Análisis exploratorio

Andrés Millán Paula Villanueva

03 de enero de 2022

1 Abstract

Se ha realizado un análisis exploratorio de un conjunto de datos. Este dataset recoge 8 indicadores económicos de 13 empresas. Tras estudiar la información recogida en el conjunto, tratando posibles valores perdidos y outliers, se han aplicado dos tipos de técnicas:

- **Análisis univariante numérico y gráfico**. En él, se ha elaborado un análisis descriptivo numérico clásico y un análisis de supuesto de normalidad.
- **Técnicas multivariantes**: se ha estudiado la correlación entre variables, la reducción de la dimensión mediante variables observables y latentes. Además, se ha estudiado la normalidad multivariante de los datos.

Finalmente, se ha construido un clasificador basado en clustering no jerárquico con el fin de estudiar cómo se agrupan las diferentes empresas. Descubrimos que existen 4 grupos diferentes.

2 Introducción

Se ha realizado el análisis exploratorio de los datos contenidos en la base de datos `DB_2`. Esta base de datos contiene un grupo constituido por 13 empresas que se ha clasificado según las puntuaciones obtenidas en 8 indicadores económicos.

Primero, se ha limpiado el dataset de cualquier anomalía posible. Hemos encontrado una instancia probablemente errónea, que contenía valores perdidos e indicadores sin ningún sentido. A continuación, se ha realizado un análisis descriptivo numérico clásico, esto es, se han obtenido las medidas de tendencia central, los cuartiles, el coeficiente de simetría, la dispersión, etc. Además, se han estudiado posibles outliers. Se ha comprobado también la normalidad de las variables individualmente mediante gráficos de normalidad.

Una vez preparado el conjunto inicial, procedemos a realizar el análisis exploratorio multivariante. Se comprobó la correlación entre las variables mediante un test de Bartlett. A continuación, se realizó un estudio de la posibilidad de reducción de la dimensión mediante variables observables, en cuyo caso se ha elegido el número óptimo de componentes principales usando distintas técnicas gráficas, y mediante variables latentes, en cuyo caso se ha elegido el número óptimo de factores a considerar. Lo siguiente fue analizar la normalidad multivariante de los datos con los tests con el paquete MVN.

Finalmente, para completar nuestro objetivo, se ha realizado un análisis cluster, es decir, un agrupamiento de los objetos formando clusters de objetos con un alto grado de homogeneidad interna y heterogeneidad. En concreto, se ha utilizado el método de las k medias, un método no jerárquico.

3 Materiales y métodos

3.1 Materiales

La base de datos elegida contiene un grupo constituido por 13 empresas que se ha clasificado según las puntuaciones obtenidas en 8 indicadores económicos:

- X1: Indicador de volumen de facturación.
- X2: Indicador de nivel de nueva contratación.
- X3: Indicador del total de clientes.
- X4: Indicador de beneficios de la empresa .
- X5: Indicador de retribución salarial de los empleados.
- X6: Indicador de organización empresarial dentro de la empresa.
- X7: Indicador de relaciones con otras empresas.
- X8: Indicador de nivel de equipamiento (ordenadores, maquinaria, etc...).

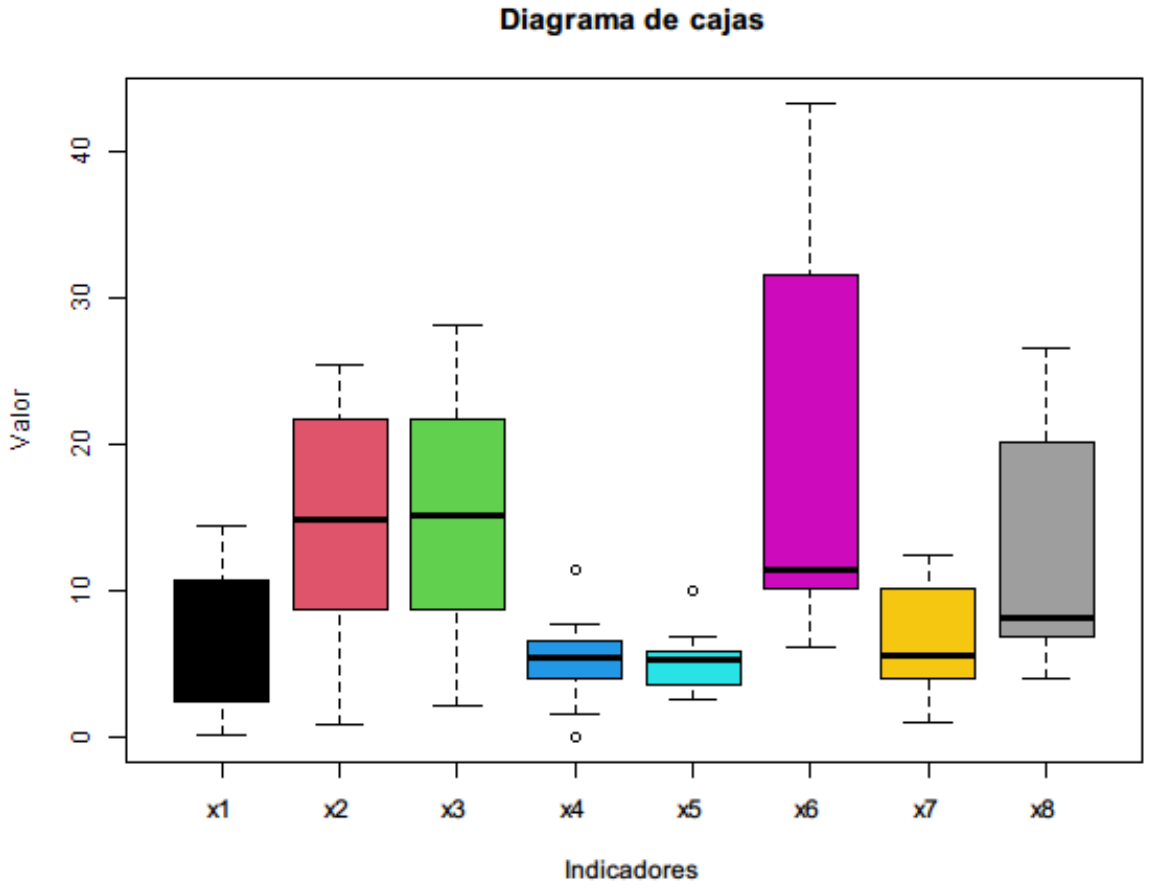
A continuación se muestra una tabla con los estadísticos descriptivos básicos.

```
##        x1               x2               x3              x4
##  Min.   : 0.128   Min.   : 0.9444   Min.   : 2.167   Min.   : 0.0299
##  1st Qu.: 2.476   1st Qu.: 8.6931   1st Qu.: 8.667   1st Qu.: 4.0673
##  Median : 7.584   Median :14.8487   Median :15.167   Median : 5.4044
##  Mean   : 6.957   Mean   :14.9138   Mean   :15.167   Mean   : 5.3976
##  3rd Qu.:10.734   3rd Qu.:21.7262   3rd Qu.:21.667   3rd Qu.: 6.5147
##  Max.   :14.364   Max.   :25.4261   Max.   :28.167   Max.   :11.3959
##        x5               x6              x7              x8
##  Min.   : 2.557   Min.   : 6.135   Min.   : 1.064   Min.   : 3.949
##  1st Qu.: 3.525   1st Qu.:10.170   1st Qu.: 3.982   1st Qu.: 6.833
##  Median : 5.336   Median :11.374   Median : 5.584   Median : 8.103
##  Mean   : 5.107   Mean   :21.006   Mean   : 6.598   Mean   :13.074
##  3rd Qu.: 5.874   3rd Qu.:31.586   3rd Qu.:10.160   3rd Qu.:20.154
##  Max.   :10.037   Max.   :43.278   Max.   :12.374   Max.   :26.571
```

```
## [1] "Desviaciones estándar:"
```

```
##        x1        x2        x3        x4        x5        x6        x7        x8
##  4.545099  8.162600  8.437954  2.823818  1.991911 13.779784  4.030662  8.249730
```

En la siguiente gráfica se muestran los diagramas de cajas de las variables.



3.2 Métodos estadísticos

En este apartado se indican las distintas técnicas estadísticas que se han utilizado.

Primero, se ha realizado un análisis numérico y gráfico de cada variable. De esta forma, mediante en visionado de la estructura del archivo de datos, se han estudiado las posibles recodificaciones y valores perdidos. También se ha realizado un análisis descriptivo numérico clásico, esto es, usando las funciones `summary`, `boxplot` y `skewness` hemos obtenido las medidas de tendencia central, dispersión, cuartiles, simetría, etc. Por otra parte, con la función `check_outliers` hemos detectado los posibles outliers mediante el método de mahalanobis. Para comprobar el supuesto de normalidad, hemos utilizado `colMeans` y para normalizar los datos hemos usado `scale`. De esta forma, podemos visualizar la normalidad con `qqplot`.

Con respecto al análisis exploratorio multivariante, se ha utilizado el test de Bartlett, `cortest.bartlett`, para estudiar la correlación entre las variables. En cuanto al Análisis de Componentes Principales, éste se ha realizado con `prcomp` y se han utilizado técnicas gráficas, tales como `ggplot` y `fviz_pca`. Sobre el AF, se han utilizado otras técnicas gráficas, como `ggcorrplot`,

`scree`, `parallel` y `diagram`, y `factanal` para realizar el test de hipótesis que constrasta si el número de factores es suficiente. Para realizar el análisis de la normalidad multivariante, se ha utilizado el paquete `MVN`. Específicamente, hemos usado dos tests diferentes: el de Henze-Zirkler y el de Royston.

Finalmente, para realizar el agrupamiento de los objetos, se ha utilizado la técnica `kmeans`, variando el número de clusters con el fin de comprobar cómo se agrupan las empresas.

4 Resultados

En este apartado se mostrarán los resultados obtenidos aplicando las técnicas mencionadas anteriormente.

4.1 Análisis exploratorio univariante

Para estudiar nuestro conjunto de datos, podemos obtener el coeficiente de simetría de la distribución estadística.

```
skewness(datos)
```

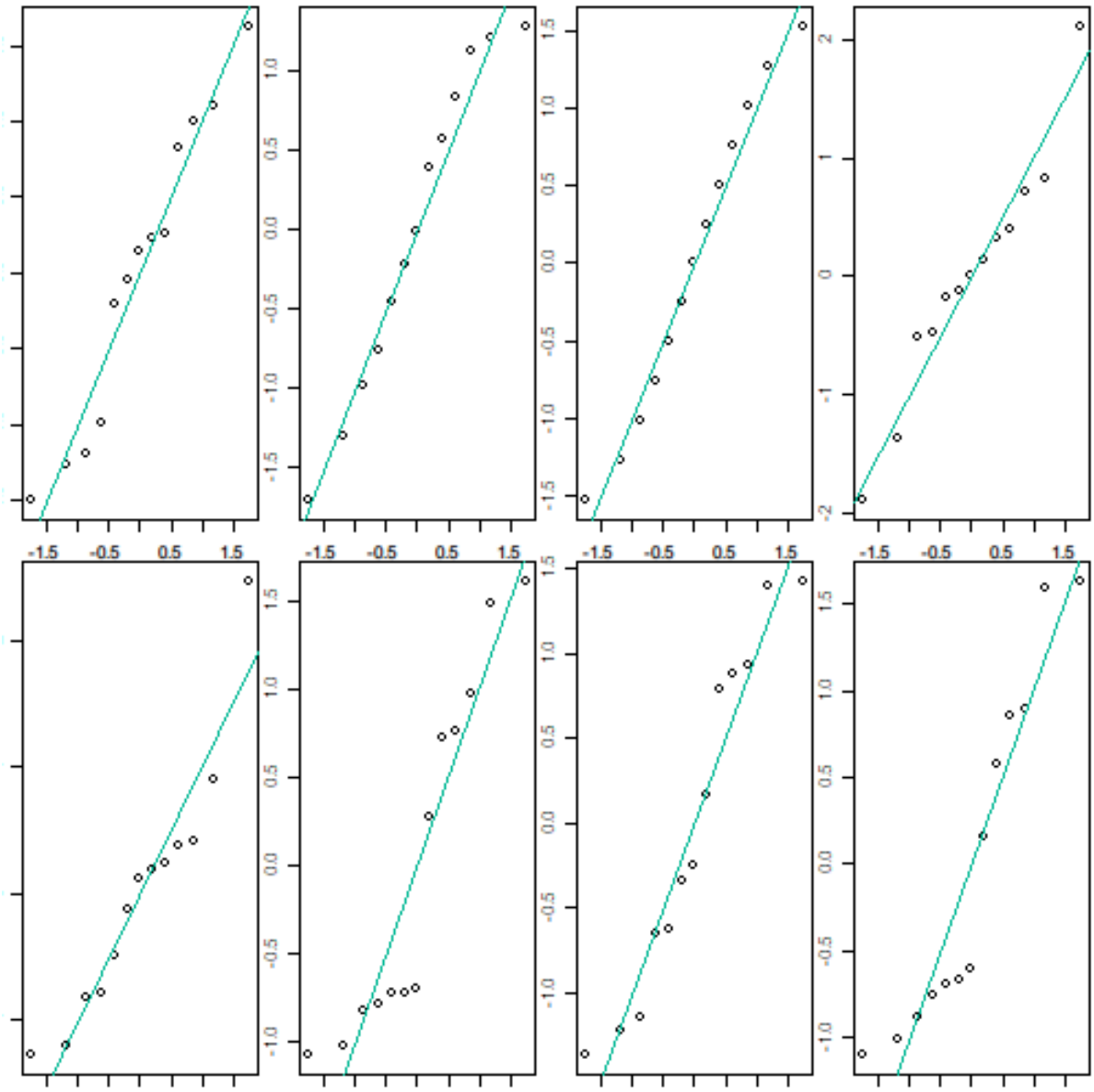
```
##            x1            x2            x3            x4            x5            x6
## -6.571128e-02 -2.178143e-01  2.441493e-16  8.611432e-02  9.539367e-01  4.214678e-01
##            x7            x8
##  1.014152e-01  4.874442e-01
```

Antes de proceder con el PCA y el AF, es necesario tratar los outliers. Este fue un punto de discusión importante, pues como vimos en la figura 1, se muestran tres outliers en las variables x4, x5. Sin embargo, utilizando el método de Mahalanobis, encontramos que no se detecta ninguno:

```
check_outliers(datos, method = "mahalanobis")
```

```
## OK: No outliers detected.
```

Tras normalizar el conjunto de datos, estudiamos cómo se distribuían las variables, produciendo el siguiente resultado:

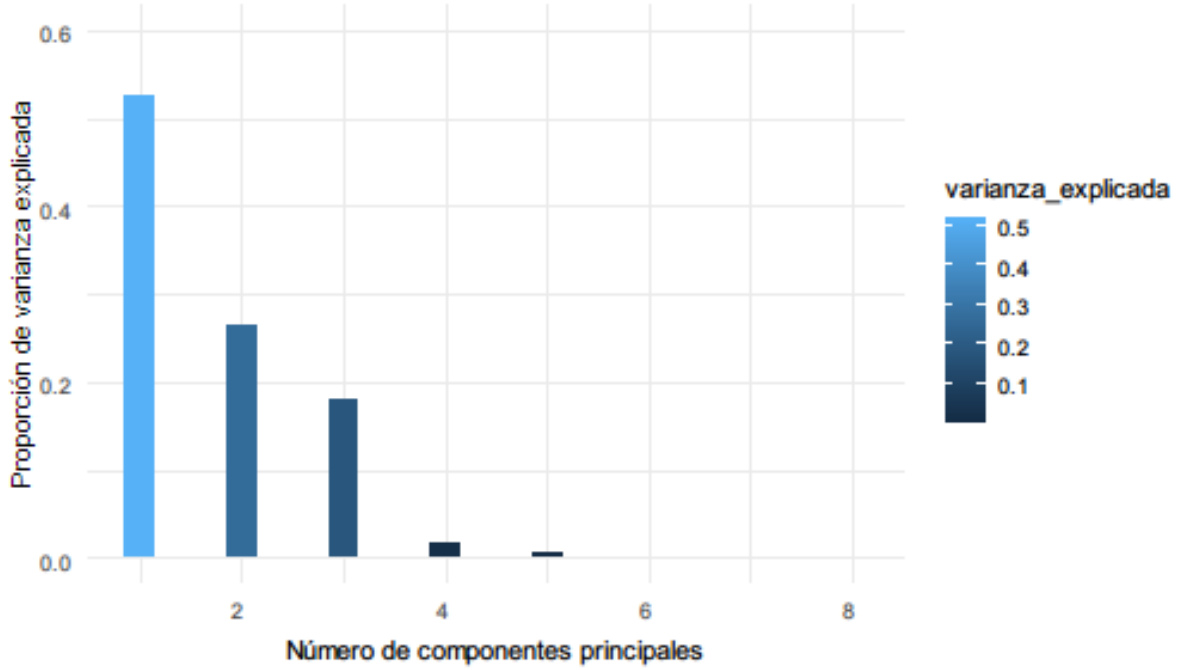


4.2 Análisis explotatorio multivariante

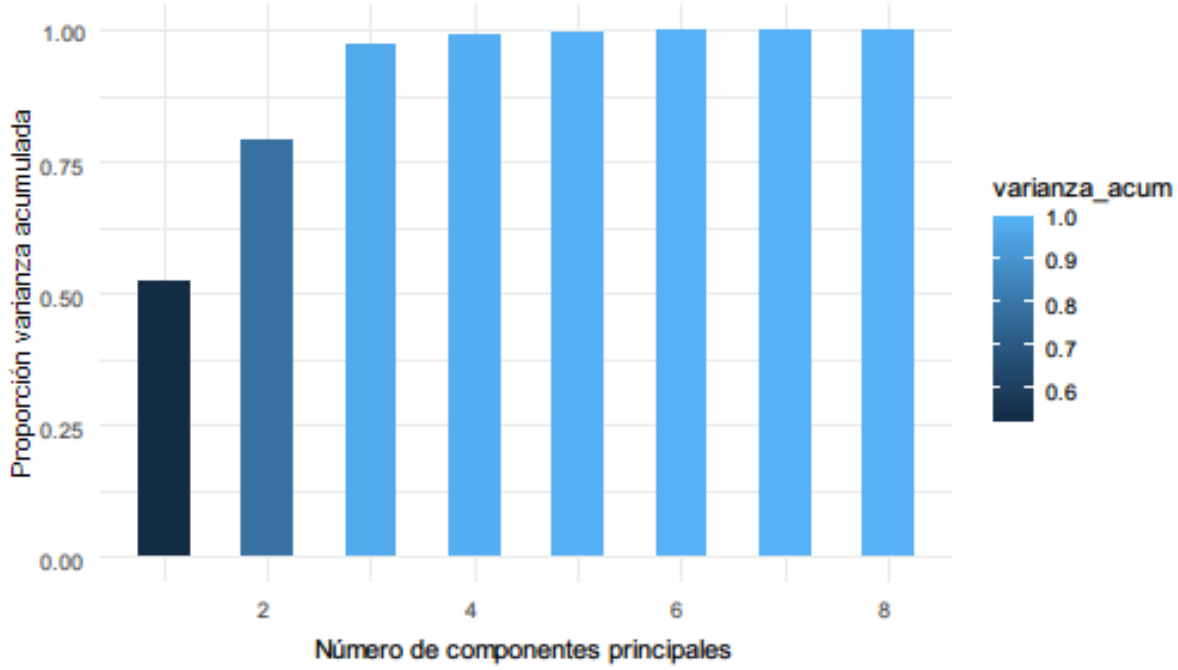
4.2.1 Análisis de componentes principales

Se ha comprobado que existe correlación entre las variables usando el test de Bartlett, pues obteníamos un `p-valor` prácticamente nulo. Esto indica que las variables están correladas, luego procederemos a realizar un Análisis de Componentes Principales (ACP).

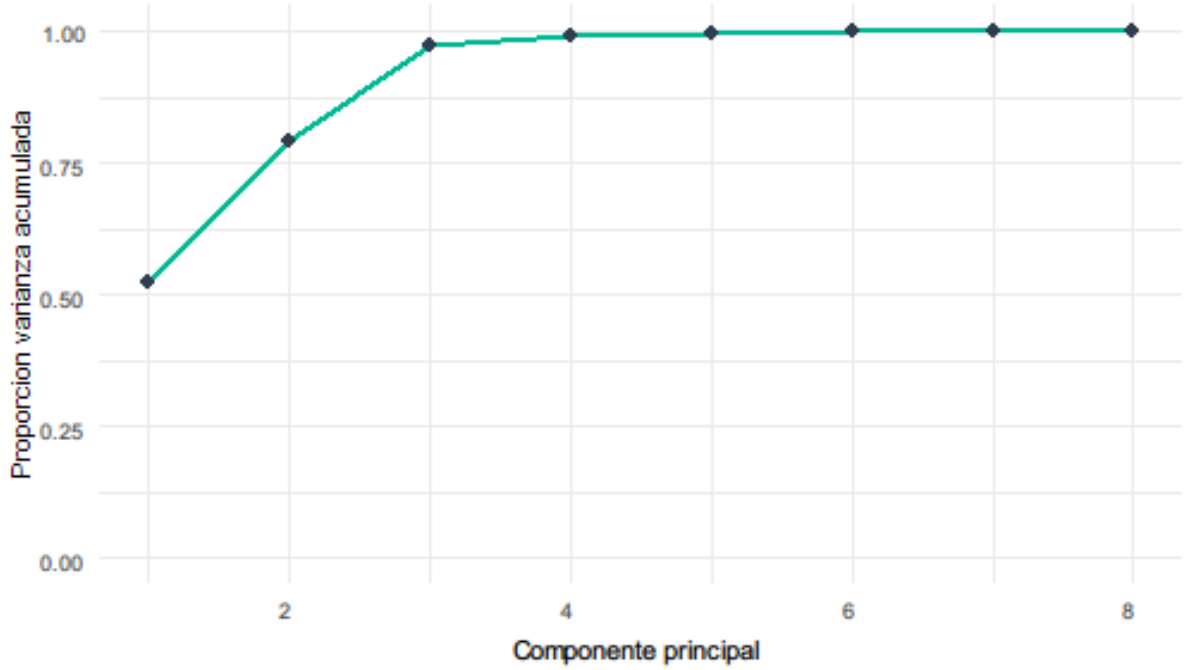
Se han obtenido las desviaciones típicas de cada componente principal y la proporción de varianza explicada y acumulada. Podemos observar en la siguiente imagen un análisis gráfico de la varianza explicada.



De la misma forma, obtenemos un análisis gráfico de la varianza acumulada.

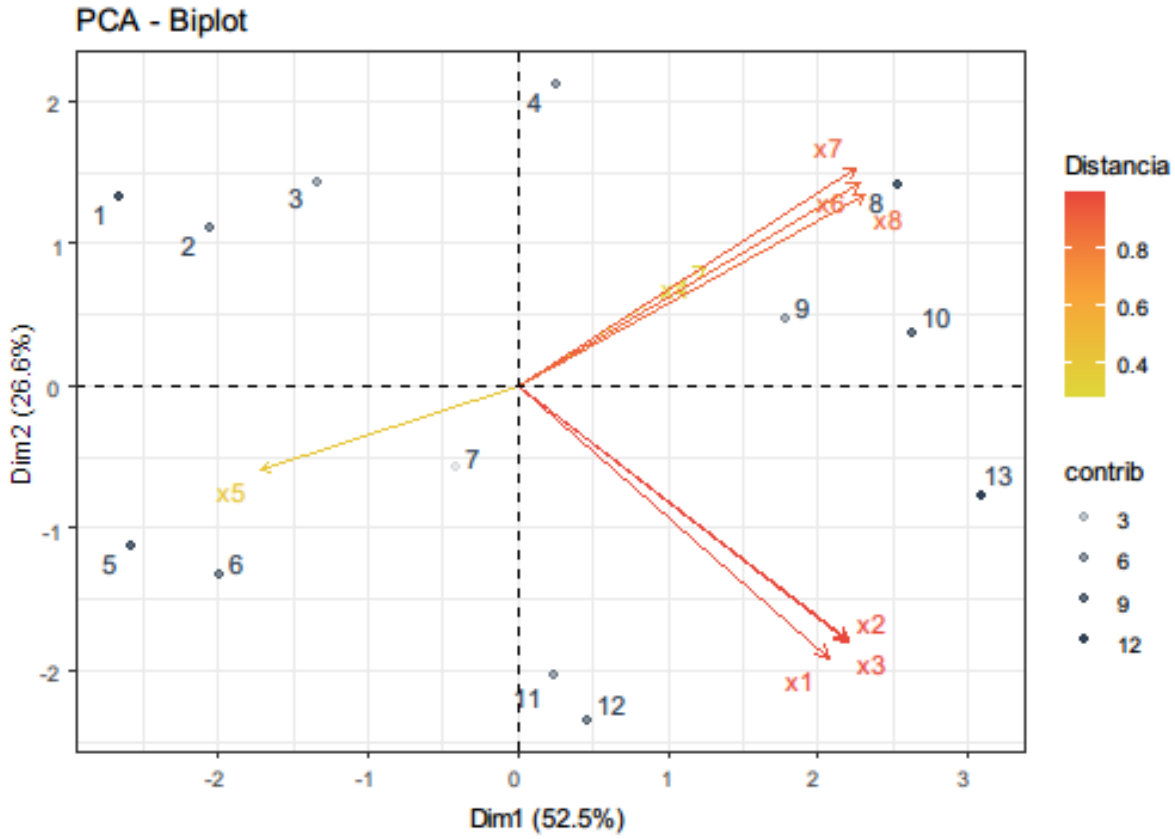


A continuación, se seleccionaron el número de componentes principales óptimo para reducir la dimensión mediante variables observables. Mediante el método del codo, se ha podido analizar gráficamente y elegir las componentes principales.

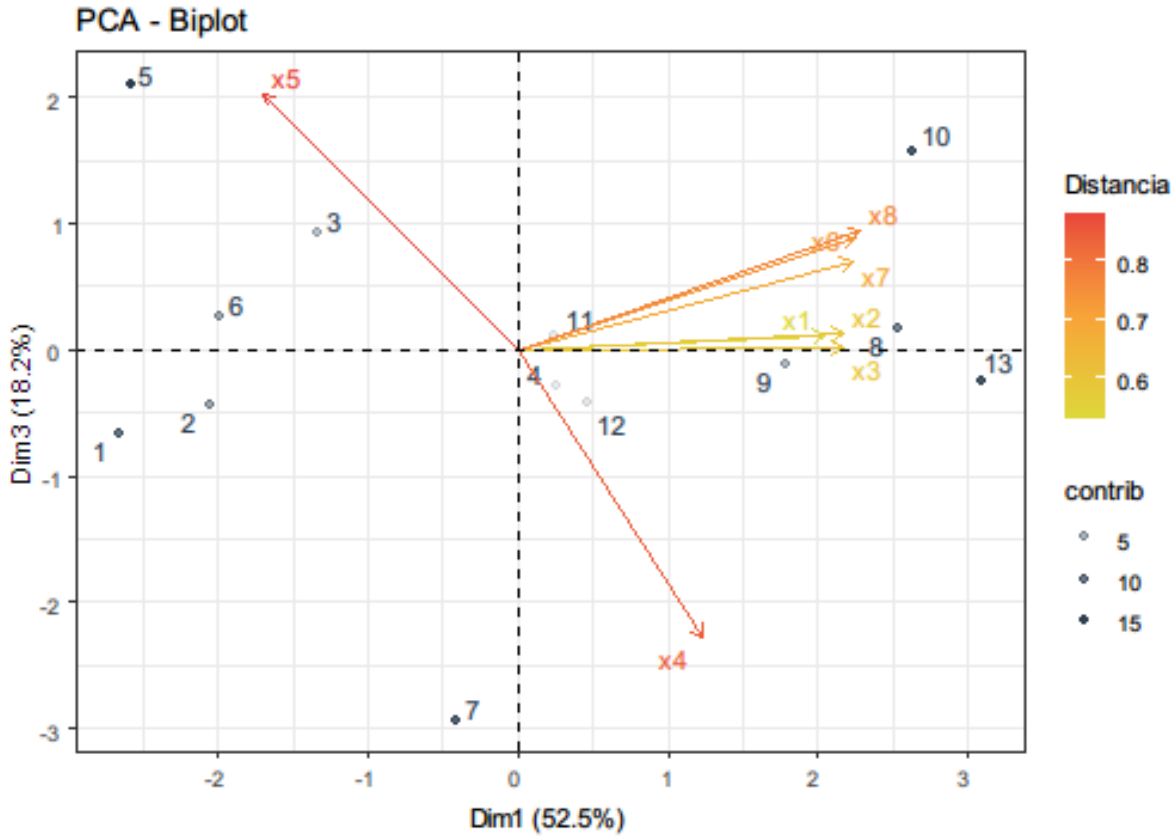


En las siguientes gráficas podremos observar la representación conjunta de variables y observaciones que relaciona visualmente las posibles relaciones entre las observaciones, las contribuciones de los individuos a las varianzas de las componentes y el peso de las variables en cada componentes principal.

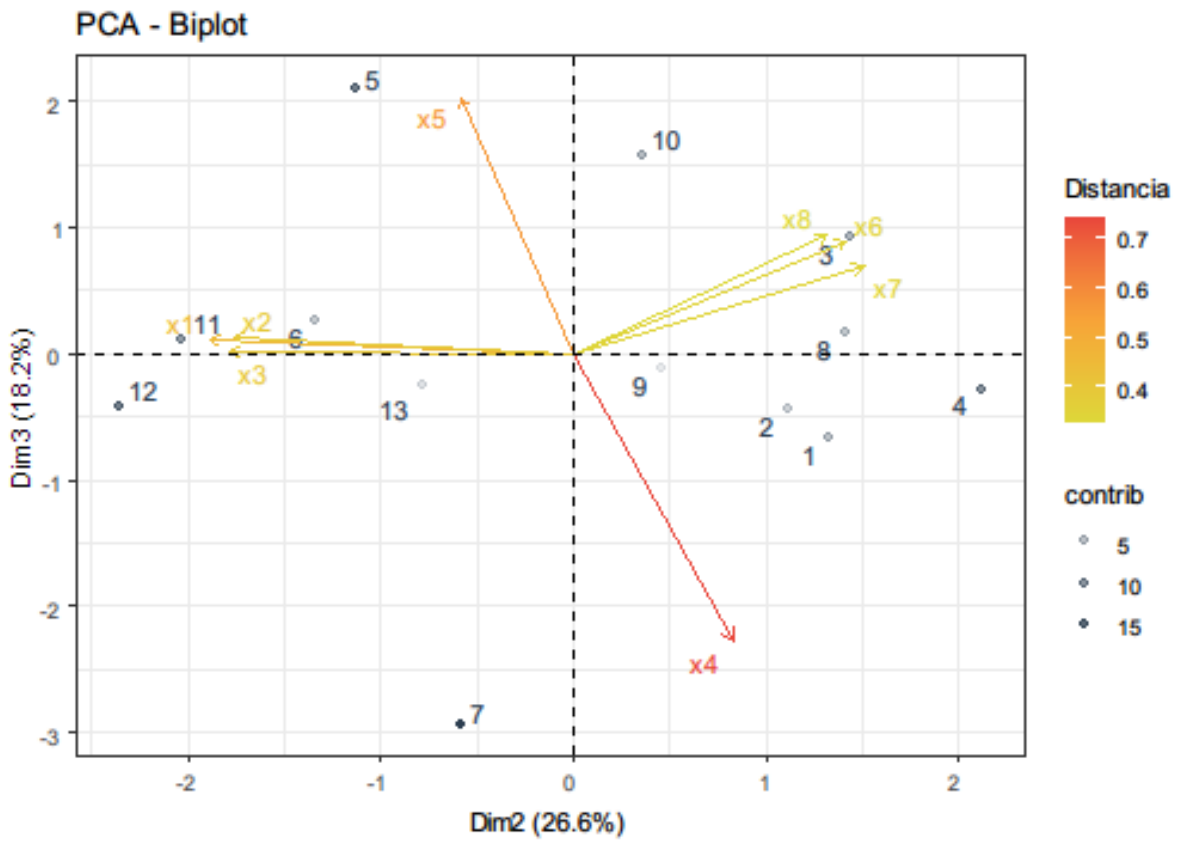
Variables y observaciones en la primera y segunda componente principal:



Variables y observaciones en la primera y tercera componente principal:

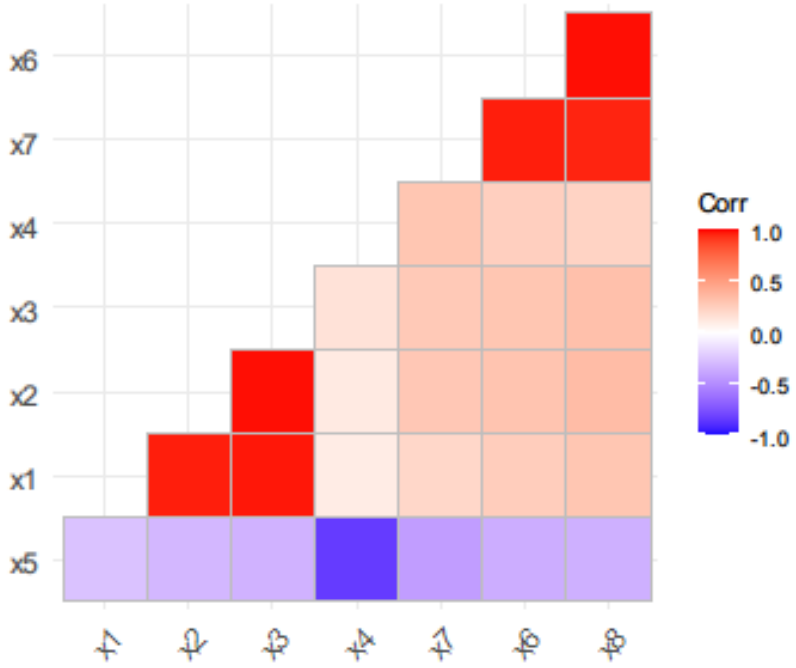


Variables y observaciones en la segunda y tercera componente principal:

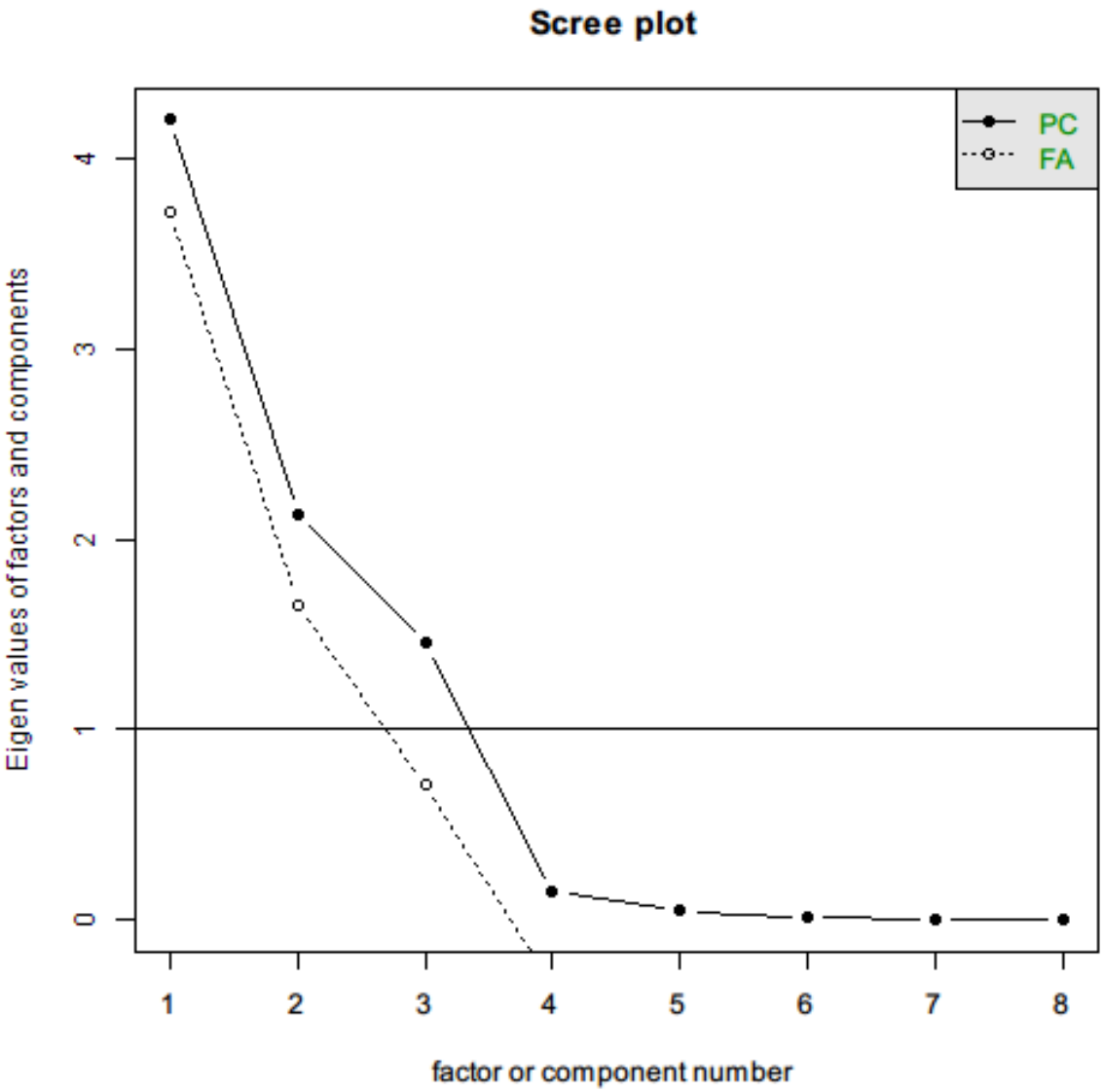


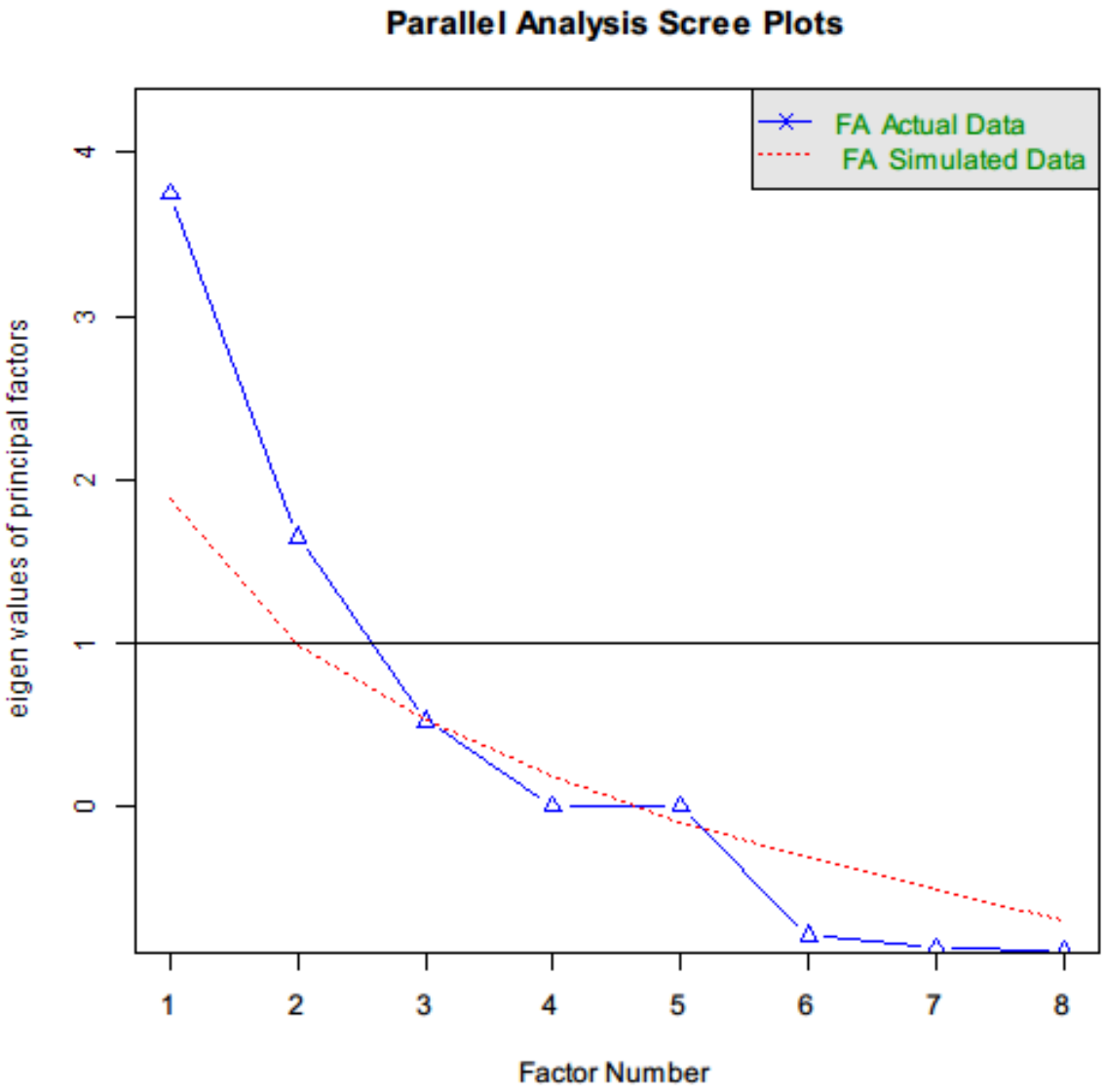
4.2.2 Análisis factorial

Tras realizar el ACP, procedemos a realizar el Análisis Factorial (AF), el cual tiene sentido porque las variables están correladas:



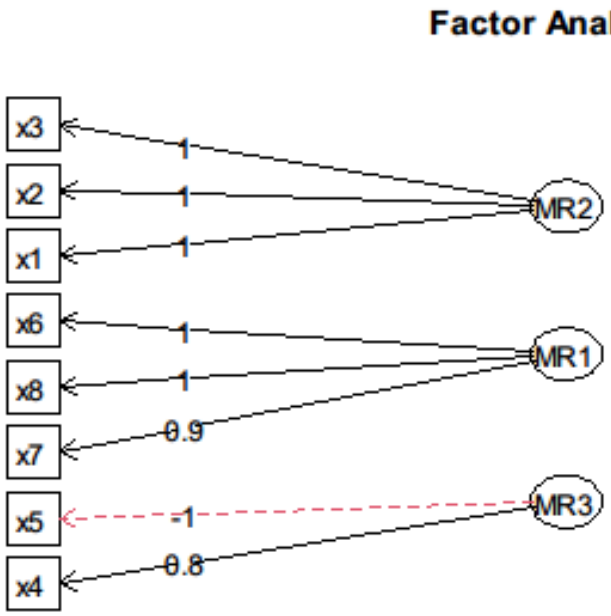
Una vez comparadas las salidas con el método del factor principal y con el de máxima verosimilitud, comparamos las comunalidades y las unicidades. De esta forma, ya podemos determinar el número óptimo de factores basándonos en los siguientes gráficos.


Scree plot
PC
FA
Eigen values of factors and components
0
1
2
3
4
1
2
3
4
5
6
7
8
factor or component number




```
## Parallel analysis suggests that the number of factors =  2  and the number of components =  NA
```

Estimamos el modelo factorial con 3 factores implementando una rotación tipo varimax para buscar una interpretación más simple.

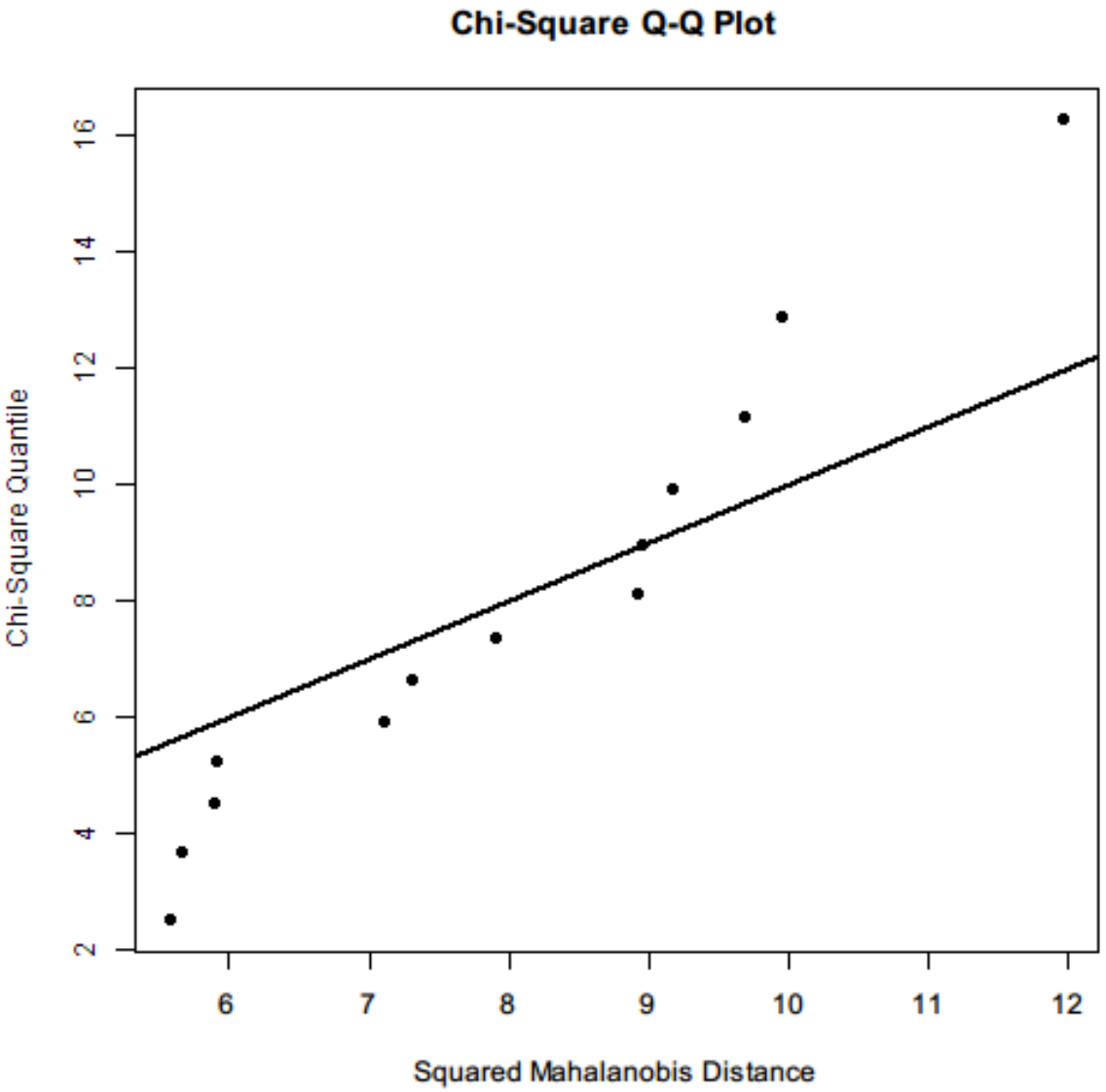


Finalmente, con el test de hipótesis contrastamos si el número de factores es suficiente, lo cual fue cierto.

4.2.3 Análisis de la normalidad multivariante

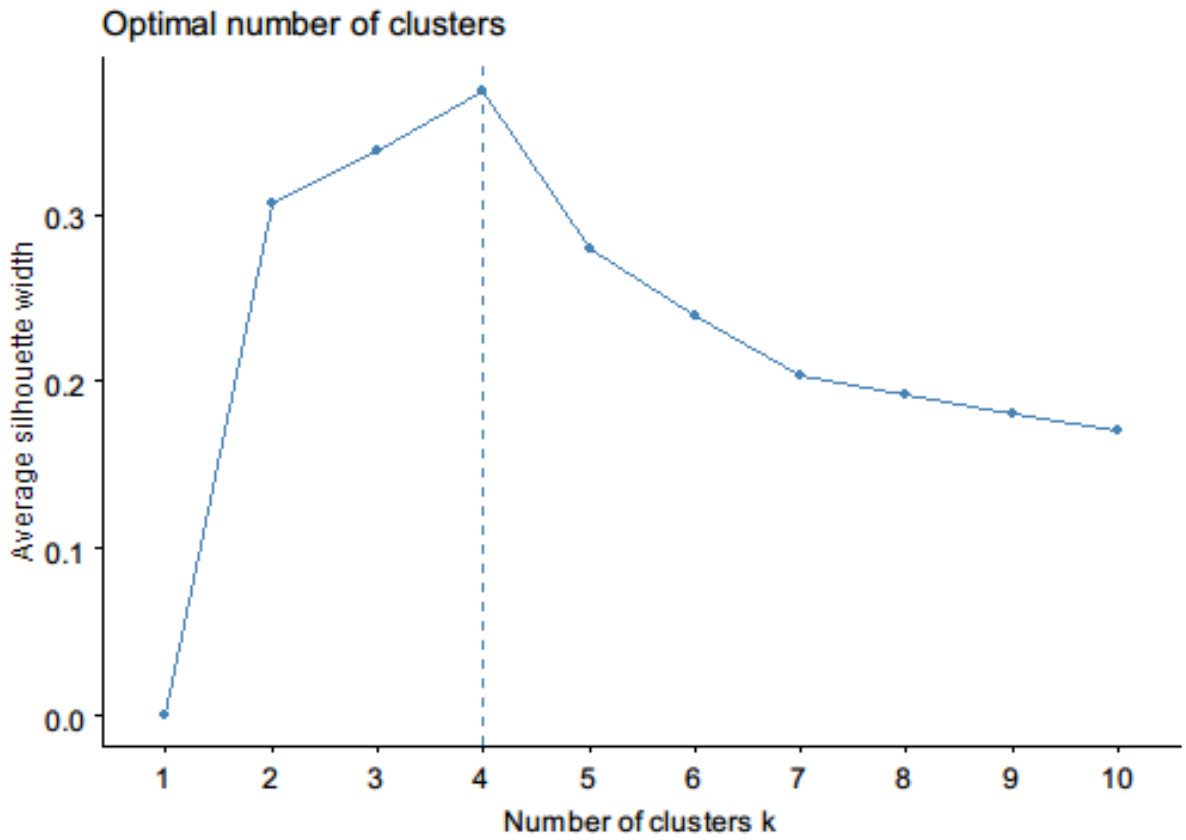
Previo al análisis cluster, se ha analizado la normalidad multivariante de los datos con los tests de Royston y de Henze-Zirkler. Sin embargo ninguno de los dos tests encuentran evidencias al 5% de significación de falta de normalidad multivariante.

Tras realizar el test de Royston, obtenemos la gráfica que se muestra a continuación.



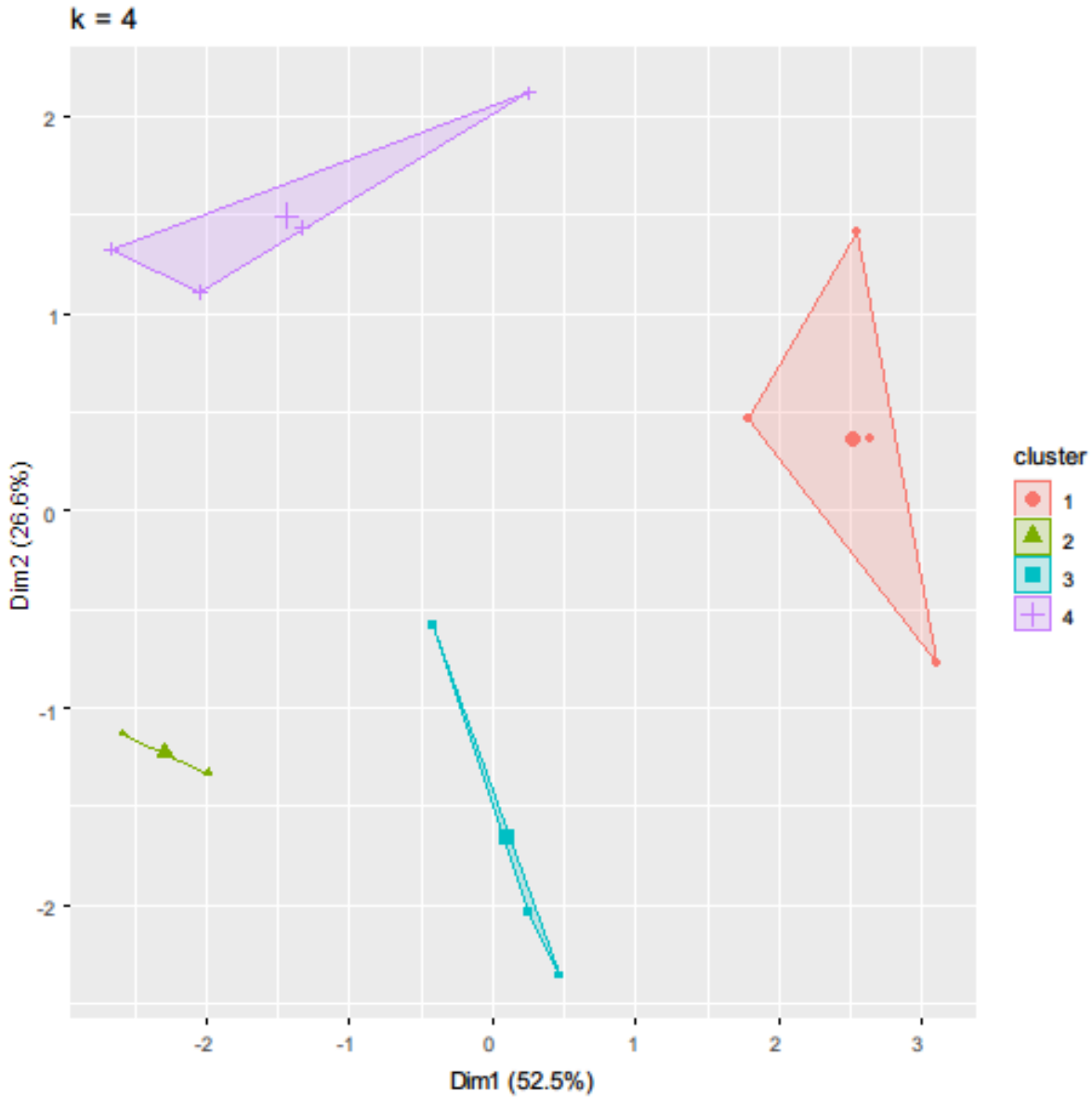
4.3 Clasificación

Para finalizar, se han aplicado técnicas de clustering. Hemos llegado a la conclusión de que el número óptimo de clusters para este dataset es 4:



K-Means produce las siguientes particiones:

p3



Este modelo nos arroja las siguientes siluetas:

```
## Error in plot(silueta): object 'silueta' not found
```

5 Discusión y conclusiones

El objetivo principal consistía en realizar un agrupamiento de los objetos formando clusters de objetos con un alto grado de homogeneidad interna y heterogeneidad entre clusters.

En primer lugar, se ha obtenido el coeficiente de simetría de cada variable en la sección 4.1. En este problema, hemos obtenido valores positivos o negativos. Un valor negativo significa que la distribución está sesgada a la derecha, mientras que un valor positivo indica que la distribución se encuentra sesgada hacia la izquierda.

Uno de los puntos más controvertidos de este trabajo es la decisión de eliminar outliers. Como podemos observar en el diagrama de cajas de la sección 3.1, se muestran tres outliers en las variables `x4` y `x5`. Sin embargo, se ha tomado la decisión de no eliminarlos, pues si no fuera así, se ha comprobado que las métricas de rendimiento bajaban.

Por otra parte, hemos observado que algunas de nuestras variables no se distribuyen con respecto a una normal. Esto queda evidenciado por la gráfica de la sección 4.1. En concreto, las variables x6 y x8 dan problemas. Realizando un test de shapiro para estas variables, se obtiene lo siguiente:

```
##
##  Shapiro-Wilk normality test
##
## data:  datos_normalizados[, "x6"]
## W = 0.84811, p-value = 0.02694
```

```
##
##  Shapiro-Wilk normality test
##
## data:  datos_normalizados[, "x8"]
## W = 0.86598, p-value = 0.04623
```

Ambos obtienen un p-valor por debajo de 0.05. Es probable que supongan una fuente de imprecisiones a la hora de pasar a la distribución normal multivariante.

En cuanto al Análisis Factorial de la sección 4.2.2, se ha tomado la decisión de estimar 3 factores apoyándose en las gráficas obtenidas. Sin embargo, el análisis paralelo sugería que se estimasen 2 factores, lo cual produciría un caso de ultra-Heywood.

Finalmente, hay otro matiz muy importante que debemos discutir. **Nuestro dataset solo tiene 13 instancias**. Hacer un análisis de calidad con tan pocos datos no es posible. En general, clustering necesita decenas de instancias como mínimo para producir resultados realistas.

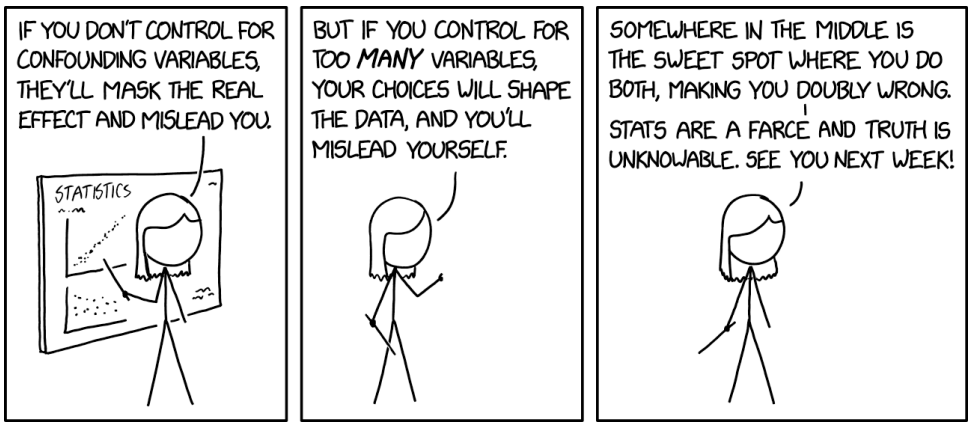


Figure 5.1: *Dos es poco, tres es mucho. Fuente: https://xkcd.com/2560/*

Para ejemplificar esto, la gráfica de las siluetas. Se genera una silueta de 0.37, la cual es considerablemente baja. Mirando los clusters, podemos ver cómo en el cluster 2 existen 2 elementos, mientras que en el 3 hay 3. Con esto no podemos llegar a ninguna conclusión.

Con el fin de encontrar agrupaciones pertinentes y producir un análisis de calidad, sería necesario ampliar el número de empresas. De esta forma, se podrían observar patrones más relevantes.

6 Distribución de las tareas entre las personas implicadas

Para realizar este trabajo, no se han distribuido las tareas, sino que ambos hemos hecho todas las tareas a la vez, supervisando el uno al otro y poniendo en común nuestro conocimiento.