Práctica ACP

Miguel López Garralón

25/10/2019

La base de datos que se trata a continuación contiene 978 casos y 11 variables, en la cual vienen reflejados los tipos de interés de bonos americanos en función del tiempo de depósito. Las principales variables son el tiempo de depósito de los bonos, desde 1 mes hasta los 10 años, y una primera variable que marca el día en el que se encuentra cada tipo de interés para cada tiempo de depósito por bono.

El objetivo de la investigación que se ha llevado a cabo es buscar una estructura subyacente a los distintos bonos en función del tiempo, con la intención de poder reducir el número de variables en función de si comparten alguna estructura que no se puede alcanzar a ver a primera vista. Por último, en caso de poder realizarse, se busca predecir el tipo de interés del bono a 10 años en función del resto de variables, ya sea teniendo todas en cuenta o únicamente las dimensiones subyacentes que se hayan podido encontrar.

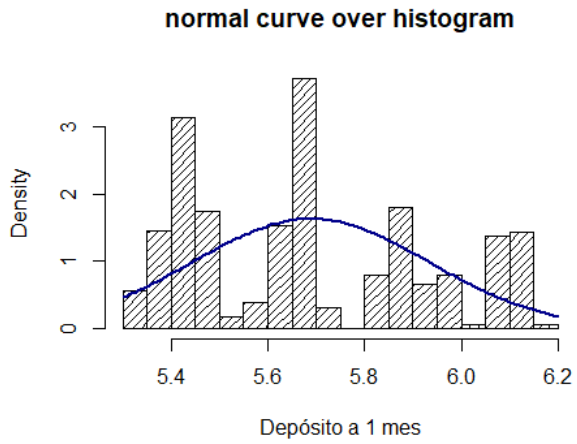
MISSING DATA

Se eliminan todos los valores que tienen algún dato NA, ya que para obtener la matriz de correlaciones es necesario que los registros posean datos y, además, que sean numéricos. Tras la eliminación de dicho valores, la muestra queda reducida a 783 casos, por lo que no se hará división en parte de train y test para predecir el valor de los depósitos a 10 años por reducirse de forma muy significatvia el tamaño muestral.

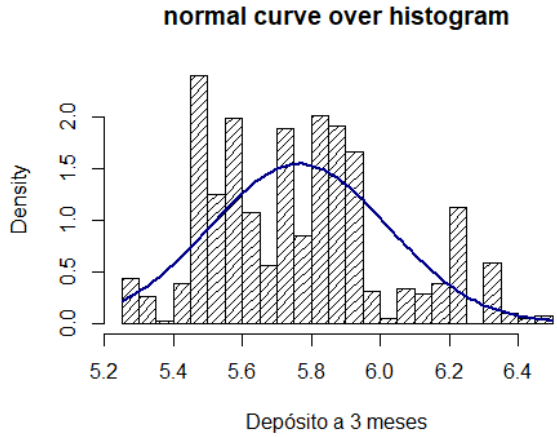
Análisis exploratorio

HISTOGRAMAS CON CURVA DE DENSIDAD

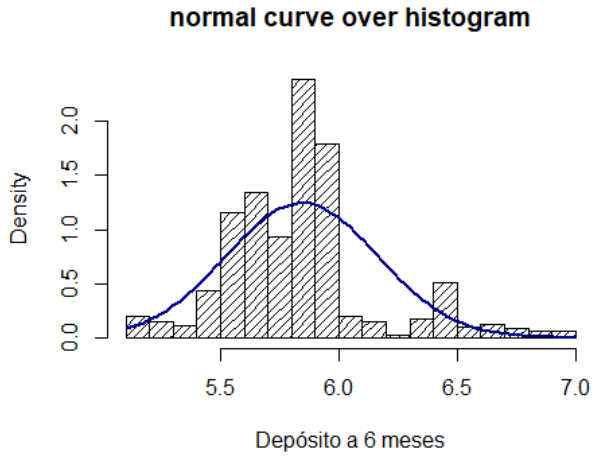
Depósito a 1 mes



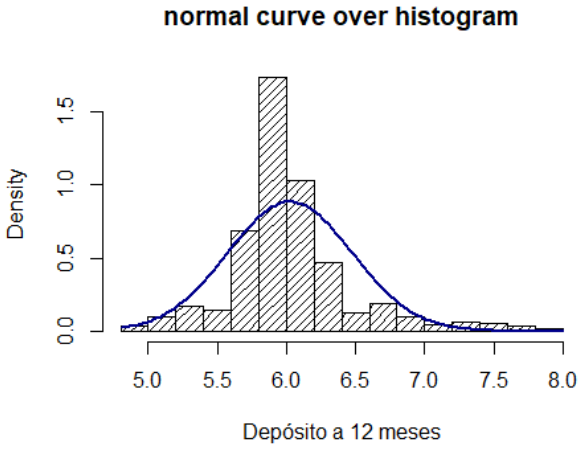
Depósito a 3 meses



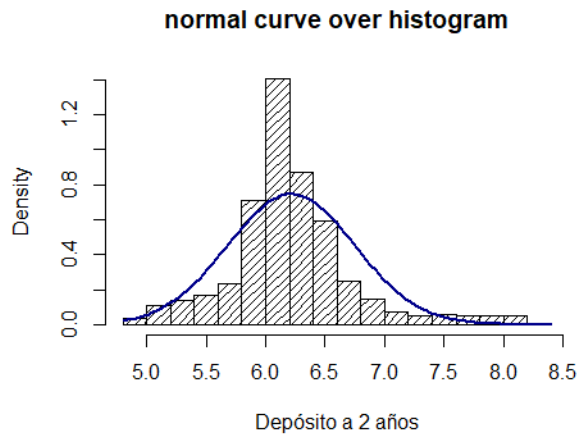
Depósito a 6 meses



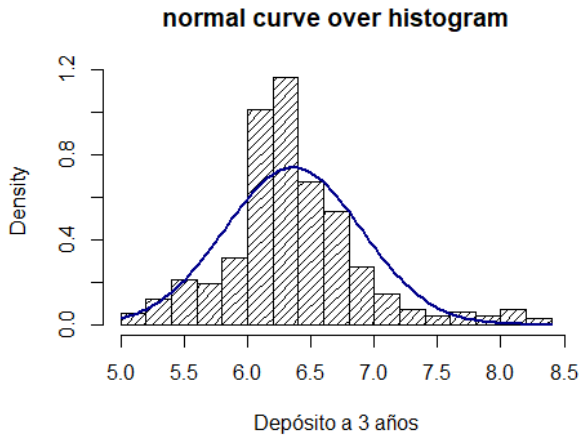
Depósito a 12 meses



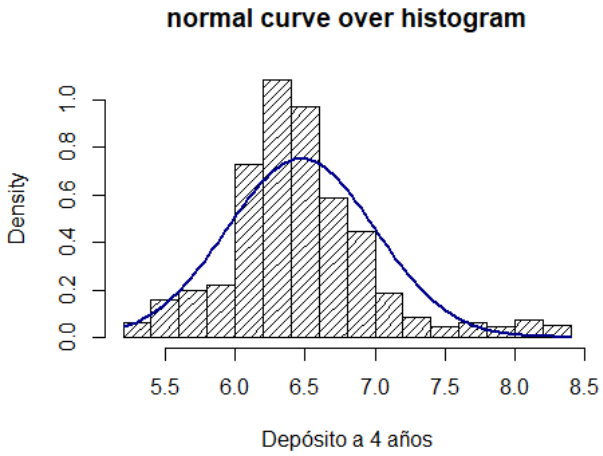
Depósito a 2 años



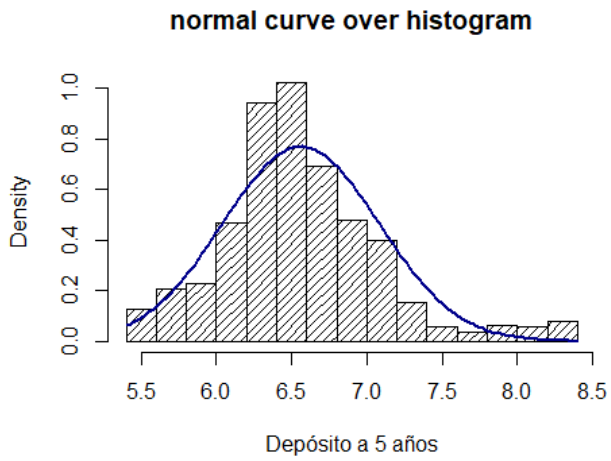
Depósito a 3 años



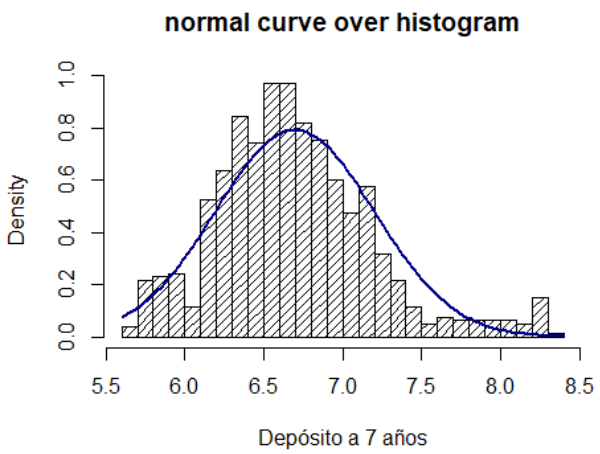
Depósito a 4 años



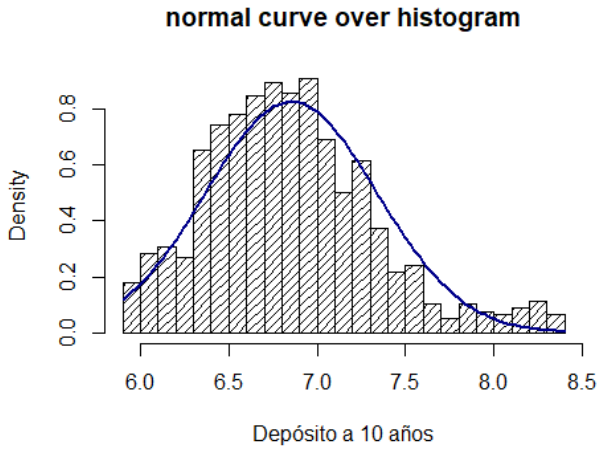
Depósito a 5 años



Depósito a 7 años



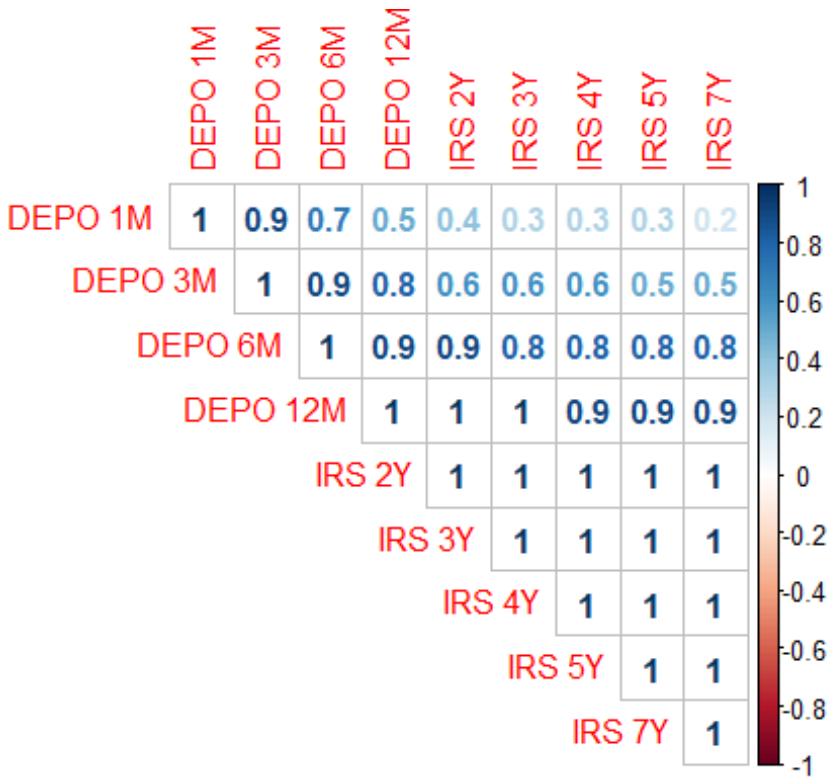
Depósito a 10 años



Con los histogramas se puede ver que las variables no siguen una distribución normal. Sin embargo, debemos pasar al análisis de la correlación entre las variables para ver la relación entre las mismas, de forma que se pudiesen llevar a cabo unas técnicas u otras.

CORRELACIÓN

En el siguiente gráfico se aprecia que hay una alta correlación entre las variables. Se observa, además, que las variables de bonos de 1 mes y 3 meses tienen unas correlaciones más bajas con el resto de bonos, sobre todo a partir de los que son a 12 meses o más (principalmente en el caso de los bonos a 1 mes). Estos resultados en el análisis de las correlaciones nos plantean la posible relación entre las variables estudiadas.



Determinante de la matriz de correlaciones

El determinante de la matriz de correlaciones muestra un resultado de 1.483166e-18, por lo que indica que existe alta multicolinealidad entre las variables. Este dato, junto a la alta correlación de las variables sugieren que un Análisis de Componentes Principales puede ser realizado. Para poder corroborar esto se realizan también las pruebas de esfericidad de Bartlett y KMO.

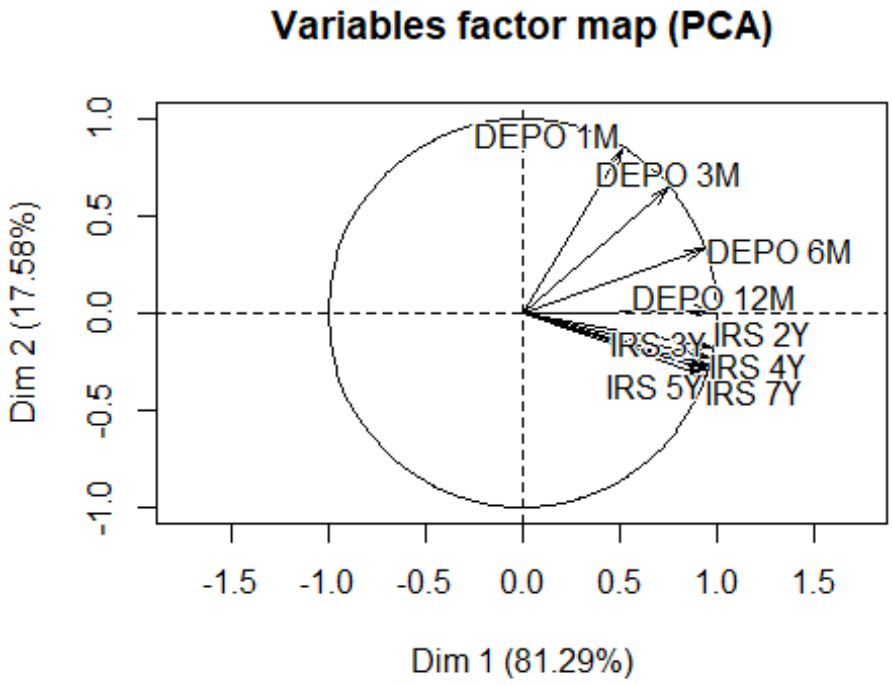
Prueba de esfericidad de Bartlett y KMO

En la prueba de efericidad de Bartlett se obtiene un p-value de 2.2e-16, por lo que se debe rechazar la hipótesis nula de varianzas homogéneas. Además, refuerza lo visto anteriormente sobre la alta correlación positiva entre las distintas variables.

En cuanto al KMO, se obtiene un valor de 0.87. Al ser este valor cercano a 1 y superior a 0.8 podemos afirmar, junto con las pruebas realizadas anteriormente, que es conveniente llevar a cabo un Análisis de Componentes Principales para las variables que nos encontramos en este data set.

Análisis de Componentes Principales (ACP)

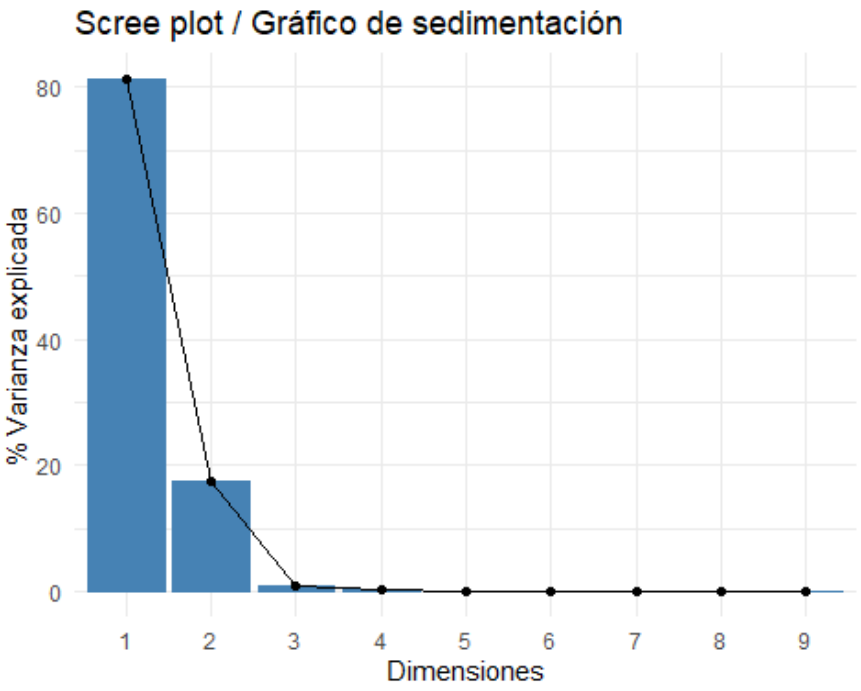
Se llevará a cabo el ACP con las variables del bono a 1 mes hasta el bono a 7 años. Aunque finalmente no se pueda predecir la variable del bono a 10 años como era la intención inicial, debido a que se ha reducido la muestra de forma considerable al eliminar los valores perdidos, se mantiene fuera dicha variable para realizar el ACP a la espera de obetener más datos y poder predecir sobre el bono a 10 años.



Según se ve en el gráfico, en el que se divide en dos dimensiones, la dimensión 1 explica un 81.29 % de la varianza, mientras la segunda dimensión explica un 17.58 % de la varianza. Por tanto, hay dos atributos (provenientes de las 9 variables numéricas) que explican el 98.87 % de la varianza.

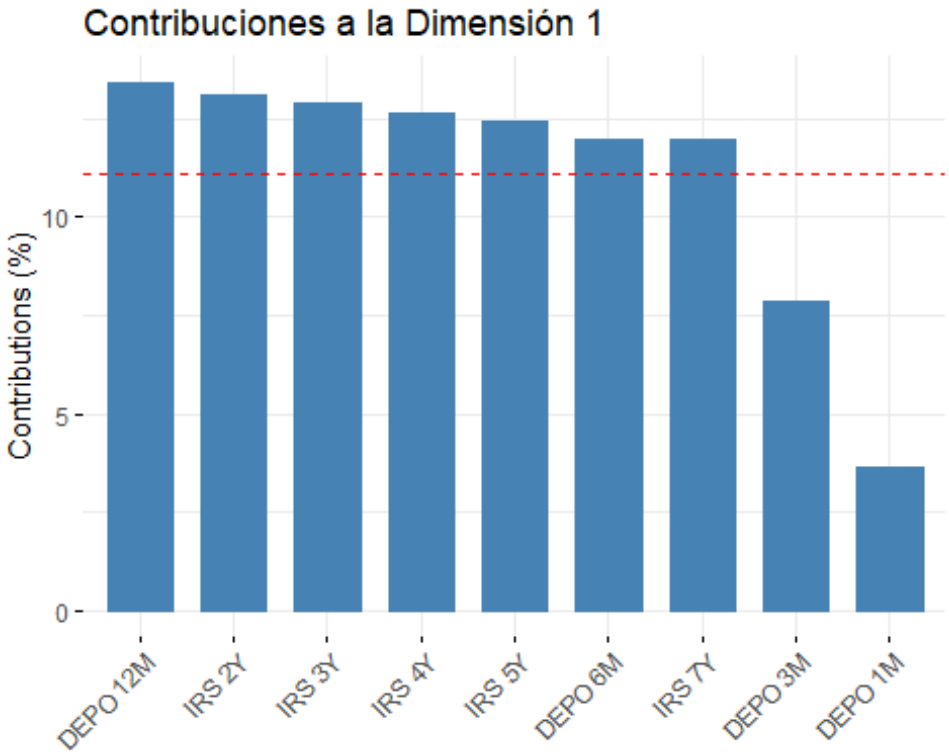
Se puede ver una primera división bastante clara en el eje vertical. En la parte superior se encuentran los depósitos desde 1 mes hasta 1 año, aunque este último aparece prácticamente sobre el eje horizontal, mientras en la parte inferior se encuentran los depósitos desde 2 años a los 7 años. Como primera aproximación, se ve una diferenciación en el corto y largo plazo de los depósitos.

Para complementar estos gráficos, se muestra a continuación el gráfico de sedimentación. En él, se puede ver de nuevo que la primera dimensión explica algo más de un 80 % de la varianza y la segunda dimensión casi el 20 %.



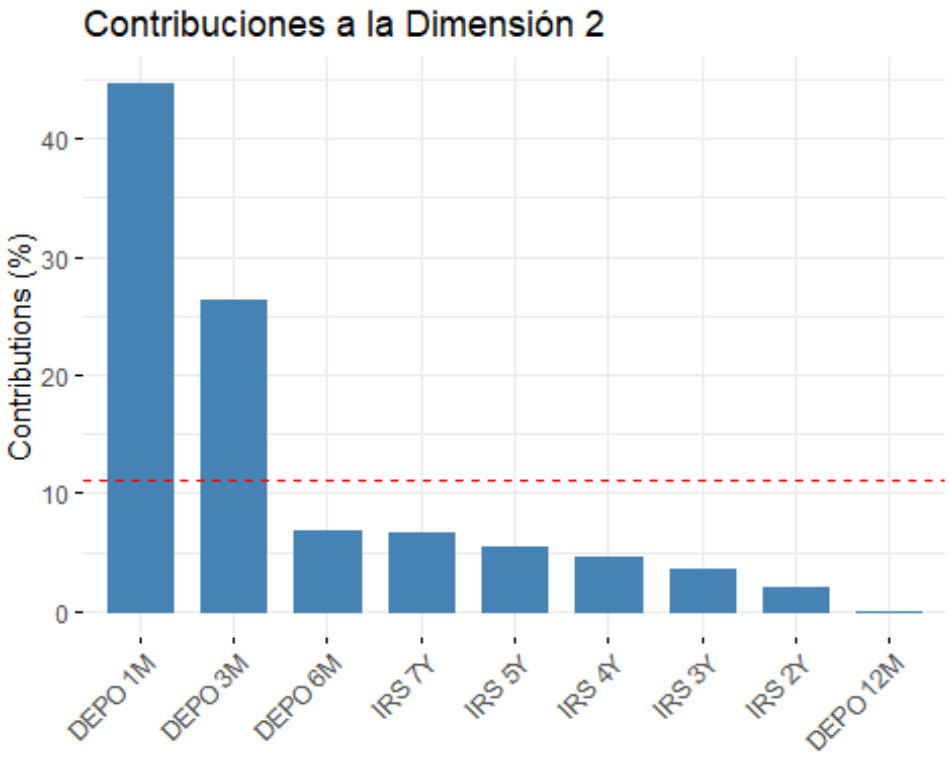
Contribución de las variables a cada dimension

Contribución de las variables a la dimensión 1



En este gráfico observamos cómo contribuye cada variable la dimensión 1. Se puede apreciar como los bonos a medio y largo plazo (del bono a 6 meses hasta el bono a 7 años) contribuyen de forma muy similar a la dimensión 1, y de forma muy superior a los bonos a corto plazo (1 mes y 3 meses).

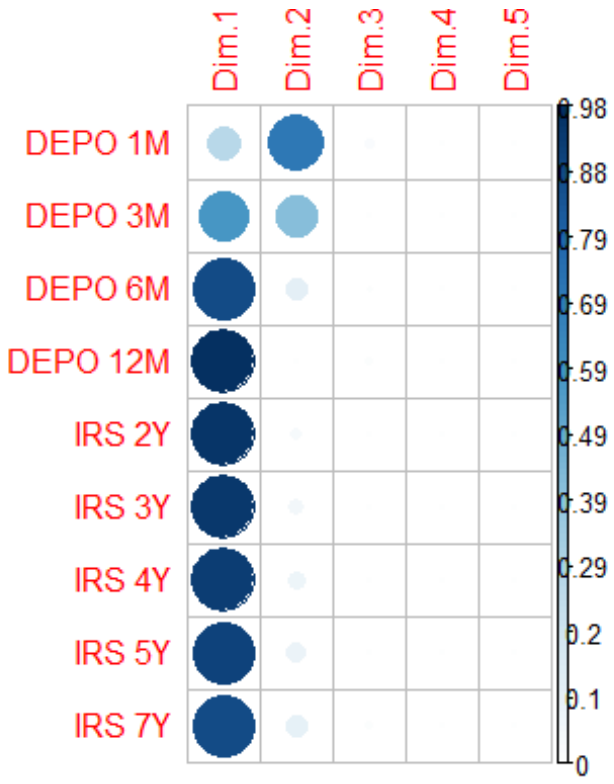
Contribución de las variables a la dimensión 2



Por otro lado, en la contribución de las varaibles a la dimensión 2 se establece de nuevo una clara diferenciación en los bonos al corto plazo (1 y 3 meses), que son los que aportan una mayor contribución a dicha dimensión, frente a los de medio y largo plazo.

Comunalidades

En el siguiente gráfico, se puede apreciar a través de un gráfico de correlaciones lo explicadas que quedan las variables por cada una de las dimensiones. Este gráfico refuerza lo señalado con anterioridad sobre la división entre el corto y el medio-largo plazo en los bonos, aunque se debe señalar que el bono a 3 meses queda también mejor explicado por la dimensión 1, aunque de forma similar en la dimensión 2. Es decir, aunque el bono a 6 meses contribuye de forma muy importante a la dimensión 2, llega a quedar más explicada incluso por la dimensión 1 que por la 2. A pesar de esto, sigue siendo más relevante que contribuya más a la dimensión 2, que lo que queda explicada dentro de cada dimensión dicha variable.



Rotación VARIMAX

La rotación varimax se realiza para facilitar la interpretación de los componentes o dimensiones obtenidas. Esta rotación puede hacer que la varianza total explicada por el conjunto de las dimensiones obtenidas cambie, e incluso que la varianza explicada por cada dimensión también varíe. De hecho, si nos fijamos en la siguiente tabla, se puede ver que la varianza total explicada por las dos dimensiones llega al 98.9 %, aunque en este caso es muy similar al 98.87 % sin rotar. Sin embargo, donde sí se ve un mayor cambio es en la varianza explicada por cada dimensión. La primera dimensión pasa a explicar un 66.8 % de la varianza, mientras anteriormente era más del 80%. En el caso de la dimensión 2, vemos que la varianza explicada ha pasado de casi el 20 % sin rotar hasta el 32.1 % con la rotación varimax.

	Dimensión 1	Dimensión 2
Proporción varianza explicada	0.668	0.321
Acumulación de varianza	0.668	0.989

Sin embargo, al hacer dicha rotación y al cambiarse las varianza explicada por cada dimension, cambia también la contribución de cada variable a las dimensiones. Se gana en interpretación en este caso, ya que en la dimension 2 las variables que más contribuyen son el bono a 1, 3 y 6 meses. Mientras que las contribuciones principals a la dimension 1 son de los bonos de 12 meses a los 7 años. De esta forma se puede ver más claramente la diferenciación entre los bonos a corto plazo y a largo plazo.

	Dimensión 1	Dimensión 2
Bono 1 a mes		0.986
Bono 3 a meses	0.358	0.929
Bono 6 a meses	0.667	0.736
Bono 12 a meses	0.866	0.482
Bono a 2 años	0.947	0.309
Bono a 3 años	0.966	0.257
Bono a 4 años	0.974	0.224
Bono a 5 años	0.977	0.199
Bono a 7 años	0.978	0.160

Predicción del bono a 10 años a través de una regresión de componentes principales

A pesar de que no se puede realizar una predicción sobre los casos solicitados debido a que con la eliminación de los casos perdidos se ha reducido la muestra a 783 registros, se ha establecido una muestra de entrenamiento de 700 registros y una de 83 de test para poder realizar una predicción del bono a 10 años a través de una regresión de componente principals.

Dicho modelo ha obtenido un MSE = 0.05667397.

Referencias

https://rpubs.com/Joaquin_AR/242707

http://www.rpubs.com/marcelo-chavez/multivariado_1

Si se desea acceder al código utilizado para elaborar el informe acceder al siguiente repositorio:
https://github.com/miguellgpm/MasterDataScienceCUNEF/tree/master/DimensionalityReduction