

Geometría euclidiana en estadística: métodos en ejes principales

CAMPO ELÍAS PARDO
Universidad Nacional de Colombia - Bogotá
Departamento de Estadística.
e-mail: cepardot@unal.edu.co *

21 de agosto de 2009

Índice

1. Introducción	2
2. Representaciones geométricas de una matriz	3
2.1. Ejemplo “café”	3
2.2. Nube de individuos (filas) N_I	4
2.2.1. Centro de gravedad	5
2.2.2. Centrado de la nube de individuos	5
2.2.3. Reducción de la nube de puntos (cambio de escala).	5
2.2.4. Inercia de la nube de individuos N_I	6
2.2.5. Búsqueda de nuevos ejes: cambio de base	7
2.3. La nube de variables (columnas) N_K	10
2.3.1. Significado del centro de gravedad y del centrado	10
2.3.2. Significado del reducido e imagen geométrica de las variables estandarizadas	11
2.3.3. Significado del ángulo entre dos flechas variables	11
2.4. Búsqueda de los ejes nuevos en \mathbb{R}^I	11
2.4.1. Círculo de correlaciones y ayudas para la interpretación.	11
3. Análisis en componentes principales generalizado o ponderado $ACP(\mathbf{X}, \mathbf{M}, \mathbf{D})$	13
3.1. Nube de filas en \mathbb{R}^K	13
3.2. Nube de columnas en \mathbb{R}^I y dualidad	14

* Conferencia para la V Jornada de Matemáticas Estadística y Didáctica. Universidad Pedagógica y Tecnológica de Colombia. Sede Duitama. Paipa, mayo 30 de 2008.

3.3.	Ayudas para la interpretación de las gráficas.	14
3.3.1.	Calidad de la representación.	14
3.3.2.	Contribución absoluta	15
3.4.	Elementos suplementarios o ilustrativos	15
3.5.	Fórmulas del ACP($\mathbf{X}, \mathbf{M}, \mathbf{D}$)	15
4.	El análisis de correspondencias simples (ACS) como un ACP($\mathbf{X}, \mathbf{M}, \mathbf{D}$)	16
4.1.	Equivalencia distribucional.	16
4.2.	Representación simultánea.	16
5.	Un programa para ejecutar los métodos en ejes principales	17

1. Introducción

La “escuela francesa de análisis de datos” utiliza la geometría euclidiana multidimensional para formalizar los métodos exploratorios multivariados básicos”. Estos métodos se utilizan para encontrar significado a la información contenida en tablas de datos grandes, por medio de su representación gráfica en espacios multidimensionales. Estas “gráficas” son abstractas pero permiten encontrar significado a la gran cantidad de cifras de una tabla mediante sus proyecciones sobre rectas y planos (Lebart, Morineau & Piron 1995).

En estas notas se muestra la lógica geométrica, apoyada con el algebra lineal, común a los métodos en ejes principales. La simbología que se adopta es la usual en muchos textos y es la siguiente: las letras mayúsculas en negrilla hacen referencia a matrices (\mathbf{A}), la minúsculas en negrilla a vectores (\mathbf{a}), las letras mayúsculas y minúsculas en itálica a variables (escalares) (A, a). En el caso de conjuntos se utiliza la misma letra mayúscula para indicar al conjunto y a su cardinalidad.

Como herramienta de cálculo se utiliza el lenguaje estadístico R (R Development Core Team 2008), junto con el paquete `ade4`: análisis de datos ecológicos y ambientales (*environnementales*) con procedimientos exploratorios euclidianos (Chessel, Dufour & Thioulouse 2004).

Obtención e instalación de R. Para que este documento tenga más vida se incluyen instrucciones en R para realizar los cálculos matriciales y obtener las gráficas. El símbolo \longrightarrow indica hacer clic en la *palabra siguiente* que debe aparecer en el menú que se presenta. Obviamente las versiones de R y los paquetes de este documento corresponden a la fecha de su edición y se podrán encontrar otras cuando se esté leyendo. Este documento da un camino posible en cada caso y supone que el lector utiliza el sistema operativo *Windows*, el buscador *Google*, y que dispone de una conexión a Internet de *banda ancha*.

- Abra *Google*
- Clic en *Mirror* y escoja uno.
- Clic en *Windows* \longrightarrow *base* \longrightarrow *R-2.7.0-win32.exe*
- Guarde el archivo en un directorio, por ejemplo *R* \longrightarrow *Ejecutar*.
- Responda a las preguntas del instalador (aceptando las sugerencias).

Instalación de paquetes. Si el equipo está en una red que requiere autenticación de *proxy* se debe hacer lo siguiente:

- Ubicar el el cursor en el acceso directo R 2.7.0 → Botón derecho del ratón → Propiedades.
- Agregar en destino el nombre del proxy y `http_proxy_user=ask`. Por ejemplo en la red de la Universidad Nacional - Bogotá
`http_proxy=http://proxy.unal.edu.co:8080/http_proxy_user=ask`
 Si R se arranca desde el menú de inicio se debe hacer el mismo procedimiento.

Para este documento se requiere instalar en R los paquetes `ade4`, `ade4TkGUI` y `scatterplot3d`, para instalarlos proceda así:

- Arranque R.
- Clic en *Paquetes* (en la barra de menú situada en la parte superior) → *Instalar paquetes*, aparece lista de *Mirrors*.
 En las redes que requieren autenticación escriba el nombre del usuario → la tecla **TAB** (\Leftrightarrow) contraseña → OK (R lo pide la primera vez en cada sesión cuando se desea instalar paquetes desde la Web).
- Seleccione un *Mirror* → OK, aparece lista de paquetes.
- Seleccione los paquetes (con la tecla *Control* oprimida cuando se quieren varios) `ade4`, `ade4TkGUI` y `scatterplot3d` → OK.

Ahora la consola de R espera comandos y los primeros usos, en la lectura de este texto, son los de calculadora graficadora y calculadora matricial. R diferencia mayúsculas y minúsculas, `<-` indica asignación (también se puede utilizar `=`) y el símbolo `#` se utiliza como comentario (el texto que aparece después de `#` se presenta pero R no lo interpreta).

2. Representaciones geométricas de una matriz

Para facilitar la visualización de los conceptos geométricos se utiliza un ejemplo pequeño, aunque los métodos en ejes principales son útiles en tablas de datos grandes.

2.1. Ejemplo “café”

En Duarte, Suarez, Moreno & Ortiz (1996) se presenta un experimento, donde se preparan tasas de café para detectar la influencia de la contaminación del grano con maíz y cebada. El experimento considera tres factores: agregado (sin, maíz, cebada), porcentaje del agregado (20% y 40%) y grado de tostación (clara, 8 minutos y oscura, 10 minutos). Entonces el experimento consta de 10 tratamientos y sobre las tasas de café de cada uno se miden propiedades químicas, físicas y sensoriales. En este ejemplo se utilizan solamente las variables físicas: *color*, *DA*: densidad aparente, *EA*: extracto acuoso (contenido de sólidos solubles). Los valores obtenidos para los 10 tratamientos se muestran en la tabla 1.

Lectura de la tabla 1 en R

- Cree del directorio *Cafe*.
- Copie la tabla en el *bloc de notes*.
- Edítela si es necesario y guárdela en directorio *Cafe* como `cafe.txt`.
- Arranque R.
- Cambie al directorio *Cafe*: Clic en *Archivo* (en la barra de menú) \rightarrow *Cambiar dir*
- Ubique el directorio \rightarrow *Aceptar*.
- `Y <- read.table('cafe.txt',header=TRUE,row.names=1);Y`
- `Y <- Y[-1]; Y # quitar la columna Cafes`

Tabla 1: Características físicas de los cafés

Cafés	Etiqueta	Color	DA	EA
ExcelsoClaro	CE	298.0	385.1	25.0
Claro40Maiz	CM40	361.0	481.3	41.0
Claro40Cebad	CC40	321.0	422.6	40.0
Claro20Maiz	CM20	335.0	444.3	33.0
Claro20Cebad	CC20	314.0	368.7	32.0
ExcelsoOscur	OE	186.0	346.6	28.0
Oscuro40Maiz	OM40	278.0	422.6	43.0
Oscuro40Ceba	OC40	238.0	403.0	42.0
Oscuro20Maiz	OM20	226.0	368.7	36.0
Oscuro20Ceba	OC20	210.0	368.7	35.0

La matriz de datos del ejemplo se identifica con \mathbf{Y} . Las 10 filas ($I = 10$) se representan como puntos en \mathbb{R}^3 ($K = 3$), imagen que se denomina *nube de filas* (figura 1). Las columnas de \mathbf{Y} representan a las variables, cada una se puede ver como un vector en \mathbb{R}^{10} . Esta geometría es abstracta pero tiene las mismas propiedades de la geometría en el plano (\mathbb{R}^2) y el espacio (\mathbb{R}^3). Los 3 vectores (*color*, *DA* y *EA*) constituyen la *nube de columnas*. En este ejemplo las filas representan “individuos” y las columnas variables continuas, entonces los vectores fila representados en \mathbb{R}^K constituyen la *nube de individuos* y los vectores columna en \mathbb{R}^I la nube de variables.

2.2. Nube de individuos (filas) N_I

En la nube de los I individuos en \mathbb{R}^K los ejes son las variables y las coordenadas de cada punto-individuo son los valores de las variables que asume (fila de \mathbf{Y}). En la figura 1 se muestra en 3D la nube de los 10 individuos del ejemplo café.

Construcción de la gráfica de la figura 1 en R. Para lo concerniente al ejemplo *Cafe* se supone que el lector tiene R abierto y que ha cambiado el directorio a la carpeta *Cafe*.

```
library(scatterplot3d) # carga el paquete
Y3D <- scatterplot3d(Y,main="Y") # grafica
cord2d<-Y3D$xyz.convert(Y) # convertir cordenadas 3D a 2D
text(cord2d,labels=row.names(Y), cex=0.8,col="blue",pos=4) # poner etiquetas
```

2.2.1. Centro de gravedad

Sobre la nube de individuos se definen el *centro de gravedad*, que generaliza el concepto de media; y la *inercia* como una medida de dispersión multivariada. El centro de gravedad de la nube de individuos es:

$$\mathbf{g} = \frac{1}{I} \sum_{i=1}^I \mathbf{y}_i = \frac{1}{I} \mathbf{Y}' \mathbf{1}_I$$

, donde $\mathbf{1}_I$ es un vector de I unos.

En R.

```
I <- nrow(Y); I
uno10 <- rep(1,10); uno10
g <- t(Y) %*% uno10 / I; g # igual a mean(Y)
# en la gráfica
Y3D$points3d(t(g), pch=19, col="darkgreen")
text(Y3D$xyz.convert(t(g)), labels="g", pos=4, col="darkgreen")
```

2.2.2. Centrado de la nube de individuos

Al restar a cada vector individuo el centro de gravedad \mathbf{g} se obtiene la matriz centrada \mathbf{Y}_C :

$$\mathbf{Y}_C = \mathbf{Y} - \mathbf{1}_I \mathbf{g}'$$

. Al representar \mathbf{Y}_C en \mathbb{R}^K el origen se traslada al centro de gravedad de N_I (figura 1).

En R.

```
Yc <- Y - uno10 %*% t(g); Yc
# en la gráfica
Yc3D <- scatterplot3d(Yc, main="Yc")
text(Yc3D$xyz.convert(Yc), labels=rownames(Yc), cex=0.8, col="blue", pos=4)
Yc3D$points3d(t(c(0,0,0)), pch=19, col="darkgreen")
text(Yc3D$xyz.convert(t(c(0,0,0))), labels="(0,0,0)", pos=1, col="darkgreen", cex=0.8)
```

2.2.3. Reducción de la nube de puntos (cambio de escala).

La matriz de varianzas y covarianzas, \mathbf{V} asociada a la tabla \mathbf{Y} es: $\mathbf{V} = \frac{1}{I} \mathbf{Y}'_C \mathbf{Y}_C$. En la diagonal de \mathbf{V} se tienen las varianzas, de modo que la suma de varianzas es igual a *traza*(\mathbf{V}).

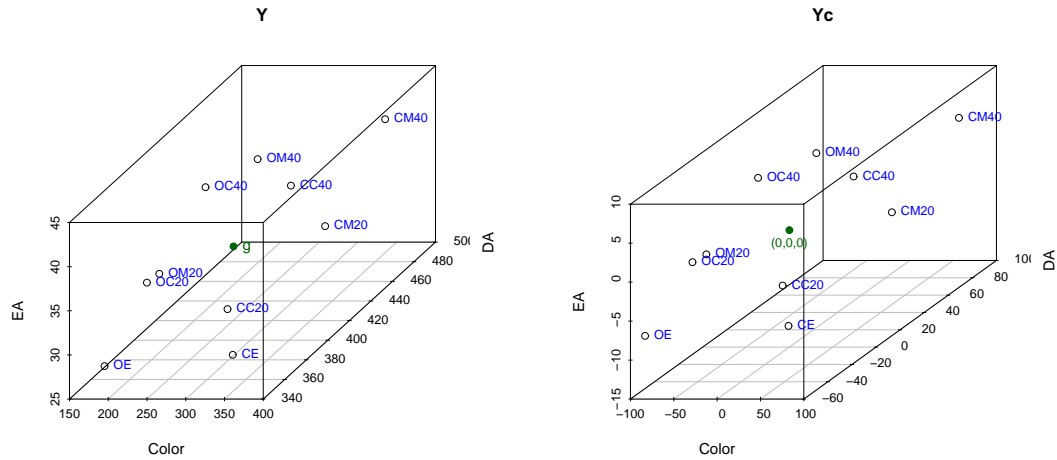
En R: `V <- t(Yc) %*% as.matrix(Yc) / I; V # = var(Y)*(I-1)/I`

La matriz normalizada \mathbf{X} (matriz de datos \mathbf{Y} centrada y reducida) tiene termino general:

$$x_{ik} = \frac{y_{ik} - \bar{y}_k}{\sigma_k}$$

, donde \bar{y}_k y σ_k , son la media y la desviación estándar de la variable k . En términos matriciales \mathbf{X} se obtiene mediante:

$$\mathbf{X} = \mathbf{Y}_C \mathbf{D}_\sigma^{-1}$$



Cafe	Color	DA	EA
CE	298	385.1	25
CM40	361	481.3	41
CC40	321	422.6	40
CM20	335	444.3	33
CC20	314	368.7	32
OE	186	346.6	28
OM40	278	422.6	43
OC40	238	403.0	42
OM20	226	368.7	36
OC20	210	368.7	35

Cafe	Color	DA	EA
CE	21.30	-16.06	-10.50
CM40	84.30	80.14	5.50
CC40	44.30	21.44	4.50
CM20	58.30	43.14	-2.50
CC20	37.30	-32.46	-3.50
OE	-90.70	-54.56	-7.50
OM40	1.30	21.44	7.50
OC40	-38.70	1.84	6.50
OM20	-50.70	-32.46	0.50
OC20	-66.70	-32.46	-0.50

Figura 1: Representación de la tabla de datos de café en 3D

donde $\mathbf{D}_\sigma = \text{diag}(\sigma_k)$. En la figura 2 se muestra \mathbf{X} para el ejemplo café. El valor que un individuo asume para una variable es la diferencia con respecto al promedio, pero ahora medida en el número de desviaciones estándar.

La matriz de correlaciones de las variables iniciales registradas en la tabla \mathbf{Y} , es la matriz de varianzas y covarianzas de \mathbf{X} :

$$\mathbf{V}_X = \frac{1}{I} \mathbf{X}' \mathbf{X}$$

Para el ejemplo la matriz de correlaciones se puede ver en la figura 4.

2.2.4. Inercia de la nube de individuos N_I .

La noción física de momento de inercia alrededor de un punto se utiliza como medida de dispersión de la nube de puntos, alrededor de su centro de gravedad y se denomina *inercia*.

Si cada individuo i se dota del peso p_i , la inercia de la nube es: $In(N_I) = \sum_{i=1}^I p_i d^2(\mathbf{i}, \mathbf{g})$

En el caso de pesos iguales para todos los individuos $p_i = 1/I$ y dado que su centro de gravedad se ha trasladado al origen, entonces:

$$In(N_I) = \frac{1}{I} \sum_{i=1}^I \sum_{k=1}^K x_{ik}^2 = \sum_{k=1}^K \frac{1}{I} \sum_{i=1}^I x_{ik}^2 = \sum_{k=1}^K \sigma_k^2 = \text{traza}(\mathbf{V}_X) = K$$

En el ejemplo la inercia total de nube es 3 (el número de variables) y está repartida equitativamente entre los tres ejes, es decir entre las tres variables.

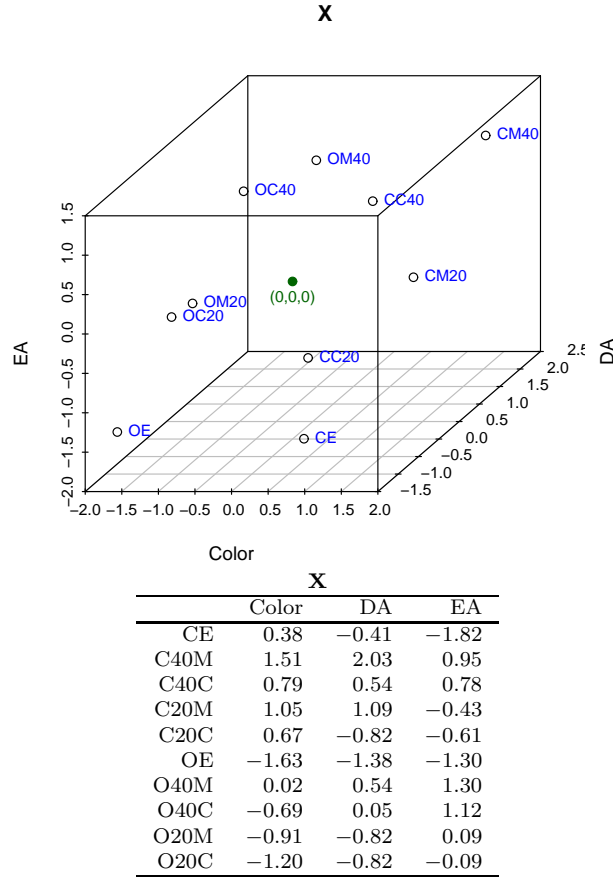


Figura 2: Nube de individuos asociada a los datos estandarizados de café

2.2.5. Búsqueda de nuevos ejes: cambio de base

El objetivo geométrico de los métodos en ejes principales es buscar un nuevo sistema de ejes de tal manera que la mayoría de la inercia se concentre en los primeros ejes. Es decir se trata de descomponer la inercia de la nube de puntos en ejes ortogonales ordenados, de tal manera que en el primer eje este la mayor inercia posible, en el segundo la mayor inercia residual posible, etc.

Se busca primero el eje de máxima inercia proyectada. Si se denota \mathbf{u} al vector unitario que la dirección del eje, la coordenada de un vector individuo \mathbf{x}_i sobre el eje es el producto punto:

$$\langle \mathbf{x}_i, \mathbf{u} \rangle = \mathbf{x}'_i \mathbf{u}$$

Las coordenadas de los I individuos sobre el eje conforman el vector \mathbf{Xu} . La inercia de la nube de individuos proyectada sobre el eje generado por \mathbf{u} es:

$$\sum_{i=1}^I p_i (\mathbf{Xu})^2 = \mathbf{u}' \mathbf{X}' \mathbf{D} \mathbf{X} \mathbf{u}$$

donde $\mathbf{D} = \text{diag}(p_i)$, es decir la matriz diagonal de los pesos de los individuos. Cuando $p_i = 1/I$ la inercia proyectada es:

$$\mathbf{u} \left(\frac{1}{I} \mathbf{X}' \mathbf{X} \right) \mathbf{u} \quad (1)$$

El problema es encontrar la dirección \mathbf{u} que maximice (1) sujeto a la restricción $\mathbf{u}' \mathbf{u} = 1$. Este problema se resuelve introduciendo un multiplicador de Lagrange λ , entonces se debe maximizar:

$$f(\mathbf{u}) = \mathbf{u} \left(\frac{1}{I} \mathbf{X}' \mathbf{X} \right) \mathbf{u} - \lambda (\mathbf{u}' \mathbf{u} - 1)$$

los puntos críticos son la solución de:

$$f'(\mathbf{u}) = 2 \left(\frac{1}{I} \mathbf{X}' \mathbf{X} \right) \mathbf{u} - 2\lambda \mathbf{u} = \mathbf{0}$$

es decir:

$$\left(\frac{1}{I} \mathbf{X}' \mathbf{X} \right) \mathbf{u} = \lambda \mathbf{u} \quad (2)$$

Las soluciones de (2) son los vectores propios asociados de los valores propios de $\left(\frac{1}{I} \mathbf{X}' \mathbf{X} \right)$. Cuál de los K valores propios escoger?. Premultiplicando por \mathbf{u} se obtiene de nuevo la cantidad que se quiere maximizar:

$$\mathbf{u}' \left(\frac{1}{I} \mathbf{X}' \mathbf{X} \right) \mathbf{u} = \lambda \mathbf{u}' \mathbf{u} = \lambda$$

y entonces la soluciones son los dos vectores propios unitarios asociados al valor propio mayor de $\left(\frac{1}{I} \mathbf{X}' \mathbf{X} \right)$, esta pareja se denota \mathbf{u}_1 y λ_1 .

Las coordenadas de los individuos sobre el eje generado por \mathbf{u} , denominado primer eje principal, se denotan por \mathbf{F}_1 y son:

$$\mathbf{F}_1 = \mathbf{X} \mathbf{u}_1$$

Para obtener el mejor plano para proyectar la nube de puntos se busca otro eje generado por un vector unitario \mathbf{u} ortogonal a \mathbf{u}_1 y que maximice la suma de cuadrados (1). El problema ahora es maximizar $\left(\frac{1}{I} \mathbf{X}' \mathbf{X} \right)$ sujeto a las restricciones $\mathbf{u}' \mathbf{u} = 1$ y $\mathbf{u}' \mathbf{u}_1 = 0$, entonces se introducen dos multiplicadores de Lagrange y la función a maximizar es:

$$f(\mathbf{u}) = \mathbf{u} \left(\frac{1}{I} \mathbf{X}' \mathbf{X} \right) \mathbf{u} - \lambda (\mathbf{u}' \mathbf{u} - 1) - \mu (\mathbf{u}' \mathbf{u}_1)$$

que tiene como primera derivada:

$$f'(\mathbf{u}) = 2 \left(\frac{1}{I} \mathbf{X}' \mathbf{X} \right) \mathbf{u} - 2\lambda \mathbf{u} - \mu \mathbf{u}_1 = \mathbf{0}$$

El segundo multiplicado debe ser $\mathbf{0}$, lo que se puede ver premultiplicando la ecuación anterior por \mathbf{u}'_1 . Entonces se obtiene de nuevo la ecuación (2) y la solución es ahora el vector propio, notado \mathbf{u}_2 , asociado al segundo valor propio más grande λ_2 .

Encontrar un subespacio 3D para proyectar la nube de puntos es, por el mismo procedimiento, introducir un tercer eje ortogonal a los dos primeros, que es el tercer vector propio \mathbf{u}_3 asociado al tercer valor propio más grande λ_3 de $\left(\frac{1}{I} \mathbf{X}' \mathbf{X} \right)$.

El rango r de $\left(\frac{1}{I} \mathbf{X}' \mathbf{X} \right)$ es el número de vectores columna linealmente independientes de \mathbf{X} (si $I > K$, más filas que columnas) y da el número de valores propios diferentes de cero. Los r vectores propios $\{\mathbf{u}_1, \dots, \mathbf{u}_s, \dots, \mathbf{u}_K\}$ constituyen una base ortonormada para el espacio de los

“individuos”, con las propiedades que se requieren para obtener las mejores proyecciones de la nube de puntos.

Las I coordenadas de los individuos sobre un eje factorial s , \mathbf{F}_s , constituyen los valores de una variable nueva denominada componente principal. Su varianza es:

$$\frac{1}{I} \sum_{i=1}^I (\mathbf{F}_s(i))^2 = \frac{1}{I} \mathbf{F}'_s \mathbf{F}_s = \mathbf{u}'_s \left(\frac{1}{I} \mathbf{X}' \mathbf{X} \right) \mathbf{u}_s = \lambda_s$$

La obtención de valores y vectores propios es un problema básico de algebra lineal, pero para su cálculo es necesario utilizar métodos numéricos.

En R

```
V <- t(X) %*% X / I; V # = cor(X)
lambda <- eigen(V)$values; round(lambda,3)
U <- eigen(V)$vectors
rownames(U) <- rownames(V)
colnames(U) <- c("Eje1", "Eje2", "Eje3"); round(U,2)
```

Los valores propios son: $\lambda_1 = 2.067$, $\lambda_2 = 0.822$ y $\lambda_3 = 0.111$ y los vectores propios: $\mathbf{u}_1 = (-0.58 \ -0.67 \ -0.46)'$, $\mathbf{u}_2 = (-0.57 \ -0.07 \ 0.82)'$ y $\mathbf{u}_3 = (0.58 \ -0.74 \ 0.35)'$. La matriz \mathbf{U} es $[\mathbf{u}_1 \ \mathbf{u}_2 \ \mathbf{u}_3]$.

Los tres vectores propios constituyen una base ortonormal, las coordenadas el nuevo sistema de ejes son $\mathbf{F} = \mathbf{XU}$. La figura 3 muestra la nube de puntos representada sobre los ejes nuevos y su proyección en el plano *Eje1 – Eje2* (primer plano factorial). Se incluyen las coordenadas sobre los nuevos ejes (matriz \mathbf{F}) y las ayudas para la interpretación del plano.

El plano construido con \mathbf{u}_1 y \mathbf{u}_2 , retiene una inercia igual a $\lambda_1 + \lambda_2$. En el ejemplo $2.067 + 0.822 = 2.889$ de inercia que es el $2.889 * 100 / 3 = 96.3\%$, es decir que se pierde poco al leer el primer plano factorial, en lugar de la representación en 3D, pero en cambio la lectura se hace mucho más fácil.

Gráfica del plano con R

```
plot(F[,1:2], las=1, sub="Plano 1-2", cex=0.8)
text(F[,1:2], label=rownames(F), col="blue", pos=4, cex=0.8)
abline(h=0, v=0, col="darkgrey")
```

El sentido de los ejes no tiene significado. Cada eje factorial se puede generar por uno de los dos vectores propios normados que definen su dirección \mathbf{u}_s o $-\mathbf{u}_s$. Escoger uno u otro significa cambiar el signo de las coordenadas, es decir un rotación de 180 grados. Esto implica que para un mismo análisis se pueden tener planos rotados, según el paquete y el procedimiento utilizado. Para comparar planos de diferentes análisis, que aparecen rotados, se pueden cambiar los signos de las coordenadas de manera conveniente, para graficarlos de nuevo.

Ayudas para la interpretación. Un plano factorial es una aproximación de la nube de puntos y como tal tendrá puntos bien representados, pero podrá contener puntos con mala calidad de proyección. Se utiliza el coseno al cuadrado que es, para un punto, el cuadrado de la relación entre la norma de la proyección sobre la norma en el espacio completo. La suma de los cosenos cuadrados sobre todos los ejes factoriales es 1. La distancia de un punto al origen, en el espacio

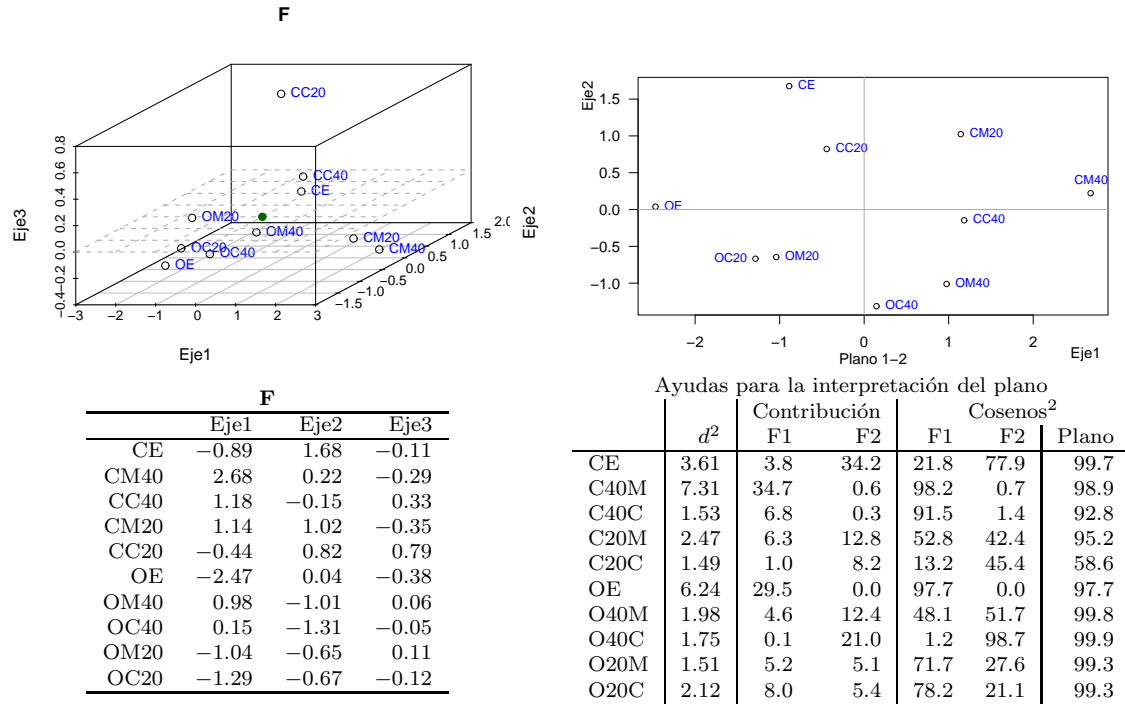


Figura 3: Nube de individuos sobre los ejes nuevos (rotada) y su proyección en el primer plano factorial

completo, es un buen complemento en la lectura de los ejes factoriales, esta dada por la norma del vector-individuo.

Para conocer los individuos que más influyen sobre la dirección de un eje factorial se utiliza el cociente de la contribución a la inercia del individuo sobre la inercia total del eje (valor propio). En la figura 3 se encuentran los valores de las ayudas a la interpretación para el ejemplo café.

2.3. La nube de variables (columnas) N_K

La nube de variables está constituida por K puntos en \mathbb{R}^I , las coordenadas de cada punto son las columnas de la matriz \mathbf{X} . El procedimiento es análogo al de la nube de individuos pero su interpretación es diferente.

2.3.1. Significado del centro de gravedad y del centrado

Sea \mathbf{Y}_k el vector columna asociado de la variable k , es decir la columna k de \mathbf{Y} , \mathbf{Y}_{C_k} y \mathbf{X}_k las columnas k de las matrices \mathbf{Y}_C y \mathbf{X} , respectivamente. El centrado de un vector \mathbf{Y}_k se logra mediante: $\mathbf{Y}_{C_k} = \mathbf{Y}_k - \bar{Y}_k \mathbf{1}_I$, donde \bar{Y}_k es la media de la variable k . El vector $\mathbf{1}_I$ es la primera bisectriz de \mathbb{R}^I y el vector $\bar{Y}_k \mathbf{1}_I$ es la homotecia (amplificación o reducción) de la primera bisectriz por la media de la variable k . Entonces todas las variables están representadas como puntos sobre la primera bisectriz (recta generada por el vector $\mathbf{1}_I$, que es un subespacio de dimensión 1). Finalmente, el vector diferencia, $\mathbf{Y}_k - \bar{Y}_k \mathbf{1}_I$, es una proyección de \mathbf{Y}_k sobre el subespacio ortogonal a la primera bisectriz. Esto significa que las variables centradas están todas en el subespacio ortogonal a la primera bisectriz y que en el proceso de centrado se pierde una dimensión.

2.3.2. Significado del reducido e imagen geométrica de las variables estandarizadas

Una columna k de \mathbf{X} se obtiene mediante $\mathbf{X}_k = \frac{1}{\sigma_k} \mathbf{Y}_{C_k}$. Esto hace que la norma al cuadrado del vector \mathbf{X}_k sea:

$$\mathbf{X}'_k \mathbf{X}_k = \frac{1}{\sigma_k^2} \mathbf{Y}'_{C_k} \mathbf{Y}_{C_k} = \frac{1}{\sigma_k^2} \frac{1}{I} (\mathbf{Y}_k - \bar{Y}_k \mathbf{1}_I)' (\mathbf{Y}_k - \bar{Y}_k \mathbf{1}_I) = \frac{1}{\sigma_k^2} \sum_{i=1}^I (y_{ik} - \bar{y}_k)^2 = I$$

Cambiando el producto escalar canónico, asociado a la matriz \mathbf{I}_{d_1} , matriz identidad de dimensión I , por el productor escalar asociado a $\frac{1}{I} \mathbf{I}_{d_1}$, se logra que la norma de cada una de las variables sea 1:

$$\langle \mathbf{X}'_k, \mathbf{X}_k \rangle_{\frac{1}{I} \mathbf{I}_{d_1}} = \mathbf{X}'_k \frac{1}{I} \mathbf{I}_{d_1} \mathbf{X}_k = \frac{1}{I} \mathbf{X}'_k \mathbf{X}_k = 1$$

y entonces, la representación de las variables estandarizadas se puede ver como flechas que terminan en el cascarón hiperesférico de radio 1 y centro en el origen.

2.3.3. Significado del ángulo entre dos flechas variables

El coseno del ángulo entre dos vectores variables es: $\langle \mathbf{X}_k, \mathbf{X}_{k'} \rangle_{\frac{1}{I} \mathbf{I}_{d_1}}$ y es exactamente igual al coeficiente de correlación entre las dos variables k y k' , ya que $\|\mathbf{X}_k\|_{\frac{1}{I} \mathbf{I}_{d_1}} = 1$ (norma=desviación estándar = 1).

Entonces el espacio de las variables estandarizadas es una representación de la matriz de correlaciones.

2.4. Búsqueda de los ejes nuevos en \mathbb{R}^I

Para poder leer la nube de variables se buscan los ejes de proyección que conservan lo más posible las normas de las variables originales. La solución se consigue con los valores y vectores propios de la matriz $\frac{1}{I} \mathbf{X} \mathbf{X}'$, esta matriz tiene K valores propios que son iguales a los valores propios de $\frac{1}{I} \mathbf{X}' \mathbf{X}$ y los restantes $I - K$ valores propios son cero. Cada valor propio λ_s tiene asociado el vector propio \mathbf{v}_s

2.4.1. Círculo de correlaciones y ayudas para la interpretación.

Un plano factorial de las variables se denomina círculo de correlaciones (figura 4), ya que es la proyección de la *hiperesfera* de correlaciones. La longitud de la proyección, sin error, de un vector-variable es 1. La calidad de la representación en el plano se observa visualmente al dibujar un círculo de radio uno en el plano factorial.

El vector de coordenadas sobre un eje s , \mathbf{G}_s , se obtiene mediante $\frac{1}{I} \mathbf{X}'_k \mathbf{v}_k$ y coincide con la correlación entre la variable k y el eje s . Los valores de la variable sintética representada por el eje s están en el vector F_s , que contiene las coordenadas de los individuos sobre el eje factorial s .

Esfera y círculo de correlaciones en R

```
G <- cor(Y,F);round(G,2)
G3D <- scatterplot3d(G,main="G",xlim=c(-1,1),ylim=c(-1,1),zlim=c(-1,1))
coord <- G3D$xyz.convert(G)
text(coord,labels=rownames(G), cex=0.8,col="blue",pos=4)
G3D$plane(0,0,0,col="darkgrey")
G3D$points3d(t(c(0,0,0)),pch=19,col="darkgreen")
cero <- G3D$xyz.convert(0, 0, 0)
for (eje in 1:3) {
  arrows(cero$x, cero$y, coord$x[eje], coord$y[eje], lwd = 2, length = 0.1)
}
# círculo de correlaciones
plot(G[,1:2],las=1,xlim=c(-1,1),ylim=c(-1,1),sub="Círculo de correlaciones")
text(G[,1:2],label=rownames(G),col="blue",pos=4,cex=0.8)
abline(h=0,v=0,col="darkgrey")
for (var in 1:3) arrows(0, 0, G[var,1],G[var,2],lty = 1,angle=15,length = 0.15)
```

La contribución de cada variable a un eje s sirve para seleccionar las variables que dan más significado al eje.

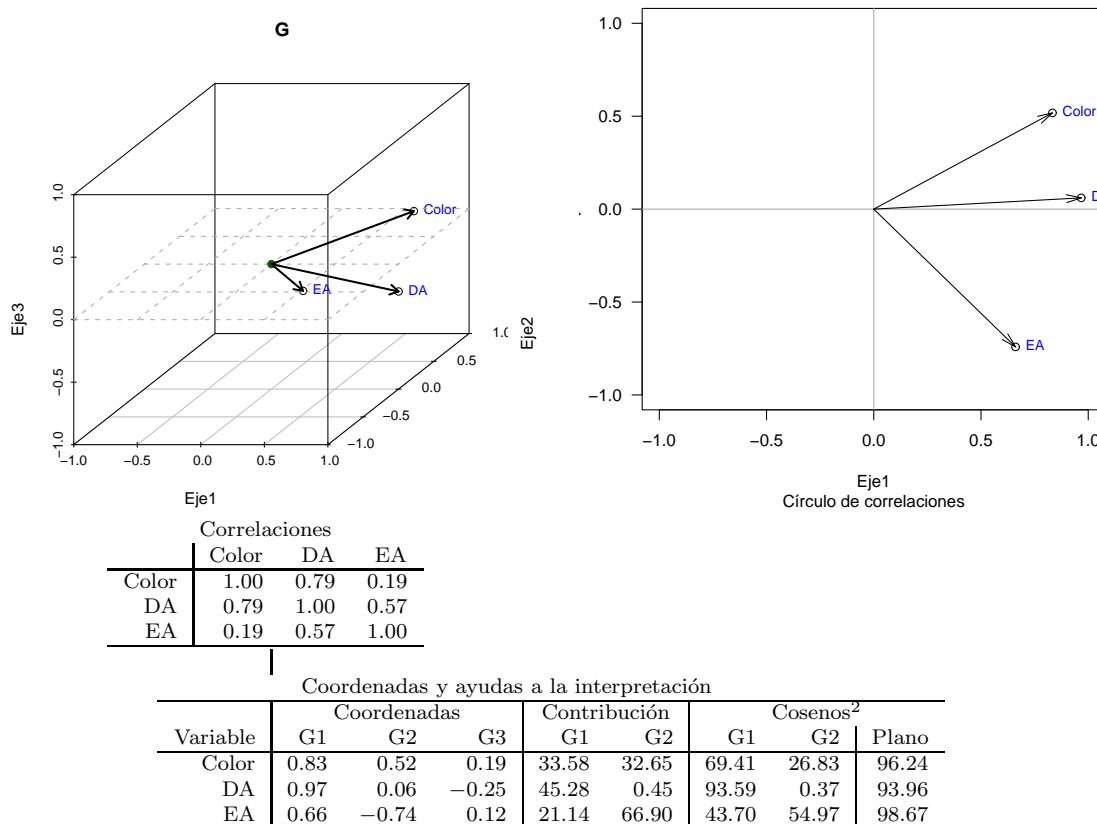


Figura 4: Esfera y círculo de correlaciones del ejemplo café

3. Análisis en componentes principales generalizado o ponderado $ACP(\mathbf{X}, \mathbf{M}, \mathbf{D})$

La sección 2 es el desarrollo del análisis en componentes principales (ACP) canónico. En el espacio de los individuos utiliza la identidad como matriz de producto interno. La generalización del ACP, que permite establecer pesos tanto para las filas como para las columnas, se constituye en un marco común para los métodos en ejes principales.

La notación $ACP(\mathbf{X}, \mathbf{M}, \mathbf{D})$ significa que se realiza un análisis en componentes principales de \mathbf{X} con pesos de las columnas dados en la matriz \mathbf{M} y pesos de las filas en \mathbf{D} , matrices que son diagonales. La matriz \mathbf{M} define el producto interno en el espacio de las filas (\mathbb{R}^K) y \mathbf{D} el producto interno en el espacio de las columnas (\mathbb{R}^I). De estos productos internos se derivan métricas o distancias, normas y ángulos que son euclidianos, aunque no canónicos. Las matrices \mathbf{M} y \mathbf{D} se suelen llamar *matrices de métricas*.

La matriz \mathbf{X} esta centrada con los pesos dados en \mathbf{D} , es decir $\mathbf{g} = \mathbf{X}'\mathbf{D}\mathbf{1}_I = \mathbf{0}$, siendo $\mathbf{1}_I$ un vector columna de I unos. Con la definición de la tripleta $(\mathbf{X}, \mathbf{M}, \mathbf{D})$ el ACP queda completamente determinando.

El ACP normado de la sección 2 se puede ver como $ACP(\mathbf{Y}_c, \text{diag}(\frac{1}{\sigma_k^2}), \frac{1}{I}\mathbf{I}_{d_I})$, la matriz a analizar es la matriz de datos centrados, la métrica $\text{diag}(\frac{1}{\sigma_k^2})$ es la matriz diagonal de las inversas de las varianzas y $\frac{1}{I}\mathbf{I}_{d_I}$ es la matriz de pesos, donde \mathbf{I}_{d_I} es la matriz identidad de dimensión I . El ACP normado también es el $ACP(\mathbf{X}, \mathbf{I}_{d_K}, \frac{1}{I}\mathbf{I}_{d_I})$ o el $ACP(\mathbf{Z}, \mathbf{I}_{d_K}, \mathbf{I}_{d_I})$, donde \mathbf{Z} es la matriz de datos normalizados dividida por \sqrt{I} , es decir que se diferencia de \mathbf{X} por la homotecia (multiplicación por) $\frac{1}{\sqrt{I}}$.

3.1. Nube de filas en \mathbb{R}^K

Con una métrica diagonal \mathbf{M} , la distancia entre dos filas i y l es:

$$d_{il}^2 = \sum_{k=1}^K m_k (x_{ik} - x_{lk})^2 \quad (3)$$

La matriz diagonal de pesos \mathbf{D} está involucrada en el cálculo del centro de gravedad y de la inercia. La inercia total es la suma ponderada de las distancias al cuadrado de los puntos-fila al centro de gravedad de la nube:

$$I = \sum_{i=1}^I p_i d^2(\mathbf{i}, \mathbf{0}) \quad (4)$$

Cada punto-fila contribuye a la inercia con el producto de su peso por el cuadrado de la distancia al centro de gravedad, el cual coincide con el origen de la representación. La contribución de un punto-fila a la inercia de la nube proyectada sobre un eje cualquiera es el peso de la fila por la coordenada al cuadrado sobre ese eje.

Lo que busca el ACP es encontrar un sistema de ejes \mathbf{u} , \mathbf{M} -ortonormales (normado $\mathbf{u}'\mathbf{M}\mathbf{u} = 1$ y ortogonal $\mathbf{u}'_s\mathbf{M}\mathbf{u}_t = 0, s \neq t$) de \mathbf{M} -inercia máxima.

Sea $F = \mathbf{X}\mathbf{M}\mathbf{u}$ el vector de las coordenadas sobre el eje definido por \mathbf{u} , entonces la inercia proyectada sobre el eje es $\sum p_i F^2 = \mathbf{u}'\mathbf{M}\mathbf{X}'\mathbf{D}\mathbf{X}\mathbf{M}\mathbf{u}$. Maximizar esta cantidad es encontrar la

dirección de mayor inercia proyectada, es decir la mejor proyección condicionada a los pesos de los puntos-fila. La solución es un vector propio \mathbf{M} -unitario \mathbf{u}_1 correspondiente al mayor valor propio λ_1 de la matriz $\mathbf{X}'\mathbf{D}\mathbf{X}\mathbf{M}$.

Se procede luego a encontrar una segunda dirección \mathbf{M} ortogonal a la primera que maximiza la inercia proyectada sobre ese eje, luego una tercera, \mathbf{M} ortogonal a las dos primeras. Encontrándose un nuevo sistema de ejes \mathbf{u}_s , con coordenadas F_s . Los \mathbf{u}_s son los vectores propios asociados a los valores propios de la matriz $\mathbf{X}'\mathbf{D}\mathbf{X}\mathbf{M}$, ordenados de mayor a menor.

La inercia proyectada sobre un subespacio de dimensión S es la suma de las inercias proyectadas sobre los ejes ortogonales que lo conforman. La aproximación de la proyección de la nube sobre ese subespacio se suele medir con el cociente de inercias:

$$\sum_{s=1}^S \lambda_s / \sum_{s=1}^K \lambda_s \quad (5)$$

3.2. Nube de columnas en \mathbb{R}^I y dualidad

En la nube de los K puntos columna en \mathbb{R}^I , las distancias están definidas por la matriz de métrica \mathbf{D} y los pesos están en la matriz \mathbf{M} . Se busca la dirección \mathbf{v} sobre la cual la $\sum_{k=1}^K m_k G_k^2$ sea máxima, donde \mathbf{G}_k es la proyección de la variable \mathbf{k} sobre la dirección \mathbf{v} . El vector de todas las proyecciones sobre \mathbf{v} es $\mathbf{X}'\mathbf{D}\mathbf{v}$ y la cantidad a maximizar es: $\mathbf{v}'\mathbf{D}\mathbf{X}\mathbf{M}\mathbf{X}'\mathbf{D}\mathbf{v}$ sujeta a la restricción $\mathbf{v}'\mathbf{D}\mathbf{v} = 1$. Sin embargo no es necesario realizar la diagonalización de la matriz $\mathbf{X}\mathbf{M}\mathbf{X}'\mathbf{D}$ pues los ejes factoriales de los dos espacios están relacionados. Sea μ_s un valor propio de la matriz $\mathbf{X}\mathbf{M}\mathbf{X}'\mathbf{D}$. Las relaciones entre los dos espacios son:

- los valores propios diferentes de cero son iguales en los dos espacios: $\mu_s = \lambda_s$
- el vector de coordenadas F_s sobre \mathbf{u}_s es un vector propio asociado a μ_s
- la \mathbf{M} -norma al cuadrado del vector F_s es λ_s
- el vector \mathbf{v}_s es igual a $\frac{1}{\sqrt{\lambda_s}} F_s$
- el vector de coordenadas G_s es igual a $\sqrt{\lambda_s} \mathbf{u}_s$

3.3. Ayudas para la interpretación de las gráficas.

3.3.1. Calidad de la representación.

Un indicador de esa calidad es el coseno que es una relación entre las magnitudes de la proyección y del vector original. Sin embargo se utiliza coseno cuadrado, ya que para un punto, la suma de los cosenos cuadrados sobre todos los ejes factoriales es 1 y su coseno cuadrado en un subespacio se obtiene sumando sus cosenos cuadrados sobre los ejes factoriales que generan al subespacio. Sobre un eje s el coseno cuadrado de un punto-fila i es:

$$\text{Cos}_s^2(i) = \frac{F_s^2(i)}{\|\mathbf{i}\|^2} \quad (6)$$

Las coordenadas de un vector fila \mathbf{i} están en la fila i de la matriz \mathbf{X} . El valor del coseno al cuadrado coincide con la relación de contribuciones del individuo i a la inercia: *contribución a la*

inercia proyectada sobre el eje s /contribución a la inercia total y se llama también contribución relativa.

3.3.2. Contribución absoluta

Otro aspecto que ayuda a la interpretación es identificar las filas que más estén contribuyendo a la inercia de una eje s . Este indicador se obtiene dividiendo la inercia proyectada del punto-fila i sobre la inercia total del eje, que es igual al valor propio. Se suele expresar en porcentaje y recibe el nombre de contribución absoluta de la fila i al eje s :

$$Con_s(i) = \frac{p_i F_s^2(i)}{\lambda_s} \quad (7)$$

Para las columnas se definen los mismos indicadores.

3.4. Elementos suplementarios o ilustrativos

Sobre los subespacios factoriales se pueden proyectar elementos que no participaron en el análisis, ya sean filas o columnas. También es posible proyectar filas *artificiales*, por ejemplo las filas promedio de grupos construidos a partir de variables nominales. La fórmulas de proyección son las mismas de los elementos activos. Es válido calcular la calidad de la representación de los elementos suplementarios. La contribución a la formación de los ejes es obviamente nula.

3.5. Fórmulas del ACP(\mathbf{X} , \mathbf{M} , \mathbf{D})

Un método específico, en ejes principales, queda completamente determinado definiendo las matrices: \mathbf{X} , que se obtiene de los datos mediante la transformación adecuada; \mathbf{M} , métrica en el espacio de las filas y pesos en el espacio de las columnas y \mathbf{D} , pesos en el espacio de las filas y métrica en el espacio de las columnas. Entonces las fórmulas del método específico se obtienen reemplazando en las fórmulas ACP(\mathbf{X} , \mathbf{M} , \mathbf{D}) (tabla 2).

Tabla 2: Fórmulas del ACP(\mathbf{X} , \mathbf{M} , \mathbf{D})

Espacio	\mathbb{R}^K	\mathbb{R}^I
Nube	N_I	N_K
Coordenadas	filas de \mathbf{X}	columnas de \mathbf{X}
Pesos	diagonal de \mathbf{D}	diagonal de \mathbf{M}
Métrica	\mathbf{M}	\mathbf{D}
Inercia	$traza(\mathbf{X}'\mathbf{D}\mathbf{X}\mathbf{M})$	$traza(\mathbf{X}\mathbf{M}\mathbf{X}'\mathbf{D})$
Valor propio	λ_s	λ_s
Vector propio	\mathbf{u}_s	\mathbf{v}_s
Coordenadas factoriales	$F_s = \mathbf{X}\mathbf{M}\mathbf{u}_s = \sqrt{\lambda_s}\mathbf{v}_s$	$G_s = \mathbf{X}'\mathbf{D}\mathbf{v}_s = \sqrt{\lambda_s}\mathbf{u}_s$
Fórmulas de transición	$F_s = \frac{1}{\sqrt{\lambda_s}}\mathbf{X}\mathbf{M}\mathbf{G}_s$ $F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_{k=1}^K x_{ik}m_k G_s(k)$	$G_s = \frac{1}{\sqrt{\lambda_s}}\mathbf{X}'\mathbf{D}\mathbf{F}_s$ $G_s(k) = \frac{1}{\sqrt{\lambda_s}} \sum_{i=1}^I x_{ik}d_i F_s(i)$

4. El análisis de correspondencias simples (ACS) como un ACP($\mathbf{X}, \mathbf{M}, \mathbf{D}$)

El ACS se utiliza para comparar perfiles fila y columna de una tabla de contingencia (TC) y para visualizar las asociaciones de las categorías fila y columna (ver por ejemplo Cabarcas & Pardo (2001)). A partir de la TC se obtiene la tabla de frecuencias relativas notada por \mathbf{F} , las notaciones del ACS se derivan de esta última tabla.

El ACS de la TC se obtiene mediante el ACP de la tabla \mathbf{X} cuyo término general está dado por (8), usando $\mathbf{D} = \mathbf{D}_{\mathbf{I}} = \text{diag}(f_{i\cdot})$, como pesos de las filas y matriz de métrica en el espacio de las columnas, y $\mathbf{M} = \mathbf{D}_{\mathbf{K}} = \text{diag}(f_{\cdot k})$, como pesos de las columnas y matriz de métrica en el espacio de las filas.

$$x_{ik} = \frac{f_{ik} - f_{i\cdot} \cdot f_{\cdot k}}{f_{i\cdot} \cdot f_{\cdot k}} \quad (8)$$

Todas las fórmulas del ACS se pueden derivar de las fórmulas correspondientes al ACP generalizado (ver tabla 2). La \mathbf{M} -distancia al cuadrado entre dos filas i y l y la \mathbf{D} -distancia al cuadrado entre las columnas k y q de \mathbf{X} son:

$$d^2(i, l) = \sum_{k=1}^K \frac{1}{f_{\cdot k}} \left(\frac{f_{ik}}{f_{i\cdot}} - \frac{f_{lk}}{f_{l\cdot}} \right)^2; \quad d^2(k, q) = \sum_{i=1}^I \frac{1}{f_{i\cdot}} \left(\frac{f_{ik}}{f_{\cdot j}} - \frac{f_{iq}}{f_{\cdot q}} \right)^2 \quad (9)$$

Las expresiones de (9) son la distancias ji-cuadrado entre los perfiles fila y columna, respectivamente, derivados de la tabla \mathbf{F} . Esta distancia tiene dos propiedades muy importantes para la interpretación de las salidas del ACS:

4.1. Equivalencia distribucional.

El ACS no se modifica si se unen dos puntos que tienen el mismo perfil, el peso del punto colapsado es la suma de los pesos de los puntos que se unen. Esto permite unir filas o columnas con perfiles parecidos, para simplificar las tablas originales. Esta propiedad hace que el ACS sea robusto ante la “arbitrariedad” en la conformación de las categorías de una variable en un estudio.

4.2. Representación simultánea.

En el ACS las relaciones de transición son:

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_{k=1}^K \frac{f_{ik}}{f_{i\cdot}} G_s(k) ; \quad G_s(k) = \frac{1}{\sqrt{\lambda_s}} \sum_{i=1}^I \frac{f_{ik}}{f_{\cdot k}} F_s(i) \quad (10)$$

Las relaciones de la expresión (10) además de hacer posible técnicamente la representación simultánea permiten su interpretación. Un punto fila se ubica en el promedio, ponderado por los valores de su perfil, de las coordenadas de todos los puntos columna, dilatado por el inverso de la norma del vector propio. Entonces el punto fila se ubica más cerca de los puntos de las columnas que más contribuyen a su perfil. La dilatación hace que la asociación más destacada sea también la más alejada.

En Pardo & Del-Campo (2007) se puede ver el ejemplo del ACS de la TC que clasifica las manzanas de Bogotá según localidades (19) y estratos (6).

5. Un programa para ejecutar los métodos en ejes principales

El ADE4 (Chessel et al. 2004) es un programa de más de 20 años de construcción y contiene, además de los métodos en ejes principales básicos, otros desarrollados para aplicaciones en ecología y medio ambiente, pero que se usan en otras áreas de aplicación. Los investigadores que desarrollaron el ADE4 decidieron programarlo también en R, como el paquete `ade4` y le construyeron un menú (paquete `ade4TkGUI`).

Los métodos están implementados sobre un ACP generalizado, que denominan diagrama de dualidad (*dudi*). Las funciones para los métodos básicos son:

dudi.pca: análisis en componentes principales (ACP)

dudi.coa: análisis de correspondencias simples (ACS)

dudi.acm: análisis de correspondencias múltiples (ACM)

dudi.pco: análisis en coordenadas principales (ACO)

Quien desee estudiar y utilizar estos métodos puede consultar la documentación disponible en mi página docente:

`www.docentes.unal.edu.co/cepardot/docs`

y utilizar el `ade4TkGUI` el cual se arranca en R mediante:

```
library(ade4TkGUI)
ade4TkGUI()
```

Para empezar se pueden obtener los resultados del ejemplo “café”, así:

- Clic en PCA.
- En Input data frame \rightarrow set \rightarrow Y.
- En Out put dudi name escribir `acp` (por ejemplo).
- Clic en Submit (abajo) \rightarrow aparece el “histograma de valores propios” y una ventana espera el número de ejes que se desean retener para el análisis \rightarrow 2 \rightarrow Choose.
- Clic en 3: `acp$eig` para ver el histograma de valores propios.
- Clic en 2: `acp$li` para ver el primer plano factorial de los cafés.
- Clic en 4: `acp$co` para ver el círculo de correlaciones.

Referencias

- Cabarcas, G. & Pardo, C.-E. (2001), ‘Métodos estadísticos multivariados en investigación social’, *Simposio de Estadística*.
*<http://www.docentes.unal.edu.co/cepardot/docs/SimposiosEstadistica/>
- Chessel, D., Dufour, A.-B. & Thioulouse, J. (2004), ‘The ade4 package-I- One-table methods’, *R News* 4, 5–10.

Duarte, R., Suarez, M., Moreno, E. & Ortiz, P. (1996), 'Análisis multivariado por componentes principales, de cafés tostados y molidos adulterados con cereales', *Cenicafé* **47**(2), 65–76.

Lebart, L., Morineau, A. & Piron, M. (1995), *Statistique exploratoire multidimensionnelle*, Dunod, Paris.

Pardo, C. E. & Del-Campo, P. C. (2007), 'Combinación de métodos factoriales y de análisis de conglomerados en R: el paquete FactoClass ', *Revista Colombiana de Estadística* **30**(2), 231–245.

*<http://www.matematicas.unal.edu.co/revcoles/>

R Development Core Team (2008), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

*<http://www.R-project.org>