Técnicas conjuntas de reducción de la dimensión y agrupamiento

TRABAJO FIN DE GRADO

Grado en Estadística

María Eugenia Irizo Cruz

Sevilla, Junio de 2022

Índice general

Prólogo

Este trabajo de fin de grado, con el título de "Técnicas conjuntas de reducción de la dimensión y agrupamiento", surgió como una propuesta por mi tutor académico ante el conocimiento de un artículo sobre este tema, y su paquete en el lenguaje de programación R para su resolución.

El periodo de investigación y redacción de este trabajo ha sido desde febrero hasta junio de 2022. Hemos trabajado para profundizar en el análisis teórico de esta técnica, de la cual no existe demasiada información, y llevar a cabo un estudio práctico a través de la generación propia de un conjunto de datos.

Quiero agradecer a mi tutor Juan Manuel Muñoz Pichardo por su orientación y dedicación. También agradecer al profesor Pedro Luis Luque Calvo, por haber estado disponible ante cualquier problema con el programa R, y por último, a mi familia, por su gran apoyo siempre.

Información sobre el autor:

- María Eugenia Irizo Cruz
- Estudiante del doble grado en Matemáticas y Estadística de la Universidad de Sevilla
- Email: meirizocruz@gmail.com

Sevilla, 17 de junio de 2022.

Resumen

Este trabajo muestra el desarrollo de una técnica simultánea de reducción de la dimensión y análisis de conglomerados, a través del lenguaje de programación R y el paquete "clustrd" de este.

Se ha divido en cuatro capítulos. El primero, de introducción, donde se recoge la necesidad y motivación de aplicar esta técnica, además de los métodos que van a usarse.

En el segundo capítulo, se desarrolla teóricamente tres diferentes métodos de resolución de la técnica estadística para conjuntos de datos continuos. Además, se expone una forma de estudiar la estabilidad de los grupos formados en la parte de análisis de conglomerados.

El tercer capítulo es un ejemplo de ilustración de esta técnica. Se genera un conjunto de datos y se les aplica la resolución de esta técnica a través del paquete ya nombrado "clustrd" de R. Se obtienen interpretaciones y conclusiones sobre estos resultados.

Por último, en el cuarto capítulo, se explican distintas funcionalidades del paquete en R con varios ejemplos y conjuntos de datos incluidos en el programa base.

Abstract

This document shows a method for that simultaneosly performs the cluster analysis and the dimension reduction, through the R programming language and its "clustrd" package.

It has been divided into four chapters. The Chapter 1 is introductory, where the application of this statistical technique is motivated.

In the second chapter, three different methods to solve the statistical technique are theoretically described for continuous data sets. In addition, a way to study the stability of the groups formed in the cluster analysis is explained.

The third chapter is an illustration of this technique. The method is applied to a generated data by the package "clustrd" from R. Interpretations and conclusions are obtained about these results.

Finally, the fourth chapter explains different functionalities of the package "clustrd" with several examples and program base data sets.

Capítulo 1

Introducción

El fin del análisis de conglomerados, o análisis cluster, es separar observaciones similares en grupos. Para analizar la similaridad entre estas observaciones se tendrá en cuenta las particularidades de los datos y las medidas de distancia correspondientes. Si se necesita una gran cantidad de variables para calcular estas disimilaridades, el trabajo de crear los grupos será complicado. Por ello, se propone un método que combine la reducción de la dimensión con el análisis de conglomerados.

Así, se usa una representación de menor dimensión de las variables originales para detectar la estructura de grupos de las observaciones, lo que evita usar variables irrelevantes a la hora de crear la agrupación. Además, la baja dimensión ayuda a obtener soluciones más simples e interpretables.

La reducción de la dimensión se realiza mediante análisis de componentes principales. Este método pretende explicar la estructura de covarianzas del conjunto de variables originales a través de otro conjunto de variables mucho menor. Cada componente principal será una combinación lineal de las variables originales que explique la mayor variabilidad.

Una forma de llevar a cabo la combinación de la reducción de la dimensión con el análisis de conglomerados es ejecutar ambos métodos secuencialmente. Es decir, se reduce la dimensión, y a estos datos transformados le aplicamos el análisis de conglomerados. A esta teoría se le denomina el método "Tandem".

Sin embargo, cada uno de estos métodos por separados, componentes principales y agrupamiento, tienen un criterio de optimización distinto. La reducción de la dimensión pretende retener la mayor varianza posible en el menor número de dimensiones posibles, mientras que, el análisis cluster pretende encontrar observaciones similares y disimilares, y distribuirlas en los grupos. Por tanto, no se tiene certeza de que los resultados obtenidos en la reducción de la dimensión, es decir, las distintas componentes principales, sean las óptimas para aplicarles el análisis de conglomerados.

Un método alternativo es relacionar ambas partes, llevar a cabo la reducción de la dimensión y el análisis de cluster simultáneamente mediante un único criterio de optimización. Se realizará un análisis de componentes principales buscando que el espacio de menor dimensión contenga la información necesaria y útil para el agrupamiento.

Este método está implementado en la librería "clustrd"(Angelos Markos [2021]) de R Program (R Core Team [2021]), y puede ser aplicado a conjuntos de datos continuos, categóricos y mixtos. En esta memoria nos centraremos en conjuntos de datos continuos.

La estrategia conjuga la reducción de la dimensión basada en componentes principales y el análisis de conglomerados basado en "k-medias" (k-medias reducidas y k-medias factoriales). También se combinan ambos métodos de "k-medias", siendo las k-medias reducidas y factoriales casos particulares.

Se ha recogido gran parte de la información del artículo "Beyond Tandem Analysis: Joint Dimension Reduction and Clustering in R", Markos et al. [2019], y el manual del paquete "clustrd", Angelos Markos [2021].

Capítulo 2

Desarrollo teórico

2.1. Métodos para datos continuos

Previamente, se va a introducir alguna notación que se usará en el desarrollo. Sea X una matriz de datos centrada y estandarizada de dimensión $n \times Q$, B una matriz de cargas por columnas $Q \times d$ ortonormal, es decir, $B^\top B = I_d$, donde d es la dimensión del espacio reducido. Además, Z_K es una matriz $n \times K$ binaria, que indica la pertenencia de las n observaciones en los K grupos(clusters), es decir, $z_{ij} = 1$ si el $i - ésimo$ caso pertenece al grupo $j - ésimo$ y $z_{ij} = 0$ en caso contrario. Por último, se usará G para denotar la matriz $K \times d$ de los centroides de cada conglomerado en el espacio reducido $d - dimensional$.

El desarrollo se va a centrar en el análisis de datos continuos. Se distinguen dos opciones para el método secuencial(tandem): $K-$ medias reducidas y $K-$ medias factoriales, de las cuáles se introducirá la función objetivo. Finalmente, se estudiará como se relacionan ambos métodos para acabar en un único algoritmo.

2.1.1. K-medias reducidas

"En el agrupamiento por $K-$ medias, las matrices Z_K y G son determinadas tal que la suma de cuadrados de la distancia euclídea entre los objetos y los centroides de los grupos a los que pertenecen, es mínima".

"Cuando se desea un agrupamiento por $K-$medias en un espacio de baja dimensión, se requiere que los K centroides se encuentren en un subespacio de dimensión R del espacio columna de X." [vea De Soete and Carroll, 1994, pág 213]

Entonces, el problema conjunto de la reducción de la dimensión y el análisis cluster se resuelve tal que la asignación de grupos y la reducción de la dimensión maximice la varianza entre los grupos en el espacio reducido. Por tanto la función objetivo que el método pretende minimizar es:

$$\phi_{RKM}(B, Z_K, G) = \|X - Z_K G B^\top\|^2$$

donde $\|\cdot\|$ es la norma Frobenius. Obsérvese que la función depende de B, Z_k y G.

Es decir, se implementa un procedimiento de mínimos cuadrados, que minimiza la función alternando entre actualizar Z_K, es decir las asignaciones de conglomerados, y calcular los nuevos centroides, G. La matriz G contiene los K centroides, que son vectores de dimensión la del espacio reducido, por tanto, para cambiarlos se actualizará también la matriz de autovectores de las componentes principales, B. Así, se está combinando la reducción de la dimensión con la asignación de grupos. Este proceso sigue hasta que no haya cambios en las matrices.

La solución de la minimización respecto a los centroides, según la notación introducida por Yamamoto y Hwang ([2014]) es, $G = (Z_K^\top Z_K)^{-1} Z_K^\top XB$, para obtener,

$$\min_{B,Z_K} \phi_{RKM}(B, Z_K) = \min_{B,Z_K} \|X - Z_K(Z_K^\top Z_K)^{-1} Z_K^\top XBB^\top\|^2$$

Tomando $P = Z_K(Z_K^\top Z_K)^{-1} Z_K^\top$, que cumple $P = P^\top$, queda,

$$\min_{B,Z_K} \phi_{RKM}(B, Z_K) = \min_{P,B} \|X - PXBB^\top\|^2$$

Desarrollando,

$$\begin{gathered}
\|X - PXBB^\top\|^2 = (X - PXBB^\top)^\top (X - PXBB^\top) = \\
X^\top X - X^\top PXBB^\top - BB^\top X^\top PX + BB^\top X^\top PPXBB^\top = \\
X^\top X - X^\top PXBB^\top - BB^\top X^\top PX + BB^\top X^\top PXBB^\top = \\
\mathrm{Tr}(X^\top X) - \mathrm{Tr}(B^\top X^\top PXB) - \mathrm{Tr}(B^\top X^\top PXB) + \mathrm{Tr}(B^\top X^\top PXBB^\top B) = \\
\mathrm{Tr}(X^\top X) - \mathrm{Tr}(B^\top X^\top PXB) - \mathrm{Tr}(B^\top X^\top PXB) + \mathrm{Tr}(B^\top X^\top PXB) = \\
\mathrm{Tr}(X^\top X) - \mathrm{Tr}(B^\top X^\top PXB)
\end{gathered}$$

Véase Relación Norma Frobenius y Traza en el Apéndice.

Minimizar ϕ_{RKM} equivale a maximizar $-\phi_{RKM}$, es decir $-\mathrm{Tr}(X^\top X)+\mathrm{Tr}(B^\top X^\top PXB)$. Por tanto, se busca el máximo $\mathrm{Tr}(B^\top X^\top PXB)$, que es la traza de la matriz de varianzas de las nuevas coordenadas. Así, se observa que la varianza entre los grupos es máxima.

"El método de De Soete y Carroll De Soete and Carroll [1994] puede fallar en encontrar un agrupamiento interesante residiendo en un subespacio de los datos, cuando los datos tienen mucha variación en las direcciones ortogonales en las que captura el agrupamiento interesante. Esto se debe a que la varianza en tales direcciones puede contribuir considerablemente a la suma de las distancias al cuadrado entre los puntos de datos y los centroides. Sin embargo, considerando que uno es sólo interesado en una representación subespacial de los datos, parece más consistente encontrar este subespacio tal que los puntos de datos proyectados (es decir, proyectados en este subespacio) tienen la suma más pequeña de distancias al cuadrado a los centroides en el mismo subespacio. En el presente trabajo, se describe un procedimiento para este mismo propósito." [vea Vichi and Kiers, 2001, pág 50]

El procedimiento que se acaba de nombrar es el expuesto a continuación.

2.1.2. K-medias factoriales

"Las coordenadas de las proyecciones sobre la base vienen dadas por los componentes recogidos en la matriz XB. Dentro de este subespacio, con estos componentes, se busca una partición tal que los objetos estén más cerca de los centroides de los grupos de objetos." [vea Vichi and Kiers, 2001, pág 55]

Entonces, este método minimiza la varianza dentro de los clusters en el espacio reducido. Son simultáneos la agrupación por K-medias con las componentes principales.

La función objetivo para las k-medias factoriales es,

$$\phi_{FKM}(B, Z_K, G) = \|XB - Z_K G\|^2$$

Es decir, el método anterior minimizaba la distancia de los puntos originales a las transformaciones de los centroides del espacio reducido, mientras que este minimiza la distancia de los datos transformados a dichos centroides.

De nuevo, se implementa un procedimiento de mínimos cuadrados, que minimiza la función alternando entre actualizar Z_K, es decir las asignaciones de conglomerados, y calcular los nuevos centroides, G.

Análogamente a las k-medias reducidas, tomando la solución $G = (Z_K^\top Z_K)^{-1} Z_K^\top XB$, tenemos,

$$\min_{B, Z_K} \phi_{FKM}(B, Z_K) = \min_{B,P} \|XB - PXB\|^2$$

Según Yamamoto and Hwang [2014], este método, al igual que antes, puede fallar al identificar la estructura de los grupos, así como un subespacio para ello óptimo, debido a la existencia de variables irrelevantes y que algunas estén correladas entre ellas.

Se propone combinar la separación de subespacios con el agrupamiento, en el que los subespacios para variables relacionadas con una estructura de conglomerados y para variables perturbadoras se obtienen por separado.

2.1.3. K-medias reducidas y K-medias factoriales

Yamamoto y Hwang ([vea Yamamoto and Hwang, 2014, pág 117]) proponen la siguiente descomposición de la función objetivo de las k-medias reducidas:

$$\|X - PXBB^\top\|^2 = \|X - XBB^\top\|^2 + \|XB - PXB\|^2$$

"Esta descomposición muestra que las k-medias reducidas pueden verse como una conciliación de las componentes principales y las k-medias factoriales. Vichi, Vicari y Kiers propusieron minimizar una combinación convexa de ellos." [vea Markos et al., 2019, pág 4]. Por consiguiente, la función objetivo es,

$$\phi_{ClusPCA}(B, Z_K) = \alpha\|X - XBB^\top\|^2 + (1 - \alpha)\|XB - PXB\|^2$$

Minimizar esta función es lo mismo que maximizar $-\phi_{ClusPCA}$.

Desarrollando,

$$-\alpha\|X - XBB^\top\|^2 - (1-\alpha)\|XB - PXB\|^2 =$$
$$-\alpha(X - XBB^\top)^\top(X - XBB^\top) - (1-\alpha)(XB - PXB)^\top(XB - PXB) =$$
$$-\alpha(X^\top X - X^\top XBB^\top - BB^\top X^\top X + BB^\top X^\top XBB^\top)$$
$$-(1-\alpha)(B^\top X^\top XB - B^\top X^\top PXB - B^\top X^\top PXB + B^\top X^\top PPXB)$$

Si tomamos trazas,

$$-\mathrm{Tr}(\alpha X^\top X) + \mathrm{Tr}(\alpha X^\top XBB^\top) + \mathrm{Tr}(\alpha BB^\top X^\top X) -$$
$$\mathrm{Tr}(\alpha BB^\top X^\top XBB^\top)) - \mathrm{Tr}((1-\alpha)(B^\top X^\top XB)) +$$
$$\mathrm{Tr}((1-\alpha)(B^\top X^\top PXB)) + \mathrm{Tr}((1-\alpha)(B^\top X^\top PXB)) - \mathrm{Tr}((1-\alpha)()B^\top X^\top PPXB)) =$$
$$-\mathrm{Tr}(\alpha X^\top X) + \mathrm{Tr}(\alpha B^\top X^\top XB) + \mathrm{Tr}(\alpha B^\top X^\top XB) - \mathrm{Tr}(\alpha B^\top X^\top XB))$$
$$-\mathrm{Tr}((1-\alpha)(B^\top X^\top XB)) + \mathrm{Tr}((1-\alpha)(B^\top X^\top PXB)) +$$
$$\mathrm{Tr}((1-\alpha)(B^\top X^\top PXB)) - \mathrm{Tr}((1-\alpha)()B^\top X^\top PXB)) =$$
$$-\mathrm{Tr}(\alpha X^\top X) + \mathrm{Tr}(\alpha B^\top X^\top XB) +$$
$$-\mathrm{Tr}((1-\alpha)(B^\top X^\top XB)) + \mathrm{Tr}((1-\alpha)(B^\top X^\top PXB)) =$$
$$\mathrm{Tr}(B^\top X^\top(1-\alpha)PXB - B^\top X^\top(1-2\alpha)IXB) - \mathrm{Tr}(\alpha X^\top X) =$$
$$\mathrm{Tr}(B^\top X^\top((1-\alpha)P - (1-2\alpha)I)XB) - \mathrm{Tr}(\alpha X^\top X)$$

En conclusión, minimizar $\phi_{ClusPCA}$ equivale a maximizar

$$\mathrm{Tr}(B^\top X^\top((1-\alpha)P - (1-2\alpha)I)XB).$$

De aquí observamos que para conocer Z_k, podemos obtener la matriz de cargas B tomando la factorización en términos de autovalores y autovectores de $X^\top((1-\alpha)P - (1-2\alpha)I)X$, eligiendo los autovectores ortonormales correspondientes a los d autovalores mayores. (véase Factorización en autovalores y autovectores en el Apéndice)

Para conocer B, es suficiente con maximizar $\mathrm{Tr}(B^\top X^\top((1-\alpha)P)XB)$, que corresponde a un método de $K - medias$ estándar, aplicado a XB. Es decir, busco optimizar B aplicando las "K-medias2 en el subespacio definido por B.

En consecuencia, conociendo α, el método se resuelve con el siguiente algoritmo de mínimos cuadrados alternado.

- Primero, se genera una agrupación inicial, Z_K (por ejemplo, por la asignación aleatoria de objetos a los grupos).
- A continuación, se obtiene la matriz de cargas B tomando la descomposición de

$$X^\top((1-\alpha)P - (1-2\alpha)I)X.$$

- Se actualiza la agrupación Z_K usando $K-medias$ a las coordenadas del espacio reducido, XB.
- Se repite el proceso usando Z_K para la matriz de agrupación hasta que ésta se mantenga constante.

"Note que, para $\alpha = 0.5$, el problema se reduce a RKM, y para $\alpha = 0$ a FKM. Cuando $\alpha = 1$ la solución es equivalente al enfoque en tándem (análisis de componentes principales seguido de $K-medias$ de las puntuaciones de los factores). La selección del modelo final se puede basar en consideraciones teóricas, por ejemplo, decidiendo a priori que el método deseado debe ser un compromis entre FKM y RKM (es decir, eligiendo $\alpha = 0.25$), o por ejemplo, justo en el medio de PCA y RKM (es decir, eligiendo $\alpha = 0.75$). En general, se pueden evaluar varios valores de alfa y el más atractivo de estos (por ejemplo, el que conduce a la interpretación más interesante) podría ser elegido. Ejemplos empíricos y basados en simulación indicaron que RKM y FKM pueden identificar correctamente grupos bien separados enmascarados por variables generadas aleatoriamente, mientras que un correspondiente enfoque secuencial (en tándem) falla (De Soete y Carroll 1994; Vichi y Kiers 2001)." [vea Markos et al., 2019, pág 5]

2.2. Estabilidad de los grupos

Se considera un algoritmo que mide la estabilidad global de la reducción de la dimensión y métodos de conglomerados conjuntamente mediante bootstrap.

El proceso es el siguiente:

- Paso 1: Generar dos muestras bootstrap S_i y T_i de tamaño n de los datos. A continuación, aplicar el método de conglomerados a ambas y obtener C_i^S y C_i^T.
- Paso 2: Asignar cada observación x_j a los centros más cercanos de C_i^S y C_i^T, mediante la distancia Euclídea, resultando las particiones $C_i^S(x_j)$ y $C_i^T(x_j)$ con $j = 1 \dots n$.
- Paso 3: Calcular el porcentaje de acuerdo entre pares ajustado, ARI.(véase Indice de Rand Ajustado en el Apéndice)

Repetimos este proceso B veces, es decir, $i = 1 \dots B$. Así, se obtendrá B valores del porcentaje ARI. Así, podremos inspeccionar este valor, sabiendo que lo ideal es que $ARI = 1$, ya que implica que el porcentaje de acuerdo entre pares es del 100 %, es decir cada observación se ha asignado al mismo grupo en ambas particiones. Si se tuviera $ARI = 0$ sería una partición aleatoria.

Capítulo 3

Ilustración de la técnica estadística y la librería clustrd

A continuación, se va a aplicar el paquete "clustrd" a un conjunto de datos generado mediante distribuciones normales.

3.1. Generación de datos

Se quiere conseguir un conjunto de datos de 200 observaciones y 20 variables, que estén repartidos 4 grupos. Para ello, se considera una matriz L, tal que ella misma por su traspuesta sea definida positiva, para así después utilizarla como matriz de varianzas y covarianzas para la distribución normal y generar las muestras para cada grupo que se desea formar. En este caso, el primer grupo tendrá 60 observaciones, el segundo 40, el tercero 55 y el último 45.

El procedimiento de generación de datos se ha implementado en una función dependiente sólo de la semilla de aleatorización con objeto de tener la posibilidad de generar diversos conjuntos de datos y, a su vez, poder mantener los datos generados en caso de ejecutar el procedimiento en diversas ocasiones.

```
gendatp1<- function(numcin) {
set.seed(numcin)
p=20
n=200
n1=60
n2=40
n3=55
n4=45

L=matrix(c(0.90, 0.90, 0.90, 0.90, 0.90, 0.90, 0.90, 0.90, 0.80, 0.80,
           0.80, 0.80, 0.80, 0.80,-0.01,-0.01,-0.01,-0.01,-0.01,-0.01,
           rep(0.92,8), rep(0.80,2), 0.90, rep(0.80,2),
           0.80, 0.80,-0.01,-0.01,-0.01,-0.01,-0.01,-0.01, 0.01,-0.01,
           -0.01, 0.01,0.01,-0.01,-0.01, 0.01, 0.90, 0.85, 0.90, 0.85,
```

```
          0.92, 0.01,-0.01,-0.01, 0.01, -0.01,-0.01, 0.01,0.01,-0.01,
          -0.01, 0.01, 0.01,-0.01,-0.01, 0.01, 0.80, 0.95, 0.80,0.95,
          0.82, 0.01,-0.01,-0.01, 0.01, -0.01,-0.01, 0.01,0.61, 0.60,
          0.70, 0.02, 0.01,-0.65,-0.74, 0.61, 0.60, 0.70, 0.02, 0.01,
         -0.65,-0.74,-0.01,-0.01,-0.01,-0.01,-0.01,-0.01,-0.61,-0.60,
         -0.70, 0.02, 0.01, 0.65, 0.74,-0.61,-0.60,-0.70, 0.02, 0.01,
          0.65, 0.74,-0.01,-0.01,-0.01,-0.01,-0.01,-0.01, 0.01, 0.01,
          0.01, 0.01,-0.01,-0.01,-0.01, 0.01, 0.01, 0.01, 0.01,-0.01,
         -0.01,-0.01, 0.01, 0.01, 0.01, 0.01,-0.01,-0.01, 0.01, 0.01,
          0.01, 0.01,-0.01,-0.01,-0.01, 0.01, 0.01, 0.01, 0.01,-0.01,
         -0.01,-0.01, 0.01, 0.01, 0.01, 0.01,-0.01,-0.01, 0.01,-0.01,
          0.01,-0.01, 0.01,-0.01, 0.01, 0.01,-0.01, 0.01,-0.01, 0.01,
         -0.01, 0.01, 0.01,-0.01, 0.01,-0.01, 0.01,-0.01, 0.01,-0.01,
          0.01,-0.01, 0.01,-0.01, 0.01, 0.01,-0.01, 0.01,-0.01, 0.01,
         -0.01, 0.01, 0.01,-0.01, 0.01,-0.01, 0.01,-0.01, 0.90, 0.90,
          0.90, 0.90, 0.90, 0.90, 0.90, 0.90, 0.80, 0.80, 0.80, 0.80,
          0.80, 0.80,-0.01,-0.01,-0.01,-0.01,-0.01,-0.01,
          rep(0.92,8), rep(0.80,2), 0.90, rep(0.80,2),
         -0.01,-0.01,-0.01,-0.01,-0.01,-0.01,0.01,-0.01,-0.01, 0.01,
          0.01,-0.01,-0.01, 0.01, 0.90, 0.85, 0.90, 0.85, 0.92, 0.01,
         -0.01,-0.01, 0.01, -0.01,-0.01, 0.01, 0.01,-0.01,-0.01,0.01,
          0.01,-0.01,-0.01, 0.01, 0.80, 0.95, 0.80, 0.95, 0.82, 0.01,
         -0.01,-0.01, 0.01, -0.01,-0.01, 0.01, 0.61, 0.60, 0.70,0.02,
          0.01,-0.65,-0.74, 0.61, 0.60, 0.70, 0.02, 0.01,-0.65,-0.74,
         -0.01,-0.01,-0.01,-0.01,-0.01,-0.01,-0.61,-0.60,-0.70, 0.02,
          0.01, 0.65, 0.74,-0.61,-0.60,-0.70, 0.02, 0.01, 0.65, 0.74,
         -0.01,-0.01,-0.01,-0.01,-0.01,-0.01, 0.01, 0.01, 0.01, 0.01,
         -0.01,-0.01,-0.01, 0.01, 0.01, 0.01, 0.01,-0.01,-0.01,-0.01,
          0.01, 0.01, 0.01, 0.01,-0.01,-0.01, 0.01, 0.01, 0.01, 0.01,
         -0.01,-0.01,-0.01, 0.01, 0.01, 0.01, 0.01,-0.01,-0.01,-0.01,
          0.01, 0.01, 0.01, 0.01,-0.01,-0.01, 0.01,-0.01, 0.01,-0.01,
          0.01,-0.01, 0.01, 0.01,-0.01, 0.01,-0.01, 0.01,-0.01, 0.01,
          0.01,-0.01, 0.01,-0.01, 0.01,-0.01, 0.01,-0.01, 0.01,-0.01,
          0.01,-0.01, 0.01, 0.01,-0.01, 0.01,-0.01, 0.01,-0.01, 0.01,
          0.01,-0.01, 0.01,-0.01, 0.01,-0.01),
        nrow = 20,ncol = 20)

# Matriz de covarianzas
S=round(L%*%t(L),4)

# Vector de medias

mu=c(rep(10,20))
mu2=mu+c(rep(2,8),rep(0,12))
mu3=mu+ c(rep(0,14),rep(2,6))
mu4=mu+c(rep(0,8),rep(2,6),rep(0,6))
```

```
# Matriz de datos
set.seed(numcin)
X_1=mvrnorm(n1,mu,S)
X_2=mvrnorm(n2,mu2,S)
X_3=mvrnorm(n3,mu3,S)
X_4=mvrnorm(n4,mu4,S)

rbind.data.frame(X_1,X_2,X_3,X_4)

}

X=gendatp1(47567)
```

X es la matriz de datos generada y almacenada como "data.frame".

```
head(round(X[,1:10],2))
```

Tabla 3.1: Primeras filas de la matriz de datos

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
9.67	10.37	10.48	9.61	10.03	9.24	9.52	11.24	10.14	9.34
8.70	8.56	8.37	9.82	9.69	11.13	11.21	8.20	6.22	6.59
10.58	14.81	15.10	13.88	12.16	10.71	8.97	11.48	16.06	14.43
9.75	6.10	6.15	5.80	7.20	7.09	8.48	8.96	4.65	3.62
10.18	8.06	8.03	8.63	8.69	9.27	9.66	8.53	7.96	4.46
11.24	13.31	13.24	13.33	11.95	11.89	10.55	10.60	12.10	13.55

```
head(round(X[,11:20],2))
```

Tabla 3.2: Primeras filas de la matriz de datos

V11	V12	V13	V14	V15	V16	V17	V18	V19	V20
8.10	8.81	7.93	8.75	10.71	10.01	9.97	9.99	9.98	10.00
8.27	7.94	9.49	11.88	9.50	10.02	10.01	10.04	10.04	9.94
13.07	11.71	9.99	6.96	8.87	9.93	10.02	9.89	10.01	9.97
3.23	4.85	4.45	7.46	8.83	10.10	10.04	10.07	10.06	10.02
4.93	5.04	5.59	5.80	10.36	10.17	10.14	10.08	10.08	10.04
13.75	12.32	12.19	12.08	8.87	9.98	9.93	10.02	9.94	9.95

Aplicando el desarrollo teórico, se va a hacer un estudio con estos datos y los distintos métodos de análisis conjunto de reducción de la dimensión y agrupamiento, se estudiará la estabilidad de cada uno y se podrán comparar los distintos resultados.

3.2. K-medias reducidas

A continuación, como los datos se han generado de forma que estén repartidos en 4 grupos con medias distintas, tendría sentido considerar las K-medias reducidas con 4 grupos. Además, el estudio se realiza para 2 dimensiones después de 10 soluciones iniciales aleatorias, y así poder dibujarlo. Se usará la rotación 'varimax' con normalización Kaiser en el método.

```
set.seed(47567)
RKM = cluspca(X, 4, 2, method = "RKM", rotation = "varimax",
scale = FALSE, nstart = 10)
```

```
summary(RKM)
```

```
## Solution with 4 clusters of sizes 58 (29%), 51 (25.5%), 47 (23.5%),
## 44 (22%) in 2 dimensions. Variables were mean centered and
## unstandardized.
##
## Cluster centroids:
##              Dim.1   Dim.2
## Cluster 1 -1.0925 -3.2856
## Cluster 2 -7.8148  4.8365
## Cluster 3  7.5119 -4.7296
## Cluster 4  2.4741  3.7771
##
## Variable scores:
##       Dim.1   Dim.2
## V1  -0.2566 -0.1140
## V2  -0.4446 -0.1438
## V3  -0.4600 -0.1630
## V4  -0.3555 -0.0399
## V5  -0.2712  0.0288
## V6  -0.1637  0.1485
## V7  -0.0721  0.2245
## V8  -0.2865 -0.0193
## V9  -0.2938  0.1299
## V10 -0.2748  0.2120
## V11 -0.1770  0.3375
## V12 -0.0927  0.3998
## V13  0.0208  0.5245
## V14  0.0856  0.4871
## V15 -0.0252  0.0747
## V16 -0.0132 -0.0490
## V17 -0.0121 -0.0454
## V18 -0.0128 -0.0467
## V19 -0.0123 -0.0475
## V20 -0.0122 -0.0453
```

```
##
## Within cluster sum of squares by cluster:
## [1]  848.2987 1232.4273  956.7397  676.0764
##  (between_SS / total_SS =  72.11 %)
## ...
##
## Objective criterion value: 4168.8901
##
## Available output:
##
##  [1] "obscoord"  "attcoord"  "centroid"  "cluster"   "criterion" "size"
##  [7] "odata"     "scale"     "center"    "nstart"
```

El vector de clusters:

```
##   [1] 1 3 2 3 3 2 3 2 3 1 1 4 1 3 1 1 1 1 2 3 1 4 3 3 4 2 1 2 2 3 3 3
##  [33] 1 4 2 2 2 3 4 4 4 3 4 1 3 4 1 3 1 2 2 4 1 4 1 4 1 1 3 2 2 2 2 1
##  [65] 1 2 1 1 3 3 1 2 3 2 2 3 1 1 2 2 1 1 1 2 2 2 3 1 4 1 1 3 1 3 1 2
##  [97] 2 1 1 1 1 3 3 1 3 1 1 1 1 1 3 2 4 1 3 2 3 2 3 2 3 3 4 1 2 4 1 2
## [129] 2 2 4 1 3 3 4 1 2 1 4 1 1 1 1 1 3 1 1 3 3 2 2 3 3 4 2 4 4 1 4 2
## [161] 4 2 2 4 3 4 3 4 3 2 4 4 2 4 4 2 3 4 4 3 1 2 4 4 4 4 2 2 3 4 2 4
## [193] 2 2 4 4 3 4 4 4
```

Del resumen se obtiene que en el primer grupo hay 58 observaciones, es decir el 29 % de ellas; en el segundo 51 (25.5 %); en el tercero 47 (23.5 %), y por último, en el cuarto están 44 de las observaciones (22 %).

La suma de cuadrados representa una medida de variación o desviación con respecto a la media. Se calcula como una suma de los cuadrados de las diferencias con respecto a la media.

Se tiene que la suma de cuadrados total se descompone en suma de cuadrados entre los grupos más suma de cuadrados dentro de los grupos.

$$SC_{Total} = SC_{Entre} + SC_{Dentro}$$

Por tanto, mientras más cercana a 1 sea la razón $\frac{SC_{Entre}}{SC_{Total}}$, mayor será la suma de cuadrados entre los grupos.

Esto quiere decir que la agrupación es bastante buena, los grupos están bien separados e identificados, es decir, no hay gran dispersión de cruces entre los grupos.

En este caso, la razón es un 72.11 %, por lo que los grupos formados por la técnica estadística están bien "separados".

De hecho, se definía el conjunto de datos con unos tamaños de grupo de, 60, 40, 55 y 45. Así, a priori se puede observar que además de que la separación sea precisa, los tamaños de los grupos que se han formado no difieren mucho de los que se definieron.

Por otra parte, la técnica estadística proporciona los siguientes coeficientes de las componentes principales:

```
RKM$attcoord
```

Tabla 3.3: Coeficientes de las componentes principales

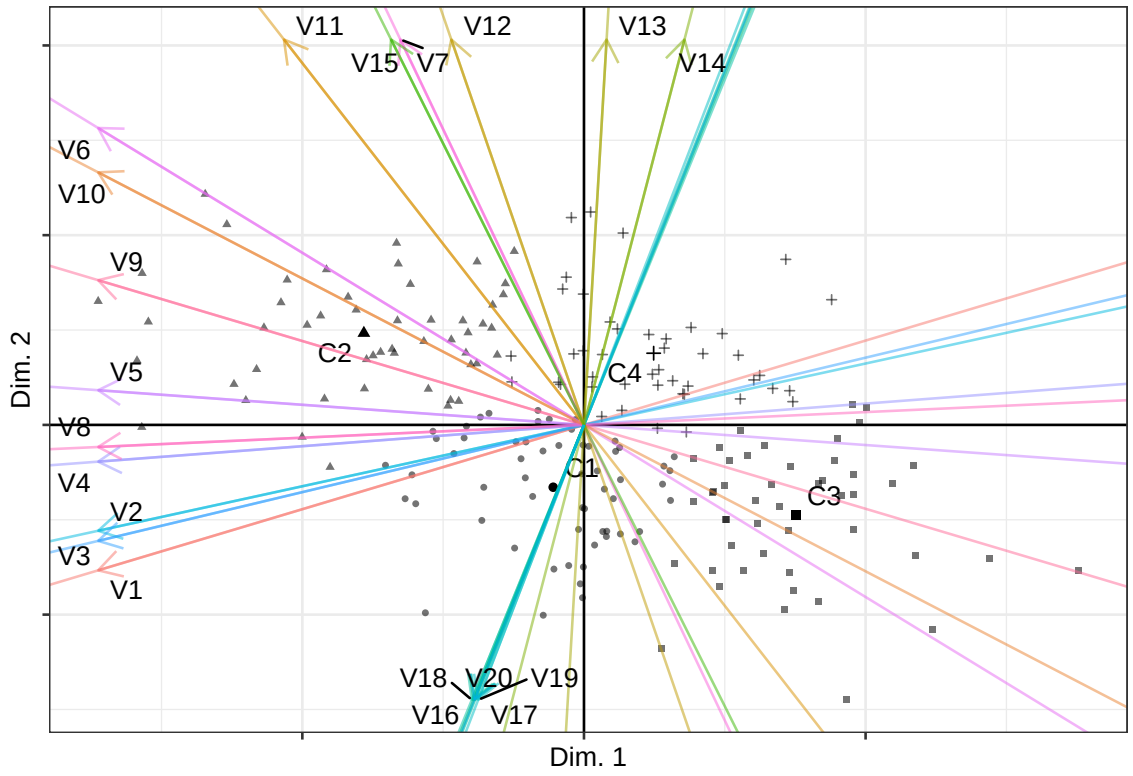
V1	-0.2566460	-0.1140482
V2	-0.4445527	-0.1437993
V3	-0.4599696	-0.1630126
V4	-0.3554852	-0.0399189
V5	-0.2712129	0.0287599
V6	-0.1637166	0.1485114
V7	-0.0720674	0.2245340
V8	-0.2864821	-0.0193395
V9	-0.2938243	0.1299055
V10	-0.2747663	0.2120189
V11	-0.1770165	0.3375437
V12	-0.0927285	0.3997820
V13	0.0207687	0.5245055
V14	0.0855849	0.4870953
V15	-0.0251771	0.0746623
V16	-0.0132401	-0.0489516
V17	-0.0121246	-0.0454287
V18	-0.0128170	-0.0466570
V19	-0.0123104	-0.0474836
V20	-0.0121651	-0.0453110

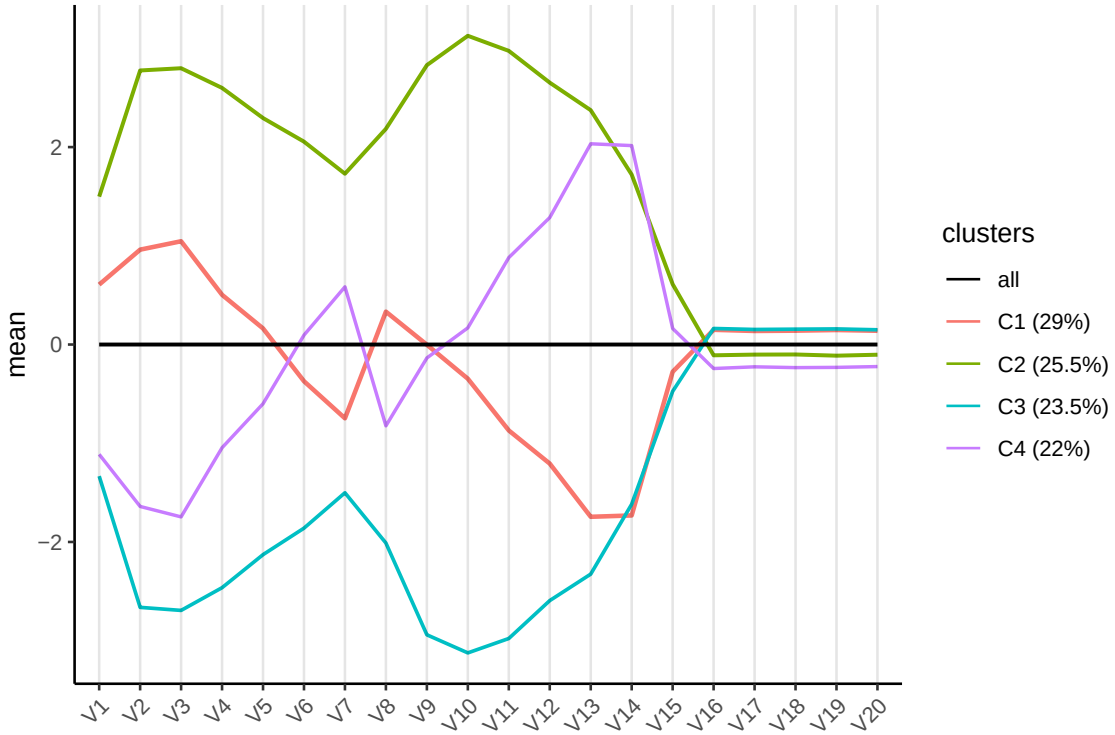
Se observa que la dimensión 1 está más correlada con las variables 2 y 3, además negativamente. Es decir, a menor valor de "V2" y "V3" mayor será el valor de la primera componente principal. Sin embargo, la dimensión 2 está correlada sobre todo con las variables 13 y 14, positivamente.

Observando el siguiente gráfico se concluye que la dimensión 1 separa los clusters 1 y 2 del 3 y 4. La dimensión 2 diferencia significativamente a los grupos 2 y 4 frente a los grupos 1 y 3. Por tanto, conjuntamente todos los grupos están separados.

Del segundo gráfico, perfiles medios de los clusters en las variables originales, se puede concluir finalmente que el están bien diferenciados el cluster 1 del 4 y el cluster 2 del 3. Se tiene que el grupo 1 se caracteriza por un crecimiento de la media de las variables "V2" y "V3" y un decrecimiento de "13" y "V14", mientras que el grupo 4 hace justo lo contrario, hay un crecimiento de la media de las variables "V13" y "V14" y un decrecimiento de "V2" y "V3". Estas variables son justos las que diferenciaban las correlaciones de las dos dimensiones. Conclusiones similares se pueden obtener de la gráfica con los grupos 2 y 3.

```
plot(RKM,cludesc = TRUE)
```

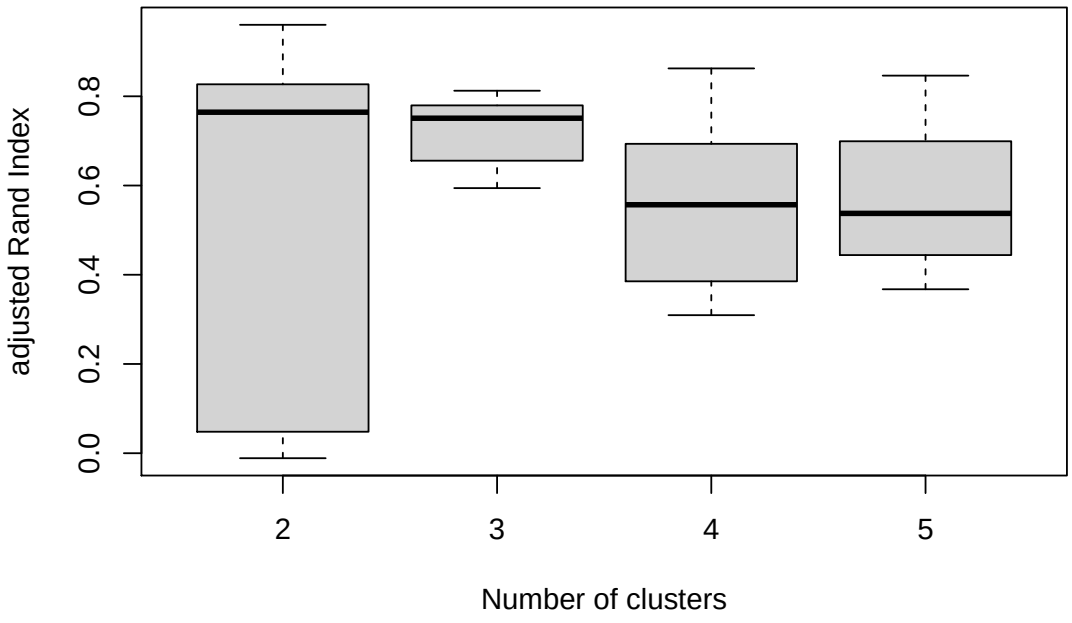




A continuación, se estudia la estabilidad como se había analizado anteriormente, buscando en un rango de 2 a 5 grupos y 2 dimensiones. Se toman 10 muestras bootstrap para ilustrar el procedimiento, aunque el número adecuado de muestras bootstrap debiera ser superior (`nboot`= 100,200...).

```
boot_RKM = global_bootclus(X, nclusrange = 2:5, ndim = 2,
method = "RKM", nboot = 10, nstart = 10, seed = 47567)
```

```
boxplot(boot_RKM$rand, xlab = "Number of clusters", ylab =
"adjusted Rand Index")
```



Anteriormente en esta memoria se había recogido que para alcanzar estabilidad y fiabilidad en los resultados, lo ideal es que el índice de Rand tomara valores cercanos a 1. En este caso, al calcularlo para cada una de las diez muestras bootstrap, y aplicarle la media, se observa que las medias de los índices que más se acercan a 1 son para 2 y 3 grupos. Sin embargo para un número de grupos de 2, los índices están más dispersos, mientras que para 3 grupos están más centrados en la mediana. Por tanto, se puede considerar que 3 grupos garantiza más estabilidad que las demás opciones.

Estas aparentes diferencias entre los grupos utilizados en la generación de datos y los resultados arriba recogidos se deben a tratar de conjugar detección de conglomerados y reducción de la dimensión. Recuérdese que por motivos de presentación de resultados e ilustración de los mismos, se han tomado sólo dos dimensiones. A través de estas dos dimensiones, las variables desde V15 a V20 no están bien representadas (véase la matriz de coeficientes que definen las componentes principales), por lo que la separación de los grupos o conglomerados no contemplan las diferencias ocasionadas por dichas varibales (véase el plot de los perfiles medios de las variables originales).

Inicialmente, se podría haber elegido el número de grupos y de dimensiones apropiado, sin dar lugar a la subjetividad, con la función “tuneclus”.

Aplicamos esta función, buscando en un rango de 3 a 5 grupos y de 2 a 5 dimensiones.

```
set.seed(47567)
bestRKM = tuneclus(X, 3:5, 2:5, method = "RKM",
criterion = "asw", dst = "low", nstart = 1, seed = 47567)
```

```
##
## The best solution was obtained for 3 clusters of sizes 83 (41.5%),
## 62 (31%), 55 (27.5%) in 2 dimensions, for an average Silhouette
## width value of 0.532. Variables were mean centered and standardized.
##
## Cluster quality criterion values across the specified range of
## clusters (rows) and dimensions (columns):
##      X2    X3    X4 X5
## 3 0.532
## 4 0.484 0.346
## 5 0.511 0.359 0.316
##
## The average Silhouette width values of each cluster are:
## [1] 0.57 0.53 0.48
##
## Cluster centroids:
##             Dim.1   Dim.2
## Cluster 1  0.5566  2.3724
## Cluster 2  2.4079 -2.1898
## Cluster 3 -3.5543 -1.1117
##
## Within cluster sum of squares by cluster:
## [1] 261.0290 226.0978 339.7259
##  (between_SS / total_SS =  69.82 %)
##
## Objective criterion value: 1033.7763
##
## Available output:
##
## [1] "clusobjbest" "nclusbest"   "ndimbest"    "critbest"    "critgrid"
## [6] "crit"        "cluasw"
```

Se obtiene que la mejor solución es para 3 grupos de tamaños 83 (41.5 %), 62 (31 %), 55 (27.5 %) en 2 dimensiones. Aunque la razón $\frac{SC_{Entre}}{SC_{Total}}$ es de 69,82 %, menor que la que habíamos obtenido anteriormente.

3.3. K-medias factoriales

Ahora que tenemos una solución, por "k-medias" reducidas, podemos usar esta como solución inicial del método de "k-medias" factoriales.

```
set.seed(47567)
FKM = cluspca(X, 4, 2, method = "FKM", rotation = "varimax",
scale = FALSE, smartStart = RKM$cluster)
```

```
summary(FKM)
```

```
## Solution with 4 clusters of sizes 70 (35%), 65 (32.5%), 34 (17%),
## 31 (15.5%) in 2 dimensions. Variables were mean centered and
## unstandardized.
##
## Cluster centroids:
##           Dim.1 Dim.2
## Cluster 1     0     0
## Cluster 2     0     0
## Cluster 3     0     0
## Cluster 4     0     0
##
## Variable scores:
##       Dim.1   Dim.2
## V1   0.0012 -0.0336
## V2  -0.3579 -0.3338
## V3   0.3243 -0.0240
## V4   0.0142 -0.3065
## V5  -0.0243  0.3100
## V6   0.0477  0.3653
## V7  -0.0049 -0.0157
## V8  -0.0003  0.0384
## V9   0.0086  0.0049
## V10  0.0030  0.3243
## V11 -0.0027  0.3108
## V12  0.0012 -0.3185
## V13 -0.0017 -0.3099
## V14 -0.0084 -0.0116
## V15  0.0003 -0.0384
## V16 -0.3887  0.3357
## V17 -0.3165 -0.2057
## V18  0.7144 -0.0720
## V19  0.0244 -0.0251
## V20 -0.0339  0.0055
##
## Within cluster sum of squares by cluster:
## [1] 0 0 0 0
##  (between_SS / total_SS =  92.53 %)
## ...
##
## Objective criterion value: 0
##
```

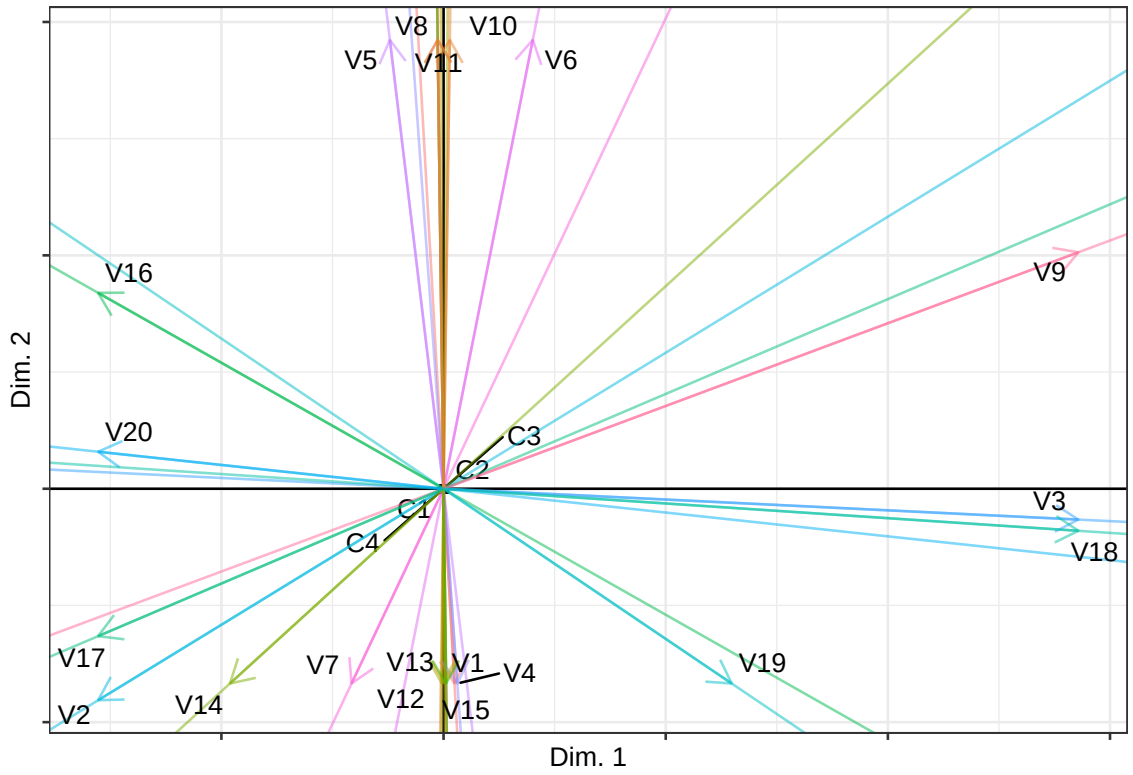
```
## Available output:
##
##  [1] "obscoord"  "attcoord"  "centroid"  "cluster"   "criterion" "size"
##  [7] "odata"     "scale"     "center"    "nstart"
```

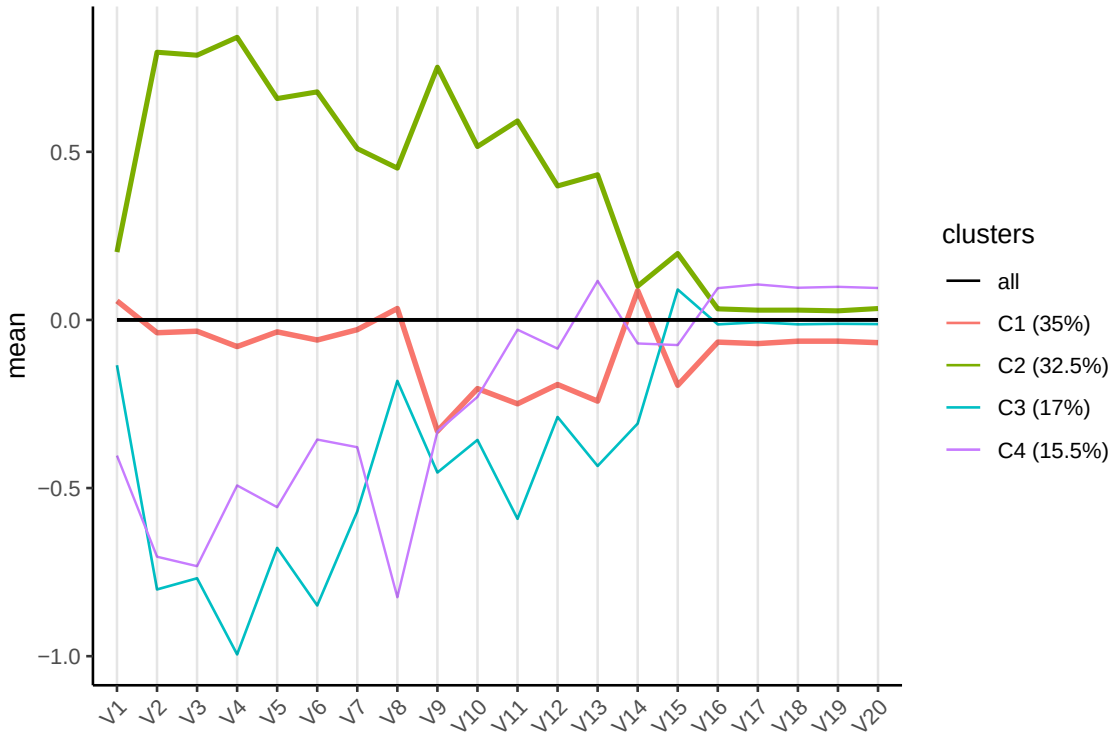
Vector de clusters:

```
##   [1] 3 4 1 1 1 4 3 1 4 1 2 2 2 3 3 1 2 4 2 4 1 2 3 2 2 4 1 2 2 1 3 3
##  [33] 3 1 4 1 2 4 2 2 3 3 4 1 3 1 1 1 1 4 2 1 1 1 2 1 2 2 4 2 1 1 2 2
##  [65] 2 1 1 1 1 1 2 1 4 1 2 1 3 1 2 2 3 3 1 3 3 2 1 2 4 2 4 2 1 1 2 2
##  [97] 2 1 2 2 1 2 3 4 2 2 2 1 3 1 1 1 3 4 4 1 1 1 3 1 4 1 3 2 3 4 2 4
## [129] 4 2 2 1 2 1 2 1 2 2 1 4 1 1 2 3 2 3 2 1 3 2 2 2 2 4 4 4 1 1 4 2
## [161] 3 3 1 2 3 2 1 2 1 1 4 4 1 1 3 3 1 2 2 3 4 2 2 1 1 4 2 1 3 1 2 1
## [193] 2 3 3 1 4 2 1 1
```

Ahora, la razón de la suma de cuadrados entre los grupos y la suma de cuadrados total es aún mayor, 92,53 %, por tanto, los grupos están bastante bien diferenciados. Sin embargo, el gráfico no es tan aclarativo como lo era el anterior.

```
plot(FKM,cludesc = TRUE)
```





3.4. K-medias reducidas y K-medias factoriales. Método Tandem

Se ha estudiado como Vichi, Vicari y Kiers propusieron como función objetivo una combinación convexa de la descomposición de Yamamoto y Hwang, así, haciendo $\alpha = 1$ se tiene el análisis tandem, análisis de componentes principales seguido de K-medias de las puntuaciones de los factores.

```
set.seed(47567)
M_Tandem = cluspca(X, 4, 2, alpha = 1)
```

```
summary(M_Tandem)
```

```
## Solution with 4 clusters of sizes 74 (37%), 54 (27%), 38 (19%),
## 34 (17%) in 2 dimensions. Variables were mean centered and
## standardized.
##
## ...
## Within cluster sum of squares by cluster:
## [1] 101.8444 315.5997 131.9907  72.9433
##  (between_SS / total_SS =  77.33 %)
## ...
##
```

De este resumen se concluye que el primer grupo tiene 74 observaciones, el segundo 54, el tercero 38, y el último 34. En este caso, la razón $\frac{SC_{Entre}}{SC_{Total}}$ es de un 77.33 %, mayor que en el método de las k-medias reducidas que era un 72.11 %.

Los coeficientes de las componentes principales son:

```
M_Tandem$attcoord
```

Tabla 3.4: Coordenadas de la componentes principales

V1	-0.2236377	0.0033313
V2	-0.2833740	-0.0256300
V3	-0.2783588	-0.0253899
V4	-0.2913200	-0.0249902
V5	-0.3163310	-0.0329390
V6	-0.2779819	-0.0263204
V7	-0.2283255	-0.0264709
V8	-0.2549555	-0.0323065
V9	-0.2751163	-0.0395986
V10	-0.2750639	-0.0384592
V11	-0.2742314	-0.0374672
V12	-0.2688589	-0.0432863
V13	-0.2280925	-0.0349258
V14	-0.1883711	-0.0167443
V15	-0.0881182	-0.2675823
V16	0.0628497	-0.4275181
V17	0.0603579	-0.4279682
V18	0.0608551	-0.4279904
V19	0.0617995	-0.4275293
V20	0.0602733	-0.4280537

Se tiene que la dimensión 1 está más correlada con las variables 4 y 5, además negativamente. Es decir, a menor valor de "V4" y "V5" mayor será el valor de la primera componente principal. Sin embargo, la dimensón 2 está correlada sobre todo con las variables 18 y 20, positivamente.

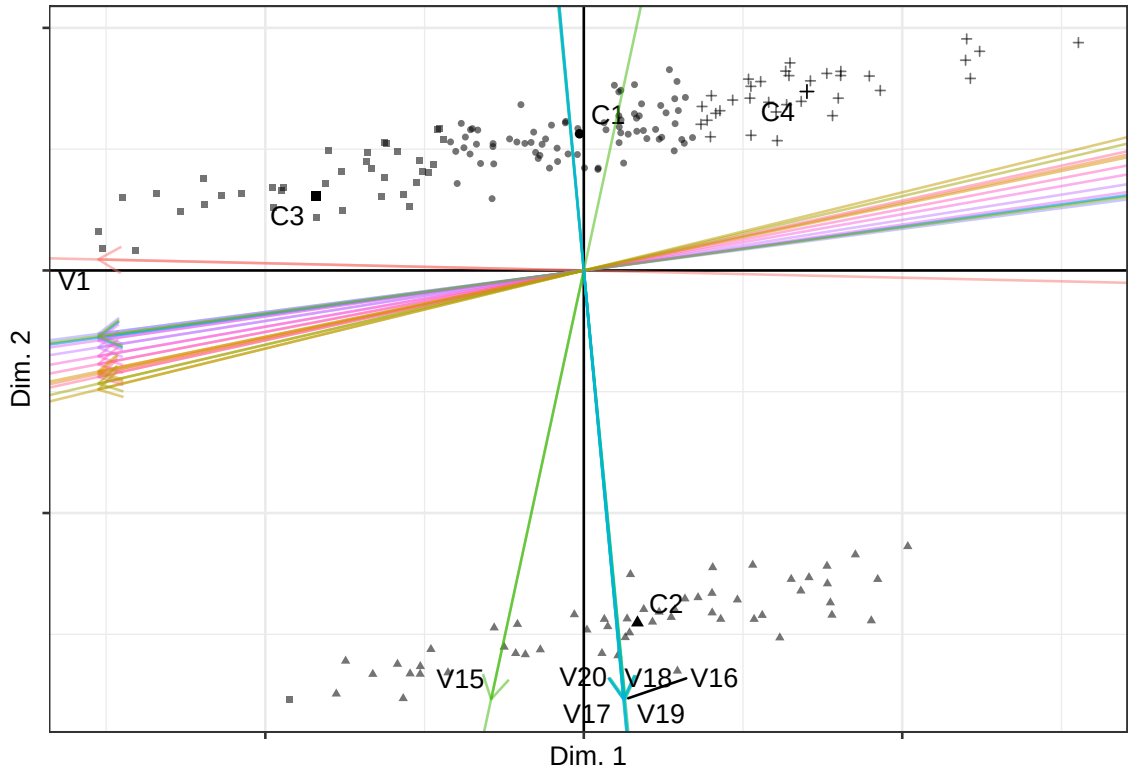
El vector de pertenencia a los grupos:

```
as.vector(M_Tandem$cluster)
```

```
##   [1] 1 4 3 4 4 3 4 3 4 1 1 1 1 4 1 1 1 1 3 4 1 1 4 4 1 1 1 3 3 4 4 4
##  [33] 1 1 1 3 3 4 1 1 1 4 1 4 4 3 1 4 1 3 3 4 1 1 1 1 1 1 4 1 3 3 3 1
##  [65] 1 3 1 1 4 4 3 3 4 3 3 4 1 3 3 3 1 1 1 3 3 1 4 1 1 1 1 4 1 4 1 3
##  [97] 3 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 2 2 2
## [129] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1
## [161] 1 3 3 1 4 1 4 3 4 3 1 1 3 1 1 1 4 1 1 4 1 3 1 1 1 1 3 3 4 1 3 1
## [193] 3 3 1 1 4 4 1 1
```

Observando el gráfico se concluye que la dimensión 2 diferencia significativamente al grupo 2 frente a los demás, mientras que la dimensión separa a los grupos 3 y 4, y el grupo 1 no está claramente diferenciado. Esto puede deberse a lo comentado anteriormente, tomando sólo dos dimensiones hay variables que no están bien representadas y el agrupamiento no contempla ciertas diferencias, o que el grupo 1 lo hayamos creado muy correlado con alguno de los demás.

```
plot(M_Tandem)
```



Capítulo 4

Implementación en R e Ilustraciones

Trabajaremos con el paquete "clustrd" que contiene desarrollos de las k-medias factoriales y las k-medias reducidas.

4.1. Datos

Para ilustrar ejemplos se va a utilizar los siguientes conjuntos de datos:

4.1.1. macro

- Descripción

Datos sobre el rendimiento económico de 20 países miembros de la OECD. (septiembre de 1999). Este rendimiento refleja la interacción de seis principales indicadores económicos: producto interno bruto (GDP), indicador adelantado (LI), tasa de desempleo (UR), tasa de interés (IR), balanza comercial (TB), tasa nacional neta de ahorros (NNS), todos estos numéricos.

- Uso

```
data(macro)
```

- Formato

Conjunto de datos con 20 observaciones y 6 variables numéricas.

Tabla 4.1: Las primeras observaciones del conjunto de datos

	GDP	LI	UR	IR	TB	NNS
Australia	4.8	8.4	8.1	5.32	0.7	4.7
Canada	3.2	2.5	8.4	5.02	1.6	5.2
Finland	3.9	-1.0	11.8	3.60	8.8	7.7
France	2.3	0.7	11.7	3.69	3.9	7.3
Spain	3.6	2.5	19.0	4.83	1.2	9.6
Sweden	4.1	1.1	8.9	4.20	7.0	4.0
USA	4.1	1.4	4.5	5.59	-1.4	7.0
Netherlands	2.9	1.6	4.2	3.69	7.0	15.8
Greece	3.2	0.6	10.3	11.70	-8.3	8.0
Mexico	2.3	5.6	3.2	20.99	0.0	12.7

4.1.2. iris

- Descripción

Datos sobre las medidas en centímetros del ancho y el largo del sépalo y el pétalo de 50 flores de 3 especies de lirios.

- Uso

```
data(iris)
```

Tabla 4.2: Las primeras observaciones del conjunto de datos

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5.0	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa

- Formato

Conjunto de datos con 150 casos y 5 variables: “Sepal.Length”, “Sepal.Width”, “Petal.Length”, “Petal.Width”, y “Species”, todas numéricas excepto la variable que representa a las especies que es un factor con 3 niveles.

4.2. Funciones

Las funciones que se describen a continuación son las que se han usado principalmente. Entre ellas están la implementación del método de las k-medias con componentes principales, funciones para medir la estabilidad, representaciones gráficas, y procesos de selección del número y dimensión de grupos para la calidad de estos.

4.2.1. cluspca

- Descripción

Implementa las k-medias factoriales y las k-medias reducidas, además de una mezcla de ambos métodos. Todos estos, combinan el análisis de componentes principales con las k-medias para la agrupación.

- Uso

```
cluspca(data, nclus, ndim, alpha = NULL, method = c("RKM","FKM"),
center = TRUE, scale = TRUE, rotation = "none", nstart = 100,
smartStart = NULL, seed = NULL)

print(x, ...)

summary(object, ...)

fitted(object, mth = c("centers", "classes"), ...)
```

- Argumentos

  - data: Conjunto de datos con variables métricas.
  - nclus: Número de clusters.
  - ndim: Dimensionalidad de la solución.
  - method: RKM para las k-medias reducidas y FKM para las k-medias factoriales, por defecto es `RKM`.
  - alpha: Se ajusta por la importancia relativa de RKM y FKM en la función objetivo; alfa = 0,5 conduce a K-medias reducidas, alfa = 0 a K-medias factoriales y alfa = 1 se reduce al método 'tandem' (Componentes principales seguido de k-medias)
  - center: Valor lógico que indica si las variables deberían estar centradas (el valor por defecto es `TRUE`).
  - scale: Valor lógico que indica si las variables deberían escalarse para tener varianza 1 (por defecto `TRUE`).
  - rotation: Método a usar para rotar los factores. Si no se requiere rotar la opción es `none`,`varimax` para la rotación 'varimax' con normalización Kaiser, y `promax` para la rotación 'promax' (por defecto `none`).

- nstart: Partición inicial (por defecto 100)
- smartStart: Solución inicial, si el valor es `NULL` se genera un vector aleatorio con los miembros de los cluster. También se puede proporcionar tal vector para la solución inicial.
- seed: Un número entero que 'set.seed()' utiliza como argumento para compensar el generador de números aleatorios cuando smartStart = NULL. El valor por defecto es `NULL`.
- ... Para más opciones véase la ayuda de R (no se usan en el desarrollo de este trabajo).

- Ejemplo con el conjunto de datos macro
  - K-medias reducidas con 3 grupos en 2 dimensiones después de 10 soluciones iniciales aleatorias.

```
data(macro)
set.seed(4756)
RKM = cluspca(macro, 3, 2, method = "RKM", rotation = "varimax",
              scale = FALSE, nstart = 10)
```

```
summary(RKM)
```

```
## Solution with 3 clusters of sizes 11 (55%), 7 (35%), 2 (10%)
## in 2 dimensions. Variables were mean centered and unstandardized.
##
## Cluster centroids:
##              Dim.1   Dim.2
## Cluster 1   1.2234 -3.4508
## Cluster 2   1.6791  5.3908
## Cluster 3 -12.6058  0.1118
##
## Variable scores:
##       Dim.1   Dim.2
## GDP -0.0361 -0.1311
## LI  -0.1330 -0.1462
## UR   0.0710 -0.5319
## IR  -0.8631 -0.1366
## TB   0.4751 -0.0963
## NNS -0.0726  0.8066
##
## Within cluster sum of squares by cluster:
## [1] 97.7489 76.0490 27.3520
##  (between_SS / total_SS =  77.39 %)
##
## Clustering vector:
##   Australia      Canada     Finland      France       Spain
```

```
##           1           1           1           1           1
##         USA Netherlands      Greece      Mexico    Portugal
##           1           2           3           3           2
##     Austria     Belgium     Denmark     Germany       Italy
##           2           2           1           1           1
##       Japan      Norway Switzerland          UK      Sweden
##           2           2           2           1           1
##
## Objective criterion value: 446.2309
##
## Available output:
##
##  [1] "obscoord"  "attcoord"  "centroid"  "cluster"   "criterion" "size"
##  [7] "odata"     "scale"     "center"    "nstart"
```

La solución contiene varios valores:

- Coordenadas de la muestra

```
RKM$obscoord
```

Tabla 4.3: Coordenadas de la muestra

Australia	-1.0584971	-5.1681273
Canada	0.4553695	-3.8977757
Finland	5.6021298	-3.7690854
France	3.0499438	-3.6177276
Spain	0.8486124	-5.9743441
Sweden	4.0050680	-5.4530046
USA	-1.7558191	-0.1176771
Netherlands	3.2318270	6.7186737
Greece	-9.8294356	-2.3310583
Mexico	-15.3822261	2.5546929
Portugal	-3.8259770	7.5930243
Austria	0.1893509	2.3840256
Belgium	2.9977218	1.7968923
Denmark	2.0810112	-2.0110873
Germany	2.0943440	-1.3093115
Italy	1.3512320	-3.2315933
Japan	2.6687983	6.9271056
Norway	2.7401596	6.8154791
Switzerland	3.7519833	5.5001930
UK	-3.2155967	-3.4092943

- Los coeficientes de las componentes principales

```
RKM$attcoord
```

Tabla 4.4: Coeficientes de las componentes principales

GDP	-0.0360671	-0.1310725
LI	-0.1329819	-0.1462009
UR	0.0710408	-0.5318615
IR	-0.8631145	-0.1366110
TB	0.4751159	-0.0963140
NNS	-0.0725701	0.8066162

- Los centroides de los grupos

```
RKM$centroid
```

Tabla 4.5: Centroides de los grupos

1.223436	-3.4508208
1.679123	5.3907705
-12.605831	0.1118173

- Pertenencia a cada grupo

```
RKM$cluster
```

```
##   Australia      Canada     Finland      France       Spain
##           1           1           1           1           1
##      Sweden         USA Netherlands      Greece      Mexico
##           1           1           2           3           3
##    Portugal     Austria     Belgium     Denmark     Germany
##           2           2           2           1           1
##       Italy       Japan      Norway Switzerland          UK
##           1           2           2           2           1
```

- Valor óptimo de la función objetivo

```
RKM$criterion
```

```
## [1] 446.2309
```

- Número de observaciones en cada grupo

```
RKM$size
```

```
## [1] 11  7  2
```

- Escala que se ha aplicado

```
RKM$scale
```

```
## [1] FALSE
```

- Centrado que se ha aplicado

```
RKM$center
```

```
## [1] TRUE
```

- Partición inicial que se ha aplicado

```
RKM$nstart
```

```
## [1] 10
```

- `odata` devuelve los datos iniciales

Es decir, hay 3 grupos, uno con 11 observaciones (55 %), otro con 7 (35 %), y por último con 2 observaciones (10 %) en 2 dimensiones. La suma de cuadrados entre los grupos por cada uno de ellos es respectivamente 97.7489, 76.0490 y 27.3520. El valor de la función objetivo, 446.2309.

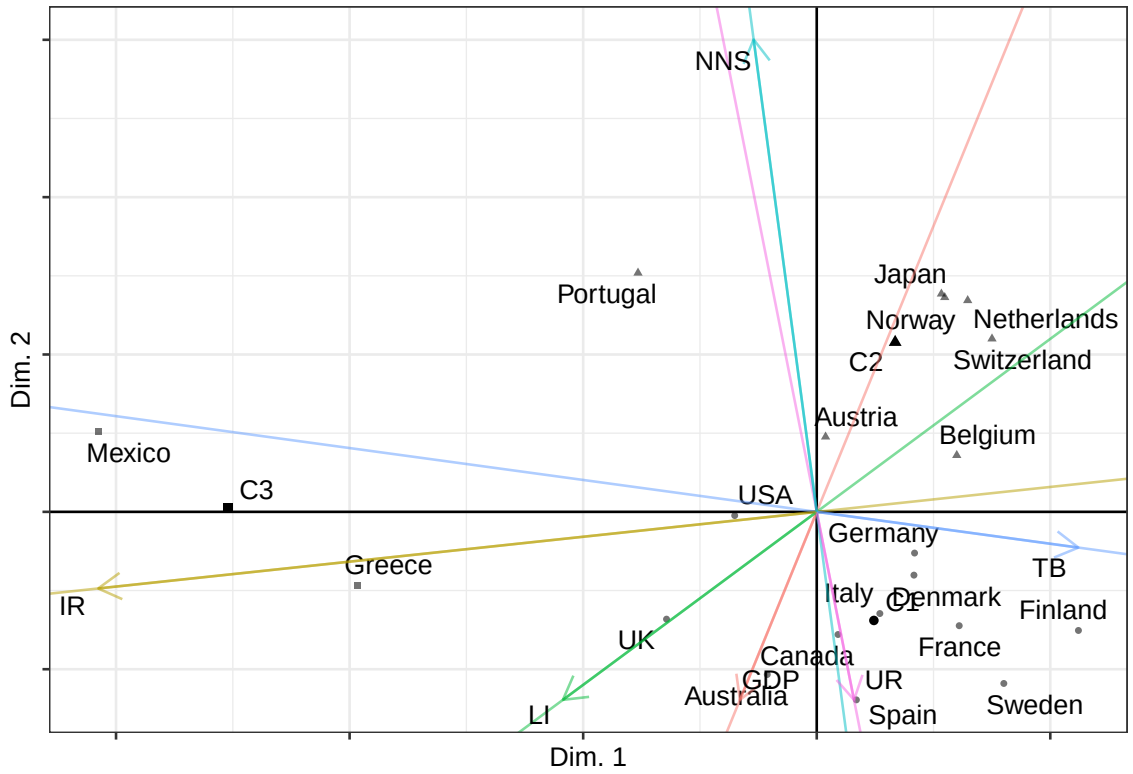
Los centroides de los grupos son los vectores (1.2234, 1.6791, -12.6058) para la dimensión 1 y (-3.4508, 5.3908, 0.1118) para la dimensión 2.

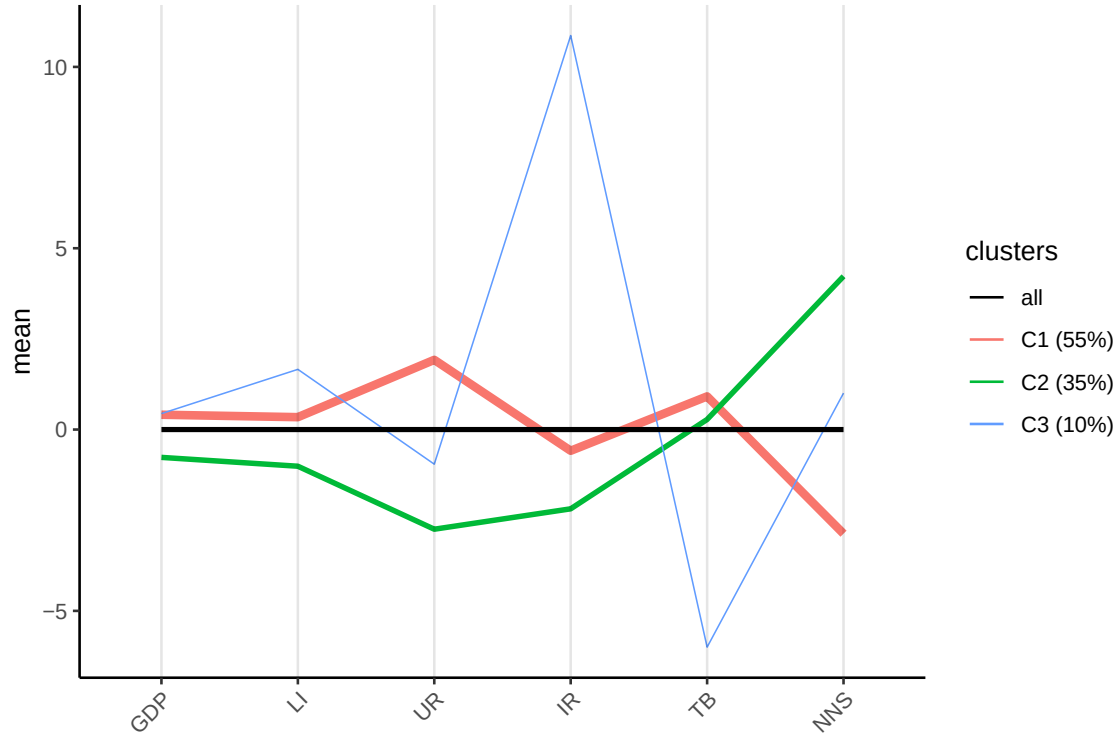
La primera componente es $-0.0361 \cdot \text{GDP} - 0.1330 \cdot \text{LI} + 0.0710 \cdot \text{UR} - 0.8631 \cdot \text{IR} + 0.4751 \cdot \text{TB} - 0.0726 \cdot \text{NNS}$.

La segunda componente es $-0.1311 \cdot \text{GDP} - 0.1462 \cdot \text{LI} - 0.5319 \cdot \text{UR} - 0.1366 \cdot \text{IR} - 0.0963 \cdot \text{TB} + 0.8066 \cdot \text{NNS}$.

Se puede obtener un diagrama de dispersión junto con un gráfico de coordenadas de los grupos que facilite su descripción, añadiendo la opción `cludesc=TRUE`.

```
plot(RKM,cludesc = TRUE)
```





Del primer gráfico se concluye que la dimensión 1 diferencia significativamente al cluster 3 de los demás, en el cual están Grecia y México. Además, esta dimensión se caracteriza principalmente por las variables IR y TB, a menor tasa de interés mayor es el valor de la componente, y a mayor balance comercial mayor es dicho valor.

La dimensión 2 diferencia significativamente al cluster 1 del 2. Está influenciada sobre todo por las variables UR y NNS, a mayor tasa de ahorro mayor es el valor de la componente, y a menor tasa de desempleo mayor es dicho valor.

El segundo gráfico describe las medias de los grupos para cada variable y representa una línea más gruesa según el tamaño de los grupos. El primer cluster se caracteriza por un pequeño crecimiento de la tasa de desempleo y un descenso de la tasa nacional neta de ahorros. El segundo contiene a las ciudades con un descenso de la tasa de desempleo y crecimiento de la tasa nacional neta de ahorros. Por último, el tercer cluster representa a las ciudades con un gran incremento de la tasa de interés, y un gran descenso de la balanza comercial.

- K-medias factoriales con 3 grupos en 2 dimensiones con solución inicial por k-medias reducidas.

Se obtienen los mismos valores que el caso anterior, esta vez para las k-medias factoriales, lo único que cambia es el método.

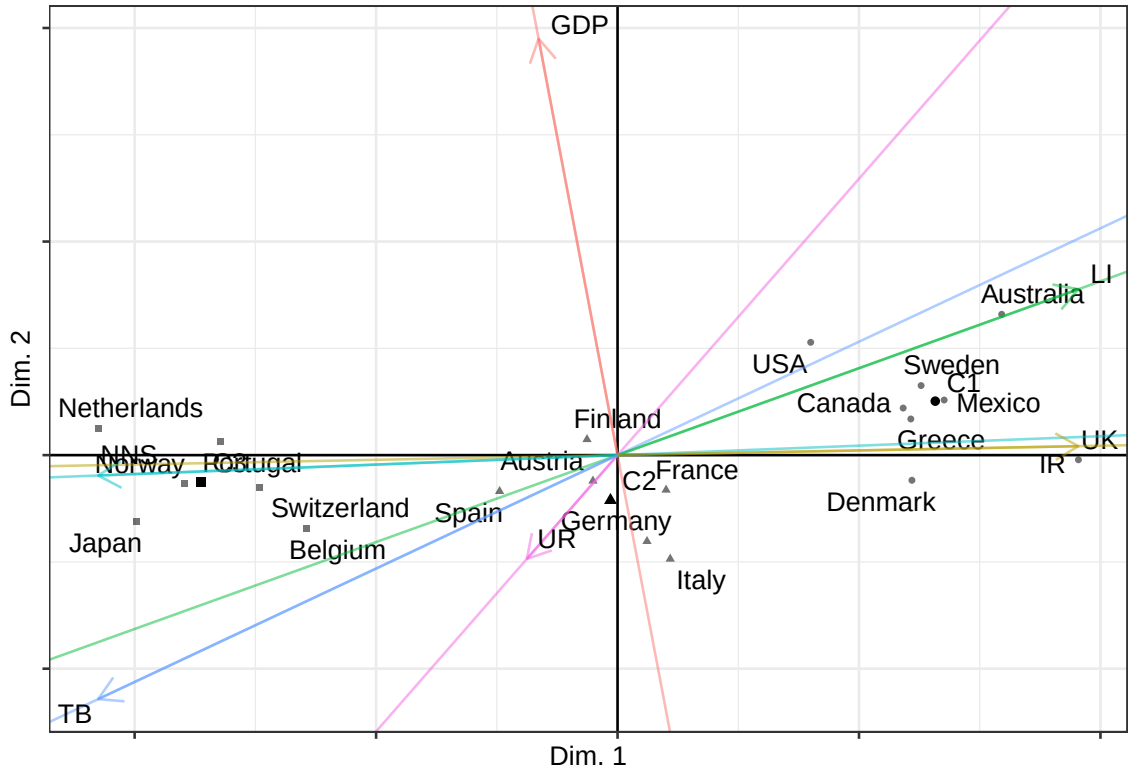
```
set.seed(4756)
FKM = cluspca(macro, 3, 2, method = "FKM", rotation = "varimax",
              scale = FALSE, smartStart = RKM$cluster)
```

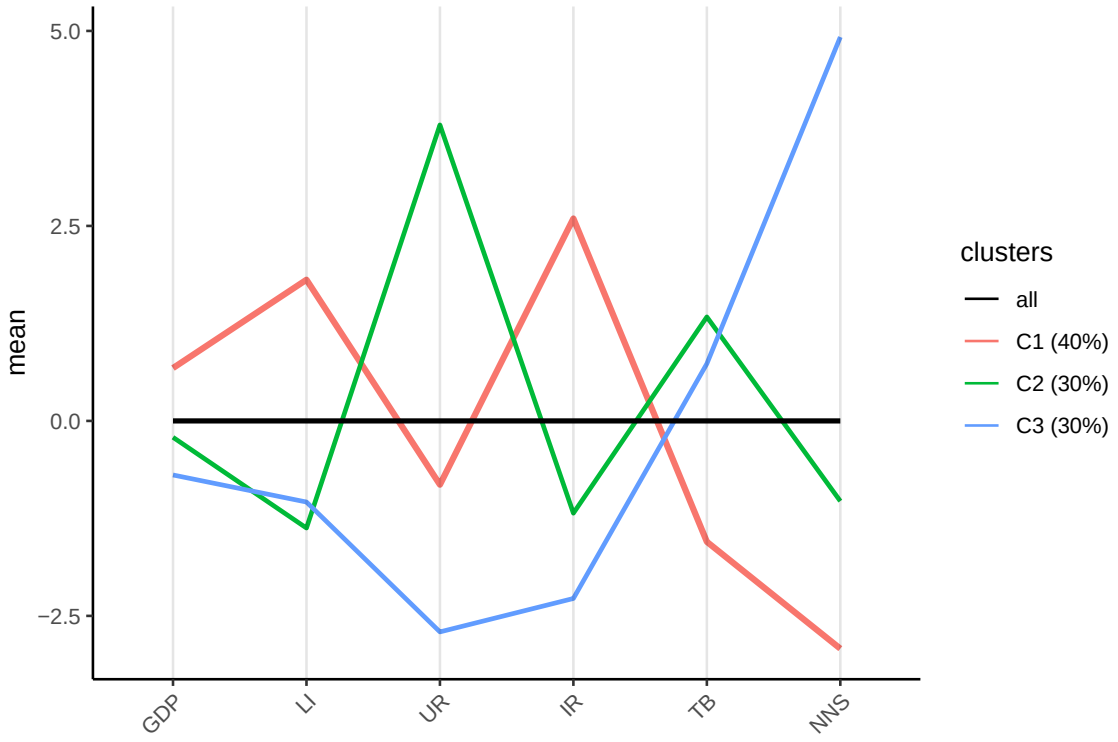
```
summary(FKM)
```

```
## Solution with 3 clusters of sizes 8 (40%), 6 (30%), 6 (30%)
## in 2 dimensions. Variables were mean centered and unstandardized.
##
## Cluster centroids:
##              Dim.1   Dim.2
## Cluster 1  3.9473  1.2634
## Cluster 2 -0.0878 -1.0575
## Cluster 3 -5.1752 -0.6270
##
## Variable scores:
##       Dim.1   Dim.2
## GDP -0.0980  0.9741
## LI   0.1268  0.0860
## UR  -0.0900 -0.1941
## IR   0.3913  0.0143
## TB  -0.0420 -0.0372
## NNS -0.9008 -0.0665
##
## Within cluster sum of squares by cluster:
## [1] 18.2995  8.4129  9.1766
##  (between_SS / total_SS =  89.54 %)
##
## Clustering vector:
```

```
##   Australia      Canada     Finland      France       Spain
##           1           1           2           2           2
##         USA Netherlands      Greece      Mexico    Portugal
##           1           3           1           1           3
##     Belgium     Denmark     Germany       Italy       Japan
##           3           1           2           2           3
## Switzerland          UK      Sweden     Austria      Norway
##           3           1           1           2           3
##
## Objective criterion value: 35.8891
##
## Available output:
##
##  [1] "obscoord"  "attcoord"  "centroid"  "cluster"   "criterion" "size"
##  [7] "odata"     "scale"     "center"    "nstart"
```

```
plot(FKM, cludesc = TRUE)
```





En este caso, el valor óptimo de la función objetivo es mucho menor, aunque en el gráfico es mucho menos aclarativo.

- Método Tandem

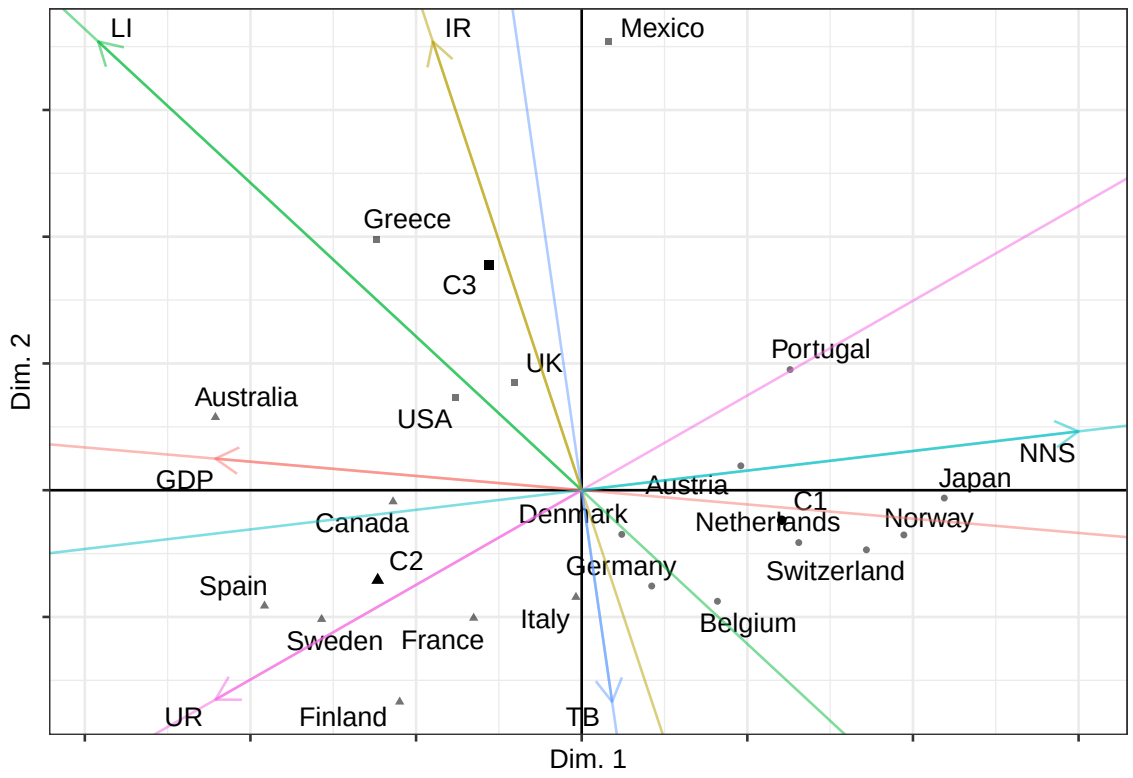
```
Tandem = cluspca(macro, 3, 2, alpha = 1, seed = 1234)
```

```
summary(Tandem)
```

```
## Solution with 3 clusters of sizes 9 (45%), 7 (35%), 4 (20%) in
## 2 dimensions. Variables were mean centered and standardized.
##
## Cluster centroids:
##              Dim.1   Dim.2
## Cluster 1  1.2074 -0.2373
## Cluster 2 -1.2317 -0.7097
## Cluster 3 -0.5612  1.7761
##
## Variable scores:
##       Dim.1   Dim.2
## GDP -0.5692  0.0640
## LI  -0.1893  0.2296
## UR  -0.4927 -0.3681
## IR  -0.1765  0.6953
## TB   0.0618 -0.5618
```

```
## NNS  0.6020  0.0931
##
## Within cluster sum of squares by cluster:
## [1] 5.9599 6.5373 6.1347
##  (between_SS / total_SS =  69.09 %)
##
## Clustering vector:
##   Australia      Canada     Finland      France       Spain
##           2           2           2           2           2
##         USA Netherlands      Greece      Mexico    Portugal
##           3           1           3           3           1
##     Belgium     Denmark     Germany       Italy       Japan
##           1           1           1           2           1
## Switzerland          UK      Sweden     Austria      Norway
##           1           3           2           1           1
##
## Objective criterion value: 53.7194
##
## Available output:
##
##  [1] "obscoord"  "attcoord"  "centroid"  "cluster"   "criterion" "size"
##  [7] "odata"     "scale"     "center"    "nstart"
```

```
plot(Tandem)
```



4.2.2. global_bootclus

- Descripción

Evalúa la estabilidad global de los métodos conjuntos de reducción de la dimensión y agrupamiento mediante bootstrap.(véase Estabilidad de los grupos)

- Uso

```
global_bootclus(data, nclusrange = 3:4, ndim = NULL,
method = c("RKM","FKM","mixedRKM","mixedFKM","clusCA","MCAk","iFCB"),
nboot = 10, alpha = NULL, alphak = NULL, center = TRUE,
scale = TRUE, nstart = 100, smartStart = NULL, seed = NULL)
```

- Argumentos
  - data: Conjunto de datos.
  - nclusrange: Número entero o vector de números enteros con el número de clusters o el rango de números de clusters (debe ser mayor que 1).
  - ndim: Dimensionalidad de la solución, si es `NULL` se usa el número de clusters - 1.
  - method: `RKM` para las k-medias reducidas, FKM para las k-medias factoriales, `mixedRKM` para las K-medias reducidas mixtas, `mixedFKM` para las k-medias factoriales mixtas.
  - nboot: Número de pares de particiones bootstrap.
  - alpha: Se ajusta por la importancia relativa de RKM y FKM en la función objetivo; alpha = 0,5 conduce a K-medias reducidas, alpha = 0 a K-medias factoriales.
  - center: Valor lógico que indica si las variables deberían estar centradas (el valor por defecto es `TRUE`).
  - scale: Valor lógico que indica si las variables deberían escalarse para tener varianza 1 (por defecto `TRUE`).
  - nstart: Partición inicial (por defecto 100)
  - smartStart: Solución inicial, si el valor es `NULL` se genera un vector aleatorio con los miembros de los cluster. También se puede proporcionar tal vector para la solución inicial.
  - seed: Un número entero que 'set.seed()' utiliza como argumento para compensar el generador de números aleatorios cuando smartStart = NULL. El valor por defecto es `NULL`.
  - ... Para más opciones véase la ayuda de R (no se usan en el desarrollo de este trabajo).
- Ejemplo con el conjunto de datos macro
  - 5 replicas bootstrap y 10 particiones iniciales

```
data(macro)
boot_RKM = global_bootclus(macro, nclusrange = 2:5,
method = "RKM", nboot = 5, nstart = 10, seed = 1234)
```

De la solución se obtiene:

- Número entero o vector de números enteros con el número de clusters o el rango de números de clusters

```
boot_RKM$nclusrange
```

```
## [1] 2 3 4 5
```

- Las particiones que resultan de aplicar el método (las que aparecen en teoría, $C_i^S(x_j), C_i^\top(x_j)$ Estabilidad de los grupos).

```
boot_RKM$clust1
```

```
boot_RKM$clust1[,,1]
```

```
##       [,1] [,2] [,3] [,4]
##  [1,]    2    3    2    2
##  [2,]    1    2    2    2
##  [3,]    1    1    1    1
##  [4,]    1    1    1    1
##  [5,]    1    1    1    1
##  [6,]    1    1    1    1
##  [7,]    2    3    2    2
##  [8,]    1    2    3    4
##  [9,]    2    3    4    5
## [10,]    2    3    4    5
## [11,]    2    3    4    5
## [12,]    1    2    3    3
## [13,]    1    2    3    3
## [14,]    1    2    2    2
## [15,]    1    2    3    3
## [16,]    1    1    1    1
## [17,]    1    2    3    4
## [18,]    1    2    3    4
## [19,]    1    2    3    4
## [20,]    2    3    2    2
```

```
boot_RKM$clust2
```

```
boot_RKM$clust2[,,5]
```

```
##       [,1] [,2] [,3] [,4]
##  [1,]    2    2    2    3
##  [2,]    2    2    2    3
##  [3,]    2    2    2    3
##  [4,]    2    2    2    3
##  [5,]    2    2    2    3
##  [6,]    2    2    2    3
##  [7,]    1    1    1    2
##  [8,]    1    1    1    1
##  [9,]    2    2    4    5
## [10,]    1    3    3    4
## [11,]    1    1    4    5
## [12,]    1    1    1    2
## [13,]    1    1    1    1
## [14,]    1    1    1    2
## [15,]    2    1    1    2
## [16,]    2    2    2    3
## [17,]    1    1    1    1
## [18,]    1    1    1    1
## [19,]    1    1    1    1
## [20,]    2    2    2    2
```

- Indices de las filas originales de los datos en las muestras bootstrap S_i, T_i. (véase Estabilidad de los grupos)

```
boot_RKM$index1
```

```
##       [,1] [,2] [,3] [,4] [,5]
##  [1,]   12   17   11   20   11
##  [2,]   15   17    4   19    1
##  [3,]    9    8    3    7    8
##  [4,]    5   17    4   17    8
##  [5,]    6    8   10    3   10
##  [6,]   16   10    6    2    5
##  [7,]    4   15    6    5   20
##  [8,]    2    3    9   15   12
##  [9,]    7    9   11   15   10
## [10,]    6   16   19   10   20
## [11,]   15    3   14   17    1
## [12,]   14   10    8    3    6
## [13,]   20   13   17   16    3
## [14,]   14    3   13    9    4
## [15,]    4   19   16   20    4
## [16,]    4   18    2   12   11
```

```
## [17,]    8    6    6    5    3
## [18,]   20   20   11   17    1
## [19,]    3    9    6    3   19
## [20,]    4    7   19   15    1
```

```
boot_RKM$index2
```

```
##       [,1] [,2] [,3] [,4] [,5]
##  [1,]    5   19   15    5   12
##  [2,]    2    6   14    6   19
##  [3,]   15    5   19   18   10
##  [4,]    8   12   19   19   13
##  [5,]   20    8    7   18   12
##  [6,]   16   16    9   11    1
##  [7,]   12    1   16   16   13
##  [8,]    3    4    8    6   10
##  [9,]    9   19    3    7   13
## [10,]   19    9   16    8    8
## [11,]    4   20    2    4   18
## [12,]    8   17    5   11    2
## [13,]   10    6   16   15    6
## [14,]   11    6    9   17   14
## [15,]    2   13    3   11    6
## [16,]   15   17    8    7    5
## [17,]   17    2    3   10   11
## [18,]    6    6    8   11   14
## [19,]   19    2   19    3    9
## [20,]    6   16    4   18   20
```

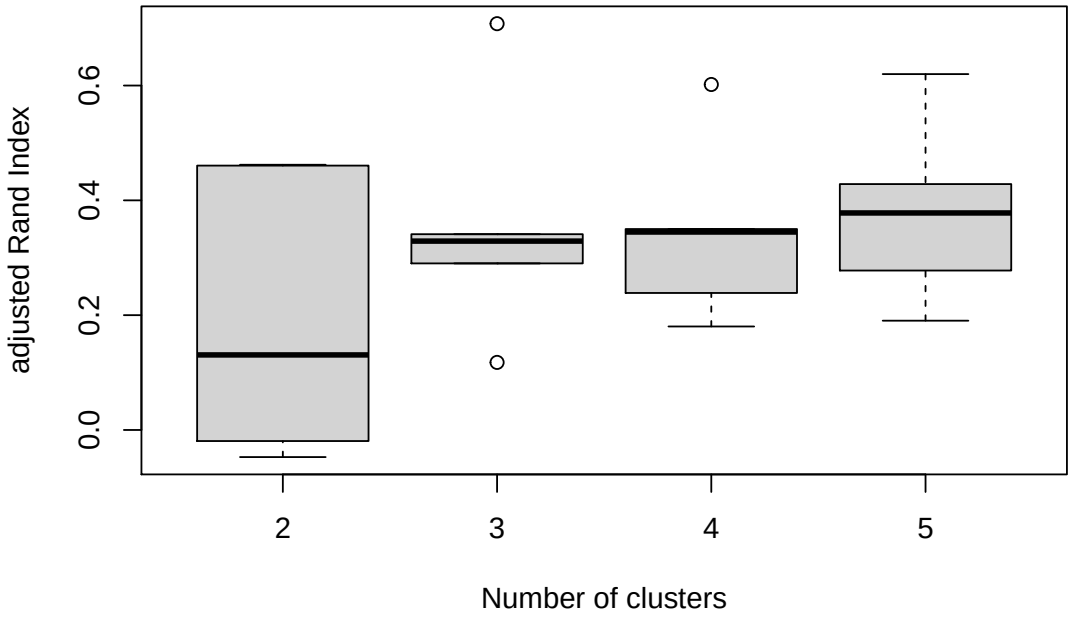
- Los valores del índice de Rand ajustado.

```
boot_RKM$rand
```

```
##                2         3         4         5
## [1,] -0.04721078 0.3410133 0.3500076 0.6199339
## [2,]  0.46052778 0.7077298 0.6018211 0.1903843
## [3,]  0.13071895 0.3289836 0.1802962 0.2777347
## [4,] -0.01931330 0.2901619 0.2386565 0.3780426
## [5,]  0.46196557 0.1176804 0.3448276 0.4283290
```

Se puede dibujar un gráfico de cajas para interpretarlo:

```
boxplot(boot_RKM$rand, xlab = "Number of clusters", ylab =
"adjusted Rand Index")
```



Lo ideal es que el índice sea 1 o muy cercano, en el gráfico se observa que para 5 de grupos, la media de los índices es mayor respecto a las demás, aunque en otros casos haya outliers más cercanos a 1.

También se puede medir la estabilidad por conglomerado, con la función análoga a esta, "local_bootclus".

4.2.3. plot.cluspca

- Descripción

Esta función crea un diagrama de dispersión de las observaciones, un círculo de correlación de las variables o un biplot de las observaciones y las variables. Opcionalmente, devuelve un gráfico de coordenadas paralelas que muestra las medias de los grupos.

- Uso

```
plot(x, dims = c(1, 2), cludesc = FALSE,
what = c(TRUE,TRUE), attlabs, max.overlaps=10, ...)
```

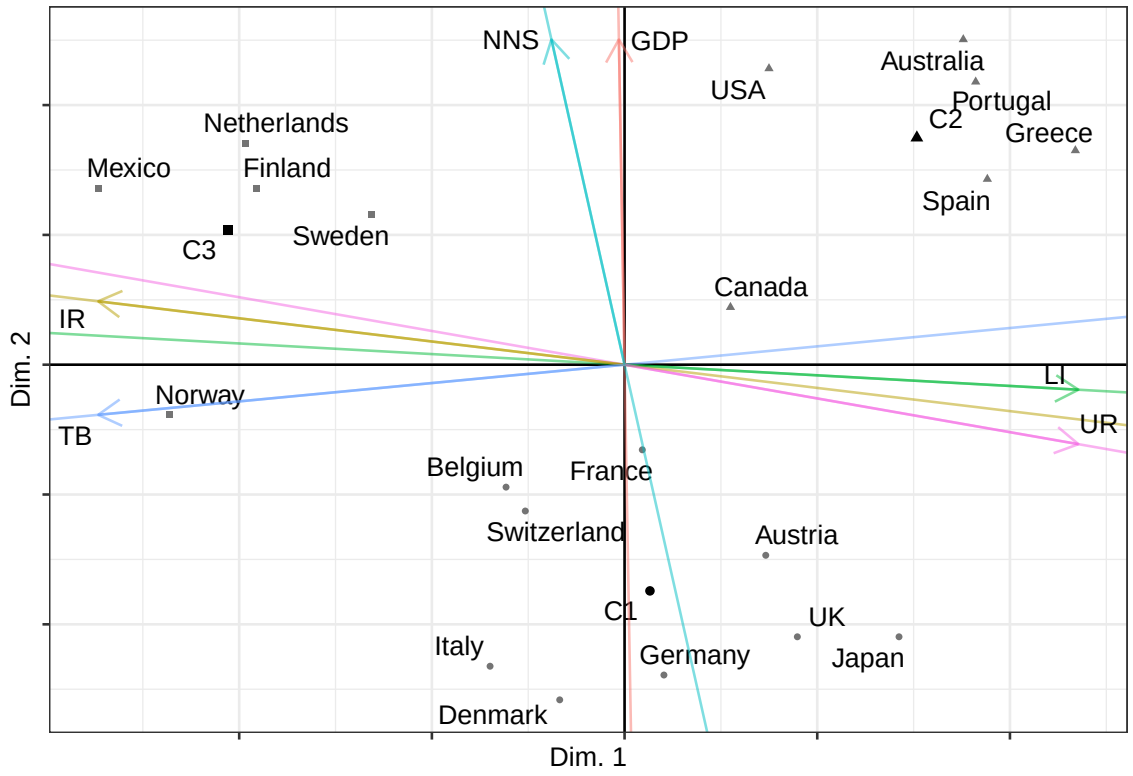
- Argumentos
  - x: Objeto que devuelve la función cluspca().
  - dims: Vector númerico de longitud 2 indicando las dimensiones a dibujar en los ejes horizontal y vertical respectivamente (por defecto se dibuja la primera dimensión en el horizontal y la segunda en el vertical).

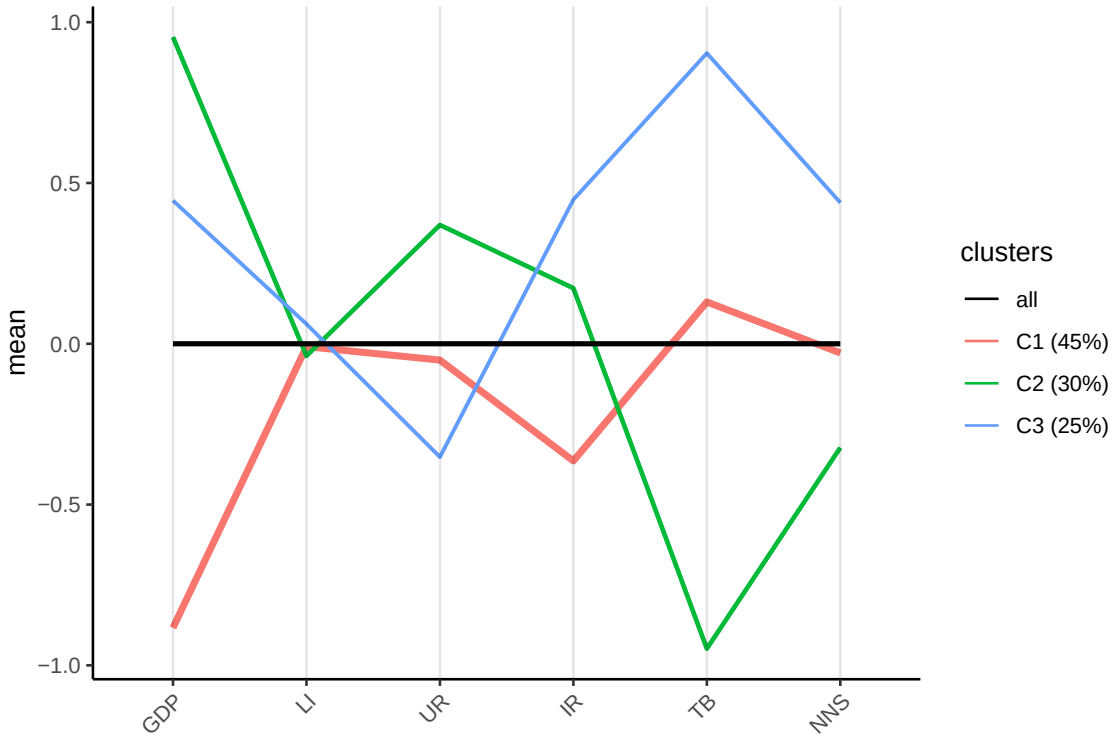
- what: Vector de 2 valores lógicos indicando el contenido de los gráficos.El primero contiene si se dibuja el diagrama de dispersión y los centroides, el segundo si se dibuja el círculo de correlación (por defecto `c(TRUE,TRUE)`).
- cludesc: Valor lógico indicando si se crea el gráficco de coordenadas paralelas (por defecto `FALSE`).
- ... Para más opciones véase la ayuda de R (no se usan en el desarrollo de este trabajo).

- Ejemplo con el conjunto de datos macro

```
data("macro")
outFKM = cluspca(macro, 3, 2, method = "FKM", rotation = "varimax")
```

```
plot(outFKM, cludesc = TRUE)
```





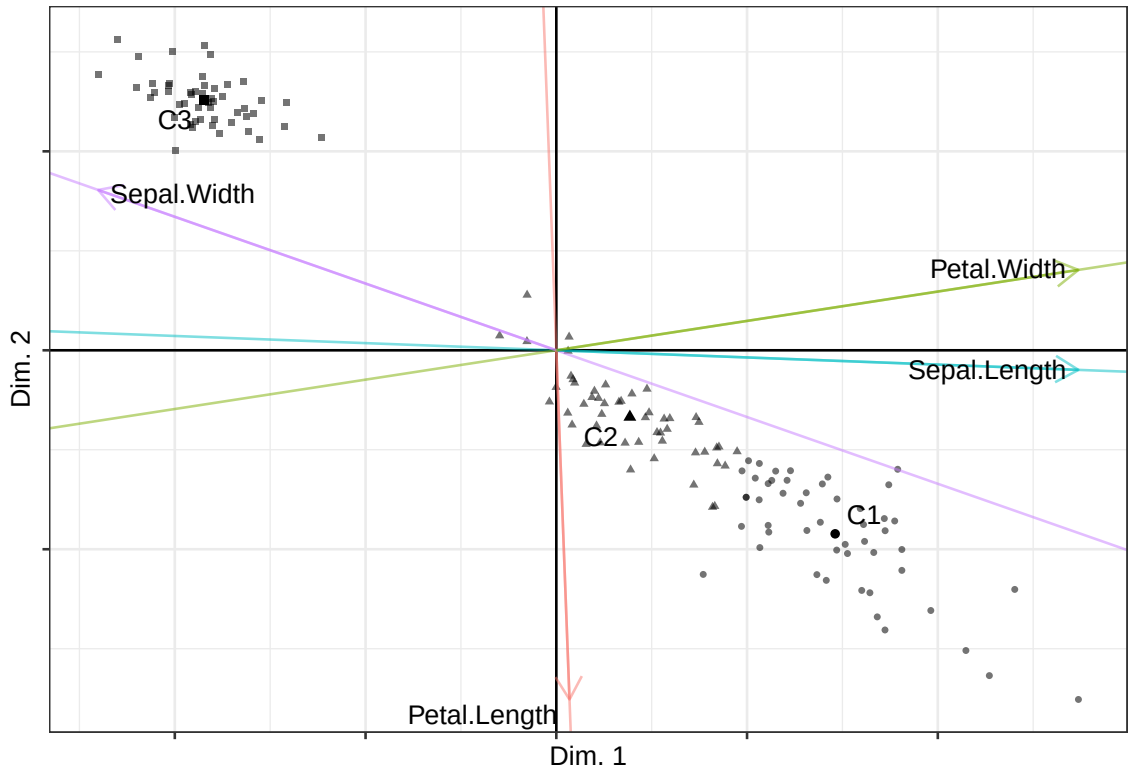
- Ejemplo con el conjunto de datos iris

```
data("iris", package = "datasets")
outclusPCA = cluspca(iris[,-5], 3, 2, alpha = 0.3, rotation = "varimax")
```

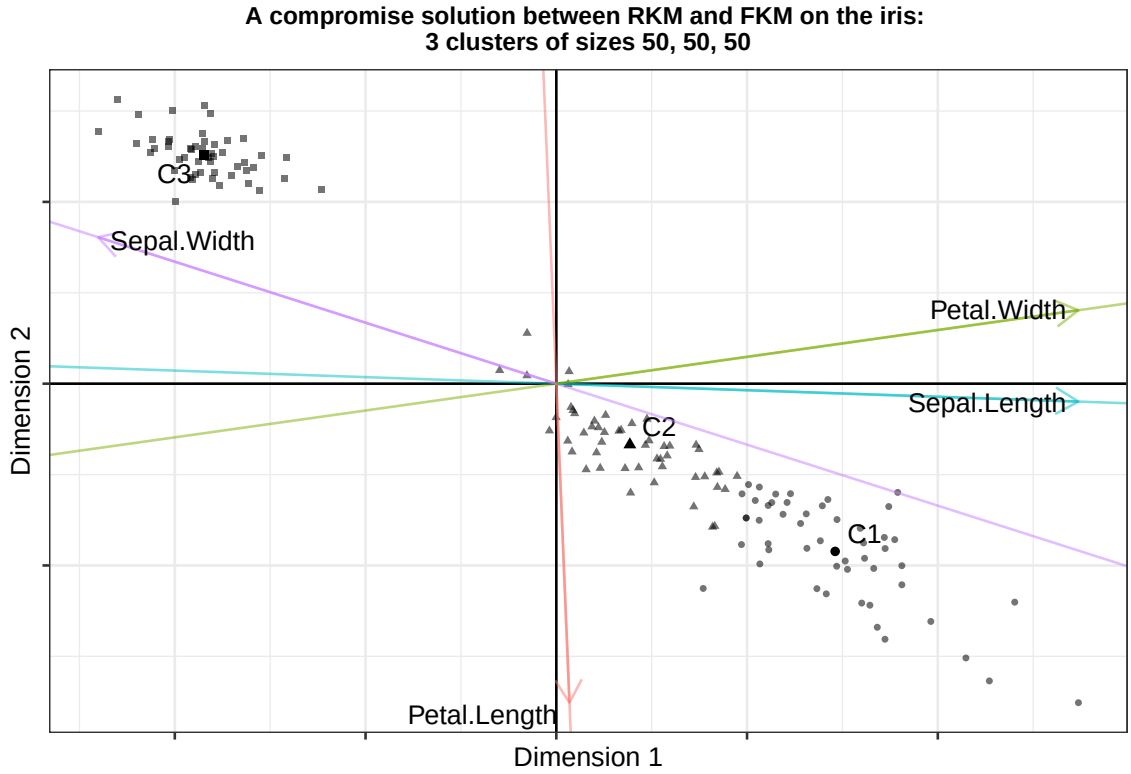
```
table(outclusPCA$cluster,iris[,5])
```

```
##
##     setosa versicolor virginica
##   1      0          6        44
##   2      0         44         6
##   3     50          0         0
```

```
map = plot(outclusPCA)$map
```




```
map + ggtitle(paste("A compromise solution between RKM and FKM on the iris:
3 clusters of sizes ", paste(outclusPCA$size,
collapse = ", "),sep = "")) + xlab("Dimension 1") + ylab("Dimension 2") +
theme(plot.title = element_text(size = 10, face = "bold", hjust = 0.5))
```



4.2.4. tuneclus

- Descripción

Ayuda en la elección del número de grupos y dimensiones apropiados para los métodos conjuntos de reducción de la dimensión y conglomerados.

- Uso

```
tuneclus(data, nclusrange = 3:4, ndimrange = 2:3,
method = c("RKM","FKM","mixedRKM","mixedFKM","clusCA","iFCB","MCAk"),
criterion = "asw", dst = "full", alpha = NULL, alphak = NULL,
center = TRUE, scale = TRUE, rotation = "none", nstart = 100,
smartStart = NULL, seed = NULL)

print(x, ...)

summary(object, ...)

fitted(object, mth = c("centers", "classes"), ...)
```

- Argumentos
  - data: Conjunto de datos.

- nclusrange: Un vector entero con el rango de números de conglomerados que se van a comparar según los criterios de validez del conglomerado.
  El número de grupos debe ser mayor que uno.
- ndimrange: Un vector entero con el rango de dimensiones que se van a comparar según los criterios de validez del conglomerado.
- method: `RKM` para las k-medias reducidas, FKM para las k-medias factoriales, `mixedRKM` para las K-medias reducidas mixtas, `mixedFKM` para las k-medias factoriales mixtas.
- criterion: Determina si se utiliza el ancho de silueta promedio, el índice de Calinski-Harabasz o el valor objetivo del método seleccionado. (`asw`,`ch` o `crit`)
- dst: Datos usados para las distancias entre objetos. `full` para los datos originales, `low` para las puntuaciones de objetos en el espacio de baja dimensión.
- alpha: Se ajusta por la importancia relativa de RKM y FKM en la función objetivo; alpha =1 se reduce a las componentes principales, alpha = 0,5 conduce a K-medias reducidas, alpha = 0 a K-medias factoriales.
- center: Valor lógico que indica si las variables deberían estar centradas (el valor por defecto es `TRUE`).
- scale: Valor lógico que indica si las variables deberían escalarse para tener varianza 1 (por defecto `TRUE`).
- rotation: Método a usar para rotar los factores. Si no se requiere rotar la opción es `none`,`varimax` para la rotación 'varimax' con normalización Kaiser, y `promax` para la rotación 'promax' (por defecto `none`).
- nstart: Partición inicial (por defecto 100)
- smartStart: Solución inicial, si el valor es `NULL` se genera un vector aleatorio con los miembros de los cluster. También se puede proporcionar tal vector para la solución inicial.
- seed: Un número entero que 'set.seed()' utiliza como argumento para compensar el generador de números aleatorios cuando smartStart = NULL. El valor por defecto es `NULL`.
- ... Para más opciones véase la ayuda de R (no se usan en el desarrollo de este trabajo).

- Ejemplo con el conjunto de datos macro.

Vamos a usar k-medias reducidas para un rango de conglomerados y dimensiones.

```
data(macro)
bestRKM = tuneclus(macro, 3:4, 2:3, method = "RKM",
criterion = "asw", dst = "low", nstart = 1, seed = 1234)
```

```
bestRKM
```

```
##
## The best solution was obtained for 4 clusters of sizes 8 (40%),
## 7 (35%), 4 (20%), 1 (5%) in 2 dimensions, for an average Silhouette
## width value of 0.447. Variables were mean centered and standardized.
##
## Cluster quality criterion values across the specified range of
## clusters (rows) and dimensions (columns):
##      X2    X3
## 3 0.438
## 4 0.447 0.366
##
## The average Silhouette width values of each cluster are:
## [1] 0.44 0.30 0.82 0.00
##
## Cluster centroids:
##              Dim.1   Dim.2
## Cluster 1  0.2276  0.7846
## Cluster 2 -1.1795  0.0504
## Cluster 3  1.8944 -0.7067
## Cluster 4 -1.1418 -3.8030
##
## Within cluster sum of squares by cluster:
## [1] 3.2991 5.0376 0.2820 0.0000
##  (between_SS / total_SS =  84.56 %)
##
## Objective criterion value: 33.3921
##
## Available output:
##
## [1] "clusobjbest" "nclusbest"   "ndimbest"    "critbest"  "critgrid"
## [6] "crit"        "cluasw"
```

La solución proporciona los siguientes valores:

- La salida de aplicar la función `cluspca()` a los valores óptimos obtenidos

```
bestRKM$clusobjbest
```

```
## Solution with 4 clusters of sizes 8 (40%), 7 (35%), 4 (20%),
## 1 (5%) in 2 dimensions. Variables were mean centered and
## standardized.
##
## Cluster centroids:
##              Dim.1   Dim.2
## Cluster 1  0.2276  0.7846
## Cluster 2 -1.1795  0.0504
## Cluster 3  1.8944 -0.7067
```

```
## Cluster 4 -1.1418 -3.8030
##
## Variable scores:
##       Dim.1   Dim.2
## GDP -0.4437  0.0233
## LI  -0.0292 -0.3927
## UR  -0.2886  0.4374
## IR  -0.4475 -0.6508
## TB   0.5334  0.1120
## NNS  0.4840 -0.4667
##
## Within cluster sum of squares by cluster:
## [1] 3.2991 5.0376 0.2820 0.0000
##  (between_SS / total_SS =  84.56 %)
##
## Objective criterion value: 33.3921
##
## Available output:
##
##  [1] "obscoord"  "attcoord"  "centroid"  "cluster"   "criterion"
##  [6] "size" "odata"     "scale"     "center"    "nstart"
```

- El número óptimo de conglomerados

```
bestRKM$nclusbest
```

```
## [1] 4
```

- El número óptimo de dimensiones

```
bestRKM$ndimbest
```

```
## [1] 2
```

- El valor óptimo de la función objetivo para ese número de grupos y dimensiones

```
bestRKM$critbest
```

```
## [1] 0.447
```

- Matriz con el número de grupos como número de filas, y número de dimensión como número de columnas, con los valores de la función objetivo para cada rango de grupos y dimensiones (los valores se calculan solo cuando el número de conglomerados es mayor que el número de dimensiones; de lo contrario, los valores en la cuadrícula se dejan en blanco)

```
bestRKM$critgrid
```

```
##      X2    X3
## 3 0.438
## 4 0.447 0.366
```

- Criterio usado

```
bestRKM$crit
```

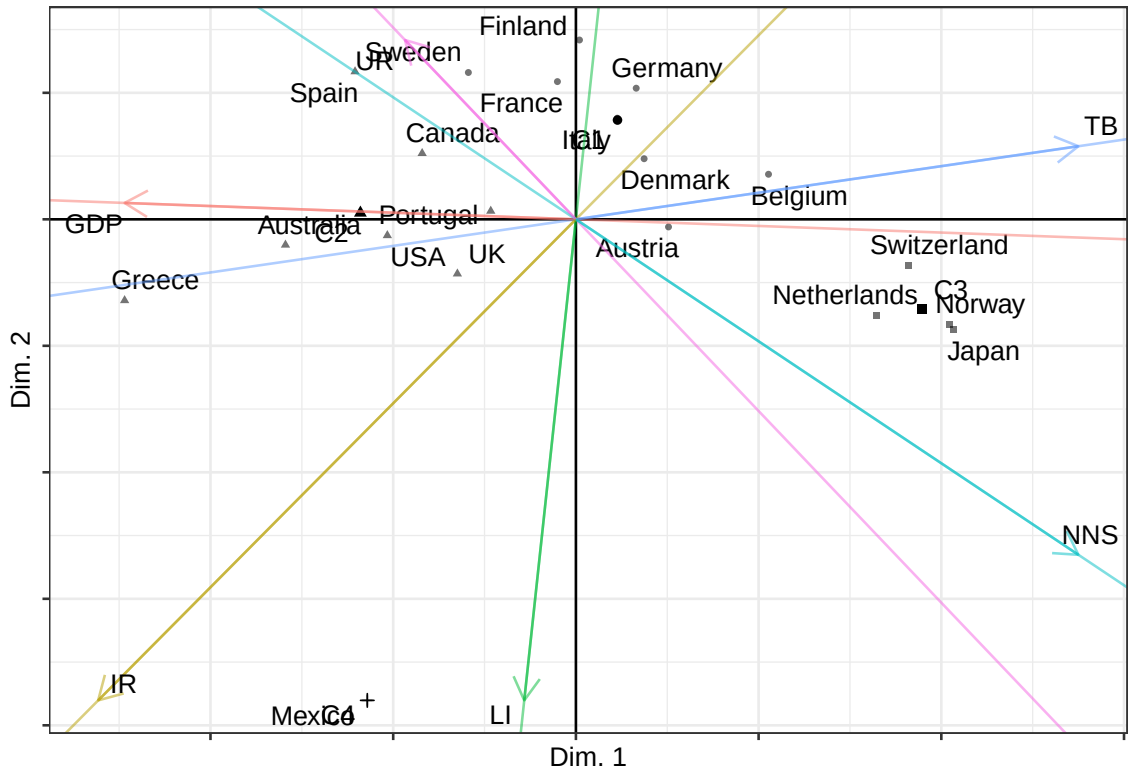
```
## [1] "asw"
```

- Si el criterio es `asw`, los valores de la anchura de la silueta media de cada cluster

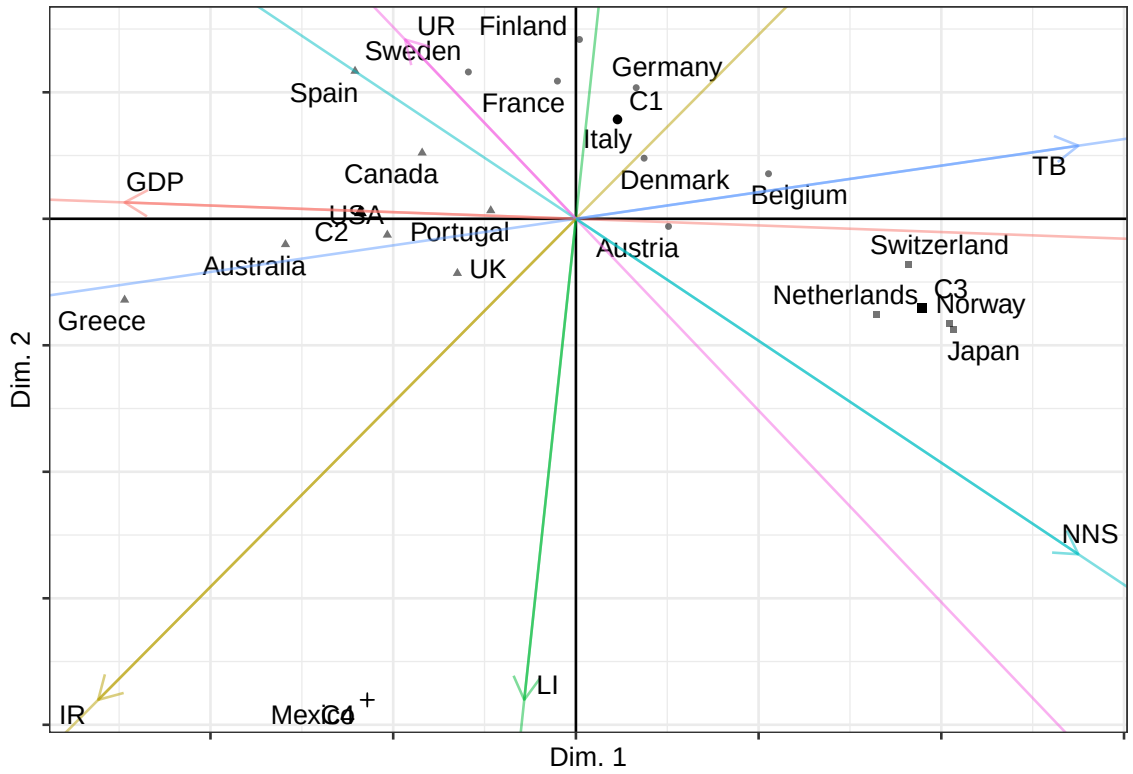
```
bestRKM$cluasw
```

```
##         1         2         3         4
## 0.4449866 0.2998814 0.8183656 0.0000000
```

```
plot(bestRKM)
```




```
## $map
```


UR
Finland
Sweden
Germany
Spain
France
C1
Italy
GDP
Canada
Denmark
Belgium
TB
USA
C2
Portugal
Australia
UK
Austria
Switzerland
Netherlands
C3
Norway
Greece
Japan
NNS
LI
IR
Mexico
C4
Dim. 2
Dim. 1

Apéndice A

Apéndice: Resultados teóricos necesarios

A.1. Relación Norma Frobenius y Traza

Sea A una matriz $n \times m$, su norma de Frobenius viene dada por,

$$\|A\|_F := (\sum_{i=1}^{n}\sum_{j=1}^{m}|a_{ij}|^2)^{1/2}\|A\|_F^2 = \sum_{i=1}^{n}(\sum_{j=1}^{m}a_{ij}a_{ij})$$

Si tenemos en cuenta,

$$A = (a_{ij})_{1\leq i\leq n, 1\leq j\leq m}A^\top = (a_{ji})_{1\leq i\leq n, 1\leq j\leq m} = (a_{ij}^*)$$

Entonces, $a_{ij} = a_{ji}^*$ y desarollando,

$$\|A\|_F^2 = \sum_{i=1}^{n}(\sum_{j=1}^{m}a_{ij}a_{ji}^*) = \sum_{i=1}^{n}(AA^\top)_{ii} = \mathrm{Tr}(AA^\top) = \mathrm{Tr}(A^\top A)$$

En conclusión,

$$\|A\|_F := (\mathrm{Tr}(A^\top A))^{1/2}$$

A.2. Factorización en autovalores y autovectores

Sea A una matriz cuadrada de dimensión $Q \times Q$, se tiene

$$Av = \lambda v$$

donde λ es un autovalor de A y v es su autovector asociado, por lo que $v^\top Av = v^\top \lambda v$. Además, traza$(A) = \sum_{i=1}^{Q}\lambda_i$.

Si se denomina a la matriz de todos los autovectores(matriz de cargas) B, y además $B^\top B = I$,

$$\mathrm{Tr}(B^\top AB) = \mathrm{Tr}(A) = \sum_{i=1}^{Q} \lambda_i$$

A.3. Indice de Rand Ajustado

Suponiendo las particiones:

$$P = \{C_1, \ldots, C_i, \ldots\}, \quad P^* = \{C_1^*, \ldots, C_j^*, \ldots\}$$

Considerando, - N el número de casos. - N_{ij} el número de casos de C_j^* que están en C_i. - N_i el número de casos de C_i. - N_j^* el número de casos de C_j^*.

El Indice de Rand Ajustado viene dado por:

$$ARI(P^*, P) = \frac{\sum_{i,j}\binom{N_{ij}}{2} - [\sum_i \binom{N_i}{2}\sum_j \binom{N_j^*}{2}]/\binom{N}{2}}{\frac{1}{2}[\sum_i \binom{N_i}{2} + \sum_j \binom{N_j^*}{2}] - [\sum_i \binom{N_i}{2}\sum_j \binom{N_j^*}{2}]/\binom{N}{2}}$$

Por tanto, $0 \leq ARI \leq 1$, - $ARI = 1$ si $P = P^*$. - $ARI \approx 0$ en caso de partición aleatoria.

Bibliografía

Michel van de Velden Angelos Markos, Alfonso Iodice D'Enza. *clustrd: Methods for Joint Dimension Reduction and Clustering*, 2021. URL https://CRAN.R-project.org/package=clustrd. R package version 1.3.9.

Geert De Soete and J. Douglas Carroll. K-means clustering in a low-dimensional euclidean space. In Edwin Diday, Yves Lechevallier, Martin Schader, Patrice Bertrand, and Bernard Burtschy, editors, *New Approaches in Classification and Data Analysis*, pages 212–219, Berlin, Heidelberg, 1994. Springer Berlin Heidelberg. ISBN 978-3-642-51175-2.

Pedro L. Luque-Calvo. *Escribir un Trabajo Fin de Estudios con R Markdown*, 2017.

Angelos Markos, Alfonso Iodice D'Enza, and Michel van de Velden. Beyond tandem analysis: Joint dimension reduction and clustering in R. *Journal of Statistical Software*, 91(10):1–24, 2019. doi: 10.18637/jss.v091.i10.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL https://www.R-project.org/.

Maurizio Vichi and Henk A. L. Kiers. Factorial k-means analysis for two-way data. *Computational Statistics & Data Analysis*, 37:49–64, 2001.

Michio Yamamoto and Heungsun Hwang. A general formulation of cluster analysis with dimension reduction and subspace separation. *Behaviormetrika*, 41(1):115–129, 2014. doi: 10.2333/bhmk.41.115.