Tema 3: Análisis de Componentes Principales

Introducción a la distribución normal multivariante

Cuando se trabaja en la vida real, la suposición más habitual es que la variable en estudio se distribuye como una normal: muchas características que se miden son la conjunción de muchas causas que actúan conjuntamente sobre el suceso. Por ejemplo, la altura de las personas se considera que se distribuye como una normal, ya que su valor es debido a múltiples causas ambientales, alimentarias y genéticas.

La justificación matemática de esto se encuentra en el *Teorema Central del Límite* que demuestra que la suma de variables independientes se distribuye en el límite como una normal.

Teorema Central del Límite

Si $X_1, \ldots, X_n$ son v.a. independientes con media μ y varianza común $\sigma^2 < \infty$, la v.a. Z definida como

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

es una v.a. cuya función de densidad se aproxima a la distribución normal cuando n es grande:

$$Z \sim N(0,1)$$

esto es,

$$\frac{X_1 + \cdots + X_n}{n} = \overline{X} \simeq N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Distribución normal bivariante

Es una generalización para vectores de v.a. del modelo normal. En el caso bivariante, la distribución normal de un vector $(X,Y)'$ de media $\mu = (\mu_1, \mu_2)'$ y matriz de covarianzas

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \operatorname{cov}(X,Y) \\ \operatorname{cov}(X,Y) & \sigma_2^2 \end{pmatrix},$$

tiene como función de densidad

$$f(x,y) = \frac{1}{\left(\sqrt{2\pi}\right)^2 \sqrt{|\Sigma|}} \exp\left\{ -\frac{1}{2} \begin{bmatrix} x-\mu_1 & y-\mu_2 \end{bmatrix} \Sigma^{-1} \begin{bmatrix} x-\mu_1 \\ y-\mu_2 \end{bmatrix} \right\},$$

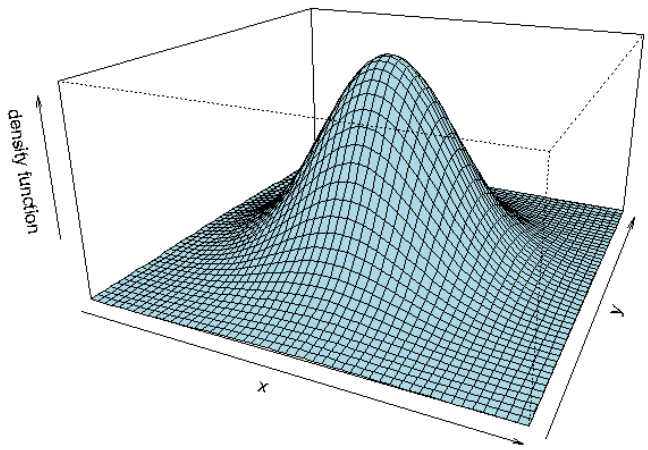
y se representa como $N(\mu, \Sigma)$,

Esta expresión se generaliza de modo inmediato al caso de un vector de v.a. con n componentes.

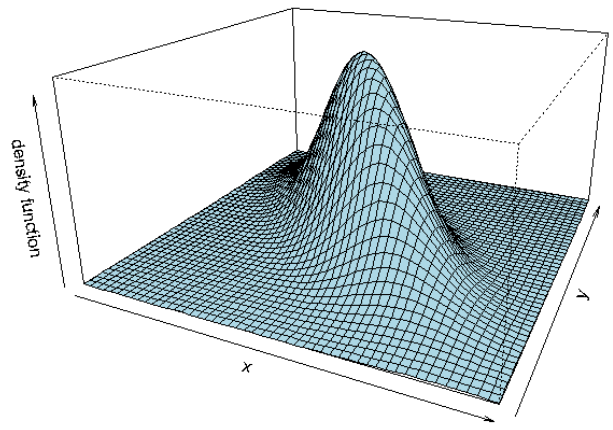
Por ejemplo, en `R` se puede dibujar la función de densidad con la siguiente secuencia de comandos:

```
library(mvtnorm)
n = 50
x = seq(-3, 3, length = n)
y = x
z = matrix(0,n,n)
sigma = diag(2)
for (i in 1:n)
    for (j in 1:n)
        z[i,j] = dmvnorm(c(x[i],y[j]),c(0,0), sigma)
    end
end
persp(x,y,z,theta=25,phi=20,zlab="density function",expand=0.5,col="blue")

# Con matriz de covarianzas diferente
```

```
z = matrix(0,n,n)
f1 = c(1,-0.75)
f2 = c(-0.675,1)
sigma = rbind(f1,f2)
for (i in 1:n)
    for (j in 1:n)
        z[i,j] = dmvnorm(c(x[i],y[j]),mean=c(0,0),sigma)
    end
end
persp(x,y,z,theta=25,phi=20,zlab="density function",expand=0.5,col="blue")
```



$N_2(\boldsymbol{\mu}, \Sigma)$ donde $\boldsymbol{\mu} = (0,0)'$, $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$



$$N_2(\boldsymbol{\mu}, \Sigma) \text{ donde } \boldsymbol{\mu} = (0,0)', \ \Sigma = \begin{bmatrix} 1 & -0{,}75 \\ -0{,}75 & 1 \end{bmatrix}$$

Propiedades

1. La distribución marginal de X es $N(\mu_1, \sigma_1)$

2. La distribución marginal de Y es $N(\mu_2, \sigma_2)$

3. La distribución de Y condicionada por $X = x$ es

$$N\left(\mu_2 + \frac{\text{cov}(X,Y)}{\sigma_1^2}(x - \mu_1); \quad \sigma_2\sqrt{1-\rho^2}\right)$$

   donde ρ es el coeficiente de correlación,

$$\rho = \frac{\text{cov}(X,Y)}{\sigma_1\sigma_2}$$

4. Si un vector aleatorio $(X,Y)'$ tiene distribución $N(\mu, \Sigma)$ y $\text{cov}(X,Y) = 0$ entonces X e Y son independientes. Como

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix},$$

sustituyendo en la expresión de la función de densidad, se obtiene que

$$f(x,y) = f(x) \cdot f(y)$$

Análisis de Componentes Principales

Introducción

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

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

$$\begin{aligned} y_j &= a_{j1}x_1 + a_{j2}x_2 + ... + a_{jp}x_p = \\ &= \mathbf{a}_j'\mathbf{x} \end{aligned}$$

siendo $\mathbf{a}_j' = (a_{1j}, a_{2j}, ..., a_{pj})$ un vector de constantes, y

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}$$

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 $\mathbf{a}_j' = (a_{1j}, a_{2j}, ..., a_{pj})$ sea

1. Es decir,

$$\mathbf{a}_j'\mathbf{a}_j = \sum_{k=1}^{p} a_{kj}^2 = 1$$

El primer componente se calcula eligiendo $\mathbf{a}_1$ de modo que y_1 tenga la mayor varianza posible, sujeta a la restricción de que $\mathbf{a}_1'\mathbf{a}_1 = 1$. El segundo componente principal se calcula obteniendo $\mathbf{a}_2$ de modo que la variable obtenida, y_2 esté incorrelada con y_1.

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

Proceso de extracción de factores:

Queremos elegir $\mathbf{a}_1$ de modo que se maximice la varianza de y_1 sujeta a la restricción de que $\mathbf{a}_1'\mathbf{a}_1 = 1$

$$Var(y_1) = Var(\mathbf{a}_1'\mathbf{x}) = \mathbf{a}_1'\Sigma\mathbf{a}_1$$

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 $\mathbf{a}_1'\Sigma\mathbf{a}_1$ sujeta a la restricción $\mathbf{a}_1'\mathbf{a}_1 = 1$.

Se puede observar que la incógnita es precisamente $\mathbf{a}_1$ (el vector desconocido que nos da la combinación lineal óptima).

Así, construyo la función L:

$$L(\mathbf{a}_1) = \mathbf{a}_1'\Sigma\mathbf{a}_1 - \lambda(\mathbf{a}_1'\mathbf{a}_1 - 1)$$

y busco el máximo, derivando e igualando a 0:

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{a}_1} &= 2\Sigma\mathbf{a}_1 - 2\lambda I\mathbf{a}_1 = 0 \implies \\ (\Sigma - \lambda I)\,\mathbf{a}_1 &= 0. \end{aligned}$$

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 $(\Sigma - \lambda I)$ tiene que ser singular. Esto implica que el determinante debe ser igual a cero:

$$|\Sigma - \lambda I| = 0$$

y 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,

$$\begin{aligned} (\Sigma - \lambda I)\,\mathbf{a}_1 &= 0 \\ \Sigma\mathbf{a}_1 - \lambda I\mathbf{a}_1 &= 0 \\ \Sigma\mathbf{a}_1 &= \lambda I\mathbf{a}_1 \end{aligned}$$

entonces,

$$\begin{aligned} Var(y_1) &= Var(\mathbf{a}_1'\mathbf{x}) = \mathbf{a}_1'\Sigma\mathbf{a}_1 = \mathbf{a}_1'\lambda I\mathbf{a}_1 = \\ &= \lambda\mathbf{a}_1'\mathbf{a}_1 = \lambda \cdot 1 = \lambda. \end{aligned}$$

Luego, para maximizar la varianza de y_1 se tiene que tomar el mayor autovalor, digamos λ_1, y el correspondiente autovector $\mathbf{a}_1$.

En realidad, $\mathbf{a}_1$ es un vector que nos da la combinación de las variables originales que tiene mayor varianza, esto es, si $\mathbf{a}_1' = (a_{11}, a_{12}, \ldots, a_{1p})$, entonces

$$y_1 = \mathbf{a}_1'\mathbf{x} = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p.$$

El segundo componente principal, digamos $y_2 = \mathbf{a}_2'\mathbf{x}$, 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:

$$\begin{aligned} Cov(y_2, y_1) &= Cov(\mathbf{a}_2'\mathbf{x}, \mathbf{a}_1'\mathbf{x}) = \\ &= \mathbf{a}_2' \cdot E\left[(\mathbf{x} - \mu)(\mathbf{x} - \mu)'\right] \cdot \mathbf{a}_1 = \\ &= \mathbf{a}_2'\Sigma\mathbf{a}_1, \end{aligned}$$

es decir, se requiere que $\mathbf{a}_2'\Sigma\mathbf{a}_1 = 0$.

Como se tenía que $\Sigma\mathbf{a}_1 = \lambda\mathbf{a}_1$, lo anterior es equivalente a

$$\mathbf{a}_2'\Sigma\mathbf{a}_1 = \mathbf{a}_2'\lambda\mathbf{a}_1 = \lambda\mathbf{a}_2'\mathbf{a}_1 = 0,$$

esto equivale a que $\mathbf{a}_2'\mathbf{a}_1 = 0$, es decir, que los vectores sean ortogonales.

De este modo, tendremos que maximizar la varianza de y_2, es decir, $\mathbf{a}_2\Sigma\mathbf{a}_2$, sujeta a las siguientes restricciones

$$\begin{aligned} \mathbf{a}_2'\mathbf{a}_2 &= 1, \\ \mathbf{a}_2'\mathbf{a}_1 &= 0. \end{aligned}$$

Se toma la función:

$$L(\mathbf{a}_2) = \mathbf{a}_2'\Sigma\mathbf{a}_2 - \lambda(\mathbf{a}_2'\mathbf{a}_2 - 1) - \delta\mathbf{a}_2'\mathbf{a}_1$$

y se deriva:

$$\frac{\partial L(\mathbf{a}_2)}{\partial \mathbf{a}_2} = 2\Sigma\mathbf{a}_2 - 2\lambda\mathbf{a}_2 - \delta\mathbf{a}_1 = 0$$

si se multiplica por $\mathbf{a}_1'$, entonces

$$2\mathbf{a}_1'\Sigma\mathbf{a}_2 - \delta = 0$$

porque

$$\begin{aligned} \mathbf{a}_1'\mathbf{a}_2 &= \mathbf{a}_2'\mathbf{a}_1 = 0 \\ \mathbf{a}_1'\mathbf{a}_1 &= 1. \end{aligned}$$

Luego

$$\delta = 2\mathbf{a}_1'\Sigma\mathbf{a}_2 = 2\mathbf{a}_2'\Sigma\mathbf{a}_1 = 0,$$

ya que $Cov(y_2, y_1) = 0$.

De este modo, $\frac{\partial L(\mathbf{a}_2)}{\partial \mathbf{a}_2}$ queda finalmente como:

$$\begin{aligned} \frac{\partial L(\mathbf{a}_2)}{\partial \mathbf{a}_2} &= 2\Sigma\mathbf{a}_2 - 2\lambda\mathbf{a}_2 - \delta\mathbf{a}_1 = 2\Sigma\mathbf{a}_2 - 2\lambda\mathbf{a}_2 = \\ (\Sigma - \lambda I)\,\mathbf{a}_2 &= 0 \end{aligned}$$

Usando los mismos razonamientos que antes, elegimos λ como el segundo mayor autovalor de la matriz Σ con su autovector asociado $\mathbf{a}_2$.

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

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

$$\mathbf{y} = A\mathbf{x}$$

donde

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix}, \quad A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix}$$

Como

$$Var(y_1) = \lambda_1$$

$$Var(y_2) = \lambda_2$$

$$\ldots$$

$$Var(y_p) = \lambda_p$$

la matriz de covarianzas de $\mathbf{y}$ será

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \lambda_p \end{pmatrix}$$

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

Se tiene que

$$\Lambda = Var(Y) = A'Var(X)A = A'\Sigma A$$

o bien

$$\Sigma = A\Lambda A'$$

ya que A es una matriz ortogonal (porque $\mathbf{a}_i'\mathbf{a}_i = 1$ para todas sus columnas) por lo que $AA' = I$.

Porcentajes de variablilidad

Vimos antes que, en realidad, cada autovalor correspondía a la varianza del componente y_i que se definía por medio del autovector $\mathbf{a}_i$, es decir, $Var(y_i) = \lambda_i$.

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

$$\sum_{i=1}^{p} Var(y_i) = \sum_{i=1}^{p} \lambda_i = traza\,(\Lambda)$$

ya que la matriz Λ es diagonal.

Pero, por las propiedades del operador traza,

$$traza(\Lambda) = traza(A'\Sigma A) = traza(\Sigma A' A) = traza(\Sigma),$$

porque $AA' = I$ al ser A ortogonal, con lo cual

$$traza(\Lambda) = traza(\Sigma) = \sum_{i=1}^{p} Var(x_i)$$

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:

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

(si multiplicamos por 100 tendremos el %).

Así, también se podrá expresar el porcentaje de variabilidad recogido por los primeros m componentes:

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

donde $m < p$.

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. La suma total de todos los autovalores es p y la proporción de varianza recogida por el autovector j-ésimo (componente) es

$$\frac{\lambda_j}{p}.$$

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:

$$\mathbf{a}_j^* = \lambda_j^{1/2}\mathbf{a}_j$$

para $j = 1, \ldots, p$.

De este modo, se suele presentar una tabla de autovectores $\mathbf{a}_j^*$ que forman la matriz factorial

$$F = (\mathbf{a}_1^*, \mathbf{a}_2^*, \ldots, \mathbf{a}_p^*)$$

Si se eleva al cuadrado cada una de las columnas y se suman los términos se obtienen los autovalores:

$$\mathbf{a}_j^{*\prime}\mathbf{a}_j^* = \lambda_j^{1/2} \cdot \lambda_j^{1/2}\mathbf{a}_j'\mathbf{a}_j = \lambda_j,$$

(porque $\mathbf{a}_j'\mathbf{a}_j = 1$).

Por otro lado, como

$$\Sigma = A\Lambda A'$$

y en SPSS presenta como matriz factorial a

$$F = A\Lambda^{1/2},$$

se tiene que

$$\Sigma = FF'.$$

Los elementos de F son tales que los mayores valores indican una mayor importancia a la hora de definir un componente.

Otra manera de verlo es considerar que como $\mathbf{y} = A\mathbf{x}$, entonces, $\mathbf{x} = A^{-1}\mathbf{y}$, de modo que

$$Cov(\mathbf{x}) = (A^{-1})'Cov(\mathbf{y})A^{-1} = A\Lambda A' = A\Lambda^{1/2}\Lambda^{1/2}A' = FF'$$

ya que al ser A ortogonal, resulta que $A^{-1} = A'$.

Así, dada la matriz factorial F, se pueden calcular las covarianzas de las variables originales, es decir, 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 Σ.

Cálculo de las covarianzas y correlaciones entre las variables originales y los factores

Como se tenía que $\mathbf{x} = A^{-1}\mathbf{y} = A'\mathbf{y}$, por ser A ortogonal, 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.

Si suponemos que las variables originales están estandarizadas: $Var(x_i) = 1$ para $i = 1, \ldots, p$, entonces

$$Cor(y_j, x_i) = \frac{\lambda_j a_{ij}}{1 \cdot \lambda_j^{1/2}} = \lambda_j^{1/2} a_{ij}.$$

De este modo, la matriz de correlaciones entre $\mathbf{y}$ y $\mathbf{x}$ es:

$$Cor(\mathbf{y}, \mathbf{x}) = \Lambda^{1/2} A' = F'$$

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 $x_1, ..., x_p$ 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

Una 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

	SO2	Neg.Temp	Empresas	Poblacion	Viento	Precip	Dias
Phoenix	10.00	70.30	213.00	582.00	6.00	7.05	36.00
Little Rock	13.00	61.00	91.00	132.00	8.20	48.52	100.00
San Francisco	12.00	56.70	453.00	716.00	8.70	20.66	67.00
Denver	17.00	51.90	454.00	515.00	9.00	12.95	86.00
Hartford	56.00	49.10	412.00	158.00	9.00	43.37	127.00
Wilmington	36.00	54.00	80.00	80.00	9.00	40.25	114.00
Washington	29.00	57.30	434.00	757.00	9.30	38.89	111.00
Jacksonville	14.00	68.40	136.00	529.00	8.80	54.47	116.00
Miami	10.00	75.50	207.00	335.00	9.00	59.80	128.00
Atlanta	24.00	61.50	368.00	497.00	9.10	48.34	115.00
Chicago	110.00	50.60	3344.00	3369.00	10.40	34.44	122.00
Indianapolis	28.00	52.30	361.00	746.00	9.70	38.74	121.00
Des Moines	17.00	49.00	104.00	201.00	11.20	30.85	103.00
Wichita	8.00	56.60	125.00	277.00	12.70	30.58	82.00
Louisville	30.00	55.60	291.00	593.00	8.30	43.11	123.00
New Orleans	9.00	68.30	204.00	361.00	8.40	56.77	113.00
Baltimore	47.00	55.00	625.00	905.00	9.60	41.31	111.00
Detroit	35.00	49.90	1064.00	1513.00	10.10	30.96	129.00
Minneapolis-St. Paul	29.00	43.50	699.00	744.00	10.60	25.94	137.00
Kansas City	14.00	54.50	381.00	507.00	10.00	37.00	99.00
St. Louis	56.00	55.90	775.00	622.00	9.50	35.89	105.00
Omaha	14.00	51.50	181.00	347.00	10.90	30.18	98.00
Alburquerque	11.00	56.80	46.00	244.00	8.90	7.77	58.00
Albany	46.00	47.60	44.00	116.00	8.80	33.36	135.00
Buffalo	11.00	47.10	391.00	463.00	12.40	36.11	166.00
Cincinnati	23.00	54.00	462.00	453.00	7.10	39.04	132.00
Cleveland	65.00	49.70	1007.00	751.00	10.90	34.99	155.00
Columbus	26.00	51.50	266.00	540.00	8.60	37.01	134.00
Philadelphia	69.00	54.60	1692.00	1950.00	9.60	39.93	115.00
Pittsburgh	61.00	50.40	347.00	520.00	9.40	36.22	147.00
Providence	94.00	50.00	343.00	179.00	10.60	42.75	125.00
Memphis	10.00	61.60	337.00	624.00	9.20	49.10	105.00
Nashville	18.00	59.40	275.00	448.00	7.90	46.00	119.00
Dallas	9.00	66.20	641.00	844.00	10.90	35.94	78.00
Houston	10.00	68.90	721.00	1233.00	10.80	48.19	103.00
Salt Lake City	28.00	51.00	137.00	176.00	8.70	15.17	89.00
Norfolk	31.00	59.30	96.00	308.00	10.60	44.68	116.00
Richmond	26.00	57.80	197.00	299.00	7.60	42.59	115.00
Seattle	29.00	51.10	379.00	531.00	9.40	38.79	164.00
Charleston	31.00	55.20	35.00	71.00	6.50	40.75	148.00
Milwaukee	16.00	45.70	569.00	717.00	11.80	29.07	123.00

Se dispone de una muestra de 41 ciudades de USA en las que se midieron diferentes variables relacionadas con la contaminación atmosférica.

Las variables son:

– Contenido en SO_2

– Temperatura anual en grados F.

– Número de empresas mayores de 20 trabajadores.

– Población (en miles de habitantes).

– Velocidad media del viento.

– Precipitación anual media.

– Días lluviosos al año.

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

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.

En la salida de R, se presentan las puntuaciones de las observaciones respecto a los tres factores combinando estos de dos en dos. Se observa que la ciudad de Chicago es un outlier y también, en menor medida, las ciudades de Phoenix y Philadelphia. Phoenix aparece como la ciudad con más calidad de vida, y Buffalo parece la más húmeda.

A continuación nos planteamos la cuestión de la relación o posible predicción de los niveles de SO_2 respecto a las variables de tipo ambiental. Se pueden representar los valores de concentración de SO_2 frente a cada uno de los tres componentes, aunque la interpretación puede ser subjetiva por la presencia de outliers. Aún así, parece que la contaminación está más relacionada con la primera componente que con las otras dos.

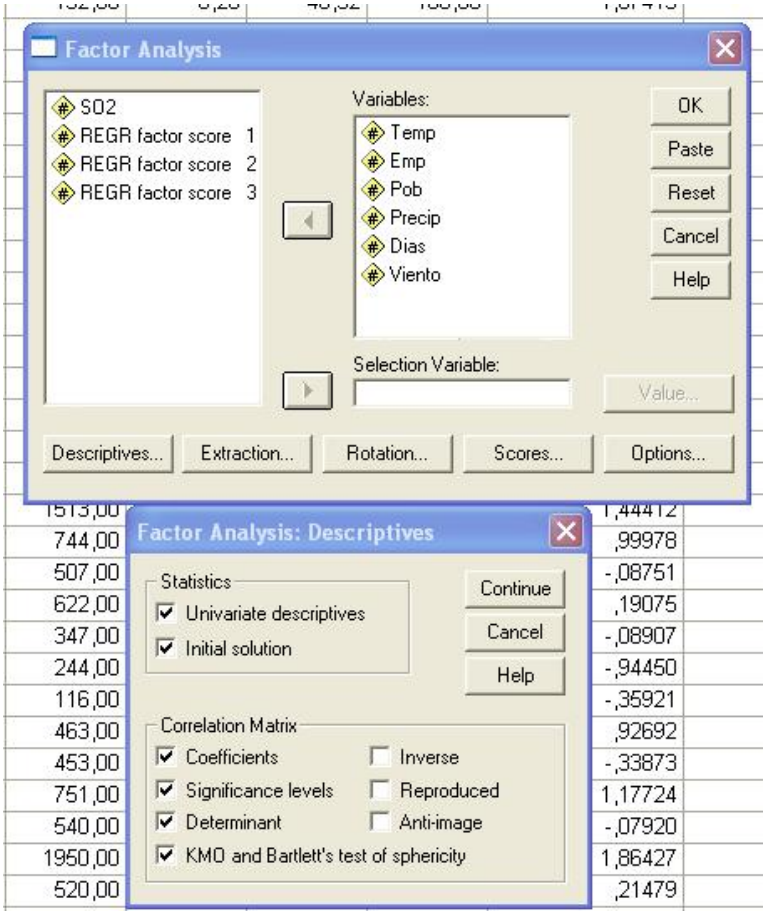
Hacemos un análisis de regresión de la variable SO_2 sobre los tres factores: claramente la cantidad de SO_2 se explica mediante el primer componente de *calidad de vida* (relacionado con el entorno humano y el clima) que cuando empeora aumenta, a su vez, la contaminación.

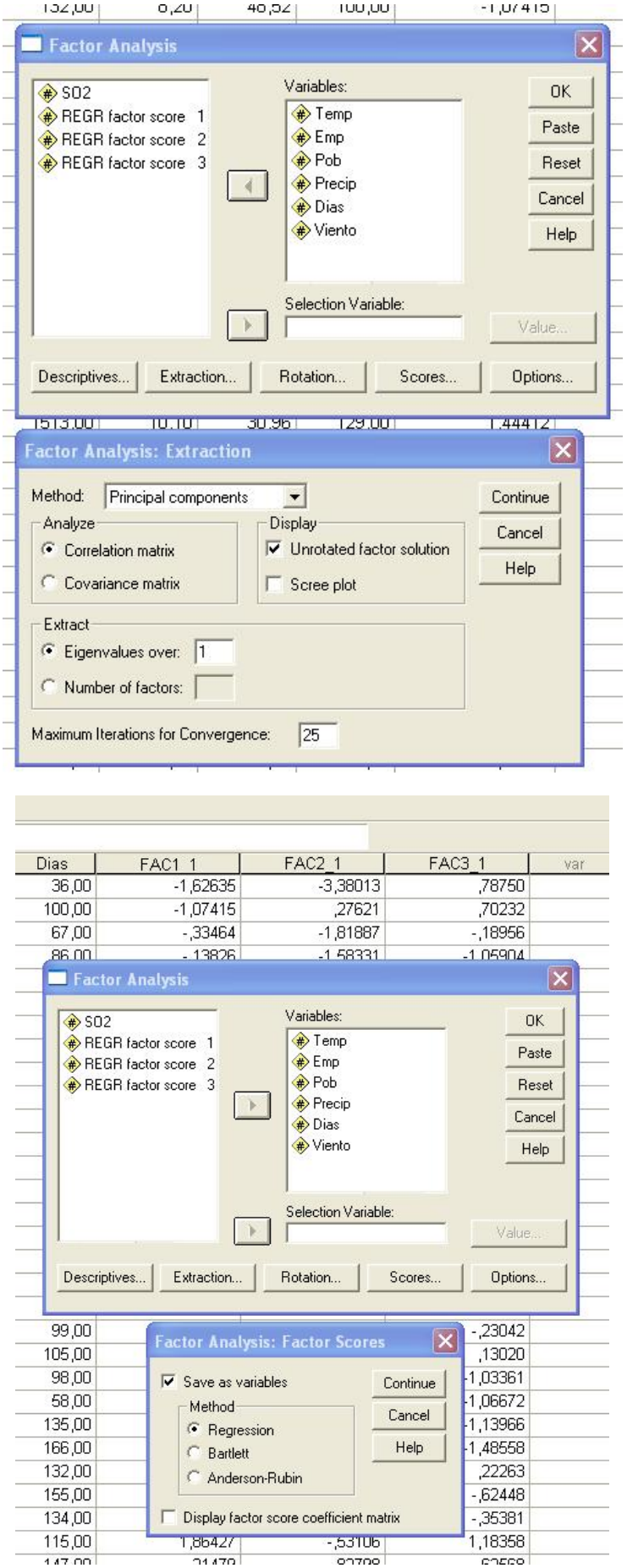
Análisis de Componentes Principales (con SPSS)

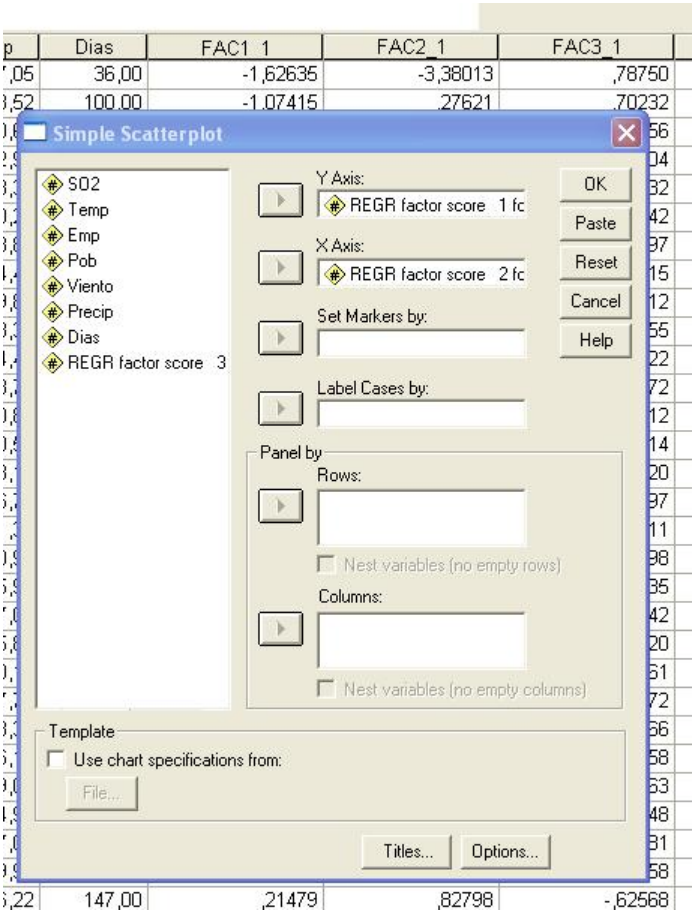
El objetivo del Análisis de Componentes Principales es identificar a partir de un conjunto de *p* variables, otro conjunto de *k* (*k*<*p*) variables no directamente observables, denominadas factores, tal que:

- *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.




Factor Analysis
SO2
REGR factor score 1
REGR factor score 2
REGR factor score 3
Variables:
Temp
Emp
Pob
Precip
Dias
Viento
OK
Paste
Reset
Cancel
Help
Selection Variable:
Value...
Descriptives...
Extraction...
Rotation...
Scores...
Options...
Factor Analysis: Extraction
Method:
Principal components
Continue
Cancel
Help
Analyze
Correlation matrix
Covariance matrix
Display
Unrotated factor solution
Scree plot
Extract
Eigenvalues over:
1
Number of factors:
Maximum Iterations for Convergence:
25
Dias
FAC1_1
FAC2_1
FAC3_1
var
36,00 -1,62635 -3,38013 ,78750
100,00 -1,07415 ,27621 ,70232
67,00 -,33464 -1,81887 -,18956
Factor Analysis
SO2
REGR factor score 1
REGR factor score 2
REGR factor score 3
Variables:
Temp
Emp
Pob
Precip
Dias
Viento
OK
Paste
Reset
Cancel
Help
Selection Variable:
Value...
Descriptives...
Extraction...
Rotation...
Scores...
Options...
99,00 -,23042
105,00 ,13020
98,00 -1,03361
58,00 -1,06672
135,00 -1,13966
166,00 -1,48558
132,00 ,22263
155,00 -,62448
134,00 -,35381
115,00 1,86427 -,53106 1,18358
Factor Analysis: Factor Scores
Save as variables
Continue
Cancel
Help
Method
Regression
Bartlett
Anderson-Rubin
Display factor score coefficient matrix



Se obtienen las medias y desviaciones estándar de cada variable:

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
Viento	9,4439	1,42864	41
Precip	36,7690	11,77155	41
Dias	113,9024	26,50642	41

Se calcula la matriz de correlaciones con la significación de cada componente:

Matriz de correlaciones(a)							
		Temp	**Emp**	**Pob**	**Viento**	**Precip**	**Dias**
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
	Viento	-,350	,238	,213	1,000	-,013	,164
	Precip	,386	-,032	-,026	-,013	1,000	,496
	Dias	-,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
	Viento	,012	,067	,091		,468	,153
	Precip	,006	,420	,436	,468		,000
	Dias	,002	,206	,397	,153	,000	
a Determinante = ,014							

Para que se pueda realizar el PCA, es necesario que las variables presenten factores comunes. Es decir, que estén muy correlacionadas entre sí. Los coeficientes de la matriz de correlaciones deben ser grandes en valor absoluto.

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 identidad.

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 si las variables están incorrelacionadas).
El test de Barlett realiza el contraste:

H_0: $|R| = 1$
H_1: $|R| \neq 1$

El determinante de la matriz da una idea de la correlación generalizada entre todas las variables.
Se basa el test en la distribución chi cuadrado donde valores altos llevan a rechazar H_0, así, la prueba de esfericidad de Bartlett contrasta si la matriz de correlaciones es una matriz identidad, que indicaría que el modelo factorial es inadecuado. Por otro lado, la medida de la adecuación muestral de Kaiser-Meyer-Olkin contrasta si las correlaciones parciales entre las variables son pequeñas:

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

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

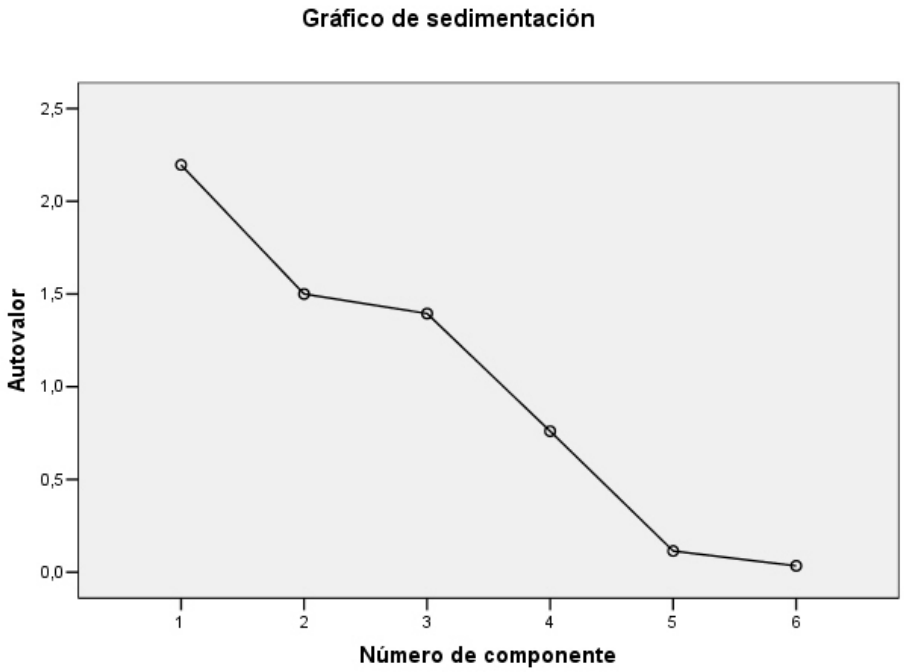
Equivale a la suma de la fila *j*-ésima de la matriz factorial. Sería igual a 0 si los factores comunes no explicaran nada la variabilidad de una variable, y sería igual a 1 si quedase totalmente explicada.

Comunalidades		
	Inicial	**Extracción**
Temp	1,000	,892
Emp	1,000	,968
Pob	1,000	,979
Viento	1,000	,424
Precip	1,000	,941
Dias	1,000	,888
Método de extracción: Análisis de Componentes principales.		

Varianza total explicada						
Componente	**Autovalores iniciales**			**Sumas de las saturaciones al cuadrado de la extracción**		
	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.						

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 usar el criterio de Kaiser: consiste en conservar aquellos factores cuyo autovalor asociado sea mayor que 1.



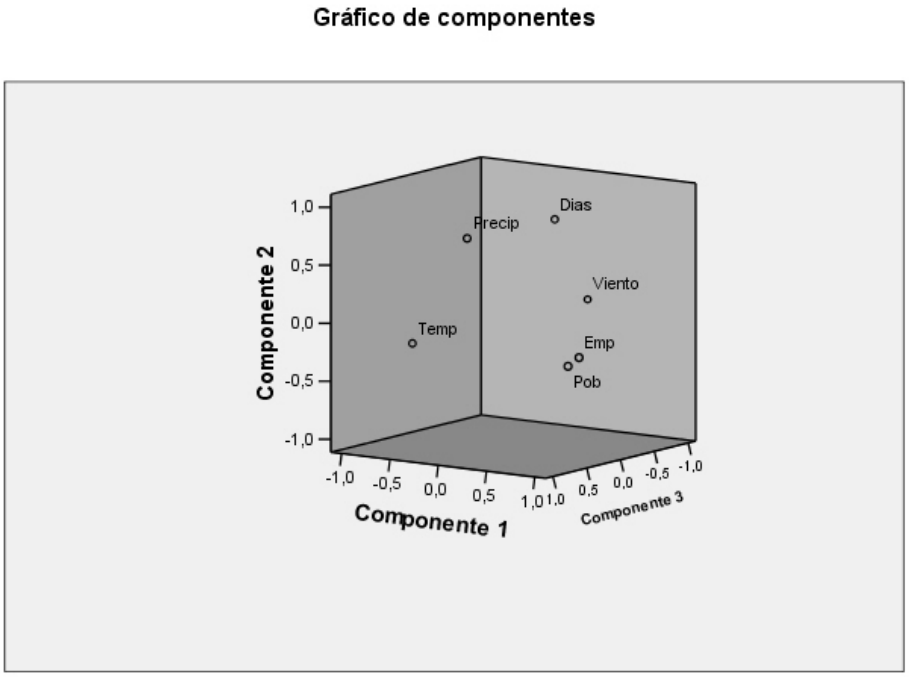
Saturaciones factoriales:

Matriz de componentes(a)			
	Componente		
	1	2	3
Temp	-,489	-,156	,793
Emp	,906	-,206	,322
Pob	,856	-,272	,414
Viento	,524	,160	-,351
Precip	-,060	,763	,596
Dias	,353	,867	-,110
Método de extracción: Análisis de componentes principales.			
a 3 componentes extraídos			

Matriz de correlaciones estimada a partir de la solución factorial. También se muestran las correlaciones residuales (la diferencia entre la correlación observada y la reproducida).

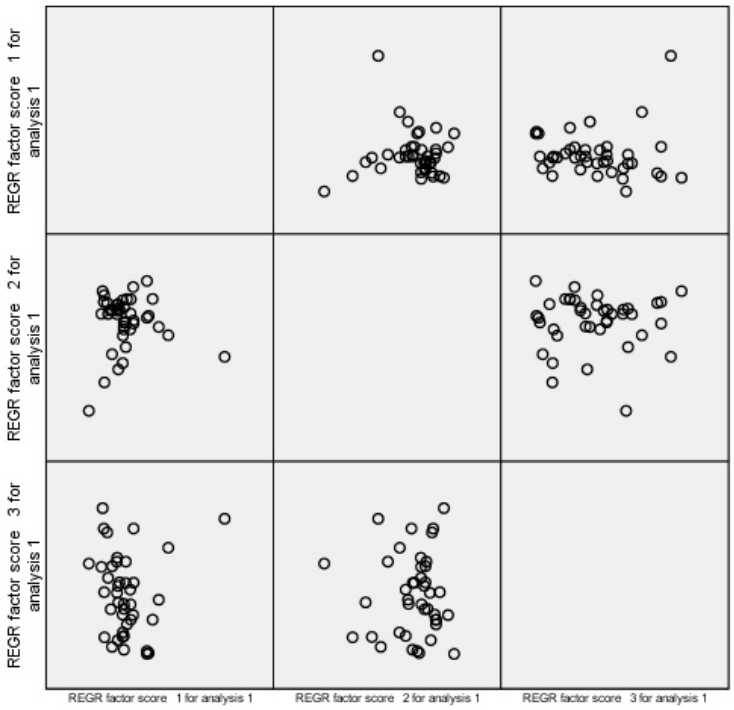
Correlaciones reproducidas							
		Temp	**Emp**	**Pob**	**Viento**	**Precip**	**Dias**
Correlación reproducida	**Temp**	,892(b)	-,155	-,048	-,560	,383	-,395
	Emp	-,155	,968(b)	,965	,329	-,020	,106
	Pob	-,048	,965	,979(b)	,260	-,013	,020
	Viento	-,560	,329	,260	,424(b)	-,119	,362
	Precip	,383	-,020	-,013	-,119	,941(b)	,574
	Dias	-,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
	Viento	,210	-,091	-,047		,106	-,198
	Precip	,003	-,013	-,013	,106		-,078
	Dias	-,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							

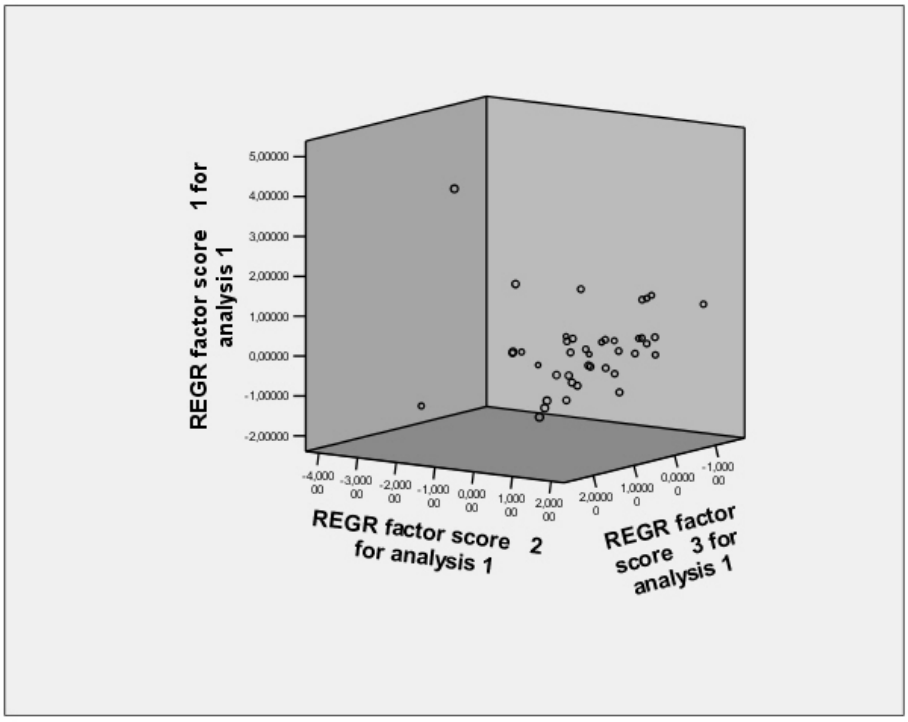
Representación tridimensional de las saturaciones factoriales para los tres primeros factores.



El cálculo de las puntuaciones factoriales consiste en pasar de la matriz original con las variables x_1, ..., x_p a la de los valores según los *k* factores

Estas puntuaciones factoriales se pueden guardar y utilizar en análisis posteriores como técnicas de regresión múltiple o en análisis de cluster.





Regresión de la variable SO_2 frente a los tres factores

Resumen del modelo				
Modelo	**R**	**R cuadrado**	**R cuadrado corregida**	**Error típ. de la estimación**
1	,647(a)	,418	,371	18,61510
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				

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						

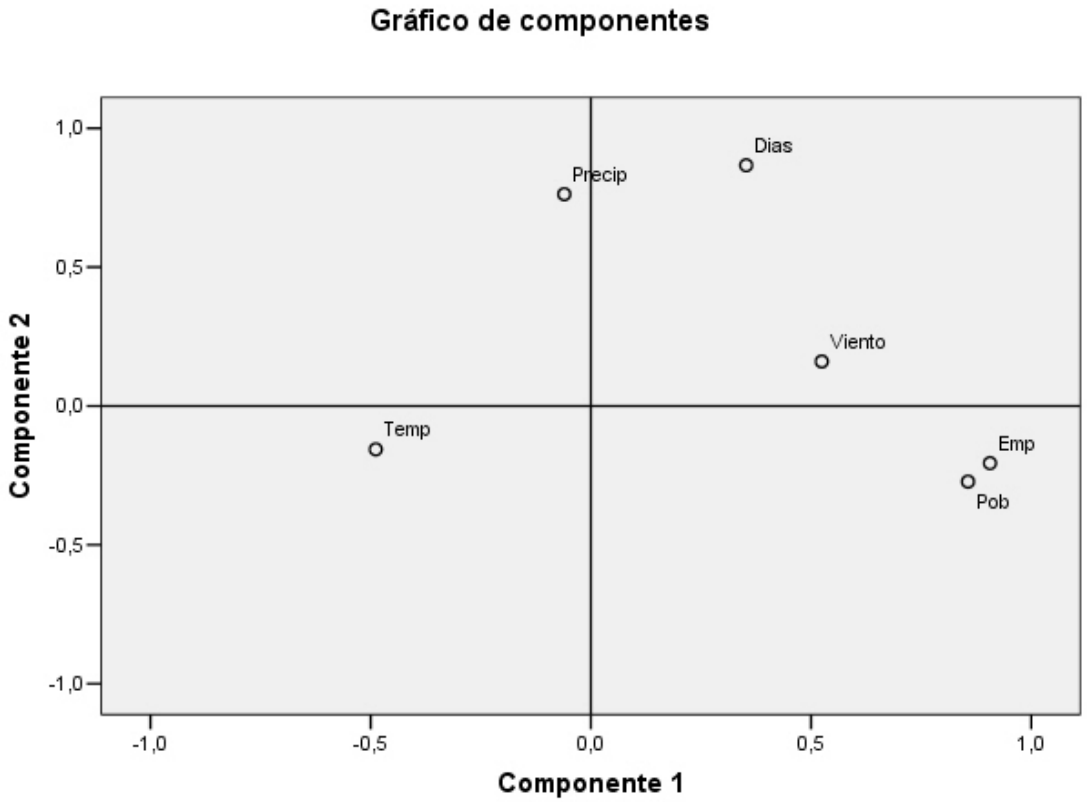
Análisis con dos factores

Matriz de componentes(a)		
	Componente	
	1	2
Temp	-,489	-,156
Emp	,906	-,206
Pob	,856	-,272
Viento	,524	,160
Precip	-,060	,763
Dias	,353	,867
Método de extracción: Análisis de componentes principales.		
a 2 componentes extraídos		

Comunalidades	
	Extracción
Temp	,263
Emp	,864
Pob	,807
Viento	,301
Precip	,586
Dias	,876
Método de extracción: Análisis de Componentes principales.	

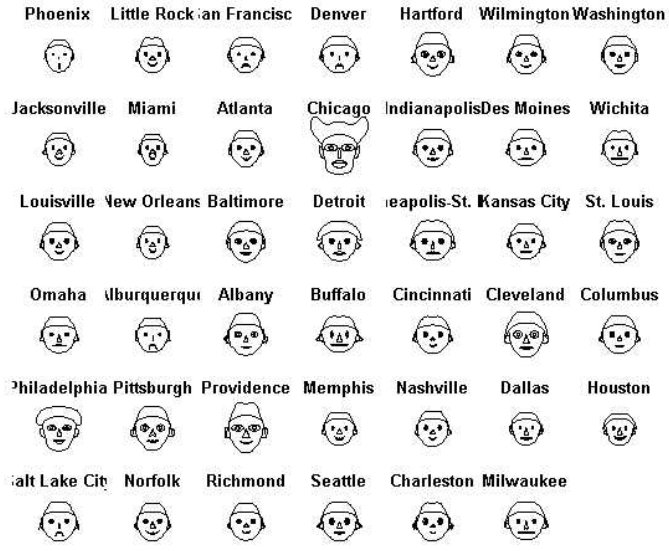
Varianza total explicada			
Componente	Sumas de las saturaciones al cuadrado de la extracción		
	Total	% de la varianza	% acumulado
1	2,196	36,603	36,603
2	1,500	24,999	61,602
Método de extracción: Análisis de Componentes principales.			

Correlaciones reproducidas							
		Temp	**Emp**	**Pob**	**Viento**	**Precip**	**Dias**
Correlación reproducida	**Temp**	,263(b)	-,411	-,376	-,281	-,090	-,308
	Emp	-,411	,864(b)	,832	,442	-,212	,141
	Pob	-,376	,832	,807(b)	,405	-,260	,066
	Viento	-,281	,442	,405	,301(b)	,090	,324
	Precip	-,090	-,212	-,260	,090	,586(b)	,640
	Dias	-,308	,141	,066	,324	,640	,876(b)
Residual(a)	**Temp**		,221	,313	-,069	,476	-,123
	Emp	,221		,123	-,204	,179	-,009
	Pob	,313	,123		-,193	,233	-,024
	Viento	-,069	-,204	-,193		-,103	-,160
	Precip	,476	,179	,233	-,103		-,144
	Dias	-,123	-,009	-,024	-,160	-,144	
Método de extracción: Análisis de Componentes principales.							
a Los residuos se calculan entre las correlaciones observadas y reproducidas. Hay 13 (86,0%) residuales no redundantes con valores absolutos mayores que 0,05.							
b Comunalidades reproducidas							

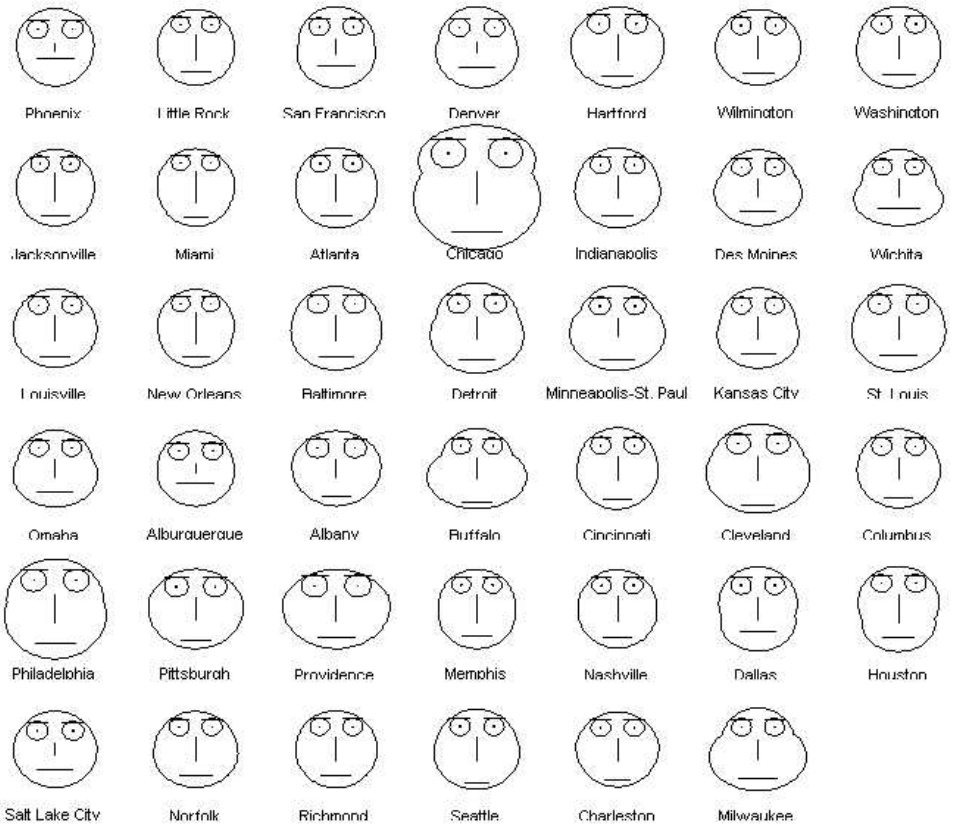


Análisis de Componentes Principales (con R)

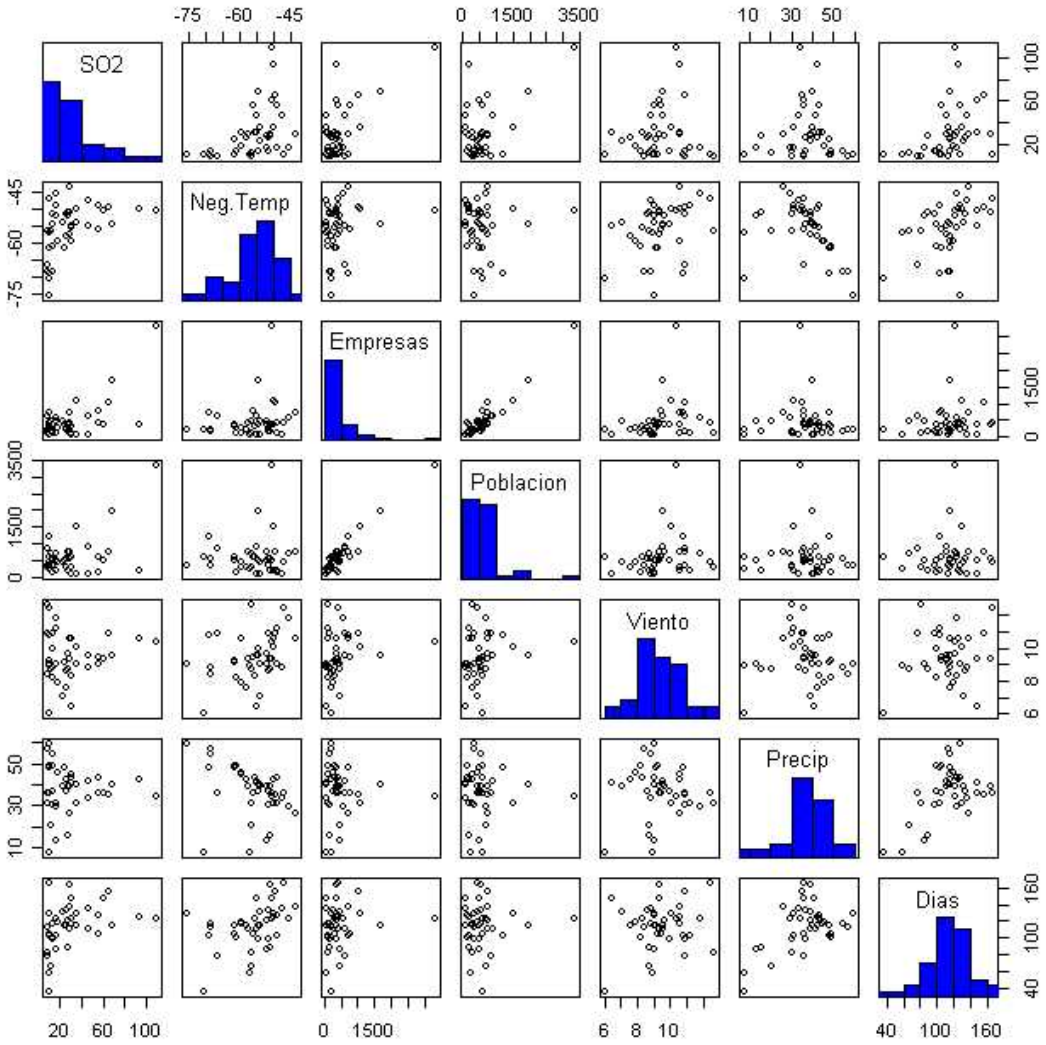
```
# Leo los datos
aire.dat <- read.table("c:\\... \\datPCA.txt",header=T)
attach(aire.dat)

dimnames(aire.dat)[[1]] <- c("Phoenix", "Little Rock", "San Francisco",
"Denver", "Hartford", "Wilmington", "Washington", "Jacksonville", "Miami",
"Atlanta", "Chicago", "Indianapolis", "Des Moines", "Wichita", "Louisville",
"New Orleans", "Baltimore", "Detroit", "Minneapolis-St. Paul", "Kansas City",
"St. Louis", "Omaha", "Alburquerque", "Albany", "Buffalo", "Cincinnati",
"Cleveland", "Columbus", "Philadelphia", "Pittsburgh", "Providence",
"Memphis", "Nashville", "Dallas", "Houston", "Salt Lake City", "Norfolk",
"Richmond", "Seattle", "Charleston", "Milwaukee")

# Hago un análisis descriptivo
summary(aire.dat)

      SO2             Neg.Temp         Empresas         Poblacion
 Min.   :  8.00   Min.   :-75.50   Min.   :  35.0   Min.   :  71.0
 1st Qu.: 13.00   1st Qu.:-59.30   1st Qu.: 181.0   1st Qu.: 299.0
 Median : 26.00   Median :-54.60   Median : 347.0   Median : 515.0
 Mean   : 30.05   Mean   :-55.76   Mean   : 463.1   Mean   : 608.6
 3rd Qu.: 35.00   3rd Qu.:-50.60   3rd Qu.: 462.0   3rd Qu.: 717.0
 Max.   :110.00   Max.   :-43.50   Max.   :3344.0   Max.   :3369.0

     Viento           Precip           Dias
 Min.   : 6.000   Min.   : 7.05   Min.   : 36.0
 1st Qu.: 8.700   1st Qu.:30.96   1st Qu.:103.0
 Median : 9.300   Median :38.74   Median :115.0
 Mean   : 9.444   Mean   :36.77   Mean   :113.9
 3rd Qu.:10.600   3rd Qu.:43.11   3rd Qu.:128.0
 Max.   :12.700   Max.   :59.80   Max.   :166.0

library(TeachingDemos)
faces(aire.dat)
```



faces2(aire.dat,nrows=7)



```
panel.hist <- function(x, ...)
{
          usr <- par("usr"); on.exit(par(usr))
          par(usr = c(usr[1:2], 0, 1.5) )
          h <- hist(x, plot = FALSE)
          breaks <- h$breaks; nB <- length(breaks)
          y <- h$counts; y <- y/max(y)
          rect(breaks[-nB], 0, breaks[-1], y, col="blue", ...)
}
pairs(aire.dat,diag.panel=panel.hist)
```



 # Calculo la matriz de correlaciones

```
cor(aire.dat[,-1])
```

```
              Temp         Emp         Pob      Viento      Precip        Dias
Temp    1.00000000 -0.19004216 -0.06267813 -0.34973963  0.38625342 -0.43024212
Emp    -0.19004216  1.00000000  0.95526935  0.23794683 -0.03241688  0.13182930
Pob    -0.06267813  0.95526935  1.00000000  0.21264375 -0.02611873  0.04208319
Viento -0.34973963  0.23794683  0.21264375  1.00000000 -0.01299438  0.16410559
Precip  0.38625342 -0.03241688 -0.02611873 -0.01299438  1.00000000  0.49609671
Dias   -0.43024212  0.13182930  0.04208319  0.16410559  0.49609671  1.00000000
```

 # Calculo los componentes principales basados en la matriz de correlaciones

```
aire.pc<-princomp(aire.dat[,-1],cor=TRUE)
summary(aire.pc,loadings=TRUE)
```

```
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4     Comp.5
Standard deviation     1.4819456 1.2247218 1.1809526 0.8719099 0.33848287
Proportion of Variance 0.3660271 0.2499906 0.2324415 0.1267045 0.01909511
Cumulative Proportion  0.3660271 0.6160177 0.8484592 0.9751637 0.99425879
                           Comp.6
Standard deviation     0.185599752
Proportion of Variance 0.005741211
Cumulative Proportion  1.000000000

Loadings:
       Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
Temp    0.330 -0.128  0.672 -0.306 -0.558 -0.136
Emp    -0.612 -0.168  0.273  0.137  0.102 -0.703
Pob    -0.578 -0.222  0.350                0.695
Viento -0.354  0.131 -0.297 -0.869 -0.113
Precip         0.623  0.505 -0.171  0.568
Dias   -0.238  0.708         0.311 -0.580
```

 # Es lo mismo que calcular los autovalores y autovectores de S

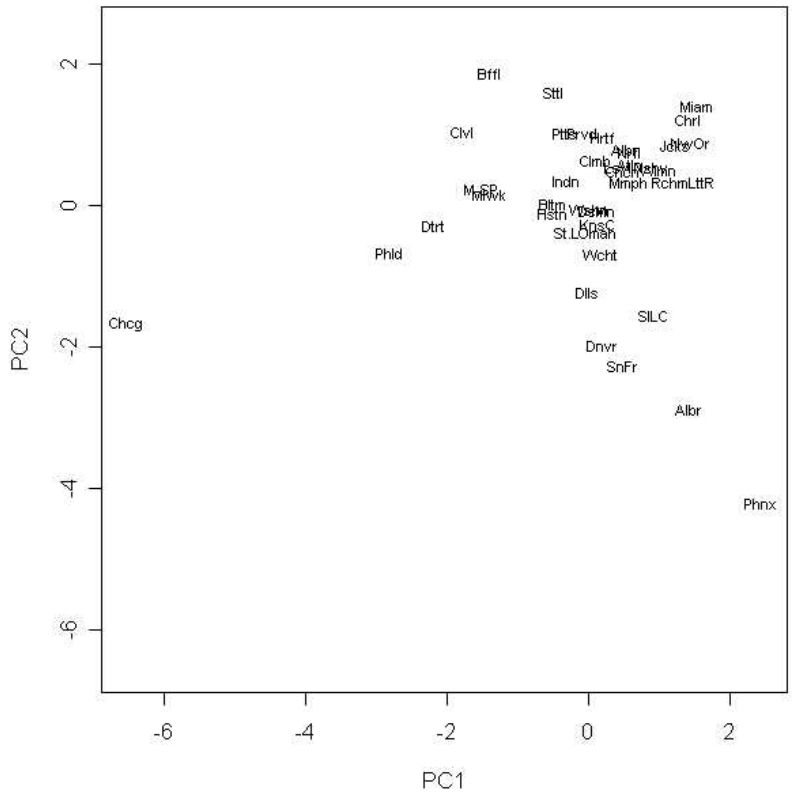
```
S = cor(aire.dat[,-1])
eigen(S)
```

```
$values
[1] 2.19616264 1.49994343 1.39464912 0.76022689 0.11457065 0.03444727

$vectors
            [,1]       [,2]        [,3]        [,4]        [,5]        [,6]
[1,]  0.32964613 -0.1275974  0.67168611 -0.30645728  0.55805638 -0.13618780
[2,] -0.61154243 -0.1680577  0.27288633  0.13684076 -0.10204211 -0.70297051
[3,] -0.57782195 -0.2224533  0.35037413  0.07248126  0.07806551  0.69464131
[4,] -0.35383877  0.1307915 -0.29725334 -0.86942583  0.11326688 -0.02452501
[5,]  0.04080701  0.6228578  0.50456294 -0.17114826 -0.56818342  0.06062222
[6,] -0.23791593  0.7077653 -0.09308852  0.31130693  0.58000387 -0.02196062
```

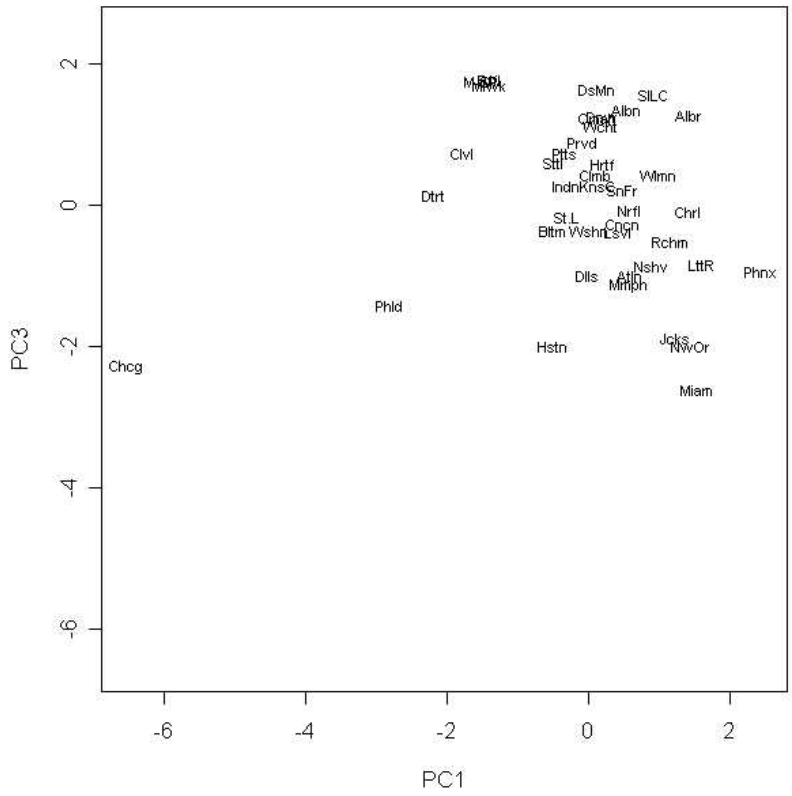
 # Las puntuaciones se obtienen mediante la orden

```
aire.pc$scores[,1:3]
```

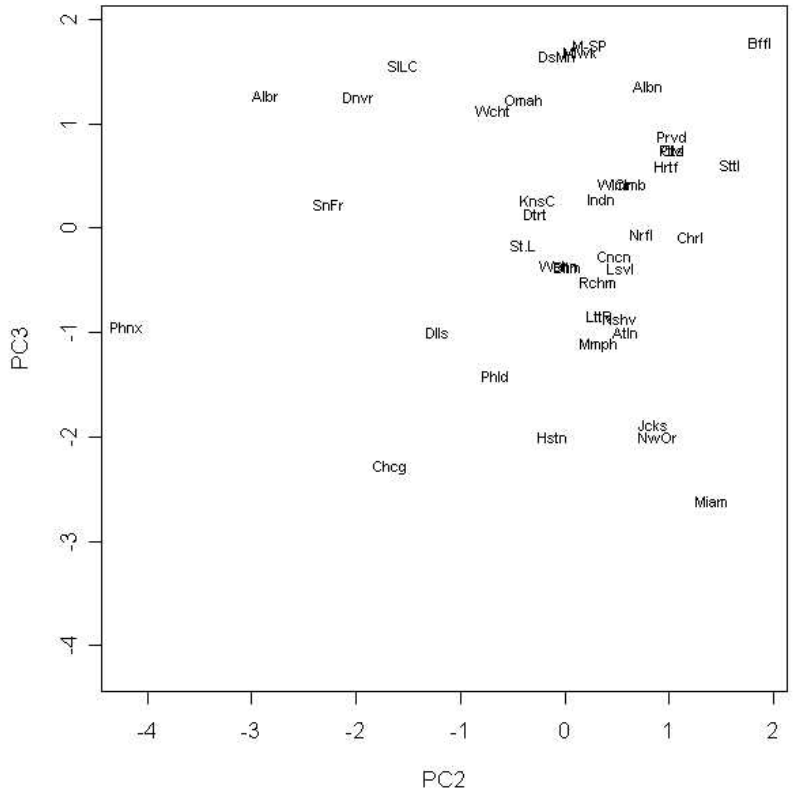
```
par(pty="s")
plot(aire.pc$scores[,1],aire.pc$scores[,2],
ylim=range(aire.pc$scores[,1]),
xlab="PC1",ylab="PC2",type="n",lwd=2)
text(aire.pc$scores[,1],aire.pc$scores[,2],
labels=abbreviate(row.names(aire.dat)),cex=0.7,lwd=2)
```



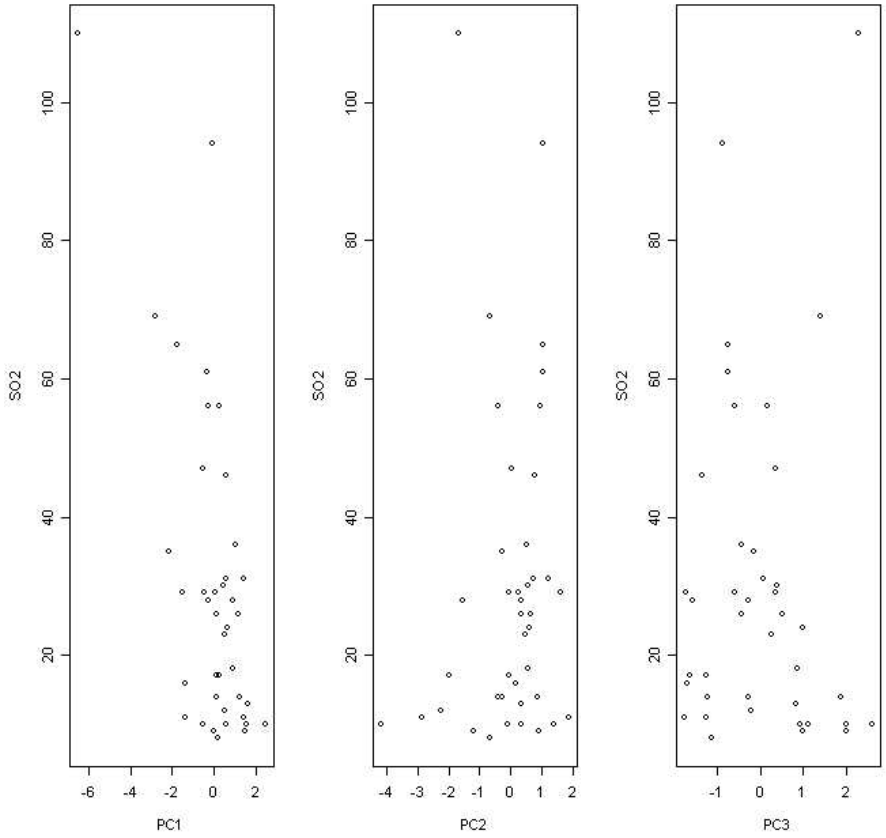
```
par(pty="s")
plot(aire.pc$scores[,1],aire.pc$scores[,3],
ylim=range(aire.pc$scores[,1]),
xlab="PC1",ylab="PC3",type="n",lwd=2)
text(aire.pc$scores[,1],aire.pc$scores[,3],
labels=abbreviate(row.names(aire.dat)),cex=0.7,lwd=2)
```



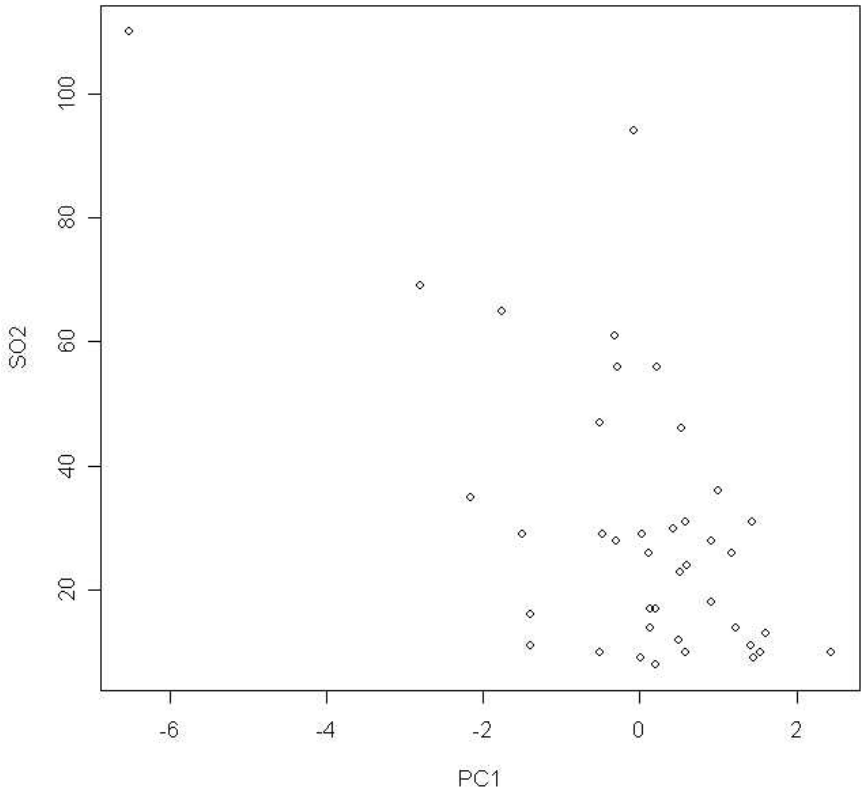
```
par(pty="s")
plot(aire.pc$scores[,2],aire.pc$scores[,3],
ylim=range(aire.pc$scores[,2]),
xlab="PC2",ylab="PC3",type="n",lwd=2)
text(aire.pc$scores[,2],aire.pc$scores[,3],
labels=abbreviate(row.names(aire.dat)),cex=0.7,lwd=2)
```



```
par(mfrow=c(1,3))
plot(aire.pc$scores[,1],SO2,xlab="PC1")
plot(aire.pc$scores[,2],SO2,xlab="PC2")
plot(aire.pc$scores[,3],SO2,xlab="PC3")
```



```
summary(lm(SO2~aire.pc$scores[,1]+aire.pc$scores[,2]+
aire.pc$scores[,3]))

Call:
lm(formula = SO2 ~ aire.pc$scores[, 1] + aire.pc$scores[, 2] +
    aire.pc$scores[, 3])

Residuals:
    Min      1Q  Median      3Q     Max
-36.420 -10.981  -3.184  12.087  61.273

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)           30.049      2.907  10.336 1.85e-12 ***
aire.pc$scores[, 1]   -9.942      1.962  -5.068 1.14e-05 ***
aire.pc$scores[, 2]    2.240      2.374   0.943    0.352
aire.pc$scores[, 3]    0.375      2.462   0.152    0.880
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.62 on 37 degrees of freedom
Multiple R-Squared: 0.4182,     Adjusted R-squared: 0.371
F-statistic: 8.866 on 3 and 37 DF,  p-value: 0.0001473


plot(aire.pc$scores[,1],SO2,xlab="PC1",ylab="SO2")
```



Análisis de Componentes Principales (con SAS)

```
/* Analisis de Componentes Principales */
options ls=80 nodate nonumber;
title 'Analisis de Componentes Principales de datos de contaminacion';
data contamino;
infile 'c:\DondeTrabajo\...\datos.txt';
/* Hay 7 variables: */
input SO2  Temp  Emp  Pob  Viento  Precip  Dias ;
run;

/* Paso un Analisis de Componentes Principales con todas las variables */
proc princomp data=contamino out=pcout;
var SO2  Temp  Emp  Pob  Viento  Precip  Dias;
run;

proc rank data=pcout out=pcout descending;
var SO2;
ranks posn;
data labels;
set pcout;
retain xsys ysys '2';
y=prin1;
x=prin2;
text=put(posn,2.);
keep xsys ysys x y text;
proc gplot data=pcout;
plot prin1*prin2 / annotate=labels;
symbol v=none;
run;

goptions reset=symbol;
proc gplot data=pcout;
plot SO2*(prin1 prin2);
run;
proc corr data=pcout;
var SO2 prin1 prin2;
run;
```

The PRINCOMP Procedure

Observations	41
Variables	7

Simple Statistics

	SO2	Temp	Emp	Pob
Mean	30.04878049	55.76341463	463.0975610	608.6097561
StD	23.47227217	7.22771596	563.4739482	579.1130234

Simple Statistics

	Viento	Precip	Dias
Mean	9.443902439	36.76902439	113.9024390
StD	1.428644249	11.77154977	26.5064189

Correlation Matrix

	SO2	Temp	Emp	Pob	Viento	Precip	Dias
SO2	1.0000	-.4336	0.6448	0.4938	0.0947	0.0543	0.3696
Temp	-.4336	1.0000	-.1900	-.0627	-.3497	0.3863	-.4302
Emp	0.6448	-.1900	1.0000	0.9553	0.2379	-.0324	0.1318
Pob	0.4938	-.0627	0.9553	1.0000	0.2126	-.0261	0.0421
Viento	0.0947	-.3497	0.2379	0.2126	1.0000	-.0130	0.1641
Precip	0.0543	0.3863	-.0324	-.0261	-.0130	1.0000	0.4961
Dias	0.3696	-.4302	0.1318	0.0421	0.1641	0.4961	1.0000

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	2.72811968	1.21578483	0.3897	0.3897
2	1.51233485	0.11736187	0.2160	0.6058
3	1.39497299	0.50298170	0.1993	0.8051
4	0.89199129	0.54521262	0.1274	0.9325
5	0.34677866	0.24649107	0.0495	0.9820
6	0.10028759	0.07477267	0.0143	0.9964
7	0.02551493		0.0036	1.0000

Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7
SO2	0.489699	0.084576	0.014350	-.404210	0.730394	0.183346	0.149529
Temp	-.315371	-.088638	0.677136	0.185228	0.162465	0.610661	-.023664
Emp	0.541169	-.225881	0.267159	0.026272	-.164101	-.042734	-.745181
Pob	0.487588	-.282004	0.344838	0.113404	-.349105	-.087863	0.649126
Viento	0.249875	0.055471	-.311265	0.861901	0.268255	0.150054	0.015765
Precip	0.000187	0.625879	0.492036	0.183937	0.160599	-.553574	-.010315
Dias	0.260179	0.677967	-.109579	-.109761	-.439970	0.504947	0.008217

The CORR Procedure

3 Variables: SO2 Prin1 Prin2

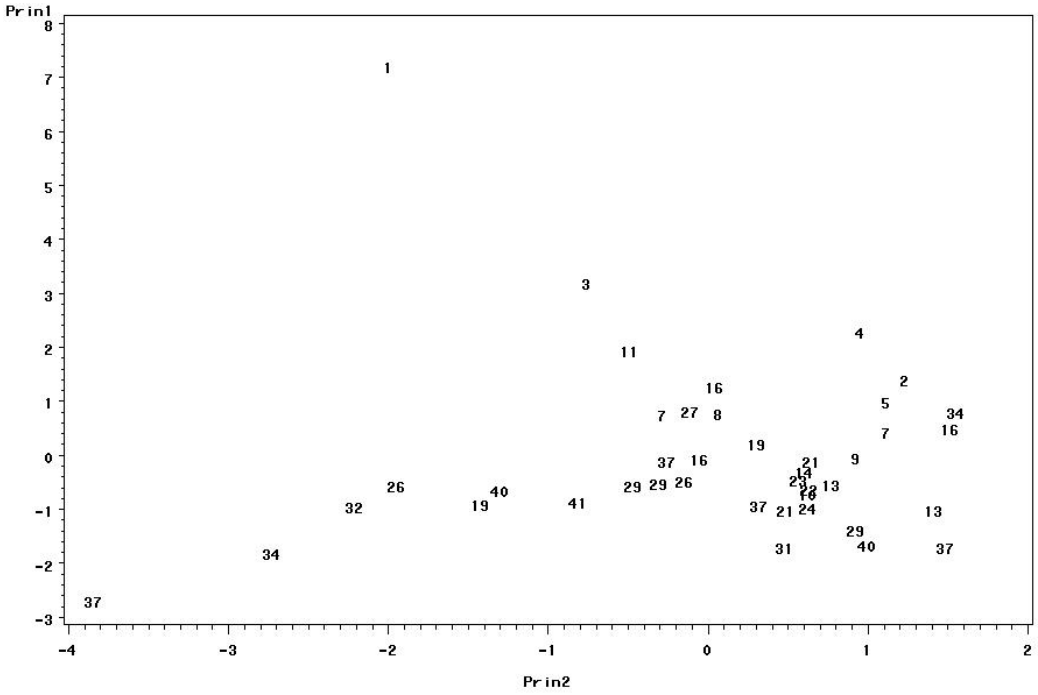
Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
SO2	41	30.04878	23.47227	1232	8.00000	110.00000
Prin1	41	0	1.65170	0	-2.68265	7.23097
Prin2	41	0	1.22977	0	-3.84369	1.54968

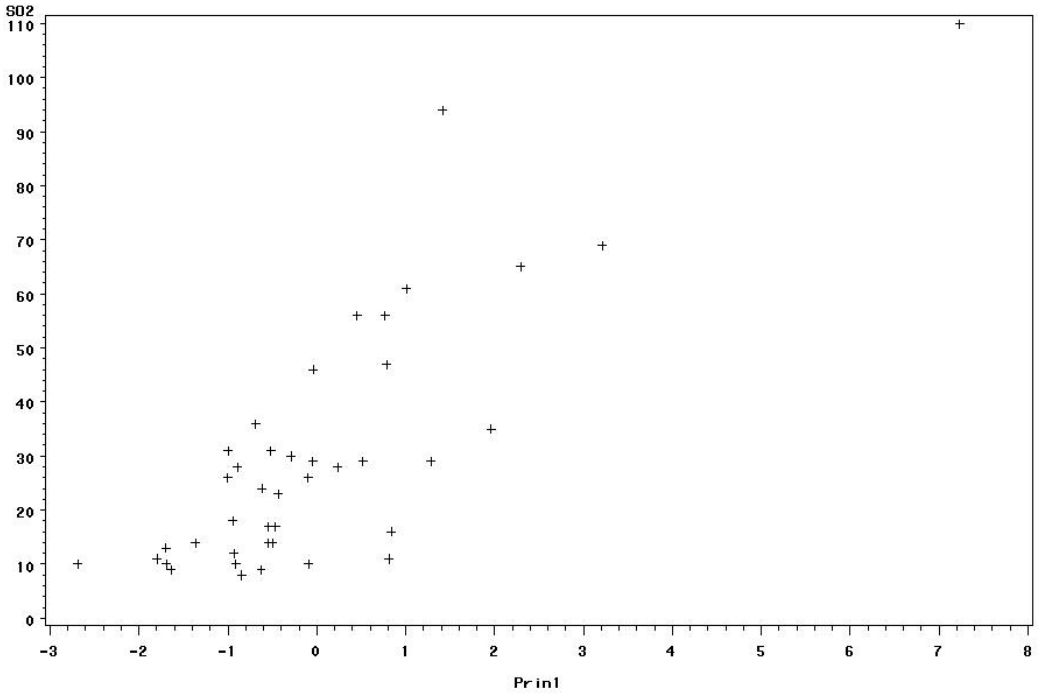
Pearson Correlation Coefficients, N = 41
Prob > |r| under H0: Rho=0

	SO2	Prin1	Prin2
SO2	1.00000	0.80884	0.10401
		<.0001	0.5175
Prin1	0.80884	1.00000	0.00000
	<.0001		1.0000
Prin2	0.10401	0.00000	1.00000
	0.5175	1.0000	

Analisis de Componentes Principales de datos de contaminacion



Analisis de Componentes Principales de datos de contaminacion



Analisis de Componentes Principales de datos de contaminacion