, A ortogonal} \Rightarrow A' A = I$$

con lo cual, $traza(\Delta) = traza(\sum) = \sum_{i=1}^{p} Var(x_i)$

Es decir, la suma de las varianzas de las variables originales y la suma de las varianzas de las componentes son iguales.

Esto permite hablar del porcentaje de varianza total que recoge un componente principal:

$$\frac{\lambda_i}{\sum_{i=1}^{p}\lambda_i} = \frac{\lambda_i}{\sum_{i=1}^{p}Var(x_i)}$$

También se podrá expresar el porcentaje de variabilidad recogido por los primeros m componentes (m < p) $\dfrac{\sum_{i=1}^{m}\lambda_i}{\sum_{i=1}^{p}Var(x_i)}$

En la práctica, al tener en principio p variables, nos quedaremos con un número mucho menor de componentes que recoja un porcentaje amplio de la variabilidad total $\sum_{i=1}^{p}Var(x_i)$

En general, no se suele coger más de tres componentes principales, a ser posible, para poder representarlos posteriormente en las gráficas.

CÁLCULO DE LOS COMPONENTES PRINCIPALES A PARTIR DE LA MATRIZ DE CORRELACIONES

Habitualmente, se calculan los componentes sobre variables originales estandarizadas, es decir, variables con media 0 y varianza 1. Esto equivale a *tomar los componentes principales*, no de la matriz de covarianzas sino *de la matriz de correlaciones* (en las variables estandarizadas coinciden las covarianzas y las correlaciones).

Así, *los componentes son autovectores de la matriz de correlaciones* y son distintos de los de la matriz de covarianzas. Si se actúa así, se da igual importancia a todas las variables originales.

En la matriz de correlaciones todos los elementos de la diagonal son iguales a 1. Si las variables originales están tipificadas, esto implica que su matriz de covarianzas es igual a la de correlaciones, con lo que la *variabilidad total (la traza) es igual al número total de variables que hay en la muestra*.

Suma total de todos los autovalores $\equiv$ p
Proporción de varianza recogida por el autovector j-ésimo (componente) $\equiv \dfrac{\lambda_j}{p}$

MATRIZ FACTORIAL

Cuando se presentan los autovectores en la salida de SPSS, se les suele multiplicar previamente por $\sqrt{\lambda_j}$ (del autovalor correspondiente), para reescalar todos los componentes del mismo modo.
Así, se calcula: $a^* = \sqrt{\lambda_j}\, a_j$ para j = 1, . . . , p.

De este modo, se suele presentar una tabla de autovectores a_j^* que forman la matriz factorial $F = (a_1^*, a_2^*, \ldots, a_p^*)$

Si se eleva al cuadrado cada una de las columnas y se suman los términos se obtienen los autovalores: $a_j^{*\circ}\, a_j^* = \sqrt{\lambda_j}\,\sqrt{\lambda_j}\, a_j^{\circ}\, a_j = \lambda_j$ puesto que $a_j^{\circ} a_j = 1$

Por otra parte, como $\sum = A \Delta A'$ y SSPS presenta como matriz factorial a $F = A \Delta^{1/2}$

Se tiene que $\boxed{\sum = F F'}$

Los elementos de F son tales que los mayores valores indican una mayor importancia en el momento de definir un componente.

Otra forma de enfocarlo consiste en considerar que como $y = Ax \Rightarrow x = A^{-1} y$

De modo que, $Cov(x) = (A^{-1})' Cov(y)\ A^{-1} = A \Delta A' = A\ \Delta^{1/2} \Delta^{1/2} A' = F F'$

Como A es ortogonal $\Rightarrow A^{-1} = A'$

Así, da la matriz factorial F, se pueden calcular las covarianzas de las variables originales, esto es, se puede recuperar la matriz de covarianzas original a partir de la matriz factorial.

Si se toma un número menor de factores (m < p), se podrá reproducir aproximadamente $\sum$

CÁLCULO DE COVARIANZAS Y CORRELACIONES ENTRE VARIABLES ORIGINALES Y LOS FACTORES

Como se tiene que $y = Ax \Rightarrow x = A^{-1} y = A'Y$ (por ser la matriz A ortogonal $\Rightarrow A^{-1} = A'$)

Entonces, $Cov(y_j, x_i) = Cov(y_j, \sum_{k=1}^{p} a_{ik} y_k) = a_{ij} Var(y_j) = \lambda_j a_{ij}$

donde, y_j es el factor j-ésimo y x_i es la variable original i-ésima.

Suponiendo que las variables originales están estandarizadas [$Var(x_i)=1$ para (i = 1, ..., p)]:

entonces, $Corre(y_j, x_i) = \frac{\lambda_j a_{ij}}{1 \cdot \sqrt{\lambda_j}} = \sqrt{\lambda_j}\, a_{ij}$

De este modo, la matriz de correlaciones entre y e x es: $\boxed{Corre(y,x) = \Delta^{1/2} A' = F'}$

con lo que la matriz factorial también mide las correlaciones entre las variables originales estandarizadas y los nuevos factores.

CAMBIOS DE ESCALAS E IDENTIFICACIÓN DE COMPONENTES

Si las variables originales ($x_1, \ldots, x_p$) están incorreladas, entonces carece de sentido calcular unos componentes principales. Si se hiciera, se obtendrían las mismas variables pero reordenadas de mayor a menor varianza.

Para saber si $(x_1, \ldots, x_p)$ están correlacionadas, se puede calcular la matriz de correlaciones aplicándose posteriormente el test de esfericidad de Barlett.

El cálculo de los componentes principales de una serie de variables $(x_1, \ldots, x_p)$ depende normalmente de las unidades de medida empleadas. Si transformamos las unidades de medida, lo más probable es que cambien a su vez los componentes obtenidos.

Una solución frecuente es usar variables $(x_1, \ldots, x_p)$ tipificadas. Con ello, se eliminan las diferentes unidades de medida y se consideran todas las variables implícitamente equivalentes en cuanto a la información recogida.

IDENTIFICACIÓN DE LOS COMPONENTES PRINCIPALES

Uno de los objetivos del cálculo de componentes principales es la identificación de los mismos, es decir, averiguar qué información de la muestra resumen. Sin embargo este es un problema difícil que a menudo resulta subjetivo.
Habitualmente, se conservan sólo aquellos componentes que recogen la mayor parte de la variabilidad, hecho que permite representar los datos según dos o tres dimensiones si se conservan dos o tres ejes factoriales, pudiéndose identificar entonces grupos naturales entre las observaciones.

Ejemplo.- Muestra de 41 ciudades de USA donde se midieron diferentes variables relacionadas con la contaminación atmosférica.

	SO2	Neg.Temp	Empresas	Poblacion	Viento	Precip	Días
Phoenix	10	70,3	213	582	6	7,05	36
Little Rock	13	61	91	132	8,2	48,52	100
San Francisco	12	56,7	453	716	8,7	20,66	67
Denver	17	51,9	454	515	9	12,95	86
Hartford	56	49,1	412	158	9	43,37	127
Wilmington	36	54	80	80	9	40,25	114
Washington	29	57,3	434	757	9,3	38,89	111
Jacksonville	14	68,4	136	529	8,8	54,47	116
Miami	10	75,5	207	335	9	59,80	128
Atlanta	24	61,5	368	497	9,1	48,34	115
Chicago	110	50,6	3344	3369	10,4	34,44	122
Indianapolis	28	52,3	361	746	9,7	38,74	121
Des Moines	17	49	104	201	11,2	30,85	103
Wichita	8	56,6	125	277	12,7	30,58	82
Louisville	30	55,6	291	593	8,3	43,11	123
New Orleans	9	68,3	204	361	8,4	56,77	113
Baltimore	47	55	625	905	9,6	41,31	111
Detroit	35	49,9	1064	1513	10,1	30,96	129
Minneapolis-St. Paul	29	43,5	699	744	10,6	25,94	137
Kansas City	14	54,5	381	507	10	37	99
St. Louis	56	55,9	775	622	9,5	35,89	105
Omaha	14	51,5	181	347	10,9	30,18	98
Alburquerque	11	56,8	46	244	8,9	7,77	58
Albany	46	47,6	44	116	8,8	33,36	135
Buffalo	11	47,1	391	463	12,4	36,11	166
Cincinnati	23	54	462	453	7,1	39,04	132
Cleveland	65	49,7	1007	751	10,9	34,99	155
Columbus	26	51,5	266	540	8,6	37,01	134
Philadelphia	69	54,6	1692	1950	9,6	39,93	115
Pittsburgh	61	50,4	347	520	9,4	36,22	147
Providence	94	50	343	179	10,6	42,75	125
Memphis	10	61,6	337	624	9,2	49,10	105
Nashville	18	59,4	275	448	7,9	46	119
Dallas	9	66,2	641	844	10,9	35,94	78
Houston	10	68,9	721	1233	10,8	48,19	103
Salt Lake City	28	51	137	176	8,7	15,17	89
Norfolk	31	59,3	96	308	10,6	44,68	116
Richmond	26	57,8	197	299	7,6	42,59	115
Seattle	29	51,1	379	531	9,4	38,79	164
Charleston	31	55,2	35	71	6,5	40,75	148
Milwaukee	16	45,7	569	717	11,8	29,07	123

Las variables son:

— Contenido en SO2
(Temp): Temperatura anual en grados F
(Emp): Número de empresas mayores de 20 trabajadores
(Pob): Población (en miles de habitantes)
(Viento): Velocidad media del viento
(Precipt): Precipitación anual media
(Días): Días lluviosos al año

En principio interesa investigar la relación entre la concentración en SO_2 y el resto de variables, utilizamos un análisis de componentes principales para eliminar relaciones entre las variables.

Se realiza un análisis de componente principales sobre todas las variables salvo SO_2.

En la salida de resultados de R se observan varias gráficas descriptivas exploratorias donde se presentan varios datos anómalos (outliers), por ejemplo Chicago.

Se obtienen los componentes principales a partir de la matriz de correlaciones para emplear las mismas escalas en todas las variables.

Los primeros tres componentes tienen todos varianzas (autovalores) mayores que 1 y entre los tres recogen el 85% de la varianza de las variables originales.

- El primer componente se le podría etiquetar como calidad de vida con valores negativos altos en empresas y población indicando un entorno relativamente pobre.

- El segundo componente se puede etiquetar como tiempo húmedo, y tiene pesos altos en las variables precipitaciones y días.

- El tercer componente se podría etiquetar como tipo de clima y está relacionado con la temperatura y la cantidad de lluvia.

Aunque no se encontrasen etiquetas claras para los componentes, siempre es interesante calcular componentes principales para descubrir si los datos se encuentran en una dimensión menor. De hecho, los tres primeros componentes producen un mapa de los datos donde las distancias entre los puntos es bastante semejante a la observada en los mismos respecto a las variables originales.

Se realiza un análisis de regresión de la variable SO_2 sobre los tres factores: claramente la cantidad de SO_2 se explica mediante el primer componente de calidad de vida (relacionado con el entorno humano y el clima) que cuando empeora aumenta, a su vez, la contaminación.

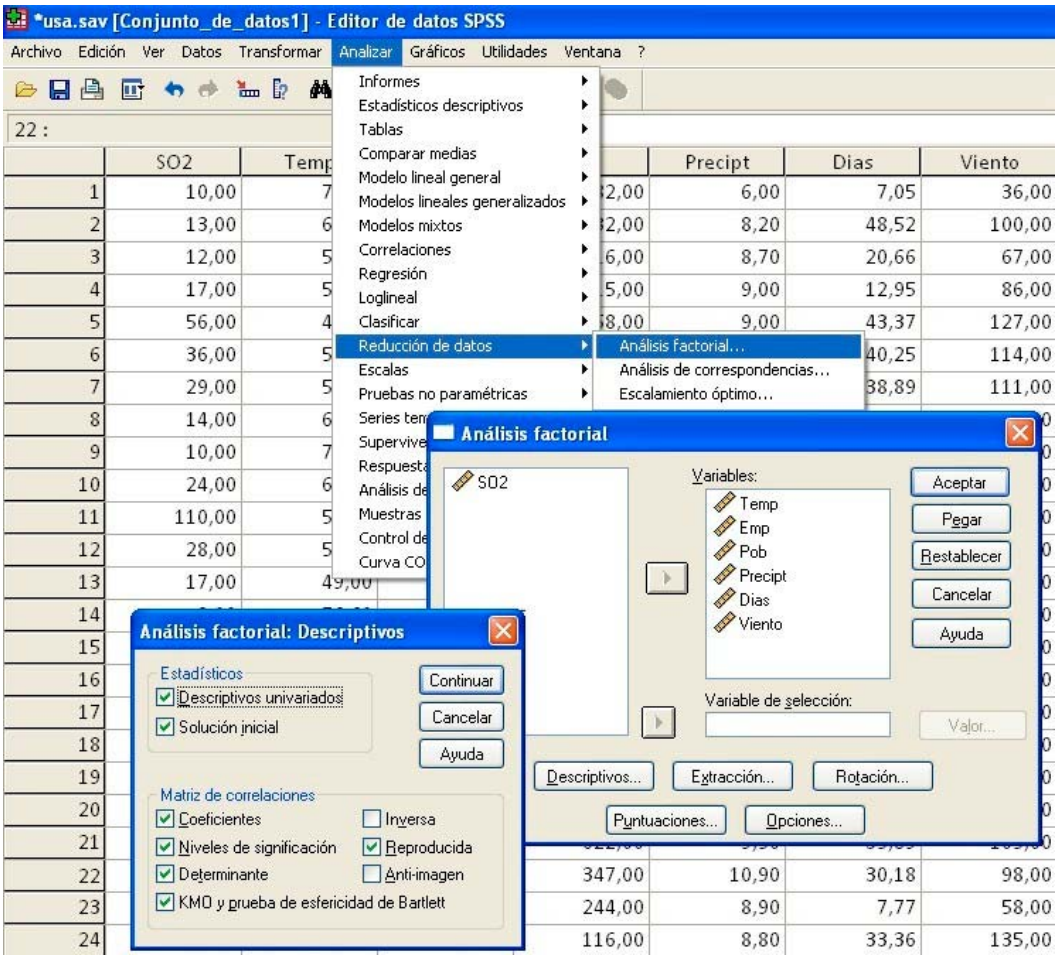
ANÁLISIS DE COMPONENTES PRINCIPALES CON SPSS

El objetivo del Análisis de Componentes Principales es identificar a partir de un conjunto de ***p*** variables, otro conjunto de ***k*** ($k < p$) variables no directamente observables, denominadas factores, tal que:

- ***k*** **sea un número pequeño**
- **Se pierda la menor cantidad posible de información**
- **La solución obtenida sea interpretable.**

Pasos en el Análisis de Componentes Principales:

- **Evaluación de lo apropiado de realizar el análisis.**
- **Extracción de los factores.**
- **Cálculo de las puntuaciones factoriales para cada caso.**



Descriptivos univariados.- Muestra, para cada variable, el número de casos válidos, la media y desviación típica.

Solución inicial.- Permite obtener las comunidades iniciales, los autovalores de la matriz analizada y los porcentajes de varianza asociada a cada valor.

MATRIZ DE CORRELACIONES

Coeficientes.- Muestra la matriz con los coeficientes de correlación entre las variables utilizadas en el análisis.

Niveles de significación.- Incluye en la matriz de correlaciones los niveles críticos asociados a este coeficiente.

Determinante.- Muestra el determinante de la matriz de correlaciones: El valor del determinante aparece en una nota a pie de tabla. Los determinantes próximos a cero están indicando que las variables utilizadas están linealmente relacionadas, lo que significa que el análisis factorial, es una técnica pertinente para analizar esas variables.

Inversa.- Muestra la inversa de la matriz de correlaciones. Esta matriz es la base para el cálculo de Comunalidades iniciales en algunos métodos de extracción y para el cálculo de la matriz anti-imagen.

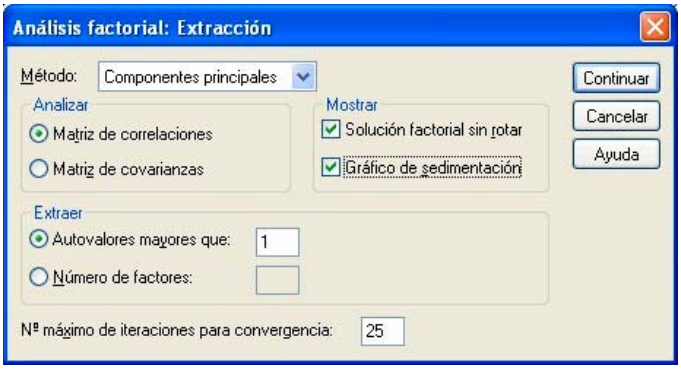
Reproducida.- Muestra la matriz reproducida. Es la matriz de las correlaciones que se obtiene a partir de la solución factorial hallada. Si el modelo es bueno y el número de factores el adecuado, la estructura factorial debe ser capaz de reproducir la matriz de correlaciones.
En la diagonal de la matriz reproducida se encuentran las Comunalidades finales.
Junto con la matriz de correlaciones reproducidas se muestra la matriz de correlaciones residuales, la cual contiene los residuos, es decir, las diferencias entre las correlaciones observadas y las correlaciones reproducidas.
Si el modelo es el correcto, el número de residuos con valores elevados debe ser mínimo.

Anti-imagen.- Muestra la matriz de covarianzas anti-imagen y la matriz de correlaciones anti-imagen. La matriz de covarianzas anti-imagen contiene los negativos de las covarianzas parciales y la matriz de correlaciones anti-imagen contiene los coeficientes de correlación parcial cambiados de signo (la correlaciones entre dos variables se parcializa teniendo en cuenta el resto de las variables incluidas en el análisis).
En la diagonal de la matriz de correlaciones anti-imagen se encuentra las medidas de adecuación muestral para cada variable. Si el modelo factorial elegido es adecuado para explicar los datos, los elementos de la diagonal del a matriz de correlaciones anti-imagen deben tener un valor próximo a 1 y el resto de elementos deben ser pequeños.

KMO y prueba de esfericidad de Bartlett.- La media de adecuación muestral KMO (Kaiser-Meyer-Olkin) contrasta si las correlaciones parciales entre las variables son suficientemente pequeñas. Permite comparar la magnitud de los coeficientes de correlación observados con la magnitud de los coeficientes de correlación parcial. EL estadístico KMO varía entre 0 y 1. Los valores pequeños indican que el análisis factorial puede no ser una buena idea, dado que las correlaciones entre los pares de variables no pueden ser explicadas por otras variables. Los menores que 0.5 indican que no debe utilizarse el análisis factorial con los datos muestrales que se están analizando.

La prueba de esfericidad de Bartlett.- Contrasta la hipótesis nula de que la matriz de correlaciones es una matriz identidad, en cuyo caso no existirían correlaciones significativas ente las variables y el modelo factorial no sería pertinente.

La opción [**Extracción**] permite controlar varios aspectos relacionados con la fase de extracción de los factores. Entre otras cosas, permite decidir que modelo factorial se desea utilizar, en qué matriz de datos basar el análisis y cuántos factores deben extraerse.



Matriz de correlaciones.- El análisis se basa en la matriz de correlaciones, en la matriz de correlaciones reducida, o en la matriz de correlaciones anti-imagen, según el método seleccionado.

Matriz de covarianza.- El análisis se basa en la matriz de varianzas covarianzas reducida, o la matriz de covarianzas anti-imagen, según el método seleccionado.

Autovalores mayores que.- Si la matriz analizada es la de correlaciones, esta opción permite utilizar el tamaño de los autovalores como un criterio para decidir si el número de factores que estarán presentes en la solución factorial. Por defecto se extraen los factores cuyos autovalores son mayores que la unidad (a este criterio se le denomina regla K1).
Si la matriz analizada es la de varianzas-covarianzas, la regla se expresa el número de veces que un autovalor debes sea mayor que el autovalor promedio de la matriz para que le correspondiente factor sea retenido en la solución.
El autovalor que actúa por defecto es 1, pero este valor puede cambiarse introduciendo otro distinto (entre cero y el número de variables) en el correspondiente cuatro de texto.

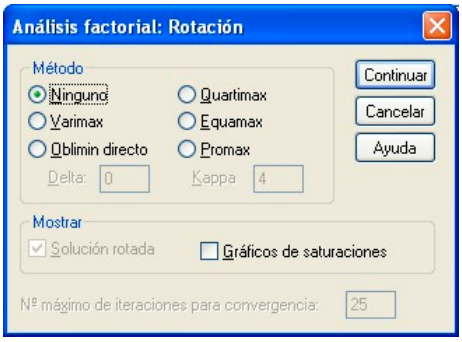
Numero de factores.- Permite especificar el número de factores exacto que se desea incluir en la solución. Se debe introducir el número en el cuadro de texto.

Solución factorial sin rotar.- Muestra las saturaciones o cargas factoriales sin rotar, las Comunalidades y los autovalores de la solución factorial.

Gráfico de sedimentación.- Muestra la representación gráfica de la magnitud de los autovalores. El corte en la tendencia descendente sirve de regla para la determinación del número de factores óptimo que deben estar presentes en la solución. Siempre se muestra la representación de los autovalores de la matriz de correlaciones (o de covarianzas) originales, independientemente del método de extracción seleccionado.

Nº de iteraciones para convergencia.- Este cuadro de texto permite establecer el número máximo de iteraciones que los algoritmos pueden realizar para encontrar una solución factorial final. El valor por defecto es 25, habitualmente suficiente para obtener una solución. Este valor puede cambiarse introduciendo un entero positivo.

La opción [**Rotación**] permite controlar que tipo de rotación llevar a cabo. Se puede definir el método de rotación que deseamos utilizar para facilitar su interpretación de la solución factorial y solicitar la representación gráfica de las saturaciones. Por defecto, no se encuentra seleccionado ningún método de rotación.



Ninguno.- No se aplica ningún método de rotación. Es la opción la que actúa por defecto. Cuando la solución consta de un único factor y no se ha marcado esta opción el visor de resultados muestra un mensaje de advertencia.

Varimax.- Método de rotación ortogonal que minimiza el número de variables que tiene saturaciones altas en cada factor. Simplifica la interpretación de los factores optimizando la solución por columna.

Quartimax.- Método de rotación ortogonal que minimiza el número de factores necesarios para explicar cada variable. Simplifica la interpretación de las variables observadas optimizando la interpretación por filas.

Equamax.- Método de rotación que es combinación del método varimax, que simplifica los factores, y el método Quartimax, que simplifica las variables. Se minimiza tanto el número de variables que saturan alto en un factor como el número de factores necesarios para explicar una variable.

Oblimin directo.- Método para la rotación oblicua (no ortogonal). Cuando delta e igual a cero (el valor por defecto), las soluciones son las más oblicuas. A medida que delta se va haciendo más negativo, los factores son menos oblicuos. Para anular el valor por defecto de delta, puede introducirse un número menor o igual a 0.8.

Delta.- El valor de delta permite controlar el grado de oblicuidad que pueden llegar a alcanzar los factores de la solución.

Promax.- Rotación oblicua que permite que los factores estén correlacionados. Puede calcularse más rápidamente que una rotación Oblimin directa, por lo que es útil para grandes conjuntos de datos.

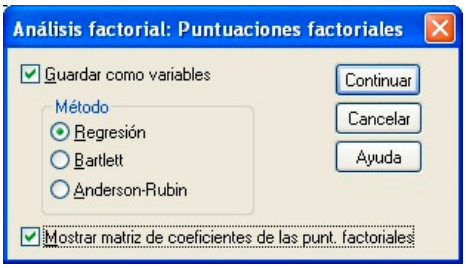
Kappa.- Parámetro que controla el cálculo de rotación de Promax. El valor por defecto es 4. Este valor es adecuado para la mayoría de los análisis.

Solución rotada.- Permite obtener una o más tablas con los resultados del proceso de rotación. Al seleccionar una rotación ortogonal, esta opción permite obtener la matriz de estructura factorial rotada y la matriz de transformación necesaria para rotar los factores a partir de la solución inicial. Además, en la tabla de porcentajes de varianza explicada aparecen columnas adicionales que contienen la varianza total explicada por los factores rotados.
Al seleccionar una rotación oblicua, esta opción permite obtener la matriz de configuración rotada, que contiene las saturaciones de las variables en los factores, y la matriz de estructura, que contiene las correlaciones entre las variables observadas y los factores (cuando la rotación es ortogonal, ambas matrices son idénticas).
Además, ofrece la matriz de correlaciones entre los factores y desecha la matriz de transformación para la rotación. En la tabla de porcentajes de varianza explicada sólo se incluyen los autovalores de los factores rotados (ya que no tiene sentido hablar de porcentajes de varianza independientes).

Gráficos de saturaciones.- Esta opción genera un gráfico de dispersión que refleja la ubicación de las variables en el espacio definido por los factores. Se trata de un gráfico de las saturaciones.
El gráfico muestra, por defecto, los tres primeros factores de la solución factorial en un gráfico tridimensional.
Si se desea representar otros factores, es necesario editar el gráfico y elegir esos otros factores.

Nº máximo de iteraciones para convergencia.- Permite determinar el número máximo de iteraciones que puede recorrer el algoritmo para la estimación de la solución rotada. Por defecto se efectúan un máximo de 25 iteraciones, lo que es suficiente para la mayoría de las situaciones.

La opción [**Puntuaciones**] se obtiene una estimación de las puntuaciones de los sujetos en cada uno de los factores resultantes de la extracción a fin de valorar la situación relativa de cada sujeto en esos 'constructor hipotéticos' capaces de resumir la información contenida en las variables originales.
El cuadro de diálogo Puntuaciones factoriales contiene las opciones que permiten solucionar las estimaciones de las puntuaciones factoriales y seleccionar el método de estimación que se desea utilizar para obtener tales estimaciones.



Señalar que por defecto se encuentra seleccionado el método de Regresión, que es el de uso más generalizado. Las opciones del método no tiene efecto alguno cuando se ha señalado componentes principales como método de extracción, ya que en ese modelo factorial las puntuaciones factoriales no son estimadas sino calculadas directamente a partir de las variables originales.

Guardar como variables.- Activando esta opción se guardan automáticamente en el Editor de datos las puntuaciones factoriales estimadas para cada sujeto en cada uno de los factores obtenidos en la solución factorial.
Para ello, el SPSS crea en el archivo de datos activo tantas variables nuevas como factores contenga la solución factorial. Si no se selecciona esta opción no es posible acceder a los métodos de estimación de las puntuaciones factoriales.

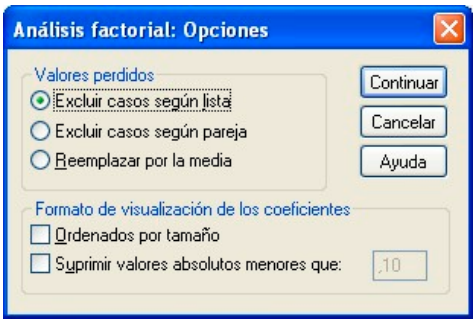
Regresión.- Método de estimación de las puntuaciones factoriales en el que las estimaciones resultantes tienen una media cero y una varianza igual al cuadrado de la correlación múltiple entre las puntuaciones factoriales estimadas y los valores factoriales verdaderos.
Las puntuaciones factoriales estimadas con este método pueden estar correlacionadas incluso cuando los factores son ortogonales.

Bartlett.- Método de estimación de las puntuaciones factoriales en el que las estimaciones resultantes tiene una media de cero.
Este método minimiza la suma de cuadrados de los factores únicos (es decir, minimiza la unicidad correspondiente a cada una de las variables incluidas en el análisis).

Anderson-Rubin.- Este método de estimación es una modificación del método de Bartlett que asegura la ortogonalidad de las puntuaciones factoriales estimadas. Las estimaciones resultantes tienen una media de cero, una desviación típica de uno y son independientes entre sí (incluso en el que se haya solicitado una solución rotada oblicua).

Mostrar matriz de coeficientes de las puntuaciones factoriales.- Esta opción permite obtener una tabla con los pesos o ponderaciones necesarios para calcular las puntuaciones factoriales a partir de las variables originales. Esta opción se encuentra desactivada por defecto. Por tanto, para obtener la matriz de coeficientes no basta con solicitar las puntuaciones factoriales.

El cuadro [**Opciones**] permite controlar algunos aspectos relacionados con le tratamiento que deben recibir los valores perdidos y el formato en las tablas de resultados que genera el Visor de resultados.



Excluir casos según la lista.- Es la opción por defecto.
Se excluyen el análisis los sujetos que tengan valores perdidos en cualquiera de las variables trasladadas a la lista de variables. Es el tratamiento más consistente de todos: sólo se incluyen en los análisis los casos completos (es decir, los casos con puntuación válida en todas las variables seleccionadas). Sin embargo, conviene tener en cuenta que esta forma de tratar los valores perdidos puede suponer la pérdida de un gran número de casos y la consiguiente reducción del tamaño efectivo de la muestra.

Excluir casos según pareja.- Los sujetos con valor perdido en una variable se excluyen del análisis sólo para el cálculo de los estadísticos en los que esté implicada esa variable.
Este método permite aprovechar más cantidad de información que la anterior, pero, puesto que no todas las correlaciones se calculan sobre el mismo número de sujetos, podrían obtenerse matrices de correlaciones inconsistentes imposibles de analizar posteriormente.

Reemplazar por la media.- Los valores perdidos de una variable se sustituyen por la media de esa variable. Si en una variable existen muy pocos casos con valor perdido, reemplazar el valor perdido por la media no constituye un problema importante. Pero en la medida de que el número de valores perdidos aumenta, la sustitución por la tiene el efecto de centrar las variables disminuyendo su variabilidad.

Ordenados por el tamaño.- Esta opción sirve para ordenar las variables de las tablas de resultados en función de la magnitud (en valor absoluto) de los coeficientes de esas tablas (saturaciones, correlaciones, etc.). La ordenación se realiza de forma ascendente: primero las variables con coeficientes más altos. Si no se marca esta opción, las tablas muestran las variables en el mismo orden en el que han sido trasladadas a la lista de Variables del cuadro de diálogo de Análisis factorial.

Suprimir valores absolutos menores que.- Esta opción permite suprimir de las tablas de resultados los coeficientes cuyo valor absoluto sea menor que el valor establecido en el cuadro de texto.
El valor por defecto es 0,10, pero este valor puede cambiarse introduciendo un valor distinto.
Esta opción es de gran ayuda: al desaparecer de la tabla los coeficientes excesivamente pequeños (en valor absoluto), se facilita notablemente la interpretación de los resultados.

Una vez señaladas las opciones, en la pantalla adjunta basta con pulsar [**Continuar**] para que el Visor SPSS nos facilite los resultados.

Se obtienen las medias y desviaciones típicas de cada variable en estudio.

Estadísticos descriptivos

	Media	Desviación típica	N del análisis
Temp	55,7634	7,22772	41
Emp	463,0976	563,47395	41
Pob	608,6098	579,11302	41
Precipt	9,4439	1,42864	41
Dias	36,7690	11,77155	41
Viento	113,9024	26,50642	41

Matriz de correlaciones[a]

		Temp	Emp	Pob	Precipt	Dias	Viento
Correlación	Temp	1,000	-,190	-,063	-,350	,386	-,430
	Emp	-,190	1,000	,955	,238	-,032	,132
	Pob	-,063	,955	1,000	,213	-,026	,042
	Precipt	-,350	,238	,213	1,000	-,013	,164
	Dias	,386	-,032	-,026	-,013	1,000	,496
	Viento	-,430	,132	,042	,164	,496	1,000
Sig. (Unilateral)	Temp		,117	,349	,012	,006	,002
	Emp	,117		,000	,067	,420	,206
	Pob	,349	,000		,091	,436	,397
	Precipt	,012	,067	,091		,468	,153
	Dias	,006	,420	,436	,468		,000
	Viento	,002	,206	,397	,153	,000	

a. Determinante = ,014

Matriz de las correlaciones con la significación de cada componente.
Para que se puede realizar el ACP, es necesario que las variables presenten factores comunes. Es decir, que estén muy correlacionadas entre sí.
Los coeficientes de la matriz de las correlaciones deben de ser grandes en valor absoluto.
Se obtienen los componentes principales a partir de la matriz de correlaciones para emplear las mismas escalas en todas las variables.

En este caso, según se observa en la parte inferior de la matriz de las correlaciones, el valor del determinante es 0,14

KMO y prueba de Bartlett

Medida de adecuación muestral de Kaiser-Meyer-Olkin.		,365
Prueba de esfericidad de Bartlett	Chi-cuadrado aproximado	159,231
	gl	15
	Sig.	,000

Test de esfericidad de Barlett.- Para comprobar que las correlaciones entre las variables son distintas de cero de modo significativo, se comprueba si el determinante de la matriz es distinto de uno, es decir, si la matriz de correlaciones es distinta de la matriz unidad.

Si las variables están correlacionadas hay muchos valores altos en valor absoluto fuera de la diagonal principal de la matriz de correlaciones, además, el determinante es menor que 1 (el máximo valor del determinante es 1 sí las variables están incorreladas).

El **test de Barlett** realiza el contraste: $\begin{cases} H_0 : |R| = 1 \\ H_1 : |R| \neq 1 \end{cases}$

El determinante de la matriz da una idea de la correlación generalizada entre todas las variables.
El test se basa en la distribución χ^2 de Pearson donde los valores altos llevan a rechazar la hipótesis nula H_0, así, la prueba de esfericidad de Barlett contrasta si la matriz de correlaciones es una matriz identidad, que indicaría que el modelo factorial es inadecuado.

Por otra parte, la medida de la adecuación muestral de **Kaiser-Meyer-Olkin** contrasta si las correlaciones parciales entre las variables son suficientemente pequeñas. El estadístico KMO varía entre 0 y 1. Los valores pequeños indican que el análisis factorial puede no ser una buena idea, dado que las correlaciones entre los pares de variables no pueden ser explicadas por otras variables. Los menores de 0,5 indican que no debe utilizarse el análisis factorial con los datos muestrales que se están analizando.

La **Comunalidad** asociada a la variable j-ésima es la proporción de variabilidad de dicha variable explicada por los k factores considerados.

Comunalidades

	Inicial	Extracción
Temp	1,000	,892
Emp	1,000	,968
Pob	1,000	,979
Precipt	1,000	,424
Dias	1,000	,941
Viento	1,000	,888

Método de extracción: Análisis de Componentes principales.

Equivale a la suma de la fila j-ésima de la matriz factorial. Sería igual a 0 si los factores comunes no explicaran nada la variabilidad de una variable, y sería igual a 1 se quedase totalmente explicada.

Varianza total explicada

	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción		
Componente	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	2,196	36,603	36,603	2,196	36,603	36,603
2	1,500	24,999	61,602	1,500	24,999	61,602
3	1,395	23,244	84,846	1,395	23,244	84,846
4	,760	12,670	97,516			
5	,115	1,910	99,426			
6	,034	,574	100,000			

Método de extracción: Análisis de Componentes principales.

La **varianza asociada** a cada factor se utiliza para determinar cuántos factores deben retenerse.

Los tres primeros factores tienen todos varianzas (autovalores) mayores que 1, y entre los tres recogen el 85% de la varianza de las variables originales.

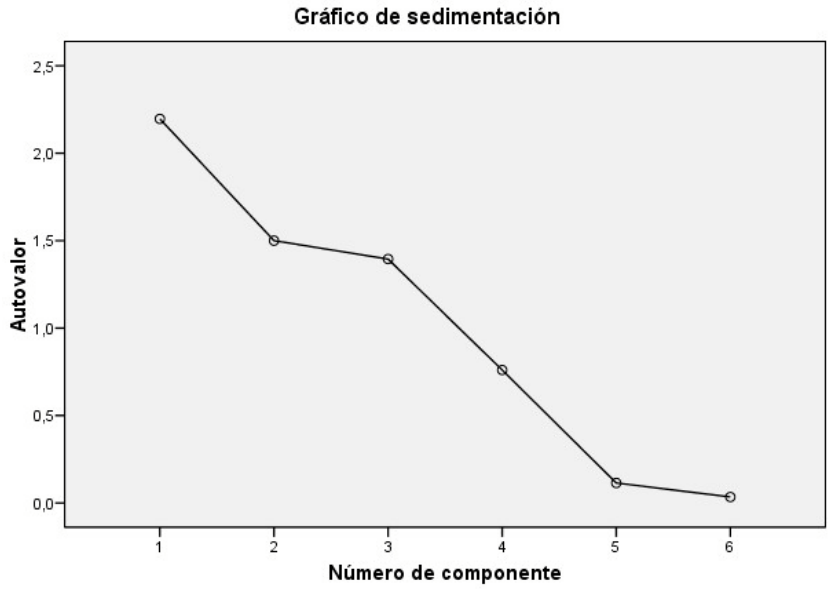
- El primer componente se le podría etiquetar como calidad de vida con valores negativos altos en empresas y población indicando un entorno relativamente pobre.

- El segundo componente se puede etiquetar como tiempo húmedo, y tiene pesos altos en las variables precipitaciones y días.

- El tercer componente se podría etiquetar como tipo de clima y está relacionado con la temperatura y la cantidad de lluvia.

Aunque no se encontrasen etiquetas claras para los componentes, siempre es interesante calcular componentes principales para descubrir si los datos se encuentran en una dimensión menor. De hecho, los tres primeros componentes producen un mapa de los datos donde las distancias entre los puntos es bastante semejante a la observada en los mismos respecto a las variables originales.

El **Gráfico de la varianza** asociada a cada factor se utiliza para determinar cuántos factores deben retenerse. Típicamente el gráfico muestra la clara ruptura entre la pronunciada pendiente de los factores más importantes y el descenso gradual de los restantes (los sedimentos)



Otra opción es utilizar el criterio de Kaiser, que consiste en conservar aquellos factores cuyo autovalor asociado sea mayor que 1.

Saturaciones factoriales:

Matriz de componentes[a]

	Componente		
	1	2	3
Temp	-,489	-,156	,793
Emp	,906	-,206	,322
Pob	,856	-,272	,414
Precipt	,524	,160	-,351
Dias	-,060	,763	,596
Viento	,353	,867	-,110

Método de extracción: Análisis de componentes principales.

a. 3 componentes extraídos

Correlaciones reproducidas

		Temp	Emp	Pob	Precipt	Dias	Viento
Correlación reproducida	Temp	,892[b]	-,155	-,048	-,560	,383	-,395
	Emp	-,155	,968[b]	,965	,329	-,020	,106
	Pob	-,048	,965	,979[b]	,260	-,013	,020
	Precipt	-,560	,329	,260	,424[b]	-,119	,362
	Dias	,383	-,020	-,013	-,119	,941[b]	,574
	Viento	-,395	,106	,020	,362	,574	,888[b]
Residual[a]	Temp		-,035	-,015	,210	,003	-,035
	Emp	-,035		-,010	-,091	-,013	,026
	Pob	-,015	-,010		-,047	-,013	,022
	Precipt	,210	-,091	-,047		,106	-,198
	Dias	,003	-,013	-,013	,106		-,078
	Viento	-,035	,026	,022	-,198	-,078	

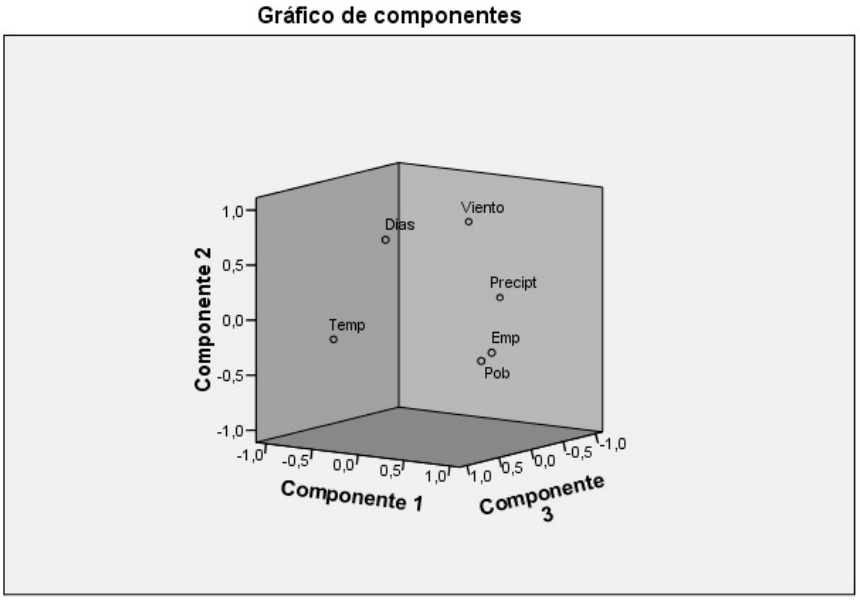
Método de extracción: Análisis de Componentes principales.

a. Los residuos se calculan entre las correlaciones observadas y reproducidas. Hay 5 (33,0%) residuales no redundantes con valores absolutos mayores que 0,05.

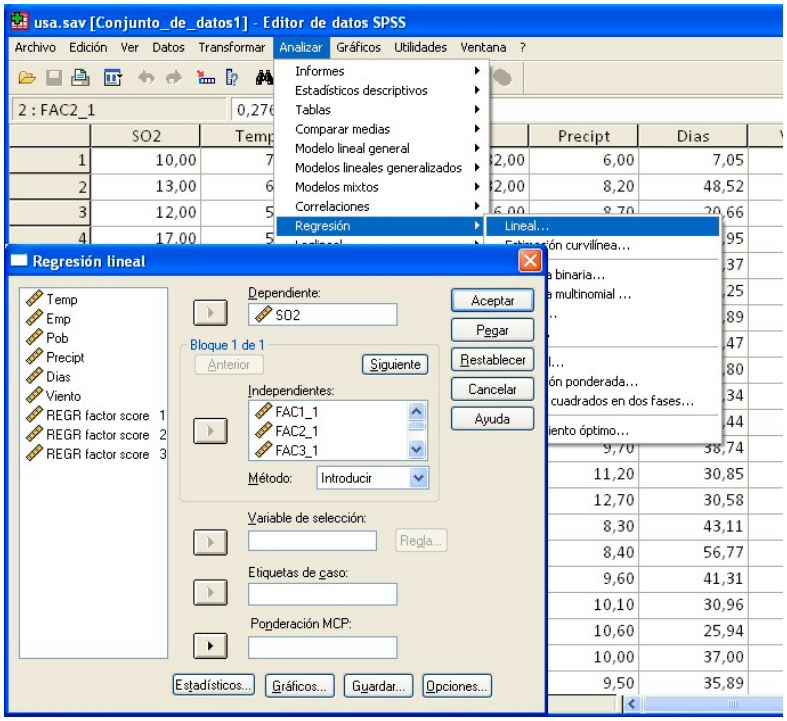
b. Comunalidades reproducidas

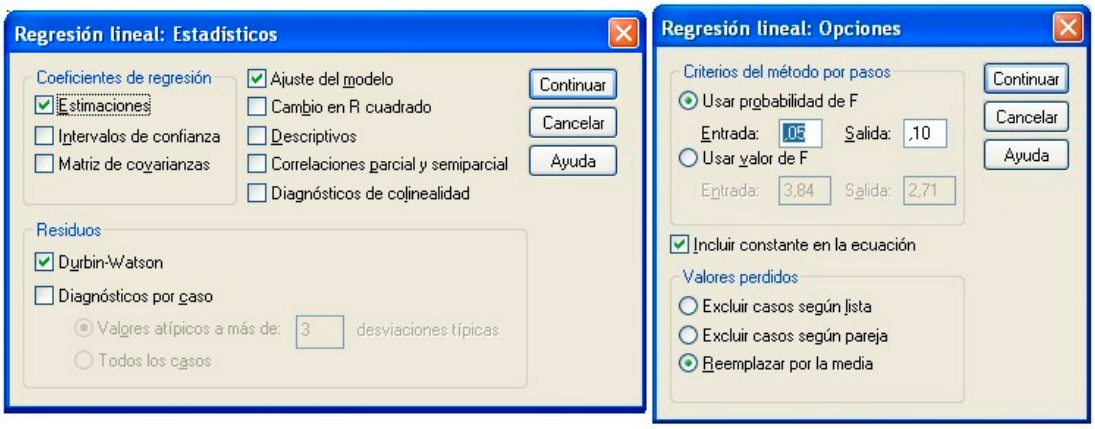
En la diagonal de la matriz reproducida se encuentran las Comunalidades finales.
Junto con la matriz de correlaciones reproducidas se muestra la matriz de correlaciones residuales, la cual contiene los residuos, es decir, las diferencias entre las correlaciones observadas y las correlaciones reproducidas.
Si el modelo es el correcto, el número de residuos con valores elevados debe ser mínimo.

Representación tridimensional de las saturaciones factoriales para los tres primeros factores:



Se realiza un **análisis de regresión** de la variable SO_2 sobre los tres factores. Para ello, en SSPS:


usa.sav [Conjunto_de_datos1] - Editor de datos SPSS
Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana ?
Informes
Estadísticos descriptivos
Tablas
Comparar medias
Modelo lineal general
Modelos lineales generalizados
Modelos mixtos
Correlaciones
Regresión
Lineal...
Regresión lineal
Temp
Emp
Pob
Precipt
Dias
Viento
REGR factor score 1
REGR factor score 2
REGR factor score 3
Dependiente:
SO2
Bloque 1 de 1
Anterior
Siguiente
Independientes:
FAC1_1
FAC2_1
FAC3_1
Método:
Introducir
Variable de selección:
Regla...
Etiquetas de caso:
Ponderación MCP:
Aceptar
Pegar
Restablecer
Cancelar
Ayuda
Estadísticos...
Gráficos...
Guardar...
Opciones...


Regresión lineal: Estadísticos
Coeficientes de regresión
Estimaciones
Intervalos de confianza
Matriz de covarianzas
Ajuste del modelo
Cambio en R cuadrado
Descriptivos
Correlaciones parcial y semiparcial
Diagnósticos de colinealidad
Continuar
Cancelar
Ayuda
Residuos
Durbin-Watson
Diagnósticos por caso
Valores atípicos a más de: 3 desviaciones típicas
Todos los casos
Regresión lineal: Opciones
Criterios del método por pasos
Usar probabilidad de F
Entrada: ,05 Salida: ,10
Usar valor de F
Entrada: 3,84 Salida: 2,71
Incluir constante en la ecuación
Valores perdidos
Excluir casos según lista
Excluir casos según pareja
Reemplazar por la media
Continuar
Cancelar
Ayuda

La salida del visor de SSPS muestra:

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación	Durbin-Watson
1	,647[a]	,418	,371	18,61510	1,926

a. Variables predictoras: (Constante), REGR factor score 3 for analysis 1, REGR factor score 2 for analysis 1, REGR factor score 1 for analysis 1

b. Variable dependiente: SO2

El estadístico de Durbin-Watson de 1,926 deja claro que la no autocorrelación de los factores.

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	9216,590	3	3072,197	8,866	,000[a]
	Residual	12821,313	37	346,522		
	Total	22037,902	40			

a. Variables predictoras: (Constante), REGR factor score 3 for analysis 1, REGR factor score 2 for analysis 1, REGR factor score 1 for analysis 1

b. Variable dependiente: SO2

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	30,049	2,907		10,336	,000
	REGR factor score 1 for analysis 1	14,917	2,943	,635	5,068	,000
	REGR factor score 2 for analysis 1	2,777	2,943	,118	,943	,352
	REGR factor score 3 for analysis 1	,448	2,943	,019	,152	,880

a. Variable dependiente: SO2

$$SO_2 = 30{,}049 + 14{,}917(\text{factor score } 1) + 2{,}777(\text{factor score } 2) + 0{,}448(\text{factor score } 3)$$

La cantidad de SO_2 se explica claramente mediante el primer componente de calidad de vida (con valores negativos altos en empresas y población indicando un entorno relativamente pobre).

ANÁLISIS ACP CON SPSS

El Análisis de Componentes Principales (**ACP**) tratará de representar ''de forma clara y ordenada", la variedad de los comportamientos observados en un conjunto de ***n*** individuos mediante un conjunto de ***p*** variables.

Buscará un nuevo sistema de ejes coordenados, ordenados (nuevas variables de referencia que llamaremos *componentes principales*) con el que poder apreciar y analizar más claramente la diversidad de comportamiento reflejada en los datos. Para ello, determinará como primer eje coordenado la nueva variable (primera componente principal) que explique la máxima variabilidad (diversidad) posible de los datos observados, para proceder secuencialmente y de forma análoga a determinar los sucesivos ejes coordenados (sucesivas componentes principales) a partir del resto de la variabilidad (diversidad) de los datos, aún no explicada por los anteriores.

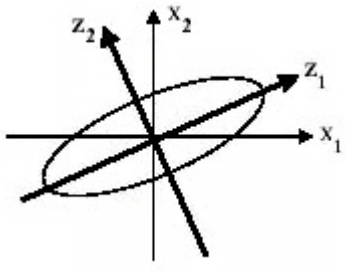
El **ACP** tratará de responder a la pregunta ¿en qué sistema de nuevos ejes coordenados podríamos apreciar de una forma más clara y ordenada la diversidad de información?

Representando por $(X_1, X_2, \cdots, X_p)$ las variables originales, el objetivo es pues, encontrar unas nuevas variables 'componentes principales', $(Z_1, Z_2, \cdots, Z_p)$, que expliquen ordenadamente y de una forma más clara la variabilidad de los datos. Parece lógico determinar la primera componente principal Z_1 como aquella que vaya en la dirección de máxima variabilidad de los datos y que, por tanto, explicará la mayor diversidad entre los datos; ya que los datos se dispersan de una forma máxima justamente en esa dirección. Esta dirección, pues, nos informará mucho del comportamiento más diversamente llamativo de esa nube de puntos.

De otra parte, obsérvese que para que estas nuevas variables de referencia (nuevo sistema de ejes coordenados) permita una representación "clara" de la realidad, deberíamos pedir lógicamente que estuviesen incorrelacionadas para que cada nueva variable informara de aspectos diferentes de la realidad y así facilitar la interpretación.
Recordemos que nubes de puntos inclinadas indicaban correlación entre variables y que nubes de puntos paralelas a los ejes indicaban incorrelación entre variables, por lo que la incorrelación entre las nuevas variables de referencia (componentes principales) se conseguirá cuando se tomen paralelas a los ejes principales de la nube de puntos. Ello nos induce a pensar que si la nube de puntos es lo suficientemente regular (aproximadamente elipsoidal), la dirección de las componentes principales deben ser ejes ortogonales.

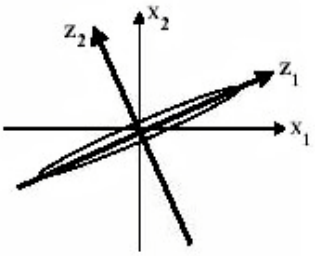
Así pues, la variable Z_2 deberá ser la variable que, siendo ortogonal a Z_1, tenga la dirección de máxima dispersión de las restantes. Así aportará una información adicional del resto de la variabilidad de los datos y que no quedaba explicada por la dirección Z_1 (nótese que existe toda una gama de individuos con un mismo valor para Z_1 que pueden presentar diferentes valores para Z_2).

El proceso se refleja en la figura:



Secuencialmente, las sucesivas componentes principales irán perdiendo importancia explicativa de la diversidad o variabilidad de los datos, ya que se extienden en direcciones de cada vez menos dispersión. Esto se acentuará más cuanto mayor sea la correlación entre las variables originales.

Cuanta mayor dependencia haya entre ellas, más alargada será la nube de puntos en alguna dirección y más estrecha en alguna dirección perpendicular [suponiendo siempre que la relación entre ellas fuera lineal].

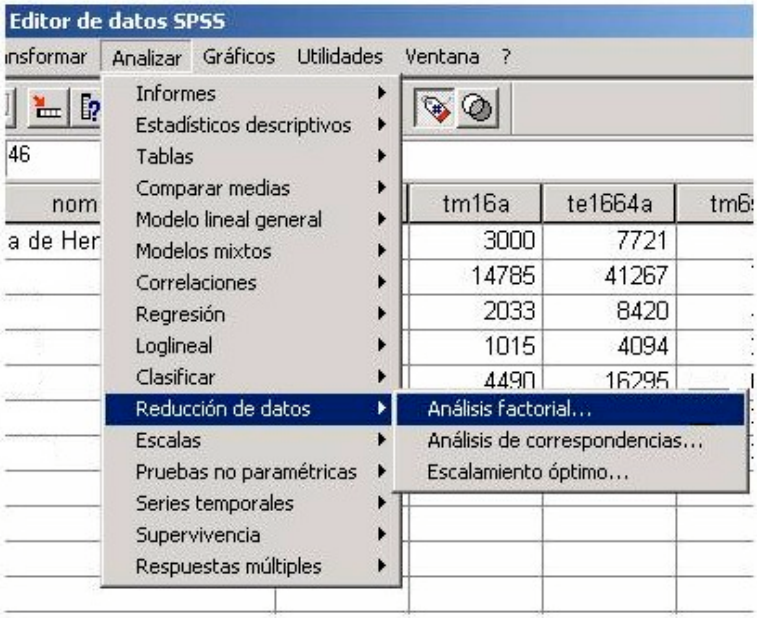


En el caso límite de que esa regresión fuera perfecta, y por tanto todos los puntos estuvieran sobre el hiperplano, la componente principal perpendicular al hiperplano no aportaría ninguna información porque no habría variabilidad en su dirección. Es en estos casos cuando vamos a conseguir una reducción efectiva de la dimensión de nuestro problema, al poder obviar o suprimir las componentes principales que no aportan información sobre la diversidad.

Así que, como consecuencia del proceso, el **ACP** no sólo encuentra ordenadamente las direcciones que mejor explican la variabilidad de esa nube de puntos, sino que también en el caso de que haya información redundante, permitirá prescindir de alguna de las últimas componentes, bien porque estrictamente no expliquen nada acerca de la variación de los datos, o bien porque expliquen una cantidad despreciable de la misma, consiguiendo simplificar el problema mediante la reducción efectiva de la dimensión del mismo.

El estudio de las **Componentes Principales** con **SPSS** se realiza a través del **Análisis Factorial**, el cual intenta identificar variables subyacentes, o factores que expliquen la configuración de correlaciones dentro de un conjunto de variables observadas.
Para que éste procedimiento estadístico tenga sentido, es necesario que entre las variables de estudio haya una estructura importante de correlación, es decir, es necesario que las variables han sido observadas estén relacionadas entre sí.

En **SPSS**, el procedimiento que permite realizar el **Análisis Factorial** se encuentran en el submenú **Reducción de datos** del menú **Analizar**: **Analizar/Reducción de datos/Análisis factorial**



Al hacer 'clic' en la opción, aparece el cuadro de diálogo adjunto donde aparecen todas las opciones que permite este procedimiento.

Se seleccionan las variables que vayan a ser incluidas en el análisis.

En la opción [Descriptivos] figuran una serie de medidas.

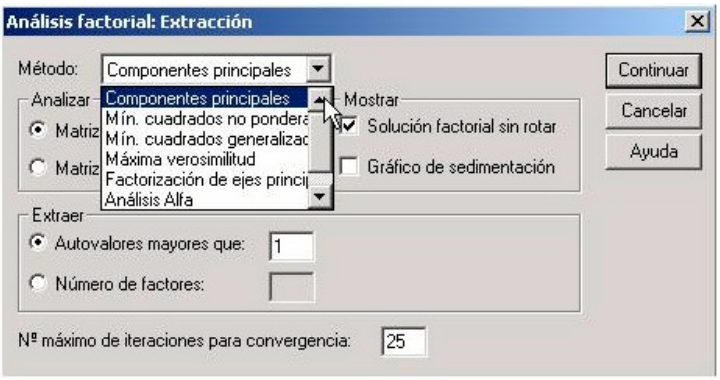
Desde la práctica, la **prueba de esfericidad de Bartlett** contrasta si la matriz de correlaciones es una matriz identidad, lo cual indicaría que el modelo factorial es inadecuado.

El **estadístico de Bartlett** se obtiene a partir de una transformación χ^2 del determinante de la matriz de correlaciones y cuanto mayor sea, y por tanto menor el nivel de significación, más improbable es que la matriz sea una matriz identidad y más adecuado resulta el análisis factorial.

- La medida de la adecuación muestral de **Kaiser-Meyer-Olkin** (***Coeficiente KMO***) contrasta si las correlaciones parciales entre las variables son pequeñas, toma valores entre 0 y 1, e indica que el análisis factorial es tanto más adecuado cuanto mayor sea su valor. Así, Kaiser propuso en 1974 el siguiente criterio para decidir sobre la adecuación del análisis factorial de un conjunto de datos:

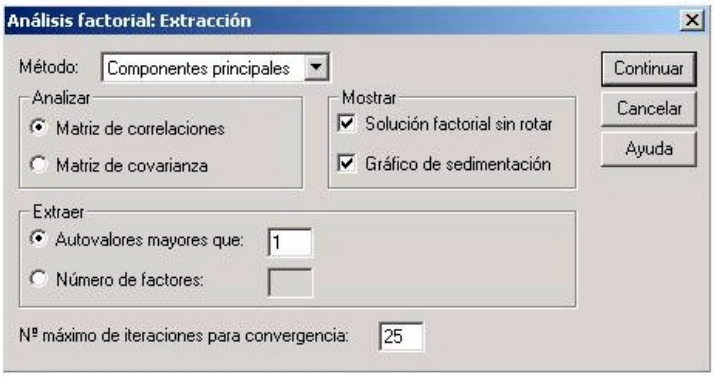
  0,9≤KMO≤1,0 = Excelente adecuación muestral.
  0,8≤KMO≤0,9 = Buena adecuación muestral.
  0,7≤KMO≤0,8 = Aceptable adecuación muestral.
  0,6≤KMO≤0,7 = Regular adecuación muestral.
  0,5≤KMO≤0,6 = Mala adecuación muestral.
  0,0≤KMO≤0,5 = Adecuación muestral inaceptable.

Los métodos de extracción de factores que realiza el SPSS son los de las **componentes principales**, **máxima verosimilitud, mínimos cuadrados no ponderados**, y algunos más.



Extracción...

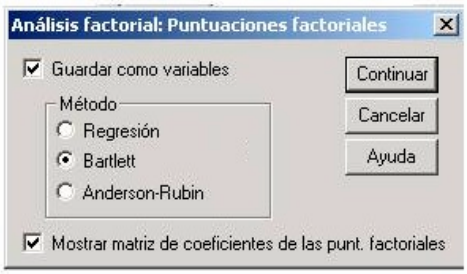
Se puede especificar que el análisis se aplique a una matriz de correlaciones o a una matriz de covarianzas. Se puede seleccionar 'a priori' el número de factores que se desea extraer, o especificar



alguna condición genérica que permita extraer sólo aquellos que verifiquen una determinada condición (usualmente se eligen aquellos factores cuyos autovalores sean superiores a la unidad).
Se puede mostrar la solución factorial sin rotar, así como el gráfico de sedimentación (criterio gráfico para la posterior decisión del número de factores a extraer).

Puntuaciones...

Se pueden almacenar las puntuaciones factoriales obtenidas a partir del análisis factorial en el área de trabajo del fichero de datos, es decir, se puede añadir ***m*** nuevas variables que representen los ***m*** factores extraídos.



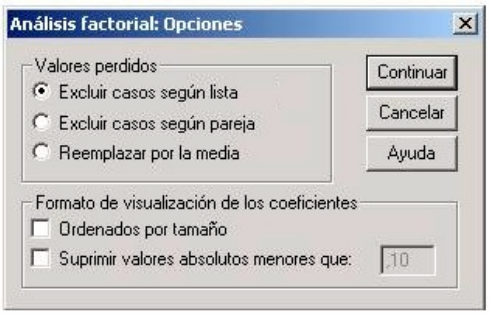
La matriz de coeficientes de las puntuaciones factoriales muestra los coeficientes por los cuales se multiplican las variables para obtener las puntuaciones factoriales.

Rotación...

Se puede elegir no rotar la solución inicial obtenida, o elegir alguno de los métodos de rotación que aparecen en las opciones del SPSS.
Además, se pueden representar las variables observadas en función de los factores extraídos, si se solicitan los Gráficos de saturaciones.

Opciones...

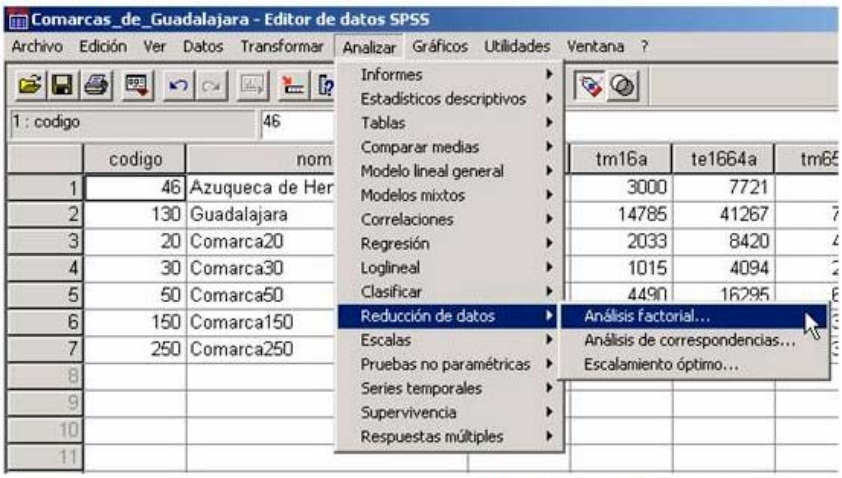
La del tratamiento de **valores perdidos**, en donde elegir 'excluir casos según lista', 'excluir casos según pareja' o 'reemplazar por la media'. Y en **Formato de visualización de los coeficientes**, se puede elegir 'Ordenados por tamaño' y 'suprimir valores absolutos menores que' en donde se puede elegir una opción numérica para eliminar aquellos valores que tengan un número menor al seleccionado.



APLICACIÓN PRÁCTICA DEL ANÁLISIS ACP CON SPSS

Como ejemplo, con el fichero de datos Comarcas de Guadalajara (Guadalajara.sav). Las variables que se incluyen en el análisis:
Lo que significa cada una de las variables son:
Agri: porcentaje de la población que trabaja en el sector agrícola.
Asal: Porcentaje de la población asalariada.
Cons: Porcentaje de la población que trabaja en el sector construcción.
Emp: Porcentaje de la población que posee su propia empresa.
Ind: Porcentaje de la población que trabaja en el sector industrial.
M16a: Porcentaje de la población de 16 años o menos.
M65a: Porcentaje de la población de 65 años o más.
Serv: Porcentaje de la población que trabaja en el sector servicios.
Tactiv: Tasa de población activa.

	codigo	nombre	total91	tm16a	te1664a	tm65a
1	46	Azuqueca de Henares	11624	3000	7721	903
2	130	Guadalajara	63649	14785	41267	7597
3	20	Comarca20	14752	2033	8420	4299
4	30	Comarca30	7232	1015	4094	2123
5	50	Comarca50	26802	4490	16295	6017
6	150	Comarca150	10324	1169	5830	3325
7	250	Comarca250	11210	1315	6189	3706
8						

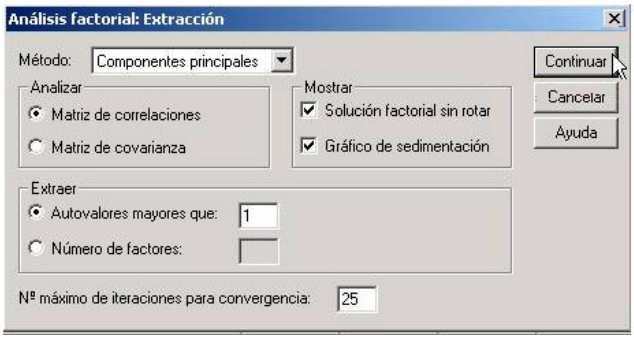


Descriptivos...

Se seleccionan: Descriptivos univariados, Solución inicial, Matriz de Coeficientes y el test KMO y prueba de esfericidad de Bartlett.

Basta presionar el botón [continuar] para proseguir con el análisis.

Extracción...



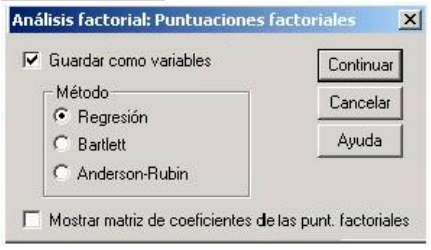
Se elige Matriz de correlaciones y Autovalores mayores que 1.

En [Mostrar], las opciones Solución inicial sin rotar y Gráfico de sedimentación.

Rotación...

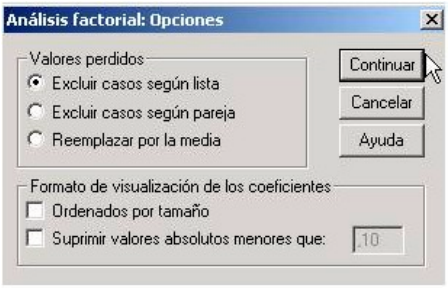
Cada uno de los métodos tiene su aplicación, y dependerá del caso en el cuál nos encontremos, para decidirse por uno u por otro método.

Puntuaciones...



Seleccionamos la opción de Guardar como variables, a través del método de Regresión. Basta dar un [clic] en el botón de [Continuar] y todo lo que hemos elegido queda guardado.

Opciones...



Se elige Excluir casos según lista.
Hacer [clic] en la opción [continuar].
Queda todo previsto para desarrollar el **ACP**.

Seleccionadas todas las opciones, se hace [clic] en el botón de [Aceptar] y SPSS genera toda la información solicitada.

El visor de resultados del SPSS muestra todos los cuadros, gráficos y resultados del análisis solicitado. En este caso, un Análisis Factorial a con el Método de Componentes Principales.

El primer cuadro presenta los **Estadísticos Descriptivos**, con la media y desviación típica de cada una de las variables en estudio.

Estadísticos descriptivos

	Media	Desviación típica	N del análisis
AGRI	19,0443	12,22738	7
ASAL	65,3586	15,14112	7
CONS	14,3286	4,18580	7
EMP	31,1314	14,09154	7
IND	24,2400	10,18694	7
M16A	16,6643	5,69981	7
M65A	23,7043	10,12656	7
SERV	42,3857	8,21876	7
TACTIV	41,01	8,465	7

Comunalidades

	Inicial	Extracción
AGRI	1,000	,963
ASAL	1,000	,988
CONS	1,000	,809
EMP	1,000	,976
IND	1,000	,927
M16A	1,000	,979
M65A	1,000	,965
SERV	1,000	,898
TACTIV	1,000	,893

Método de extracción: Análisis de Componentes principales

Las **Comunalidades** aparecen al principio, y son muy altas (cercanas a 1), con lo cual se afirma que las variables quedan muy bien explicadas a través de las componentes extraídas.

En el cuadro de la varianza total explicada de cada componente y cuáles son las componentes que han sido extraídas (aquellas cuyos autovalores superan la unidad).

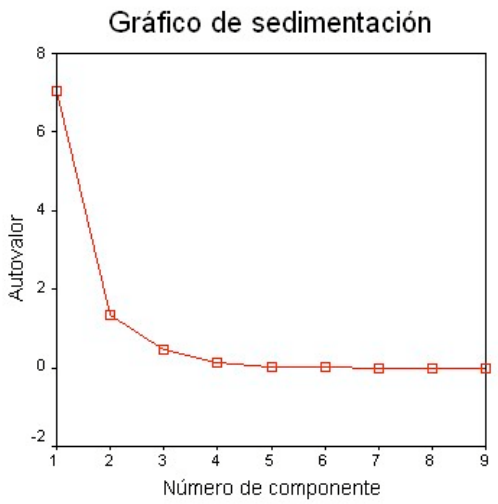
Varianza total explicada

	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción		
Componente	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	7,045	78,277	78,277	7,045	78,277	78,277
2	1,354	15,046	93,323	1,354	15,046	93,323
3	,448	4,978	98,300			
4	,117	1,295	99,596			
5	,030	,330	99,926			
6	,007	,074	100,000			
7	6,435E-16	7,150E-15	100,000			
8	9,622E-17	1,069E-15	100,000			
9	-1,302E-16	-1,447E-15	100,000			

Método de extracción: Análisis de Componentes principales.

Entre las dos primeras componentes extraídas se acumula el 93,323% de la variabilidad de las variables originales.

En el **Gráfico de Sedimentación** (herramienta gráfica para la decisión del número de componentes que hay que seleccionar) se visualiza que la selección de dos primeras componentes parece ser adecuada, pues a partir de la tercera componente no es muy acusada la pendiente de la representación gráfica de los autovalores.



Matriz de componentes (a)

	Componente	
	1	2
AGRI	-,974	-,121
ASAL	,993	,038
CONS	-,458	,774
EMP	-,986	-,053
IND	,869	,414
M16A	,980	,137
M65A	-,975	-,119
SERV	,606	-,729
TACTIV	,945	-,023

Método de extracción: Análisis de componentes principales.
(a) 2 componentes extraídos

La **Matriz de Componentes** que aparece en la salida se denomina **Matriz de Cargas** o **Saturaciones Factoriales**, indica la carga de cada variable en cada factor, de forma que los factores con pesos factoriales más elevados en términos absolutos indican una relación estrecha con las variables.

Se puede expresar cada variable en función de los factores.- Haciendo una combinación lineal de ellos utilizando sus cargas factoriales respectivas. De este modo, se puede expresar la variable **Agri** en función de las dos componentes extraídas: **Agri = – 0,974 F_1 – 0,121 F_2**

A partir de las **Cargas Factoriales** se calcula la **Comunalidad** de cada una de las variables, por ejemplo, para la variable Agri: **Comunalidad (Agri) = $(– 0,974)^2 + (– 0,121)^2 = 0,963317$**
Indicando qué cantidad de información original se conserva (96,33%).

El **Gráfico de Saturaciones** (gráfico de componentes principales) visualiza la representación gráfica de la matriz de componentes analizados.
De la representación se extrae la explicación de los factores subyacentes, de manera que:

- El **primer factor** es un factor de tipo económico-demográfico, que se opone a las variables **Agri, Cons, Emp y M65a** al resto.

- El **segundo factor** es un factor de tipo ocupacional, y separa los sectores en los que trabaja la población.