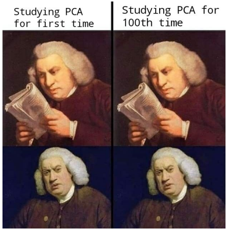
Estadística Aplicada III

Componentes Principales

Jorge de la Vega Góngora

Departamento de Estadística,
Instituto Tecnológico Autónomo de México

Semana 6



Conceptos

Introducción

- Hay conjuntos de datos que tienen un gran número de variables, y entender sus relaciones o sacar conclusiones de estos datos es complejo.
  - **Evaluación de vehículos a través de sus características**. Supongamos que a potenciales compradores de vehículos se les pide evaluar, en una escala Likert de 1 a 5, la importancia de 50 características de coches: color, consumo de combustible, tipo de motor, etc.
  - **Rendimientos de mercado**. Rendimientos diarios de los últimos 20 años en las 30 acciones en el Índice Dow Jones.
  - **Mediciones de inteligencia**. 100 medidas de inteligencia para una muestra de 1,000 estudiantes de preparatoria.
  - **Datos de clientes y sus compras**. Se tienen 200 variables demográficas de la base de datos de clientes de Amazon México con historia de compras y devoluciones de todos los productos que han comprado en 5 años.

Introducción

- El Análisis de Componentes Principales (ACP o PCA en inglés) encuentra combinaciones lineales de variables que mejor explican la estructura de variación de las variables.
- Aplicaciones:
  - **Reducción de dimensión**: para resumir un conjunto grande de variables en un conjunto más pequeño; creación de índices e indicadores
  - **Interpretación de datos**: encontrar las características que explican variación
  - **Visualización**
  - como un **paso intermedio en el análisis de datos**: una vez obtenidas las ACP, se pueden aplicar otras técnicas de análisis multivariado a las ACP como si fueran los datos originales.
- Para realizar ACP, *no* se requiere hacer el supuesto de normalidad para los vectores aleatorios, pero si los datos son normales, entonces es posible hacer inferencia (es decir: estimación, algunas pruebas de hipótesis e intervalos de confianza).

Planteamiento del problema I

- Dado un vector aleatorio $\mathbf{x} = (x_1, \dots, x_p) \in \mathbb{R}^p$, se buscan p combinaciones lineales de sus componentes

$$
\begin{aligned}
y_1 &= a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p = \mathbf{a}_1'\mathbf{x} \\
y_2 &= a_{21}x_1 + a_{22}x_2 + \cdots + a_{2p}x_p = \mathbf{a}_2'\mathbf{x} \\
&\vdots \\
y_p &= a_{p1}x_1 + a_{p2}x_2 + \cdots + a_{pp}x_p = \mathbf{a}_p'\mathbf{x}
\end{aligned}
$$

  tal que la variabilidad de k de esas p combinaciones sea aproximadamente la variabilidad de todo $\mathbf{x}$, con $k << p$ (k mucho menor que p).
- Los componentes del vector $\mathbf{a}_k$ son las *cargas (loadings)* para la k-ésima CP.

Planteamiento del problema II

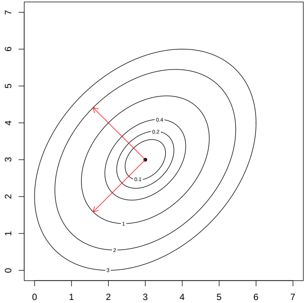
- Las combinaciones y_i tienen las propiedades:

$$\mathsf{Var}(y_i) = \mathbf{a}_i'\Sigma\mathbf{a}_i \text{ y}$$

$$\mathsf{cov}\,(y_i, y_j) = \mathbf{a}_i'\Sigma\mathbf{a}_j$$

  Las CP son las combinacionales lineales, no correlacionadas entre sí, cuya varianza es tan grande como sea posible.

- Geométricamente, la transformación representa la selección de un nuevo sistema de coordenadas, obtenido por rotación y traslación del sistema original a ejes en donde se maximiza la varianza, en cada dirección.



Planteamiento del problema III

- Las primeras k componentes principales expanden un subespacio que contiene *la mejor visualización posible en k dimensiones* de los datos: una proyección o 'sombra' vista en la dirección de más información.

- Se define la *rotación de componentes principales*, como la transformación:

$$\mathbf{y} = \mathbf{P}'(\mathbf{x} - \boldsymbol{\mu})$$

donde $\mathbf{P} = [\mathbf{e}_1 \cdots \mathbf{e}_p]$ es la matriz de eigenvectores de $\boldsymbol{\Sigma}$ tal que $\boldsymbol{\Sigma} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}'$, y que bajo dicha transformación, los contornos de la distribución de las variables transformadas se pueden expresar como

$$\mathbf{y}'\boldsymbol{\Lambda}^{-1}\mathbf{y} = k^2.$$

Componentes Principales poblacionales I

- Sea $\mathbf{x}_{p\times 1}$ un vector aleatorio cuya matriz de covarianzas es $\boldsymbol{\Sigma}$ con eigenvalores $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_p \geq 0$.
- Sean $\mathbf{a}_1, \ldots, \mathbf{a}_p \in \mathbb{R}^p$ los coeficientes de p combinaciones lineales $Y_i = \mathbf{a}_i\mathbf{x}$. Entonces $\text{cov}\,(Y_i, Y_j) = \mathbf{a}_i'\boldsymbol{\Sigma}\mathbf{a}_j$. Queremos que las componentes principales sean ortogonales entre sí: $\text{cov}\,(Y_i, Y_j) = 0$

Problema de optimización

Las CPs serán las combinaciones lineales no correlacionadas ($Y_i \perp Y_j$) tales que $\text{Var}(Y_1), \ldots, \text{Var}(Y_p)$ es la máxima posible. Se agrega la restricción adicional de que $||\mathbf{a}_i|| = 1$, para no incrementar de forma arbitraria $\text{Var}(Y_i)$.
El problema consiste en resolver el siguiente sistema de problemas de optimización:

$$
\begin{aligned}
\mathbf{a}_1 &= \arg\max\left\{\mathbf{a} : ||\mathbf{a}|| = 1, \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a} = \text{Var}(Y_1)\right\} \\
\mathbf{a}_i &= \arg\max\left\{\mathbf{a} : ||\mathbf{a}_i|| = 1, \mathbf{a}_i'\boldsymbol{\Sigma}\mathbf{a}_k = 0 \text{ si } k < i, \mathbf{a}_i'\boldsymbol{\Sigma}\mathbf{a}_i = \text{Var}(Y_i)\right\}, \quad i = 2, \ldots, p.
\end{aligned}
$$

Componentes Principales poblacionales II

Demostración.

Supongamos que $\mathbf{\Sigma} = \mathbf{P\Lambda P'}$ es la descomposición espectral de tal manera que $\mathbf{\Sigma}^{1/2} = \mathbf{P}\Lambda^{1/2}\mathbf{P}'$, donde la matriz $\mathbf{P}$ cumple con la condición $\mathbf{P'P} = \mathbf{PP'} = \mathbf{I}$. Sea $\mathbf{y} = \mathbf{P'a}$. Entonces, si $\mathbf{a} \neq \mathbf{0}, \mathbf{y} \neq \mathbf{0}$ y

$$\begin{aligned}
\frac{\mathbf{a'\Sigma a}}{\mathbf{a'a}} = \frac{\mathbf{a'}\mathbf{\Sigma}^{1/2}\mathbf{\Sigma}^{1/2}\mathbf{a}}{\mathbf{a'PP'a}} &= \frac{\mathbf{a'P}\Lambda^{1/2}\mathbf{P'P}\Lambda^{1/2}\mathbf{P'a}}{\mathbf{y'y}} \\
&= \frac{\mathbf{y'}\Lambda\mathbf{y}}{\mathbf{y'y}} \\
&= \frac{\sum_{i=1}^{p} \lambda_i y_i^2}{\sum_{i=1}^{p} y_i^2} \\
&\leq \lambda_1 \frac{\sum_{i=1}^{p} y_i^2}{\sum_{i=1}^{p} y_i^2} = \lambda_1
\end{aligned}$$

Si se toma $\mathbf{a} = \mathbf{e}_1$ entonces $\mathbf{y} = \mathbf{P}'\mathbf{e}_1 = \mathbf{e}_1$, entonces $\frac{\mathbf{e}_1'\boldsymbol{\Sigma}\mathbf{e}_1}{\mathbf{e}_1'\mathbf{e}} = \mathbf{e}_1\boldsymbol{\Sigma}\mathbf{e}_1 = \lambda_1$.
Por otra parte, $\mathbf{y} = \mathbf{P}'\mathbf{x} \Rightarrow \mathbf{P}\mathbf{y} = \mathbf{a} = \sum_{i=1}^{p} y_i\mathbf{e}_i$. Si $\mathbf{a} \perp \mathbf{e}_1, \ldots \mathbf{e}_k$ entonces

$$0 = \mathbf{e}_i'\mathbf{a} = y_1\mathbf{e}_1'\mathbf{e}_1 + y_2\mathbf{e}_2'\mathbf{e}_2 + \cdots + y_p\mathbf{e}_p'\mathbf{e}_p = y_i, \quad i \leq k$$

Por lo tanto el cociente se puede escribir como:

$$\frac{\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}}{\mathbf{a}'\mathbf{a}} = \frac{\sum_{i=k+1}^{p} \lambda_i y_i^2}{\sum_{i=k+1}^{p} y_i^2}$$

Si tomamos $y_{k+1} = 1$ y el resto de las $y_{k+i} = 0$ para $i = 2, \ldots, p$, se obtiene el máximo establecido.

$\square$

- Entonces la solución al problema de optimización está dada por los vectores y valores propios de la matriz $\boldsymbol{\Sigma}$:

$$\mathbf{a}_i^* = \mathbf{e}_i \text{ y } \mathrm{Var}(Y_i) = \lambda_i, \quad i = 1, \ldots, p$$

Componentes Principales poblacionales IV

- Podemos escribir la matriz de covarianzas $\mathbf{\Sigma}$ como:

$$\mathbf{\Sigma} = \mathbf{P}\mathbf{\Delta}\mathbf{P}' = \sum_{i=1}^{p} \lambda_i \mathbf{e}_i \mathbf{e}_i'$$

- Noten que:
  - Var$(Y_i) = \mathbf{e}_i'\mathbf{\Sigma}\mathbf{e}_i = \mathbf{e}_i'\mathbf{P}\mathbf{\Delta}\mathbf{P}'\mathbf{e}_i = \lambda_i$
  - Las variables son ortogonales: cov $(Y_i, Y_j) = \mathbf{e}_i'\mathbf{\Sigma}\mathbf{e}_j = \mathbf{e}_i'\mathbf{P}\mathbf{\Delta}\mathbf{P}'\mathbf{e}_j = 0$ si $i \neq j$.
  - Las componentes principales quedan determinadas salvo el signo de las cargas, ya que sólo se pide que $||\mathbf{a}_i|| = 1$
  - Para calcular las componentes principales, no se requiere tener la muestra, basta tener una matriz de covarianza o una matriz de correlaciones.

Solución alternativa (multiplicadores de Lagrange) I

- Una forma alternativa de resolver el problema de optimización con restricciones es utilizar los multiplicadores de Lagrange, que también nos será útil en otros problemas de optimización.
- Para $i = 1$, maximizamos la función

$$f(\mathbf{a}_1, \lambda) = \mathbf{a}_1' \boldsymbol{\Sigma} \mathbf{a}_1 - \lambda(\mathbf{a}_1' \mathbf{a}_1 - 1)$$

donde λ es el multiplicador de Lagrange. Derivando respecto a $\mathbf{a}_1$ e igualando a 0 se obtiene el sistema:

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

Por lo tanto, λ es un eigenvalor y $\mathbf{a}_1$ es un eigenvector. ¿Cuál de los eigenvectores? Notemos que:

$$\mathbf{a}_1' \boldsymbol{\Sigma} \mathbf{a}_1 = \mathbf{a}_1' \lambda \mathbf{a}_1 = \lambda ||\mathbf{a}_1|| = \lambda$$

Entonces la función se maximiza cuando $\lambda^* = \lambda_1$, el eigenvalor más grande, y en ese caso, $\mathbf{a}_1 = \mathbf{e}_1$.

Solución alternativa (multiplicadores de Lagrange) II

- Para las componentes subsecuentes, se agrega adicionalmente la restricción de ortogonalidad $\mathbf{a}_i\mathbf{a}_j = 0$, utilizando un segundo multiplicador y tomando las derivadas correspondientes:

$$f(\mathbf{a}_i, \lambda, \phi) = \mathbf{a}_i'\mathbf{\Sigma}\mathbf{a}_i - \lambda(\mathbf{a}_i'\mathbf{a}_i - 1) - \phi(\mathbf{a}_i'\mathbf{a}_j)$$

Ejemplo: calificaciones I

Los siguientes datos simulan la calificación de 20 estudiantes universitarios en 5 materias: matemáticas, literatura, física, estadística y filosofía (son datos simulados

```
          mat      lit       fis       est       fil
 [1,] 70.18746 77.01845  81.73103  45.24811  91.96414
 [2,] 68.15747 69.07357  44.74910  60.87647  87.88669
 [3,] 56.28669 76.62567  43.42675  53.39355  92.56447
 [4,] 64.00832 69.40469  76.68948  76.80231  92.58300
 [5,] 72.94545 73.67401  40.64696  91.32753  91.20121
 [6,] 73.89794 78.13169  59.42369  94.32252  89.79415
 [7,] 57.91924 76.56222  64.65050  84.71381  92.29882
 [8,] 66.36324 75.63921  53.97583  60.37583 100.10260
 [9,] 53.73327 79.49119  46.44771  81.25490  83.08027
[10,] 67.43522 78.73110  73.10455  45.07361  88.63022
[11,] 81.01780 70.73130  51.98725  77.61844  86.91823
[12,] 77.55782 79.61027  53.30887  41.39145  88.04259
[13,] 67.61766 84.84283  87.35908  49.03109  89.34772
[14,] 79.87445 80.92463 102.75534  65.62993  94.77674
[15,] 77.41390 73.10028  70.11639  40.20128  91.97479
[16,] 70.89347 72.82243  75.72685  93.45413  81.90177
[17,] 60.45056 81.81044  41.95576  40.40346  80.96661
[18,] 68.04850 71.20457  70.65794  61.39224  93.46463
[19,] 79.25521 78.37728  47.08211  48.96723  89.74084
[20,] 74.82979 76.74219  65.81975 100.45173  95.13514
```

De acuerdo a lo visto, las cargas de las componentes principales corresponden a los vectores propios de la matriz Σ. Usando la estimación plug-in, tenemos

Ejemplo: calificaciones II

```
S <- var(X)
v <- eigen(S)
(lambdas <- v$values) # valores propios correspondientes a las componentes principales

[1] 415.25375 299.92983  59.94395  20.87397  13.64242

e <- v$vectors
round(e,4)           # cargas

        [,1]   [,2]    [,3]    [,4]    [,5]
[1,]  0.0150 0.1431  0.9775 -0.1410  0.0617
[2,]  0.0703 0.0306 -0.1282 -0.4686  0.8707
[3,]  0.1123 0.9791 -0.1446 -0.0330 -0.0825
[4,] -0.9911 0.1153 -0.0107 -0.0391  0.0534
[5,]  0.0001 0.0819  0.0834  0.8706  0.4780
```

Entonces las componentes principales se pueden calcular como las combinaciones lineales con pesos dados por $\mathbf{e}_i$. Los scores obtenidos son:

Ejemplo: calificaciones III

```
(y <- X %*% e)
```

```
            [,1]      [,2]     [,3]     [,4]     [,5]
 [1,] -29.19455 105.17294 54.10778 29.60464 111.0186
 [2,] -49.42532  69.89877 57.98228 30.67453 105.9128
 [3,] -41.80441  66.65705 46.07145 33.21745 113.7008
 [4,] -61.66100 102.80763 49.48657 33.51660 106.4039
 [5,] -79.66977  70.49016 62.61736 29.67480 113.7620
 [6,] -80.20198  89.37657 60.11266 25.48983 115.6399
 [7,] -70.44183  91.25687 44.24958 30.86180 113.5397
 [8,] -47.45767  79.81977 55.07670 38.20036 116.5693
 [9,] -68.91517  71.77150 41.68268 22.78997 112.7427
[10,] -29.91132  96.09223 52.16912 26.57925 111.4488
[11,] -64.89580  80.72712 69.03507 26.34824 107.9837
[12,] -28.27143  77.71380 64.80466 25.02544 113.9955
[13,] -31.80038 110.77652 49.52050 23.68892 116.1593
[14,] -46.61359 129.84298 60.05319 27.36660 115.7150
[15,] -25.66413  94.13480 63.40844 31.01158 108.7475
[16,] -77.92940 103.99931 54.84970 21.02668 105.6677
[17,] -28.66926  63.52367 48.86176 20.65984 112.3568
[18,] -46.87865  95.83150 54.31574 33.67194 108.3171
[19,] -36.53991  72.83497 67.58303 26.75192 114.7567
[20,] -85.64166  96.87377 60.65663 30.20875 116.8398
```

y vemos que la varianza de los scores coincide con los valores propios, por ejemplo

Ejemplo: calificaciones IV

```r
apply(y,2,var)
```

```
[1] 415.25375 299.92983  59.94395  20.87397  13.64242
```

```r
v$values # valores propios.
```

```
[1] 415.25375 299.92983  59.94395  20.87397  13.64242
```

Observaciones I

- Las CP $Y_1, \ldots, Y_p$ tienen la misma varianza total que las de las variables originales $x_1, \ldots, x_p$:

$$\sum_{i=1}^{p} \text{Var}(x_i) = \text{tr}(\mathbf{\Sigma}) = \sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{n} \text{Var}(Y_i)$$

- La proporción de la varianza total explicada o debida a la $j-$ésima componente principal esta dada por:

$$r_i = \frac{\lambda_i}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} \quad i = 1, \ldots, p.$$

- Si $\sum_{i=1}^{k} r_i \approx 1$ para alguna $k < p$ la varianza total se puede atribuir a las primeras cuantas componentes, entonces esas componentes pueden "sustituir" a las p variables sin mucha pérdida de información.
- Es importante notar que las componentes principales están determinadas salvo por el signo de las cargas, es decir, podemos obtener como cargas $\mathbf{a}$ o $-\mathbf{a}$.

Ejemplo: calificaciones (cont)

```r
# Varianza total
vary <- apply(y,2,var)    # varianzas
sum(vary)

[1] 809.6439

sum(diag(S))           # varianza total (traza de S)

[1] 809.6439

vary/sum(vary)          # Contribuciones de cada componente principal a la varianza

[1] 0.51288441 0.37044659 0.07403742 0.02578167 0.01684991

cumsum(vary)/sum(vary)  # porcentaje acumulado de varianza explicada:

[1] 0.5128844 0.8833310 0.9573684 0.9831501 1.0000000
```

La primera componente explica 51 % de la varianza, las dos primeras 88 % y así sucesivamente.

Contribución de las variables a las CPs

Las componentes del eigenvector $\mathbf{e}_i = (e_{i1}, \dots, e_{ip})$ ponderan la importancia de las variables originales en la $i-$ésima CP, es decir, son proporcionales a la correlación entre la $i-$ésima componente y la $k-$ésima variable:

Teorema

Si $Y_i = \mathbf{e}_i'\mathbf{x}, i = 1, 2, \dots, p$ son las CP, entonces $\rho_{Y_i,x_k} = e_{ik}\frac{\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}$.

Demostración.

Si $\mathbf{a}_k = \mathbf{1}_k$, entonces podemos escribir $x_k = \mathbf{a}_k'\mathbf{x}$. Así que:

$$\text{cov}\,(x_k, Y_i) = \text{cov}\left(\mathbf{a}_k'\mathbf{x}, \mathbf{e}_i'\mathbf{x}\right) = \mathbf{a}_k'\mathbf{\Sigma}\mathbf{e}_i = \mathbf{a}_k'\lambda_i\mathbf{e}_i = \lambda_i\mathbf{a}_k'\mathbf{e}_i = \lambda_i e_{ik}$$

Entonces

$$\rho_{Y_i,x_k} = \frac{\text{cov}\,(Y_i, x_k)}{\sqrt{\text{Var}(Y_i)}\sqrt{\text{Var}(x_k)}} = \frac{\lambda_i e_{ik}}{\sqrt{\lambda_i}\sqrt{\sigma_{kk}}} = e_{ik}\frac{\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}$$

Por lo tanto, cada e_{ik} es proporcional a la correlación entre Y_i y x_k.

□

Comunalidades

- Otra medida relevante es cuál es la contribución de cada CP a la explicación de la varianza de cada variable x_i.
- La porción de la varianza de la variable i explicada por las CPs es llamada la i**-ésima comunalidad**. Esta comunalidad es simplemente la suma de los cuadrados de las cargas de la variable i en cada una de las CPs consideradas:

$$h_i^2 \stackrel{def}{=} \sum_{l=1}^{p} e_{li}^2$$

- Las comunalidades tienen una intuición más clara que se verá cuando revisemos análisis factorial.

Ejemplo calificaciones (cont.)

- Las correlaciones de las componentes con cada variable:

```
cor(y,X)  # Cuando tenemos los datos.

             mat        lit          fis          est          fil
[1,]  0.03818296  0.3353994  0.133427274 -0.995056381 0.0002600249
[2,]  0.30984401  0.1240556  0.988707326  0.098359112 0.3069950919
[3,]  0.94617080 -0.2323000 -0.065282602 -0.004076347 0.1396703388
[4,] -0.08053956 -0.5012815 -0.008792261 -0.008793326 0.8604557197
[5,]  0.02851227  0.7529084 -0.017772560  0.009713136 0.3819189664

e[1,1]*sqrt(v$values[1])/sqrt(diag(S)[1]) # de acuerdo a la fórmula, si no tenemos los datos

       mat
0.03818296
```

- Las comunalidades para cada número de componentes:

```
h2 <- apply(e,2,function(x)cumsum(x^2))
h2

             [,1]       [,2]      [,3]      [,4]        [,5]
[1,] 0.0002246504 0.02048085 0.9555980 0.0198836 0.003812883
[2,] 0.0051669653 0.02141697 0.9720218 0.2395065 0.761887798
[3,] 0.0177762043 0.97999851 0.9929322 0.2405957 0.768697399
[4,] 0.9999999965 0.99328585 0.9930464 0.2421216 0.771546160
[5,] 1.0000000000 1.00000000 1.0000000 1.0000000 1.000000000
```

Herramientas en R I

- hay métodos de componentes principales en varios paquetes de R:
  - La función `princomp` calcula las componentes principales, usando una matriz de datos $\mathbf{X}$ o directamente las matrices de correlación o covarianza. El argumento `cor` controla si se usa correlación o covarianza, y también se le pueden pasar estimaciones robustos de la varianza.
  - La función `prcomp` también está disponible sin necesidad de cargar ningún paquete. Sólo acepta la matrix $\mathbf{X}$. El cálculo con esta función es menos robusto que con la función anterior.
  - `FactoMiner`: Exploratory data analysis methods to summarize, visualize and describe datasets.
  - `ade4`: Tools for multivariate data analysis.
  - `ca`: Computation and visualization of simple, multiple and joint correspondence analysis.
  - `MASS`: Functions and datasets to support Venables and Ripley, “Modern Applied Statistics with S”(4th edition, 2002).
  - `Exposition`: descriptive (i.e., fixed-effects) multivariate analysis with the singular value decomposition.
  - `factoextra` que permite en general visualizar y extraer información de análisis de datos exploratorio multivariado.

Ejemplo calificaciones (cont.)

- Aplicando la función `princomp` obtenemos la siguiente salida. Para facilitar la interpretación, quita las cargas que son menores a 10 %.

```
m <- princomp(X)
summary(m, loadings=T)
```

```
Importance of components:
                           Comp.1     Comp.2     Comp.3     Comp.4     Comp.5
Standard deviation     19.8617991 16.8799686 7.54630713 4.45311948 3.60004211
Proportion of Variance  0.5128844  0.3704466 0.07403742 0.02578167 0.01684991
Cumulative Proportion   0.5128844  0.8833310 0.95736842 0.98315009 1.00000000

Loadings:
    Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
mat         0.143  0.978  0.141
lit               -0.128  0.469  0.871
fis  0.112  0.979 -0.145
est -0.991  0.115
fil                      -0.871  0.478
```

ACP en la práctica: consideraciones

- El reescalamiento de las variables puede generar un cambio fundamental en los resultados del ACP. El ACP está tratando de explicar la variación en covarianza (**S**) o en correlación (**R**).
    - Si las unidades de las p variables son comparables, usar la covarianza puede ser más informativa, porque las unidades de medición se conservan.
    - Si las unidades de las p variables no son comparables, la correlación puede ser más informativa porque las unidades de medición se remueven.
- Recordar que la solución del problema de optimización puede cambiar de signo, porque toma en cuenta la dirección de los vectores, no su sentido.

Ejemplo 2: Creación de un Score para evaluación (Mardia) I

- Se tienen las calificaciones de 88 estudiantes en diferentes materias (algunos exámenes son a libro abierto y otros a libro cerrado). Las variables son:
  - `MC`: Mecánica (libro cerrado)
  - `VC`: Álgebra lineal (libro cerrado)
  - `LO`: Álgebra moderna (libro abierto)
  - `NO`: Análisis (libro abierto)
  - `SO`: Estadística (libro abierto)

```
library(factoextra)
```

```
Cargando paquete requerido: ggplot2
Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
X <- read.table("https://raw.githubusercontent.com/jvega68/EA3/master/datos/Mardia_Kent_Bibby/openclosedbook.dat",
header = T)
head(X)
```

```
  MC VC LO NO SO
1 77 82 67 67 81
2 63 78 80 70 81
3 75 73 71 66 81
4 55 72 63 70 68
5 63 63 65 70 63
6 53 61 72 64 73
```

Ejemplo 2: Creación de un Score para evaluación (Mardia) II

- Notemos que en este caso las unidades de las calificaciones son similares, por lo que podemos trabajar directamente con la matriz **S**. También al no tener unidades, podemos trabajar con la matriz de correlación **R**.
- Nos interesa, a partir de las 5 evaluaciones, una evaluación global que preserve en la mayor medida posible la variabilidad del grupo de estudiantes y poderlos separar adecuadamente por su desempeño.
- Una posibilidad es hacer un promedio de las observaciones, pero así se ponderan del mismo modo todas las variables que conforman la calificación sin importar su variabilidad.
- Aplicando la descomposición en componentes principales:

```
z <- princomp(X)  #función de componentes principales
summary(z)
```

```
Importance of components:
                          Comp.1     Comp.2      Comp.3     Comp.4     Comp.5
Standard deviation     26.061142 14.1355705 10.12760414 9.14706148 5.63807655
Proportion of Variance  0.619115  0.1821424  0.09349705 0.07626893 0.02897653
Cumulative Proportion   0.619115  0.8012575  0.89475453 0.97102347 1.00000000
```

Ejemplo 2: Creación de un Score para evaluación (Mardia) III

- Noten que $\text{tr}(\mathbf{\Sigma}) = \sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{n} \text{Var}(Y_i)$, y entonces la proporción de la varianza total explicada o debida a la $j-$ésima componente principal es

$$\frac{\lambda_i}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} \quad i = 1, \ldots, p.$$

- De acuerdo a los resultados anteriores, la primera componente explica cerca del 64 % de la variabilidad total, mientras que las primeras dos componentes cubren 78.4 % de la variabilidad.
- para obtener las combinaciones lineales (o sea, los vectores $\mathbf{e}_1, \ldots \mathbf{e}_p$ para cada componente), podemos ver los ponderadores o cargas (*loadings*):

Ejemplo 2: Creación de un Score para evaluación (Mardia) IV

```
summary(z, loadings = T)

Importance of components:
                          Comp.1     Comp.2      Comp.3     Comp.4     Comp.5
Standard deviation     26.061142 14.1355705 10.12760414 9.14706148 5.63807655
Proportion of Variance  0.619115  0.1821424  0.09349705 0.07626893 0.02897653
Cumulative Proportion   0.619115  0.8012575  0.89475453 0.97102347 1.00000000

Loadings:
   Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
MC  0.505  0.749  0.300  0.296
VC  0.368  0.207 -0.416 -0.783  0.189
LO  0.346        -0.145        -0.924
NO  0.451 -0.301 -0.597  0.518  0.286
SO  0.535 -0.548  0.600 -0.176  0.151
```

Entonces las combinaciones lineales de las componentes principales son:

$$
\begin{aligned}
Z_1 &= +0.505MC + 0.368VC + 0.346LO + 0.451NO + 0.535SO \\
Z_2 &= +0.749MC + 0.207VC - 0.076LO - 0.301NO - 0.548SO \\
Z_3 &= +0.3MC - 0.416VC - 0.145LO - 0.597NO + 0.6SO \\
Z_4 &= +0.296MC - 0.783VC - 0.003LO + 0.518NO - 0.176SO \\
Z_5 &= +0.079MC + 0.189VC - 0.924LO + 0.286NO + 0.151SO
\end{aligned}
$$

Ejemplo 2: Creación de un Score para evaluación (Mardia) V

por ejemplo, $\mathbf{e}_1 = (0.505, 0.368, 0.346, 0.451, 0.535)$.

- Por otro lado, las componentes tienen varianzas dadas por:

$$(\lambda_1^2, \lambda_2^2, \lambda_3^2, \lambda_4^2, \lambda_5^2) = (26.06, 14.14, 10.13, 9.15, 5.64)^2 = (679.18, 199.81, 102.57, 83.67, 31.79)$$

respectivamente.

- La comunalidad para la primera variable está dada por:

$$h_1^2 = 0.505^2 + 0.749^2 + 0.3^2 + 0.296^2 + 0.079^2 \approx 1$$

(no es exactamente 1 por redondeo) La comunalidad para la primera variable tomando sólo las dos primeras CP es:

$$h_1^2 = 0.505^2 + 0.749^2 = 0.816026$$

Es decir, las primeras dos componentes explican el 81.6 % de la varianza de la primera variable.

Ejemplo 2: Creación de un Score para evaluación (Mardia) VI

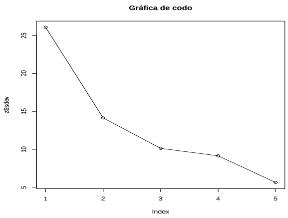
- Interpretación de las componentes:
  - la primera da un peso positivo a cada variable, por lo que representa una calificación promedio, pero ponderada de acuerdo a la variabilidad de cada evaluación.
  - La segunda componente se puede interpretar como un contraste entre las calificaciones a libro abierto y las que son a libro cerrado.
- En la interpretación, es relevante fijarse en la magnitud absoluta de los coeficientes (ya que el signo no importa) o los contrastes (cambios de signo) entre las diferentes variables.

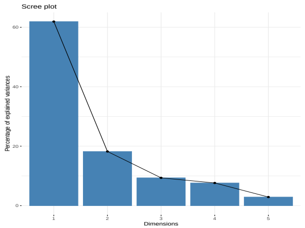
Elección del número de componentes a considerar

- Sólo en el caso de vectores con distribución normal se puede usar una guía formal para determinar cuántas CPs son significativas. En general, se tienen las siguientes guías:
  - Retener las CPs que acumulen un umbral, por ejemplo, 80 %, de la variación total.
  - Retener las CPs cuyos eigenvalores sean mayores que el promedio $\bar{\lambda}$. Para una matriz de correlación, este promedio es 1.
  - Usar la gráfica de codo (*scree plot*): $\{i, \lambda_i\}$ y buscar el "doblez" natural entre valores grandes y valores pequeños de λ_i. La gráfica de codo para estos datos está dada a continuación. En esta gráfica no está claro dónde se da el doblez, que podría ser en la segunda o tercera componente. Vemos dos versiones de la gráfica de codo:

```r
plot(z$sdev, type = "o", main = "Gráfica de codo")       # versión simple
```

```r
fviz_screeplot(z)  #versión "bonita"
```

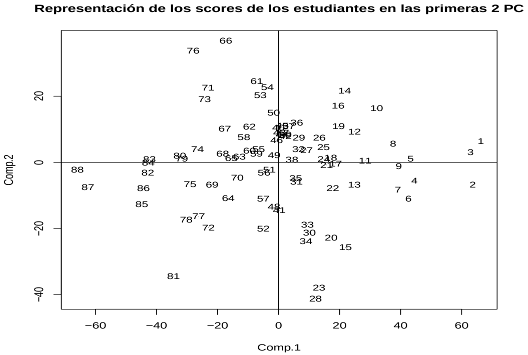




Biplots I

- Un **biplot** es una gráfica que muestra las proyecciones de los datos en las componentes principales. Puede ser de individuos, de variables o de ambas.
- Podemos utilizar las dos primeras componentes principales para visualizar los datos y asociar las evaluaciones a los diferentes alumnos para visualizar su distribución en la dirección de esas componentes. Esta es una gráfica de individuos.

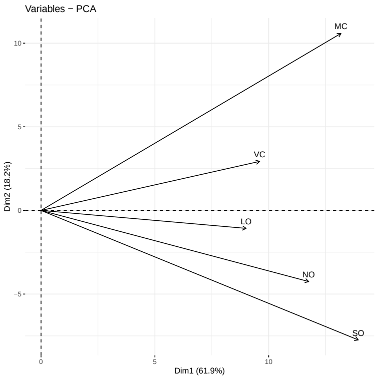
```
plot(z$scores[,1:2], pch = 16, cex = 0.1, main = "Representación de los scores de los estudiantes en las primeras 2 PC")
text(z$scores, labels = 1:88, cex = 0.9) # Pon el número de alumno en el punto del i-ésimo score
abline(h=0); abline(v=0)
```



- Podemos visualizar las CP como proyecciones en el espacio de las dos primeras componentes principales. Lo que se grafica son las cargas que tienen las variables en las dos primeras componentes. Esta es una gráfica de variables

```
fviz_pca_var(z)
```

Biplots III

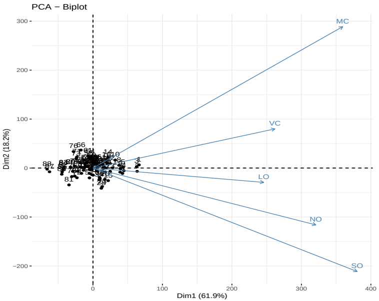


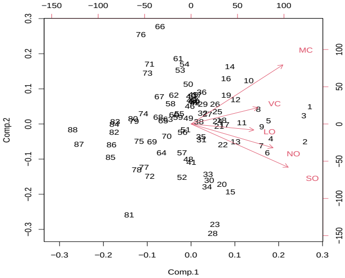
Biplots IV

- Podemos visualizar tanto individuos como variables en la gráfica

```r
fviz_pca_biplot(z) # ggplot2
```

```r
biplot(z) # tradicional
```

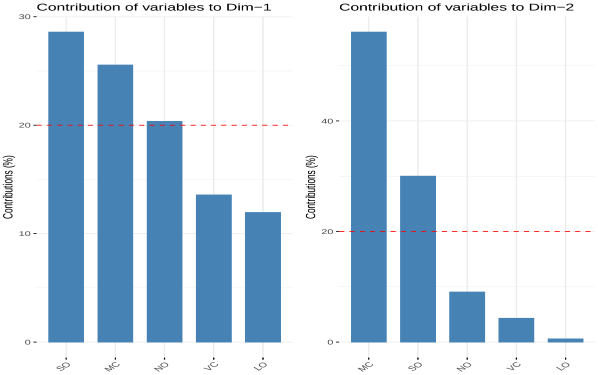




- También podemos ver la contribución de cada variable original a la CP (es decir, son las comunalidades individuales, o los valores de las cargas cuadradas):

```r
library(gridExtra)
gr1 <- fviz_contrib(z, choice = "var", axes = 1)
gr2 <- fviz_contrib(z, choice = "var", axes = 2)
grid.arrange(gr1, gr2, ncol = 2)
```

Biplots V



La comunalidad h^2 correspondería a las sumas de las contribuciones de cada variable en cada una de las gráficas. Por ejemplo, SO tiene una comunalidad de 28 % + 30 % $\approx$ 50 % en las dos primeras componentes.

Ejemplo 3: Datos demográficos USA I

El archivo `USDemographics.csv` tiene 6 variables demográficas para un grupo de ciudades americanas:

- % de negros
- % de hispanos
- % de asiáticos
- edad mediana
- tasa de desempleo
- mediana del ingreso per capita

Ejemplo 3: Datos demográficos USA II

```r
options(width=130)
datos <- read.csv("https://raw.githubusercontent.com/jvega68/EA3/master/datos/Winston/USDemographics.csv")
head(datos)
```

```
  No_ciudad      Ciudad per_black per_hispanic per_asian median_age unemployment_rate income_percapita100
1         1 Albuquerque         3           35         2         32                 5                  18
2         2     Atlanta        67            2         1         31                 5                  22
3         3      Austin        12           23         3         29                 3                  19
4         4   Baltimore        59            1         1         33                11                  22
5         5      Boston        26           11         5         30                 5                  24
6         6   Charlotte        32            1         2         32                 3                  20
```

```r
X <- datos[,-c(1,2)]     # quita las dos primeras columnas de datos
(z <- prcomp(scale(X))) # la función scale estandariza los datos
```

```
Standard deviations (1, .., p=6):
[1] 1.3940729 1.2343421 1.1225968 0.7418743 0.6896837 0.4966845

Rotation (n x k) = (6 x 6):
                            PC1         PC2       PC3         PC4        PC5         PC6
per_black            0.06746821 -0.73005663 0.2362843 -0.09965218  0.1537334 -0.61079502
per_hispanic        -0.38424095  0.47572235 0.4203843  0.21397861 -0.3013892 -0.55919755
per_asian            0.42894849  0.46655289 0.1508061 -0.54373988  0.4688343 -0.24521514
median_age           0.55795618  0.07848629 0.1446866  0.77534418  0.2422613 -0.04173095
unemployment_rate   -0.11286739 -0.12220948 0.8351779 -0.07604415  0.1305849  0.50196475
income_percapita100  0.58288591 -0.04403077 0.1620124 -0.20417697 -0.7681029  0.01967245
```

El resumen de las componentes principales es:

Ejemplo 3: Datos demográficos USA III

```r
summary(z)
```

```
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6
Standard deviation     1.3941 1.2343 1.1226 0.74187 0.68968 0.49668
Proportion of Variance 0.3239 0.2539 0.2100 0.09173 0.07928 0.04112
Cumulative Proportion  0.3239 0.5778 0.7879 0.87961 0.95888 1.00000
```

- La primera componente principal es de la forma:

$$PC1 = 0.07z_{black} - 0.38z_{Hispanic} + 0.43z_{Asian} + 0.56z_{MedianAge} - 0.11z_{unrate} + 0.58z_{Income}$$

  Los coeficientes son las *cargas* o *loadings*. $PC1$ explica el 32 % de la variabilidad de los datos estandarizados.
- Esta variable representa a las personas asiáticas de edad con altos ingresos.

Ejemplo 3: Datos demográficos USA IV

- La segunda componente principal es de la forma:

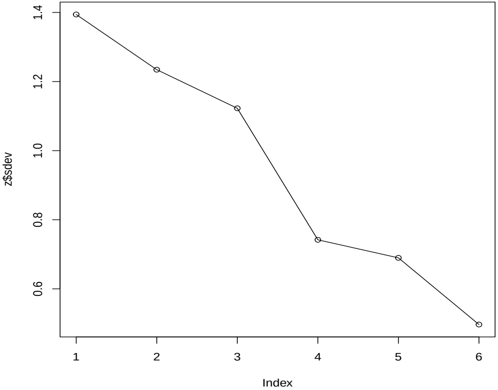
$$PC2 = -0.73 z_{black} + 0.48 z_{Hispanic} + 0.47 z_{Asian} + 0.08 z_{MedianAge} + 0.12 z_{unrate} - 0.04 z_{Income}$$

  Los coeficientes son las *cargas* o *loadings*. $PC1$ explica el 25 % de la variabilidad de los datos estandarizados y las dos primeras explican el 32 % +25 % = 57 % de la variabilidad.
- Esta variable representa un componente altamente negro, no hispánico y no asiático.

La gráfica de codo sugiere que tres variables pueden ser relevantes:

```
plot(z$sdev, type="o")
```

Ejemplo 3: Datos demográficos USA V



Ejemplo 4: Visualización de `iris` usando PCA I

- Podemos interpretar las componentes como proyecciones en las direcciones de máxima variabilidad de los datos.
- Las primeras CP son útiles para revelar estructura en los datos.

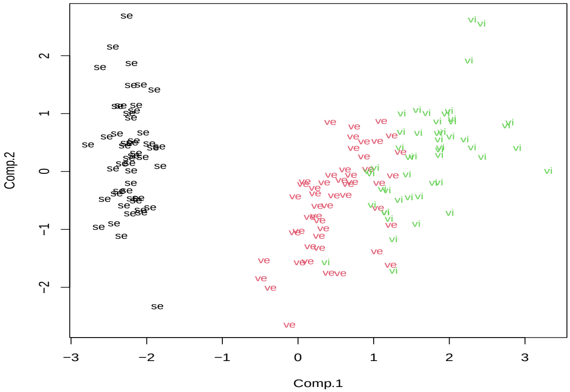
```
data("iris")
head(iris) #La última columna son las etiquetas de las variables

  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa

iris.pc <- princomp(iris[,-5],cor=T)
summary(iris.pc,loadings=T)

Importance of components:
                          Comp.1    Comp.2     Comp.3      Comp.4
Standard deviation     1.7083611 0.9560494 0.38308860 0.143926497
Proportion of Variance 0.7296245 0.2285076 0.03668922 0.005178709
Cumulative Proportion  0.7296245 0.9581321 0.99482129 1.000000000

Loadings:
             Comp.1 Comp.2 Comp.3 Comp.4
Sepal.Length  0.521  0.377  0.720  0.261
Sepal.Width  -0.269  0.923 -0.244 -0.124
Petal.Length  0.580        -0.142 -0.801
Petal.Width   0.565        -0.634  0.524
```

Ejemplo 4: Visualización de `iris` usando PCA II

- Graficamos las primeras dos componentes. La primera componente separa claramente los datos en relación al ancho del pétalo. En este caso ya lo sabíamos por conocer previamente los datos, pero en general una CP puede ayudar a identificar cuando los datos forman grupos.

```
plot(iris.pc$scores[,1:2], type="n")
text(iris.pc$scores[,1:2],cex=0.8,labels=substr(iris[,5],1,2),
     col = as.numeric(factor(iris[,5])))
```

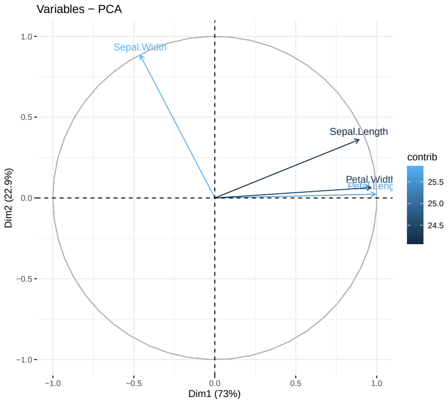
Ejemplo 4: Visualización de `iris` usando PCA III



Ejemplo 4: Visualización de `iris` usando PCA IV

- visualización de las componentes en el espacio de las dos primeras CP:

```
fviz_pca_var(iris.pc,col.var="contrib")
```

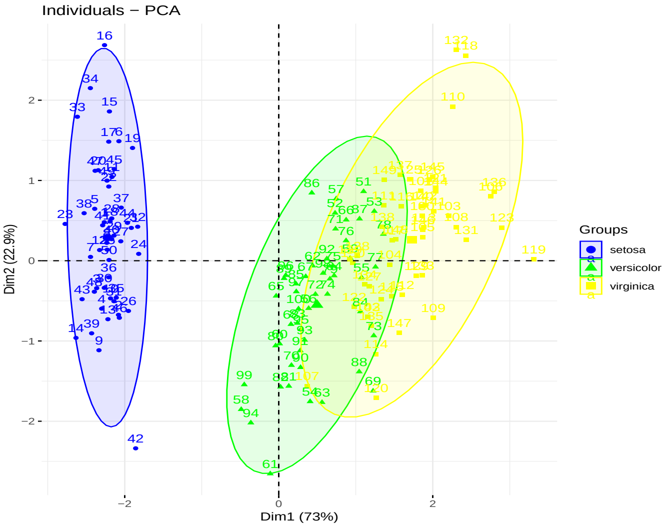


Ejemplo 4: Visualización de `iris` usando PCA V

- Podemos ver de otra manera las contribuciones de los diferentes grupos

```
fviz_pca_ind(iris.pc,habillage = iris[,5], palette=c("blue","green","yellow"),
             addEllipses = T)
```

Ejemplo 4: Visualización de `iris` usando PCA VI



Transformaciones de los datos para CP I

- Usualmente la matriz $\mathbf{X}_{n\times p}$ se transforma a una matriz de covarianzas Σ o de correlaciones ρ, que es lo único necesario para estimar las CP[1].
- Las CP pueden ser sensibles a outliers, por lo que se recomienda usar estimaciones robustas de Σ.
- Sin embargo, el análisis de CP **no es invariante** ante cambios de escala, por lo que las estimaciones obtenidas de ρ o de Σ pueden ser diferentes, y no hay transformación directa fácil entre ellas.

Ejemplo. [(Mardia 1979)]

Supongamos que se tienen tres variables dadas por:

- $x_1 =$ peso, dado en lbs
- $x_2 =$ altura, dado en pies
- $x_3 =$ edad (en años)

Supongamos que queremos realizar un cambio de escala de las variables a kg, cm y décadas, respectivamente.
Hay dos maneras de poder hacer el cambio de escala:

Transformaciones de los datos para CP II

1. Multiplicar las variables por los factores de conversión (0.453592, 30.48 y 0.1) y hacer PCA en la matriz obtenida de las variables rescaladas.
2. Llevar a cabo el PCA en la matriz de covarianzas de las variables originales y multiplicar los elementos de los componentes relevantes por los respectivos factores de conversión.

Entonces no se llega al mismo resultado por los dos métodos. En general, no se llegará al mismo resultado si se usa la matriz **S** o la matriz **R**. Algunas de las diferencias son:

- El porcentaje de varianza explicado por las CP de **R** diferirá del porcentaje basado en **S**.
- Los coeficientes de las CP de **R** difieren también de las de **S**
- Aún si se expresan las componentes de **R** en términos de las variables originales, no serán iguales a las de **S**

- En general, se recomienda:
  - Usar **S** cuando las unidades entre las variables son similares
  - Usar **R** cuando las varianzas en las variables originales son muy diferentes. En este caso, la interpretación de las CP puede ser más sencilla.Las CP de **R** sí son invariantes ante cambios de escala, porque lo es **R**.

□

Transformaciones de los datos para CP III

- Cuando las variables tienen diferentes unidades y escalas, se usa ρ para hacer las variables comparables.
- Usualmente se busca como primera aproximación para la interpretación de las CP, ver qué coeficientes de la respectiva componente son mayores (positivos o negativos), y también pensar en los cambios de signo de las variables (lo que define *contrastes* entre variables).

[1] Aunque se requieren los datos para poder graficar los *scores*, que son los valores específicos de las componentes para cada item en la muestra

Ejemplo 5: Datos de pobreza del CONEVAL I

- A partir de los datos publicados por el CONEVAL en el Anexo de entidades federativas, se puede utilizar el PCA para reducir la dimensión de un problema.
- El archivo `ConevalPobreza2016.csv` resume la información para 32 entidades de la República Mexicana y se cuenta con 15 variables que corresponden a diferentes poblaciones en situaciones de pobreza.
- En este caso todas las variables están en las mismas unidades (miles de personas), por lo que se puede utilizar directamente la matriz de covarianzas.
- Se considerarán tres casos: los datos transformados a logaritmos, los datos originales usando covarianzas y los datos originales, usando correlaciones, para compara los resultados.

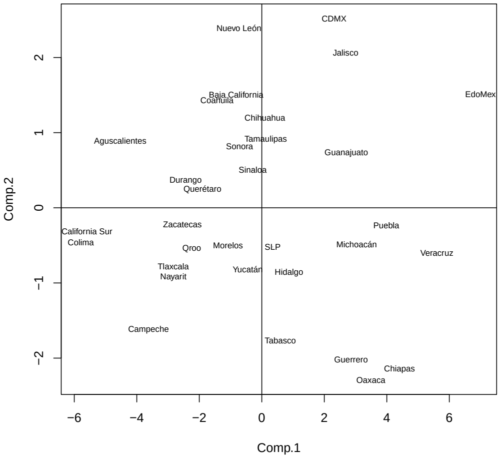
Ejemplo 5: Datos de pobreza del CONEVAL II

```r
X <- read.csv("https://raw.githubusercontent.com/jvega68/EA3/master/datos/ConevalPobreza2016.csv")
vars <- names(X)
vars #variables originales en el archivo
```

```
 [1] "Estado"
 [2] "Población.en.situación.de.pobreza"
 [3] "Población.en.situación.de.pobreza.moderada"
 [4] "Población.en.situación.de.pobreza.extrema"
 [5] "Población.vulnerable.por.carencias.sociales"
 [6] "Población.vulnerable.por.ingresos"
 [7] "Población.no.pobre.y.no.vulnerable"
 [8] "Población.con.al.menos.una.carencia.social"
 [9] "Población.con.al.menos.tres.carencias.sociales"
[10] "Rezago.educativo"
[11] "Carencia.por.acceso.a.los.servicios.de.salud"
[12] "Carencia.por.acceso.a.la.seguridad.social"
[13] "Carencia.por.calidad.y.espacios.en.la.vivienda"
[14] "Carencia.por.acceso.a.los.servicios.básicos.en.la.vivienda"
[15] "Carencia.por.acceso.a.la.alimentación"
[16] "Población.con.ingreso.inferior.a.la.línea.de.bienestar.mínimo"
[17] "Población.con.ingreso.inferior.a.la.línea.de.bienestar"
```

```r
names(X) <- paste0("x",1:17) #cambio los nombres para hacerlos más manejables
x <- X[,-c(1:2)] #quitamos los nombres de los estados y una varriable que es combinación lineal de dos que ya están
pc.coneval <- princomp(log(x)) #se considera el log de los datos.
plot(pc.coneval$scores[,1:2], pch = 16, cex = 0.1); abline(h = 0); abline(v = 0)
text(pc.coneval$scores[,1:2], labels = X[,1], cex = 0.7)
```

Ejemplo 5: Datos de pobreza del CONEVAL III



- Para interpretar las componentes, podemos ver los pesos de las variables:

```r
options(width=120)
summary(pc.coneval,loadings = T)
```

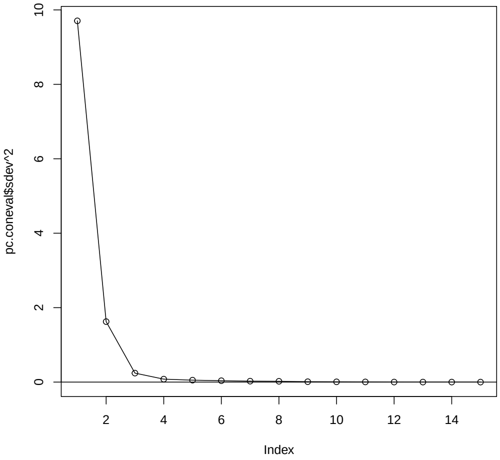
```
Importance of components:
                          Comp.1    Comp.2     Comp.3      Comp.4      Comp.5      Comp.6      Comp.7      Comp.8
Standard deviation     3.1153533 1.2747763 0.49164136 0.282307188 0.228391276 0.193604790 0.158762682 0.146467942
Proportion of Variance 0.8215432 0.1375573 0.02046033 0.006746207 0.004415448 0.003172839 0.002133598 0.001815938
Cumulative Proportion  0.8215432 0.9591005 0.97956083 0.986307036 0.990722484 0.993895323 0.996028921 0.997844859
                             Comp.9      Comp.10      Comp.11      Comp.12      Comp.13      Comp.14      Comp.15
Standard deviation     0.1032845533 0.0857215376 0.0695824723 0.0369712507 2.426869e-02 2.276924e-02 1.132292e-02
Proportion of Variance 0.0009029975 0.0006220076 0.0004098411 0.0001157029 4.985498e-05 4.388466e-05 1.085258e-05
Cumulative Proportion  0.9987478563 0.9993698639 0.9997797049 0.9998954078 9.999453e-01 9.999891e-01 1.000000e+00

Loadings:
    Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10 Comp.11 Comp.12 Comp.13 Comp.14 Comp.15
x3   0.260         0.247         0.195  0.211  0.303  0.214          0.257   0.367   0.508   0.403           0.146
x4   0.361 -0.370  0.253 -0.108 -0.284 -0.430 -0.316  0.178 -0.131   0.435                           0.199        
x5   0.175  0.276 -0.293         0.304 -0.109 -0.125  0.184  0.186   0.345  -0.619   0.299                  -0.104
x6   0.172  0.451  0.338  0.515 -0.225         0.227        -0.258   0.191  -0.208  -0.295          -0.217        
x7   0.133  0.529 -0.225  0.178 -0.289 -0.282 -0.381  0.112  0.267  -0.192   0.407   0.128                        
x8   0.239  0.113        -0.114  0.151  0.101                0.233                  -0.249  -0.288           0.819
x9   0.293 -0.156 -0.246 -0.108                     -0.126           0.110   0.302          -0.106  -0.822        
x10  0.254                              0.260 -0.263 -0.824          0.225                   0.102   0.205        
x11  0.247  0.239 -0.144 -0.354         0.306 -0.248  0.181 -0.706  -0.202                                        
x12  0.252               -0.214  0.185  0.198         0.245  0.370                  -0.619   0.200          -0.409
x13  0.267        -0.352 -0.184 -0.650  0.103  0.484         0.135          -0.209                   0.113        
x14  0.313 -0.401 -0.365  0.676  0.165  0.201         0.149         -0.141                           0.161        
x15  0.231        -0.119         0.375 -0.637  0.380 -0.240 -0.210  -0.238                   0.153   0.202        
x16  0.313 -0.135  0.455                      -0.207         0.207  -0.591  -0.324   0.179   0.210  -0.239        
x17  0.265         0.250                       0.175                           0.148   0.211  -0.780   0.159  -0.342
```

Ejemplo 4: Datos de pobreza del CONEVAL II

- Las dos primeras CP explican casi el 96 % de la variabilidad de los datos.
- La primera componente parece hacer un promedio ponderado sobre todas las variables de pobreza, generando un rankeo de las entidades de acuerdo a la multidimensionalidad de la pobreza.
- La segunda variable hace un contraste entre dos conjuntos de variables: $A = \{x_7, x_6, x_5\}$ que corresponde a la población que no es pobre y no es vulnerable, vulnerable por ingresos o vulnerable por carencias sociales, y el conjunto $B = \{x_{14}, x_4, x_9\}$ que son: carencia de servicios básicos de vivienda, pobreza extrema y al menos 3 carencias sociales. Entonces en conjunto parece que la segunda componente contrasta áreas urbanas con respecto a áreas rurales.
- La gráfica de codo nos da el tamaño de los eigenvalores:

```
plot(pc.coneval$sdev^2,type="o")
abline(h=0); abline(v=0)
```

Ejemplo 4: Datos de pobreza del CONEVAL III



Inferencia

Inferencia I

- La relación entre los eigenvalores de la matriz poblacional $\boldsymbol{\Sigma}$ y los de la matriz muestral **S** se da en el siguiente teorema:

Teorema (Anderson, 1963)

Sea $\boldsymbol{\Sigma} > 0$ con eigenvalores distintos y sea $\mathbf{U} \sim \frac{1}{m} W_p(\boldsymbol{\Sigma}, m)$ con descomposiciones $\boldsymbol{\Gamma}\Delta\boldsymbol{\Gamma}'$ y $\mathbf{PLP}'$ respectivamente. Entonces:

$$\sqrt{m}(\mathbf{l} - \boldsymbol{\lambda}) \to \mathcal{N}_p\left(0, 2\Delta^2\right)$$

donde $\mathbf{l} = diag(\mathbf{L})$ y $\boldsymbol{\lambda} = diag(\boldsymbol{\Delta})$.

Inferencia II

- Para probar la significancia de componentes grandes, se prueba H_0: últimos k eigenvalores son pequeños e iguales $(\lambda_{p-k+1} = \lambda_{p-k+2} = \cdots = \lambda_p)$ La estadística de prueba (LRT) es:

$$u = \left(n - \frac{2p+11}{6}\right)\left(k\log(\bar{\lambda}) - \sum_{i=1}^{p}\log(\lambda_i)\right) \sim \chi^2_v.$$

  donde $v = \frac{1}{2}(k-1)(k+2)$ y $\bar{\lambda} = \sum_{i=p-k+1}^{p}\frac{\lambda_i}{k}$.
- El procedimiento comienza evaluando $H_{02} : \lambda_{p-1} = \lambda_p$. Si se acepta, se evalúa $H_{03} : \lambda_{p-2} = \lambda_{p-1} = \lambda_p$, y así sucesivamente hasta que se rechace la primera hipótesis.