Máster en Ciencias Actuariales y Financieras

TRABAJO FINAL DE MÁSTER

ANÁLISIS DE COMPONENTES INDEPENDIENTES APLICADO A SERIES FINANCIERAS

Autor: Esteban Requena Cadena

Tutor: Salvador Torra Porras

Curso: 2º año del Máster en Ciencias Actuariales y Financieras

UNIVERSITAT DE BARCELONA
Facultat d'Economia i Empresa

Màster de Ciències Actuarials i Financeres

Facultad de Economía y Empresa

Universidad de Barcelona

Trabajo Final de Máster

Máster en Ciencias Actuariales y Financieras

ANÁLISIS DE COMPONENTES INDEPENDIENTES APLICADO A SERIES FINANCIERAS

Autor: Esteban Requena Cadena

Tutor: Salvador Torra Porras

“El contenido de este documento es de exclusiva responsabilidad del autor, quien declara que no ha incurrido en plagio y que la totalidad de referencias a otros autores han sido expresadas en el texto”

RESUMEN

El trabajo realizado tiene como objetivo recoger diversas técnicas clásicas de análisis de datos multivariantes utilizadas con frecuencia en el análisis financiero y compararlas con el análisis de componentes independientes y otras técnicas similares que se están desarrollando en la actualidad. De esta forma, después de realizar un estudio teórico y definir el funcionamiento de las técnicas, se procederá a la aplicación práctica de estos métodos de análisis sobre una base de datos de precios diarios de diez empresas tecnológicas con la finalidad de ver hasta qué punto el análisis de componentes independientes y las nuevas técnicas mejoran los métodos clásicos.

Palabras Clave: Reducción de dimensión, agrupación de variables, factorización, componentes principales/independientes.

ABSTRACT

The objective of this work is to collect several classic multivariate data analysis techniques frequently used in financial analysis and compare them with the analysis of independent components and other similar techniques that are currently being developed. In this way, after explaining the theoretical part and defining all the techniques of multivariate analysis, we will proceed to the practical application of all these methods of analysis on a database of daily prices of ten technological companies to see in what the independent component analysis and the new techniques improve classical methods.

Keywords: Dimension reduction, variable grouping, factoring, principal/independent components.

ÍNDICE

1. INTRODUCCIÓN

El análisis de componentes independientes (ICA, *Independent Component Analysis*) es un método que permite el tratamiento multivariante de series temporales. El objetivo principal de los métodos de análisis de datos multivariante es mostrar cuales son las características que presentan nuestras observaciones y pueden ser útiles para encontrar proyecciones de los datos que permitan estudiar el comportamiento de nuestras observaciones. Dentro de los métodos de análisis multivariante encontramos métodos clásicos como son el análisis de componentes principales (PCA) o el análisis factorial (FA), y otros métodos que no son tan frecuentes pero que se están comenzando a utilizar en muchos campos como el *NonNegative Matrix Factorization* (NMF) o el propio ICA.

El ICA se utiliza con frecuencia en muchos campos como la medicina o la ingeniería, pero no está tan desarrollado en el ámbito financiero, donde predominan métodos como el PCA y el FA. Por este motivo, una vez explicado el funcionamiento del ICA se enfocará el trabajo a aplicar esta metodología sobre una serie financiera de datos reales.

El objetivo final de este trabajo es aplicar el análisis de componentes independientes (ICA) sobre un conjunto de datos multivariantes que se corresponden con diez series temporales de rendimientos financieros, con la finalidad de ver si este tipo de análisis es útil en el mercado financiero y si mejora los métodos clásicos que se utilizan actualmente. Para ello, en la primera parte del trabajo se hará una explicación teórica de cuál es el funcionamiento del ICA relacionándolo con otros métodos de análisis multivariante más frecuentes en finanzas, y en la segunda parte del trabajo se aplicarán estos razonamientos teóricos sobre una serie de datos financiera real a través del programa RStudio[1] y las librerías disponibles en este sobre el ICA y el resto de los métodos que se estudien.

El trabajo quedará estructurado de la siguiente forma:

En primer lugar, se hará una breve introducción a las técnicas de análisis de datos multivariante y seguidamente una explicación del modelo de análisis de componentes independientes (ICA). El objetivo de esta primera parte es presentar la definición, ámbito de aplicación y estimación del modelo ICA, teniendo en cuenta su relación con el resto de los métodos de análisis multivariante clásicos (PCA, FA).

En segundo lugar, se aplicarán las técnicas de análisis multivariante clásicas juntamente con el método de análisis de componentes independientes (ICA) sobre nuestra base de datos de 10 activos tecnológicos, de tal manera que, a partir de la aplicación de estos métodos mediante los paquetes que proporciona la herramienta RStudio podremos analizar en profundidad los resultados obtenidos de los distintos métodos estudiados.

Seguidamente se hará una comparativa de nuestro modelo de análisis ICA con los otros métodos de análisis multivariante clásicos estudiados como son el PCA y el FA. También se estudiará el modelo de factorización de matrices no negativas (*Nonnegative matrix factorization*, NMF) y su aportación al ámbito financiero.

Por último, en las conclusiones se comprobará si la aplicación del método de análisis multivariante ICA sobre nuestra base de datos de 10 activos tecnológicos está justificada

[1] Programa RStudio: descarga de software libre https://www.rstudio.com/products/rstudio/download/

y se compararán los resultados obtenidos con los de las técnicas clásicas de análisis multivariante estudiadas (PCA y FA[2]).

En cuanto a la parte práctica del trabajo para la obtención de resultados se utilizará el programa RStudio y se emplearán las guías que ofrecen los distintos paquetes de análisis de datos multivariantes como el de ICA ("FastICA"), NNMF ("NMF") u otros métodos clásicos estudiados en el trabajo. El soporte teórico para aplicar estos métodos de análisis multivariante serán artículos y publicaciones sobre el análisis de componentes independientes (ICA) y sobre otros métodos clásicos como el PCA y FA o alternativos como el NMF.

La muestra elegida como series temporales de datos financieros son los precios diarios de cierre de 10 empresas tecnológicas extraídos de la página web *yahoo finance*[3]. El motivo de escoger 10 empresas del sector tecnológico es que todas trabajan y compiten en el mismo mercado y guardan algún tipo de relación entre ellas, ya sea porque algunas de las empresas fabrican componentes o software para las otras, o bien porque comercializan el mismo tipo de productos en el mercado y son rivales directos.

Las empresas tecnológicas seleccionadas para realizar nuestro estudio han sido las siguientes: IBM, Qualcomm (QCOM), Sony (SNE), Microsoft (MSFT), Apple (AAPL), Intel (INTC), Nokia (NOK), Nintendo (NTDOY), Canon (CAJ) y Google (GOOGL). En el primer anexo del trabajo se realiza una breve descripción de la procedencia y ámbito de negocio de las empresas seleccionadas. Además, en el segundo anexo también se recoge un gráfico que representa los rendimientos de las 10 empresas expresados en términos relativos, es decir, como la diferencia logarítmica de los precios diarios.

[2] El NMF de igual manera que el resto de los métodos realiza una factorización de matrices, aún así el NMF no se puede comparar directamente con el resto de los métodos ya que no reduce la dimensión de los datos en nuevos componentes o factores, sino que los agrupa en función de características similares.
[3] *Yahoo finance:* https://es.finance.yahoo.com/

2. ANÁLISIS DE DATOS MULTIVARIANTES

2.1 Contexto Teórico

La gran cantidad de datos que tenemos a nuestro alcance en muchos campos como son la medicina, la ingeniería, las finanzas y la economía crece continuamente. La elevada cantidad de información que nos proporcionan los datos es muy diversa y tenemos que aprender a procesarla y analizarla de la mejor manera posible.

Según Cuadras (1981): "El análisis multivariante es la rama de la estadística y del análisis de datos que estudia, interpreta y elabora el material estadístico sobre un conjunto de n>1 variables, que pueden ser cuantitativas, cualitativas o una mezcla".

Por lo que un conjunto de n>1 elementos de estudio dan lugar a una población que puede ser medida en función de varias variables. Por ejemplo, si estudiamos los rasgos faciales de un conjunto de personas utilizamos variables como el color de ojos y del cabello, tono de piel, longitud de la nariz, etc. Si describimos en qué situación política se encuentra un país tendremos en cuenta variables como el régimen implantado en función de si es una dictadura o una democracia, el número de partidos políticos que se presentan y la participación de los ciudadanos en las elecciones. En estos dos ejemplos tendremos una gran cantidad de datos de una población a estudiar y más de una variable de referencia a analizar estadísticamente.

Daniel Peña (2002) resume los objetivos principales del análisis de datos multivariantes de una población como: "el estudio estadístico de diversas variables medidas en elementos de una población. Pretende los siguientes objetivos:

1. Resumir el conjunto de variables en unas pocas variables nuevas variables, construidas como transformaciones de las originales, con la mínima pérdida de información.
2. Encontrar grupos en los datos si existen.
3. Clasificar nuevas observaciones en grupos definidos.
4. Relacionar dos conjuntos de variables."

Si explicamos más detalladamente estos objetivos encontramos que en primer lugar resumir el conjunto de variables de estudio es necesario para simplificar la información de la que disponemos, pero sin que se produzcan pérdidas de esta información que nos hagan elaborar análisis erróneos. Un ejemplo de esta simplificación podría ser el IPC (Índice de Precios de Consumo), que se trata de un índice en que se valoran los precios de un determinado conjunto de bienes y servicios sobre una base de presupuestos familiares. De esta forma se consigue resumir una gran cantidad de variables como son los precios y los presupuestos familiares en un solo indicador que es el IPC. Además, disponer de este tipo de indicadores también nos permite representar gráficamente la información para hacerla más visual y establecer comparaciones en el tiempo. Por lo tanto, el análisis de datos multivariantes nos proporcionada métodos empíricos para determinar cuántas variables son estrictamente necesarias utilizar como indicadoras para representar una realidad más compleja.

El segundo objetivo principal es encontrar si existen grupos en los datos con la finalidad de clasificar las observaciones y dividirlas en función de las características propias de

cada una de ellas. Un ejemplo sería un estudio que hiciese una empresa en función del tipo de clientes que existen en un determinado país (procedencia, rango de edad, aficiones) para comercializar un nuevo producto.

Seguidamente, el tercer objetivo guarda mucha relación con el anterior, ya que si encontramos grupos existentes en nuestros datos podremos clasificar nuestras observaciones en grupos definidos con unas características propias.

Finalmente, el cuarto objetivo hace referencia a la estructura de dependencia de las variables estudiadas. La relación existente entre dos o más variables es la que proporciona la capacidad de agruparlas en otras nuevas variables indicadoras que no son visibles *a priori*. Un ejemplo sería el anteriormente mencionado indicador del IPC, que relaciona dos variables conectadas entre sí y con cierto grado de dependencia que son los precios de un predeterminado conjunto de bienes y servicios y el presupuesto familiar.

Por lo tanto, uno de los objetivos principales del análisis de datos multivariantes es la reducción de la dimensión de los datos con el fin de mostrarnos cuales son las características más relevantes de nuestras observaciones. Algunos de los ejemplos clásicos de este análisis de datos multivariantes son el análisis de componentes principales (PCA) y el análisis factorial (FA). El análisis de componentes independientes (ICA), que estudiaremos en el trabajo, también es otro método de análisis multivariante, aunque menos empleado en finanzas. Dichos métodos de análisis multivariante de datos nos permiten realizar una factorización de matrices que supone la posibilidad de reducir la dimensión de los datos para estudiar su comportamiento.

Los métodos de análisis multivariante tienen un amplio abanico de aplicaciones en casi todos los campos. Primeramente, se emplearon en el campo de la ciencia para solucionar problemas de clasificación en Biología, pero se extendieron hacia otros campos como el de la Psicometría, la Economía y las Ciencias Sociales en general. Actualmente se utilizan con mucha frecuencia en campos como la Ingeniería y las TIC (Tecnologías de la Información y la Comunicación) para resumir la información disponible, elaborar métodos de clasificación y para el reconocimiento de patrones en los datos. Algunos ejemplos concretos de sus aplicaciones en distintas ciencias son los siguientes:

En Economía estos métodos de análisis multivariante permiten cuantificar cual es el desarrollo de un país, ver cuál es la diferencia entre ingresos y gastos de las familias, comprender cuál es el comportamiento que tienen los consumidores en función de sus características y tipologías, entre otros. En la Ingeniería, uno de los ámbitos con más aplicación, se emplea con mucha frecuencia para diseñar programas y máquinas inteligentes que permitan clasificar los datos de forma interactiva aprendiendo del entorno en el que se encuentran. Otro aspecto de la Ingeniería con mucho desarrollo del análisis multivariante es el desarrollo de inteligencia artificial. Finalmente, y en el campo de la Medicina, este tipo de análisis ha ayudado a construir procesos automáticos de ayuda al diagnóstico y dentro de la de la psicología a interpretar resultados de las pruebas de aptitudes de los pacientes.

2.2 Utilidad e introducción al Análisis de Componentes Independientes

El análisis de componentes independientes es un método estadístico empleado en el análisis de datos multivariantes y su origen está relacionado con la búsqueda de una solución al problema de separación ciega de señales (BSS, *Blind Signal Separation*). El problema consiste en obtener señales a partir de las fuentes originales que intervienen en una mezcla, sin tener información previa sobre las ponderaciones de la mezcla. Este problema se da con frecuencia en el campo de la ingeniería con el procesamiento de señales y en su aplicación práctica en sistemas de reconocimiento de voz, procesamiento de señales médicas, de telecomunicaciones, entre otros. Para entender mejor este problema de separación ciega de señales y su solución a través del ICA se explicará el ejemplo conocido como efecto "*cocktail-party*" en el apartado 2.2.1.

El ICA surge como una metodología alternativa a los métodos de análisis de datos multivariantes clásicos. El objetivo principal del análisis de componentes independientes es conseguir factores latentes independientes que se hayan generado a partir de los datos observados. Estos factores latentes independientes reciben el nombre de componentes independientes y nos permiten reducir la dimensión de los datos.

El análisis de componentes independientes puede ser aplicado en numerosos ámbitos de muchos campos, se usa con mucha frecuencia en el campo de la ingeniería sobre todo para el análisis y gestión de señales digitales en tareas que van desde la eliminación de ruido de imágenes, eliminación de las interferencias en canales de comunicación y procesamiento y preprocesamiento en el campo de la acústica y la imagen en general.

También se usa constantemente en otros ámbitos como la medicina o la bioingeniería, donde por ejemplo se emplea para estudiar la actividad del cerebro analizando la información que viene dada por las distintas señales que se recogen de un encefalograma u otro tipo de análisis.

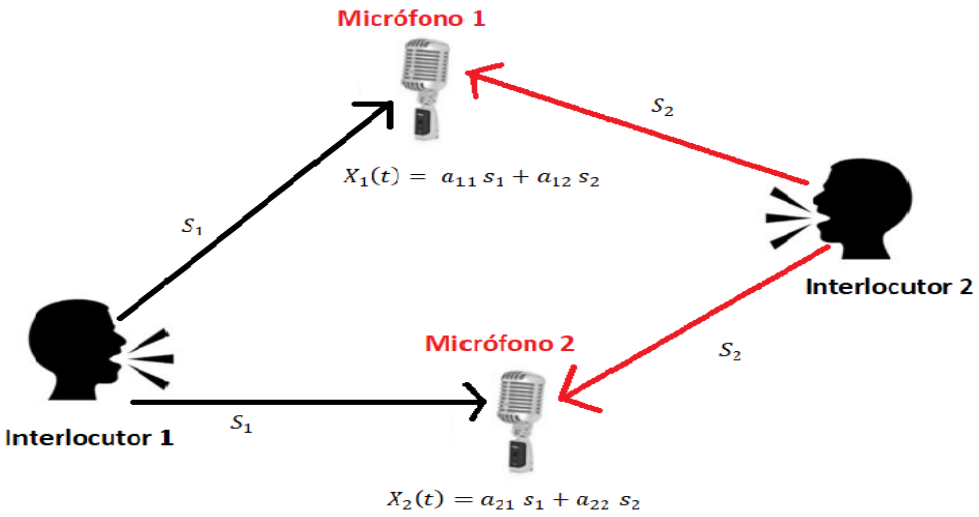
Alternativamente, en otros campos como la geografía, el ICA es muy útil para el análisis de datos sísmicos y otro tipo de señales meteorológicas.

En el aspecto financiero, el ICA se utiliza en muchos casos por delante de métodos clásicos como son los anteriormente mencionados análisis de componentes principales (PCA) y análisis factorial (FA) debido a la necesidad de adaptarse a los cambios de dinámica que se producen en los mercados financieros. Por lo que la finalidad de utilizar el análisis de componentes independientes en el ámbito financiero es buscar factores latentes de riesgo que sean independientes.

Cabe destacar que la implementación del ICA y de cualquier tipo de análisis de datos multivariantes a prácticamente cualquier disciplina no hubiera sido posible sin el desarrollo y las mejoras computacionales de los últimos años y los programas que existen de gestión y análisis de grandes bases de datos.

2.2.1 Cocktail-Party problem

Tal y como se ha mencionado en el apartado anterior el análisis de componentes independientes tiene su origen en buscar una solución al problema de separación ciega de señales. Un ejemplo habitual que se utiliza para entender fácilmente este problema y su solución a partir del análisis de componentes independientes es el *Cocktail-party problem* o *Cocktail-party effect*, definido por primera vez por Colin Cherry (1953).



Esquema 1: *Cocktail-party problem.* **Fuente:** Elaboración propia basado en Comon, J., & Jutten, C. (2010). *Handbook of Blind Source Separation: Inde-pendent Component Analysis and Applications.* Academic Press.

El funcionamiento del ejemplo definido por Colin Cherry se basa en lo siguiente:

Se colocan dos micrófonos en distintos lugares de una habitación donde hay dos personas hablando al mismo tiempo. Los micrófonos registran dos señales distintas que llamaremos $X_1(t)\ y\ X_2(t)$. Si nos fijamos en el esquema 1 dichas señales tienen una amplitud de $X_1\ y\ X_2$. $La\ t$ representa el índice temporal. Cada una de estas señales grabadas es una suma ponderada de las señales que emiten las dos personas al hablar que denominamos $S_1(t)\ y\ S_2(t)$. Podemos expresar este escenario mediante la ecuación lineal (1):

$$X_1(t) = a_{11}\, s_1 + a_{12}\, s_2$$

$$X_2(t) = a_{21}\, s_1 + a_{22}\, s_2 \qquad (1)$$

Donde $a_{11}, a_{12}, a_{21}, a_{22}$ representan parámetros que dependen de las distancias de los micrófonos a las personas.

Por lo que el problema es estimar las dos señales originales que emiten las dos personas al hablar $S_1(t)\ y\ S_2(t)$ utilizando únicamente las señales grabadas $X_1(t)\ y\ X_2(t)$, teniendo en cuenta que para este ejemplo no se consideran los retrasos u otros factores específicos de la transmisión de señales.

Si conociéramos los parámetros a_{ij}, que representa el parámetro de distancia entre el micrófono y las personas, podríamos resolver la primera ecuación lineal por los métodos clásicos, pero el hecho es que si no podemos conocer a_{ij} el problema se hace mucho más difícil de resolver.

Una posible solución a este problema sería usar la información que proporcionan las propiedades estadísticas de las señales $S_i(t)$ para estimar a_{ij}. Resulta que si asumimos que $S_1(t)\ y\ S_2(t)$ son estadísticamente independientes en cada instante t se pueden hallar de una forma sencilla los parámetros a_{ij} y recuperar las señales originales para resolver el problema. Asumiendo la independencia estadística de $S_1(t)\ y\ S_2(t)$ y mediante la aplicación del análisis de componentes independientes conseguimos desgranar las señales de cada micrófono y podemos conocer que interlocutor habla en cada momento del tiempo t y cuál es su mensaje o señal original.

3. ANÁLISIS DE COMPONENTES INDEPENDIENTES

3.1 Definición del modelo

El análisis de componentes independientes es una técnica estadística que se comenzó a aplicar en muchos campos de estudio a partir de los años 80. Inicialmente, este método se empleó en el campo de la ciencia para solucionar problemas de clasificación en Biología, pero se extendió hacia otros campos como el de la Psicometría, la Economía y las Ciencias Sociales en general.

Posteriormente, aproximadamente a mediados de los años 90 fue cuando se definió formalmente el método ICA. Según indicó Comon (1994) el objetivo principal del análisis ICA es que las señales observadas se transformen linealmente en componentes que son independientes entre sí. Actualmente se utiliza con mucha frecuencia en Ingeniería y las TIC para resumir la información disponible, elaborar métodos de clasificación y para el reconocimiento de patrones en los datos.

El modelo y la notación empleada por Ester González (2011) en su tesis doctoral para representar el método de análisis de componentes independientes es la siguiente:

En primer lugar se define $x = (x_1, \dots, x_m)'$ como un vector de observaciones de dimensión m. A través del método ICA se asume que x es un vector que se genera de forma lineal por un conjunto de r, con la restricción $r \leq m$, y con una distribución de los componentes a analizar que debe ser mutuamente independiente y no Gaussiana. El modelo ICA se representa de la siguiente forma:

$$\boldsymbol{x = As} \qquad (2)$$

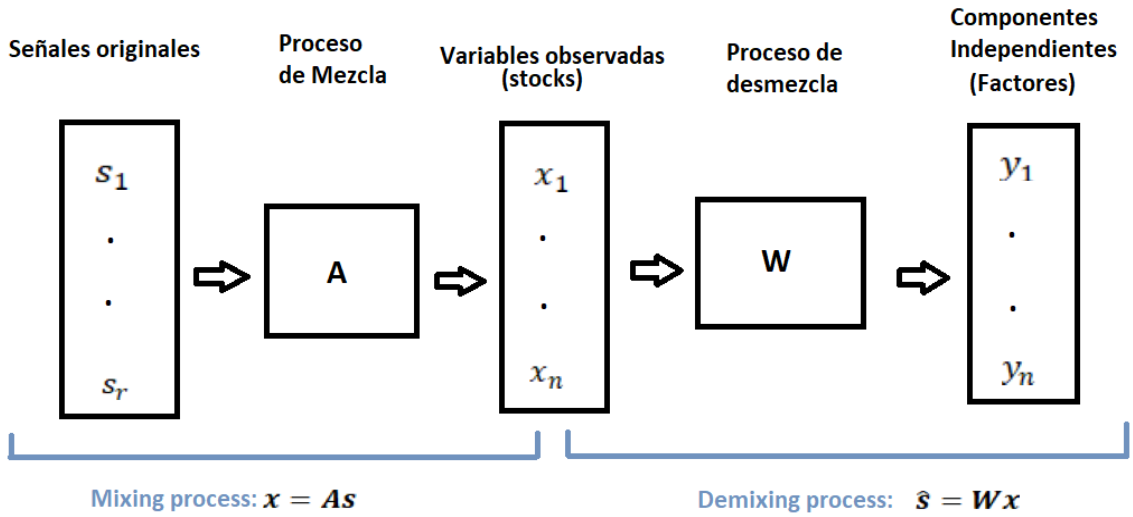
De la fórmula anterior (2) se deriva que $\boldsymbol{A}$ es una matriz $m\ x\ r$ desconocida de parámetros constantes, que recibe el nombre de matriz de mezclas.
Por otro lado, se define $s = (s_1, \dots, s_r)'$ como el vector de componentes subyacentes no Gaussianos y que son mutuamente independientes, que son los llamados componentes independientes (IC).

Si tomamos una muestra aleatoria de x, $(x_1, \dots, x_n)$, a partir de la aplicación del método ICA se logra estimar la matriz $\boldsymbol{A}$ y la $\boldsymbol{s}$, pero únicamente de las observaciones.

De esta forma, después de estimar la matriz $\boldsymbol{A}$, el ICA se centra en encontrar la transformación lineal de las observaciones:

$$\boldsymbol{\hat{s} = Wx} \qquad (3)$$

De la ecuación anterior se deriva que la matriz de separación $\boldsymbol{W}$ de dimensión $r\ x\ m$, se obtiene como la inversa de la matriz $\boldsymbol{A}$ ($\boldsymbol{W = A^{-1}}$) y los componentes de $\hat{\boldsymbol{s}}$ (los componentes independientes) se vuelven lo más independiente posible. La representación esquemática del funcionamiento sería la siguiente:



Esquema 2: Funcionamiento detallado del ICA. **Fuente:** Elaboración propia. Basado en la Tesis de Ester González. "Independent Component Analysis for Time Series" (2011).

Pero asumir la independencia estadística del vector $\boldsymbol{s}$ de componentes subyacentes no es suficiente para garantizar la fiabilidad del modelo (ecuación 2). Para ello Comon (1994) establece las siguientes suposiciones que permiten que el método ICA sea más fiable:

1. La primera suposición establece una restricción que indica que el número de variables observadas no puede ser mayor al número de componentes independientes, por lo que $r \leq m$. Adicionalmente, la matriz de mezclas $\boldsymbol{A}$ es cuadrada, ya que el rango de $\boldsymbol{A}$ es igual a r.

2. Las señales originales que se recuperan con el análisis de componentes independientes son estadísticamente independientes. Adicionalmente, las varianzas de los componentes independientes se fijan para ser iguales a una sola, siendo esta varianza: $var(s_{t1}) = \boldsymbol{I}_r$. A pesar de establecer una varianza equivalente para los componentes independientes, estos componentes siguen siendo indeterminados y manteniendo su signo.

3. No puede existir más de un componente independiente que siga una distribución Gaussiana. La existencia de más componentes Gaussianos hace que las observaciones sean cada vez más Gaussianas, con la consecuencia de que los componentes no podrán ser separados en el análisis (en base a lo que establece el Teorema Central del Límite, la suma de un conjunto de variables independientes tiende a una distribución Gaussiana).

Adicionalmente, bajo el supuesto de Gaussianidad, el análisis de componentes independientes y el análisis de componentes principales son equivalentes. Este supuesto justifica la aplicación del ICA cuando los datos no se distribuyen según una normal multivariante.

Si se asumen los tres supuestos anteriores se obtienen los componentes independientes que permiten maximizar la independencia y la distancia a la distribución normal.

3.2 Estimación del modelo

Dada la dificultad en la estimación del análisis de componentes independientes, la mayoría de los métodos propuestos para solucionar el método ICA incorporan restricciones que hacen más sencilla su resolución. Una de las restricciones más importantes se basa en imponer la ortogonalidad de la matriz de mezclas, $\boldsymbol{A}$, la cual es una matriz cuadrada de orden *m* (Ester González (2011)).

Ester González (2011) establece en su tesis sobre el análisis de componentes independientes que: "la solución del ICA se limita a un espacio de matrices ortogonales y el número de parámetros a estimar se reduce de m^2 en $\boldsymbol{A}$ hasta $\frac{m(m+1)}{2}$ en la nueva matriz ortogonal de mezclas".

Siguiendo el modelo y la notación empleada por Ester González (2011) en su tesis para la estimación del método ICA:

A partir de la ortogonalidad se consigue estandarizar las observaciones y esta restricción se introduce con el fin de transformar las observaciones originales en un nuevo conjunto de observaciones con media cero e idéntica matriz de varianzas y covarianzas, de tal forma que la fórmula es:

$$\boldsymbol{z} = \boldsymbol{V}\boldsymbol{x} \qquad (4)$$

Donde $\boldsymbol{V}$ es una matriz $m\ x\ m$ tal que $E\{\boldsymbol{z}\} = 0\ y\ \boldsymbol{V_z} = E\{\boldsymbol{z}\boldsymbol{z}'\} = \boldsymbol{I_m}$.

El método ICA emplea el análisis de componentes principales para para realizar la estandarización multivariante de **x**, es decir, para realizar el paso previo conocido como blanqueo y técnicas de reducción de dimensión de los datos. La forma de estandarizar los datos originales (**x**) a partir de la notación propuesta por Ester González (2011) es la siguiente:

Siendo $\boldsymbol{V_x} = E\{\boldsymbol{x}\boldsymbol{x}'\}$ la matriz de covarianzas de $\boldsymbol{x}$**.** La descomposición del valor propio de $\boldsymbol{V_x}$ viene dada por $\boldsymbol{V_x} = \boldsymbol{Q}\boldsymbol{D}\boldsymbol{Q}'$ donde $\boldsymbol{Q}_{m\ x\ m}$ es la matriz ortogonal de valores propios expresados en columnas y $\boldsymbol{D}_{m\ x\ m} = diag(d_1, \dots, d_m)$, con $d_1 \geq \cdots \geq d_m$es la matriz de valores propios.
Si utilizamos la fórmula $\boldsymbol{V} = \boldsymbol{D}^{-1/2}\boldsymbol{Q}'$ en la ecuación (4), las observaciones originales, $\boldsymbol{x}$, serán multivariantes estandarizadas y el método ICA original (ecuación 2) se puede expresar con datos estandarizados, $\boldsymbol{z}$, como:

$$\boldsymbol{z} = \widetilde{\boldsymbol{A}}\boldsymbol{s} \qquad (5)$$

Donde consideramos que $\widetilde{\boldsymbol{A}} = \boldsymbol{V}\boldsymbol{A} = \boldsymbol{D}^{-1/2}\boldsymbol{Q}'\boldsymbol{A}$ es la nueva matriz de mezclas de dimensión $m\ x\ m$ que se considera ortogonal.

Adicionalmente, a partir de la ortogonalidad se logra reducir la dimensión de los datos desde *m* hasta *r*, descartando los $m - r$ valores propios más pequeños de $\boldsymbol{V_x}$.

A través de combinación lineal de las observaciones estandarizadas se consigue estimar los componentes independientes (***y***)**:**

$$\boldsymbol{y} = \boldsymbol{W_z} \qquad (6)$$

De la ecuación anterior (6), $\boldsymbol{W}$ es una matriz ortogonal de dimensión $m\ x\ m$ ($r\ x\ r$ si la dimensión se ha reducido), donde los componentes son máximamente independientes.

3.3 Principios para la estimación del ICA

Para el análisis y estimación de los componentes independientes Hyvärinen et al. (2001) distinguen tres principios fundamentales en función del método empleado para calcular la independencia estadística:

1. **Maximización de la no Gaussianidad**. Existen numerosas formas de estimar los componentes independientes a través del análisis ICA. Una de las metodologías más conocidas es a través de la no Gaussianidad o no Normalidad de las observaciones. El objetivo de estos métodos es maximizar la independencia estadística de los componentes independientes mediante la maximización de su no Gaussianidad. A continuación, se muestran diversos métodos para estimar la no Gaussianidad:

   - El primer método implementado por Delfosse y Loubaton (1995) mesura la no Gaussianidad a partir de la estimación del coeficiente de kurtosis. La kurtosis de los componentes independientes viene dada por la siguiente fórmula:

     $$kurt(y_i) = E\{y_i^4\} - 3E\{y_i^2\}^2 \qquad (7)$$

     Donde la kurtosis de y_i puede tomar valores de signo positivo o negativo y será igual a 0 siempre que y_i sea Gaussiana. El problema principal de aplicar esta metodología es que la kurtosis puede ser muy sensible a los datos que son atípicos o *outliers*, lo cual provoca que esta forma de estimar la no Gaussianidad no sea muy robusta en términos estadísticos.

   - La segunda forma de mesurar la no Gaussianidad, que permite solventar el problema de los valores atípicos que presenta la kurtosis, hace referencia a un concepto llamado entropía. La entropía de una variable aleatoria se puede definir como el grado de información que dan las observaciones de la variable, de tal forma que cuanto más impredecibles y desestructuradas sean las variables, mayor será su entropía. Cover and Thomas (2001) definen que, para una determinada matriz de covarianzas, la distribución que tiene la entropía más alta es la distribución Gaussiana. Por lo tanto, el principio de maximizar la no Gaussianidad en este caso se corresponde con minimizar la entropía.

     Debido a que la entropía presenta ciertas limitaciones, como por ejemplo el hecho de no mantenerse invariante a las transformaciones lineales, se acepta la utilización de la negentropía como mesura de no Gaussianidad. La negentropía se considera una mesura de distancia a la distribución Normal (Gaussiana) definida de la siguiente forma por Cover and Thomas (2001):

$$J(\boldsymbol{y}) = H(\boldsymbol{y_{Gauss}}) - H(\boldsymbol{y}) \qquad (8)$$

De donde se deriva que $\boldsymbol{y}$ es un vector aleatorio y $\boldsymbol{y_{Gauss}}$ es un vector aleatorio cuya matriz de covarianzas es igual a la de $\boldsymbol{y}$.

Teniendo en cuenta que la negentropía no varía frente a las transformaciones lineales, que siempre es no negativa y cero si $\boldsymbol{y}$ es un vector Gaussiano se puede considerar un buen índice para estimar la no Gaussianidad. Como principal limitación, a pesar de ser la negentropía uno de los mejores métodos para estimar la no Gaussianidad, es muy difícil de aplicar en cuanto a tiempo de computación y por ello se usan aproximaciones como veremos en la parte práctica del trabajo utilizando el paquete de RStudio 'FastICA', cuyo fundamento de cálculo de la no Gaussianidad se basa en la negentropía.

- En tercer lugar, y como alternativa a los dos métodos anteriores para mesurar la no Gaussianidad, se pueden utilizar los cúmulos de orden superior. Estos cúmulos de orden superior se parecen mucho a los momentos de orden superior y los dos dan la misma información (Ester González (2011)). El método JADE definido por Cardoso y Souloumiac (1993) y presente en librerías del programa Rstudio, es un proceso para estimar el ICA que calcula los componentes independientes maximizando su no Gaussianidad utilizando los cúmulos de cuarto orden.

2. **Minimización de la información mutua**. Este principio hace referencia a la mesura y estimación de la dependencia estadística entre diversas variables aleatorias. La información mutua en el ICA tiene en cuenta toda la estructura de dependencia entre las variables y no sólo la matriz de covarianzas como ocurre en el análisis de componentes principales que veremos en los siguientes apartados.

   La información mutua (I) de un vector aleatorio r-dimensional $\boldsymbol{y} = (y_1, \dots, y_r)'$ se define por Cover y Thomas (2001) como:

   $$I(y_1, \dots, y_r) = \sum_{i=1}^{r} H(y_i) - H(\boldsymbol{y}) \qquad (9)$$

   Dado que el concepto de información mutua también está recogido por la entropía, el algoritmo FastICA permite recoger la información mutua usando diferentes aproximaciones.

3. **Maximización de la verosimilitud**. El principio de máxima verosimilitud puede aplicarse en el análisis ICA para estimar los componentes independientes (Gaeta and Lacome (1990) and Pham et al. (1992)). Dicho proceso equivale a la minimización de la información mutua.

Aunque los tres principios fundamentales anteriores hacen referencia a características muy diferentes existen relaciones entre ellos que permiten unificarlos. En primer lugar, existe una equivalencia matemática entre el principio de la información mutua y el de la máxima verosimilitud Cardoso (1997). Otra equivalencia es la que define Lee (2000) que afirma que la maximización de la negentropía (que se corresponde con maximizar la no

Gaussianidad) tiene propiedades equivalentes en los otros principios. A partir de las equivalencias definidas por los autores anteriores es posible unificar los tres principios en un solo método que permita estimar los componentes independientes, como veremos en la parte aplicada del trabajo con el programa RStudio y las librerías del ICA.

3.4 Comparación con los métodos de análisis multivariante clásicos

El análisis de componentes independientes se puede comparar con otros métodos de análisis multivariante clásicos utilizados para reducir la dimensión de los conjuntos de datos originales, como es el caso del análisis de componentes principales (PCA) o el análisis factorial (FA).

El análisis de componentes independientes, al igual que el análisis de componentes principales y el análisis factorial, se centran en encontrar una representación fidedigna de los datos a partir de la proyección de las observaciones en un espacio de menor dimensión. Es por este motivo que el ICA en muchos casos se considera una extensión del PCA y el FA. Cabe destacar que la definición de representación fidedigna de los datos es muy distinta si utilizamos un método de análisis u otro.

El PCA tiene como objetivo obtener componentes principales que no estén correlacionados entre sí. Estos componentes principales serán estadísticamente independientes solo si las observaciones tienen una distribución Normal (Gaussiana). Sin embargo, al aplicar el análisis ICA los componentes que se obtienen son estadísticamente independientes. Por este motivo el ICA se puede considerar una generalización del PCA, dado que los componentes independientes se pueden estimar a partir de la rotación de los componentes principales del PCA que los hacen lo más independientes posible (Ester González (2011)).

Adicionalmente, en el PCA la forma de ordenar los componentes principales se establece en base a sus varianzas. Se establece que el primer componente principal es aquel que recoge la máxima varianza posible, el segundo componente principal recoge la máxima varianza en el subespacio ortogonal restante, y la misma casuística con el resto de los componentes. Por lo tanto, las direcciones determinadas por el PCA, los llamados componentes principales, son ortogonales los unos con los otros y dirigidos a la máxima varianza (gráfico 1). Mientras que las direcciones que determina el ICA, los llamados componentes independientes, no son ortogonales y ofrecen una representación distinta a la que proporciona el método PCA (gráfico 1).

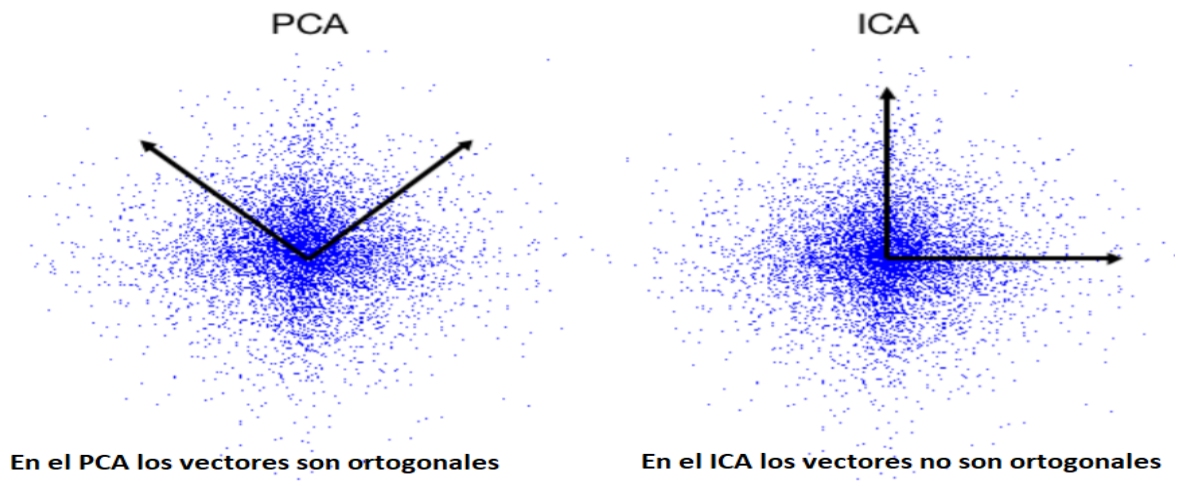


Gráfico 1: Comparativa ortogonalidad PCA/ICA. **Fuente:** Jon Shlens. *A tutorial on PCA*

La forma de ordenar los componentes independientes en el ICA se puede realizar a partir de la correlación existente entre los componentes principales y los componentes independientes (Daniel Peña, 2002). A partir del criterio planteado por Daniel Peña los componentes independientes se ordenan a partir del criterio de máxima correlación, de tal forma que el primer componente independiente es el que está máximamente correlacionado al primer componente principal, el segundo componente independiente tendrá máxima correlación con el segundo componente principal, y así sucesivamente.

En base a lo que establece J. Cardoso (1989) en su artículo *Source separation using higher order moments*, otra diferencia fundamental entre el PCA y el ICA es que para el PCA solo se estiman estadísticos de segundo orden, mientras que para el ICA se utilizan estadísticos de un orden superior a dos para la separación de las señales. Cabe destacar que la mayor parte de la información importante está contenida en los estadísticos de orden superior y que a efectos prácticos, las variables aleatorias raramente estarán distribuidas de forma Gaussiana o normal.

El ICA también presenta algunas similitudes con el análisis factorial (FA) en referencia a la ecuación inicial del método ICA (ecuación 2), ya que se asemeja mucho a la ecuación del método clásico FA. En el método FA se asume que los componentes subyacentes (o factores) no están correlacionados, mientras que el ICA asume la no Gaussianidad de los factores y su independencia estadística. Por lo tanto, el ICA puede entenderse como un método FA no Gaussiano que utiliza estadísticos de orden superior (Hyvärinen and Kano (2003)).

El ICA selecciona diferentes enfoques para procesar los datos en función de su estructura y revela distribuciones y estructuras más profundas de los datos. En general, el ICA se puede considerar una técnica mucho más adecuada a la naturaleza de los datos financieros que los métodos clásicos por su capacidad de encontrar factores o componentes cuando los métodos clásicos no se ajustan a la realidad de los datos.

A efectos prácticos, los métodos PCA, FA e ICA pueden ser implementados con una gran variedad de algoritmos diferentes que están disponibles en las librerías y paquetes del programa RStudio. En este trabajo se aplicarán algunos de estos algoritmos sobre una base de datos financiera para ver de una forma más sencilla y visual cuales son las diferencias y similitudes entre los métodos clásicos y el ICA. Además, también se aplicará

el método alternativo NMF que, de igual manera que los métodos PCA, FA y ICA, realiza una factorización de matrices. El motivo principal de aplicar el método NMF en este trabajo es que, a pesar de ser un método que no genera componentes o factores subyacentes, sí que permite agrupar las variables en función de características similares.

3.5 Modelo alternativo: *Nonnegative Matrix Factorization*

La factorización matricial no negativa (NMF) es una metodología empleada para el análisis multivariante de datos. Se puede comparar en algunos aspectos con otras técnicas de análisis como las que hemos visto anteriormente, ya sea el PCA o el ICA. Pero en este caso, el NMF difiere de los anteriores porque impone y requiere que la factorización de las matrices sea no negativa.

La restricción del método NMF a utilizar matrices de datos no negativas se debe a los motivos siguientes definidos por Lee and Seung (1999):

- En primer lugar, en ciertos campos o ramos de estudio el conjunto de observaciones que se emplea no puede ser negativo, un ejemplo claro de esta casuística se encuentra en las aplicaciones de ingeniería, donde las técnicas de procesado de imagen no pueden contener datos negativos.

- Adicionalmente, el hecho de requerir que las bases de datos sean no negativas se puede emplear para defender que las partes se unen de forma aditiva y no mediante diferencias.

El modelo NMF permite una representación adecuada de los datos a la vez que muestra ciertas características de las observaciones que no se identifican *a priori*. El NMF tiene como objetivo establecer una función lineal de los datos, que deben ser no negativos y se define por Z. Cazalet y T. Roncalli (2011) de la siguiente forma:

$$A \approx BC \qquad (10)$$

Donde *B* y *C* son dos matrices no negativas con dimensiones $m\ x\ n\ \ y\ \ n\ x\ p$. Uno de los principales inconvenientes de este método es que las dimensiones de *m*, *n* y *p* pueden ser muy grandes y resulta muy complicado estimar esta fórmula en un tiempo razonable a nivel computacional, igual que ocurre con los métodos clásicos y el ICA.

Uno de los primeros algoritmos implementados para desarrollar el método NMF fue propuesto por Lee and Seung (1999). Esta primera implementación del método era sencilla pero muy útil para aplicar con éxito el reconocimiento de patrones comunes. A partir de este primer algoritmo sencillo se ha logrado pulir y mejorar el análisis y la aplicación del método NMF.

El algoritmo propuesto por Lee and Seung (2001) para la factorización no negativa de matrices del método NMF permite medir la calidad de la factorización a partir de la función de coste *f*. La optimización de la función de coste *f* es la siguiente:

$$\{\hat{B}, \hat{C}\} = \arg\min f(A, BC) \qquad sujeto\ a \begin{cases} B \geq 0 \\ C \geq 0 \end{cases} \qquad (11)$$

Donde las matrices *B* y *C* deben ser estrictamente positivas (no negativas).

Lee and Seung (2001) consideraron en su estudio dos tipos de funciones de coste. La primera es la que seguía la norma de Frobenious y la segunda es la divergencia de Kullback-Leibler.

En este trabajo solo se definirá la primera función de coste que sigue la norma de Frobenious:

$$f(A, BC) = \sum_{i=1}^{m} \sum_{j=1}^{p} (A_{i,j} - (BC)_{ij})^2 \qquad (12)$$

Como se ha mencionado anteriormente, si el conjunto de observaciones es muy grande, el tiempo de computación para calcular la función de coste y obtener una solución para la ecuación 12 sería demasiado elevado. Para ello se han introducido técnicas que permiten estimar un resultado óptimo en un tiempo razonable, con el fin de resolver los problemas del método NMF con grandes bases de datos. En la parte práctica del trabajo se estimará y analizará nuestra base de datos financiera a partir de este método NMF.

De igual manera que el ICA, el NMF también es una técnica que se puede emplear no tan solo para el análisis de datos financieros, sino también para el procesamiento de imagen y sonido en el campo de la ingeniería. También es utilizado con frecuencia en otros campos como la medicina, en aplicaciones como el análisis de señales cerebrales o del sistema nervioso.

En el campo de las finanzas, el método NMF es muy útil para identificar patrones del mercado de valores, ya que nos permite visualizar características comunes en la dinámica del precio de las acciones. Esta aplicación la veremos en la parte práctica del trabajo aplicando la librería y los paquetes disponibles en RStudio sobre el NMF, el paquete denominado NMF.

A modo de resumen, el método NMF se centra en buscar una función lineal en las observaciones, que además deben ser no negativas. Si comparamos el resto de las técnicas de análisis multivariante clásicas (PCA y FA) y el ICA con el método NMF se observa que, el método NMF también logra reducir la dimensión de los datos, pero sin generar componentes o factores latentes. Adicionalmente, a través del método NMF, la reducción de dimensión de las bases de datos permite identificar características propias de los datos que no se observaban inicialmente.

Una diferencia destacable entre el método NMF frente a los métodos clásicos y el ICA se encuentra en las restricciones que se fijan para cada técnica y los resultados obtenidos. De este modo, para el método NMF es necesario que las matrices de datos sean estrictamente no negativas. Sin embargo, para los métodos PCA, FA e ICA no es necesario ningún tipo de restricción en el signo de las bases de datos, pudiendo ser positivas o negativas.

Finalmente, se identifica otra diferencia relevante entre el método NMF y el ICA. Para aplicar la técnica NMF no se tiene en cuenta la dependencia estadística que existe entre las variables aleatorias estimadas (Lee and Seung (1999)). Mientras que si aplicamos el método ICA es requisito indispensable que las variables sean estadísticamente independientes, como se ha visto reflejado en el apartado 3.3 de este trabajo. Por lo que estas restricciones también suponen un problema en la aplicación del método ICA y habrá que analizar si permiten que este sea un método útil a la práctica para el ámbito financiero.

3.6 Taxonomía

A continuación, se muestra un cuadro resumen que recoge las características principales de tres de los métodos estudiados PCA, FA, ICA[4]. Las ventajas que se muestran en el cuadro son aquellas que diferencian unos métodos de otros, dado que todos los métodos comparten ventajas básicas como permitir la reducción de dimensión de los datos o permitir factorizar los datos en componentes o factores comunes.

Método	Definición	Ventajas	Desventajas
PCA	Busca proyectar los datos en la dirección de la varianza máxima. Todos los componentes principales se ordenan en función de sus varianzas. Siendo el primer componente principal el que define la dirección de la máxima varianza posible, el segundo componente principal define la dirección de la máxima varianza en el subespacio restante, y así sucesivamente.	- Determina en pocos factores (componentes principales) la variabilidad que contienen los datos. - Las variables obtenidas son independientes (bajo el supuesto de normalidad) e incorrelacionadas.	- Se limita a tener en cuenta la información estadística de segundo orden. - No podemos afirmar que las componentes principales son independientes (si no hay normalidad), solo podemos afirmar que están incorrelacionados.
FA	Se basa en el análisis de la relación de una gran cantidad de indicadores y variables con el fin de representar la estructura subyacente. Además, el FA asume que los componentes subyacentes no están correlacionados, lo que hace posible reducir los datos a un número más manejable de variables agregadas.	- Reduce el riesgo de contar varias veces atributos que están muy correlacionados. - Tiene en cuenta el error de medición, mientras que el PCA no.	- Los pesos de los factores varían mucho si añadimos más variables a nuestro estudio. - Difícil interpretación gráfica de los resultados finales obtenidos.
ICA	Método analítico para datos multivariantes que tiene como objetivo principal encontrar componentes que sean estadísticamente independientes y no Gaussianos a partir de dichos datos (como mucho una componente Gaussiana).	- Usa información estadística de un orden más elevado que el PCA para separar las señales. - Técnica mucho más potente que los métodos clásicos por su capacidad de encontrar factores o componentes cuando los métodos clásicos (PCA, FA) fallan.	- Requiere la no Gaussianidad de los datos. - No da ninguna información sobre la varianza de la señal original ni de su signo (se busca la varianza unitaria de las características iniciales).

[4] En este caso no se compara el método NMF porque se trata de un método de análisis multivariante que no tiene como fin la reducción de dimensión de los datos en nuevos componentes o factores latentes como sí ocurre con el PCA, FA y ICA.

4. APLICACIÓN DE LA METODOLOGÍA ICA A SERIES FINANCIERAS

4.1 Introducción del ICA a las series financieras

A diferencia de los datos estadísticos estáticos que se comparan a la práctica en muchos ámbitos como la economía, la industria y los mercados, también existe una gran cantidad de series de datos dinámicos con atributos multivariantes, que constituyen los datos multivariantes, como es el caso de los precios diarios de cierre de mercado de las empresas que estudiaremos en este apartado.

La dinámica y el comportamiento de la demanda de mercado se encuentra en continuo cambio especialmente en finanzas. En este apartado aplicaremos el ICA sobre diez series de datos financieros. Este método se ha convertido en una de las técnicas de descomposición de señales más importantes en los últimos años con el fin de extraer información potencial del mercado para mejorar las predicciones y hacerlas más acordes a la dinámica de cambios del mercado.

El análisis de datos dinámicos es uno de los campos más ampliamente estudiados en estadística. Hay una gran cantidad de información acerca de varios modelos de predicción, como por ejemplo el de redes neuronales o el modelo autorregresivo integrado de media móvil (ARIMA), pero hay que tener en cuenta que estos métodos solo se basan en los datos históricos.

Aunque fijarse en los datos históricos es muy importante también hay que tener en cuenta los componentes o características propias de los datos, ya que en muchos casos son decisivas en el comportamiento de los activos en el mercado y sin ellos la precisión del análisis de datos se ve muy mermada. Para solucionar las limitaciones de los métodos tradicionales se necesita que los análisis de datos multivariantes tengan en cuenta lo máximo posible las características o factores propios de los datos.

Para ello ha surgido como herramienta el ICA, que es un método estadístico que permite encontrar componentes subyacentes de variables aleatorias o señales. El ICA define un modelo para los datos multivariantes observados que generalmente viene dado por muestras de grandes bases de datos, en nuestro caso diez series de datos temporales financieros que vienen dadas por los precios de cierre diarios de diez empresas tecnológicas a lo largo de un año.

En los siguientes apartados se aplicará la metodología PCA, FA, ICA y NMF sobre una base de datos de los rendimientos de 10 activos financieros y se explicará paso a paso todos los procedimientos llevados a cabo para obtener una conclusión del estudio. También se elaborará una comparativa general entre los métodos de análisis multivariante que realizan reducción de dimensión en componentes o factores latentes, es decir, se compararán los métodos PCA, FA y ICA. Para realizar todos estos análisis se empleará la herramienta RStudio a partir de un código y diferentes paquetes que permiten aplicar los métodos anteriormente mencionados.

4.2 Técnicas de análisis de datos multivariantes clásicas: PCA Y FA

4.2.1 Pre-procesado de nuestra base de datos

Para realizar el análisis de los estadísticos clásicos principales, tanto del análisis de componentes principales (PCA) como del análisis factorial (FA), de una base de datos formada por los precios diarios de 10 empresas del sector tecnológico es necesario realizar un pre-procesado de los datos para hacer posible su análisis.

Analizando los precios de la base de datos de los 10 activos se observa que los precios de los 10 activos no son estacionarios en media ni en varianza por lo que es necesario aplicar transformaciones sobre la serie con el fin de conseguir esta estacionariedad que haga que ninguna variable con mayor varianza o media predomine por encima del resto de variables.

Para ello se aplica una primera transformación a los precios de las 10 empresas calculando su rentabilidad absoluta a partir de la diferencia de precios diarios ($R_{abs} = P_t - P_{t-1}$) con lo que gráficamente observamos que se ha conseguido estabilizar las 10 series de los activos en media. Dado que se observa que las series obtenidas siguen siendo no estacionarias en varianza es necesario realizar otra transformación con el fin de estabilizarlas en varianza. La transformación que aplicaremos sobre nuestras series será aplicar el logaritmo a los rendimientos absolutos anteriores ($R_{cont} = \ln(P_t - P_{t-1})$), de esta forma se obtiene la rentabilidad relativa continua que hace que nuestras series de datos de los 10 activos sean estacionarias tanto en media como en varianza y podamos trabajar con ellas para aplicar los métodos de análisis de datos multivariantes que veremos en los siguientes apartados.

Todos estos cálculos de pre-procesado de datos han sido aplicados sobre nuestra base de datos de los 10 activos tecnológicos con el fin de poder realizar el análisis multivariante en RStudio.

4.2.2 Análisis de Componentes Principales (PCA)

El análisis de componentes principales (PCA) es una técnica de análisis de datos multivariantes que permite la reducción de dimensión de los datos, perdiendo la menor cantidad de información posible. Cuando disponemos de un gran número de variables cuantitativas posiblemente correlacionadas, como es el caso de nuestra base de datos de los rendimientos diarios de 10 empresas tecnológicas durante un año, el PCA nos permite reducirlas a un número menor de variables transformadas, que son los llamados componentes principales. Cada componente principal generado por el PCA es una combinación lineal de las variables originales y permiten explicar una gran parte de la variabilidad de los datos. Cabe destacar que dichos componentes principales serán independientes y no correlacionados entre sí. Por lo tanto, el PCA también nos será muy útil a la hora de aplicar el análisis de componentes independientes (ICA), ya que el PCA se puede utilizar como fase previa al ICA para el centrado y blanqueamiento de los datos observados.

Para poder calcular los componentes principales del PCA es importante la unidad de medida empleada en las variables que se observan. Por lo tanto, es importante que antes de aplicar el PCA se estandaricen las variables para que tengan media 0 y sean estacionarias en varianza como se ha explicado en el apartado anterior 4.2.1, ya que de lo contrario las variables con una mayor varianza serían las que predominarían sobre el resto de las variables estudiadas.

En primer lugar, una vez aplicados todos los procesos previos sobre los datos, tenemos que determinar cuál es el número de componentes principales que se encuentran en nuestra base de datos. Para ello existen dos formas: una es sin inferencia estadística y la segunda es a partir del contraste de raíces características no relevantes. Pero para utilizar el segundo método de contraste de raíces para determinar el número de componentes principales es necesario que las variables originales de nuestra base de datos se distribuyan según una normal multivariante.

Aplicando el paquete de RStudio MVN se puede comprobar si nuestros datos se distribuyen según una normal multivariante. Este paquete realiza diversas pruebas sobre nuestros datos para comprobar la normalidad multivariante, entre ellos el Test de Mardia y el Test de Henze-Zirkler. Aplicando ambos test sobre nuestra base de datos de 10 activos tecnológicos se rechaza la hipótesis nula de que los datos siguen una normal multivariante. Por lo tanto, no podemos determinar el número de componentes principales a partir del contraste de raíces características no relevantes y será necesario utilizar el método sin inferencia estadística.

Si determinamos el número de componentes principales a partir de la inferencia estadística se necesita tanto la matriz de varianzas y covarianzas como la matriz de correlaciones. Generalmente, se pueden obtener tantas componentes principales distintas como variables haya disponibles en nuestra base de datos. La elección se realiza de tal manera que la primera componente principal sea la que mayor varianza recoja; la segunda debe recoger la máxima variabilidad que no recoja la primera, y así sucesivamente eligiendo al final un número de componentes principales que recoja un porcentaje suficiente de varianza total.

Utilizando la matriz de correlaciones de nuestros datos, el número de componentes principales se puede determinar a partir de la función con matrices eigen que ofrece RStudio. A partir de esta función se obtiene un vector de valores propios de la matriz de correlaciones y el número de componentes principales vendrá determinado por los valores que excedan la unidad ($x > 1$) en dicho vector.

```
eigen() decomposition
$values
 [1] 5.0409680 0.9339124 0.8526047 0.6997569 0.5950797 0.4985375 0.4657806 0.4011508 0.3393223 0.1728872
```

Figura 1: Vector de valores propios a partir de la matriz de correlaciones para estimar los PCs. **Fuente:** Elaboración propia

En nuestra base de datos de 10 empresas tecnológicas se observa que solo hay un valor del vector que exceda la unidad (figura 1), aunque el segundo valor está muy próximo a uno y se podría considerar para nuestro estudio. Por lo tanto, en nuestra base de datos utilizando solo la matriz de correlaciones para estudiar el número de componentes

principales se llega a la conclusión de que solo hay un componente principal o como mucho dos si consideramos que el segundo se aproxima mucho a la unidad.

Si utilizamos el análisis de la matriz de varianzas y covarianzas para determinar el número de componentes principales el funcionamiento es distinto. En este caso utilizando la función eigen de RStudio sobre la matriz de varianzas de nuestra base de datos obtendremos un vector de valores propios distinto al que obteníamos con la matriz de correlaciones. En este caso el número de componentes principales vendrá determinado por el número de valores del vector que excedan la media del conjunto ($x > \bar{x}$).

```
eigen() decomposition
$values
 [1] 16.7606311  4.1400519  2.7034582  2.2281782  2.0155074  1.7901225  1.4406823  1.1038448  0.6953625  0.5496788
 mean
   3.342752
```

Figura 2: Vector de valores propios a partir de la matriz de varianzas para estimar los PCs. **Fuente:** Elaboración propia

Si nos fijamos en los resultados obtenidos sobre nuestra base de datos (figura 2) se observa que hay dos valores del vector de valores propios que exceden la media del conjunto (la media es 3,34). Por lo tanto, la interpretación es que el número de componentes principales que se observan en nuestra base de datos es de dos utilizando la matriz de varianzas como fuente del análisis, resultado muy similar al que obtenemos usando la matriz de correlaciones.

Además, RStudio también dispone de una función llamada *princomp* a partir de la cual podemos aplicar el análisis de componentes principales con la finalidad de saber cuál es la importancia de cada componente en términos de varianza. Es decir, a partir de esta función se puede representar la proporción de varianza explicada de cada componente y cuanto acumulan conjuntamente entre ellos.

```
Importance of components:
                          Comp.1     Comp.2     Comp.3     Comp.4     Comp.5     Comp.6     Comp.7     Comp.8
Standard deviation     2.2452100 0.96639141 0.92336596 0.83651472 0.77141406 0.70607188 0.68248122 0.63336468
Proportion of Variance 0.5040968 0.09339124 0.08526047 0.06997569 0.05950797 0.04985375 0.04657806 0.04011508
Cumulative Proportion  0.5040968 0.59748804 0.68274851 0.75272420 0.81223216 0.86208591 0.90866397 0.94877905
                           Comp.9    Comp.10
Standard deviation     0.58251375 0.41579706
Proportion of Variance 0.03393223 0.01728872
Cumulative Proportion  0.98271128 1.00000000
```

Figura 3: Importancia de cada componente y proporción de varianza explicada. **Fuente:** Elaboración propia

Si aplicamos la función *princomp* sobre nuestra base de datos (figura 3) se observa que la primera componente tiene una proporción de varianza explicada de aproximadamente el 50% y si añadimos la segunda componente acumulan un 59,7% de varianza explicada entre los dos. Por lo que, si tenemos en cuenta nuestros análisis aplicados a la matriz de varianzas y a la matriz de correlaciones realizados anteriormente, llegamos a la conclusión de que nuestro modelo tenía dos componentes principales que explican

aproximadamente un 60% de la varianza total. Podríamos llegar a plantearnos incluir una tercera componente, ya que este 60% de varianza explicada se encuentra muy en el límite de lo comúnmente aceptado en el PCA aplicado a finanzas, esto supondría que la varianza explicada de nuestro modelo alcanzaría el 70% aproximadamente. Pero dado que el análisis de los vectores de valores propios de las matrices de correlaciones y varianzas nos llevaban a la conclusión de que solo había dos componentes principales y la proporción de varianza explicada por las dos componentes es suficiente (aunque muy justa), nos quedaremos con dos componentes para nuestro estudio.

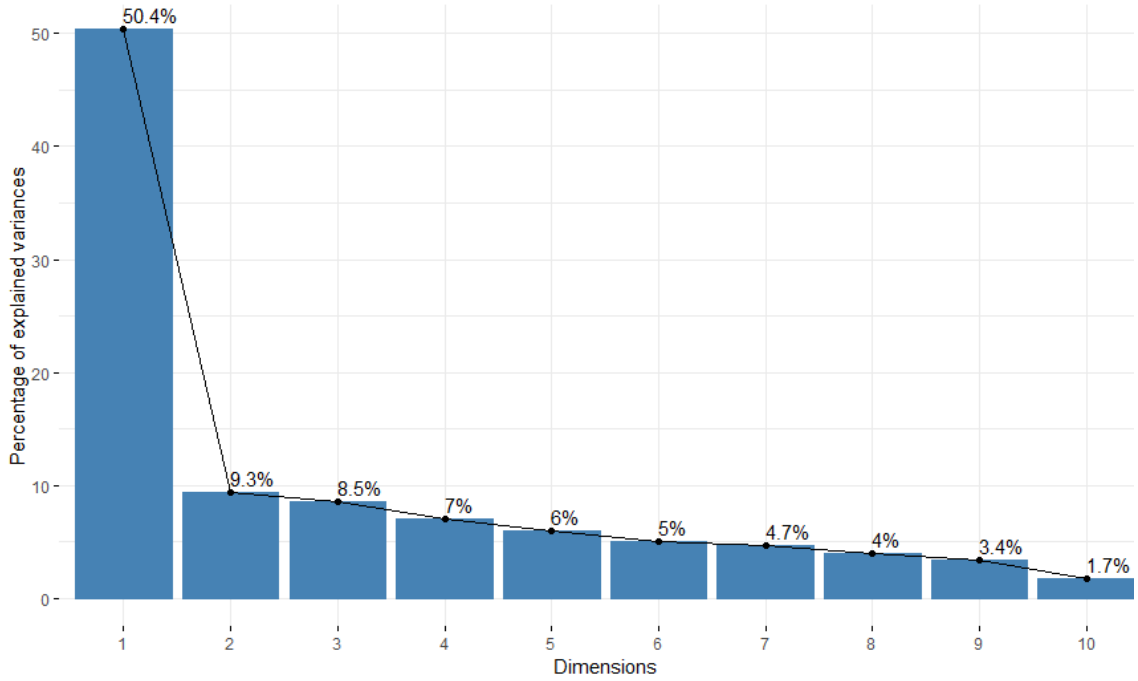


Gráfico 2: Porcentaje de varianza explicada por cada componente. **Fuente:** Elaboración propia

A través del paquete FactoMiner de RStudio podemos obtener gráficos muy representativos que nos ayuden a entender más fácilmente cuales son los componentes principales que tienen un mayor porcentaje de varianza explicada. Por ejemplo, el gráfico 2 se ha obtenido a partir de esta librería de RStudio y nos permite ver que el primer componente acumula la mayor parte de varianza explicada de todo el modelo con un 50,4%, mientras que el segundo componente suma un 9,3% de varianza explicada al modelo, cantidad levemente superior a la que aporta la tercera componente y que nos podría llevar a pensar que la tercera componente también es importante para representar la varianza explicada del modelo.

Con otros paquetes de RStudio, como el corrplot o el ggplot2 combinados con el FactoMiner anterior, se pueden obtener también gráficos y diagramas muy representativos que nos permiten ver que variables de nuestra base de datos de 10 empresas tecnológicas tienen más incidencia sobre cada una de las componentes principales estimadas del modelo.

El siguiente gráfico, el gráfico 3, nos muestra cual es el impacto que tiene cada componente (eje horizontal) sobre cada una de las empresas estudiadas en el trabajo (eje vertical). El impacto viene medido a partir de círculos azules de distintas tonalidades, dichas tonalidades azules se van oscureciendo en función de si se trata de un impacto muy

fuerte de la componente sobre la empresa, o bien se va aclarando el tono azul si se trata de un impacto más leve de la componente sobre la empresa.

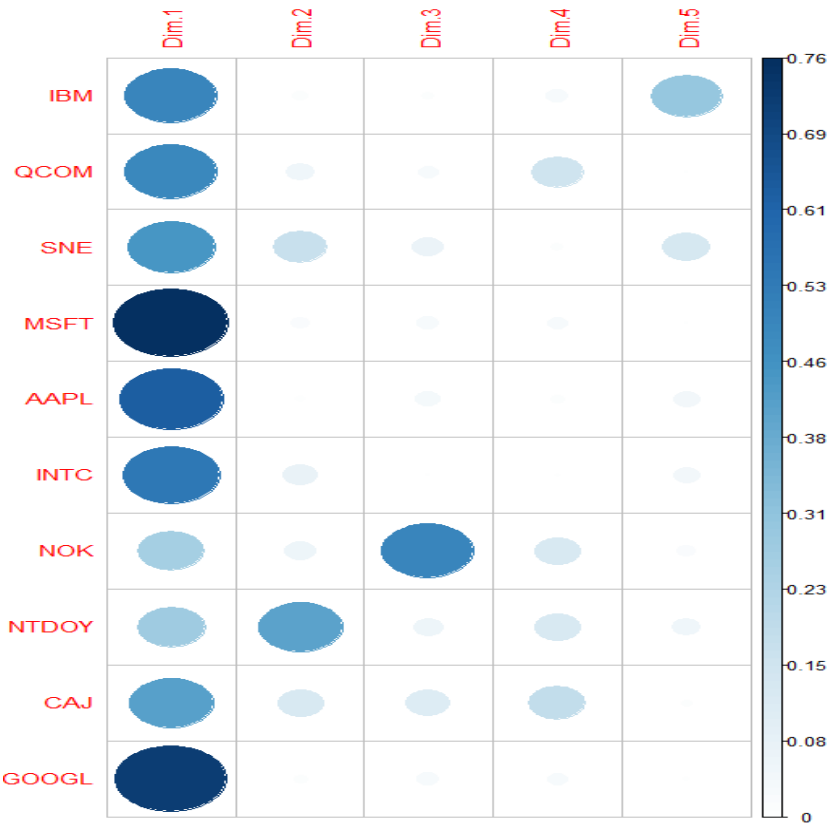


Gráfico 3: Impacto de cada componente sobre cada una de las empresas. **Fuente:** Elaboración propia

En referencia a nuestra base de datos en el gráfico 3 se observa que la primera componente principal tiene mucho impacto sobre todas las empresas estudiadas, siendo Microsoft (MSFT) la que más impacto recibe de la primera componente, seguida de Google (GOOGL), Apple (AAPL) e Intel (INTC), que también reciben mucho impacto de la primera componente principal. La segunda componente principal tiene mucho menos efecto sobre las empresas respecto a la primera, y esto tiene sentido dado que la primera componente recoge un 50% de la varianza explicada del modelo y la segunda componente solo recoge un 9% aproximadamente. Aun así, la segunda componente principal también tiene impacto sobre la empresa Nintendo (NTDOY), Sony (SNE) y Canon (CAJ) aunque mucho menor que el impacto que produce la primera componente. También es interesante destacar que, aunque consideremos que el modelo aplicado a nuestra base de datos queda explicado por únicamente dos componentes principales, en el gráfico 3 se observa como la tercera componente principal tiene mucho impacto sobre la empresa Nokia (NOK) y la quinta componente principal sobre IBM.

Utilizando solo las dos componentes principales que hemos considerado para nuestro modelo existe otra forma de ver gráficamente que empresas se ven más afectadas por cada componente y en qué cantidad. En el siguiente gráfico, el gráfico 4, se muestra en un diagrama que tiene como ejes la componente principal uno (eje horizontal o Dim 1) y la componente principal dos (eje vertical o Dim 2). Dentro de estos ejes se representa mediante flechas a las 10 empresas estudiadas y se posicionan en función de que componente les afecta más. Además, estas flechas tienen diferentes colores en función de si el efecto de la componente principal es mayor o menor.

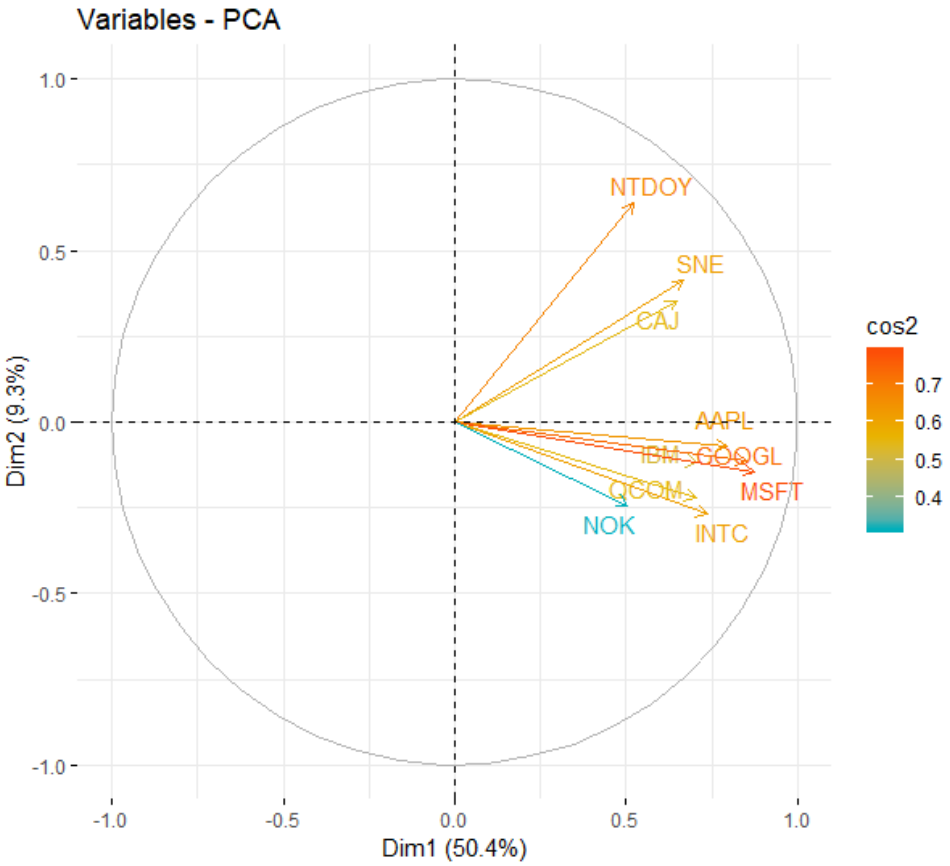


Gráfico 4: Diagrama que relaciona las dos componentes principales con su peso en cada empresa.
Fuente: Elaboración propia

En el gráfico 4 observamos como la primera componente principal (Dim 1) es la que tiene efecto sobre todas las empresas, ya que todas se encuentran en la posición derecha del gráfico circular que representa algún tipo de impacto de la primera componente. Las empresas con un mayor impacto de la primera componente son Microsoft (MSFT), Google (GOOGL) y Apple (AAPL), ya que las flechas están representadas con colores más anaranjados que muestran un mayor impacto. Mientras que Nokia (NOK) es la que claramente tiene menos impacto de la componente principal uno, ya que la flecha de color azul representa un menor impacto. Esto puede ser debido a que, como hemos visto en el gráfico 3, la empresa Nokia (NOK) está muy afectada por la tercera componente principal que no se considera en nuestro modelo.

En cuanto al impacto de la segunda componente principal, que se representa en la parte superior del gráfico circular, vemos como Nintendo (NTDOY), Sony (SNE) y Canon (CAJ) son las empresas que reciben una mayor incidencia de la segunda componente principal, de igual forma que observábamos en el gráfico 3 anterior. A pesar de que estas empresas reciben incidencia de la segunda componente principal, cabe destacar que también la reciben de la primera componente principal, dado que se encuentran localizadas en la parte superior derecha del gráfico circular.

4.2.3 Análisis Factorial (FA)

El otro método estadístico clásico de análisis multivariante de datos que aplicaremos sobre nuestra base de datos de 10 activos tecnológicos es el análisis factorial (FA). Esta técnica de análisis propone un modelo explícito en el cual se explican las variables observadas a partir de unos factores comunes y únicos, pero que no son observables.

Para realizar el análisis factorial es necesario realizar dos contrastes previamente. En primer lugar, es necesario realizar un análisis ex-ante a partir del contraste de esfericidad de Bartlett. Este contraste se utiliza para evaluar la aplicabilidad del análisis factorial a las variables que se estudian.

El objetivo para este contraste es rechazar la hipótesis nula, con lo que habrá suficiente multicolinealidad[5] entre las variables para que se pueda aplicar el análisis factorial sobre nuestra base de datos (esto significa que el p-value<0,05).

```
        Bartlett's Test of Sphericity

Call: bart_spher(x = activos)

     X2 = 1149.681
     df = 45
p-value < 2.22e-16
```

Figura 4: Resultados del Test de esfericidad de Bartlett. **Fuente:** Elaboración propia

Aplicando el Test de Bartlett sobre nuestra base de datos de rendimientos de 10 empresas tecnológicas mediante la función de RStudio *bart_spher* que proporciona el paquete REdas se observa que el p-value obtenido es inferior a 0,05, por lo que se rechaza la hipótesis nula y se acepta la hipótesis alternativa de que nuestra matriz es distinta a la matriz identidad. De esta forma, a partir de este test confirmamos que el análisis factorial es aplicable sobre nuestra base de datos y que hay multicolinealidad entre nuestras variables.

El segundo paso para el análisis factorial es realizar un estudio ex-post mediante la medida de bondad del ajuste que se basa en otro contraste distinto al del análisis ex-ante. Para realizar este contraste se utilizará la función de RStudio *factanal*, que nos permitirá saber cuántos factores comunes son necesarios para explicar nuestro modelo. Mediante la función de RStudio podemos ir probando con distinto número de factores comunes hasta encontrar una cantidad de factores que logre aceptar la hipótesis nula que plantea el modelo y que por lo tanto represente un número de factores válido.

También cabe destacar que el número de factores comunes que existe en nuestro modelo no puede exceder el número de variables que se estudian, en nuestro caso el número de factores comunes no puede ser superior a 10, que son las diez empresas tecnológicas que estudiamos. Además, para que el número de factores comunes a emplear sea válido también se tiene que cumplir una restricción que plantea el modelo de análisis factorial, la restricción es la siguiente:

[5] Multicolinealidad: Fuerte correlación de las variables explicativas del modelo.

$$p(p+1)/2 \geq p(m+1) \quad (13)$$

Donde p son las variables y m representa el número de factores comunes.

En primer lugar, si tenemos en cuenta la restricción anterior (13) que plantea el análisis factorial, en nuestro modelo no pueden proponerse más de 4 factores comunes, ya que poniendo 5 factores se incumple la restricción. Por lo tanto, se podrían proponer para nuestro modelo de 10 empresas tecnológicas de 1 a 4 factores comunes para comprobar su validez mediante el p-value en la función factanal de RStudio.

Si realizamos esta comprobación se observa que utilizando un factor para explicar nuestro modelo se rechaza la hipótesis nula de que el modelo que utiliza factor es válido, por lo que también tenemos que desechar la opción de que un factor común es suficiente.

Si comprobamos el p-value utilizando 2, 3 o 4 comprobamos que en todos los casos se acepta la hipótesis nula de que nuestro modelo utilizando esa cantidad de factores es válido. Por lo tanto, tendremos que seleccionar entre dos, tres o cuatro factores comunes para plantear un modelo con nuestra base de datos. Una forma adecuada de seleccionar los factores es teniendo en cuenta la varianza explicada utilizando cada opción factores.

```
Call:
factanal(x = activos, factors = 4, rotation = "varimax")

Uniquenesses:
  IBM  QCOM   SNE  MSFT  AAPL  INTC   NOK NTDOY   CAJ GOOGL
 0.57  0.00  0.44  0.12  0.41  0.49  0.65  0.65  0.47  0.23

Loadings:
      Factor1 Factor2 Factor3 Factor4
MSFT   0.83
AAPL   0.60    0.34
GOOGL  0.75    0.31
SNE            0.64
CAJ            0.58            0.34
QCOM   0.33            0.91
NOK                            0.52
IBM    0.46
INTC   0.50                    0.38
NTDOY  0.34    0.49

               Factor1 Factor2 Factor3 Factor4
SS loadings       2.43    1.43    1.20    0.89
Proportion Var    0.24    0.14    0.12    0.09
Cumulative Var    0.24    0.39    0.51    0.60

Test of the hypothesis that 4 factors are sufficient.
The chi square statistic is 7.19 on 11 degrees of freedom.
The p-value is 0.784
```

Figura 5: Resultado del análisis factorial con 4 factores comunes. **Fuente:** Elaboración propia

Utilizando dos factores comunes para nuestro modelo se conseguía una varianza explicada total del 40% y utilizando tres factores se conseguía una del 51%. Mientras que utilizando cuatro factores se alcanza un 60% de varianza explicada por el modelo (figura 5). Por lo tanto, dado que las tres opciones se pueden aceptar como cantidad de factores comunes para nuestro modelo, nos quedaremos con 4 factores comunes, que además de ser válidos permiten representar una mayor proporción de varianza explicada para nuestro modelo de 10 activos tecnológicos.

Si nos fijamos en los resultados obtenidos del análisis factorial sobre nuestra base de datos con 4 factores comunes (figura 5) se observa que el primer factor es el que mayor varianza explicada acumula. También se observa en las cargas (loadings) de la tabla que este primer factor tiene incidencia sobre casi todas las empresas, en especial sobre Microsoft (MSFT), Google (GOOGL) y Apple (AAPL). El segundo factor también acumula una gran parte de varianza explicada, teniendo sus mayores cargas en empresas como Sony (SNE), Canon (CAJ) y Nintendo (NTDOY), seguido del cuarto factor común que tiene carga sobre todo en Nokia (NOK). Sin embargo, en el factor común tres se observa en las cargas que toda su incidencia recae en exclusiva sobre la empresa Qualcomm (QCOM), a pesar de que esta empresa también tiene incidencia del primer factor. Si hacemos una comparación a primera vista con el resultado obtenido del PCA vemos una similitud muy clara, ya que los dos primeros factores del FA afectan en mayor medida a las mismas empresas que tienen una mayor afectación de las dos componentes principales que hemos extraído con el PCA.

4.3 Análisis de componentes independientes (ICA)

El análisis de componentes independientes (ICA) es otro método efectivo para la extracción de señales y reducción de dimensión de los datos. El ICA puede utilizarse como una extensión del análisis de componentes principales (PCA) o del análisis factorial (FA).

Tanto el PCA, el FA como el ICA se centran en encontrar una representación fidedigna de los datos a partir de la proyección de las observaciones en un espacio de menor dimensión. El objetivo principal del PCA es proyectar los datos para maximizar la varianza explicada del modelo y extraer componentes que no estén correlacionados entre sí, con el fin de reducir la dimensión de los datos.

Por otro lado, en el método ICA se buscan componentes que sean estadísticamente independientes y además en la estimación del ICA se utilizan estadísticos de orden superior a los del PCA. Adicionalmente, mientras que en el PCA los datos analizados pueden seguir o no una distribución Normal, en el método ICA se requiere que nuestros datos sean estrictamente no-Gaussianos, es decir, que no sigan una distribución Normal. Por lo que el ICA se puede considerar una técnica muy potente para encontrar factores o componentes ocultos en nuestra base de datos cuando los métodos clásicos PCA y FA fallan.

De igual manera que para aplicar los métodos clásicos es necesario realizar un pre-procesado de los datos con la finalidad de que nuestra base de datos sea estacionaria en media y en varianza. Por este motivo, para el análisis ICA también trabajaremos con el logaritmo de los rendimientos de las 10 empresas tecnológicas estudiadas, es decir, con su rentabilidad relativa continua como hemos visto en el apartado 4.2.1.

4.3.1 Validación del ICA

Para motivar el uso del análisis de componentes independientes sobre nuestra base de datos se van a aplicar dos test de normalidad multivariante, el Test de Mardia y el Test de Henze-Zirkler, que nos permitirán ver si nuestra base de datos se distribuye según una normal multivariante (si se acepta la hipótesis nula) o no (si se rechaza la hipótesis nula).

Como ya hemos visto en el apartado del PCA para aplicar estos test de normalidad multivariante se utiliza el paquete de RStudio MVN. Si comprobamos los resultados obtenidos sobre nuestra base de datos de 10 activos tecnológicos se observa que:

```
res1<-mvn(activos, mvnTest = "mardia");res1$multivariateNormality
           Test         Statistic              p value Result
Mardia Skewness 847.602907686418 1.88451991511825e-74     NO
Mardia Kurtosis 33.5702645318511                    0     NO
            MVN              <NA>                 <NA>     NO
res2<-mvn(activos,mvnTest="hz");res2$multivariateNormality
          Test       HZ p value MVN
Henze-Zirkler 1.323589       0  NO
```

Figura 6: Test de Mardia y Test de Henze-Zirkler. **Fuente:** Elaboración propia

Aplicando ambos test sobre nuestra base de datos se rechaza la hipótesis nula de que los datos siguen una normal multivariante (p-value<0,05). Por lo tanto, nuestros datos no se distribuyen según una normal multivariante y se distribuyen de una forma no-Gaussiana, por lo que es muy razonable aplicar el análisis de componentes independientes (ICA) con el fin de extraer características interesantes de nuestros datos.

Que el ICA sea un método muy adecuado para aplicar como técnica de reducción de dimensión sobre nuestros datos se debe a que, como hemos definido en el apartado anterior, el método ICA tiene como objetivo maximizar la no-Gaussianidad de nuestros datos para obtener componentes que sean independientes, por lo que requiere que nuestra base de datos se distribuya de forma no-Gaussiana (como mucho solo puede haber una señal Gaussiana).

Por lo que el uso del ICA como método para reducir la dimensión de nuestra base de datos y extraer componentes o factores de datos relevantes queda totalmente justificado.

4.3.2 Implementación del algoritmo FastICA sobre las series financieras

Para aplicar el análisis de componentes independientes (ICA) sobre nuestra base de datos utilizaremos el algoritmo FastICA que facilita el paquete de RStudio fastICA.

Una vez realizado el pre-procesado de los datos con la finalidad de que nuestra base de datos sea estacionaria en media y en varianza y comprobada la validez del método ICA se puede proceder con la implementación de esta técnica sobre la base de datos de rendimientos.

A partir de la factorización de las matrices definidas en el apartado 3.1 se consigue construir y estimar el método de análisis de componentes independientes ICA para la posterior interpretación de los resultados.

Para mesurar la no Gaussianidad en el algoritmo FastICA se utiliza la negentropía definida en los principios para la estimación del ICA del apartado 3.3 de este trabajo a partir de la ecuación (8), definida por Cover and Thomas (2001).

Para aplicar la función FastICA en Rstudio se deben tener en cuenta dos parámetros que es necesario establecer en el código y que selecciona el usuario en función de lo que quiera estudiar.

El primer parámetro es el número de componentes que se extraerán mediante el método ICA. Cabe destacar que no existe un método objetivo para escoger el número de componentes a estimar para el análisis ICA, por lo que dependerá del problema a analizar y del juicio de la persona que realice el análisis. Si explican suficiente varianza y el objetivo es visualizar fácilmente los datos lo más frecuente es no utilizar más de 2 o 3 componentes para facilitar la representación gráfica y su interpretación. En nuestro caso, con nuestra base de datos de 10 activos tecnológicos utilizaremos dos componentes para el análisis ICA, dado que mediante el PCA hemos concluido que este número de componentes era suficiente para explicar gran parte de la varianza del modelo.

El segundo parámetro que tenemos que introducir en el código de la función FastICA es el tipo de algoritmo que utilizamos para la ortogonalización[6]. A partir de la ortogonalización se consigue que el algoritmo se ejecute correctamente varias veces con el fin de estimar más de una componente independiente a partir de los vectores iniciales.

Existen dos métodos posibles que permiten implementar la ortogonalización en el algoritmo FastICA de Rstudio, la ortogonalización deflacionaria y la ortogonalización simétrica o paralela.

Para evitar las limitaciones de la ortogonalización deflacionaria, para el código FastICA aplicado sobre nuestra base de datos utilizaremos el método de ortogonalización simétrica pero también se realizará una comparación con los resultados obtenidos con la ortogonalización deflacionaria.

Si aplicamos el paquete FastICA a través de la herramienta de RStudio utilizando dos componentes para nuestra estimación y el algoritmo de ortogonalización simétrico

[6] Ortogonalización: Es el algoritmo empleado para crear un conjunto ortonormal de vectores en el mismo subespacio vectorial a partir de un conjunto de vectores inicial (Técnica de Gram-Schmidt).

obtenemos que el RStudio realiza una estimación del análisis ICA sobre nuestra base de datos. En primer lugar, se obtiene la matriz de datos pre-procesados (matriz X) aplicándoles las técnicas de centrado y blanqueo a los rendimientos relativos de los 10 activos que teníamos inicialmente a partir de la matriz que proyecta los datos en la dirección de la máxima varianza, es decir, la matriz utilizada en la estimación del método PCA (matriz K).

Seguidamente, la herramienta estima la matriz W de desmezcla para maximizar la aproximación mediante la negentropía para que los componentes estimados sean incorrelacionados. De esta forma, esta aplicación nos permite obtener la matriz de señales estimada (la matriz S) que representa las dos componentes independientes que hemos estimado.

Por lo tanto, a partir de esta herramienta que ofrece RStudio obtenemos todas las matrices que necesitamos para el análisis de componentes independientes, lo único que tenemos que decidir es el número de componentes que utilizamos y el algoritmo de ortogonalización que veamos más convenientes. Si observamos el gráfico 5 se muestra un conjunto de tres gráficos donde en el primero vemos una representación de los rendimientos relativos de los activos sin pre-procesado, en el segundo se observan las dos componentes principales obtenidas a partir de la técnica PCA y en el tercero se muestran las dos componentes independientes obtenidas a partir de los resultados del análisis ICA sobre nuestra base de datos.

Cabe destacar que en el gráfico 5, en los gráficos que representan los componentes del PCA e ICA, en el eje horizontal se representa la primera componente principal y en el eje vertical la segunda componte. Hay que tener en cuenta que las componentes independientes generadas por el método ICA no están ordenadas como sí que lo están las componentes principales del método PCA en función de la máxima varianza que recogen, por lo que las dos primeras componentes del PCA y el ICA no son directamente comparables.

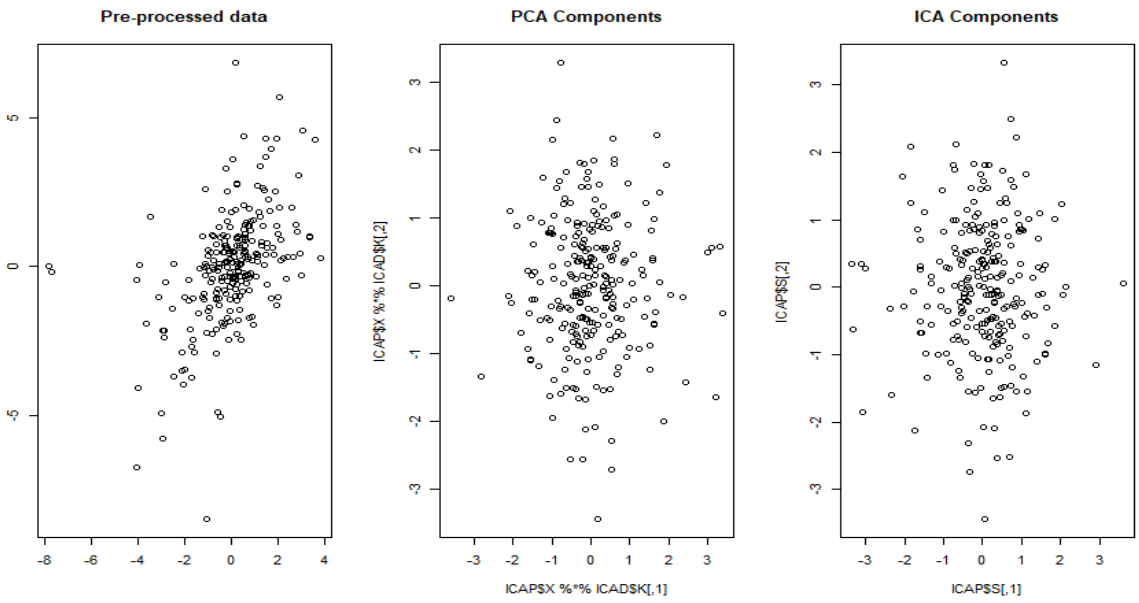


Gráfico 5: Representación de los resultados obtenidos del análisis ICA. **Fuente:** Elaboración propia

Si comparamos ambos gráficos del PCA y el ICA con los datos pre-procesados de dos de nuestros activos vemos que se mueven aproximadamente de la misma forma, teniendo en cuenta que el eje de los datos pre-procesados es distinto ya que existen rendimientos muy atípicos a lo largo de un año que el PCA y el ICA no recogen para su estimación de componentes. También hay que destacar que el gráfico de los componentes del PCA que se observa en el gráfico 5 se ha representado a partir de la multiplicación matricial de la matriz de datos pre-procesados (la matriz X) por la matriz obtenida antes del blanqueo que proyecta la dirección de los componentes principales (la matriz K), por lo que los resultados obtenidos en el gráfico se asemejan a los obtenidos utilizando el análisis de componentes principales (PCA) realizado en el apartado 4.2.2, pero no tienen por qué ser iguales.

Si hacemos una representación gráfica de los dos componentes independientes generados en nuestro análisis de componentes independientes a lo largo del año del año (250 días financieros) su distribución es la que podemos observar en el gráfico 6.

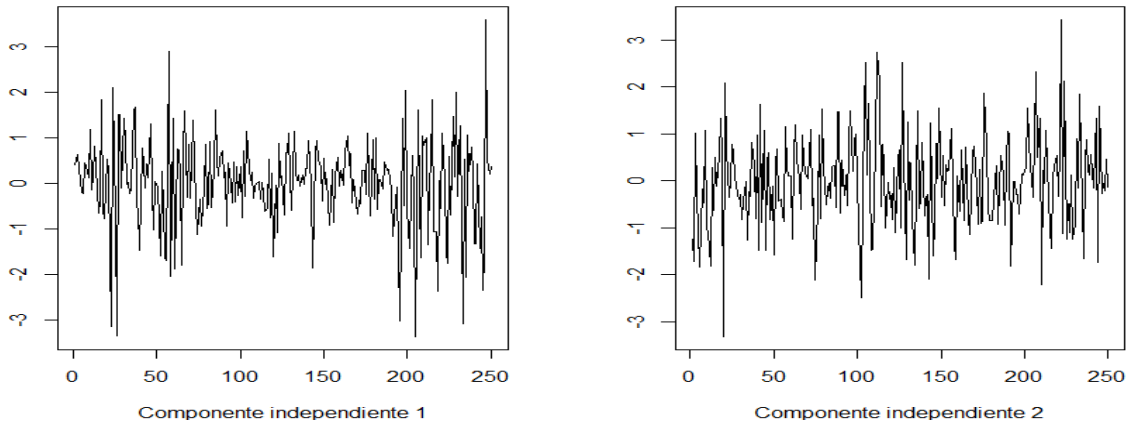


Gráfico 6: Componentes independientes extraídas del análisis ICA. **Fuente:** Elaboración propia

A modo comparativo en el gráfico 7 se observa cómo se comportan dos de las componentes independientes del ICA, las componentes principales del PCA y dos activos de la matriz de datos pre-procesados (matriz X) como referencia de nuestros activos (IBM y Qualcomm). Los ejes horizontales representan un año en días financieros y el eje vertical representa el rendimiento sintético de los factores latentes de riesgo.

En primer lugar, cabe destacar que el rango del eje vertical de rendimientos de la matriz X de datos pre-procesados es mucho mayor debido a que con el PCA y el ICA no se recogen los *outliers*[7] o datos residuales que son extremos de nuestra serie de datos, dado que nuestro objetivo es reducir la dimensión de nuestra base de datos.

Si nos fijamos en los gráficos que representan las componentes del ICA y del PCA se observa que los movimientos más abruptos en el componente principal del PCA también suponen un movimiento abrupto del componente independiente del ICA, pero de signo contrario. Esto ocurre tanto en la primera como la segunda componente principal y puede suceder debido a que el método ICA no da información alguna sobre la varianza de la señal original ni del signo de la misma. Sin embargo, esto no supone un problema dado

[7] Outliers: datos atípicos o poco frecuentes en un determinado estudio.

que mediante el pre-procesado de datos lo que hacemos es estandarizar las variables de modo que su varianza sea unitaria.

El hecho de realizar el proceso que estandariza las variables estudiadas para que tengan varianza unitaria provoca que, a la hora de determinar la varianza explicada del modelo, se reparta toda la varianza explicada proporcionalmente entre el número de componentes que hemos determinado para aplicar el algoritmo FastICA.

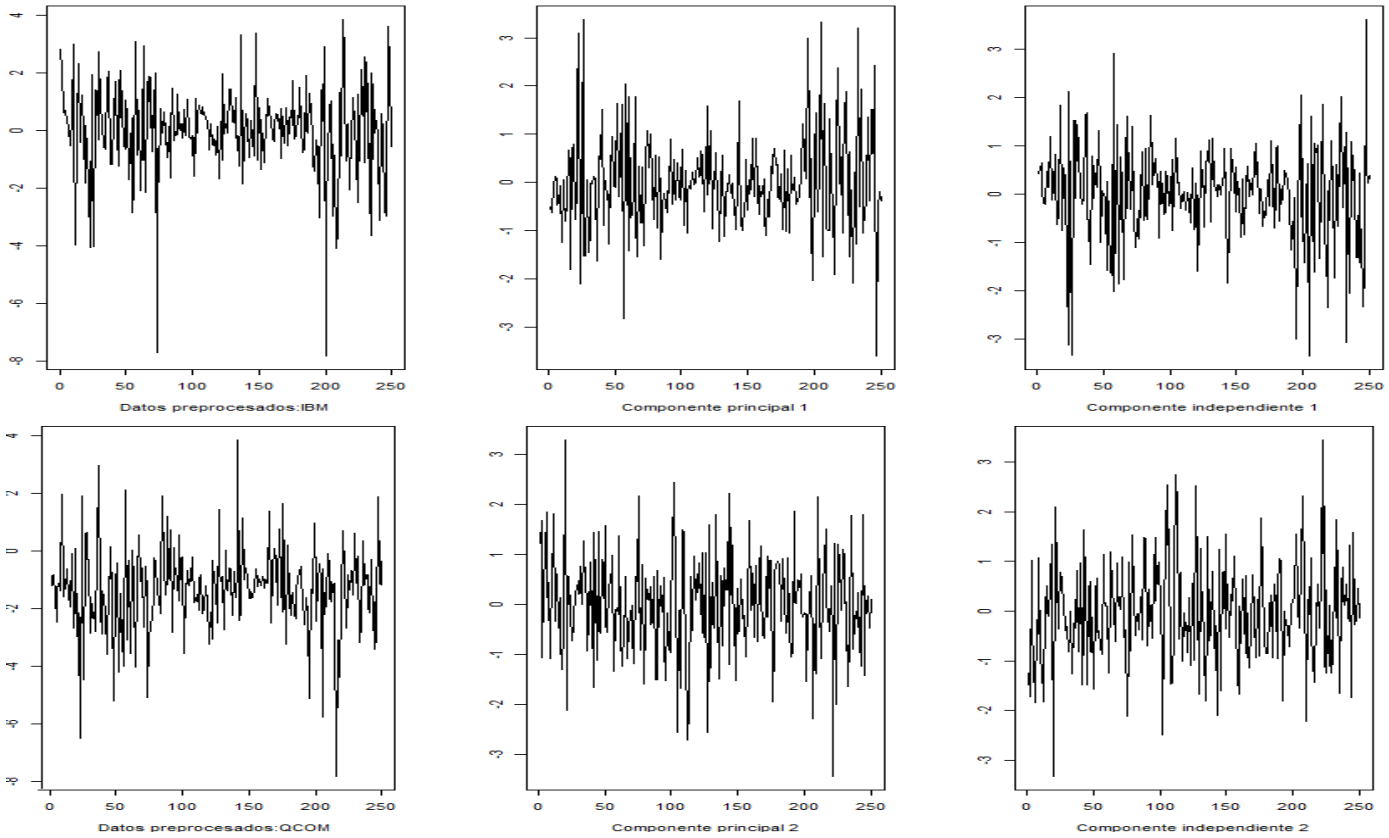


Gráfico 7: Comparativa componentes del PCA e ICA. **Fuente:** Elaboración propia

A nivel comparativo podemos aplicar el mismo algoritmo FastICA sobre nuestra base de datos, pero en este caso incorporando tres componentes para el estudio, dado que en el análisis PCA recomendábamos escoger dos componentes pero tres componentes también podría ser una opción válida.

Si nos fijamos en el gráfico 8, vemos como al incorporar una componente más a nuestro análisis los datos se dispersan un poco más y se aproximan más al cuadrante de la primera componente (parte derecha del eje horizontal). Por lo que la primera componente toma un mayor peso si incorporamos una componente más al estudio. La incorporación de tres componentes a nuestro modelo también provoca que la segunda componente tenga un menor peso sobre nuestro modelo, ya que muchos de los puntos que se encontraban en la parte superior de cuadro (por encima del 0 en el eje vertical) se han desplazado hacia el centro del cuadro (punto 0,0 en el gráfico).

Aun así, a nivel gráfico utilizando tres componentes no sería del todo correcta esta representación dado que no podemos localizar la tercera componente utilizando un gráfico de esta escala.

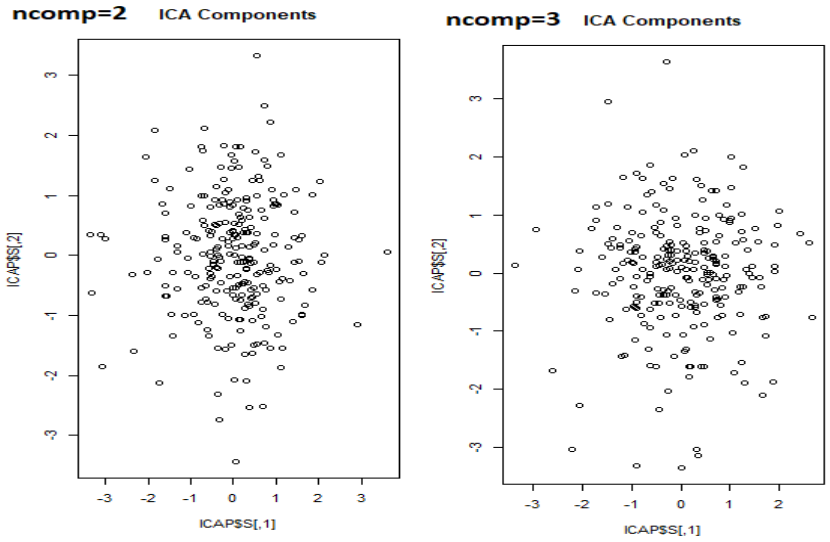


Gráfico 8: Comparativa entre el gráfico ICA seleccionando 2 o 3 componentes. **Fuente:** Elaboración propia

A continuación, se compararán los resultados obtenidos en nuestra base de datos utilizando los distintos algoritmos de ortogonalización para nuestro modelo. En el gráfico 9, los gráficos de la primera fila representan el primer componente independiente estimado a partir del algoritmo de ortogonalización paralelo en el primer caso y en el segundo caso mediante el algoritmo de deflación. La segunda fila se interpreta de la misma forma, pero sobre la segunda componente independiente de nuestro modelo.

La diferencia principal entre estas proyecciones radica en que en el algoritmo deflacionario acumula el error de estimación de los primeros vectores o componentes para los siguientes, ya que se estiman las componentes independientes una a una y cada vez que se ejecuta el algoritmo en cada iteración para estimar las otras componentes se eliminan las proyecciones de los componentes anteriores. Mientras que en la ortogonalización simétrica los vectores se estiman de forma paralela y no acumula los errores de estimación de proyecciones anteriores dado que se componen a la vez.

A nivel matricial se ha realizado un cálculo en RStudio que confirma que la matriz de datos pre-procesados (matriz X) blanqueada y centrada, y la matriz K obtenida previa al blanqueo de los datos son exactamente iguales, por lo que ambos algoritmos de ortogonalización parten de la misma base y la única diferencia radica en si se estiman los componentes a la vez (forma simétrica) o uno por uno (forma deflacionaria).

A nivel gráfico la interpretación es mucho más compleja, pero se observa que para la primera componente el algoritmo deflacionario presenta datos mucho más abruptos al principio y al final de la serie de un año (gráfico 9), mientras que con el algoritmo simétrico o paralelo los datos se mueven más o menos dentro de un mismo rango del eje vertical, sin excesivas caídas o subidas. El hecho de que la primera componente independiente presente esta estructura un poco más estable dentro del eje vertical con el algoritmo simétrico viene motivada principalmente porque en este caso los vectores se estiman todos a la vez de forma paralela, sin acumular errores de iteraciones anteriores.

Además, también debemos destacar el hecho de que si solo se estiman dos componentes independientes el algoritmo no tiene que realizar excesivas iteraciones para estimar el

resto de las componentes independientes, y por lo tanto no se eliminan proyecciones de muchos componentes previamente estimados.

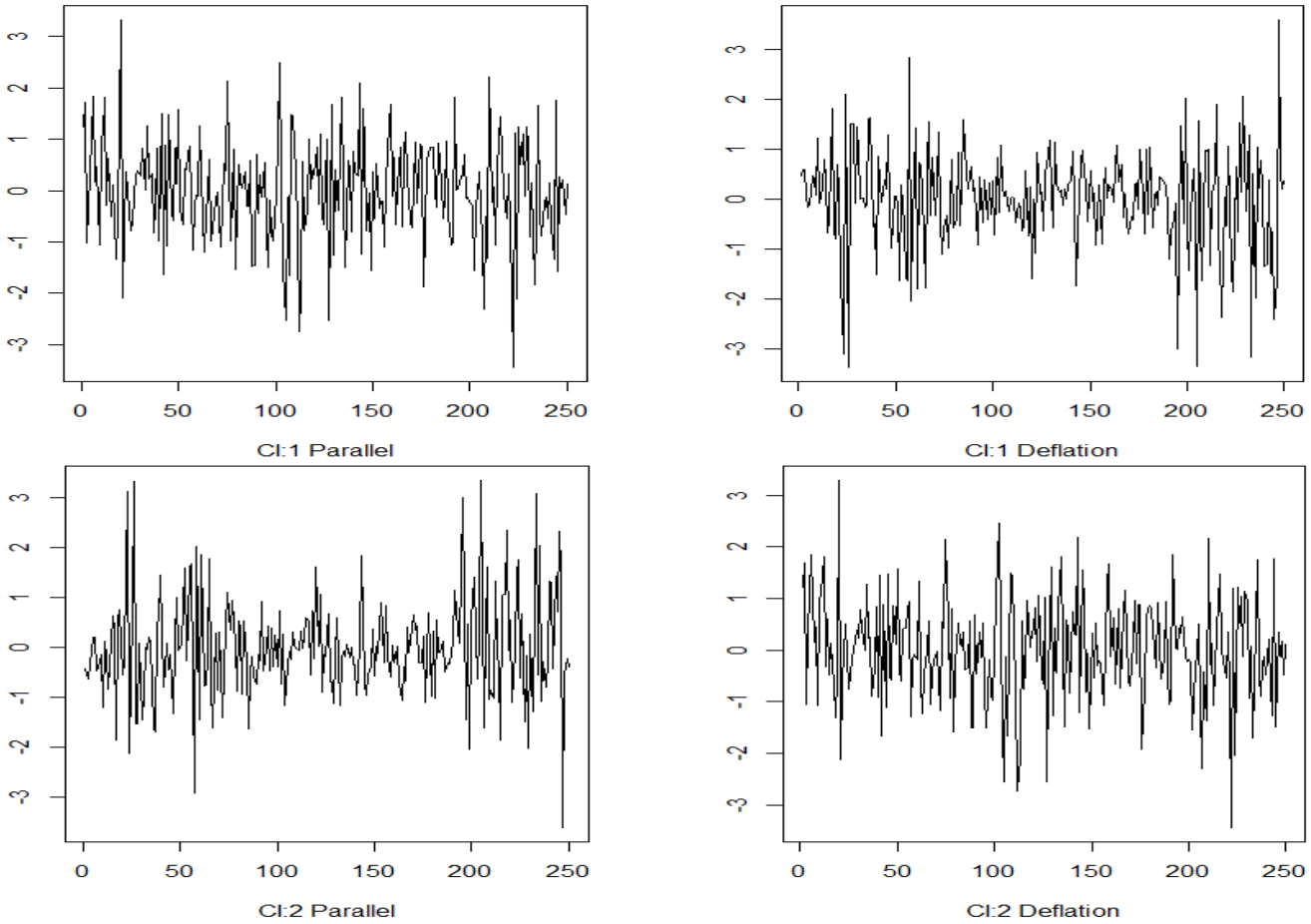


Gráfico 9: Comparativa componentes independientes en función del tipo de ortogonalización. **Fuente:** Elaboración propia

4.4 Modelo alternativo: Factorización no negativa de matrices (NMF)

El método de factorización no negativa de matrices (NMF) es una herramienta utilizada para el análisis de datos multivariantes. Este método guarda alguna relación con otras técnicas clásicas de análisis multivariante como el análisis de componentes principales (PCA) o el análisis de componentes independientes (ICA), ya que en el fondo todos estos métodos realizan una factorización de matrices. Pero en este apartado se comprobará si lo resultados finales obtenidos del análisis NMF permiten la reducción de la dimensión de los datos y si es comparable con el resto de los métodos aplicados.

En este apartado utilizaremos el método NMF para identificar patrones y estimar factores comunes sobre nuestra base de datos de 10 activos tecnológicos con el fin de lograr una agrupación y representación adecuada de los datos.

Como su nombre indica, la diferencia principal del NMF frente a los métodos clásicos es que impone la no negatividad de las matrices, por lo que la matriz de rendimientos[8] relativos que utilizamos para el análisis de nuestra base de datos tendrá que ser modificada, dado que presenta una gran cantidad de rendimientos negativos. En nuestro caso, utilizaremos la matriz de rendimientos de los activos en valor absoluto como veremos más adelante.

Para aplicar el método NMF sobre nuestra base de datos se ha utilizado el paquete de RStudio NMF, que facilita la implementación de los algoritmos necesarios para realizar el análisis, visualizar datos a partir de gráficos y comparar rendimientos.

Como se ha definido en el apartado 3.5 de este trabajo, el objetivo principal del modelo NMF es que la función de coste *f* cuantifique la calidad de la factorización encontrando el mínimo local entre las matrices *B* y *C*, la optimización debe ser la siguiente:

$$\{\hat{B}, \hat{C}\} = \arg\min f(A, BC) \qquad sujeto\ a \begin{cases} B \geq 0 \\ C \geq 0 \end{cases} \qquad (11)$$

Donde las matrices *B* y *C* deben ser estrictamente positivas (no negativas).

Para aplicar el paquete NMF de RStudio a nuestra base de datos debemos tener en cuenta un paso previo en el que establecemos una semilla antes de iniciar el algoritmo a partir de la cual se inicia el proceso de iteración. Este paso previo se debe a que no hay un algoritmo para encontrar el mínimo global, por lo que es necesario escoger una semilla de inicio para que los resultados sean significativos. Existen varios métodos para computar un punto de partida de la semilla razonable. Para nuestra base de datos utilizaremos el método de semilla del ica, que usa los resultados de un análisis ICA (del paquete de RStudio fastICA) con solo la parte positiva de los resultados para inicializar los factores.

Una vez tenemos la semilla fijada para nuestro estudio debemos seleccionar alguno de los 11 algoritmos disponibles que ofrece el paquete NMF en RStudio. En nuestro estudio utilizaremos distintos algoritmos a modo comparativo.

Para ejecutar el comando que realiza el análisis NMF también es necesario establecer varios parámetros en el código. En primer lugar, se debe determinar la matriz que se estudia que debe ser no negativa, en nuestro caso utilizaremos la matriz de rendimientos relativos de los 10 activos tecnológicos en valor absoluto, de esta forma la matriz será completamente positiva y podremos aplicar el análisis NMF y los datos se representarán como la acumulación total de elementos. Otra opción válida sería utilizar únicamente los rendimientos que fueran positivos para el análisis, pero resulta mucho menos realista en nuestro caso, ya que existe una gran cantidad de valores negativos en los rendimientos de nuestros activos y se desecharían demasiados datos para el análisis.

En segundo lugar, para completar el comando NMF es necesario establecer el rango de factorización. Determinar este rango es bastante complicado, ya que depende en gran parte del criterio del analista. Aun así, existen varias maneras de determinar el rango de factorización óptimo a partir del estudio de diversos gráficos (gráfico 10). Estos gráficos representan mesuras de calidad de los resultados que permiten escoger el mejor valor para el rango de acuerdo con este criterio. A continuación, veremos las técnicas que han

[8] Pese a ser más adecuado utilizar precios y no rentabilidades para las estimaciones del método NMF dada la restricción de utilizar valores estrictamente no negativos, en el estudio de este trabajo se ha realizado sobre los rendimientos para mantener la homogeneidad con el análisis realizado con el resto de métodos clásicos (PCA y FA) y el ICA.

propuesto diversos autores para interpretar los gráficos y escoger cual es el rango óptimo para nuestros datos:

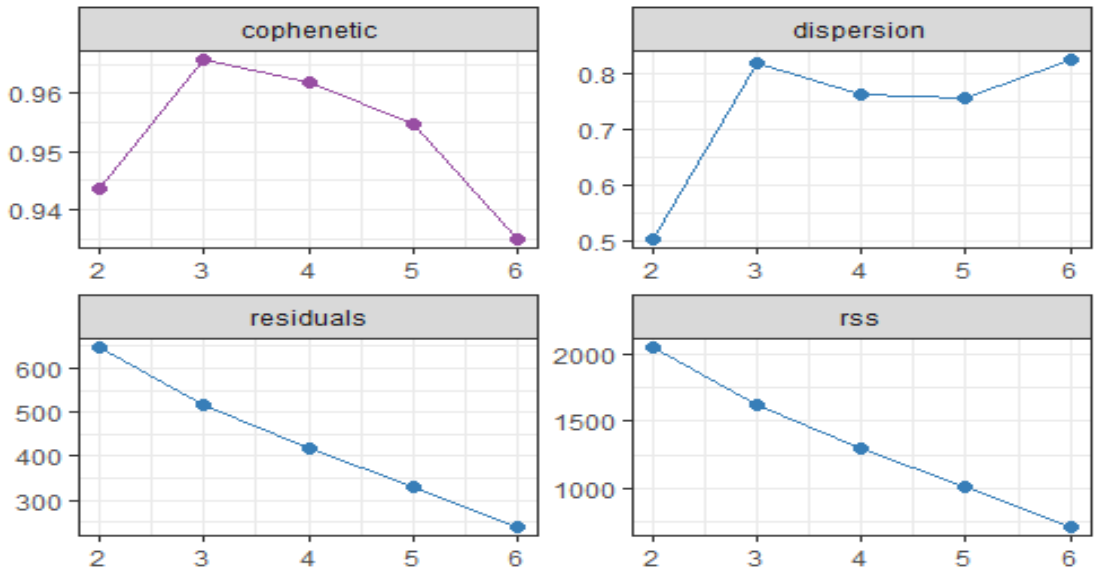


Gráfico 10: Resultados del análisis gráfico del rango de factorización. **Fuente:** Elaboración propia

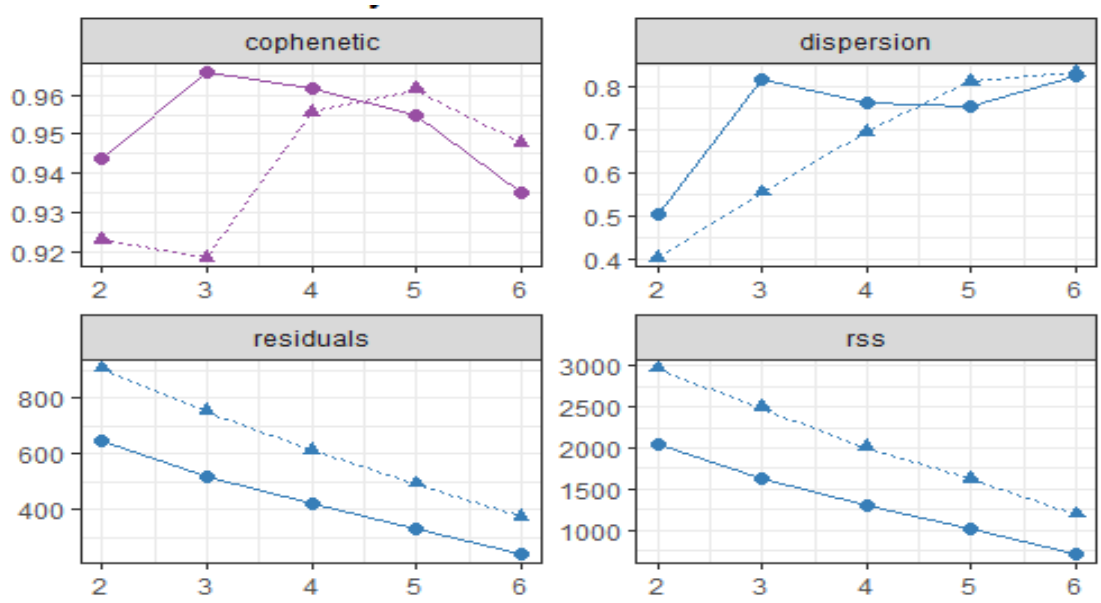


Gráfico 11: Resultados del análisis gráfico del rango de factorización comparando con una matriz de datos generada aleatoriamente (línea discontinua). **Fuente:** Elaboración propia

- La primera técnica definida por Brunet (2004) se basa en: "seleccionar el primer valor de rango óptimo en el que el gráfico del coeficiente de cophenetic comienza a decrecer". En el caso de nuestra base de datos sería en el punto del rango 3.

- La segunda técnica propuesta por Hutchins (2008) consiste en: "escoger el primer valor donde el gráfico de la curva RSS presenta un punto de inflexión". En el

estudio sobre nuestra matriz de datos el punto de inflexión es muy difícil de observar, pero en cualquier caso donde más cambia la tendencia es en el rango 3.

- Finalmente, la última técnica definida por Frigyesi (2008) considera que: "el rango viene determinado por el valor más pequeño al cual el decrecimiento de la curva del RSS es menor que el decrecimiento de la curva RSS obtenida de una matriz de datos aleatoria". En nuestra base de datos (gráfico 11) se observa como el punto en el que la curva RSS original de nuestros datos (línea continua) decrece menos que la curva RSS generada aleatoriamente (línea discontinua) es en el rango 3.

Por lo tanto, las tres formas anteriores de determinar el rango óptimo nos llevan a concluir que el rango de nuestra base de datos de 10 activos tecnológicos es tres. Cabe destacar que estas estimaciones se han hecho a partir de 10 iteraciones de cada valor del rango para obtener una estimación robusta como indican Brunet (2004) y Hutchins (2008).

Una vez tenemos nuestra matriz de datos en valor absoluto (no negativa) y sabemos el rango de factorización solo nos queda determinar el algoritmo o los algoritmos que vamos a emplear para estimar el método NMF. Para comparar los resultados de diferentes algoritmos de una forma equitativa y justa es necesario establecer una semilla, en nuestro caso fijaremos la semilla aleatoria 123456.

Si utilizamos los siguientes algoritmos (Brunet, lee, nsNMF) para determinar el NMF podremos determinar cuál es el método más adecuado para aplicar a nuestra base matriz de datos (figura 6). La forma más adecuada de comparar los diferentes métodos que permiten realizar el análisis NMF es a partir del error de estimación (gráfico 12).

	method	seed	rng	metric	rank	sparseness.basis	sparseness.coef	silhouette.coef	silhouette.basis	residuals
brunet	brunet	random	1	KL	3	0.2459191	0.6773261	0.8985676	0.7075719	516.3413
lee	lee	random	1	euclidean	3	0.2714996	0.6311133	0.8581165	0.6714745	776.4381
nsNMF	nsNMF	random	1	KL	3	0.2979072	0.9170490	0.9721876	0.6986556	584.0659

	niter	cpu	cpu.all	nrun
brunet	530	NA	NA	1
lee	510	NA	NA	1
nsNMF	560	NA	NA	1

Figura 7: Resultados del análisis NMF con 3 algoritmos distintos (Brunet, lee, nsNMF). **Fuente:** Elaboración propia

Dado que el algoritmo NMF se crea a partir del tracking error, se puede realizar un gráfico de los errores (gráfico 12). Cada error está normalizado por lo que el primer valor equivale a 1 y la iteración se para cuando el criterio de convergencia queda satisfecho.

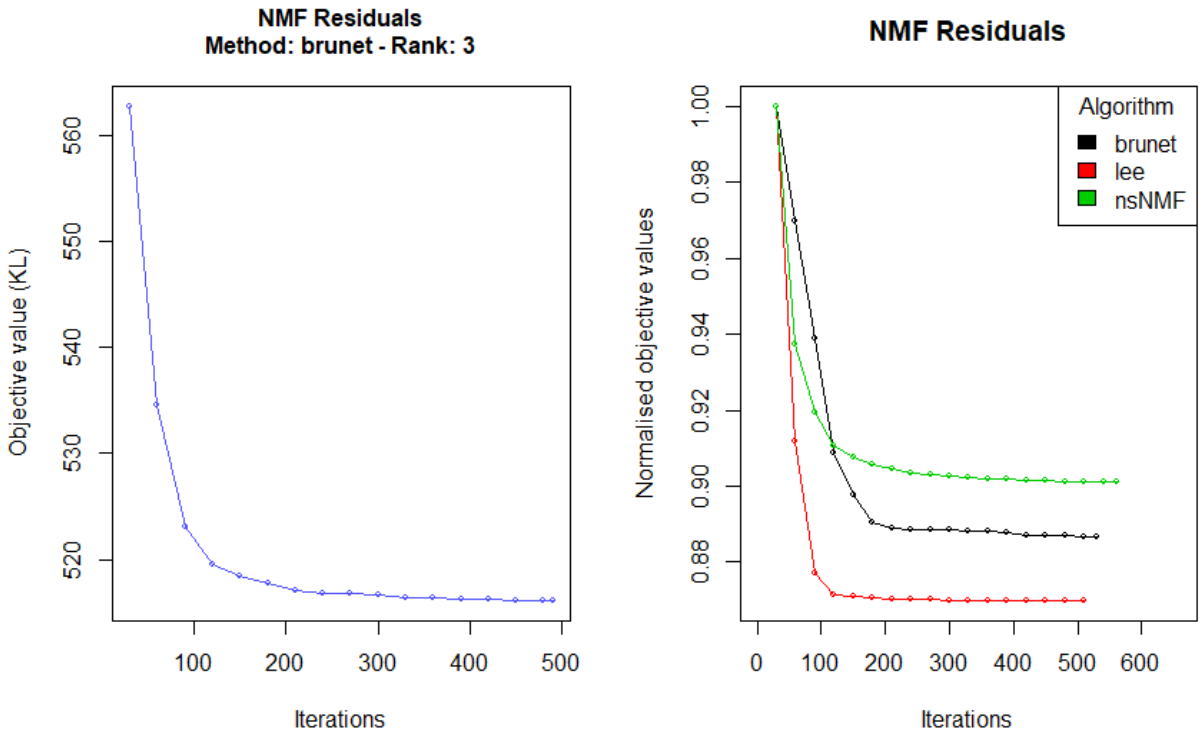


Gráfico 12: Error para un solo método del NMF (izquierda) y para múltiples métodos (derecha). **Fuente:** Elaboración propia

En el gráfico 12 se observa la evolución del error para un solo método en la izquierda y para múltiples métodos en el gráfico de la derecha en función de las iteraciones realizadas. Lo primero que observamos en el gráfico de la derecha es que tanto el método brunet como el nsNMF (línea negra y verde respectivamente) convergen mucho más despacio que el método lee, dado que éstos convergen a las 150 iteraciones aproximadamente, mientras que el método lee a las 100 iteraciones. En cualquier caso, el método lee (línea roja) tiene el error euclídeo más bajo de los tres casos por lo que es el método que escogeremos para realizar nuestro análisis NMF.

El siguiente gráfico crea una jerarquía de agrupaciones en nuestra base de datos a partir de un algoritmo que queda representado en forma de árbol (gráfico 13). Este algoritmo depende directamente del rango de factores que hemos decidido anteriormente (3 factores) y del método de análisis NMF que hemos empleado. A partir de este gráfico se obtiene una estructura de árbol donde las ramas más bajas hacen referencia a nuestros activos tecnológicos estudiados individualmente y a mesura que nos acercamos a la parte más alta del gráfico de árbol vemos distintas agrupaciones de todos los activos. Este gráfico que agrupa todos los activos estudiados recibe el nombre de dendrograma y ofrece una forma sencilla de interpretar los resultados en caso de utilizar múltiples métodos (en el gráfico vienen representados por las bases). En nuestro ejemplo hemos utilizado los métodos que vienen por defecto.

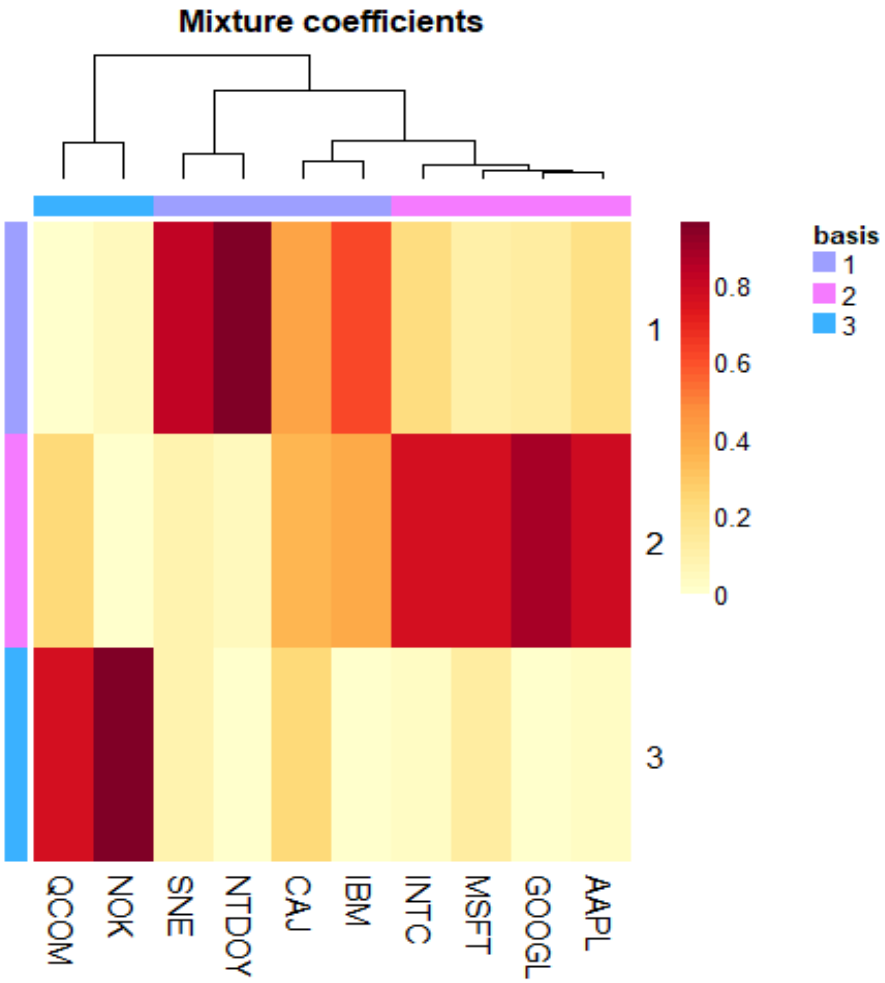


Gráfico 13: Dendrograma de nuestra base de datos de 10 activos tecnológicos. **Fuente:** Elaboración propia

El dendrograma anterior muestra los grupos que se forman al crear conglomerados de observaciones en cada iteración y cuál es el nivel de similitud. Los grupos creados se muestran en la parte superior del gráfico a partir de la forma de árbol y el nivel de similitud de los grupos se puede determinar a partir del mapa de calor que vemos en la parte inferior del gráfico, indicando los colores más rojizos una mayor similitud y los colores más claros y amarillos una menor similitud.

El dendrograma obtenido a partir de nuestra base de datos obtuvo una partición final de 4 conglomerados (4 últimas ramas del árbol). El primer conglomerado, situado en el extremo izquierdo del gráfico, se compone de las empresas Qualcomm (QCOM) y Nokia (NOK). El segundo conglomerado, inmediatamente a la derecha, se compone de las empresas Sony (SNE) y Nintendo (NTDOY). Más a la derecha, el tercer grupo se compone de las empresas Canon (CAJ) e IBM. El cuarto conglomerado, situado en el extremo derecho del dendrograma, se compone de Intel (INTC), Microsoft (MSFT), Google (GOOGL) y Apple (AAPL).

A medida que nos fijamos en la parte superior del dendrograma vemos que existen menos conglomerados finales, aunque el nivel de similitud entre ellos también es menor, dado que cuanto más abajo se corte el dendrograma mayor similitud habrá entre las

observaciones a coste de que se recojan más conglomerados finales y no se consiga reducir tanto la dimensión. Si tenemos en cuenta los conglomerados superiores en nuestra base de datos podemos agrupar tres conjuntos distintos. El primer conjunto, localizado en el extremo izquierdo, sería el mismo formado por las empresas Qualcomm (QCOM) y Nokia (NOK). El segundo conglomerado, inmediatamente a la derecha, también sería el mismo y se compondría de las empresas Sony (SNE) y Nintendo (NTDOY). Finalmente, el tercer grupo, situado en el extremo derecho, agruparía el tercer y cuarto grupo anteriormente mencionados formado por las empresas Canon (CAJ), IBM, Intel (INTC), Microsoft (MSFT), Google (GOOGL) y Apple (AAPL).

Por último, si quisiéramos agrupar nuestra base de datos en los dos conglomerados más elevados del árbol el primer grupo sería el mismo formado por las empresas Qualcomm (QCOM) y Nokia (NOK), mientras que el segundo grupo comprendería al resto de empresas de nuestro estudio. Lo cual nos indica que, en nuestro estudio, las empresas que tienen una menor similitud con el resto son Qualcomm (QCOM) y Nokia (NOK) aunque entre ellas sí que presentan una gran similitud, debido a que en el mapa de calor tienen un color rojo muy oscuro. Esto puede ser debido a que son dos empresas destinadas principalmente a la telefonía móvil (Qualcomm realiza procesadores para los teléfonos móviles y Nokia se dedica principalmente a realizar y comercializar dispositivos completos).

El otro gran grupo con un elevado nivel de similitud es el formado por las empresas Intel (INTC), Microsoft (MSFT), Google (GOOGL) y Apple (AAPL), dado que además de que forman todas juntas un conglomerado del dendrograma también tienen un elevado nivel de similitud entre ellas (el mapa de calor es muy rojizo). Esta similitud es debida a que las cuatro empresas se dedican al negocio de la tecnología y desarrollo informático y prácticamente compiten y trabajan en el mismo mercado, dado que Intel (INTC) se dedica a fabricar procesadores para otras compañías como Microsoft (MSFT), mientras que Google (GOOGL) fabrica sistemas operativos y multitud de aplicaciones para los ordenadores y teléfonos móviles, y Apple (AAPL) opera comercializando ordenadores y dispositivos móviles propios que compiten directamente con el resto de empresas estudiadas.

Finalmente, si analizamos el resto de las empresas, IBM y Canon (CAJ) guardan una mayor similitud con el grupo anteriormente mencionado que el resto de las empresas, ya que una rama intermedia del dendrograma los une, aunque el mapa de calor indica una similitud intermedia (dado que el color es anaranjado). De igual forma las empresas Sony (SNE) y Nintendo (NTDOY) están conectadas con una rama muy alta del dendrograma de las empresas anteriormente mencionadas (Canon, IBM, Intel, Microsoft, Google y Apple) con las que presenta una similitud intermedia. Las dos empresas Sony (SNE) y Nintendo (NTDOY) sí que presentan una similitud muy alta entre ellas dado al color rojo que muestra el mapa de calor y esto puede ser debido a que operan en mercados muy similares, dado que Nintendo (NTDOY) se dedica fundamentalmente a la comercialización de consolas y videojuegos y Sony también tiene una rama específica para eso, aunque Sony (SNE) opera en muchos más mercados como el de la fotografía por lo que también guarda alguna similitud con Canon (CAJ), o el mercado de la telefonía móvil por lo que también guarda similitudes con Apple (AAPL) o Google (GOOGL).

Por lo tanto, el método de factorización NMF también es un método muy útil de análisis de datos multivariantes, pero en la aplicación realizada en este estudio no se reduce la

dimensión de los datos de nuestras series financieras en nuevos factores o componentes que recojan las características principales de los datos, como sí hacen los métodos clásicos PCA y FA o el método ICA. El método NMF se centra en estructurar y agrupar adecuadamente los datos estudiados en conglomerados o grupos representados en el dendrograma en función de su similitud (gráfico 13), pero no es comparable con los métodos anteriormente estudiados en términos del resultado obtenido.

Aun así, el método NMF[9] también es un método de factorización de matrices igual que los métodos estudiados anteriormente y es interesante incluirlo en este trabajo como alternativa a la metodología empleada por el PCA, FA o el ICA, y por su capacidad agrupadora de datos con características similares teniendo en cuenta su grado de similitud.

4.5 Comparación de los resultados entre los métodos PCA, FA, ICA

A continuación, se establece una comparación de los pesos o cargas (*loadings*) que suponen para nuestra muestra de 10 empresas tecnológicas los cuatro componentes o factores calculados mediante las tres técnicas PCA, FA y ICA.

En la Figura 8 se establece una comparativa entre las cargas de los componentes o factores de los métodos clásicos PCA y FA. El motivo de escoger cuatro componentes principales y no dos para comparar el PCA con el FA, como indicaba el análisis que realizamos sobre nuestra base de datos, es que en el análisis factorial, como hemos comprobado anteriormente en el apartado 4.2.3, hay restricciones en cuanto al número de factores que se deben utilizar. Por lo que para facilitar el análisis comparativo hemos escogido cuatro componentes principales y cuatro factores para el FA.

Cargas con el PCA

Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4
IBM	0.315	0.121		0.206
QCOM	0.313	0.231	0.165	0.477
SNE	0.297	-0.426	-0.268	0.112
MSFT	0.389	0.148	0.181	-0.188
AAPL	0.354		0.209	-0.134
INTC	0.328	0.279		
NOK	0.225	0.251	-0.765	-0.422
NTDOY	0.233	-0.660	0.256	-0.423
CAJ	0.288	-0.365	-0.363	0.521
GOOGL	0.378	0.115	0.180	-0.184

Cargas con el FA

Loadings:

	Factor1	Factor2	Factor3	Factor4
MSFT	0.83			
AAPL	0.60	0.34		
GOOGL	0.75	0.31		
SNE		0.64		
CAJ		0.58		0.34
QCOM	0.33		0.91	
NOK				0.52
IBM	0.46			
INTC	0.50			0.38
NTDOY	0.34	0.49		

Figura 8: Cargas de los componentes del PCA y de los factores del FA. **Fuente:** Elaboración propia

Si nos fijamos en las cargas que se producen en las cuatro componentes del PCA vemos como la primera componente es la que recoge mayores cargas positivas en la mayoría de las empresas analizadas, teniendo un mayor peso la primera componente en las empresas

[9] Se ha introducido la técnica NMF a este trabajo como un paso previo para futuras ampliaciones del estudio de este método y su aplicación y aportación a las finanzas, dada su capacidad para generar características latentes de los datos que no son visibles *a priori*

IBM, Microsoft (MSFT), Apple (AAPL), Intel (INTC) y Google (GOOGL). La segunda componente tendría mayores cargas sobre la empresa Sony (SNE) y Nintendo (NTSOY), la tercera componente es la que tiene una mayor carga sobre la empresa Nokia (NOK) y finalmente, la cuarta componente tiene un mayor peso en las empresas Qualcomm (QCOM) y Canon (CAJ). Si comparamos con el análisis realizado en el apartado 4.2.2, donde solo se han utilizado dos componentes principales para el estudio del PCA, vemos como las empresas afectadas por la primera y segunda componente son las mismas, pero en este caso algunas de las empresas recogidas por esas dos componentes pasan a tener más peso en las componentes tres y cuatro. A nivel visual es muy útil utilizar un gráfico de tipo red neuronal sin la parte oculta para ver de forma representativa cuales son los componentes que tienen más peso sobre cada activo tecnológico estudiado (gráfico 14).

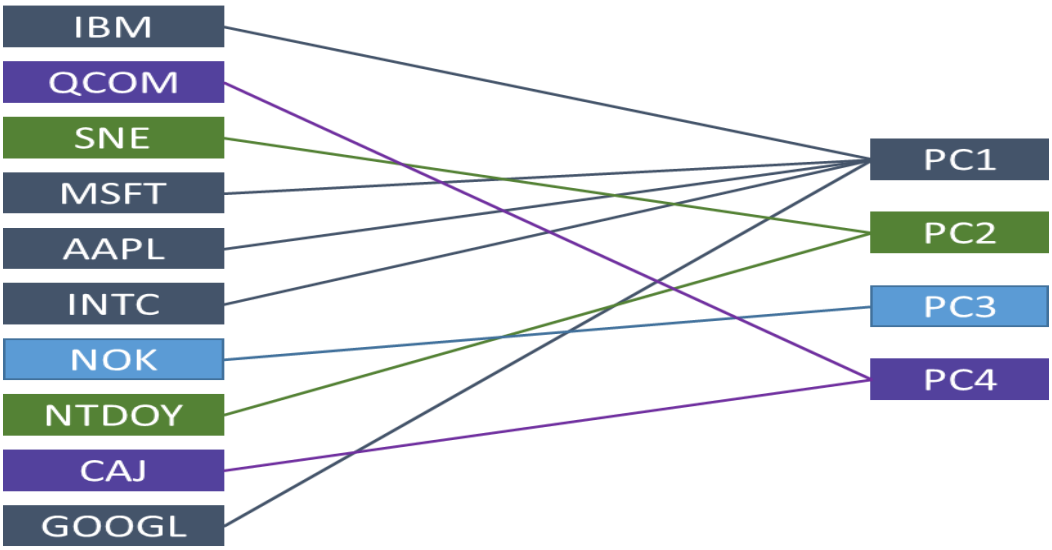


Gráfico 14: Cargas de los componentes principales (PC) en el PCA sobre los 10 activos en forma de red neuronal sin parte oculta. **Fuente:** Elaboración propia

En el caso de las cargas que se producen en el método de análisis FA, de igual forma que hemos analizado en el apartado 4.2.3, se observa que el primer factor es el que tiene incidencia sobre casi todas las empresas, en especial sobre Microsoft (MSFT), Google (GOOGL) y Apple (AAPL), seguido de Intel (INTC) y IBM. El segundo factor tiene sus mayores cargas en empresas como Sony (SNE), Canon (CAJ) y Nintendo (NTDOY), seguido del cuarto factor común que tiene carga de forma exclusiva en Nokia (NOK). Sin embargo, en el factor común tres se observa en las cargas que toda su incidencia recae en exclusiva sobre la empresa Qualcomm (QCOM), a pesar de que esta empresa también tiene incidencia del primer factor.

Si hacemos una comparación con el resultado del PCA vemos una similitud muy clara, ya que los dos primeros factores del FA afectan en mayor medida a las mismas empresas que tienen una mayor afectación de las dos componentes principales que hemos extraído con el PCA. Además, tanto en FA como en PCA, la empresa Nokia queda explicada por una única componente principal (componente 3 en el PCA) o por un único factor (factor 4 en el FA). De igual forma que con el PCA, si utilizamos un gráfico de tipo red neuronal sin la parte oculta se puede ver de forma más clara los factores que tienen más peso sobre cada activo estudiado (gráfico 15).

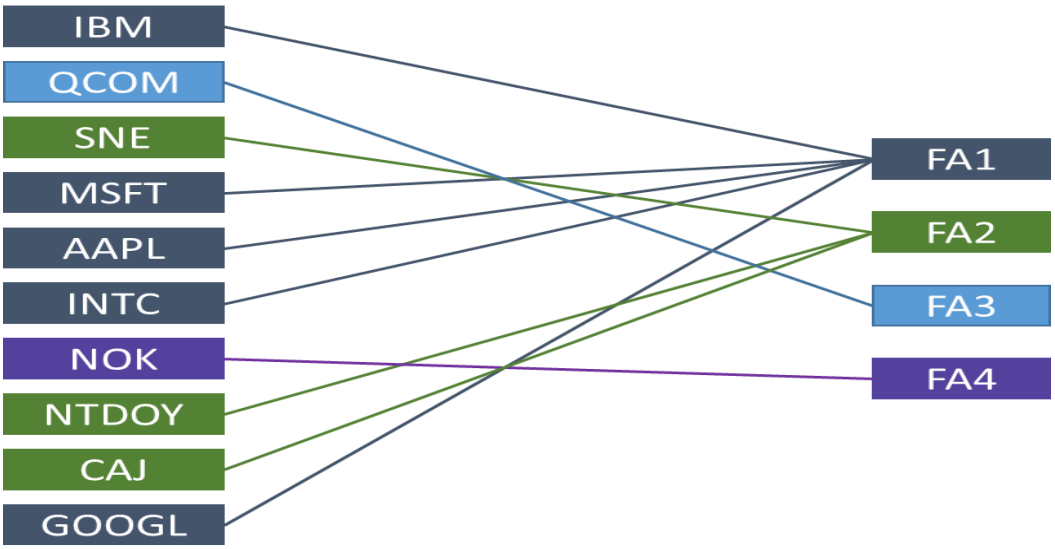


Gráfico 15: Cargas de los factores (FA) en el método de análisis factorial sobre los 10 activos en forma de red neuronal sin parte oculta. **Fuente:** Elaboración propia

Si hacemos referencia a los pesos de los componentes en el método ICA en comparación con los del PCA debemos tener en cuenta que las cargas que se obtienen del ICA estiman componentes que son incorrelacionados y además independientes, mientras que el PCA solo garantiza componentes principales que pueden ser independientes o no. Además, los componentes que se obtienen del ICA no están ordenados como sí que lo están los del PCA en función de la máxima varianza explicada, por lo que para comparar ambos métodos será necesario ordenar los componentes independientes del ICA.
El criterio para ordenar los componentes independientes para el análisis ICA es muy distinto al del PCA, ya que en el ICA la forma de ordenar los componentes independientes es a partir de la correlación existente entre los componentes principales y los componentes independientes (Daniel Peña, 2002). De tal forma que el primer componente independiente es el que está máximamente correlacionado al primer componente principal, el segundo componente independiente tendrá máxima correlación con el segundo componente principal, y así sucesivamente.

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
Comp.1	0.07473116	-0.17325930	-0.25338313	-0.03038910	0.915158366	-0.027617992	0.069814545	-0.11761864	-0.16596129	-0.1214125
Comp.2	-0.31586525	0.32264144	0.13525866	0.70062228	0.022189125	0.055892708	0.316995393	-0.32094345	-0.26133661	-0.1075230
Comp.3	-0.25667566	0.48391737	-0.10569783	-0.02488939	0.104583583	-0.175473487	-0.780549350	-0.11685902	0.01700388	-0.1523777
Comp.4	0.25598671	-0.60218905	0.08158622	0.35142628	-0.137189675	0.144936506	-0.431168381	-0.42811893	0.02715325	-0.1786771

Figura 9: Correlación entre los componentes principales del PCA y los componentes independientes del ICA. **Fuente:** Elaboración propia

Una vez analizada la correlación entre los componentes principales y los componentes independientes en la figura 9, donde las columnas representan las 10 componentes independientes estimadas y las filas representan las cuatro componentes principales que analizamos en este caso, podemos ordenar las componentes independientes. Se observa que el componente independiente que mayor correlación tiene con la primera componente principal es el quinto (V5) por lo que es la primera componente independiente. La

segunda componente independiente sería la cuarta (V4), ya que es la que tiene una mayor correlación con la segunda componente principal del PCA. Utilizando la misma lógica la tercera componente independiente sería la séptima (V7) en la figura 9 y la cuarta componente independiente sería la segunda (V2) en la figura 9. Por lo que el orden de las componentes independientes es V5, V4, V7, V2.

Una vez ordenados los componentes independientes podemos proceder a estimar los pesos o cargas de cada uno de los componentes. Para determinar los pesos en el ICA se utiliza la matriz estimada de desmezcla (matriz W), que podría considerarse una matriz de cargas o *loadings*, ya que hace la misma función que la matriz de pesos en el PCA.

De esta manera, a partir de la conexión establecida con el método PCA hemos conseguido analizar los pesos (figura 10). En este caso, la primera componente independiente tiene un mayor peso sobre las empresas IBM, Nintendo (NTDOY) y Google (GOOGL), por lo que solo coincide en cuanto a mayor peso con la primera componente del PCA en IBM y Google. La segunda componente del ICA tiene un mayor efecto sobre Qualcomm (QCOM), mientras que la tercera componente tiene un mayor impacto sobre Microsoft (MSFT), Sony (SNE) y Apple (AAPL). Finalmente, la cuarta componente independiente tiene un mayor peso sobre Intel (INTC), Nokia (NOK) y Canon (CAJ).

	IC1(V5)	IC2(V4)	IC3(V7)	IC4(V2)
IBM	-0.913003941	0.077174767	-0.05903354	0.10886592
QCOM	0.009741479	-0.845690901	-0.31731235	0.00265676
SNE	0.039274478	0.189716284	-0.82354516	0.20476168
MSFT	0.042508439	-0.202313391	0.28657628	0.21005782
AAPL	-0.115782329	-0.008437441	-0.26549123	0.01599290
INTC	-0.183559083	-0.394905655	0.02931958	-0.43366043
NOK	-0.040858643	-0.019713275	0.08196077	0.18164530
NTDOY	-0.268969076	0.027692583	0.05330650	-0.07344369
CAJ	0.015469085	0.205973338	-0.15342031	-0.82144840
GOOGL	0.204026678	0.048226585	-0.17616229	0.02461509

Figura 10: Cargas de los componentes independientes ordenados en el ICA (Matriz W). **Fuente:** Elaboración propia

Por lo tanto, a nivel comparativo hay muchas diferencias con los resultados obtenidos en los pesos del ICA respecto a los del PCA y FA. Mientras que las cargas del PCA y el FA sobre las empresas para cuatro factores o componentes son muy parecidas, en el ICA los pesos son muy distintos y los componentes afectan a empresas distintas. La diferencia principal es que en el ICA el primer componente independiente no tiene tanto peso sobre las empresas como sí lo tiene el primer factor del FA o el primer componente del PCA, que recogen más de la mitad de la varianza explicada del modelo.

Además, cabe destacar que en el análisis ICA los pesos quedan mucho más repartidos entre las cuatro componentes independientes, y no es la primera componente la que recoge la mayor parte del impacto sobre las variables.

Si nos fijamos en el gráfico con forma de red neuronal del ICA (gráfico 16) vemos como los pesos de las componentes independientes que se han generado no tienen ninguna relación con los que se obtienen de los componentes de los métodos PCA y el FA.

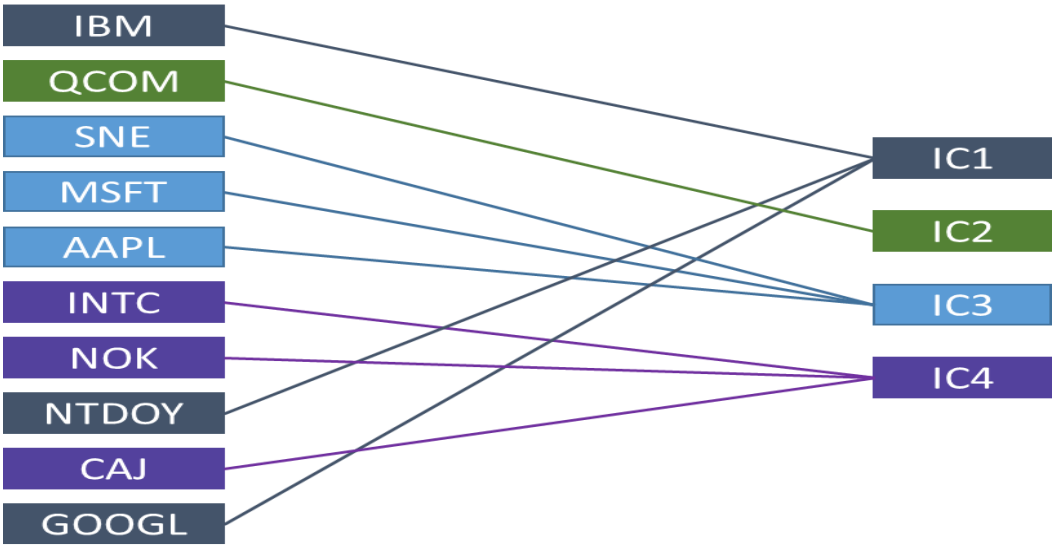


Gráfico 16: Cargas de los componentes independientes ordenados (IC) en el ICA sobre los 10 activos en forma de red neuronal sin parte oculta. **Fuente:** Elaboración propia

Por lo tanto, en el ICA se consigue realizar la extracción de señales y la reducción de dimensión de los datos en componentes independientes. Tanto el PCA, el FA y el ICA tienen el objetivo principal de transformar linealmente las señales observadas en componentes o factores. La diferencia clave se encuentra en el tipo de componentes que se obtienen.

En el ICA se buscan componentes que sean estadísticamente independientes y además los algoritmos para realizar el ICA utilizan estadísticos de orden superior a los del PCA. Por este motivo, las restricciones que se aplican en la metodología ICA nos llevan a que los resultados obtenidos sobre nuestra base de datos de 10 activos tecnológicos sean muy distintos a los obtenidos con el PCA y el FA en cuanto a pesos e impacto de cada una de las componentes sobre las 10 empresas tecnológicas.

Los resultados obtenidos de los análisis multivariantes clásicos PCA y FA parecen tener más sentido a efectos prácticos, ya que las empresas que recogen sus componentes o factores tienen características similares y siguen patrones de comportamiento parecidos a nivel de mercado. Sin embargo, el sentido de las componentes independientes que genera el ICA no parece tan claro a nivel interpretativo, ya que a pesar de que las 10 empresas estudiadas son del ámbito tecnológico, el impacto de las componentes se produce sobre empresas bastante diferentes dentro del sector.

Por ejemplo, IBM, Nintendo (NTDOY) y Google (GOOGL), recogidas por la primera componente independiente del ICA, son empresas dedicadas al sector de la tecnología, pero trabajan con productos muy distintos. Sin embargo, las empresas recogidas por la primera componente principal del PCA o el FA son IBM, Microsoft (MSFT), Apple (AAPL), Intel (INTC) y Google (GOOGL), que son empresas tecnológicas destinadas principalmente al mercado de la informática en cuanto a la producción y comercialización de software y hardware, por lo que guardan una mayor relación entre ellas. Aun así, la aplicación del ICA queda justificada, ya que nuestras series de datos financieros no seguían una normal multivariante y el método ICA es una buena opción de análisis cuando partimos de la no-Gaussianidad de los datos.

5. CONCLUSIONES

A través de este trabajo se ha podido realizar una comparativa de tres métodos de análisis multivariante clásicos, los cuales nos permiten resumir un conjunto de variables en unas pocas variables nuevas con el fin de simplificar y reducir la dimensión de los datos. Además, se ha analizado un método alternativo, el NMF, que nos permite encontrar grupos con características similares.

El primer método de análisis multivariante que se ha definido en el trabajo es el análisis de componentes independientes (ICA), un método que surge como una alternativa a los métodos de análisis clásicos PCA y FA. El ICA es un método muy empleado con frecuencia en diversos campos como la ingeniería y la medicina. En el ámbito financiero se utiliza con frecuencia cuando los resultados de los métodos de análisis multivariante clásicos PCA y FA no son muy realistas o fallan.

En el ICA, al igual que en los métodos clásicos PCA y FA, se basa en la idea de encontrar una representación fidedigna de los datos a partir de la proyección de las observaciones en un espacio de menor dimensión. Es por este motivo que el ICA en muchos casos se considera una extensión del PCA y el FA. Pero el concepto de representación fidedigna de los datos difiere mucho si utilizamos un método de análisis u otro.

El PCA tiene como objetivo obtener componentes principales que no estén correlacionados entre sí. Estos componentes principales serán estadísticamente independientes solo si las observaciones tienen una distribución Normal (Gaussiana). Sin embargo, al aplicar el análisis ICA los componentes que se obtienen son estadísticamente independientes.

Adicionalmente, en el PCA la forma de ordenar los componentes principales se establece en base a sus varianzas. Se establece que el primer componente principal es aquel que recoge la máxima varianza posible, el segundo componente principal recoge la máxima varianza en el subespacio restante, y la misma casuística con el resto de los componentes. Mientras que la forma de ordenar los componentes independientes en el ICA se puede realizar a partir de la correlación existente entre los componentes principales y los componentes independientes (Daniel Peña, 2002). A partir del criterio planteado por Daniel Peña los componentes independientes se ordenan a partir del criterio de máxima correlación, de tal forma que el primer componente independiente es el que está máximamente correlacionado al primer componente principal, el segundo componente independiente tendrá máxima correlación con el segundo componente principal, y así sucesivamente.

En cuanto al método de análisis clásico FA también se pueden extraer similitudes con el método ICA. En el método de análisis FA se asume que los factores subyacentes no están correlacionados, de igual forma que en el PCA. Por otro lado, en el ICA además de ser componentes incorrelacionados también se asume la no-Gaussianidad de los factores y su independencia estadística (Hyvärinen and Kano (2003)).

Además, en este trabajo también se ha definido un método alternativo para el análisis de datos multivariantes, el NMF. Este método, de igual manera que el ICA, el PCA y el FA se basa en la factorización de matrices para su análisis, pero en este caso las matrices de datos son estrictamente no negativas. El método NMF permite mostrar características

relevantes de los datos que no son observables *a priori*. Su inclusión en este trabajo es interesante ya que este método consigue agrupar variables con características similares teniendo en cuenta cuál es su grado de similitud.

Por lo que se refiere a la parte empírica del trabajo a partir de la aplicación de estos cuatro métodos de análisis multivariante sobre nuestras series de datos de rendimientos de 10 activos tecnológicos se han obtenido los siguientes resultados:

A partir del análisis PCA se llegaba a la conclusión de que se podía reducir la dimensión de nuestras series de 10 activos tecnológicos a dos componentes principales que lograban explicar un 60% de varianza. Un porcentaje de varianza explicada que se encuentra muy en el límite de lo aceptado en finanzas, pero que estaba respaldado por el análisis de los vectores de valores propios de la matriz de varianzas y de la matriz de correlaciones de nuestros datos que confirmaban que dos componentes principales son suficientes.

En cuanto al análisis del FA partir del estudio ex-post de nuestras series de datos, y teniendo en cuenta las restricciones que plantea el modelo en cuanto al límite de factores comunes a representar, se llega a la conclusión de que el número de factores óptimo para nuestro ejemplo y que logra un mayor porcentaje de varianza explicada es cuatro. A partir de estos 4 factores comunes obtenidos se lograba un 60% de varianza explicada del modelo.

Después de demostrar que nuestros datos no se distribuyen según una normal multivariante mediante los test de normalidad de Mardia y Henze-Zirkler, la aplicación del método ICA queda totalmente justificada, ya que este método de análisis multivariante parte de la premisa de maximizar la no-Gaussianidad de nuestros datos. Para la estimación del ICA sobre nuestras series de datos hemos partido de utilizar dos componentes para nuestro estudio como nos indicaba el método PCA. En nuestro estudio hemos visto que los pesos y las cargas de estos dos componentes independientes son muy distintos a los que se obtienen del PCA y del FA.

El último método de análisis de datos multivariante aplicado ha sido el NMF. A partir de la aplicación de este método sobre nuestra base de datos se obtiene una agrupación en conglomerados de los activos que tienen características similares. Estos conglomerados representados por un dendrograma se agrupan en función del grado de similitud que tengan las variables entre sí, de tal manera que se forman familias con características similares a partir de una representación en forma de árbol. En el estudio realizado en este trabajo, el método NMF no permite una reducción de dimensión de los datos en nuevos componentes o factores comunes, pero sí que logra agruparlos en función de características similares y de su nivel de similitud, por lo que es un método alternativo muy a tener en cuenta.

Finalmente se ha establecido una comparación entre los pesos o cargas que tienen los componentes o factores de los distintos métodos de reducción de dimensión aplicados PCA, FA y ICA. Para establecer esta comparación se ha fijado el mismo número de factores o componentes a estudiar por cada uno de los métodos.

En cuanto a los pesos de las cuatro componentes principales del PCA sobre nuestras series de datos se observa que la primera componente es la que recoge los mayores pesos en gran parte de las 10 empresas estudiadas. De tal forma que la primera componente tiene un gran peso especialmente en las empresas IBM, Microsoft (MSFT), Apple (AAPL), Intel (INTC) y Google (GOOGL). La segunda componente tendría mayores cargas sobre

las empresas Sony (SNE) y Nintendo (NTSOY), la tercera componente es la que tiene una mayor carga sobre la empresa Nokia (NOK) y finalmente, la cuarta componente tiene un mayor peso en las empresas Qualcomm (QCOM) y Canon (CAJ).

En cuanto al análisis de las cargas en el método FA los resultados obtenidos son muy similares a los obtenidos con el método PCA. El primer factor común es el que tiene una mayor incidencia sobre casi todas las empresas de igual forma que con el primer componente del PCA y afectando a las mismas empresas. El segundo factor cambia respecto al PCA, ya que este afecta a las tres empresas Sony (SNE), Canon (CAJ) y Nintendo (NTDOY). En el caso del tercer factor del FA solo tiene incidencia sobre la empresa Qualcomm (QCOM) y el cuarto factor solo sobre la empresa Nokia (NOK). Por lo tanto, encontramos similitudes especialmente en el primer factor del FA y el primer componente del PCA, pero también en el hecho de que la empresa Nokia (NOK) siempre viene recogida por un único factor o componente, por lo que es una empresa con características diferentes al resto.

Los resultados obtenidos de estos dos métodos tienen bastante sentido, ya que las empresas que recogen sus dos primeros factores o componentes principales son las mismas: IBM, Microsoft (MSFT), Apple (AAPL), Intel (INTC) y Google (GOOGL). Estas empresas se dedican sobre todo al mercado de la informática en cuanto a la producción y comercialización de software y hardware, por lo que guardan una mayor relación entre ellas.

En cuanto a los pesos o cargas obtenidos de los cuatro componentes independientes del método ICA estudiados en nuestro ejemplo los resultados son muy distintos a los de los métodos clásicos PCA y FA. En el método ICA la primera componente independiente tiene mayores pesos sobre las empresas IBM, Nintendo (NTDOY) y Google (GOOGL), que son empresas dedicadas al sector de la tecnología, pero trabajan con productos muy distintos. Por lo tanto, tienen mucho más sentido las empresas que recogen el primer factor del FA y la primera componente principal del PCA, ya que son empresas que compiten y guardan una relación de similitud mayor entre ellas.

La única similitud del ICA con los métodos clásicos anteriores es recoger a Qualcomm (QCOM) en un solo componente como hace el FA en su análisis. Sin embargo, la empresa Nokia (NOK), que tanto en el PCA como el FA quedaba recogida exclusivamente por un solo componente o factor, en el método ICA comparte el cuarto componente independiente con otras empresas que son Intel (INTC) y Canon (CAJ), hecho que no tiene mucho sentido, ya que tampoco guardan mucha relación en sus ámbitos de negocio.

Aun así, la aplicación del ICA queda justificada, ya que nuestras series de datos financieros no seguían una normal multivariante y el método ICA es una buena opción de análisis cuando partimos de la no-Gaussianidad de los datos, pero los resultados obtenidos parecen mucho más interpretables a efectos prácticos en el PCA o el FA e incluso en el método alternativo NMF.

6. BIBLIOGRAFÍA

Álvarez, D.A., and E. Giraldo (2008). *ICA aplicado a la extracción de características en imágenes.* Scientia et Technica Año XIV, No40. Universidad Tecnológica de Pereira.

Black, A. D., and Andreas S. Weigend (1997). *A First Application of Independent Component Analysis to Extracting Structure from Stock Returns*. International Journal of Neural Systems, Vol. 8, No5.

Cadavid, A.C., J.K. Lawrence, A. Ruzmaikin. *Principal Components and Independent Component Analysis of Solar Space Data*. Departament of Physics and Astronomy, California State University, Northridge.

Cardoso, J.F. (1989). *Source separation using higher order moments.* In International Conference on Acoustics, Speech and Signal Processing, pages 2109-2112

Cardoso, J.F. (1997). *Infomax and maximum likelihood for source separation.* IEEE Letters on Signal Processing 4, 112-114.

Cardoso, J.F. (1998). *Blind signal separation: statistical principles*. Proceedings of the IEEE, Vol. 86, pp. 2009-2025. October 1998.

Cazalet, Z., and T. Roncalli (2011). *Nonnegative Matrix Factorization and Financial Applications*. Research & Development.

Comon, P. (1994). *Independent component análisis- a new concept?* Signal Processing 36,287-314.

Comon, J., & Jutten, C. (2010). *Handbook of Blind Source Separation: Inde-pendent Component Analysis and Applications*. Academic Press.

Cover, T. M., and J. A. Thomas (2001). *Elements of Information Theory*. John Wiley.

Cuadras, C. M. (1981). *Métodos de Análisis Multivariante*. EUNIBAR, Barcelona, 1981. 642pp. 2a. edic., PPU, Barcelona, 1991, 3a ed. EUB, Barcelona. 1996.

Delfosse, N., and P. Loubaton (1995). *Adaptative blind separation of independent sources: a deflation approach.* Signal Processing *45*, 59-83.

Gaeta, M., and J.L. Lacoume (1990). *Source separation without prior knowledge: the maximum likelihood solution*. Proceedings of the EUSIPCO, 621-624.

Gaujoux, R. (February 18, 2018). *An introduction to NMF package. Version 0.20.6*

González, E. (2011). *Independent Component Analysis for Time Series* (Tesis doctoral). Universidad Carlos III de Madrid.

Hyvärinen, A. (1999). *Fast and robust fixed-point algorithms for independent componente analysis*. IEEE Transactions on Neural Networks, *10, 626-634*

Hyvärinen, A., and E. Oja (2000). *Independent Component Analysis: Algorithms and Applications.* Neural Networks, 13(4-5):411-430.

Hyvärinen, A. (2001). *Blind source separation by nonstationarity of variance: A cumulant- based approach*. IEEE Transactions on Neural Networks 12.

Hyvärinen, A., and K. Kano. (2003). *Independent component análisis for non-normal factor analysis.* Tokyo: Springer Verlag.

Lee, D.D., and H.S Seung (1999), *Learning the Parts of Objects by Non-Negative Matrix Factorization*, Nature, 401, pp. 788-791.

Lee, D.D., and H.S Seung (2001), *Algorithms for Non-negative Matrix Factorization, Advances in Neural Information Processing Systems*, 13, pp.556-562.

Lee, T.W., M. Girolami, A. J. Bell, and T. J. Sejnowski (2000). *A unifying information theoretic framework for independent component analysis*. Computers and Mathematics with Applications 11, 1-21.

Martínez, C.G. (2018). *Máquinas de vector soporte*. Página web: https://rpubs.com/Cristina_Gil/PCA

Martos, G. (2014). *Reduciendo la dimensión: Componentes Principales*. Técnicas de investigación, Web: https://rpubs.com/gabrielmartos/multivPCA.

Muñoz, J., J. Rivera, E. Duque (2008). *Análisis de Componentes Principales e Independientes aplicados a la reducción de ruido en señales electrocardiográficas*. Scientia et Technica Año XIV, No39. Universidad Tecnológica de Pereira.

Peña, D. (2002). *Análisis de Datos Multivariantes.* Article of University Carlos III de Madrid.

Pham, D. T., P. Garrat, and C. Jutten (1992). *Separation of a mixture of independent sources through a maximum likelihood approach*. Proceeding EUSIPCO, 771-774.

Shlens, J. (2005). *A tutorial on principal component analysis. Derivation, discusion and singular value decomposition. (Version 1)*

Vicente Villardón, J.L. *Introducción al análisis de clúster*. Departamento de Estadística, Universidad de Salamanca.

Zamora, R., and J. Esnaola (2015). *Análisis Factorial y Análisis de Componentes Principales.* Ayudantía Estadística IV, Sociología, Universidad de Chile.

Webgrafía

La base de datos de precios diarios de las 10 empresas tecnológicas estudiadas y la información sobre estas empresas ha sido extraída de *Yahoo Finance*, páginas web:

-IBM. https://es.finance.yahoo.com/quote/IBM?p=IBM&.tsrc=fin-srch

-Qualcomm. https://es.finance.yahoo.com/quote/QCOM?p=QCOM&.tsrc=fin-srch

-Microsoft. https://es.finance.yahoo.com/quote/MSFT?p=MSFT&.tsrc=fin-srch

-Sony. https://es.finance.yahoo.com/quote/SNE?p=SNE&.tsrc=fin-srch

-Apple. https://es.finance.yahoo.com/quote/AAPL?p=AAPL&.tsrc=fin-srch

-Intel. https://es.finance.yahoo.com/quote/INTC?p=INTC&.tsrc=fin-srch

-Canon. https://es.finance.yahoo.com/quote/CAJ?p=CAJ&.tsrc=fin-srch

-Nintendo. https://es.finance.yahoo.com/quote/NTDOY?p=NTDOY&.tsrc=fin-srch

-Google. https://es.finance.yahoo.com/quote/GOOGL?p=GOOGL&.tsrc=fin-srch

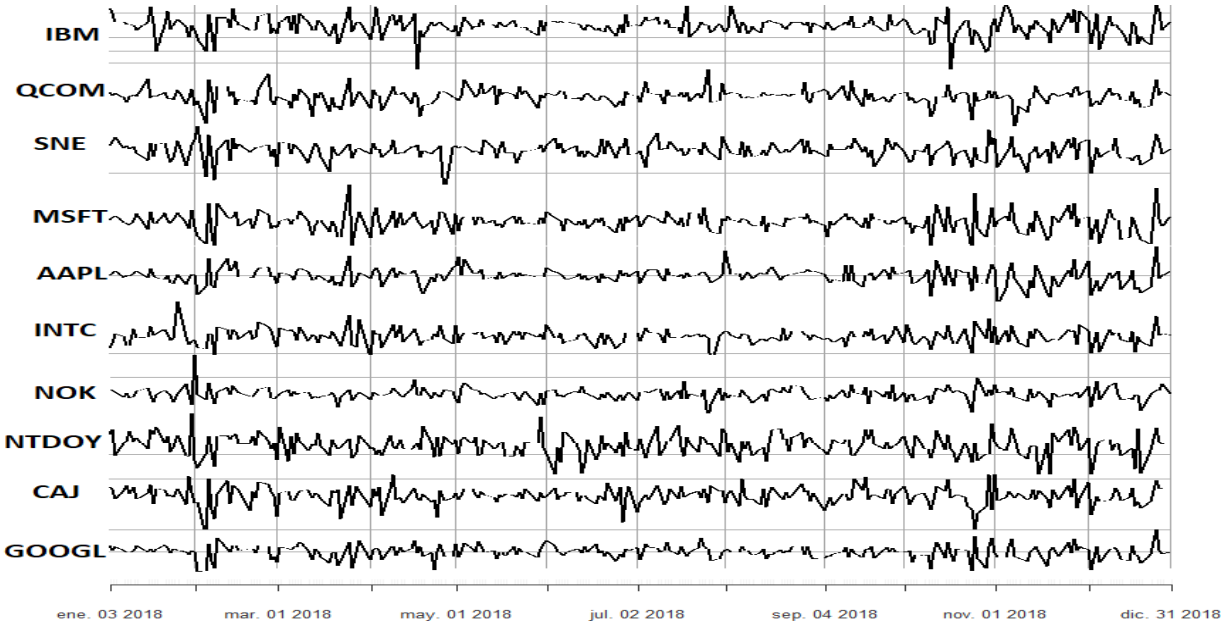
-Nokia. https://es.finance.yahoo.com/quote/NOK?p=NOK&.tsrc=fin-srchç

7. ANEXOS

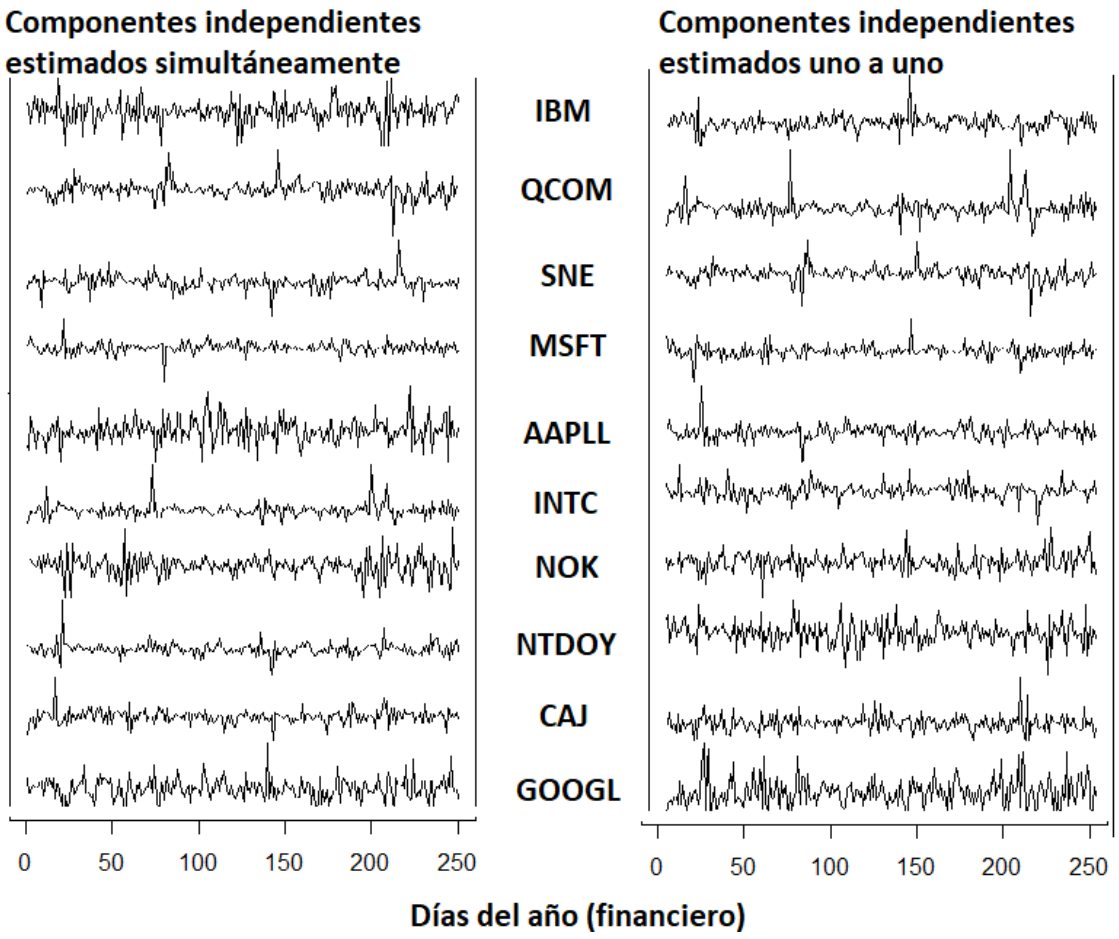
Anexo 1. Descripción de la actividad de las 10 empresas tecnológicas estudiadas.

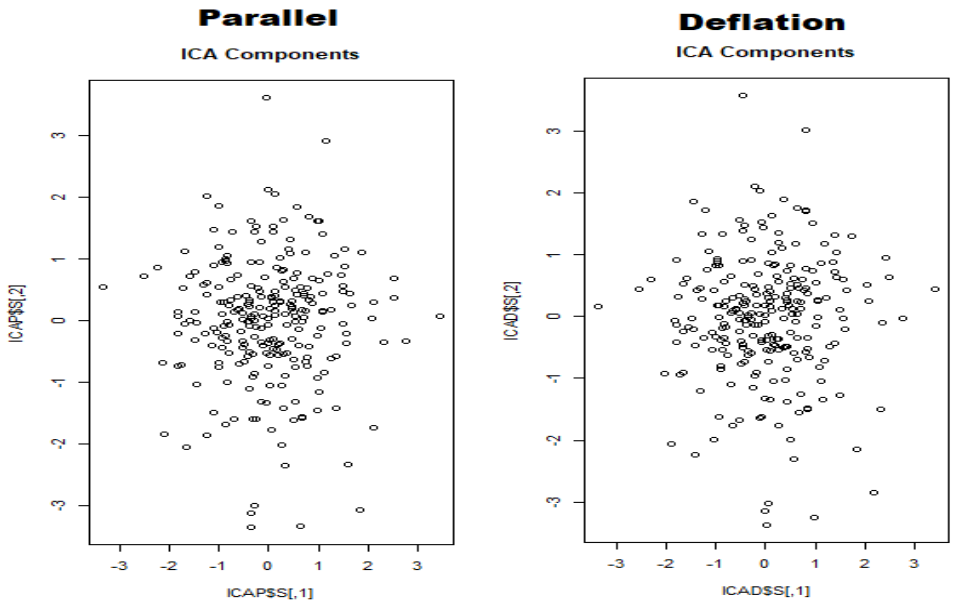
La empresa IBM, que es una multinacional estadounidense dedicada principalmente a la fabricación y comercialización de hardware y software tecnológico para ordenadores y otro tipo de tecnologías. La segunda empresa seleccionada ha sido Qualcomm (QCOM) que principalmente se encarga de producir y diseñar procesadores y tecnología para dispositivos móviles. Otra empresa de nuestra base de datos será Sony (SNE), una multinacional japonesa que se dedica a la fabricación de todo tipo de electrónica de consumo, desde audio y video, hasta computación, fotografía y videojuegos. La empresa Microsoft (MSFT) con sede en Estados Unidos, que se encarga de desarrollar, manufacturar, licenciar y proveer de soporte para el software de ordenadores personales, servidores, dispositivos electrónicos, entre otros. Otra empresa de nuestra base de datos es la multinacional estadounidense Apple (AAPL) que produce y diseña equipos electrónicos, tanto ordenadores como teléfonos y todo tipo de software. La siguiente empresa es Intel (INTC), otra compañía estadounidense que se encarga de la fabricación de procesadores para la mayoría de los ordenadores personales del mundo. Otra empresa utilizada en nuestra base de datos es la multinacional Finlandesa Nokia (NOK), que se orienta principalmente a la fabricación de teléfonos móviles. Nintendo (NTDOY) es otra de las 10 empresas escogidas, esta empresa tiene su sede en Japón y se dedica principalmente al mercado de videojuegos y a la electrónica de consumo, especialmente de videoconsolas. La siguiente compañía escogida es Canon (CAJ), procedente de Japón y especializada en productos de captura y reproducción de imágenes, lo que incluye fotografía, video, fotocopiadoras e impresoras. Finalmente, la última empresa escogida para nuestra base de datos es Google (GOOGL) que es una multinacional estadounidense cuya especialización son los servicios relacionados con Internet, software, dispositivos electrónicos y otro tipo de tecnologías.

Anexo 2. Gráfico de los rendimientos relativos continuos de los 10 activos.



Anexo 3. Gráficos de los componentes independientes estimados mediante el algoritmo parallel o el deflation.





Anexo 4. Comandos y paquetes utilizados en RStudio para el análisis de nuestra base de datos

```
install.packages("quantmod")
library(quantmod)

##10 activos utilizados para nuestro estudio
getSymbols("IBM",from ="2018-01-01", to = "2019-01-01")
head(IBM)
chartSeries(IBM)
chartSeries(IBM[,4], type="lines")
r1 = diff(log(IBM[,4]))*100
r1
r1=r1[-1]
length(r1)

getSymbols("QCOM",from ="2018-01-01", to = "2019-01-01")
head(QCOM)
chartSeries(QCOM)
chartSeries(QCOM[,4], type="lines")
r2 = diff(log(QCOM[,4]))*100
r2
r2=r2[-1]
length(r2)

getSymbols("SNE",from ="2018-01-01", to = "2019-01-01")
head(SNE)
chartSeries(SNE)
chartSeries(SNE[,4], type="lines")
r3 = diff(log(SNE[,4]))*100
r3
r3=r3[-1]
length(r3)

getSymbols("MSFT",from ="2018-01-01", to = "2019-01-01")
head(MSFT)
chartSeries(MSFT)
chartSeries(MSFT[,4], type="lines")
r4 = diff(log(MSFT[,4]))*100
r4
r4=r4[-1]
length(r4)

getSymbols("AAPL",from ="2018-01-01", to = "2019-01-01")
head(AAPL)
chartSeries(AAPL)
chartSeries(AAPL[,4], type="lines")
r5 = diff(log(AAPL[,4]))*100
r5
r5=r5[-1]
length(r5)
```

```
getSymbols("INTC",from ="2018-01-01", to = "2019-01-01")
head(INTC)
chartSeries(INTC)
chartSeries(INTC[,4], type="lines")
r6 = diff(log(INTC[,4]))*100
r6
r6=r6[-1]
length(r6)

getSymbols("NOK",from ="2018-01-01", to = "2019-01-01")
head(NOK)
chartSeries(NOK)
chartSeries(NOK[,4], type="lines")
r7 = diff(log(NOK[,4]))*100
r7
r7=r7[-1]
length(r7)

getSymbols("NTDOY",from ="2018-01-01", to = "2019-01-01")
head(NTDOY)
chartSeries(NTDOY)
chartSeries(NTDOY[,4], type="lines")
r8 = diff(log(NTDOY[,4]))*100
r8
r8=r8[-1]
length(r8)

getSymbols("CAJ",from ="2018-01-01", to = "2019-01-01")
head(CAJ)
chartSeries(CAJ)
chartSeries(CAJ[,4], type="lines")
r9 = diff(log(CAJ[,4]))*100
r9
r9=r9[-1]
length(r9)

getSymbols("GOOGL",from ="2018-01-01", to = "2019-01-01")
head(GOOGL)
chartSeries(GOOGL)
chartSeries(GOOGL[,4], type="lines")
r10 = diff(log(GOOGL[,4]))*100
r10
r10=r10[-1]
length(r10)

activo<-cbind(r1,r2,r3,r4,r5,r6,r7,r8,r9,r10)
activos<-as.matrix.data.frame(activo)
```

```
colnames(activos)<-
c("IBM","QCOM","SNE","MSFT","AAPL","INTC","NOK","NTDOY","CAJ","GOO
GL")
head(activos)

er<-colMeans(activos)
cor.mat<-cor(activos)
cov.mat<-cov(activos)

#paquete MVN para saber si mis activos siguen una normal multivariante

install.packages("MVN")
library(MVN)

res1<-mvn(activos, mvnTest = "mardia");res1$multivariateNormality
res2<-mvn(activos,mvnTest="hz");res2$multivariateNormality
res3<-mvn(activos,mvnTest = "royston");res3$multivariateNormality
res4<-mvn(activos,mvnTest = "dh");res4$multivariateNormality
res5<-mvn(activos,mvnTest = "energy");res5$multivariateNormality

#matriz de correlaciones, test de esfericidad de Bartlett y descomposicion de Eigen

install.packages("REdaS")
library(REdaS)
install.packages("grid")
library(grid)


corr <- cor(activos);corr
View(activos)

##test de bartlett: Test de Esfericidad de Bartlett: Comprueba que la matriz de
correlaciones se ajuste a la matriz identidad ( I), es decir ausencia de correlación
significativa entre las variables. Esto significa que la nube de puntos se ajustara a una
esfera perfecta, expresando así la hipótesis nula por:
#Ho: R = I es decir, que el determinante de la matriz de correlaciones es 1.

bart_spher(activos)
myEig <- eigen(corr);myEig
myEig$vectors[,1]
View(activos)



##ANÁLISIS PCA

##MATRIZ DE CORRELACIONES##
myEig <- eigen(corr);myEig
myEig$vectors[,1]
View(activos)
```

```
##MATRIZ DE VARIANZAS##
var(activos)
eigen(var(activos))
mean(eigen(var(activos))$values)

# Pricipal Components Analysis
# entering raw data and extracting PCs
# from the correlation matrix

fit <- princomp(activos, cor=TRUE)
summary(fit)

loadings(fit) # pc loadings

plot(fit,type="lines") # scree plot
plot(fit,type="barplot") # scree plot
fit$scores # the principal components
biplot(fit)


##PCA con Factor Miner, Proporciona otros gráficos

install.packages("FactoMineR")

library("FactoMineR")

res.pca <- PCA(activos, graph = FALSE)
print(res.pca)

install.packages("ggplot2")
install.packages("factoextra")

library(ggplot2)
library(factoextra)

eig.val <- get_eigenvalue(res.pca)
eig.val

fviz_eig(res.pca, addlabels = TRUE, ylim = c(0, 50))
var <- get_pca_var(res.pca)
var

# Coordinates
head(var$coord)
# Cos2: quality on the factore map
head(var$cos2)
# Contributions to the principal components
head(var$contrib)
# Coordinates of variables
```

```
head(var$coord, 4)

fviz_pca_var(res.pca, col.var = "black")

head(var$cos2, 4)

install.packages("corrplot")
library(corrplot)
corrplot(var$cos2, is.corr=FALSE)

fviz_cos2(res.pca, choice = "var", axes = 1:2)
fviz_pca_var(res.pca, col.var = "cos2",
            gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
            repel = TRUE # Avoid text overlapping
            )

 fviz_pca_var(res.pca, alpha.var = "cos2")
head(var$contrib, 4)
corrplot(var$contrib, is.corr=FALSE)
fviz_contrib(res.pca, choice = "var", axes = 1, top = 10)
fviz_contrib(res.pca, choice = "var", axes = 2, top = 10)
fviz_pca_var(res.pca, col.var = "contrib",
            gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07")
            )

            fviz_pca_var(res.pca, alpha.var = "contrib")
res.desc <- dimdesc(res.pca, axes = c(1,2), proba = 0.05)
# Description of dimension 1
res.desc$Dim.1
res.desc$Dim.2


##ANÁLISIS FACTORIAL##

##EX ANTE##
library(grid)
library(REdaS)
# correlation matrix
cor(activos)
# bartlett's test
bart_spher(activos)

# Maximum Likelihood Factor Analysis
# entering raw data and extracting 2 factors,
# with varimax rotation

##EX POST##
fit <- factanal(activos, 4,rotation="varimax")
print(fit, digits=2, cutoff=.3, sort=TRUE)
fit$scores
```

```
##Dos factores son suficientes, uno no

# plot factor 1 by factor 2

load <- fit$loadings[,1:2]
plot(load,type="n") # set up plot
text(load,labels=names(activos),cex=.7) # add variable names



###ANALISIS ICA####
install.packages("ica")
library(ica)
install.packages("fastICA")
library(fastICA)

##Usamos el algoritmo parallel para estimar los componentes a la vez (es más correcto) con 2 componentes que nos dice el PCA (proceso de blanqueamiento)


ICAP<-fastICA(activos, n.comp = 2,alg.typ = "parallel");View(ICAP)#componentes estimados a la vez
View(ICAD$X);View(ICAD$K);View(ICAD$W);View(ICAD$A);View(ICAD$S)


par(mfrow=c(1,3))
plot(ICAP$X,main="Pre-processed data")
plot(ICAP$X%*%ICAP$K,main="PCA Components")
plot(ICAP$S,main="ICA Components")

corrica<-cor(ICAP$S)#matriz independiente de mis datos para extraer componentes es ICA$S
corrica

##Vemos la diferencia gráfica de componentes estimados a la vez (ICAP) y componentes estimados uno a uno (ICAD) con 2 comp

ICAP<-fastICA(activos, n.comp = 2,alg.typ = "parallel");View(ICAP)#componentes estimados a la vez

ICAD<-fastICA(activos, n.comp = 2, alg.typ = "deflation")#COMPONENTES ESTIMADOS 1 POR 1

a<-diff(ICAD$X-ICAP$X)#iguales
b<-diff(ICAD$K-ICAP$K)#iguales

##Comparación señal PCA-ICA

PCAs<-ICAP$X%*%ICAP$K
```

PCAs

```
par(mfcol = c(1, 3))
plot (ICAP$X[,1], type = "l",xlab = "Datos preprocesados:IBM", ylab = "")
plot (PCAs[,1], type = "l",xlab = "Componente principal 1", ylab = "")
plot (ICAP$S[,1], type = "l",xlab = "Componente independiente 1", ylab = "")

par(mfcol = c(1, 3))
plot (ICAP$X[,2], type = "l",xlab = "Datos preprocesados:QCOM", ylab = "")
plot (PCAs[,2], type = "l",xlab = "Componente principal 2", ylab = "")
plot (ICAP$S[,2], type = "l",xlab = "Componente independiente 2", ylab = "")

par(mfcol = c(1, 2))
plot (ICAP$S[,1], type = "l",xlab = "Componente independiente 1", ylab = "")
plot (ICAP$S[,2], type = "l",xlab = "Componente independiente 2", ylab = "")

#Diferencia gráfica de componentes estimados a la vez (ICAP) y componentes
estimados uno a uno (ICAD)

par(mfcol = c(1, 2))
plot (ICAP$S[,1], type = "l",xlab = "CI:1 Parallel", ylab = "")
plot (ICAD$S[,1], type = "l",xlab = "CI:1 Deflation", ylab = "")

plot(ICAP$S[,2], type = "l", xlab = "CI:2 Parallel", ylab = "")
plot(ICAD$S[,2], type = "l", xlab = "CI:2 Deflation", ylab = "")


par(mfrow=c(1,3))
plot(ICAD$X,main="Pre-processed data")
plot(ICAD$X%*%ICAD$K,main="PCA Components")
plot(ICAD$S,main="ICA Components")

##Análisis de la correlación entre los componentes de PCA e ICA

cor(fit$scores,ICAP$S)



## ANÁLISIS NMF##

install.packages("NMF")
library(NMF)


nmfSeed("ICA")

##implementación del método NMF con todos los algoritmos

nmfAlgorithm(all=TRUE)
meth<-nmfAlgorithm(version="R")
meth<-c(names(meth),meth)
```

```
meth

#Nuestra matriz de datos debe ser no negativa

activos2<-abs(activos)

#Buscamos cual es el rango que aplicar al método NMF

estim.r<-nmf(activos2,2:6,nrun=10,seed=123456)

#Estimación del rango, mesuras de calidad de 10 veces de cada valor de r

plot(estim.r)
v.random<-randomize(activos2)
estim.r.random<-nmf(v.random, 2:6, nrun=10, seed=123456)
plot(estim.r,estim.r.random)

## El plot anterior Nos lleva a la conclusión que el rango es 3 (punto en el que decrece
el cophenetic)

##Aplicación del método NMF con distintos algoritmos

res<-nmf(activos2,3,meth,seed=123456)

res.multi.method<-nmf(activos2,3,list("brunet","lee","ns"),seed=123456,.options="t")

#Comparación de los algoritmos

compare(res.multi.method)

res<-nmf(activos2,3,.options="t")

#Error track
par(mfcol = c(1, 2))
plot(res)   #para una sola tirada del NMF
plot(res.multi.method)   #para multiples tiradas del NMF


#Mapa de calor

install.packages("gridBase")
library(gridBase)

layout(cbind(1,2))

#Componentes básicos
basismap(res,subsetRow=TRUE)

#Mixtura de coeficientes
coefmap(res)
```