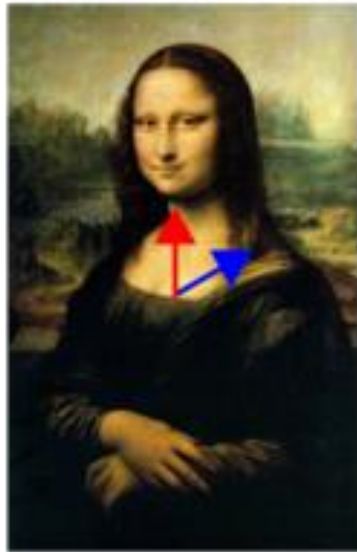


Reducción de dimensionalidad

Dora Suárez
Universidad del Rosario

Preliminares – Vectores propios

En álgebra lineal, los vectores propios o autovectores de un operador lineal son los vectores no nulos que, cuando son transformados por el operador, dan lugar a un múltiplo escalar de sí mismos, con lo que no cambian su dirección. “autoespacio”



$$\mathbf{A}\mathbf{v} = c\mathbf{v}, \quad \mathbf{v} \neq \mathbf{0}, c \in \mathbb{K},$$

Métodos no supervisionados

Métodos

Matriz de datos

$$X = \begin{bmatrix} x_{11} & \dots & x_{1p} \\ \dots & \dots & \dots \\ x_{n1} & \dots & x_{np} \end{bmatrix}$$

n : Número de datos

p : Número de variables

Matriz de datos centrada

$$X - 1'\bar{X} = \begin{bmatrix} x_{11} - \bar{X}_1 & \dots & x_{1p} - \bar{X}_p \\ \dots & \dots & \dots \\ x_{n1} - \bar{X}_1 & \dots & x_{np} - \bar{X}_p \end{bmatrix}$$

n : Número de datos

p : Número de variables

Matriz de datos estandarizada

$$\tilde{X} = \begin{bmatrix} \frac{x_{11} - \bar{X}_1}{s_1} & \dots & \frac{x_{1p} - \bar{X}_p}{s_p} \\ \dots & \dots & \dots \\ \frac{x_{n1} - \bar{X}_1}{s_1} & \dots & \frac{x_{np} - \bar{X}_p}{s_p} \end{bmatrix}$$

n : Número de datos

p : Número de variables

Datos

```
library(FactoClass)  
data("cafe")  
cafe
```



Descripción de los datos – Matriz de correlación

```
library(GGally)
```

```
plot(cafe, pch = 20) # Diagramas de dispersión
```

```
ggpairs(cafe, lower = list(continuous = "smooth"),  
        diag = list(continuous = "bar"), axisLabels  
= "none")
```


Descripción de los datos – Caras de Chernoff

```
library(aplpack)
```

```
faces(cafe)
```

```
faces(longley)
```

Variables latentes

Variables que no se pueden observar directamente

Ej.

Felicidad

Calidad de vida

Análisis Factorial

Análisis Factorial

Es un método estadístico utilizado para describir la variabilidad entre variables observadas y correlacionadas en términos de un número potencialmente menor de variables no observadas llamadas factores.

$$z_a = \sum_p l_{ap} F_p + e_a$$

z_a : Variable a-ésima estandarizada

l_{ap} : Peso del p-ésimo factor en la conformación de la a-ésima variable

e_a : Error asociado

¿Cuándo se utiliza?

Existe un gran número de variables

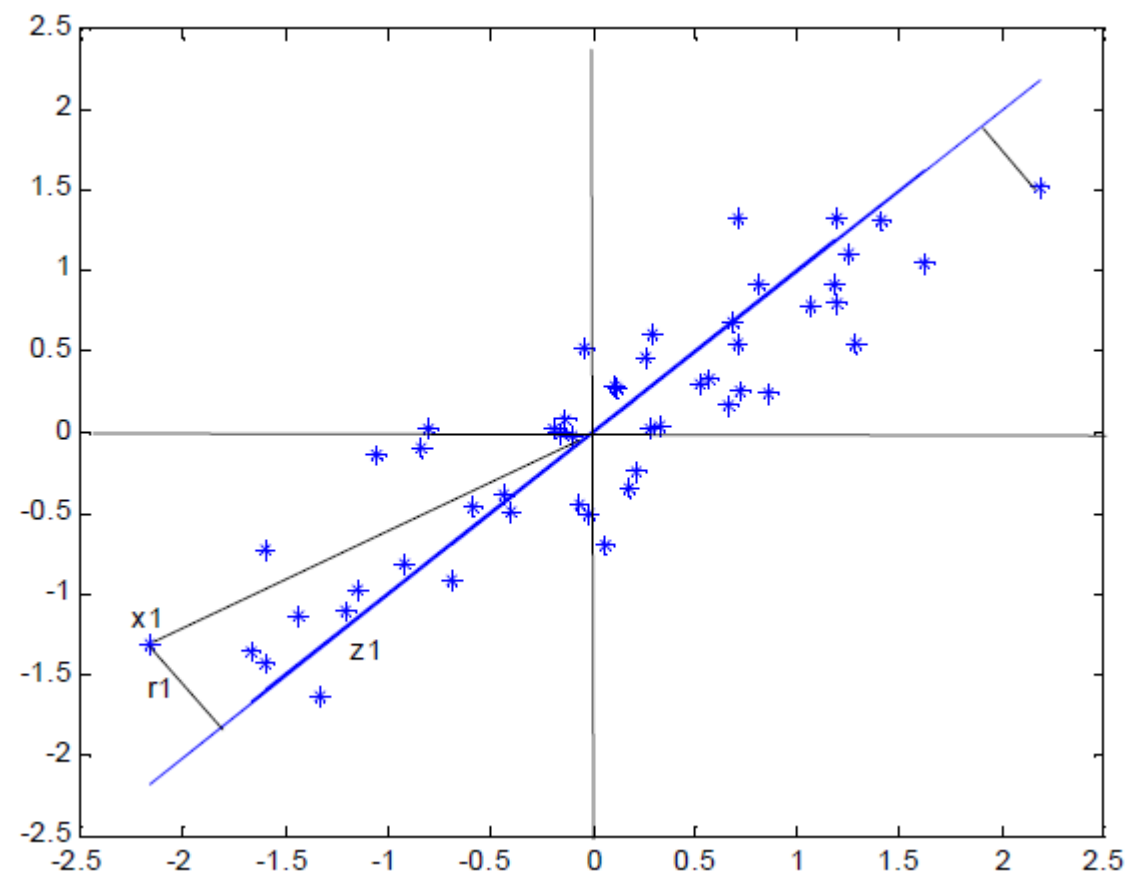
Existe alta correlación entre las variables

Las variables se distribuyen de forma normal*

Las relaciones entre las variables son lineales

Análisis de componentes principales

- Permite representar óptimamente en un espacio de dimensión pequeña, observaciones de un espacio general p -dimensional. En este sentido componentes principales es el primer paso para identificar posibles variables latentes o no observadas, que están generando la variabilidad de los datos.
- Permite transformar las variables originales, en general correladas, en nuevas variables incorreladas, facilitando la interpretación de los datos.



$$\mathbf{x}_i' \mathbf{x}_i = z_i^2 + r_i^2,$$

Minimizar

Maximizar (varianza)

$$\sum_{i=1}^n \mathbf{x}_i' \mathbf{x}_i = \sum_{i=1}^n z_i^2 + \sum_{i=1}^n r_i^2.$$

$$d_{ij}^2 = \mathbf{x}_i' \mathbf{x}_j$$

↑
Distancias entre los
puntos

$$\hat{d}_{ij}^2 = (z_i - z_j)^2$$

↑
Distancias entre los
puntos proyectados

Minimizar:

$$D = \sum_i \sum_j (d_{ij}^2 - \hat{d}_{ij}^2)$$

Objetivo

Encontrar una proyección en una dimensión $r < p$ tal que:

$$D = \sum_i \sum_j (d_{ij}^2 - \hat{d}_{ij}^2)$$

Sea mínimo y tal que los factores se haga la proyección sean linealmente independientes entre ellos

Objetivo

Primer componente principal

$$Var(\mathbf{z}_1) = \frac{1}{n} \mathbf{z}_1' \mathbf{z}_1 = \frac{1}{n} \mathbf{a}_1' \mathbf{X}' \mathbf{X} \mathbf{a}_1 = \mathbf{a}_1' \mathbf{S} \mathbf{a}_1$$

Para que la solución exista, se añade la restricción de que la norma del vector \mathbf{a} sea uno:

$$M = \mathbf{a}_1' \mathbf{S} \mathbf{a}_1 - \lambda(\mathbf{a}_1' \mathbf{a}_1 - 1) \quad \leftarrow \text{Función objetivo}$$

Calculo de los componentes

$$\frac{\partial M}{\partial \mathbf{a}_1} = 2\mathbf{S}\mathbf{a}_1 - 2\lambda\mathbf{a}_1 = 0$$

$$\mathbf{S}\mathbf{a}_1 = \lambda\mathbf{a}_1$$

Vector propio de la
matriz de covarianzas
asociado al valor propio
 λ

Ejemplo:

$$S = \begin{bmatrix} 0.35 & 0.15 & -0.19 \\ 0.15 & 0.13 & -0.03 \\ -0.19 & -0.03 & 0.16 \end{bmatrix}$$

$$\lambda_1 = 0.521$$

$$\lambda_2 = 0.113$$

$$\lambda_3 = 6.51 \times 10^{-3}$$

$$\begin{aligned} 0 &= |\mathbf{S} - \lambda \mathbf{I}| = \\ &= \left| \begin{bmatrix} 0.35 & 0.15 & -0.19 \\ 0.15 & 0.13 & -0.03 \\ -0.19 & -0.03 & 0.16 \end{bmatrix} - \begin{bmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{bmatrix} \right| = \\ &= 0,000382 - 0,0628\lambda + 0,64\lambda^2 - \lambda^3 \end{aligned}$$

$$\mathbf{S}\mathbf{a}_1 = \lambda_1 \mathbf{a}_1$$

$$\begin{bmatrix} 0.35 & 0.15 & -0.19 \\ 0.15 & 0.13 & -0.03 \\ -0.19 & -0.03 & 0.16 \end{bmatrix} \begin{bmatrix} a_{11} \\ a_{12} \\ a_{13} \end{bmatrix} = 0.521 \times \begin{bmatrix} a_{11} \\ a_{12} \\ a_{13} \end{bmatrix}$$

$$\begin{bmatrix} -0.171a_{11} + 0.15a_{12} - 0.19a_{13} \\ 0.15a_{11} - 0.391a_{12} - 0.03a_{13} \\ -0.19a_{11} - 0.03a_{12} - 0.361a_{13} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\{a_{11} = x, \ a_{12} = 0.427x, \ a_{13} = -0.562x\}$$

$$\mathbf{a}_1 = \begin{bmatrix} -0.817 \\ -0.349 \\ 0.459 \end{bmatrix}$$

$$Z_1 = -0.817X_1 - 0.349X_2 + 0.459X_3$$

Segundo componente

$$\phi = \mathbf{a}'_1 \mathbf{S} \mathbf{a}_1 + \mathbf{a}'_2 \mathbf{S} \mathbf{a}_2 - \lambda_1 (\mathbf{a}'_1 \mathbf{a}_1 - 1) - \lambda_2 (\mathbf{a}'_2 \mathbf{a}_2 - 1)$$

$$\frac{\partial \phi}{\partial \mathbf{a}_1} = 2\mathbf{S} \mathbf{a}_1 - 2\lambda_1 \mathbf{a}_1 = 0$$

$$\frac{\partial \phi}{\partial \mathbf{a}_2} = 2\mathbf{S} \mathbf{a}_2 - 2\lambda_2 \mathbf{a}_2 = 0$$

propio de la
covarianzas
valor propio
 λ

$$\mathbf{S}\mathbf{a}_1 = \lambda_1\mathbf{a}_1,$$

$$\mathbf{S}\mathbf{a}_2 = \lambda_2\mathbf{a}_2$$



Primer y segundo
vector propio asociados
a λ_1, λ_2

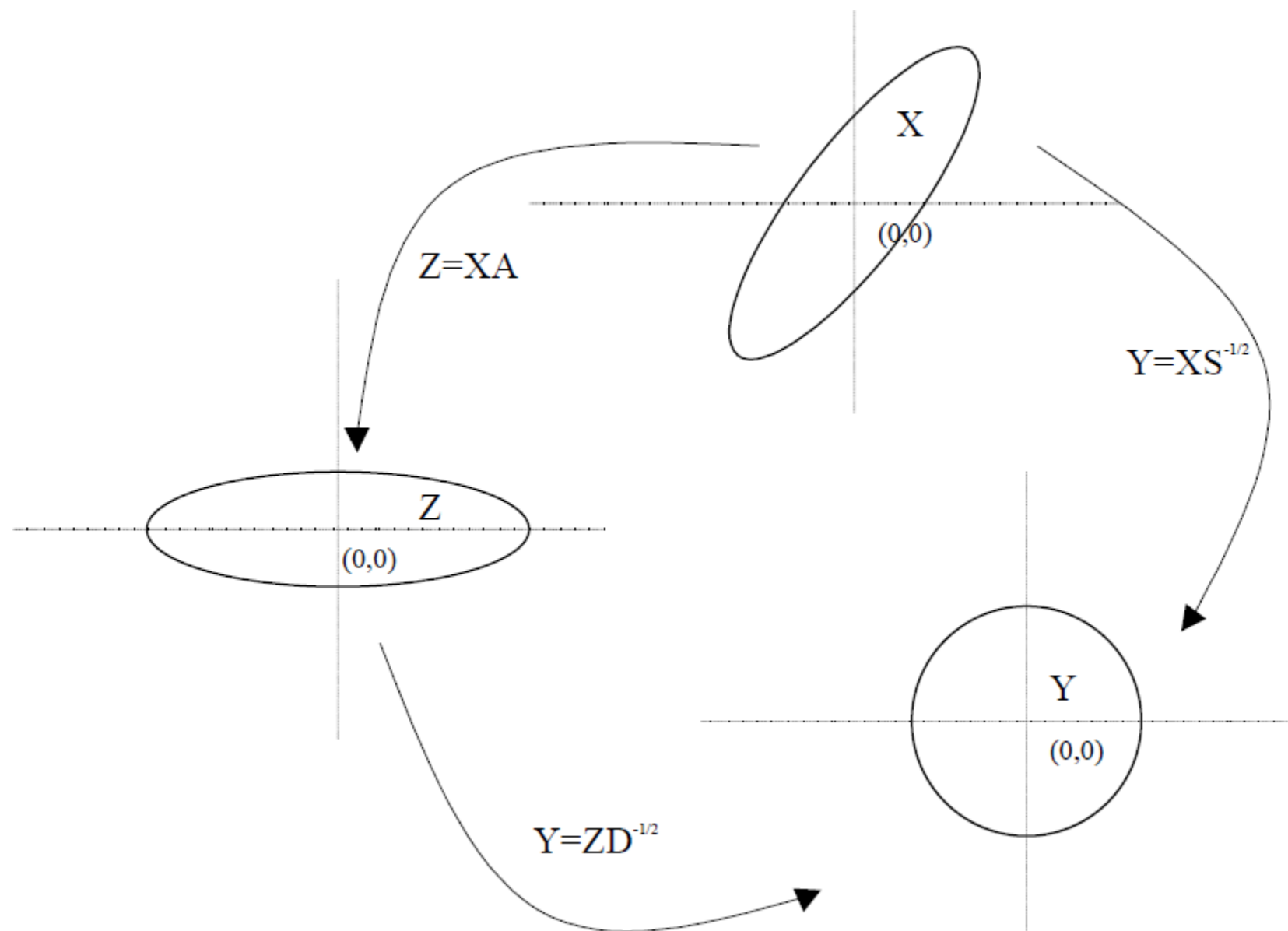
Generalización

- Para r componentes, los coeficientes asociados corresponden a los vectores propios
- La varianza de cada uno de los vectores es el valor propio asociado
- Se desea retener la mayor cantidad de variabilidad con el menor número de variables

Estandarización multivariada

$$\mathbf{Y}_c = \mathbf{ZD}^{-1/2} = \mathbf{XAD}^{-1/2}$$

Utilizar la matriz de correlaciones (ACP normado) tiene el efecto de controlar la escala, cuando se des-escalan las variables originales, magnitudes de las variables son más homogéneas y las proyecciones quedan con la misma escala de la variable original



Interpretaciones importantes

- El ángulo entre dos vectores en una proyección dada es interpretada como la correlación entre las dos variables sobre dicha proyección
- Las contribuciones son el % de aporte de variabilidad de una variable original en la formación de una componente específica
- Una forma de seleccionar el número de componentes puede basarse en la cantidad de variabilidad acumulada hasta determinado factor

Análisis de Componentes Principales en R

Análisis de correspondencias

Análisis de correspondencias

Es una técnica descriptiva o exploratoria cuyo objetivo es resumir una gran cantidad de datos en un número reducido de dimensiones, con la menor pérdida de información posible.

- El análisis de correspondencias es una técnica descriptiva para representar tablas de contingencia, es decir, tablas donde recogemos las frecuencias de aparición de dos o más variables cualitativas en un conjunto de elementos.


Análisis de correspondencias simples

Asociación entre categorías de columnas o filas: Medir la asociación de solo una fila o columna, para ver, por ejemplo, si las modalidades de una variable pueden ser combinadas.

Asociación entre categorías de filas y columnas: Estudiar si existe relación entre categorías de las filas y columnas.

C. ojos	Color del pelo					total
	rubio	pelirrojo	castaño	oscuro	negro	
claros	688	116	584	188	4	1580
azules	326	38	241	110	3	718
castaños	343	84	909	412	26	1774
oscuros	98	48	403	618	85	1315
total	1455	286	2137	1391	118	5387

Frecuencias relativas

$$\sum_{i=1}^I \sum_{j=1}^J f_{ij} = 1$$


$$\mathbf{R} = \mathbf{D}_f^{-1} \mathbf{F}$$

- R: Frecuencias relativas condicionales al total de la fila
- D: Matriz diagonal con las frecuencias relativas de las filas en la diagonal
- F: Matriz cuyas sumas por fila suman 1

Para obtener comparaciones razonables entre estas frecuencias relativas tenemos que tener en cuenta la frecuencia relativa de aparición del atributo que estudiamos.

Matriz de distancias – Distancia chi-cuadrado

$$D^2(\mathbf{r}_a, \mathbf{r}_b) = \sum_{j=1}^J \left(\frac{f_{aj}}{f_{a.}} - \frac{f_{bj}}{f_{b.}} \right)^2 \frac{1}{f_{.j}} = \sum_{j=1}^J \frac{(r_{aj} - r_{bj})^2}{f_{.j}}$$

$$D^2(\mathbf{r}_a, \mathbf{r}_b) = (\mathbf{r}_a - \mathbf{r}_b)' \mathbf{D}_c^{-1} (\mathbf{r}_a - \mathbf{r}_b)$$

Procedimiento

- Caracterizar las filas por sus frecuencias relativas condicionadas, y considerarlas como puntos en el espacio.
- Definir la distancia entre los puntos por la distancia χ^2 , que tiene en cuenta que cada coordenada de las filas tiene distinta precisión.
- Proyectar los puntos sobre las direcciones de máxima variabilidad, teniendo en cuenta que cada fila tiene un peso distinto e igual a su frecuencia relativa.

El procedimiento operativo para obtener la mejor representación bidimensional de las filas de la tabla de contingencia es:

- (1) Calcular la matriz $\mathbf{Z}'\mathbf{Z}$ y obtener sus vectores y valores propios.
- (2) Tomar los dos vectores propios, $\mathbf{a}_1, \mathbf{a}_2$, ligados a los mayores valores propios menores que la unidad de esta matriz.
- (3) Calcular las proyecciones $\mathbf{D}_f^{-1}\mathbf{F}\mathbf{D}_c^{-1/2}\mathbf{a}_i, i = 1, 2$, y representarlas gráficamente en un espacio bidimensional.

Generalización - ACM

Proximidad entre individuos en términos de parecidos: Dos individuos se parecen si pertenecen casi las mismas categorías. Es decir, dos individuos están próximos se han elegido globalmente las mismas categorías.

Generalización - ACM

Proximidad entre categorías de variables diferentes en términos de asociación: Son cercanos puesto que globalmente están presentes en los mismos individuos. Es decir, dos categorías están próximas si han sido elegidas globalmente por el mismo conjunto de individuos.

Generalización - ACM

Proximidad entre categorías de una misma variable en términos de parecido: (a) Son excluyente por construcción. (b) Si son cercanas es porque los individuos que las poseen presentan casi el mismo comportamiento en las otras variables.

Individuos	Género	Años	Ingreso
1	Mujer	5	Medio
2	Mujer	3	Alto
3	Hombre	4	Bajo
4	Mujer	1	Bajo
5	Mujer	2	Medio
6	Hombre	5	Alto
7	Mujer	2	Medio
8	Hombre	3	Bajo
9	Hombre	1	Alto
10	Mujer	4	Medio

Género		Años					Ingresos		
Mujer	Hombre	1	2	3	4	5	Bajo	Medio	Alto
1	0	0	0	0	0	1	0	1	0
1	0	0	0	1	0	0	0	0	1
0	1	0	0	0	1	0	1	0	0
1	0	1	0	0	0	0	1	0	0
1	0	0	1	0	0	0	0	1	0
0	1	0	0	0	0	1	0	0	1
1	0	0	1	0	0	0	0	1	0
0	1	0	0	1	0	0	1	0	0
0	1	1	0	0	0	0	0	0	1
1	0	0	0	0	1	0	0	1	0

MATRIZ DE BURT

$$Z'.Z =$$

Género

Años

Ingresos

Género		Años					Ingresos		
M	H	1	2	3	4	5	B	M	A
M	6 0	1	2	1	1	1	1	4	1
H	0 4	1	0	1	1	1	2	0	2
1	1 1	2	0	0	0	0	1	0	1
2	2 0	0	2	0	0	0	0	2	0
3	1 1	0	0	2	0	0	1	0	1
4	1 1	0	0	0	2	0	1	1	0
5	1 1	0	0	0	0	2	0	1	1
B	1 2	1	0	1	1	0	3	0	0
M	4 0	0	2	0	1	1	0	4	0
A	1 2	1	0	1	0	1	0	0	3

El procedimiento operativo para obtener la mejor representación bidimensional de las filas de la tabla de contingencia es:

- (1) Calcular la matriz $\mathbf{Z}'\mathbf{Z}$ y obtener sus vectores y valores propios.
- (2) Tomar los dos vectores propios, $\mathbf{a}_1, \mathbf{a}_2$, ligados a los mayores valores propios menores que la unidad de esta matriz.
- (3) Calcular las proyecciones $\mathbf{D}_f^{-1}\mathbf{F}\mathbf{D}_c^{-1/2}\mathbf{a}_i, i = 1, 2$, y representarlas gráficamente en un espacio bidimensional.