Modelos Estadísticos para la Regresión y la Clasificación

Clase 10: Análisis de Componentes Principales

Mathias Bourel

Instituto de Matemática y Estadística Prof. Rafael Laguardia (IMERL)
Facultad de Ingeniería, Universidad de la República, Uruguay

2 de octubre de 2024

Enfoque geométrico

Consideramos la matriz de datos $X \in \mathcal{M}_{n\times p}$ **centrada**, es decir la media de cada columna es 0 y suponemos que todas las variables son continuas. Tenemos entonces una nube de puntos que representa a los n individuos en $\mathbb{R}^p$.

Enfoque geométrico

Consideramos la matriz de datos $X \in \mathcal{M}_{n\times p}$ **centrada**, es decir la media de cada columna es 0 y suponemos que todas las variables son continuas. Tenemos entonces una nube de puntos que representa a los n individuos en $\mathbb{R}^p$.

Queremos encontrar un subespacio de dimensión menor que p (la cantidad de variables) que represente de manera adecuada los datos. Más precisamente querremos encontrar un subespacio de dimensión menor que p tal que cuando proyectamos los individuos sobre él, la estructura de la nube de puntos se distorciona lo menos posible.

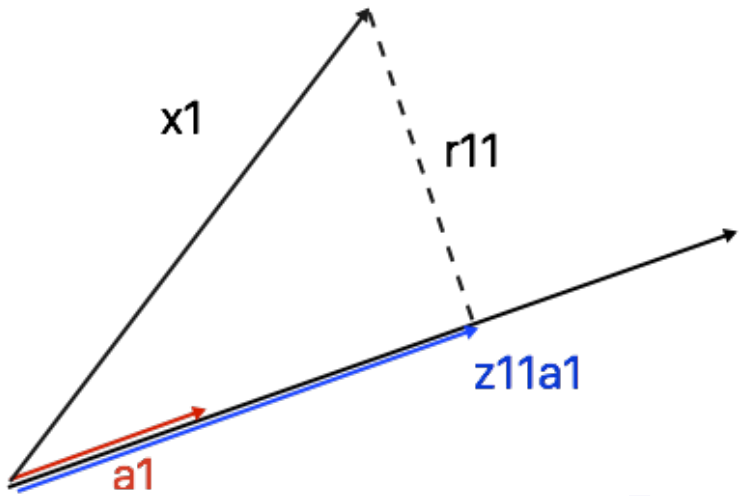
Enfoque geométrico

Consideramos la matriz de datos $X \in \mathcal{M}_{n\times p}$ **centrada**, es decir la media de cada columna es 0 y suponemos que todas las variables son continuas. Tenemos entonces una nube de puntos que representa a los n individuos en $\mathbb{R}^p$.

Queremos encontrar un subespacio de dimensión menor que p (la cantidad de variables) que represente de manera adecuada los datos. Más precisamente querremos encontrar un subespacio de dimensión menor que p tal que cuando proyectamos los individuos sobre él, la estructura de la nube de puntos se distorciona lo menos posible.

Consideremos una recta por el origen (subespacio de dimensión 1) generada por un vector $a_1 \in \mathbb{R}^p$ unitario. Si consideramos un individuo $\mathbf{x}_1$ su proyección sobre el subespacio generado por a_1 es

$$z_{11}a_1 = \frac{\mathbf{x}_1' a_1}{||a_1||^2} a_1 = \mathbf{x}_1' a_1 a_1 = a_1' \mathbf{x}_1 a_1$$



Enfoque geométrico

Llamaremos a r_{11} la distancia del individuo $\mathbf{x}_1$ al eje generado por a_1. Lo que hacemos por el individuo 1 lo podemos hacer para los n individuos y considerar $\mathbf{x}_1, \ldots, \mathbf{x}_n$ y $r_{11} \ldots, r_{n1}$.

Enfoque geométrico

Llamaremos a r_{11} la distancia del individuo $\mathbf{x}_1$ al eje generado por a_1. Lo que hacemos por el individuo 1 lo podemos hacer para los n individuos y considerar $\mathbf{x}_1, \ldots, \mathbf{x}_n$ y $r_{11} \ldots, r_{n1}$.

El objetivo consiste en minimizar la suma de las distancias $\sum_{i=1}^{n} r_{i1}^2 = \sum_{i=1}^{n} r_{i1}^2 = \sum_{i=1}^{n} ||\mathbf{x}_i - z_{i1} a_1||^2$,

es decir buscar la recta que minimiza la suma de los cuadrados de las proyecciones ortogonales de los puntos sobre ella.

Enfoque geométrico

Llamaremos a r_{11} la distancia del individuo $\mathbf{x}_1$ al eje generado por a_1. Lo que hacemos por el individuo 1 lo podemos hacer para los n individuos y considerar $\mathbf{x}_1, \ldots, \mathbf{x}_n$ y $r_{11} \ldots, r_{n1}$.

El objetivo consiste en minimizar la suma de las distancias $\sum_{i=1}^{n} r_{i1}^2 = \sum_{i=1}^{n} r_{i1}^2 = \sum_{i=1}^{n} ||\mathbf{x}_i - z_{i1} a_1||^2$,

es decir buscar la recta que minimiza la suma de los cuadrados de las proyecciones ortogonales de los puntos sobre ella.

Por el teorema de Pitágoras tenemos para cada individuo que

$$\mathbf{x'}_i \mathbf{x}_i = z_{i1}^2 + r_{i1}^2$$

Enfoque geométrico

Llamaremos a r_{11} la distancia del individuo $\mathbf{x}_1$ al eje generado por a_1. Lo que hacemos por el individuo 1 lo podemos hacer para los n individuos y considerar $\mathbf{x}_1, \ldots, \mathbf{x}_n$ y $r_{11} \ldots, r_{n1}$.

El objetivo consiste en minimizar la suma de las distancias $\sum_{i=1}^{n} r_{i1}^2 = \sum_{i=1}^{n} r_{i1}^2 = \sum_{i=1}^{n} ||\mathbf{x}_i - z_{i1} a_1||^2$, es decir buscar la recta que minimiza la suma de los cuadrados de las proyecciones ortogonales de los puntos sobre ella.

Por el teorema de Pitágoras tenemos para cada individuo que

$$\mathbf{x'}_i \mathbf{x}_i = z_{i1}^2 + r_{i1}^2$$

Entonces sumando se tiene

$$\sum_{i=1}^{n} \mathbf{x'}_i \mathbf{x}_i = \sum_{i=1}^{n} z_{i1}^2 + \sum_{i=1}^{n} r_{i1}^2$$

Como el término $\sum_{i=1}^{n} \mathbf{x'}_i \mathbf{x}_i$ es constante, minimizar $\sum_{i=1}^{n} r_{i1}^2$ equivale a maximizar $\sum_{i=1}^{n} z_{i1}^2$ que no es otra cosa que la varianza muestral **de los datos proyectados** dado que los datos son centrados.

Enfoque geométrico

Llamaremos a r_{11} la distancia del individuo $\mathbf{x}_1$ al eje generado por a_1. Lo que hacemos por el individuo 1 lo podemos hacer para los n individuos y considerar $\mathbf{x}_1, \ldots, \mathbf{x}_n$ y $r_{11} \ldots, r_{n1}$.

El objetivo consiste en minimizar la suma de las distancias $\sum_{i=1}^{n} r_{i1}^2 = \sum_{i=1}^{n} r_{i1}^2 = \sum_{i=1}^{n} ||\mathbf{x}_i - z_{i1} a_1||^2$, es decir buscar la recta que minimiza la suma de los cuadrados de las proyecciones ortogonales de los puntos sobre ella.

Por el teorema de Pitágoras tenemos para cada individuo que

$$\mathbf{x'}_i \mathbf{x}_i = z_{i1}^2 + r_{i1}^2$$

Entonces sumando se tiene

$$\sum_{i=1}^{n} \mathbf{x'}_i \mathbf{x}_i = \sum_{i=1}^{n} z_{i1}^2 + \sum_{i=1}^{n} r_{i1}^2$$

Como el término $\sum_{i=1}^{n} \mathbf{x'}_i \mathbf{x}_i$ es constante, minimizar $\sum_{i=1}^{n} r_{i1}^2$ equivale a maximizar $\sum_{i=1}^{n} z_{i1}^2$ que no es otra cosa que la varianza muestral **de los datos proyectados** dado que los datos son centrados.

En efecto

$$\sum_{i=1}^{n} z_{i1} = \sum_{i=1}^{n} a_1' \mathbf{x}_i = a_1' \left(\sum_{i=1}^{n} \mathbf{x}_i \right) = a_1' \overline{\mathbf{x}} = 0$$

Objetivos

- Reducir el número de variables sin perder (demasiada) información: al proyectar los n individuos sobre un espacio de dimensión l con $l < p$ tal que la dispersión en el espacio proyectado sea máxima.
- Simplificar la descripción del conjunto de datos. Analizar la estructura y relación de las observaciones y de las variables.
- Las componentes principales deben tener varianza máxima (mayor información relacionado con mayor variabilidad).

Para eso:

- Cada componente principal z es una combinación lineal de las variables originales.

$$\text{Probabilidad}: \ z_j = a_{j1}x_1 + a_{j2}x_2 + \cdots + a_{jp}x_p = \quad \forall\, j = 1, \ldots, l, \ l < p$$

$$\text{Estadística}: \ z_j = a_{j1}x_1 + a_{j2}x_2 + \cdots + a_{jp}x_p = \mathbf{X}a_j = \begin{pmatrix} \mathbf{x_1'}a_j \\ \mathbf{x_2'}a_j \\ \vdots \\ \mathbf{x_n'}a_j \end{pmatrix} = \begin{pmatrix} z_{1j} \\ z_{2j} \\ \vdots \\ z_{nj} \end{pmatrix}$$

- Las componentes principales son no correlacionadas dos a dos, y de esta manera eliminamos información repetida:

$$x_1, \ldots, x_p \text{ correladas} \rightarrow z_1, \ldots, z_l \textbf{ incorreladas}$$

Cálculo de las componentes

1. Vamos a imponer que $||a_j'|| = 1 \ \ \forall j = 1, \ldots, p$
2. Vamos a buscar a_1 tal que z_1 tenga la mayor varianza y $||a_1|| = 1$.
3. Vamos a buscar a_2 tal que z_2 sea incorrelada con z_1, con varianza menor que z_1 y $||a_2|| = 1$.
4. Vamos a buscar a_3 tal que z_3 sea incorrelada con z_1 y z_2, con varianza menor que z_1 y z_2 y $||a_2|| = 1$.
5. ...

Teorema de la esfera unidad

Sea Σ la matriz de covarianzas de $\mathbf{X}$. Habitualmente se usa la matriz de correlaciones ya que se estandariza los datos (cada columna tiene media cero y desvío 1).
Como las variables originales tienen media cero entonces el vector $z_1 = \mathbf{X}a_1$ tiene también media cero y su varianza es $Var(z_1) = \frac{1}{n} z_1' z_1 = \frac{1}{n} a_1' \mathbf{X}'\mathbf{X} a_1 = a_1' \Sigma a_1$. Para maximizar $Var(z_1)$ de manera que $||a_1|| = 1$. Para ello vamos a usar el teorema de la esfera unidad.

Cambio de notación: Veremos el teorema de la esfera unidad en general para el caso de una matriz $A \in \mathcal{M}_{p\times p}$.

Teorema de la esfera unidad

Si A es una matriz real simétrica y $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ a sus valores propios y sea u_1 vector propio asociado a λ_1 tal que $||u_1|| = 1$, entonces $\max\limits_{||x||=1} x'Ax = \lambda_1$ y además $\operatorname{Argmax}\limits_{||x||=1} x'Ax = u_1$

Al ser A diagonalizable en una base ortonormal, tenemos que exite B ortogonal ($B' = B^{-1}$) y D diagonal tal que $A = BDB'$ y por lo tanto

$$x'Ax = x'BDB'x = (B'x)'D(B'x) = y'Dy = \lambda_1 y_1^2 + \lambda_2 y_2^2 + \cdots + \lambda_p y_p^2 \leq \lambda_1(y_1^2 + y_2^2 + \cdots + y_p^2) = \lambda_1$$

Esto es porque al ser $y = B'x$ entonces

$$||y||^2 = ||B'x||^2 = (B'x)'(B'x) = x'BB'x = x'x = ||x||^2$$

Esta cota superior λ_1 se alcanza en u_1. En efecto si $x = u_1$ entonces

$$x'Ax = u_1'Au_1 = u_1'\lambda_1 u_1 = \lambda_1 u_1'u_1 = \lambda_1 ||u_1||^2 = \lambda_1$$

Teorema

Si A es una matriz real simétrica y $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ a sus valores propios y sea u_1 vector propio asociado a λ_1 tal que $||u_1|| = 1$ y u_2 vector propio asociado a λ_2 tal que $||u_2|| = 1$, entonces $\underset{||x||=1,\, x \perp u_1}{\text{máx}} x'Ax = \lambda_2$ y además $\underset{||x||=1,,x \perp u_1}{\text{Argmax}} x'Ax = u_2$

Como en la demostración anterior hacemos el cambio $y = B'x$ obteniendose que $x'Ax = y'Dy$ y la condición $||x|| = 1$ se transforma en $||y|| = 1$. La condición $x \perp u_1$ implica que $y_1 = 0$. En efecto $x \perp u_1$ equivale a $By \perp u_1$ o sea $(By)'u_1 = 0$, es decir $y'B'u_1 = 0$ pero $B'u_1 = (1, 0, \ldots, 0)'$ entonces $y \perp B'u_1$ son aquellos y tales que $y_1 = 0$. Por lo que

$$\underset{||x||=1,\, x \perp u_1}{\text{máx}} x'Ax = \underset{||y||=1,\, y_1=0}{\text{máx}} y'Dy = \underset{||y||=1,\, y_1=0}{\text{máx}} \lambda_2 y_2^2 + \cdots + \lambda_p y_p^2 \leq \lambda_2(y_2^2 + \ldots y_p^2) = \lambda_2$$

Además la cota superior λ_2 se alcanza en u_2 ya que $||u_2|| = 1$, $u_2 \perp u_1$ y $u_2'Au_2 = u_2'\lambda_2 u_2 = \lambda_2$

(CUIDADO: cambio de notación!!)
Repetimos este procedimiento p veces, obteniendo los vectores $a_1 = u_1, a_2 = u_2, \ldots, a_p = u_p$ y se considera la matriz ortogonal $A = \begin{pmatrix} | & | & & | \\ a_1 & a_2 & \ldots & a_p \\ | & | & & | \end{pmatrix}$

Como $z_1 = \mathbf{X}a_1, \quad z_p = \mathbf{X}a_p$, entonces:

Relación entre las viejas y las nuevas variables

- Observar que se puede escribir (poniendo las caracteristicas en filas):

$$\begin{pmatrix} - & z_1 & - \\ - & z_2 & - \\ & \vdots & \\ - & z_p & - \end{pmatrix} = \begin{pmatrix} a_{11} & a_{21} & \dots & a_{p1} \\ a_{12} & a_{22} & \dots & a_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1p} & a_{2p} & \dots & a_{pp} \end{pmatrix} \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_p' \end{pmatrix}$$

$$Z' = A'\mathbf{X}'$$

- O si no:

$$\begin{pmatrix} | & | & & | \\ z_1 & z_2 & \dots & z_p \\ | & | & & | \end{pmatrix} = \begin{pmatrix} | & | & & | \\ x_1 & x_2 & \dots & x_p \\ | & | & & | \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \dots & a_{pp} \end{pmatrix}$$

$$Z = \mathbf{X}A$$

A las columnas de Z se le llaman *componentes principales* de $\mathbf{X}$.

- Como $Cov(z_j, x_i) = Cov(z_j, \sum_{k=1}^{p} a_{ik} z_k) = a_{ij} Var(z_j) = \lambda_j a_{ij}$

entonces la correlación, si X está estandarizada, es:

$$Cor(z_j, x_i) = \frac{\lambda_j a_{ij}}{\sqrt{\lambda_j}} = \sqrt{\lambda_j} a_{ij} = d_{ij}$$

- Las componentes principales proveen una representación en pocas dimensiones de la nube de puntos, tanto de los individuos que de las variables.
- Se obtienen dos tipos de representaciones gráficas complementarias, que pueden ser condensadas en una sola eventualmente (biplot):
  1. Los planos principales (2D) o los espacios principales (3D). Los planos principales tienen como referencial las componentes principales y se puede ver las coordenadas de los individuos en cada componente, las principales agrupaciones y dispersiones.
  2. Los círculos de correlaciones: allí se puede ver las correlaciones entre las variables originales y las componentes principales normalizadas. Se observan las agrupaciones entre variables y sus proximidades a las componentes principales.
     La coordenada de la variable x_j en la componente $z_k = Xa_k$ es $d_{jk} = r(x_j, z_k)$
- IMPORTANTE: Para interpretar estos gráficos, siempre se debe tener en cuenta que son representaciones y simplificaciones de la realidad (la verdadera nube).

- Cada eje de $\mathbb{R}^p$ representa una de las p variables.
- $\mathbf{z_n'} = \mathbf{x_n'A} = \mathbf{P(x_n)}$ son las coordenadas del individuo $\mathbf{x_n'}$ en el nuevo sistema de referencia determinado por las componentes principales.
- Podemos entonces pensar que "proyectamos" la nube de la población dada por $\mathbf{X}$ sobre un subespacio de dimensión la cantidad de componentes principales que retendremos.
- Como $Var(z_1) = \lambda_1$, $Var(z_2) = \lambda_2$, $\ldots$, $Var(z_p) = \lambda_p$ y son incorreladas:

$$\Sigma_Z = Var(Z) = \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & \lambda_p \end{pmatrix} \underbrace{=}_{Z=\mathbf{X}A} A' Var(\mathbf{X}) A.$$

Entonces:

$$\mathbf{\Sigma} = A \Sigma_Z A'$$

$$\sum_{i=1}^{p} Var(z_i) = \sum_{i=1}^{p} \lambda_i = tr\,(\Sigma_Z) = tr\,(A' \Sigma_X A) = tr\,(\Sigma_X A A') = tr\,(\Sigma_X)$$

Porcentajes de variabilidad

Porcentaje de variabilidad de la variable i:

$$\frac{Var(z_i)}{\sum_{i=1}^{p} Var(z_i)} = \frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i} \qquad \left(\text{con matriz correlaciones } \frac{\lambda_i}{p}\right)$$

Porcentaje de variabilidad de las m primeras variables i:

$$\sum_{j=1}^{m} \frac{\lambda_j}{\sum_{i=1}^{p} \lambda_i} \qquad \text{donde } m < p$$

Nos quedamos con un número mucho menor de componentes que recogen un porcentaje amplio de la variabilidad total (fijada por el usuario). En general no se elige más de 3.

Consideraciones finales

1. En general se suele calcular las componentes principales sobre variables originales estandarizadas (media 0 y varianza 1). Tomo entonces las componentes principales sobre la matriz de correlaciones y se le da la misma importancia a todas las variables.
2. Si las variables $x_1, \ldots, x_p$ ya son incorreladas, entonces no tiene sentido hacer componentes principales. Si se hace se obtiene las mismas variables ordenadas de mayor a menor varianza. Para ver eso se hace el test de esfericidad de Bartlett (package psych) o el indice de Kayser-Meyer-Olkin (KMO).
3. Si Σ tiene un valor propio con multiplicidad mayor que 1 se toma vectores propios ortgonales en el subespacio propio correspondiente.
4. Se conservan en general dos o tres componentes.

Inercia de una nube de puntos

- La inercia de un punto i al punto A se define como

$$I(i, A) = d^2(i, A) \times p_i$$

  donde d es una distancia.
- La inercia de una nube de puntos N al punto A es

$$I(N, A) = \sum_{i=1}^{n} d^2(i, A) \times p_i$$

- La inercia de la nube de puntos N al centro de gravedad G es

$$I = \sum_{i=1}^{n} d^2(i, G) \times p_i = \sum_{i=1}^{n} p_i \sum_{j=1}^{p} (x_{ij} - \bar{x}_j)^2 = \sum_{j=1}^{p} Var(x_j)$$

I es un indicador de la cantidad de información, de la dispersión o de la forma de la nube respecto del centro de gravedad: más I será grande, más la nube será dispersa alrededor del centro de gravedad.

- Si $I = 0$ entonces todos los individuos son idénticos.
- Si la matriz es centrada el centro de gravedad de los individuos es el origen.
- Puesto que $\sum_{j=1}^{p} Var(x_j) = tr(\Sigma)$ si las variables son centradas y reducidas entonces $I = p$.
- Para obtener la inercia en R: sum(diag(cov(X))).

Componentes principales

El nuevo sistema de coordenadas pasa por el centro de gravedad G de la nube de puntos.

Si X es centrada entonces $S = \Sigma$ y si X es centrada y reducida entonces $S = R$.

Queremos buscar el eje a_1 de manera que cuando proyectamos la nube de puntos sobre él la inercia sea máxima.

Sea $z_1 = (c_{11}, c_{21}, \ldots, c_{i1}, \ldots, c_{n1})$ el vector de coordenadas de la proyección ortogonal de los individuos sobre el eje a_1:

$$c_{i1} = \langle \mathbf{x}_i, a_1 \rangle \quad \text{y en definitiva} \quad z_1 = Xa_1$$

Si I_1 es la inercia de la proyección de la nube de puntos:

$$I_1 = \sum_{i=1}^{n} p_i d^2(\mathbf{x}_i, G) = \sum_{i=1}^{n} p_i c_{i1}^2 = z_1' P z_1 = Var(z_1) = a_1' X' P X a_1 = a_1' \Sigma a_1$$

siendo $S = X'PX$ la matriz de varianzas y covarianzas de X con la matriz de pesos P.

Maximizamos I_1 y por lo tanto $a_1' S a_1$ sujeto a $a_1' a_1 = 1$ cuya solución consiste en tomar a_1 un vector propio unitario asociado al mayor valor propio λ_1 de S.

Entonces:

- $I_1 = a_1' S a_1 = a_1' \lambda_1 a_1 = \lambda_1$
- El vector de coordenadas de los n puntos de la nube sobre el primer eje es $z_1 = Xa_1$.
- $\bar{z}_1 = \mathbf{0}$ y $Var(z_1) = z_1' P z_1 = a_1' \Sigma a_1 = \lambda_1$.

Plan

1. Ejemplo completo
   - Proyección de los individuos y variables sobre el plano principal
   - Interpretación de los ejes: contribuciones
   - Calidad de la representación

Ejemplo (con princomp)

```
> #Consumición anual en franco de 8 tipo de comida/bebida (variables) por 8 categorias socio-profesionales.
>
> #Variables: 1 Pan común, 2 Otro tipo de pan, 3 Vino común, 4 Otro tipo de vino,
> # 5 Papas, 6 Vegetales, 7 Uva, 8 Plato preparado
>
> #Individus 1 Productor rural, 2 Asalariado rural, 3 Profesional independiente,
> #4 Ejecutivo superior, 5 Ejecutivo medio, 6 Empleado, 7 Obrero, 8 Desocupado
>
>
> X=t(matrix(c(167,1,163,23,41,8,6,6,162,2,141,12,40,12,4,15,119,6,69,56,39,5,
+ 13,41,87,11,63,111,27,3,18,39,103,5,68,77,32,4,11,30,111,4,72,66,34,6,10,28,130,
+ 3,76,52,43,7,7,16,138,7,117,74,53,8,12,20),nrow=8) )
> colnames(X)=c("PC", "OP", "VC","OV","P","Veg","Uva", "Platos")
> rownames(X)=c("PRodRu", "Asalrur","Prof","Ejsup","Ejmoy","Emp","Obr", "Des")
> X
         PC OP  VC  OV  P Veg Uva Platos
PRodRu  167  1 163  23 41   8   6      6
Asalrur 162  2 141  12 40  12   4     15
Prof    119  6  69  56 39   5  13     41
Ejsup    87 11  63 111 27   3  18     39
Ejmoy   103  5  68  77 32   4  11     30
Emp     111  4  72  66 34   6  10     28
Obr     130  3  76  52 43   7   7     16
Des     138  7 117  74 53   8  12     20
>
> princomp(X,cor=T)
Call:
princomp(x = X, cor = T)

Standard deviations:
   Comp.1       Comp.2       Comp.3       Comp.4       Comp.5       Comp.6       Comp.7       Comp.8
2.491575e+00 9.379133e-01 6.449505e-01 5.535835e-01 4.104162e-01 1.344162e-01 5.870919e-02 2.257138e-08

 8  variables and  8 observations.
```

La salida es $\sqrt{\lambda_k}$ de cada una de las componentes

Ejemplo (con princomp)

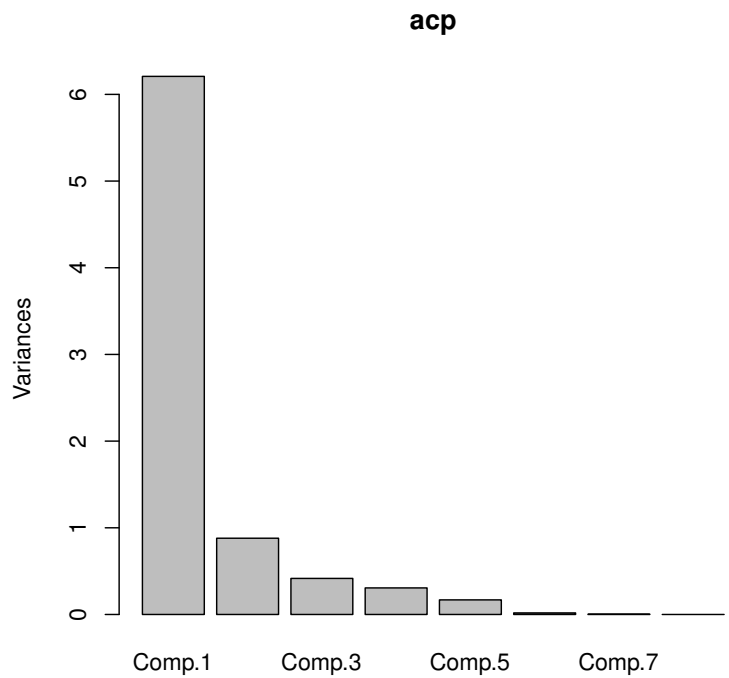
```
> acp=princomp(X,cor=T)
> summary(acp)
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4     Comp.5      Comp.6       Comp.7       Comp.8
Standard deviation     2.4915752 0.9379133 0.64495048 0.55358348 0.41041625 0.134416177 0.0587091892 2.257138e-08
Proportion of Variance 0.7759934 0.1099602 0.05199514 0.03830683 0.02105519 0.002258464 0.0004308461 6.368342e-17
Cumulative Proportion  0.7759934 0.8859535 0.93794867 0.97625550 0.99731069 0.999569154 1.0000000000 1.000000e+00
```

- En el primer renglón aparece los $\sqrt{\lambda_k}$
- En el segundo renglón la proporción de varianza $\frac{I_k}{I}$
- En el tercer renglón esta proporción cumulada.
- $I = \sum \lambda_k = 8$.

Elección de la cantidad de ejes

Esencialmente hay dos criterios:

- Criterio del codo. Se seleccionan los ejes antes del decaimiento menor de la varianza.



- Criterio de Kaiser. Se seleccionan los ejes cuya inercia es mayor a la inercia media I/p. Si la matriz es centrada y reducida, se retienen los ejes cuyos valores propios son mayores que 1.

En la practica se retienen los ejes que el usuario sabe interpretar.

Ejemplo (con princomp)

Las coordenadas de los vectores propios (son los a_k).

```
> loadings(acp)

Loadings:
       Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
PC      0.391 -0.138  0.162 -0.119 -0.294  0.398 -0.107  0.729
OP     -0.349 -0.441  0.320 -0.218  0.265  0.521  0.423 -0.118
VC      0.349 -0.202  0.681        -0.246 -0.465  0.254 -0.180
OV     -0.374 -0.260         0.397  0.346 -0.423         0.575
P       0.246 -0.744 -0.558        -0.176 -0.108        -0.135
Veg     0.365 -0.128        -0.519  0.669 -0.185 -0.313
Uva    -0.373 -0.326  0.254        -0.272        -0.766 -0.159
Platos -0.362        -0.162 -0.708 -0.333 -0.360  0.225  0.219

               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000
Proportion Var  0.125  0.125  0.125  0.125  0.125  0.125  0.125  0.125
Cumulative Var  0.125  0.250  0.375  0.500  0.625  0.750  0.875  1.000
>
```

Estas coordenadas están expresadas en el referencial original, el de las viejas variables.

Ejemplo (con princomp)

Las coordenadas de los individuos sobre los ejes (los $c_{ik} = \mathbf{x_i}' a_k$):

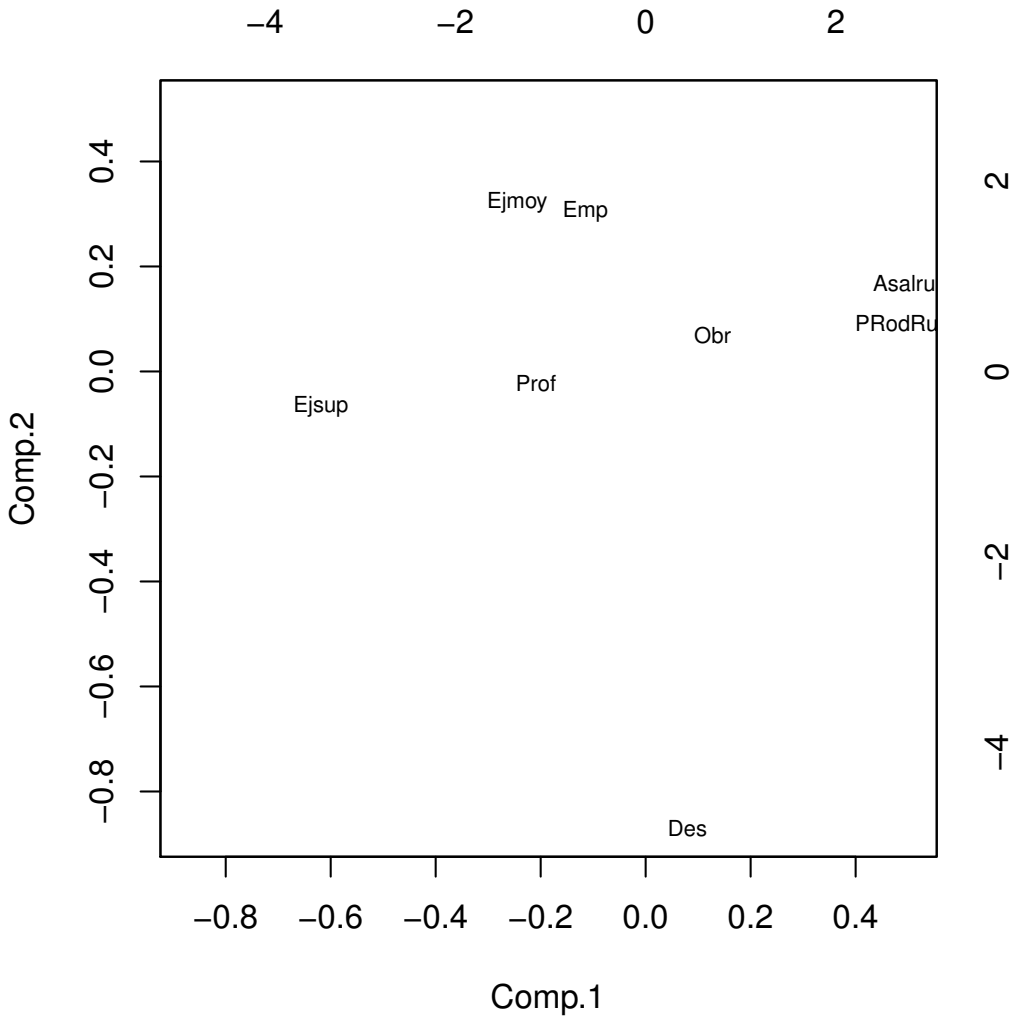
```
> acp$scores
            Comp.1      Comp.2     Comp.3      Comp.4      Comp.5       Comp.6       Comp.7        Comp.8
PRodRu   3.3715788  0.24581608  0.8395890  0.62172682 -0.57655700  0.021785204 -0.022311599  1.498801e-15
Asalrur  3.5217117  0.44739860  0.3515271 -0.91617942  0.49365924  0.004996441  0.031635398 -2.220446e-15
Prof    -1.4720309 -0.05851415 -0.5529570 -0.85448454 -0.74930243  0.058582704 -0.004644449 -1.609823e-15
Ejsup   -4.3587865 -0.17610682  1.0291875 -0.01517950  0.25877162  0.126514563 -0.015935125 -1.443290e-15
Ejmoy   -1.7180777  0.85664744 -0.1746349  0.41188554 -0.03988644 -0.139997633  0.116965074 -2.595146e-15
Emp     -0.8065346  0.80852679 -0.3448490  0.06912202  0.20594611 -0.195710003 -0.109502701  9.339751e-15
Obr      0.8991001  0.18303912 -0.9776683  0.55082419  0.29317809  0.233721990 -0.005887291 -3.441691e-15
Des      0.5630391 -2.30680707 -0.1701944  0.13228491  0.11419083 -0.109893267  0.009680694  4.024558e-16
```

Estas coordenadas están expresadas en el nuevo referencial, el de las componentes principales.

Ejemplo

Coordenadas de los individuos sobre los ejes factoriales:

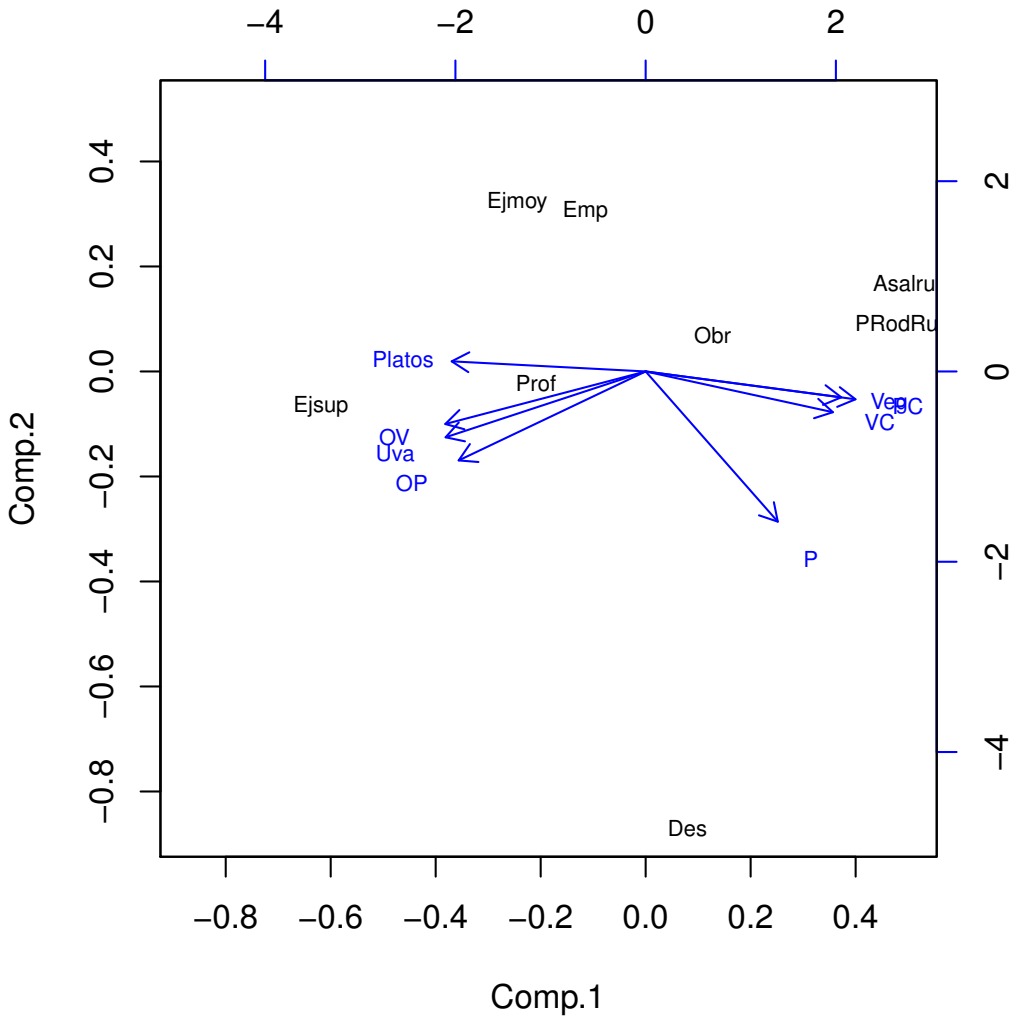
```
> biplot(acp, cex=0.7,col=c(1,0))
```



Ejemplo

Coordenadas de los individuos sobre los ejes factoriales y proyección de las variables originales:

```
> biplot(acp, cex=0.7,col=c(1,4))
```



Ejemplo

Las coordenadas de las variables sobre los ejes factoriales es

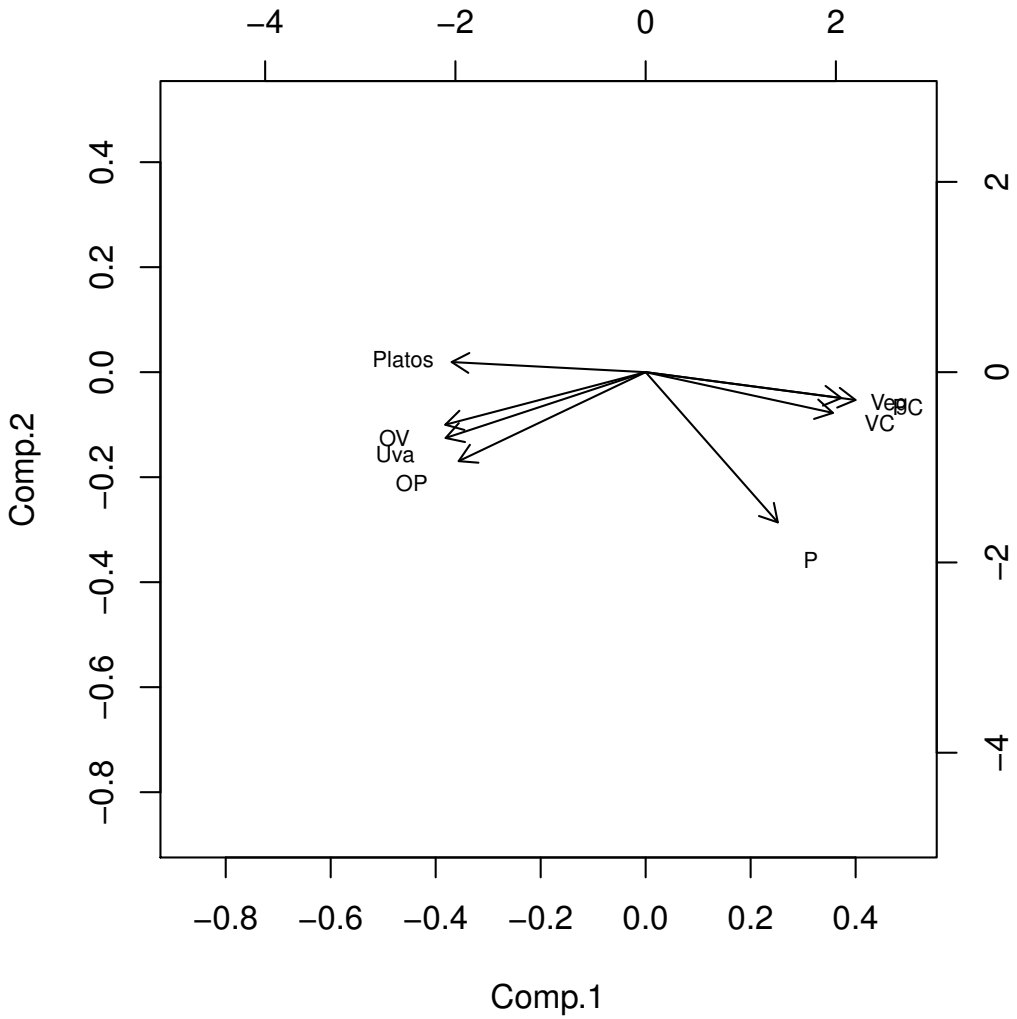
$$d_{jk} = r(x_j, z_k)$$

```
> cor(X,acp$scores)
           Comp.1      Comp.2      Comp.3      Comp.4     Comp.5       Comp.6       Comp.7       Comp.8
PC      0.9749797 -0.12926598  0.10429757 -0.06606998 -0.1206810  0.053463727 -0.006277188 -0.050156302
OP     -0.8687483 -0.41323074  0.20635173 -0.12063082  0.1089416  0.069991068  0.024838650 -0.123257316
VC      0.8700402 -0.18916036  0.43897378  0.01598936 -0.1008460 -0.062470185  0.014907633  0.015617014
OV     -0.9309151 -0.24414749  0.04739248  0.21952071  0.1418418 -0.056840057  0.001957682  0.032209797
P       0.6138529 -0.69764474 -0.35966296  0.04096049 -0.0721205 -0.014482890  0.005485109 -0.082240510
Veg     0.9089814 -0.12007291  0.02089707 -0.28724855  0.2746472 -0.024859138 -0.018382280 -0.022042647
Uva    -0.9294859 -0.30574089  0.16397854 -0.03526677 -0.1114413  0.002186234 -0.044965573 -0.003946343
Platos -0.9011429  0.04710881 -0.10428318 -0.39199413 -0.1366334 -0.048422745  0.013207567 -0.023521898
```

Ejemplo

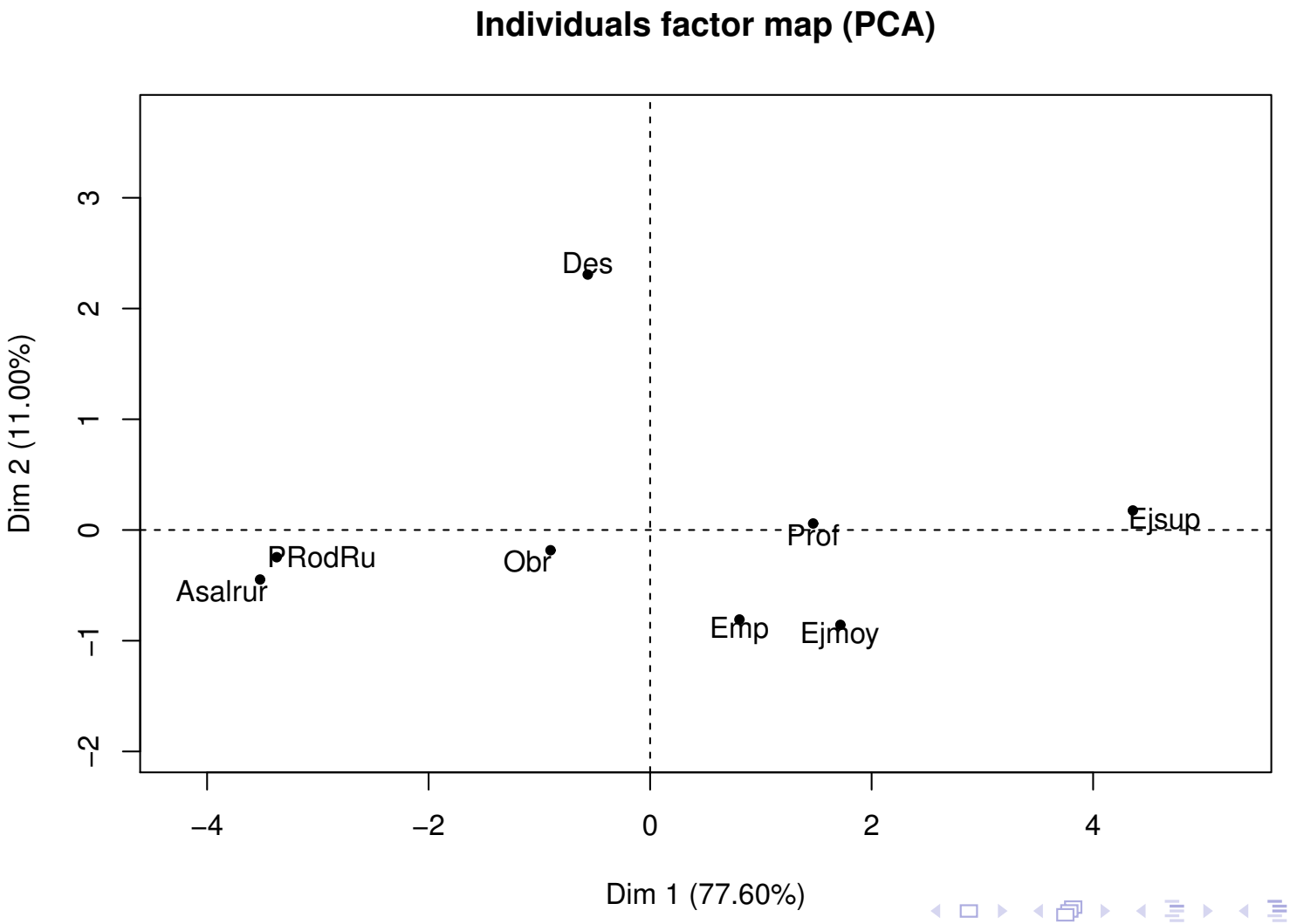
```
> biplot(acp, cex=0.7,col=c(0,1))
```

(con $c(0,1)$ activo las variables y escondo los individuos)



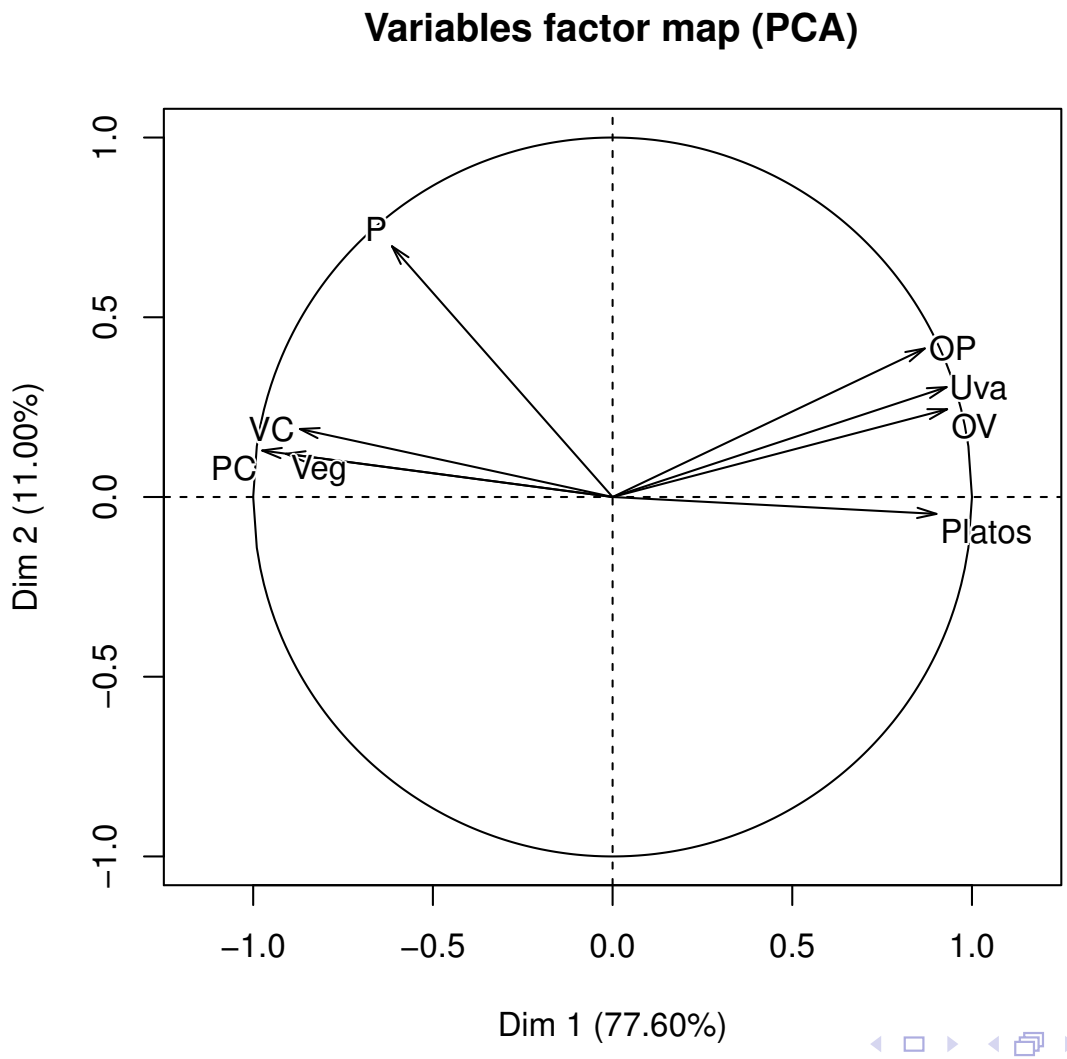
Mismo ejemplo con FactoMineR

```
> library(FactoMineR)
> acp1=PCA(X)
```



Mismo ejemplo con FactoMineR

```
> library(FactoMineR)
> acp1=PCA(X)
```



Mismo ejemplo con FactoMineR

Valores propios y varianza explicada por los ejes.

```
> acp1$eig
        eigenvalue percentage of variance cumulative percentage of variance
comp 1 6.207946839             77.59933549                           77.59934
comp 2 0.879681393             10.99601741                           88.59535
comp 3 0.415961123              5.19951404                           93.79487
comp 4 0.306454670              3.83068337                           97.62555
comp 5 0.168441497              2.10551872                           99.73107
comp 6 0.018067709              0.22584636                           99.95692
comp 7 0.003446769              0.04308461                          100.00000
```

observar que son bien los cuadrados de los valores que se obtienen con la función princomp.

Las coordenadas de los individuos sobre los ejes (los $c_{ik} = \mathbf{x_i}' a_k$):

```
> acp1$ind$coord
             Dim.1       Dim.2      Dim.3       Dim.4       Dim.5
PRodRu  -3.3715788 -0.24581608  0.8395890 -0.62172682  0.57655700
Asalrur -3.5217117 -0.44739860  0.3515271  0.91617942 -0.49365924
Prof     1.4720309  0.05851415 -0.5529570  0.85448454  0.74930243
Ejsup    4.3587865  0.17610682  1.0291875  0.01517950 -0.25877162
Ejmoy    1.7180777 -0.85664744 -0.1746349 -0.41188554  0.03988644
Emp      0.8065346 -0.80852679 -0.3448490 -0.06912202 -0.20594611
Obr     -0.8991001 -0.18303912 -0.9776683 -0.55082419 -0.29317809
Des     -0.5630391  2.30680707 -0.1701944 -0.13228491 -0.11419083
```

Mismo ejemplo con FactoMineR

Coordenadas de las variables sobre los ejes factoriales (los d_{jk}):

```
> acp1$var$coord
             Dim.1       Dim.2       Dim.3       Dim.4      Dim.5
PC      -0.9749797  0.12926598  0.10429757  0.06606998  0.1206810
OP       0.8687483  0.41323074  0.20635173  0.12063082 -0.1089416
VC      -0.8700402  0.18916036  0.43897378 -0.01598936  0.1008460
OV       0.9309151  0.24414749  0.04739248 -0.21952071 -0.1418418
P       -0.6138529  0.69764474 -0.35966296 -0.04096049  0.0721205
Veg     -0.9089814  0.12007291  0.02089707  0.28724855 -0.2746472
Uva      0.9294859  0.30574089  0.16397854  0.03526677  0.1114413
Platos   0.9011429 -0.04710881 -0.10428318  0.39199413  0.1366334
```

Interpretación de los ejes

Para cada eje seleccionado y cada nube, miramos:

- ¿Cuáles son las variables que más contribuyen a la formación del eje?
- ¿Qué individuos participan más en la formación del eje?

Herramienta de medición: contribuciones de puntos (individuos, si no son anónimos y variables) a la inercia de este eje.

Son los puntos cuya contribución es mayor que el promedio lo que hace posible dar significado al eje.

Interpretación de los ejes

Para cada eje que se retiene y cada nube, miramos cuales son los variables que participan más a la formación del eje y cuales son los individuos que contribuyen más a la formación del eje.

Se mide esta contribución respecto de la inercia del eje. Si estos individuos tienen una contribución superior a la media, los mismos dan un sentido al eje.

bf La contribución del individuo i a la construcción del eje k es la inercia del individuo i dividido la inercia del eje k:

$$ctr_k(i) = \frac{I(i)}{I_k} = \frac{p_i c_{ik}^2}{\lambda_k}$$

La suma de las contribuciones da 1. Se suelen retener los individuos cuya contribución es mayor que $1/n$ en valor absoluto. Si todos los individuos tienen igual peso, entonces retenemos los individuos tales que $|c_{ik}| > \sqrt{\lambda_k}$

Interpretación de los ejes

Para cada eje que se retiene y cada nube, miramos cuales son los variables que participan más a la formación del eje y cuales son los individuos que contribuyen más a la formación del eje.

La contribución de la variable x_j a la construcción del eje k es

$$ctr_k(x_j) = \frac{I(x_j)}{I_k} = \frac{d_{jk}^2}{\lambda_k} = \frac{(\sqrt{\lambda_k} a_{jk})^2}{\lambda_k} = a_{jk}^2$$

La suma de las contribuciones da 1. Se suelen retener las variables cuya contribución es mayor que $1/p$ en valor absoluto, es decir tales que $|a_{jk}| > 1/\sqrt{p}$

Si la matriz de datos es estandarizada, son las variables proximas al borde de la circunferencia que contribuyen más a la construcción del eje, puesto que

$$d_{jk}^2 = r_{x_j,z_k}^2$$

Interpretación de los ejes, ejemplo

Interpretación del primer eje. Contribución de los individuos. $c_{i1} > \sqrt{\lambda_1} = 2{,}491575$

```
> acp1$eig
        eigenvalue percentage of variance cumulative percentage of variance
comp 1 6.207946839             77.59933549                           77.59934
comp 2 0.879681393             10.99601741                           88.59535
comp 3 0.415961123              5.19951404                           93.79487
comp 4 0.306454670              3.83068337                           97.62555
comp 5 0.168441497              2.10551872                           99.73107
comp 6 0.018067709              0.22584636                           99.95692
comp 7 0.003446769              0.04308461                          100.00000

> sqrt(acp1$eig[1,1])
[1] 2.491575
> acp1$ind$contrib
            Dim.1       Dim.2      Dim.3        Dim.4      Dim.5
PRodRu  22.889096  0.85862826 21.1831585 15.766778699 24.6686523
Asalrur 24.972938  2.84428985  3.7134275 34.237720045 18.0848730
Prof     4.363107  0.04865263  9.1884034 29.781885764 41.6653666
Ejsup   38.255441  0.44069383 31.8307079  0.009398506  4.9692885
Ejmoy    5.943573 10.42770773  0.9164726  6.919852720  0.1180624
Emp      1.309809  9.28909571  3.5736754  0.194884165  3.1475171
Obr      1.627714  0.47607181 28.7236966 12.375700062  6.3785789
Des      0.638321 75.61486019  0.8704580  0.713780038  0.9676612
```

Los valores de la tabla anterior son porcentajes y se calculan como:

```
>a$ind$coord[,1]^2/(8*a$eig[1])
```

PRodRu, Asalrur, Ejsup son los individuos que contribuyen más a la construcción del primer eje.

Interpretación de los ejes, ejemplo

Interpretación del primer eje. Contribución de las variables ($1/\sqrt{8} \approx 0{,}35$).

```
> a$var$contrib #devuelve porcentajes, la suma de cada columna es 100
           Dim.1      Dim.2      Dim.3       Dim.4     Dim.5
PC      15.312396  1.8995166  2.6151444  1.42443318  8.646269
OP      12.157378 19.4115329 10.2367831  4.74843319  7.045927
VC      12.193565  4.0675682 46.3259598  0.08342491  6.037657
OV      13.959573  6.7760892  0.5399656 15.72478652 11.944270
P        6.069888 55.3277802 31.0984451  0.54747464  3.087936
Veg     13.309507  1.6389460  0.1049827 26.92461130 44.781774
Uva     13.916742 10.6262892  6.4642966  0.40584971  7.372980
Platos 13.080951  0.2522777  2.6144227 50.14098653 11.083187

> sqrt(a$var$contrib[,1])# Son los loadings de la primera componente (en val. abs)
      PC       OP       VC       OV        P      Veg      Uva   Platos
3.913106 3.486743 3.491929 3.736251 2.463714 3.648220 3.730515 3.616760

> a$var$coord #los d_{jk}
            Dim.1       Dim.2       Dim.3       Dim.4      Dim.5
PC     -0.9749797  0.12926598  0.10429757  0.06606998  0.1206810
OP      0.8687483  0.41323074  0.20635173  0.12063082 -0.1089416
VC     -0.8700402  0.18916036  0.43897378 -0.01598936  0.1008460
OV      0.9309151  0.24414749  0.04739248 -0.21952071 -0.1418418
P      -0.6138529  0.69764474 -0.35966296 -0.04096049  0.0721205
Veg    -0.9089814  0.12007291  0.02089707  0.28724855 -0.2746472
Uva     0.9294859  0.30574089  0.16397854  0.03526677  0.1114413
Platos  0.9011429 -0.04710881 -0.10428318  0.39199413  0.1366334
```

Contribuyen más PC, VC por un lado OP, OV, Uva, Platos por otro a la construcción del primer eje.

Interpretación de los ejes, ejemplo

La primer componente mide la repartición de la consumición entre alimentos básicos (PC,VC,Veg) y alimentos más refinados (OP, OV, Uva, Platos) y contraponen los ejecutivos superiores a los trabajadores rurales.

El segundo eje es más característico de la consumición de papas, comida generalmente consumida por inactivos.

Coseno y correlación

- Si suponemos que $\overline{x_j} = \overline{x_k} = 0$ y que $s_j = s_k = 1$ (matriz centrada y reducida) entonces:

$$cos(x_j, x_k) = x_j' x_k = \sum_{i=1}^{n}(x_{ij} - \overline{x_j})(x_{ik} - \overline{x_k}) = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x_j})(x_{ik} - \overline{x_k})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x_j})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ik} - \overline{x_k})^2}} = r_{jk}$$

- Si suponemos que $\overline{x_j} = \overline{x_k} = 0$ se tiene que:

$$cos(x_j, x_k) = \frac{x_j' x_k}{||x_j'||||x_k||} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x_j})(x_{ik} - \overline{x_k})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x_j})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ik} - \overline{x_k})^2}} = r_{jk}$$

Le podemos dar pesos a las observaciones. Si $\mathbf{p}$ es el vector de probabilidades de la transparencia anterior, considero $P = diag(\mathbf{p})$. Si todos tienen el mismo peso $P = \frac{1}{n} I_n$ donde I_n es la identidad. Podemos definir las mismas nociones:

- El producto escalar entre dos variables centradas es su covarianza: $\langle x_j, x_k \rangle_P = x_j' P x_k = Cov(x_j, x_k)$ $\quad$ $||x_j||_P^2 = x_j' P x_j = Var(x_j)$
- El producto escalar entre dos variables centradas y reducidas es el coeficiente de correlación $\langle x_j, x_k \rangle_P = x_j' P x_k = r_{jk}$

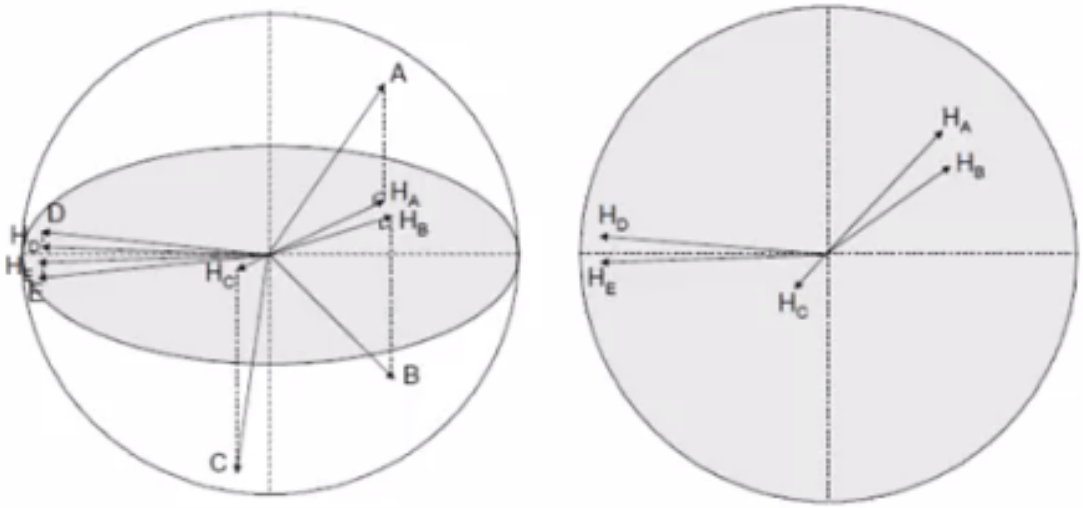
Variables e individuos bien representadas - Calidad de la proyección

Una vez que se interpretan los ejes, uno puede mirar los gráficos y analizar con mayor precisión las proximidades entre los puntos.

Las proximidades entre los puntos observados en un eje o un plano factorial deben corresponder a la realidad (y no deben ser creadas artificialmente por la proyección).
Para poder interpretar las proximidades entre puntos, deben estar bien representados en el eje o el plano factorial.

Se dice que un punto está bien representado en un eje o un plano factorial si está cerca de su proyección en el eje o el plano. Si está lejos, se dice que está pobremente representado. El indicador de la calidad de la representación será el ángulo formado entre el punto y su proyección en el eje.

Variables e individuos bien representadas - Calidad de la proyección



$r(A, B) = \cos(A, B)$ y si $\cos(A, B) \approx \cos(H_A, H_B)$ si las variables están bien proyectadas.
Sólo las variables bien proyectadas (cerca del eje y del borde del círculo) pueden ser correctamente interpretadas.
Lo mismo en cuanto a los individuos. Si dos individuos están mal proyectados, quizás estén lejos en el espacio de partido.

Calidad de representación de los puntos

Calidad de representación de los individuos.

Se mide si $\mathbf{x}_i$ es próximo a su proyección sobre el eje o el plano factorial con el angulo que forman. Sobre el eje k-esimo la calidad de representación de $\mathbf{x_i}$ con peso p_i es

$$cal_k(i) = \cos^2(\theta_{ik}) = \frac{p_i c_{ik}^2}{||\mathbf{x}_i||^2}$$

Cuando este coseno está cerca de 1, es decir el angulo es 0 o π el individuo está bien representado. En caso contrario el coseno está cerca de cero y el individuo pobremente representado.
Sobre el plano factorial determinado por a_{k_1} y a_{k_2}, la calidad de la representación es

$$cal_{k_1,k_2}(i) = cal_{k_1}(i) + cal_{k_2}(i)$$

Calidad de representación de los puntos

Calidad de representación de las variables.

De la misma manera la calidad de la representación de la variable x_j sobre el eje k es:

$$cal_k(x_j) = \cos^2(\theta_{jk}) = \frac{d_{jk}^2}{||x_j||^2}$$

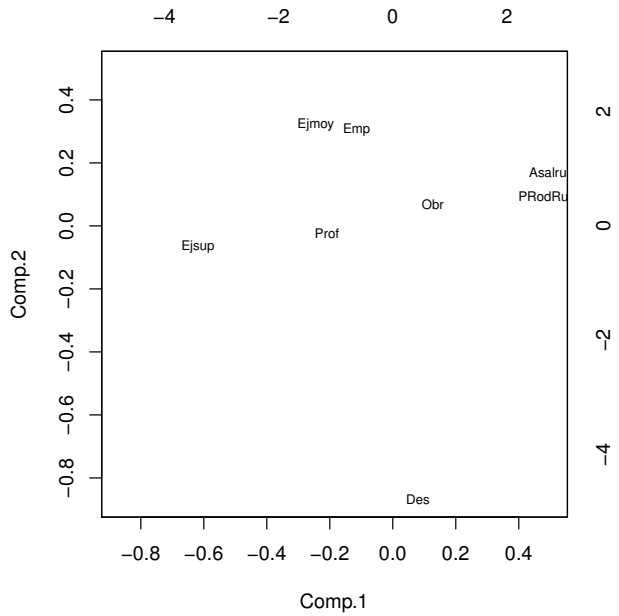
y si la matriz de datos es centrada reducida $d_{jk}^2 = r_{jk}^2$.

```
> (a$var$coord[,1])^2==a$var$cos2[,1]
    PC     OP     VC     OV      P    Veg    Uva Platos
  TRUE   TRUE   TRUE   TRUE   TRUE   TRUE   TRUE   TRUE
```

- Una variable estará mejor representada en un eje si estáa cerca del borde del círculo de correlaciones y el eje, mientras que estará pobremente representada si se encuentracerca del origen.
- Solamente las variables bien representadas pueden ser interpretadas!
- Las variables que contribuyen más a la construcción del eje son aquellas que están mejor representadas y al revez: aquellas que contribuyen menos son las que no tienen una buena representación. Esto es porque $cal_k(x_j) = \frac{d_{jk}^2}{||x_j||^2}$ y $ctr_k(x_j) = \frac{d_{jk}^2}{\lambda_k}$.

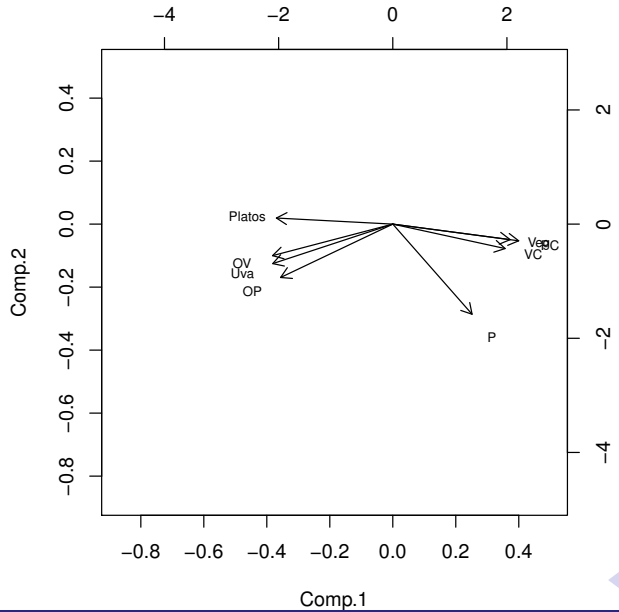
ejemplo

```
> a$ind$cos2
             Dim.1        Dim.2       Dim.3        Dim.4        Dim.5
PRodRu  0.88444010 0.0047013477 0.054844774 3.007468e-02 0.0258634411
Asalrur 0.89805821 0.0144939287 0.008947765 6.077961e-02 0.0176462090
Prof    0.57459845 0.0009079299 0.081079929 1.936150e-01 0.1488829157
Ejsup   0.94181776 0.0015374041 0.052507908 1.142223e-05 0.0033194718
Ejmoy   0.75288231 0.1871740871 0.007778640 4.327076e-02 0.0004057814
Emp     0.42778496 0.4299008580 0.078205503 3.142044e-03 0.0278924499
Obr     0.36060411 0.0149452245 0.426380807 1.353445e-01 0.0383422502
Des     0.05551846 0.9319290611 0.005072836 3.064650e-03 0.0022836142
```



ejemplo

```
> a$var$cos2
            Dim.1      Dim.2        Dim.3        Dim.4       Dim.5
PC      0.9505854 0.01670969 0.0108779840 0.0043652420 0.014563904
OP      0.7547236 0.17075964 0.0425810378 0.0145517953 0.011868265
VC      0.7569700 0.03578164 0.1926979829 0.0002556595 0.010169921
OV      0.8666029 0.05960800 0.0022460470 0.0481893426 0.020119107
P       0.3768154 0.48670819 0.1293574416 0.0016777616 0.005201366
Veg     0.8262471 0.01441750 0.0004366874 0.0825117286 0.075431090
Uva     0.8639440 0.09347749 0.0268889606 0.0012437454 0.012419159
Platos 0.8120585 0.00221924 0.0108749822 0.1536593947 0.018668686
```



Referencias

1. Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, An Introduction to Statistical Learning with Applications in R, Springer, 2013.
2. Daniel Peña, Análisis Multivariante, Mac Graw Hill, 2002.
3. FactoMineR `https://www.youtube.com/watch?v=CTSbxU6KLbM&list=PLnZgp6epRBbTsZEFXi_p6W48HhNyqwxIu&index=3`