Universidad de Sevilla

Facultad de Matemáticas

Trabajo Fin de Grado:

Técnicas de reducción de la dimensión para datos con mixtura de variables cualitativas y cuantitativas: PCAMIX.

Autor: Elena Domínguez Domínguez

Tutor: Juan Manuel Muñoz Pichardo

18 de junio de 2018

Resumen

En este trabajo se aborda el problema de la reducción de la dimensión de conjuntos de datos con mixtura de variables cualitativas y cuantitativas. Con este objetivo se describe el método PCAMIX como una unión de análisis de componentes principales (ACP) y análisis de correspondencia múltiple (ACM), pudiendo tener variables cualitativas y cuantitativas y aplicar un solo método para analizarlas.

En el primer capítulo se explica con detalle teóricamente este método y su aplicación a un conjunto de datos, el procedimiento original propuesto por Kiers [6], una reformulación de ese procedimiento a través de una Descomposición en Valores Singulares (DVS) y la rotación varimax en este método.

A continuación, en el segundo capítulo se incluye la descripción de la implementación del algoritmo en R-Programm. Se recoge un breve manual del paquete `PCAmix` y sus funciones asociadas, incluyendo una ilustración de todas estas funciones aplicadas a un conjunto de datos.

Abstract

In this project the problem of the reduction of the dimension of data sets with mixture of qualitative and quantitative variables is addressed. With this objective, the PCAMIX method as a union of principal component analysis (PCA) and multiple correspondence analysis (MCA), and may have qualitative and quantitative variables and apply a single method to analyze them.

In the first chapter this method and its application to a data set is explained in detail, the original procedure proposed by Kiers [6], a reformulation of that procedure through a Decomposition in Singular Values (DSV) and the varimax rotation in this method.

Next, the description of the implementation of the algorithm in R-Programm is included in the second chapter. A brief manual of the `PCAmix` package and its associated functions is collected, including an illustration of all these functions applied to a data set.

Índice general

Introducción

El principal objetivo de este trabajo es describir el método del PCAMIX como método de reducción de la dimensión para datos con mixtura de variables cualitativas y cuantitativas, incluyendo ilustraciones y su obtención a través de R-Programm.

Kaiser[5] introdujo el criterio varimax para lograr la estructura simple por rotación en el análisis de componentes principales (ACP). Este criterio tiene como objetivo maximizar la suma entre las columnas de los elementos al cuadrado de la matriz de cargas. Esta matriz juega un papel importante en la interpretación de los resultados, ya que contiene las correlaciones entre las variables y las componentes principales. La idea es rotar la matriz de cargas y las componentes principales estandarizadas para agrupar las variables de manera que: tener altas cargas en las mismas componentes, cargas moderadas en otras componentes e insignificantes en las componentes restantes. Por lo tanto, la varianza se redistribuye a lo largo de los ejes recién rotados.

A pesar de la estrecha relación entre ACP y el análisis de correspondencia múltiple (ACM), la rotación en ACM no ha recibido mucha atención. Sin embargo, Kiers [6] maneja la rotación ortogonal en el contexto general de rotación en PCAMIX, un método de componentes principales para la mezcla o mixtura de variables cualitativas y cuantitativas. Las rotaciones ortogonales se usan mucho en la práctica, ya que las cargas rotadas ortogonalmente pueden ser interpretadas directamente como correlaciones entre las variables y las componentes rotadas.

Recordemos primero los objetivos del ACP y ACM por separado, ya que estos

son casos particulares de PCAMIX en el que solo se usan variables cuantitativas o variables cualitativas, respectivamente.

Análisis de componentes principales

El objetivo del ACP es determinar un espacio de dimensión reducida que represente adecuadamente un conjunto de n observaciones p-dimensionales.

El ACP pretende sustituir las variables originales por un número pequeño de combinaciones lineales de éstas, de manera que las nuevas variables sean incorreladas y que la pérdida de información sea lo menos posible entre el conjunto de datos inicial y el conjunto de datos transformados.

Primero aclararemos la notación que vamos a utilizar. Sea $\mathbf{A}$ un matriz de dimensiones $n \times p$ la escribiremos por filas y columnas de manera que:

$$\mathbf{A} = \begin{bmatrix} \underline{a}'_{[1]} \\ \underline{a}'_{[2]} \\ \vdots \\ \underline{a}'_{[n]} \end{bmatrix} = \left[\underline{a}_1 | \underline{a}_2 | \cdots | \underline{a}_p\right]$$

Sea $\underline{x}_{[1]}, \ldots, \underline{x}_{[n]}$ una realización muestral de un vector $\underline{X} = (X_1, \ldots, X_p)' \sim (\mu, \Sigma)$. Sea $\mathbf{X} = \begin{bmatrix} \underline{x}'_{[1]} \\ \vdots \\ \underline{x}'_{[n]} \end{bmatrix}$ la matriz de datos muestrales y $\underline{x}_1, \ldots, \underline{x}_p$ las columnas de $\mathbf{X}$. Calculamos $\hat{\Sigma}$ la matriz de varianzas-covarianzas muestrales:

$$\hat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} (\underline{x}_i - \overline{\underline{x}})(\underline{x}_i - \overline{\underline{x}})'$$

Sean $\hat{\underline{e}}_1, \hat{\underline{e}}_2, \ldots, \hat{\underline{e}}_p$ los autovectores unitarios ortogonales de $\hat{\Sigma}$, entonces se define la

k-ésima componente principal muestral como:

$$\underline{\hat{e}}_k'\underline{X}, \quad k = 1, \ldots, p.$$

Luego, tenemos que las variables originales $\underline{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix}$ se han transformado en unas nuevas variables $\underline{Y} = \begin{pmatrix} \underline{\hat{e}}_1'\underline{X} \\ \vdots \\ \underline{\hat{e}}_p'\underline{X} \end{pmatrix}$. Y los valores correspondientes a la k-ésima componente principal son:

$$\underline{y}_{(k)} = \begin{pmatrix} y_{1k} \\ \vdots \\ y_{nk} \end{pmatrix} = \begin{pmatrix} \underline{\hat{e}}_k'\underline{x}_1 \\ \vdots \\ \underline{\hat{e}}_k'\underline{x}_n \end{pmatrix} = \mathbf{X}\underline{\hat{e}}_k, \quad k = 1, \ldots, p$$

Así el coeficiente de correlación muestral entre $\underline{y}_{(k)}$ y $\underline{x}_{(h)}$ viene dado por

$$r(\underline{y}_{(k)}; \underline{x}_{(h)}) = \frac{\hat{e}_{kh}\sqrt{\hat{\lambda}_h}}{\hat{\sigma}_{\underline{x}_{(k)}}}$$

siendo $\hat{\lambda}_h$ el autovalor h-ésimo de $\hat{\Sigma}$ y $\hat{\sigma}_{\underline{x}_{(k)}}$ la raíz cuadrada de la varianza de muestral de $\underline{x}_{(k)}$.

Sea $\underline{z}_{[1]}, \ldots, \underline{z}_{[n]}$ los datos estandarizados de la realización muestral. Definimos la matriz $\mathbf{Z}$ de dimensiones $n \times p$ como

$$\mathbf{Z} = \begin{bmatrix} \underline{z}_{[1]}' \\ \underline{z}_{[2]}' \\ \vdots \\ \underline{z}_{[n]}' \end{bmatrix} = [\underline{z}_1 | \underline{z}_2 | \cdots | \underline{z}_p]$$

Las componentes principales basadas en los datos estandarizados vienen dadas por $\underline{\tilde{e}}_k'\mathbf{Z}$, donde $\underline{\tilde{e}}_k$ es el autovector unitario de R asociado al k-ésimo autovalor ($\tilde{\lambda}_k$) mayor de dicha matriz, siendo

$$R = \frac{1}{n-1}\sum_{i=1}^{n} \underline{z}_{[i]}\underline{z}_{[i]}' = \frac{1}{n-1}\mathbf{Z}'\mathbf{Z}$$

la matriz de correlación muestral de los datos originales.

Además, dado que $\mathrm{R}\underline{\tilde{e}}_k = \tilde{\lambda}_k\underline{\tilde{e}}_k$ entonces

$$\frac{1}{n-1}\mathbf{Z}'\mathbf{Z}\underline{\tilde{e}}_k = \tilde{\lambda}_k\underline{\tilde{e}}_k \quad \Rightarrow \quad \frac{1}{n-1}\mathbf{Z}\mathbf{Z}'\underbrace{(\mathbf{Z}\underline{\tilde{e}}_k)}_{\underline{\tilde{w}}_k} = \tilde{\lambda}_k(\mathbf{Z}\underline{\tilde{e}}_k)$$

$$\Rightarrow \quad (\tilde{\lambda}_k, \underline{\tilde{w}}_k) \quad \text{autovalor-autovector de la matriz } \mathbf{Z}\mathbf{Z}'.$$

Además, las puntuaciones o scores de las CP en los datos muestrales son

$$\underline{\tilde{e}}_k'\underline{z}_{[i]} = \underline{z}_{[i]}'\underline{\tilde{e}}_k \quad k = 1, \ldots, p,\ i = 1, \ldots, n$$

Por tanto, las puntuaciones en la k-ésima CP es $\mathbf{Z}\underline{\tilde{e}}_k = \underline{\tilde{w}}_k$.

En general, se tiene

$$\tilde{\mathbf{W}} = \mathbf{Z}[\underline{\tilde{e}}_1|\ldots|\underline{\tilde{e}}_p] = [\underline{\tilde{w}}_1|\ldots|\underline{\tilde{w}}_p].$$

Análisis de correspondencia múltiple

El análisis de correspondencia múltiple (ACM) es una técnica de análisis de datos para datos categóricos nominales, utilizado para detectar y representar estructuras subyacentes en un conjunto de datos. Se basa en representar los datos como puntos en un espacio euclídeo de baja dimensión. ACM puede ser visto como una extensión del análisis de correspondencia simple (AC) para un conjunto grande de variables categóricas.

El ACM se lleva a cabo aplicando el algoritmo de AC a la matriz de indicadores o tabla de Burt formada a partir de estas variables categóricas. Una matriz de indicadores es una matriz de "individuos × variables", donde las filas representan a los individuos y las columnas son indicadores binarios que representan a las categorías de las variables. Analizar la matriz de indicadores permite la representación directa de los individuos como puntos en un espacio geométrico. La tabla de Burt es la matriz simétrica que contiene las tabulaciones cruzadas para cada pareja de variables categóricas y es el análago de la matriz de covarianzas para variables continuas.

En la aproximación mediante la matriz indicadora, las asociaciones entre variables son representadas gráficamente como "mapas", facilitando la interpretación de la estructura de los datos. Igual que en el análisis de componentes principales, el primer eje es la dimensión más importante, el segundo eje la segunda más importante, y así sucesivamente, en relación a la cantidad de inercia explicada. El número de ejes a retener se determina calculando valores propios modificados.

Ya que el ACM se entiende como una extensión del AC, recogeremos a continuación este método.

Consideremos dos variables categóricas A con n modalidades, y B con p modalidades. Consideremos la tabla de contingencia asociada a una muestra de tamaño N. Sea $F = (f_{ij})_{i=1,\dots,n,j=1,\dots,p}$ la matriz de frecuencias relativas asociada a la tabla de contingencia $(n \times p)$:

$$F = \begin{pmatrix} f_{11} & \cdots & f_{1j} & \cdots & f_{ip} \\ \vdots & & \vdots & & \vdots \\ f_{i1} & \cdots & f_{ij} & \cdots & f_{ip} \\ \vdots & & \vdots & & \vdots \\ f_{n1} & \cdots & f_{nj} & \cdots & f_{np} \end{pmatrix} \quad \begin{pmatrix} f_{1.} \\ \vdots \\ f_{i.} \\ \vdots \\ f_{n.} \end{pmatrix}$$

$$(f_{,1} \quad \cdots \quad f_{.j} \quad \cdots \quad f_{.p})$$

Sean D_n y D_p las matrices diagonales asociadas a las distribuciones marginales

$$D_n = diag(f_{1.}, \ldots, f_{n.})$$
$$D_p = diag(f_{,1}, \ldots, f_{.p})$$

Las distribuciones condicionadas por filas y columnas son las filas de las matrices

$$\text{Perfiles filas:} \quad M_r = D_n^{-1} F$$
$$\text{Perfiles columnas:} \quad M_c = D_p^{-1} F'$$

El objetico del AC es representar los perfiles filas y columnas en un número menor de dimensiones (generalmente 2) de forma que los perfiles próximos en la métrica $ji-cuadrado$ tengan representaciones próximas en la distancia euclídea. El objetivo es el mismo que en el ACP con la peculiaridad de que los "elementos" del AC son distribuciones y la distancia que se utiliza es la distancia $ji-cuadrado$. Esta distancia es una distancia entre distribuciones de probabilidad.

Cada uno de estos métodos están también implementados en R, el ACP mediante la función `princomp` y AC mediante la función `ca`.

Luego, una vez explicados brevemente estos métodos, en el primer capítulo del trabajo introduciremos el método PCAMIX de manera teórica y también hablaremos de la rotación ortogonal en este método. Y una vez explicado teóricamente este método pasaremos a ilustrarlo con el software R, haciendo previamente un manual de como funciona este método en el programa.

Capítulo 1

El método PCAMIX

Kiers [6] introduce la rotación ortogonal en PCAMIX, un método de componentes principales para una mixtura de variables cualitativas y cuantitativas. El PCAMIX incluye el análisis de componentes principales ordinario y el análisis de correspondencia múltiple, ya que estamos tratando con variables cualitativas y cuantitativas. En este capítulo presentamos el método PCAMIX donde las componentes principales y las cargas al cuadrado son obtenidas de una descomposición en valores singulares. Las cargas de las variables cuantitativas y las coordenadas principales de las categorías de las variables cualitativas son también obtenidas directamente.

Definimos primero la notación usada para presentar el método.

- n denota el número de observaciones, p_1 el número de variables cuantitativas, p_2 el número de variables cualitativas y $p = p_1 + p_2$ el número total de variables.
- $\mathbf{z}_j$ es el vector columna que contiene los valores estandarizados de las n observaciones de la variable j si la j-ésima variable es cuantitativa.
- $\mathbf{G}_j$ es la matriz de dimensiones $n \times m_j$ de indicadores de la variable j si ésta es cualitativa con m_j categorias, y $\mathbf{D}_j$ es matriz diagonal de dimensiones $m_j \times m_j$ de frecuencias observadas. Veremos un poco más adelante un ejemplo de como se construyen estas matrices.

- $m = m_1 + \cdots + m_{p_2}$ es el número total de categorias de las p_2 variables cualitativas.

- $\mathbf{G} = (\mathbf{G}_1|\cdots|\mathbf{G}_j|\cdots|\mathbf{G}_{p_2})$ es la matriz de dimensiones $n \times m$ de las variables indicadoras de las m categorías de las p_2 variables cualitativas y
  $\mathbf{D} = \text{diag}\,(\mathbf{D}_1, \ldots, \mathbf{D}_j, \ldots, \mathbf{D}_{p_2})$ la matriz de dimensiones $m \times m$ diagonal de las frecuencias de las m categorías.

- $\mathbf{J} = \mathbf{I}_n - \mathbf{1}\mathbf{1}'/n$ es el operador de centralizazción donde $\mathbf{I}_n$ es la matriz identidad $n \times n$ y $\mathbf{1}$ es el vector de orden n de unos. Este operador actúa sobre una matriz $\mathbf{X}$ $n \times p$ y nos queda:

$$(\mathbf{I}_n - \mathbf{1}\mathbf{1}'/n)\,\mathbf{X} = \mathbf{X} - \mathbf{1}\underbrace{\frac{1}{n}\mathbf{1}'\mathbf{X}}_{\overline{\underline{x}}'} = \begin{bmatrix} \underline{x}_1' \\ \vdots \\ \underline{x}_n' \end{bmatrix} - \mathbf{1}\overline{\underline{x}}' = \begin{bmatrix} \underline{x}_1' - \overline{\underline{x}}' \\ \vdots \\ \underline{x}_n' - \overline{\underline{x}}' \end{bmatrix}$$

Veámos un ejemplo de la construcción de las matrices $\mathbf{G}_j$ y $\mathbf{D}_j$.

Ejemplo 1.1. Supongamos que tenemos $n = 6$ observaciones de una variable cualitativa, como es el estado civil de una persona. Esta variable tiene $m = 4$ modalidades: soltero = 1, casado = 2, divorciado = 3 y viudo = 4. Y se observan $n = 6$ casos, (Soltero, Divorciado, Casado, Soltero, Divorciado, Divorciado), así las frecuencias observadas de estas modalidades son:

Estado civil	Frecuencia observada
Casado	1
Soltero	2
Divorciado	3
Viudo	0

Las matrices $\mathbf{G}$ y $\mathbf{D}$ asociadas a esta variable son:

$$\mathbf{G} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \quad \mathbf{D} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

En las dos secciones siguientes, se dará dos formulaciones del método PCAMIX y se destacará sus principales diferencias.

1.1. Procedimiento original del PCAMIX

Se describe en este apartado el procedimiento PCAMIX propuesto por Kiers [6]. Supongamos que k es el número de componentes seleccionadas en PCAMIX. El procedimiento determina la matriz $\mathbf{X}$ de dimensiones $n \times k$ que contiene los valores estandarizados de las componentes, la varianza de cada componente y la matriz $\mathbf{C}$ de dimensiones $n \times k$ de las cargas al cuadrado. Estas cargas están definidas como las correlaciones al cuadrado de las variables cuantitativas con las componentes PCAMIX y como el coeficiente de correlación para las variables cualitativas. Este procedimiento se lleva a cabo en los siguientes pasos:

1. Para $j = 1, \ldots, p$: se calcula la matriz de cuantificación de dimensiones $n \times n$, $\mathbf{S}_j$ con:

$$\begin{cases} \mathbf{S}_j = \frac{1}{n}\mathbf{z}_j\mathbf{z}_j' & \text{si la variable } j \text{ es cuantitativa.} \\ \mathbf{S}_j = \mathbf{J}\mathbf{G}_j\mathbf{D}_j^{-1}\mathbf{G}_j'\mathbf{J} & \text{si la variable } j \text{ es cualitativa.} \end{cases}$$

2. Calcular la matriz de dimensiones $n \times n$, $\mathbf{S} = \sum_{j=1}^{p} \mathbf{S}_j$. En el caso de que sólo

tengamos variables cuantitativas la matriz $\mathbf{S}$ sería la matriz de correlación.

3. Realizar la descomposición en autovalores-autovectores de $\mathbf{S}$. La matriz $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_k]$ de los valores estandarizados de las componentes viene dada por los primeros k autovectores $\mathbf{x}_k$ de $\mathbf{S}$ normalizados a n (tal que $\mathbf{X}'\mathbf{X} = n\mathbf{I}_k$).

4. Para $l = 1, \ldots, k$: calcular la varianza de la l-ésima componente, dada por $\mathbf{x}_l'\mathbf{S}\mathbf{x}_l$ donde $\mathbf{x}_l$ denota la columna l-ésima de $\mathbf{X}$.

5. Calcular la $(p \times k)$-matriz $\mathbf{C} = (c_{jl})$ de las cargas al cuadrado de las p variables en las k componentes con $c_{jl} = \frac{1}{n}\mathbf{x}_l'\mathbf{S}_j\mathbf{x}_l$. Para variables cuantitativas, c_{jl} es la correlación al cuadrado entre la variable j y la componente l, y para variables cualitativas, c_{jl} es el coeficiente de correlación entre la variable j y la componente l.

Cuando todas las variables son cuantitativas (resp. cualitativas), este procedimiento es equivalente al análisis de componentes principales (ACP), (resp.análisis de correspondencia múltiple, ACM). Pero las cargas (correlaciones entre las variables y las componentes) y las coordenadas principales de las categorias (los baricentros de los valores de la componentes) no se obtienen directamente y deben ser calculados posteriormente si los deseamos. Desde un punto de vista práctico, este procedimiento requiere la construcción y el almacenamiento de p matrices de dimensión $n \times n$, que puede conducir a problemas de tamaño de memoria cuando n y p aumentan.

1.2. Descomposición en valores singulares (DVS) basada en el procedimiento PCAMIX

Una DVS de una matriz real o compleja $\mathbf{M} \in \mathbb{R}^{n\times m}$ es una factorización del tipo $\mathbf{M} = \mathbf{U}\Phi\mathbf{V}'$ con $\mathbf{U} \in \mathbb{R}^{m\times m}$, $\mathbf{V} \in \mathbb{R}^{n\times n}$ ortogonales y $\Phi \in \mathbb{R}^{m\times n}$ una matriz formada con los Valores Singulares de $\mathbf{M}$ en su diagonal principal ordenados de

mayor a menor. El i-ésimo Valor Singular es la raiz cuadrada del i-ésimo autovalor de $\mathbf{M}$. Obsérvese que

$$\mathbf{M}'\mathbf{M} = \mathbf{V}\Phi\mathbf{U}'\mathbf{U}\Phi\mathbf{V}' = \mathbf{V}\Phi^2\mathbf{V}'$$
$$\mathbf{M}\mathbf{M}' = \mathbf{U}\Phi\mathbf{V}'\mathbf{V}\Phi\mathbf{U}' = \mathbf{U}\Phi^2\mathbf{U}'$$

Así, considerando $\underline{v}_j$ una columna de $\mathbf{V}$, dado que $\mathbf{V}'\mathbf{V} = \mathbf{I}_n$, entonces $\mathbf{V}'\underline{v}_j$ es un vector columna con un 1 un la posición j-ésima, luego

$$\mathbf{M}'\mathbf{M}\underline{v}_j = \mathbf{V}\Phi^2 \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} = \mathbf{V} \begin{bmatrix} 0 \\ \vdots \\ \phi_j^2 \\ \vdots \\ 0 \end{bmatrix} = \phi_j^2\underline{v}_j \Rightarrow \quad \begin{array}{l} \underline{v}_j \text{ autovector de } \mathbf{M}'\mathbf{M} \\ \text{con autovalor asociado } \phi_j^2 \end{array}$$

Considerando ahora $\underline{u}_j$ una columna de $\mathbf{U}$, dado que $\mathbf{U}'\mathbf{U} = \mathbf{I}_m$, entonces $\mathbf{U}\underline{u}_j$ es un vector columna con un 1 en la posición j-ésima, luego

$$\mathbf{M}'\mathbf{M}\underline{u}_j = \mathbf{U}\Phi^2 \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} = \mathbf{U} \begin{bmatrix} 0 \\ \vdots \\ \phi_j^2 \\ \vdots \\ 0 \end{bmatrix} = \phi_j^2\underline{u}_j \Rightarrow \quad \begin{array}{l} \underline{u}_j \text{ autovector de } \mathbf{M}\mathbf{M}' \\ \text{con autovalor asociado } \phi_j^2 \end{array}$$

El objetivo de este apartado es obtener la matriz de las cargas al cuadrado con el cálculo de una sola matriz obtenida de la DVS de $\mathbf{Z}$.

Este procedimiento se lleva a cabo de acuerdo a los siguientes pasos:

1. Determinar la matriz de dimensiones $n \times (p_1 + m)$, $\mathbf{Z} = \frac{1}{\sqrt{n}}(\mathbf{Z}_1|\mathbf{Z}_2)$, donde:

   - $\mathbf{Z}_1 = (\mathbf{z}_1|\cdots|\mathbf{z}_j|\cdots|\mathbf{z}_{p_1})$ es la matriz de dimensiones $n \times p_1$ de los valores estandarizados de las n observaciones en las p_1 variables cuantitativas,

(como en el ACP).

- $\mathbf{Z}_2$ es la matriz de dimensiones $n \times m$ obtenida recodificando $\mathbf{G}$ de la siguiente manera:
  $\mathbf{Z}_2 = \mathbf{JGD}^{-1/2}$, (como en el ACM).

2. Realizar la DVS de $\mathbf{Z}$:

$$\mathbf{Z} = \mathbf{U}\Lambda^{1/2}\mathbf{V}', \tag{1.1}$$

donde $\mathbf{U}'\mathbf{U} = \mathbf{V}'\mathbf{V} = \mathbf{I}_r$, Λ es la matriz diagonal de autovalores (en orden descendente) y r es el rango de $\mathbf{Z}$.

3. Calcular la matriz de dimensiones $n \times k$ de los valores estandarizados de las componentes:

$$\mathbf{X} = \sqrt{n}\mathbf{U}_k \tag{1.2}$$

donde $\mathbf{U}_k$ denota la matriz de las primeras k columnas de $\mathbf{U}$.

4. Para $l = 1, \ldots, k$, la varianza de la l-ésima componente es dada por el l-ésimo autovalor en Λ.

5. Calcular la matriz:

$$\mathbf{A} = \mathbf{V}_k\Lambda_k^{1/2}. \tag{1.3}$$

donde $\mathbf{V}_k$ denota la matriz de las primeras k columnas de $\mathbf{V}$ y Λ_k es la matriz diagonal de los k autovalores más grandes.

6. Escribir $\mathbf{A} = \begin{pmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \end{pmatrix}$, que es la concatenación de una matriz de dimensiones $p_1 \times k$, $\mathbf{A}_1$ y una matriz de dimensiones $m \times k$, $\mathbf{A}_2$.

   - La matriz $\mathbf{A}_1$ contiene las cargas de las variables cuantitativas, correlaciones entre las variables y las componentes.
   - La matriz $\mathbf{DA}_2$ contiene las coordenadas principales de las categorias de las variables cualitativas.

- Calcular la matriz $\mathbf{C}$ de las cargas al cuadrado de las p variables en las k componentes. Esta matriz se obtiene de la matriz $\mathbf{A}$ así:

$$\begin{cases} c_{jl} = a_{jl}^2 & \text{si la variable } j \text{ es cuantitativa,} \\ c_{jl} = \sum_{s \in \mathbf{I}_j} a_{sl}^2 & \text{si la variable } j \text{ es cualitativa} \end{cases}$$

  donde $\mathbf{I}_j$ es el conjunto de los índices de las filas de $\mathbf{A}$ asociados con las categorías de la variable cualitativa j. Para simplificar la notación, denotaremos de ahora en adelante $c_{jl} = \sum_{s \in \mathbf{I}_j} a_{sl}^2$ para ambas variables, cuantitativas y cualitativas, con $\mathbf{I}_j = \{j\}$ en el caso cuantitativo.

Señalar que la matriz $\mathbf{X}$ de los valores estandarizados de las componentes es obtenida mediante una Descomposición en Valores Singulares de la matriz de datos recodificada $\mathbf{Z}$, mientras que ésta es obtenida de una descomposición en autovalores de la matriz $\mathbf{S}$ (suma de las matrices cuantificadas $\mathbf{S}_j$) en el planteamiento original de Kiers [6]. Además, la matriz $\mathbf{C}$ de las cargas al cuadrado (correlaciones al cuadrado o coeficiente de correlación entre las variables y las componentes) es calculada aquí de una sola matriz $\mathbf{A}$ obtenida con la DVS de $\mathbf{Z}$, mientras que ésta es calculada de dos matrices $\mathbf{X}$ y $\mathbf{S}_j$ en el planteamiento original de Kiers [6].

Al contrario que el planteamiento original del PCAMIX, este procedimiento ofrece simultáneamente las cargas de las variables cuantitativas y las coordenadas principales de las categorías de las variables cualitativas.

1.3. Rotación varimax en PCAMIX

1.3.1. El problema de optimización

Rotación ortogonal. Como indican Eckart y Young [4], de la Descomposición en Valores Singulares en (1.1) y las definiciones de las matrices $\mathbf{X}$ y $\mathbf{A}$ dadas en (1.2) y (1.3), la matriz $\mathbf{XA}'$ es una aproximación de mínimos cuadrados de rango k de $\mathbf{Z}$. Sea

ahora $\mathbf{T}$ una matriz de rotación ortogonal: $\mathbf{TT}' = \mathbf{T}'\mathbf{T} = \mathbf{I}_k$. Sea $\widetilde{\mathbf{X}} = \mathbf{XT}$ y $\widetilde{\mathbf{A}} = \mathbf{AT}$. Como $\mathbf{XA}' = \widetilde{\mathbf{X}}\widetilde{\mathbf{A}}'$, esta aproximación no es única sobre rotaciones ortogonales.

Esta falta de unicidad puede utilizarse para mejorar la interpretación de las soluciones originales. Para simplificar las interpretaciones, las matrices $\mathbf{X}$ y $\mathbf{A}$ son rotadas de manera que cuando consideremos una variable, algunas cargas al cuadrado son grandes (cercanas a 1) y el mayor número posible son cercanas a cero.

El problema varimax. En ACP, como $\widetilde{\mathbf{A}}$ contiene las cargas de las variables después de la rotación, el problema varimax se puede formular como:

$$\max_{\mathbf{T}\, ortonormal} f(\mathbf{T}), \tag{1.4}$$

donde

$$f(\mathbf{T}) = \sum_{l=1}^{k}\sum_{k=1}^{p}(\tilde{a}_{jl}^2)^2 - \frac{1}{p}\sum_{l=1}^{k}\left(\sum_{j=1}^{p}\tilde{a}_{jl}^2\right)^2 \tag{1.5}$$

es la función varimax que mide la simplicidad de las componentes después de la rotación.

En el planteamiento de la DVS en PCAMIX, la función varimax f es definida reemplazando en (1.5) los términos $\tilde{a}_{jl}^2$ por $\tilde{c}_{jl}$, donde $\tilde{c}_{jl} = \sum_{s\in\mathbf{I}_j}\tilde{a}_{sl}^2$ son las cargas al cuadrado después de la rotación:

$$f(\mathbf{T}) = \sum_{l=1}^{k}\sum_{k=1}^{p}(\tilde{c}_{jl})^2 - \frac{1}{p}\sum_{l=1}^{k}\left(\sum_{j=1}^{p}\tilde{c}_{jl}\right)^2 \tag{1.6}$$

Se debe señalar que las cargas al cuadrado después de la rotación, $\tilde{c}_{jl}$, son las correlaciones al cuadrado (respectivamente, razones de correlación) entre las variables cuantitativas (respectivament, cualitativas) y las componentes rotadas, siendo la definición de razón de correlación:

Definición 1. Sean dos variables (X, Y) definidas sobre una población, siendo X cuantitativa e Y cualitativa con colección de modalidades $\mathcal{M}$. Se considera una

muestra de tamaño n, $\{(x_i, y_i) : i = 1 \dots n\}$ y sean:

- n_s la frecuencia observada de la modalidad o categoría $s \in \mathcal{M}$
- $\overline{x}$ la media muestra de X y $\overline{x}_s$ la media muestral de X entre los individuos con modalidad $s \in \mathcal{M}$.

Se define la razón de correlación como la proporción de variabilidad de X "explicada" por las categorías de Y

$$\eta^2(X|Y) = \frac{\sum_{s\in\mathcal{M}} n_s(\overline{x}_s - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Podemos observar que esta definición esta basada en la descomposición de la variabilidad de X entre los grupos determinados por las modalidades de Y. En la expresión original de Kiers [6] de la función varimax en PCAMIX, las cargas al cuadrado después de la rotación, $\tilde{c}_{jl}$, son dadas por $\frac{1}{n}\tilde{\mathbf{x}}_l'\mathbf{S}_j\tilde{\mathbf{x}}_l$, donde $\tilde{\mathbf{x}}_l$ denota la columna l-ésima de $\widetilde{\mathbf{X}}$. Por tanto, la función varimax (1.6) se convierte en:

$$f(\mathbf{T}) = \sum_{l=1}^{k}\sum_{k=1}^{p}\left(\frac{1}{n}\tilde{\mathbf{x}}_l'\mathbf{S}_j\tilde{\mathbf{x}}_l\right)^2 - \frac{1}{p}\sum_{l=1}^{k}\left(\sum_{j=1}^{p}\frac{1}{n}\tilde{\mathbf{x}}_l'\mathbf{S}_j\tilde{\mathbf{x}}_l\right)^2. \tag{1.7}$$

El procedimiento iterativo de optimización. Como una solución directa de (1.4) para el óptimo $\mathbf{T}$ no es factible, puede usarse para el PCAMIX un procedimiento iterativo de optimización sugerido por Kaiser [5] para ACP.
La idea es considerar en cada iteración una rotación plana en la que la matriz de rotación $\mathbf{T}$ solo depende de un ángulo θ. Este procedimiento rota pares de dimensiones de la siguiente manera: las rotaciones en un único plano son aplicadas a las dimensiones 1 y 2, 1 y 3,. . . , 1 y k, 2 y 3,. . . ,(k-1) y k, iterativamente hasta que el proceso converge, es decir, hasta que se obtengan $k(k-1)/2$ rotaciones sucesivas que proporcionen un ángulo de rotación igual a cero.

La clave de este procedimiento de rotación es la definición del paso de rotación en un único plano. Veámos primero detalladamete el cálculo de un ángulo óptimo para

la rotación plana y luego veremos el procedimiento iterativo de rotación completo en más de dos dimensiones.

1.3.2. Rotación plana

Cualquier rotación plana en $\mathbb{R}^2$ se obtiene con una matriz de rotación $\mathbf{T}$ definida por:

$$\mathbf{T} = \begin{bmatrix} \cos\theta & -\operatorname{sen}\theta \\ \operatorname{sen}\theta & \cos\theta \end{bmatrix} \tag{1.8}$$

donde θ es el ángulo de rotación. El problema de rotación varimax (1.4) se reescribe como:

$$\max_{\theta\in\mathbb{R}} f(\theta)$$

Veamos a continuación como se realizaría la rotación plana usando el procedimiento DVS de PCAMIX.

La función varimax $f(\mathbf{T})$ definida con la DVS en (1.6) se escribe:

$$\begin{aligned} f(\theta) = &\sum_{j=1}^{p}\left(\sum_{s\in\mathbf{I}_j}\tilde{a}_{s1}^2\right)^2 + \sum_{j=1}^{p}\left(\sum_{s\in\mathbf{I}_j}\tilde{a}_{s2}^2\right)^2 - \frac{1}{p}\left(\sum_{j=1}^{p}\sum_{s\in\mathbf{I}_j}\tilde{a}_{s1}^2\right)^2 \\ &- \frac{1}{p}\left(\sum_{j=1}^{p}\sum_{s\in\mathbf{I}_j}\tilde{a}_{s2}^2\right)^2 \end{aligned} \tag{1.9}$$

con

$$\tilde{a}_{s1} = a_{s1}\cos(\theta) + a_{s2}\operatorname{sen}(\theta) \quad \text{y} \quad \tilde{a}_{s2} = -a_{s1}\operatorname{sen}(\theta) + a_{s2}\cos(\theta). \tag{1.10}$$

Esta función es igual que (ver Apéndice de la página 49):

$$f(\theta) = f(0) + \frac{\rho}{4p}(\cos(4\theta - \psi) - \cos\psi) \tag{1.11}$$

donde ρ y ψ son definidas por:

$$\rho = (a^2 + b^2)^{1/2}, \quad \cos\psi = b/\rho, \quad \operatorname{sen}\psi = a/\rho \tag{1.12}$$

con a y b dados por:

$$\begin{aligned} a &= 2p\sum_{j=1}^{p} u_j v_j - 2\sum_{j=1}^{p} u_j \sum_{j=1}^{p} v_j, \\ b &= p\sum_{j=1}^{p}(u_j^2 - v_j^2) - \left(\sum_{j=1}^{p} u_j\right)^2 + \left(\sum_{j=1}^{p} v_j\right)^2, \end{aligned} \tag{1.13}$$

donde u_j y v_j están definidas por:

$$u_j = \sum_{s\in\mathbf{I}_j}(a_{s1}^2 - a_{s2}^2) \quad \text{y} \quad v_j = 2\sum_{s\in\mathbf{I}_j} a_{s1}a_{s2}. \tag{1.14}$$

La función f obtenida en (1.11) es máxima para $\cos(4\theta - \psi) = 1 \Leftrightarrow 4\theta - \psi = 2k\pi$, así, los ángulos óptimos son:

$$\theta = \frac{\psi}{4} + k\frac{\pi}{2}, \quad k \in \mathbb{Z} \tag{1.15}$$

1.3.3. Procedimiento iterativo de rotación

Consideramos ahora el caso en que el número k de dimensiones en la rotación es mayor que dos. El procedimiento iterativo de rotación proporciona la matriz $\widetilde{\mathbf{X}}$ de los valores de las componentes estandarizadas rotadas y la matriz $\widetilde{\mathbf{A}}$ que es usada para obtener las cargas al cuadrado rotadas, las cargas rotadas (correlaciones) de las variables cuantitativas y las coordenadas principales rotadas de las categorias. Este procedimiento se lleva a cabo en los siguientes pasos:

1. Inicialización: $\widetilde{\mathbf{X}} = \mathbf{X}$ y $\widetilde{\mathbf{A}} = \mathbf{A}$ donde la matriz $n \times k$ $\mathbf{X}$ y la matriz $(p_1 + m) \times k$ $\mathbf{A}$ vienen dadas por la Descomposición en Valores Singulares basada en el

procedimiento PCAMIX dado en la Sección 1.2.

2. Para $l = 1, \ldots, k-1$ y $t = (l+1), \ldots, k$, calcular para el par de dimensiones (l, t):

   - el ángulo de rotación $\theta = \psi/4$ con ψ definido en (1.12), en concreto:

$$\psi = \begin{cases} \arccos\left(\frac{b}{\sqrt{a^2+b^2}}\right) & \text{si } a \geq 0, \\ -\arccos\left(\frac{b}{\sqrt{a^2+b^2}}\right) & \text{si } a \leq 0. \end{cases} \tag{1.16}$$

     donde a y b están definidos en (1.13).

   - la matriz de rotación $\mathbf{T} = \begin{bmatrix} \cos\theta & -\operatorname{sen}\theta \\ \operatorname{sen}\theta & \cos\theta \end{bmatrix}$,

   - las matrices $\widetilde{\mathbf{X}}$ y $\widetilde{\mathbf{A}}$ actualizadas con la rotación de sus l-ésima y t-ésima columnas.

3. Repetir el paso anterior hasta que los $k(k-1)/2$ ángulos θ sean iguales a cero.

4. Calcular:

   - la matriz $\widetilde{\mathbf{C}}$ con $\tilde{c}_{jl} = \sum_{s \in \mathbf{I}_j} \tilde{a}_{sl}^2$.

   - la matriz $\widetilde{\mathbf{A}}_1$ de las primeras p_1 filas de $\widetilde{\mathbf{A}}$ que contiene las cargas rotadas de las variables cuantitativas.

   - la matriz $\widetilde{\mathbf{A}}_2$ de las últimas m filas de $\widetilde{\mathbf{A}}$ y la matriz $\mathbf{D}\ \widetilde{\mathbf{A}}_2$ que contiene las coordenadas principales rotadas de las categorías de las variables cualitativas.

Capítulo 2

PCAMIX en R

En este capítulo, el método PCAMIX que hemos descrito en el Capítulo 1 se aplica a un conjunto de datos quel se compone de n observaciones descritas por $p = p_1 + p_2$ variables, siendo p_1 el número de variables cuantitativas y p_2 variables cualitativas, y m el número de categorías de las p_2 variables cualitativa. Este algoritmo mezcla ACP (caso cuantitativo) y ACM (caso cualitativo).

El paquete implementado en R denominado **PCAmixdata**[2] amplía los métodos de análisis multivariante ordinarios para incorporar este tipo de datos. Los métodos claves incluídos en el paquete son: análisis de componentes principales para datos mixtos (`PCAmix`), rotación ortogonal para `PCAmix`, y análisis factorial múltiple para datos mixtos. En esta memoria no se incluye descripción alguna de este último método, por lo que tampoco será tratada su implementación en R.

Para ilustrar el método PCAMIX y la rotación varimax asociada consideramos el conjunto de datos `gironde`, que se describe a continuación. Este conjunto de datos, incluído en el paquete arriba citado, está compuesto por cuatro tablas de datos, cada una caracteriza las condiciones de vida en 542 municipios de la región de Gironde en el suroeste de Francia. La primera tabla de datos describe los 542 municipios con 9 variables numéricas relacionadas con las condiciones de empleo, la segunda tabla describe los municipios con 5 variables (2 categóricas y 3 numéricas) relacionadas con

las condiciones de la vivienda, la tercera con 9 variables categóricas relacionadas con servicios (restaurantes, doctores, correos,...) y la última con 4 variables numéricas relacionadas con las condiciones ambientales. Por tanto, un total de 27 variables divididas en 4 grupos (Employment, Housing, Services, Enviroment).

2.1. Manual e ilustración del paquete `PCAmix`

Vamos a ilustrar el `PCAmix` con la tabla de datos `housing` del conjunto de datos `gironde`. Esta tabla contiene $n = 542$ municipios descritos en $p_1 = 3$ variables numéricas (`density`, `primaryes`, `owners`) y $p_2 = 2$ variables categóricas (`houses`, `council`) con un total de $m = 4$ categorías. En concreto, trabajaremos con la tabla de datos mixtos `housing`.

```
R> library("PCAmixdata")
R>data("gironde")
R>head(gironde$housing)
```

	density	primaryres	houses	owners	council
ABZAC	131.70	88.77	inf 90%	64.23	sup 5%
AILLAS	21.21	87.52	sup 90%	77.12	inf 5%
AMBARES-ET-LAGRAVE	531.99	94.90	inf 90%	65.74	sup 5%
AMBES	101.21	93.79	sup 90%	66.54	sup 5%
ANDERNOS-LES-BAINS	551.87	62.14	inf 90%	71.54	inf 5%
ANGLADE	63.82	81.02	sup 90%	80.54	inf 5%

Vamos a realizar un análisis de componentes principales usando la función `PCAmix`, que realiza el análisis del componentes principales de un conjunto de individuos (observaciones) descrito por una mezcla de variables cualitativas y cuantitativas. `PCAmix` incluye análisis ordinario de componentes principales (PCA) y análisis de correspondencia múltiple (MCA) como casos particulares.

- **Orden**

```
PCAmix(X.quanti = NULL, X.quali = NULL, ndim = 5, rename.level =
       FALSE, weight.col.quanti = NULL,weight.col.quali = NULL,
       graph = TRUE)
```

- **Argumentos**

`X.quant`	matriz númerica de datos, o un objeto que pueda ser usado como una matriz (vector numérico o un data frame con todas las columnas numéricas).
`X.quali`	matriz categórica de datos, o un objeto que pueda ser usado como una matriz (como un vector de caracteres, un factor o un data frame con todas las columnas de factores).
`ndim`	número de dimensiones guardadas en los resultados (por defecto 5).
`rename.level`	boleano, si es TRUE todos las categorías de las variables cualitativas se renombran de la siguiente manera: "nombre_variable" = "nombre_categoría". Así se evita que las categorías tengan nombres idénticos.
`weight.col.quanti`	vector de los pesos de las variables cuantitativas.
`weight.col.quali`	vector de los pesos de las variables cualitativas.
`graph`	boleano, si es TRUE se muestran los siguientes gráficos para las dos primeras dimensiones de `PCAmix`: mapa de componentes de los individuos, gráfico de las cargas al cuadrado de todas las variables (cuantitativas y cualitativas), gráfico del círculo de correlación (si se han incluído variables cuantitativas), mapa de componentes de las categorías (si las variables cualitativas se han incluído).

Sobre la orden anterior se han de ralizar diversas consideraciones:

- Si no se especifica `X.quali` (es decir, NULL), solo se incluyen las variables

cuantitativas, entonces se realiza ACP estándar. Si `X.quanti` es NULL, solo se incluyen las variables cualitativas, entonces se realiza ACM estándar.

- Si faltan valores, estos se reemplazan por las medias para las variables cuantitativas y por ceros en la matriz indicadores en las variables cualitativas.

- `PCAmix` calcula las cargas al cuadrado que almacena en unamatriz denominada `sqload`. Las cargas al cuadrado para una variables cualitativa son los coeficientes de correlacion entre la variables y las componentes principales. Para una variable cuantitativa, las cargas al cuadrado son las correlaciones al cuadrado entre la variable y las componentes principales.

- Cuando todas las p variables son cualitativas, las coordenadas de los factores (valores) de las n observaciones son iguales a las coordenadas de los factores (valores) del ACM estándar multiplicadas por la raíz cuadrada de p y los autovalores son iguales a los autovalores usuales del ACM multiplicados por p.

- Cuando todas las variables son cuantitativas, PCAmix proporciona exactamente los mismos resultados que ACP estándar.

La función crea un objeto que contiene los siguientes resultados:

`eig` una matriz que contiene los autovalores, los porcentajes de varianza y los porcentajes acumulados de varianza.

`ind` una lista que contiene resultados para los individuos (observaciones):
- `$coord`: coordenadas de los factores (valores) de los individuos,
- `$contrib`: contribuciones absolutas de los individuos.
- `$contrib.pct`: contribuciones relativas de los individuos,
- `$cos2`: cosenos cuadrados de los individuos.

`quanti` una lista que contiene resultados para las variables cuantitativas:
- `$coord`: coordenadas de los factores (valores) de las variables cuantitativas,
- `$contrib`: contribuciones absolutas de las variables cuantitativas,
- `$contrib.pct`: contribuciones relativas de las variables cuantitativas (en porcentaje),
- `$cos2`: cosenos cuadrados de las variables cuantitativas.

`levels` una lista que contiene resultados para las categorías de las variables cualitativas:
- `$coord`: coordenadas de los factores (valores) de las categorías,
- `$contrib`: contribuciones absolutas de las categorías.
- `$contrib.pct`: contribuciones relativas de las categorías (en porcentaje),
- `$cos2`: cosenos cuadrados de las categorías.

`quali` una lista que contiene resultados para las variables cualitativas:
- `$contrib`: contribuciones absolutas de las variables cualitativas (suma de las contribuciones absolutas de las categorías),
- `$contrib.pct`: contribuciones relativas de las variables cualitativas (suma de las contribuciones relativas de las categorías).

`sqload` matriz de dimensión (p,`ndim`) que contiene las cargas al cuadrado de las variables cualitativas y cuantitativas.

`coef` coeficientes de las combinaciones lineales utilizadas para construir las componentes principales de `PCAmix`, y para predecir las coordenadas (valores) de las nuevas observaciones en la función `predict.PCAmix`.

`M` vector de los pesos de las columnas usado en la DVS.

2.1.1. Funciones asociadas a `PCAmix`

Asociadas a esta orden básica del `PCAmix` del paquete existen otras funciones que permiten profundizar e interpretar los resultados. La función `plot.PCAmix` permite analizar gráficamente algunos de estos resultados.

Esta orden muestra los gráficos de salida de `PCAmix` y `PCArot`. Los individuos (observaciones), las variables cuantitativas y las categorías de las variables cualita-

tivas se trazan como puntos usando sus coordenadas (valores). Todas las variables (cuantitativas y cualitativas) se trazan como puntos en el mismo gráfico utilizando sus cargas al cuadrado.

- **Orden**

```
## método S3 para la clase 'PCAmix'
plot(x, axes = c(1, 2), choice = "ind", label = TRUE,
     coloring.ind = NULL, col.ind = NULL, coloring.var = FALSE,
     lim.cos2.plot = 0, lim.contrib.plot = 0, posleg = "topleft",
     xlim = NULL, ylim = NULL, cex = 1, leg = TRUE, main = NULL,
     cex.leg = 1, ...)
```

- **Argumentos**

`x`	objeto de clase `PCAmix` obtenido con la función `PCAmix` o `PCArot`.
`axes`	vector de longitud 2 que especifica las componentes a dibujar.
`choice`	gráfico que dibujará: • "“ind ” para el mapa de componentes de los individuos, • “cor” para el círculo de correlación si las variables cuantitativas están disponibles en los datos. • “levels” para el mapa de componentes de las categorias (si las variables cualitativas están disponibles en los datos) • “sqload” para la gráfica de las cargas al cuadrado de todas las variables.
`label`	boleano, si es FALSE las etiquetas de los puntos no están dibujadas.
`coloring.ind`	variable cualitativa como un vector de caracteres o un factor de tamaño n (número de individuos). Los individuos están coloreados de acuerdo con los niveles de esta variable. Si es NULL, los individuos no se colorean.

`col.ind`	vector de colores, de tamaño el número de categorías de `coloring.ind`.Si es NULL, los colores se eligen de manera automática.
`coloring.var`	boleano, si es TRUE las variables en el gráfico de las cargas al cuadrado se colorean de acuerdo con su tipo (cuantitativo o cualitativo).
`lim.cos2.plot`	valor entre 0 y 1. Los puntos con coseno cuadrado por debajo de este valor no se dibujan.
`lim.contrib.plot`	valor entre 0 y 100. Los puntos con contribuciones relativas (en porcentaje) por debajo de este valor no se dibujan.
`posleg`	posición de la leyenda.
`xlim`	vector numérico de longitud 2, que da el rango de coordenadas de x. Si es NULL (por defecto) el rango se define de manera automática (recomendado).
`ylim`	vector numérico de longitud 2, que da el rango de coordenadas de y. Si es NULL (por defecto) el rango se define de manera automática (recomendado).
`cex`	coeficiente de la función `par` en el paquete **graphics**.
`leg`	boleano, si es TRUE se muestra una leyenda.
`main`	una cadena con el título del gráfico a dibujar.
`cex.leg`	valor numérico que da la cantidad por la cual la leyenda debe ser ampliada. El valor predeterminado es 0.8.
...	argumentos para pasar a los métodos, como los parámetros gráficos.

Las observaciones pueden ser coloreadas de acuerdo con los niveles de una variable cualitativa. Las observaciones, las variables cuantitativas y las categorías se pueden seleccionar de acuerdo con su coseno cuadrado (`lim.cos2.plot`) o su contribución relativa (`lim.contrib.plot`) para el mapa de componentes. Solo se dibujan puntos

con coseno cuadrado o contribución relativa mayor que un umbral dado. Hay que tener en cuenta que la contribución relativa de un punto al mapa de componentes (un plano) es la suma de las contribuciones absolutas a cada dimensión dividida por la suma de los autovalores correspondientes.

A continuación vamos a explicar en detalle cómo la función `predict.PCAmix` puede ser útil para predecir las coordenadas (valores) de las observaciones no utilizadas en `PCAmix`.

Esta función calcula los valores de nuevas observaciones en las componentes principales de `PCAmix`. Si las componentes se han rotado, esta función calcula los valores de las nuevas observaciones en las componentes principales rotadas. En otras palabras, esta función está proyectando las nuevas observaciones en las componentes principales de `PCAmix` (o `PCArot`) obtenidas previamente en un conjunto de datos separado. Hay que tener en cuenta que las nuevas observaciones deben describirse con las mismas variables que las utilizadas en `PCAmix` (o `PCArot`).

- **Orden**

```
## método S3 para clase 'PCAmix'
predict(object, X.quanti = NULL, X.quali = NULL,
        rename.level = FALSE, ...)
```

- **Argumentos**

`object`	objeto de clase `PCAmix` obtenido con la función `PCAmix` o `PCArot`.
`X.quanti`	matrix de datos numéricos o un objeto que pueda ser usado como dicha matriz (como un vector numérico o un marco de datos con todas las columnas numéricas).
`X.quali`	matriz de datos categóricos, o un objeto que pueda ser usado como dicha matriz (como un vector de caracteres o un marco de datos con todas las columnas de factores).

`rename.level`	boleano, si es TRUE todas las categorías de las variables cualitativas se renombran de la siguiente manera: "nombre_variable" = "nombre_categoría". Esto evita tener nombres idénticos para las categorías.
...	otros argumentos para pasar a los métodos. Se ignoran en esta función.

La función devuelve la matriz de los valores de las nuevas observaciones en las componentes principales o en las componentes principales rotadas de `PCAmix`.

La función `supvar.PCAmix` calcula las coordenadas de las variables suplementarias (numéricas o categóricas) en las componentes de un objeto de clase `PCAmix`. Más precisamente, esta función crea un objeto R de clase `PCAmix` que incluye las coordenadas suplementarias. Veamos detalladamente como actúa esta función.

- **Orden**

```
## método S3 para clase 'PCAmix'
supvar(obj, X.quanti.sup = NULL, X.quali.sup = NULL,
       rename.level = FALSE, ...)
```

- **Argumentos**

`obj`	objeto de clase `PCAmix`.
`X.quanti.sup`	matrix de datos numéricos.
`X.quali.sup`	matriz de datos categóricos.
`rename.level`	boleano, si es TRUE todas las categorías de las variables cualitativas se renombran de la siguiente manera: "nombre_variable" = "nombre_categoría". Esto evita tener nombres idénticos para las categorías.
...	otros argumentos

2.1.2. Ilustración: housing

En este apartado se ilustra una aplicación de la técnica PCAMIX y su obtención en R-Programm a través del conjunto de datos `housing` incluido en `gironde`.

En primer lugar, la función `splitmix` divide una matriz de datos mixtos en dos conjuntos de datos: uno con las variables numéricas y otro con las variables categóricas.

```
R> split<-splitmix(gironde$housing)
R> X1<-split$X.quanti
R> X2<-split$X.quali
```

Aplicamos la orden `PCAmix` al conjunto de datos objetivo, obteniéndose:

```
R> res.pcamix <-PCAmix(X.quanti = X1,X.quali = X2,rename.level =
   TRUE, graph = FALSE)
R> res.pcamix$eig
      Eigenvalue Proportion Cumulative
dim 1  2.5268771  50.537541   50.53754
dim 2  1.0692777  21.385553   71.92309
dim 3  0.6303253  12.606505   84.52960
dim 4  0.4230216   8.460432   92.99003
dim 5  0.3504984   7.009968  100.00000
```

La suma de los autovalores es igual a la inercia total $p_1 + m - p_2 = 5$ y las dos primeras dimensiones recuperan el 71 % de la inercia total. Vamos a visualizar estas dos dimensiones en 4 gráficos diferentes. Para ello usaremos la función `plot.PCAmix`.

```
R> plot(res.pcamix, choice = "ind", coloring.ind = X2$houses,
   label =FALSE, posleg = "bottomright" , main =
   "(a) Observaciones")
R> plot(res.pcamix, choice = "levels", xlim = c(-1.5,2.5),
```

```
   main = "(b) Levels")
R> plot(res.pcamix, choice = "cor", main = "(c) Variables
   numéricas")
R> plot(res.pcamix, choice = "sqload", coloring.var = T, leg =
   TRUE, posleg = "topright", main="(d) Todas las variables")
```

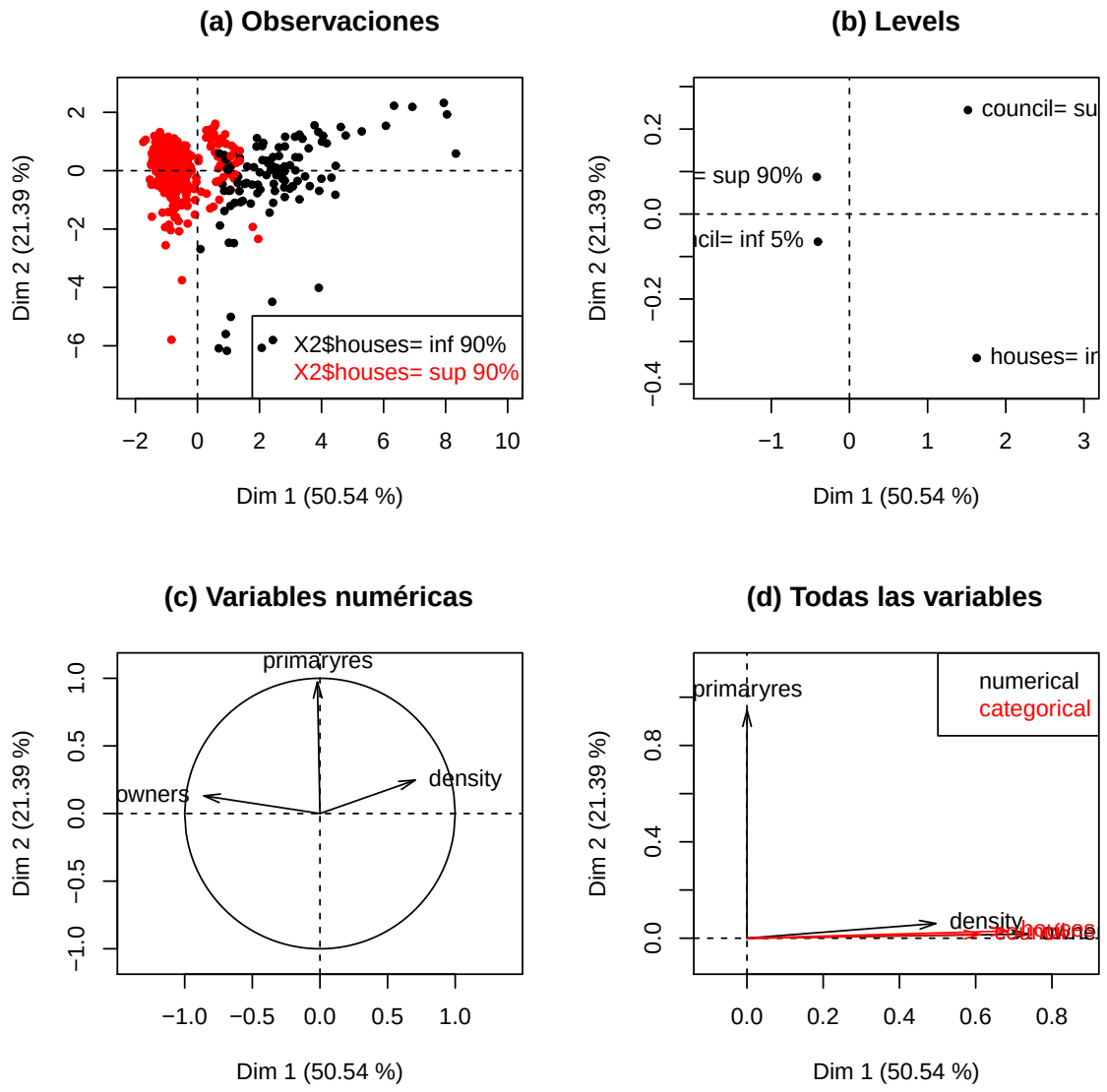

Figura 2.1: Gráficos de salida del `PCAmix` aplicado a la tabla de datos `housing`.

La Figura 2.1 (a) muestra el mapa de las componentes principales donde los municipios (las observaciones) están coloreadas por su porcentaje de viviendas (menos del 90 %, más del 90 %). La primera dimensión (lado izquierdo) destaca a los municipios con grandes proporciones de propiedades de propiedad privada. El mapa de niveles en la Figura 2.1 (b) confirma esta interpretación y sugiere que los municipios con una alta proporción de viviendas (a la izquierda) tienen un bajo porcentaje de viviendas municipales. El círculo de correlación en la Figura 2.1 (c) nos indica que la densidad de población está negativamente correlacionada con el porcentaje de propietarios de viviendas y que estas dos variables discriminan a los municipios en la primera dimensión.

La Figura 2.1 (d) dibuja las variables (categóricas y númericas) usando las cargas al cuadrado como coordenadas. Para las variables numéricas las cargas al cuadrado son correlaciones al cuadrado y para las variables categóricas son coeficientes de correlación. En ambos casos, miden el vínculo entre las variables y las componentes principales. Se observa que las dos variables númericas, `density` y `owners`, y las dos variables categóricas, `houses` y `council`, están vinculadas a la primera componente. Por el contrario, la variable `primaryres` es claramente ortogonal a estas variables y está asociada a la segunda componente. Hay que tener en cuenta que estos enlaces no muestran una asociación positiva ni negativa, y los mapas de la Figura 2.1 (b) y (c) son necesarios para una interpretación más precisa.

En resumen, los municipios a la derecha del mapa de componentes principales tienen una proporción relativamente alta de viviendas municipales y un pequeño porcentaje de viviendas de propiedad privada, y la mayoría de los alojamientos se alquilan. Por otro lado, los municipios en el lado izquierdo están compuestos en su mayoría por propietarios que viven en su residencia principal. El porcentaje de residencias primarias también tiene un papel estructurante en la caracterización de los municipios de esta región de Francia al definir claramente la segunda dimensión.

Ahora veremos en un ejemplo cómo la funcíon `predict.PCAmix` predice los valores de los municipios del conjunto test en las dos primeras componentes principales obtenidas con el conjunto training. Aquí, 100 municipios se muestrean al azar (conjunto test) y los 442 municipios restantes (conjunto training) se utilizan para realizar `PCAmix`.

```
R> set.seed(10)
R> test <- sample(1:nrow(gironde$housing), 100)
R> train.pcamix <- PCAmix(X1[-test,], X2[-test,], ndim = 2,
   graph = FALSE)
R> pred <- predict(train.pcamix, X1[test,], X2[test,])
R> head(pred)
                              dim1        dim2
MAZION                   -0.4120140  0.03905247
FLAUJAGUES               -0.6881160 -0.33163728
LATRESNE                  0.7447583  0.65305517
SAINT-CHRISTOLY-DE-BLAYE -0.7006372 -0.33216807
BERSON                   -1.1426625  0.33607088
CHAMADELLE               -1.3781919  0.24609791
```

Estas coordenadas ajustadas pueden usarse para dibujar los 100 municipios suplementarios en el mapa de los otros 442 municipios (ver Figura 2.2)

```
R> plot(train.pcamix, axes = c(1,2), label = FALSE, main =
  "Mapa de observaciones")
R> points(pred, col = 2, pch = 16)
R> legend("bottomright", legend = c("train","test"), fill = 1:2,
   col = 1:2)
```

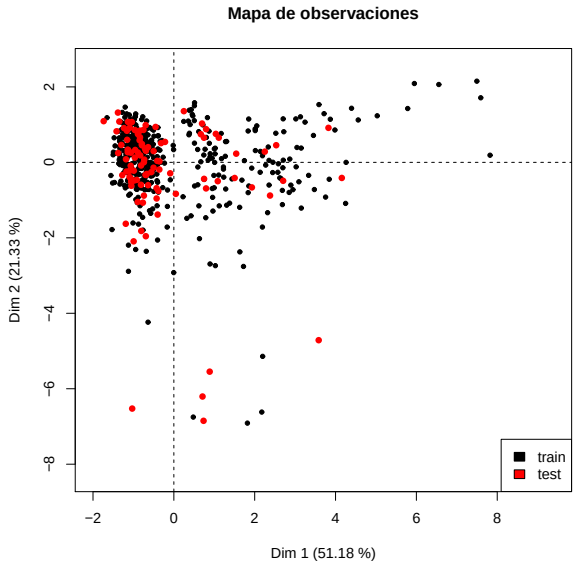

Figura 2.2: Proyección de 100 municipios suplementarios (en rojo) en el mapa de componentes principales de los otros 442 municipios (en negro).

Veámos un ejemplo de cómo actúa la función `supvar.PCAmix`, considerando la variable numérica `building` del conjunto de datos `environment` y la variable categórica `doctor` del conjunto de datos `services` como variables suplementarias.

```
R> X1sup <- gironde$environment[,1,drop = FALSE]
R> X2sup <- gironde$services[,7,drop = FALSE]
R> res.sup <- supvar(res.pcamix, X1sup, X2sup, rename.level
   = TRUE)
R> res.sup$quanti.sup$coord[,1:2,drop = FALSE]
              dim1      dim2
building 0.6945295 0.1884711
R> res.sup$levels.sup$coord[,1:2]
                     dim1         dim2
doctor=0       -0.44403187 -0.006224754
doctor=1 to 2   0.07592759 -0.112352412
doctor=3 or +   1.11104073  0.099723319
```

Las coordenadas de la variable numérica suplementaria `building` siguen siendo correlaciones. Por ejemplo, la correlación entre `building` y la primera componente principal es igual a 0.69. Las coordenadas de las categorías de las variables categóricas suplementarias siguen siendo baricentros. Por ejemplo, la coordenada -0.44 de la categoría `doctor = 0` es el valor medio de los municipios con 0 médicos en la primera componente principal estandarizada. Probablemente, la mayoría estará en la parte izquierda del mapa de componentes principales. Los gráficos de salida que incluyen estas variables suplementarias y las originales se pueden obtener como anteriormente conla función `plot.PCAmix` (ver Figura 2.3).

```
R> plot(res.sup, choice = "cor", main = "Variables numéricas")
R> plot(res.sup, choice = "levels", main = "Categorías",
   xlim = c(-2,2.5))
```

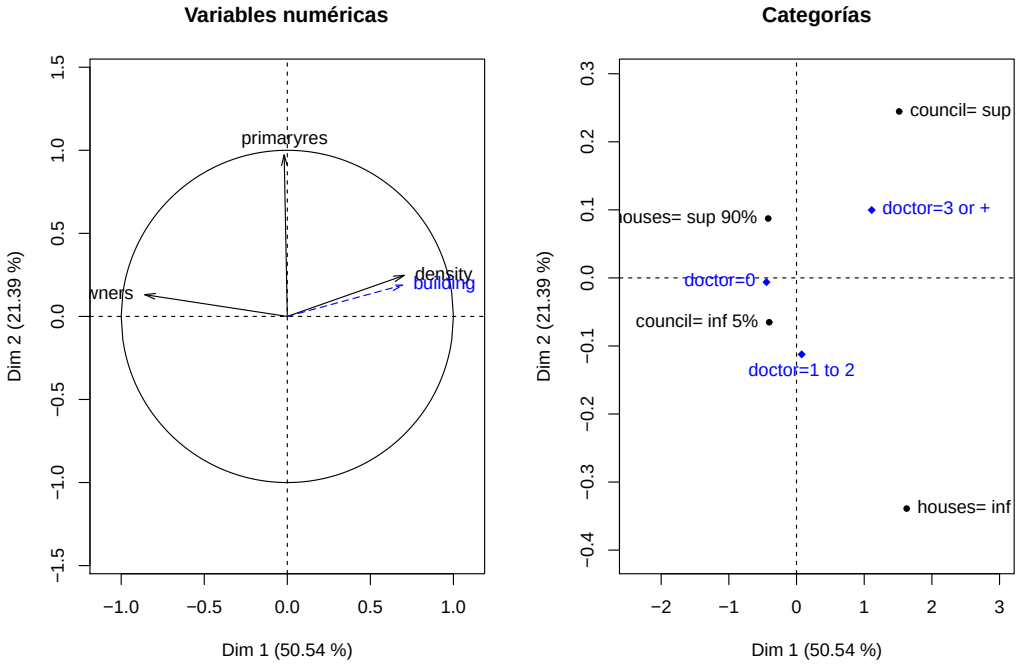

Figura 2.3: En azul, proyección de la variable numérica suplementaria `building` (izquierda) y proyección de las categorías de la variable categórica suplementaria `doctor` (derecha).

2.2. Manual e ilustración del paquete `PCArot`

Vamos ilustrar ahora el procedimiento `PCArot` con la tabla de datos mixtos `housing` que hemos utilizado en la sección (2.1). Veamos primero como actúa esta función y más adelante, un ejemplo sobre nuestro conjunto de datos.

Esta función realiza la rotación ortogonal en `PCAmix` mediante la maximización de la función varimax expresada en términos de las cargas al cuadrado del `PCAmix` (coeficientes de correlación para variables cualitativas y correlaciones al cuadrado para variables cuantitativas). `PCArot` incluye la rotación varimax ordinaria en el análisis de componentes principales ACP y una rotación de tipo varimax en el análisis de correspondencia múltiple ACM en casos especiales.

- **Orden**

```
PCArot(obj, dim, itermax = 100, graph = TRUE)
```

- **Argumentos**

`obj`	objeto de clase `PCAmix`.
`dim`	número de componentes principales rotadas.
`itermax`	número máximo de iteraciones en el algoritmo de optimización práctica de Kaiser basado en rotaciones planarias sucesivas por pares.
`graph`	boleano, si es TRUE, se muestran los siguientes gráficos para las dos primeras dimensiones después de la rotación: diagrama de los individuos (coordenadas de los factores), gráfico de las variables (cargas al cuadrado), diagrama del círculo de correlación (si hay variables cuantitativas disponibles), diagrama del mapa de componentes de las categorías (si las variables cualitativas están disponibles).

Si no se especifica `X.quali` (es decir, NULL) en el paso anterior de `PCAmix`, solo se incluyen variables cuantitativas y se realiza la rotación varimax estándar de

ACP. Si `X.quali` es NULL, solo se incluyen las variables cualitativas y se realiza la rotación de tipo varimax en ACM. Hay que tener en cuenta que p_1 es el número de variables cuantitativas, p_2 es el número de variables cualitativas y m es el número total de categorías de las p_2 variables cualitativas.

Esta función crea un objeto que contiene los siguientes resultados:

`eig`	varianzas de las `ndim` dimensiones después de la rotación.
`ind$coord`	una $n\times$ `dim` matriz cuantitativa que contiene las coordenadas (valores) de las n individuos en las `dim` componentes principales rotadas.
`quanti$coord`	una $p_1\times$ `dim` matriz cuantitativa que contiene las coordenadas (cargas) de las p_1 variables cuantitativas después de la rotación. Las coordenadas de las variables cuantitativas después de la rotación son correlaciones con las componentes principales rotadas.
`levels$coord`	una $m\times$ `dim` matriz cuantitativa que contiene las coordenadas de las m categorías en las `dim` componentes principales rotadas.
`quali$coord`	una $p_2\times$ `dim` matriz cuantitativa que contiene las coordenadas de las p_2 variables cualitativas en las `dim` componentes principales rotadas. Las coordenadas de las variables cualitativas después de la rotación son el coeficiente de correlación con las componentes principales rotadas.
`coef`	coeficientes de las combinaciones lineales utilizadas para construir las componentes principales rotadas del `PCAmix`.
`theta`	ángulo de rotación si `dim` es igual a 2.
`T`	matriz de rotación.

2.2.1. Ilustración: `housing`

Crearemos primero un marco de datos sin los primeros diez municipios (que los utilizaremos para predecir).

```
R> library("PCAmixdata")
```

```
R> data("gironde")
R> train <- gironde$housing[-c(1:10),]
R> split <- splitmix(train)
R> X1 <- split$X.quanti
R> X2 <- split$X.quali
R> res.pcamix <- PCAmix(X.quanti = X1, X.quali = X2, rename.level
   = TRUE, graph = FALSE)
R> res.pcamix$eig
      Eigenvalue Proportion Cumulative
dim 1  2.5189342  50.378685   50.37868
dim 2  1.0781913  21.563825   71.94251
dim 3  0.6290897  12.581794   84.52430
dim 4  0.4269180   8.538361   93.06267
dim 5  0.3468667   6.937335  100.00000
```

Las primeras $q = 3$ componentes principales de `PCAmix` recuperan el 84.5 % de la inercia total. Para mejorar la interpretación de estas 3 componentes sin afectar adversamente la proporción de inercia explicada, realizaremos una rotación usando la función `PCArot`.

```
R> res.pcarot<-PCArot(res.pcamix,dim=3,graph=FALSE)
R> res.pcarot$eig #varianza de las CP rotadas
         Variance Proportion
dim1.rot 1.919546   38.39092
dim2.rot 1.057868   21.15737
dim3.rot 1.248801   24.97601
```

La extensión de la proporción de varianza en las tres dimensiones se modifica, pero las componentes principales rotadas todavía contienen el 84.5 % de la inercia total:

```
R> sum(res.pcarot$eig[,2])
```

```
[1] 84.5243
```

La rotación también modifica las cargas al cuadrado con una asociación más clara despúes de la rotación entre la tercera componente principal y la variable `density`. De hecho, la correlación al cuadrado entre `density` y la tercera CP es igual a 0.39 antes de la rotación y aumenta a 0.9 después de la rotación.

```
R> res.pcamix$sqload[,1:3]
                  dim 1      dim 2       dim 3
density    0.4947711471 0.07081212 0.388696551
primaryres 0.0002426258 0.93841462 0.022365474
owners     0.7339206974 0.02431834 0.002697125
houses     0.6779904323 0.03359359 0.030895246
council    0.6120093292 0.01105260 0.184435327
R> res.pcarot$sqload
               dim1.rot   dim2.rot   dim3.rot
density    4.234771e-02 0.01205796 0.89987415
primaryres 3.187229e-05 0.95565170 0.00533914
owners     4.817170e-01 0.03308744 0.24613176
houses     6.336158e-01 0.02516749 0.08369595
council    7.618338e-01 0.03190379 0.01375971
```

Debido a que la rotación mejora la interpretación de la tercera componente principal, vamos a dibujar las observaciones y las variables en las dimensiones 1 y 3.

```
R> plot(res.pcamix, choice="ind", axes=c(1,3),label=FALSE,
   main="Observaciones antes de la rotación")
R> plot(res.pcarot, choice="ind", axes=c(1,3), label=FALSE,
   main="Observaciones después de la rotación")
R> plot(res.pcamix, choice="sqload", axes=c(1,3),
   main="Variables antes de la rotación", coloring.var=TRUE,
```

```
   leg=TRUE)
R> plot(res.pcarot, choice="sqload", axes=c(1,3),
   main="Variables después de la rotación", coloring.var=TRUE,
   leg=TRUE)
```

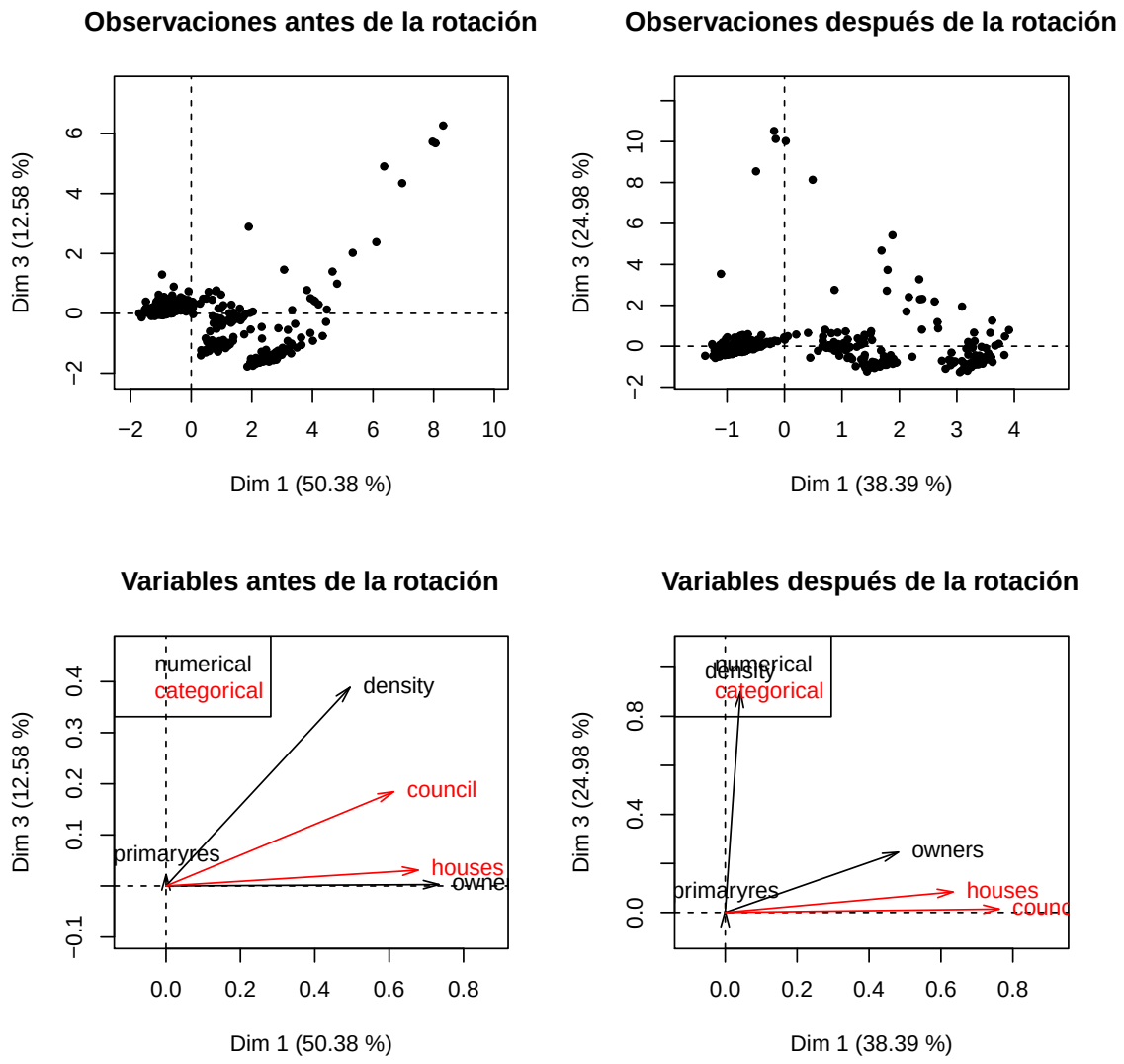

Figura 2.4: Gráficos de salida del `PCAmix` aplicados a la tabla de datos `housing` (sin las 10 primeras filas) antes de la rotación (izquierda) y después de la rotación con`PCArot` (derecha).

La Figura 2.4 muestra como la variable `density` se vincula más claramente despúes de la rotación a la tercera componente principal. De hecho, despúes de la rotación, las coordenadas de la variable `density` en el eje y es igual a 0.9 (la correlación al cuadrado entre `density` y la tercera componente principal rotada). Los municipios en la parte superior de la gráfica de las observaciones despúes de la rotación se caracterizan por su densidad de población. Hay que tener en cuenta que la ventaja de utilizar rotación en este conjunto de datos es bastante limitada.

Vamos a predecir los valores de los 10 primeros municipios de la tabla de datos `housing` en las componentes principales rotadas de `PCArot`.

```
R> test <- gironde$housing[1:10,]
R> splitnew <- splitmix(test)
R> X1new <- splitnew$X.quanti
R> X2new <- splitnew$X.quali
R> pred.rot <- predict(object = res.pcarot, X.quanti = X1new, X.quali
   = X2new)
R> pred.rot
                     dim1.rot   dim2.rot    dim3.rot
ABZAC                3.2685436  0.3494533 -0.85177749
AILLAS              -0.7235629  0.1200285 -0.22254455
AMBARES-ET-LAGRAVE   2.8852451  0.9823515 -0.03451571
AMBES                1.7220716  1.1590890 -0.78227835
ANDERNOS-LES-BAINS   0.3423361 -2.6886415  0.90574890
ANGLADE             -0.9131248 -0.4514258 -0.20108349
ARBANATS            -0.6653760  0.4217893  0.13105217
ARBIS               -0.7668742  0.3099338 -0.23304721
ARCACHON             1.8825083 -4.4533014  2.36935740
ARCINS              -0.6896492  0.2060403 -0.09049882
```

Estas coordenadas predichas pueden usarse para dibujar los 10 municipios suplementarios en el mapa de las componentes principales rotadas de los otros 532 municipios (Figura 2.5).

```
R> plot(res.pcarot, axes = c(1,3), label = FALSE, main = "Mapa de
   observaciones despúes de la rotación")
R> points(pred.rot[,c(1,3)], col = 2, pch = 16)
R> legend("topright", legend = c("train","test"), fill = 1:2,
   col = 1:2)
```

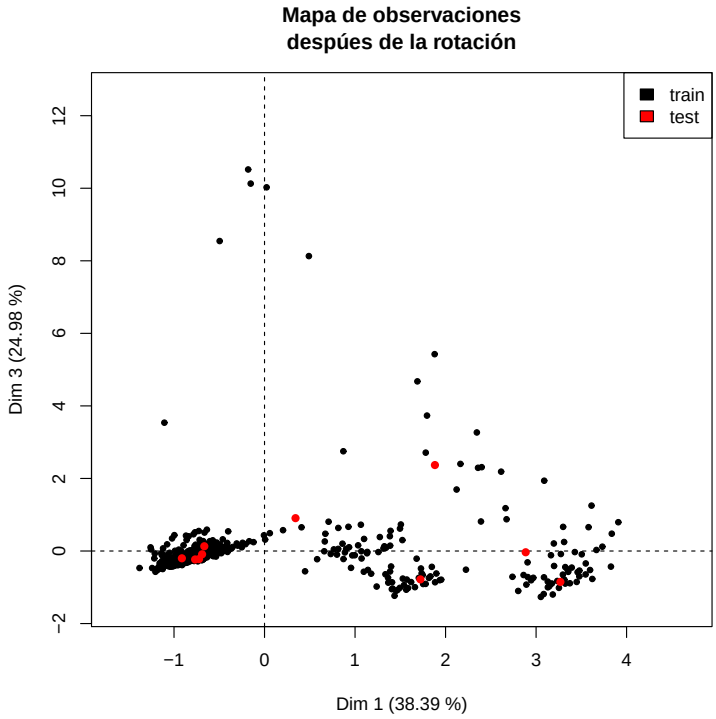

Figura 2.5: Proyección de 10 municipios suplementarios (en rojo) en el mapa después de la rotación.

Apéndice

Se define los números complejos:

$$a_s \stackrel{\text{def}}{=} a_{s1} + ia_{s2}, \quad \tilde{a}_s \stackrel{\text{def}}{=} e^{-i\theta} a_s = \tilde{a}_{s1} + i\tilde{a}_{s2},$$
$$t_j \stackrel{\text{def}}{=} \sum_{s \in \mathbf{I}_j} a_s^2 = u_j + iv_j, \quad \tilde{t}_j \stackrel{\text{def}}{=} \sum_{s \in \mathbf{I}_j} \tilde{a}_s^2 = e^{-2i\theta} t_j = \tilde{u}_j + i\tilde{v}_j,$$

donde $\tilde{a}_{s1}$, $\tilde{a}_{s2}$ han sido definidas en (1.10), u_j, v_j en (1.14) y $\tilde{u}_j$, $\tilde{v}_j$ son dados por la misma fórmula que u_j, v_j pero con una tilde sobre a_{s1} y a_{s2}.

Introducimos ahora una función varimax de valores complejos, $F(\theta)$, del ángulo de rotación θ dada por:

$$F(\theta) \stackrel{\text{def}}{=} p \sum_{j=1}^{p} \tilde{t}_j^2 - \left(\sum_{j=1}^{p} \tilde{t}_j \right)^2 = e^{-4i\theta} F(0),$$

donde $F(0)$ viene dado por la fórmula para $F(\theta)$ después de suprimir las "tildes" de los parámetros. $F(\theta)$ se puede descomponer de la forma:

$$\begin{aligned} F(\theta) = & \underbrace{p \sum_{j=1}^{p} (\tilde{u}_j^2 - \tilde{v}_j^2) - \left(\sum_{j=1}^{p} \tilde{u}_j \right)^2 + \left(\sum_{j=1}^{p} \tilde{v}_j \right)^2}_{g(\theta)} \\ & + \underbrace{2i \left\{ p \sum_{j=1}^{p} \tilde{u}_j \tilde{v}_j - \sum_{j=1}^{p} \tilde{u}_j \sum_{j=1}^{p} \tilde{v}_j \right\}}_{ih(\theta)}. \end{aligned} \tag{2.1}$$

Comparando con las fórmulas (1.12), (1.13) y (1.14) definiendo, a, b, ρ y ψ nos comprueba que :

$$F(0) = g(0) + ih(0) = b + ia = \rho e^{i\psi}.$$

Por tanto, obtenemos:

$$F(\theta) = \rho e^{i(\psi - 4\theta)} = \rho\{\cos(4\theta - \psi) - i\,\mathrm{sen}(4\theta - \psi)\}.$$

Pero tomando derivadas de la función varimax $f(\theta)$ definida en (1.9) nos da, usando el hecho de que $a'_{s1}(\theta) = a_{s2}(\theta)$ y que $a'_{s2}(\theta) = -a_{s1}(\theta)$:

$$\begin{aligned} pf'(\theta) &= 2\left\{p\sum_{j=1}^{p}\tilde{u}_j\tilde{v}_j - \sum_{j=1}^{p}\tilde{u}_j\sum_{j=1}^{p}\tilde{v}_j\right\} \\ &= h(\theta) = -\rho\,\mathrm{sen}(4\theta - \psi) \qquad (2.2) \\ &= a\cos 4\theta - b\,\mathrm{sen}\,4\theta, \qquad (2.3) \end{aligned}$$

y (2.2) prueba (1.11) por integración.

Bibliografía

[1] Chavent M., Kuentz-Simonet V., Labenne A., Liquet B., Saracco J., *Multivariate Analysis of Mixed Data Version 3.0.* Package 'PCAmixdata' (2017).

[2] Chavent M., Kuentz-Simonet V., Labenne A., Saracco J., *Multivariate Analysis of Mixed Data:The R Package PCAmixdata.* 11 Diciembre 2017.

[3] Chavent M., Kuentz-Simonet V., Saracco J., *Orthogonal rotation in PCAMIX.* Advances in Classification and Data Analysis, vol. 6, pp. 131-146, 2011.

[4] Eckart C., Young G., *The aproximation of one matrix by another of lower rank.* Psychometrika, 1: 211-218, 1936.

[5] Kaiser H. F., *The varimax criterion for analytic rotation in factor analysis.* Psychometrika, 23, 187-200, 1958.

[6] Kiers Henk A. L., *Simple structure in component analysis techniques for mixtures of qualitative and quantitative variables.* Psychometrika vol.56, nº 2, 197-212, June 1991 .