TRABAJO FIN DE MÁSTER

DEPARTAMENTO DE ESTADÍSTICA

MÁSTER EN ANÁLISIS AVANZADO DE DATOS MULTIVARIANTES

"ANÁLISIS DE COMPONENTES PRINCIPALES *SPARSE*

Formulación, algoritmos e implicaciones en el análisis de datos "

Nerea González García

Alejandra Taborda Londoño

Tutoras

Mª Purificación Galindo Villardón

Ana Belén Nieto Librero

2015

Dpto. de Estadística

Universidad de Salamanca

Mº Purificación Galindo Villardón

Profesora titular del Departamento de Estadística de la Universidad de Salamanca

Ana Belén Nieto Librero

Profesora asociada del Departamento de Estadística de la Universidad de Salamanca

---

CERTIFICA que **D.ª Alejandra Taborda Londoño y D.ª Nerea González García** han realizado en la Universidad de Salamanca, bajo su dirección, el trabajo que para optar título de Máster en Análisis Avanzado de Datos Multivariantes presenta con el título ***Análisis de Componentes Principales Sparse: formulación, algoritmos e implicaciones en el análisis de datos***, autorizando expresamente su lectura y defensa.

Y para que conste, firma el presente certificado en Salamanca a 14 de julio de 2015.

Mª Purificación Galindo Villardón

Ana Belén Nieto Librero

AGRADECIMIENTOS

Nos gustaría aprovechar este espacio para mostrar nuestros más profundos agradecimientos a todas aquellas personas que, de alguna manera, forman parte de este trabajo.

En primer lugar, nuestros agradecimientos más sinceros a nuestras tutoras, Puri y Ana, por abrirnos las puertas del departamento y confiar en nosotras. No sólo por su dirección, sino también por sus ganas de trabajar, por guiarnos, y por compartir sus conocimientos para formarnos como profesionales de esta área.

Agradecer a todos los profesores del departamento por sus conocimientos y su dedicación a lo largo del Máster, así como a nuestros compañeros por estar con nosotras en este camino.

A la Universidad de Salamanca, por acogernos y permitir formar parte de sus proyectos e incrementar nuestras habilidades.

Personalmente, a mí, Alejandra, me gustaría dar las gracias a:

> A mi familia y demás seres queridos que con su apoyo tanto moral como económico me ayudaron a alcanzar esta meta. Especialmente, a mi madre porque con sus consejos y amor ha logrado compartirme su fuerza, tenacidad y perseverancia.
>
> A mi excelente compañera Nerea por dedicación y empeño en la elaboración de este trabajo de investigación. Principalmente, por haberme brindado su amistad y su paciencia.
>
> A mis profesores de Colombia porque han sido los primeros en creer en mis capacidades y me han brindado su ayuda para labrar mi camino profesional.
>
> A Instituciones como el FONDE DE EPM, ICETEX y el Banco Santander que han financiado mis estudios y me han brindado la oportunidad de educarme.

A mí, Nerea, me gustaría dar las gracias:

> A mi compañera de trabajo, Ale, que a lo largo de estos meses se ha convertido en amiga. Por su ayuda y comprensión, por nuestras cabezonerías, por compartir conmigo todos sus conocimientos y haber recorrido este camino junto a mí.

A mis padres, por apoyarme en todo lo que hago, por creer en mis ilusiones y hacerme creer en mí. Gracias por respaldar las decisiones que he tomado hasta el momento y que me han llevado hasta aquí. Gracias a mi enana favorita, Emma, porque, aunque a tu manera, me ayudas aportándome cosas diferentes.

A mi kuadrilla y amigas, que con sus *pinchos y cañas* apoyan cada una de mis locuras, algunas sin sentido, estando a mi lado y en la distancia.

Y sobre todo, gracias, gracias, gracias, a ti. Por escucharme cuando lo necesito. Por confiar en mí cuando yo dudo de mí misma. Por estar a mi lado durante todas las fases que tengo a lo largo de los días, ya sean de alegría, nervios, estrés o completa felicidad, jeje. Porque eres mi mejor apoyo y acompañante de vida. Ahora sí, nos hemos ganado unos buenos batidos!

RESUMEN

El Análisis de Componentes principales es una de las técnicas más implementada en las etapas de pre-procesamiento o de reducción de la dimensión de matrices de datos. Su principal función es proyectar los datos de entrada en nuevas direcciones, conocidas como componentes principales (PCs), que absorban la mayor cantidad de información posible y así, poder eliminar aquellas variables que aporten menos variabilidad. Sin embargo, la interpretación de las PCs es complicada, pues resultan de la combinación lineal de todas las variables originales. Es por ello que surgen distintas formas de enfrentar esta problemática; como los conocidos métodos de rotación.

En este trabajo, se presenta el Análisis de Componentes Principales *Sparse (SPCA)* como otra forma de solventar esta dificultad. Es un método de selección de variables características, intentando que gran parte de las cargas que definen las PCs sean nulas (cargas *sparse).* A partir de la búsqueda de bibliografía relevante, se redactará el estado del arte del SPCA, integrando los enfoques de maximización de la varianza y minimización del error en el SPCA. Profundamente, se enfocará la técnica a partir de la reformulación del PCA como problema de minimización del error, aprovechando los desarrollos de los modelos de regresión lineal e integrando restricciones típicas de estos, como la penalización *Elastic net,* para mejorar el análisis de datos. Se comienza entonces con la formulación del SPCA, los algoritmos e implicaciones en el análisis de datos, comparando diferencias entre las componentes principales clásicas, las soluciones rotadas y las soluciones *sparse*.

Palabras clave: *Ridge*, *Lasso*, *Elastic net*, *sparse*.

NOTACIÓN

A continuación se presenta la notación matemática básica usada para el desarrollo del presente trabajo.

Parámetros

n: Número de observaciones

p: Número de variables

r: Número reducido de componentes principales.

Índices

$i = \{1, \dots, n\}$: La i −ésima observación X.

$j = \{1, \dots, r, \dots, k\}$: La j −ésima variable de X, o j −ésima componente principal.

Matrices

X: Matriz de datos centrada por columnas de orden $(n \times p)$.

$S = \frac{1}{n} X^T X$: Matriz de covarianza muestral de X.

$R = XX^T$: Matriz de correlación muestral de X. También puede definirse como: $R = \hat{\Sigma} = X^T X$

$X = UDV^T$: Descomposición en valores singulares (SVD) de X.

U: Matriz $n \times n$ con los vectores propios de X, asociados a una base ortonormal de las Columnas de X.

D: Matriz diagonal singular $n \times p$ con los valores singulares de X

$V =$ Matriz de cargas de las componentes principales

V^T: Matriz transpuesta de $p \times p$ con los vectores propios de X, asociados a una base ortonormal de las filas de X.

$Y = XV = UD$: Matriz de componentes principales.

Vectores

x_i: El i −ésimo vector fila de la matriz X.

y_i: La i −ésima Componente principal de X.

$a_i = b_i = v_i = \beta_i$: El vector de saturaciones a_i de la $i - ésima$ PC, los coeficientes de regresión b_i, las cargas v_i de la $i - ésima$ PC y los coeficientes penalizados β_i que presentan una interpretación equivalentes entre sí.

CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

1. INTRODUCCIÓN

El Análisis de Componentes principales (*Principal Component Analysis*, PCA) puede definirse como una técnica de extracción de características, implementada como etapa de pre-procesamiento o de reducción de la dimensión de matrices de datos. Su principal función es proyectar los datos de entrada en nuevas direcciones que absorban la mayor cantidad de información posible y así, poder eliminar aquellas variables que menos aporten (Sanchez Mangos, 2012).

Aunque es sabido que el PCA en sus orígenes surgió como una técnica propia, sus desarrollos más recientes se han dado entorno al Aprendizaje máquina (*Machine Learning,* ML), a partir de los modelos de regresión. El concepto de ML hace referencia a la familia de técnicas que permiten mejorar un sistema en el tiempo, a partir de la experiencia. Estas técnicas están compuestas por los diferentes tipos de clasificación como el análisis clúster, los métodos reducción de la dimensión, la selección de modelos como los de regresión lineal y el pre procesamiento y extracción de características (Nilsson, 1996).

El PCA es una técnica simple que se utiliza tanto para la reducción de la dimensión como para la extracción de características, sin importar el uso posterior de los datos. Esta es la razón por la cual la técnica es ampliamente utilizada en diferentes áreas del conocimiento y con diferentes enfoques. Por ello, se ha convertido en una de las técnicas de análisis estadístico multivariante más recurridas.

Por este motivo, diversos autores han propuesto mejoras para intentar paliar las deficiencias de la propia técnica, haciendo hincapié en su deficiencia principal: la interpretación de las componentes principales, dado que cada una de ellas se forma como combinación lineal de todas las variables originales.

Con el objetivo de resolver el problema y facilitar la interpretación de las componentes principales, han surgido distintas técnicas a lo largo de los años: técnicas de rotación, técnicas que restringen el valor que se puede asignar a las cargas de las componentes principales, técnicas de reducción del valor de las variables; y no sólo de dimensionalidad, y técnicas de penalización a partir de modelos de regresión lineal múltiple, que producen modelos interpretables introduciendo términos de penalización en el problema de optimización.

Ahora bien, estas técnicas no siempre proporcionan los resultados esperados. A esto se suma la necesidad primordial del análisis de los llamados *big data*. Todo ello hace que algunas de estas técnicas se queden cortas para las finalidades buscadas.

Desde el último enfoque de las técnicas de penalización es desde el que se desarrolla el Análisis de Componentes Principales *Sparse (Sparse Principal Component Analysis,* SPCA), como un resurgir del PCA con los avances en los modelos de regresión. Se introduce un nuevo método para producir componentes principales modificadas con cargas *sparse;* es decir, componentes principales modificadas en las que se fuerza a que los vectores de proyección tengan gran cantidad de sus cargas nulas. Es una herramienta que permite reducir la dimensión de matrices de datos, haciendo selección de variables (incluso cuando el número de variables es muy superior al número de observaciones) y mejorando la interpretación de los datos originales.

Las dos vertientes encontradas en el PCA (búsqueda de las componentes principales a través del subespacio de mejor ajuste a los datos (Pearson, 1901) y búsqueda de las componentes principales a través de la maximización de la información de los datos originales absorbida por ellas (Hotelling, 1933)) vuelven a aparecer en el caso del SPCA, donde surgen algoritmos que resuelven el problema desde ambas vías. Gracias a los avances en los modelos de regresión, en este trabajo se abordará la resolución del SPCA por medio de la solución del algoritmo *Elastic net (*técnica de regresión penalizada que surge como combinación de otros dos tipos de regresión: las regresiones *Ridge* y *Lasso* (Zou et al., 2006), recogiendo sus ventajas y remediando sus defectos).

Gracias a la reformulación del PCA como un problema de optimización de tipo regresión; es decir, como un problema de minimización del error, se facilitará la modificación del comportamiento del algoritmo mediante la adición de nuevos términos, surgidos para problemas de regresión y ya aplicables sobre el PCA modificado como un problema de regresión. Las cargas *sparse* se obtendrán más tarde imponiendo la restricción *Elastic net* en los coeficientes de regresión. Se propondrán algoritmos eficientes para adaptar los modelos SPCA para datos en los que el número de variables sea muy superior al número de observaciones en el estudio (Zou et al., 2006).

La Figura 1 recoge la redefinición del PCA a lo largo del tiempo, hasta llegar al SPCA actual. Se observa la versatilidad de la técnica, que ha ido modificándose o empleándose de diferentes formas debido a su aplicación en todo tipo de investigaciones científicas, desde las estadísticas hasta las biológicas o geológicas, en las que es deseable ajustar un conjunto de datos de gran tamaño en un plano menor (Jolliffe, 2002).

Una vez definidos los objetivos del trabajo y la metodología empleada para alcanzarlos en los capítulos 2 y 3, se comenzará, en el capítulo 4, con una breve introducción a las componentes principales, desde sus distintos enfoques, así como de las técnicas de rotación más empleadas en la resolución práctica. El capítulo 5 introducirá la teoría de los modelos de regresión necesarios de entender para la introducción del Análisis de Componentes Principales *Sparse:* partiendo del modelo de regresión ordinaria, se llegará a las penalizaciones *Ridge, Lasso y Elastic net*. En el capítulo 6 se presenta el Análisis de Componentes Principales *Sparse,* desde el punto de vista de un problema de mínimos cuadrados. Una vez reformulado el PCA como un problema de minimización del error, se añadirá a este la penalización presentada *Elastic net,* obteniendo un análisis de componentes principales *sparse* con cargas nulas. El capítulo 7 recoge las posibilidades de software disponibles para la resolución de esta técnica y dos casos prácticos en los que se verá la resolución del SPCA sobre una matriz de datos en estudio. Todo ello para acabar con una conclusión final en el capítulo 8.

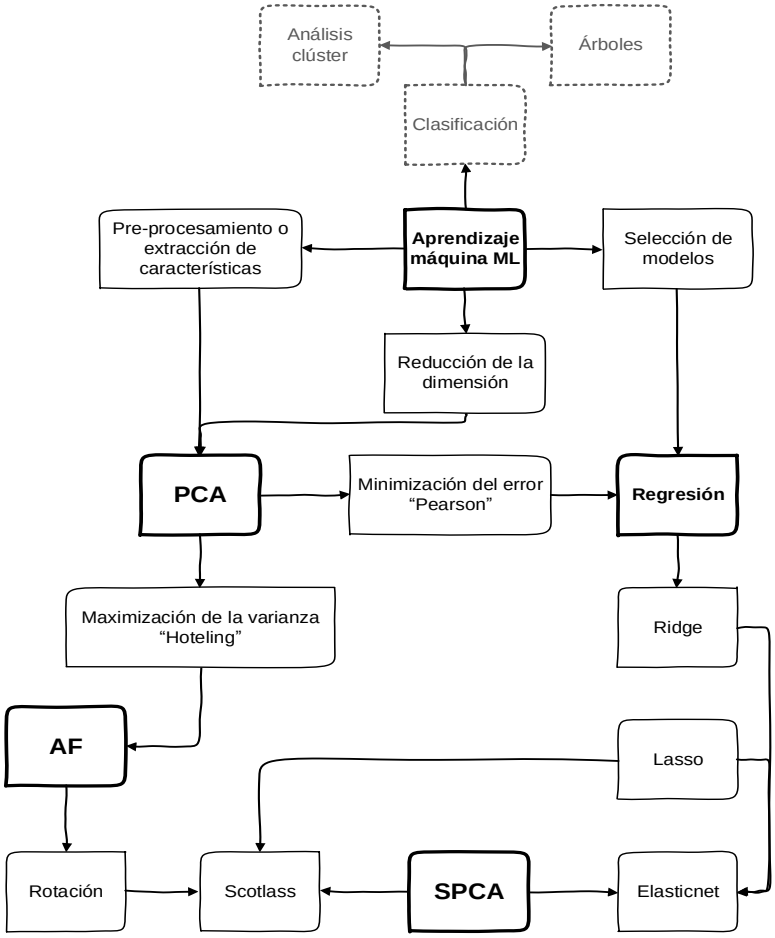

Figura 1. Introducción al PCA

2. OBJETIVOS

Los objetivos de este trabajo de fin de máster son:

2.1. Objetivo general

Presentar la técnica conocida como Análisis de Componentes Principales *Sparse*.

Integrar los enfoques de maximización de la varianza y minimización del error en el Análisis de Componentes Principales *Sparse* realizando una introducción a su formulación, los algoritmos existentes para su resolución e implicaciones en el análisis de datos.

2.2. Objetivos específicos

- Identificar el estado del arte de la literatura científica sobre el Análisis de Componentes *Sparse.*
- Comparar los diferentes enfoques de formulación del Análisis de Componentes principales *Sparse.*
- Ejemplificar las soluciones obtenidas mediante la versión clásica del Análisis de Componentes principales, los métodos de rotación y las diferentes versiones *Sparse*.

3. METODOLOGÍA Y MÉTODOS

3.1. Metodología

La metodología diseñada para alcanzar los objetivos de investigación propuestos se basa en dos etapas, así:

Etapa I: Revisión Bibliográfica

Esta etapa comenzó con la búsqueda de bibliografía relevante en las herramientas SCOPUS y Web of Science, así como Google académico. Lo cual permitió identificar los artículos seminales en el Análisis de Componentes Principales y realizar el estudio exhaustivo de los mismos. La misma búsqueda permitió redactar el estado del arte en el Análisis de Componentes Principales *Sparse* desde los enfoques de maximización de la varianza y minimización del error. Se realizó una breve descripción de los diferentes algoritmos para el cálculo de las soluciones *sparse*.

Etapa II: Ejemplificación

Esta comienza con la búsqueda de las librerías del programa R que permitan ejecutar las diferentes versiones del Análisis de Componentes Principales *Sparse,* así como el Análisis de Componentes Principales clásico y la rotación más empleada en el análisis de datos, la rotación VARIMAX, y comparar los resultados obtenidos con cada una. En dicha comparación se ejemplificaran las diferencias entre las componentes principales clásicas, las soluciones rotadas y las diferentes componentes *sparse*.

3.2. Métodos

El desarrollo de este Trabajo de Fin de Máster se basa fundamentalmente en los siguientes métodos empíricos, estadísticos y teóricos.

Los métodos empíricos

Los datos recolectados para la etapa de ejemplificación se obtuvieron en la base de datos BEDCA (Base de Datos Española de Composición de Alimentos), que han sido obtenidos de distintas fuentes que incluyen laboratorios, industria alimentaria y publicaciones científicas o valores directamente calculados.

Los métodos estadísticos

Los métodos estadísticos implementados para describir e interpretar las relaciones entre los datos se basan principalmente en las técnicas de regresión lineal y el análisis de componentes principales, ambas agrupadas en la estadística multivariante.

Los métodos teóricos

Los métodos aplicados implícitamente a lo largo de este trabajo, que permitieron el desarrollo de la teoría científica y la interpretación conceptual de los datos empíricos encontrados, fueron el Análisis y síntesis, la Inducción y deducción y el análisis histórico desde el enfoque sistémico. Cada uno de estos métodos cumple funciones gnoseológicas determinadas, por lo que en el proceso de realización de la investigación se complementan entre sí.

4. INTRODUCCIÓN AL ANÁLISIS DE COMPONENTES PRINCIPALES

4.1. Breve historia del PCA

Según la literatura, puede decirse que la primera publicación científica relacionada con el PCA fue realizada por Pearson (1901) en *Philosophical Magazine*. Para un conjunto de datos dado, éste hacía hincapié en las propiedades de modelado del espacio de mejor ajuste. La definición propuesta por el autor estaba desarrollada en el contexto al Análisis de regresión lineal, dado que su tratamiento encontraba la línea de mejor ajuste de los datos, por medio de los mínimos cuadrados, en el plano de mejor representación.

Más tarde, en los años treinta, Hotelling (1933) presentó en la revista *Journal of Educational Psychology,* la idea de tomar nuevas variables conocidas como componentes principales, combinaciones lineales de las originales, haciendo hincapié en la variabilidad absorbida por ellas. Evidenciaba además la relación existente entre las cargas de las componentes principales y los valores propios de la matriz de covarianzas, obtenidos por medio de los Multiplicadores de *Lagrange*. Mostró cómo las cargas asociadas a las componentes principales eran justamente los vectores propios de la matriz de covarianzas.

En relación con los primeros acercamientos al PCA, Bro y Smilde (2014) mencionaron la existencia de una diferencia conceptual fundamental entre los enfoques de Pearson y Hotelling, importante de entender. En el enfoque de Hotelling, las componentes principales definen nuevos ejes. Debido a que son tomadas en su dirección específica, la cantidad de varianza explicada por las componentes está dada en orden decreciente; es decir, la primera componente es la encargada de explicar la mayor variabilidad, la segunda componente es la segunda con mayor variabilidad sin contar la primera y así sucesivamente (esto es conocido como **propiedad de eje principal**). Por otro lado, el enfoque de Pearson define un subespacio de proyección, más importante que los ejes como tal, pues simplemente dichos ejes están dados como la base del subespacio (Berge & Kiers, 1997).

En la misma década que Hotelling, Eckart y Young (1936) publicaron un primer documento sobre el método de Descomposición en Valores Singulares (*Singular Value Descomposition*, SVD) en la revista *Psychometrika*. Esta descomposición permitía

descomponer una matriz de datos en tres matrices diferentes. La primera de ellas con los vectores propios ortogonales a las filas de la matriz original, una segunda matriz diagonal con los valores singulares de la matriz original, y una tercera con los vectores propios ortogonales a las columnas de la matriz original. Esta descomposición abrió las puertas a la continuación de estudios del PCA, pues su aplicación en la técnica supuso un gran paso hacia adelante en los algoritmos de solución.

Adicionalmente a los documentos base de Pearson o Hotelling, algunos autores consideraban la existencia de otros documentos seminales. Este fue el caso de Rao (1964), quien sugirió que Frisch (1929) adoptó un enfoque similar al de Pearson. A su vez, Jolliffe (2002) mencionó que una nota al pie en el artículo de Hotelling (1933) permitía deducir que Thurstone en 1931 estaba trabajando en un área de estudio similar a la de éste. Aun así, este documento, citado también por Bryant y Atchley (1975) está más relacionado con el Análisis Factorial en lugar de con el PCA.

Durante los 25 años siguientes a los documentos de Pearson y Hotelling, como puede observarse en la Figura 2, aparecieron una pequeña cantidad de trabajos de diferentes aplicaciones del PCA. Los propios autores afirmaban que la ejecución manual de la técnica, aplicada a una matriz de datos inicial, era muy costosa.

El auge de su aplicación vino relacionado con la evolución de la informática. Jolliffe (2002) afirmó que, debido al gran poder computacional que los algoritmos de solución del PCA requerían cuando las matrices de datos eran grandes (principalmente con un elevado número de variables), la expansión de su uso como técnica de reducción de la dimensión coincidió con el auge de los ordenadores. En la Figura 2 se corrobora la teoría expuesta por Jolliffe (2002).

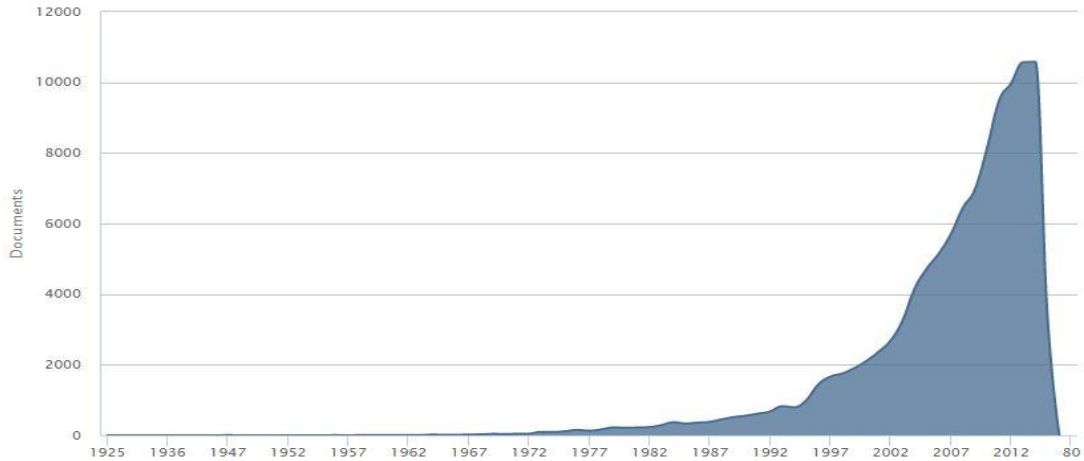

Figura 2. Documentos por año de publicación

Fuente: (ELSEIVER, 2015)

Este mismo autor indicó la aparición de cuatro grandes trabajos en el inicio de la expansión del interés por el PCA, que han llegado a ser importantes referencias dentro de este campo. El primero de ellos, escrito por Anderson (1963), es el más teórico de los cuatro. En él se discute el muestreo de distribuciones asintóticas de coeficientes y varianza de las componentes principales, construidas bajo los principios del trabajo de Girshick (1939). El segundo es el artículo de Rao (1964), que se destaca por el gran número de ideas concernientes al uso, la interpretación y las extensiones del PCA. El tercero es el documento de Gower (1966), el cual discutía los puntos de unión entre el PCA y otras técnicas estadísticas, además de proporcionar una mejor interpretación de la técnica desde un punto de vista geométrico. El cuarto se corresponde con Jeffers (1967). Éste dio un gran ímpetu al lado práctico de la técnica mediante la comparación de dos casos de estudio, en los que el uso del PCA iba más allá de una simple reducción de dimensiones.

La investigación en el tema está aún abierta y con mucho futuro por delante, como se ilustra en las Figura 2 y Figura 3. La herramienta SCOPUS permite identificar 111.673 documentos publicados entre los años 1929 y principios del 2015, que incluyen las palabras *Principal Component Analysis* en su título, resumen o palabras clave. De estos, 10.579 fueron publicados en el año 2014.

Es preciso resaltar que los tres primeros desarrollos en el PCA se expusieron, en los años 30 del siglo pasado, en revistas de psicología, intentando reivindicar la misma importancia de la medición cuantitativa de la inteligencia humana y la personalidad, como del interés contemporáneo en los genes, el procesamiento de señales, el análisis de las grandes masas de datos climáticos, financieros o los generados por Internet (Trendafilov, 2014).

Sin embargo, el PCA es una técnica no supervisada que no tiene en cuenta el objetivo de los datos, lo que le da una gran versatilidad y ha facilitado su uso en tantas y diversas áreas del conocimiento. En la Figura 3 se evidencia la implementación del PCA sobre todo en las áreas de Ingeniería, Ciencias Agrícolas, Biológicas y Medicina (ELSEIVER, 2015).

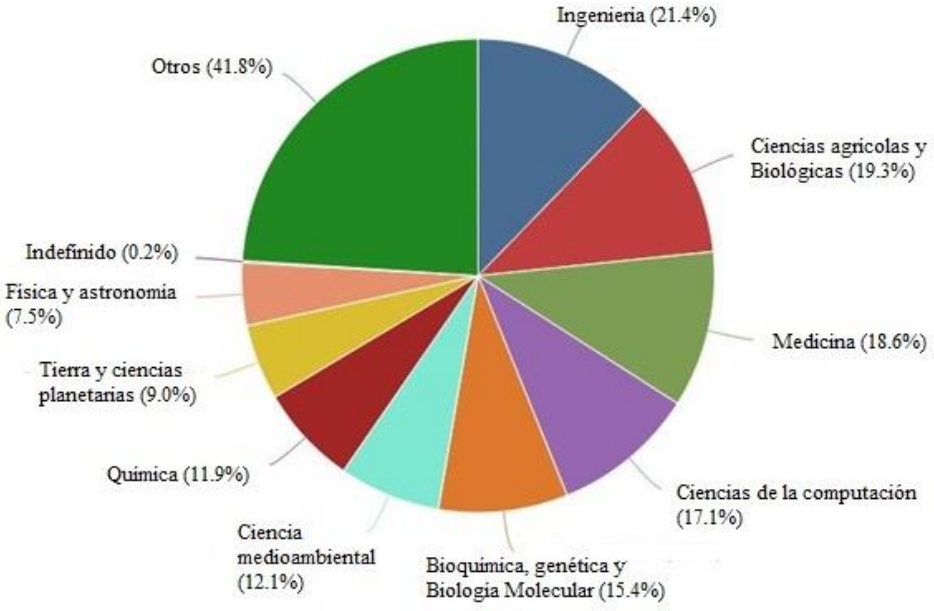

Figura 3. Documentos por área de conocimiento

Fuente: (ELSEIVER, 2015)

4.2. Definición del PCA

Al recoger información de una muestra en una matriz datos, lo más frecuente es considerar el mayor número posible de variables, con el fin de capturar la mayor parte de información, por desconocimiento del comportamiento de la población o simplemente para un uso exploratorio. Sobre todo hoy en día, con la necesidad de analizar los *big data*, concepto que hace referencia a las matrices de gran escala generadas principalmente por los sistemas informáticos. Sin embargo, al tomar un conjunto grande de variables, dadas las interacciones entre ellas, se hace aún más evidente la dificultad de visualizar el comportamiento de la muestra.

Queda con ello clara la importancia de la reducción del número de variables, procurando capturar la mayor cantidad de información posible. El concepto de mayor cantidad de información se relaciona con el de mayor variabilidad o varianza absorbida. Cuanto mayor sea la variabilidad absorbida de los datos, se considera que existe una mayor cantidad de información.

La reducción de la dimensión, es decir, poder describir con precisión los valores de un conjunto de p variables por medio de un pequeño subconjunto $r < p$ de ellas, con una pérdida pequeña de información, se convierte en un problema central en el análisis de datos multivariantes.

El Análisis de Componentes Principales es una de las técnicas multivariantes más conocida y eficaz para reducir la dimensión de una matriz de datos de alta dimensión. Es utilizada extensamente en el procesamiento de datos (Jolliffe, 2002).

Como es sabido, en el campo de la Estadística Descriptiva existen dificultades para ajustar modelos en presencia de colinealidad, dado que entre los supuestos de estos modelos se supone que las variables son independientes entre sí, lo cual comúnmente no sucede en la práctica. A esto se une el hecho de que, aunque el coeficiente de correlación sea muy bajo o casi nulo entre las variables de unos datos en estudio, no se garantiza que ambas variables no tengan otro tipo de asociación, pues el coeficiente de correlación sólo está midiendo el grado de asociación lineal.

Autores como Bro y Smilde (2014) consideran que en el análisis univariante, se descuidan explícitamente las covariaciones entre variables, lo que puede dar lugar a ignorar características importantes de las unidades muestrales.

A partir de esto, el PCA se define como una técnica para aprovechar las relaciones existentes entre las variables, desapareciendo el problema de la colinealidad, evitando perder información y pudiendo así explicar mejor la variabilidad de los datos.

La obtención de cada una de las compontes principales (*Principal component*s, PCs) puede realizarse por varios métodos alternativos:

1. Buscando el subespacio de mejor ajuste por el método de los mínimos cuadrados (minimizando la suma de cuadrados de las distancias de cada punto al subespacio) (Pearson, 1901).
2. Buscando aquella combinación lineal de las variables que maximiza la variabilidad (Hotelling, 1933).

Según Zou y Hastie (2005) El éxito del PCA se debe a las siguientes propiedades óptimas:

1) Las componentes principales capturan secuencialmente la máxima variabilidad entre las columnas de X, lo que garantiza que haya una mínima pérdida de información.
2) Las componentes principales son no correlacionadas, por lo que se puede hablar de una componente principal sin hacer referencia a otras. El PCA permite transformar las variables originales, en general correlacionadas, en nuevas variables no correlacionadas, facilitando la interpretación de los datos

Sin embargo, el PCA sufre del hecho de que cada componente principal es una combinación lineal de todas las variables originales, por lo que es a menudo difícil interpretar los resultados.

4.3. Entendiendo la combinación lineal

Bro y Smilde (2014) muestran cómo a partir de dos variables con una correlación claramente observable, la representación gráfica de ellas puede ignorar una fuerte relación entre ambas. Es por ello que estos autores mencionan la posible utilización de la media o suma de ambas variables como una nueva variable en sustitución de las originales. Esta reemplazaría las dos originales de forma que no se perdería información, ya que siempre sería posible pasar del promedio o la suma, al valor original de ambas variables.

Por ejemplo, si las variables estuvieran relacionadas directamente con un peso de 0.5 cada una, como sucede en un promedio común, se podría crear un vector unitario que

represente la nueva variable (combinación lineal de las iniciales), pudiendo obtenerse las coordenadas de dicho punto.

Si se aplica esto al caso en que se desea obtener la combinación lineal de más de dos variables, que estén correlacionadas en diferente medida, el PCA proporciona una solución a dicho problema: proporciona los pesos necesarios que permiten obtener la nueva variable que mejor explica la variación en el conjunto de datos en un cierto sentido. Esta nueva variable, incluidos los pesos que la definen, recibe el nombre de primera **componente principal**.

La combinación lineal (el nombre formal que recibe dicho promedio ponderado) es la forma de combinar las variables originales de manera lineal, recogiendo la información que estas proporcionan. En el caso del PCA, estas nuevas variables reciben el nombre de componentes principales.

Sea X una matriz de tamaño $n\,x\,p$, donde n es el número de objetos, observaciones o muestras y p el número de variables. Si las columnas de esta matriz; es decir, cada una de las variables, se denotan por vectores x_j , $j = 1, \dots, p$, se puede definir el nuevo vector y como la combinación lineal de todas las x_j, el cual es ahora un nuevo vector en el mismo espacio que las variables.

Se llama combinación lineal de las variables originales x_j a:

Ecuación 1

$$y = v_1 \times x_1 + \cdots + v_p \times x_p$$

donde $v = (v_1, \dots, v_p)$ es un vector de tamaño $p \times 1$ de pesos ponderados, formado por los elementos v_j $(j = 1, \dots, p)$ y conocido como **vector de cargas**. Es importante tener en cuenta que su dimensión o tamaño es $n \times 1$. Asumiendo que X es una matriz centrada por columnas, la Ecuación 1 se puede reescribir de forma matricial, así:

Ecuación 2

$$\boldsymbol{y = Xv}$$

PCA como modelo

Ahora bien, la matriz X es la que contiene la información relevante de los datos iniciales. Por ello, parece razonable capturar tanta información (variabilidad) como sea posible de ella en la componente principal y. Si la variabilidad expresada en y es alta, dicha combinación lineal resumirá de forma adecuada la información de las variables x_j. De esta forma, las p variables de X pueden resumirse en una sola, y, que retiene la mayor

parte de la información relevante de X. Sin embargo, si el porcentaje de variación explicada por una única componente principal es demasiado pequeño, quiere decir que la combinación lineal de las variables x_j a través del vector de cargas no es suficientemente buena para explicar los datos. Es decir, una componente no es suficiente para explicar la variabilidad de X. Será entonces necesario seleccionar un mayor número de componentes principales (PCs); esto es, un mayor número de nuevas variables. Usualmente, se escogen las primeras r PCs ($r \ll \min(n, p)$) para representar los datos, consiguiendo una gran reducción de la dimensión. Si la nube de los n puntos que representan las filas de la matriz X está contenida en un subespacio de dimensión r, será posible reconstruir las posiciones de los n puntos a partir de esos nuevos r ejes.

Siguiendo con la definición vista, cada una de las nuevas variables de X, $y_j, j \in \{1, \dots, p\}$, llamadas componentes principales, se pueden definir como una combinación lineal de las variables observadas x_j ponderadas por un vector de cargas v_j, así.

$$y_j = v_{1j}x_1 + \cdots + v_{pj}x_p$$

De esta forma, las coordenadas del individuo i en la j-ésima PC vienen dadas por:

$$y_{ij} = x_{i1}v_{1j} + \cdots + x_{ip}v_{pj}$$

En forma matricial:

$$Y_{nxp} = X_{nxp}\, V_{pxp}$$

Y_{nxp} es la matriz que contiene las puntuaciones de cada uno de los individuos sobre las componentes y V_{pxp}, conocida como matriz de cargas, es la matriz que contiene los coeficientes de las combinaciones lineales en columnas.

Ahora, como se ha dicho, para lograr la reducción de la dimensionalidad se eligen las primeras r componentes principales, con lo que:

$$Y_{nxr} = X_{nxp}\, V_{pxr}$$

4.3.1. Modelo de maximización de la varianza

Este es el enfoque de cálculo de las componentes principales proporcionado por (Hotelling, 1933). Con la definición de este modelo se buscan componentes principales que sean combinación lineal de las variables originales, no correlacionadas y que absorban la máxima información posible de matriz original X. Esta información viene dada por la varianza explicada por las nuevas variables (PCs). Usualmente, la variabilidad de Y se ha medido como la trata de la matriz de covarianzas de Y.

Como la intención del investigador es: dadas unas variables originales (x_j), encontrar unas nuevas variables (y_j recogidas en la matriz Y) que recojan la mayor parte de la información proporcionada por las variables iniciales (recogida en la matriz de covarianzas de las nuevas variables; es decir, por la matriz de covarianzas de la matriz Y), el problema se traduce en maximizar las varianzas que conforman la matriz de covarianzas eligiendo unas cargas $v_1, \dots, v_p$ óptimas:

$$\begin{gathered}\max\ \mathrm{tr}(\mathrm{Cov}_{YY}) \\ \text{s.a.}\ \ V^T V = I\end{gathered}$$

La restricción $V^T V = I$ garantiza que las PCs sean ortogonales.

Además, se sabe que si X es una matriz estandarizada, es decir, centrada por filas, $X^T 1_n = 0_p$, y por columnas (variables), de longitud uno, $\mathrm{diag}(X^T X) = 1_p$, se puede demostrar que los resultados obtenidos de la matriz de covarianzas son proporcionales a su estimación, así:

$$S = \mathrm{Cov}_{XX} = \frac{1}{n} X^T X \leftrightarrow \hat{S} = X^T X$$

Se sabe además que la variabilidad de la matriz Y viene dada por $||Y||^2$, que por definición de la norma de Frobenius,

$$||Y||^2 = \mathrm{tr}(YY^T)$$

Así que se tiene:

Ecuación 3

$$\begin{gathered}\max \mathrm{tr}(\mathrm{Cov}_{YY}) \\ \text{s.a.}\ V^T V = I\end{gathered} \leftrightarrow \begin{gathered}\max \mathrm{tr}(YY^T) \\ \text{s.a.}\ V^T V = I\end{gathered} \leftrightarrow \begin{gathered}\max ||Y||^2 \\ \text{s.a.}\ V^T V = I\end{gathered}$$

Además, se puede reescribir la matriz de covarianzas de Y en términos de la matriz de covarianzas de la matriz original:

Ecuación 4

$$\mathrm{Cov}_{YY} = \frac{1}{n} Y^T Y = \frac{1}{n} (XV)^T XV = \frac{1}{n} V^T X^T XV = V^T \frac{1}{n} X^T XV = V^T SV$$

El problema de la Ecuación 3 se reformula a partir de la Ecuación 4, así:

Ecuación 5

$$\begin{gathered}\max ||Y||^2 \\ \text{s.a.}\ V^T V = I\end{gathered} \leftrightarrow \begin{gathered}\max V^T SV \\ \text{s.a.}\ V^T V = I\end{gathered} \leftrightarrow \begin{gathered}\max V^T \hat{S} V \\ \text{s.a.}\ V^T V = I\end{gathered} \leftrightarrow \begin{gathered}\max V^T X^T XV \\ \text{s.a.}\ V^T V = I\end{gathered}$$

Se puede demostrar cómo el óptimo v (estandarizado) se corresponde con el primer vector propio (el asociado al mayor valor propio) de la matriz de covarianzas $\frac{1}{n}X^TX$ o de la correspondiente estimación X^TX.

4.3.2. Modelo a través de la Descomposición en Valores Singulares de X

Gracias a los desarrollos de Eckart y Young (1936), se sabe que, como el PCA busca las combinaciones lineales de las variables originales, las PCs pueden ser calculadas a través de la Descomposición en Valores Singulares de la matriz de datos original. La SVD de X es:

$$X = UDV^T$$

donde U es una matriz unitaria de tamaño $n \times n$. Las n columnas de U son los vectores singulares de izquierda de X , que a su vez son los vectores propios de XX^T. La matriz D es una matriz diagonal de tamaño $n \times p$, que contiene las raíces cuadradas de los valores propios no nulos de XX^Ty X^TX. Es una matriz diagonal, $D = \mathrm{diag}(\lambda_1, \lambda_2, \dots, \lambda_p)$, con $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ los valores singulares de X. La matriz V^T es la matriz traspuesta de V, que es una matriz de tamaño $p \times p$. Las p columnas de V son los vectores singulares de derecha de X, que a su vez son los vectores propios de X^TX.

Las componentes principales de X pueden definirse como:

$$y = x_j v_j = u_i \lambda_i$$

En la primera igualdad las componentes principales son calculadas como vectores y_j generados por la proyección de los datos de entrada en la dirección de los vectores columna v_j. En la segunda igualdad las componentes se calculan como vectores y_i generados por la proyección de los vectores fila u_i, escalados por los valores singulares λ_i.

En forma matricial:

$$Y = UD = XV$$

La varianza de cada componente principal viene dada por:

$$var(Y_i) = \lambda_i^2 / n$$

Como los valores singulares se presentan en orden decreciente, la primera componente principal de X, y_1, tiene la mayor varianza entre todas las combinaciones lineales

normalizadas de las columnas de X, y_2, tiene la segunda mayor varianza y así sucesivamente.

4.3.3. Modelo de minimización del error de reconstrucción

Otra forma de enfocar el problema del PCA es como un modelo de minimización del error de reconstrucción de X; es decir, como un problema de mínimos cuadrados (Pearson, 1901). La definición de la regresión puede utilizarse para juzgar la calidad de la síntesis de información en Y, pero también destaca una interpretación importante del PCA, donde este puede ser visto como una actividad de modelización.

Se desea estimar la matriz de cargas que definen las componentes principales, minimizando el error de reconstrucción; esto es, minimizando la diferencia entre los datos de la matriz original, y los que se obtendrían proyectando las nuevas variables (PCs) en el espacio original:

Ecuación 6

$$\min \left\| X - \hat{X} \right\|^2$$
$$\text{s. a. } V^T V = I$$

Siendo $\hat{X}$ la matriz de coordenadas de las proyecciones sobre el subespacio de las componentes, en el sistema de referencia original. Si $Y = X\,V$, entonces, las nuevas coordenadas $\hat{X}$ de las proyecciones son:

$$\hat{X} = YV^T = XVV^T$$

El problema de optimización de la Ecuación 6, queda reescrito, así:

Ecuación 7

$$\min \| X - XVV^T \|^2$$
$$\text{s. a. } V^T V = I$$

Los valores de V, se pueden establecer a partir de la Ecuación 7 haciendo mínimo el error entre la matriz original $X\ y$ y su estimada $\hat{X}$. Es decir, resolviendo un problema de mínimos cuadrados.

Se calcula de forma sucesiva el subespacio generado por los r vectores $v_1, v_2, \dots, v_r$, ortogonales dos a dos, que conforman el subespacio de mejor ajuste a la matriz de datos inicial. A través de los multiplicadores de Lagrange, puede demostrarse que estos vectores se corresponden con los vectores propios de la matriz $X^T X$ asociados a los valores propios $\lambda_1, \lambda_2, \dots, \lambda_r$ de la matriz respectivamente, de forma que $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r$.

Por tanto, las componentes principales, combinación lineal de las variables originales, poseen varianza progresivamente decreciente, pues el mayor valor propio es el que mayor varianza recoge (la varianza absorbida por cada eje viene dada por el cuadrado de los valores propios asociados a los vectores directores que los conforman).

4.3.1. Equivalencia entre los modelos

La solución obtenida buscando el subespacio de mejor ajuste es equivalente a la obtenida maximizando la información absorbida por las componentes principales; esto es, ambos problemas de optimización son equivalentes.

Es lógico que:

$$X = \hat{X} + E$$

Y, por tanto,

$$E = X - \hat{X}$$

Sea $||X||^2$ la variabilidad de los datos originales, $||\hat{X}||^2$ la variabilidad del subespacio estimado y $||X - \hat{X}||^2$ la variabilidad del error de reconstrucción.

Para probar que ambos problemas de optimización son equivalentes, dado que $||X||^2$ es un valor fijo, basta con demostrar la igualdad:

$$||X||^2 = ||X - \hat{X}||^2 + ||\hat{X}||^2$$

Como $||X||^2$ no varía, cuanto más grande sea la variabilidad del espacio estimado, $||\hat{X}||^2$, menor será la variabilidad del error de reconstrucción $||X - \hat{X}||^2$.

Sea $||X||^2$ la norma de Frobenius de X, por definición, $||X||^2 = tr(XX^T)$. Entonces:

Ecuación 8

$$\|X - \hat{X}\|^2 + \|\hat{X}\|^2 = \mathrm{tr}\left((X - \hat{X})(X - \hat{X})^T\right) + \mathrm{tr}\left(\hat{X}\hat{X}^T\right)$$

$$= \mathrm{tr}\left((X - XVV^T)\left(X^T - (XVV^T)\right)^T\right) + \mathrm{tr}(XVV^T(XVV^T)^T)$$

$$= \mathrm{tr}\left((X - XVV^T)(X^T - VV^TX^T)\right) + \mathrm{tr}(XVV^TVV^TX^T)$$

$$= \mathrm{tr}(XX^T - XVV^TX^T - XVV^TX^T + XVV^TVV^TX^T) + \mathrm{tr}(XVV^TVV^TX^T)$$

Dado que $V^TV = I$, y aplicando las propiedades de la traza de una matriz:

$$\|X - \hat{X}\|^2 + \|\hat{X}\|^2 = \mathrm{tr}\left(XX^T - 2XVV^TX^T + XVV^TX^T\right) + \mathrm{tr}\left(XVV^TX^T\right)$$

$$= \mathrm{tr}(XX^T - XVV^TX^T) + \mathrm{tr}(XVV^TX^T)$$

$$= \mathrm{tr}(XX^T) - \mathrm{tr}(XVV^TX^T) + \mathrm{tr}(XVV^TX^T)$$

$$= \mathrm{tr}(XX^T) = ||X||^2$$

Con ello, queda demostrada la igualdad y, consecuentemente, la equivalencia de ambos problemas de optimización, ya que si $||X||^2$ es fijo, si $||\hat{X}||^2$ aumenta, es obvio que $||X - \hat{X}||^2$ disminuirá. Por tanto, si el valor de $||\hat{X}||^2$ se hace máximo, el valor de $||X - \hat{X}||^2$ es mínimo.

Es claro que:

$$\|X - \hat{X}\|^2 = \mathrm{tr}\left(\left(X - XVV^T\right)\left(X^T - \left(XVV^T\right)\right)^T\right) = \mathrm{tr}\left(XX^T - XVV^TX^T\right)$$

$$= \mathrm{tr}\left(XX^T\right) - \mathrm{tr}\left(XVV^TX^T\right)$$

Como $\mathrm{tr}(XX^T)$ es un valor fijo y no influye en la minimización del error, minimizar $\|X - \hat{X}\|^2$ equivale a maximizar:

$$\max\ tr(XVV^TX^T) = \max\ tr(V^TX^TXV) = \max\ tr(Y^TY) = \max\ \|Y\|^2$$

donde $\|Y\|^2$ se corresponde con la variabilidad de las componentes principales.

Por tanto:

$$\begin{matrix}\max \|Y\|^2 \\ \text{s. a. } V^TV = I\end{matrix} \quad \leftrightarrow \quad \begin{matrix}\min \left\|X - \hat{X}\right\|^2 \\ \text{s. a. } V^TV = I\end{matrix}$$

4.4. Visualización e interpretación de los resultados del PCA

La interpretación de las componentes principales se favorece representando las proyecciones de las observaciones (puntuaciones) sobre un espacio de dimensión dos, definido por ejes que se corresponden con parejas de componentes principales.

El PCA se interpreta considerando las magnitudes de las cargas en las componentes (la matriz V), pues estas indican la intensidad con la que cada una de las variables originales contribuye a la formación de la PC.

4.4.1. Visualización e interpretación de las puntuaciones

Las puntuaciones de los individuos en las PCs pueden ser visualizadas mediante gráficos de dispersión, consiguiendo así la interpretación en base a las nociones esenciales de dichos gráficos: los individuos que están cerca en términos de distancia,

serán similares en lo que representen las componentes definidas por los vectores de carga. Además, la proyección de cualquier observación sobre una componente se corresponderá directamente con el valor de la componente para dicha observación.

Si (y solo sí) las dos componentes representan la totalidad o casi la totalidad de la variación en los datos, es posible evaluar las similitudes y diferencias entre las muestras en cuanto a los datos en bruto. Si dos componentes explican la totalidad o la mayor parte de la variación en los datos, un gráfico de dispersión de las puntuaciones reflejará directamente distancias en términos de los datos originales, si los resultados se muestran en la misma escala.

4.4.2. Visualizar e interpretar las cargas

Las cargas de la matriz V se definen como el valor que representa cada variable en una PC específica, pues son las encargadas de dar forma a la combinación lineal de la que resulta cada PC en particular. Dado que cada componente principal proviene de la combinación lineal de las variables originales, la matriz de cargas de las componentes (que cuantifican la correlación entre las variables originales y las componentes principales) posee un papel fundamental en el análisis de su significado (Trendafilov, 2014).

La cantidad de varianza explicada por las PCs (o ejes) puede interpretarse como la cantidad de información recogida en el plano de representación. Las cargan pueden observarse en gráficos de líneas, aunque también es factible hacer gráficos de dispersión. El gráfico de dispersión es a menudo útil para encontrar patrones de variación.

En cuanto a la nube de puntos, las distancias sólo se conservan en el gráfico de dispersión, si las dos cargas se representan en la misma escala.

Los factores de carga pueden ser considerados a su vez como los coeficientes resultantes de una ecuación de regresión múltiple, donde las variables originales serían las variables dependientes y las componentes las independientes. Como estas últimas se definen no correlacionadas, los coeficientes no dependen uno del otro y representan la contribución única de cada componente o la correlación de la variable con cada una de las componentes. Será importante tener en cuenta tanto la magnitud de la carga cómo los signos de ellas, para conocer la forma que tienen de relacionarse. La interpretación será más sencilla cuando muchas de las cargas que conformen una componente sean nulas y el resto cercanas a uno, pues así podrá atribuirse un

significado a la componente. Objetivo que se logrará con el Análisis de Componentes Principales *Sparse*.

4.4.3. Visualización a través de Biplots

Análogamente, existe una forma de visualizar conjuntamente las filas y las variables de X por medio de los **Biplots**. Esta técnica fue desarrollada originalmente por Gabriel (1971), con contribuciones posteriores de Galindo (1986), Gower (1995), entre otros.

Los métodos Biplot son una representación gráfica a baja dimensión de una matriz de datos de datos multivariantes X_{nxp} (n individuos y p variables) (Gabriel 1971).

Desde el punto de vista algebraico, el Biplot se basa en el mismo principio sobre el que se sustentan la mayoría de las técnicas factoriales de reducción de la dimensión: la Descomposición en Valores Singulares (SVD) de una matriz. La diferencia fundamental es que, en el Biplot de Gabriel (1971) se trata de reproducir el dato y se incorpora una representación conjunta de individuos y variables.

Utilizando la SVD se aproxima la matriz de datos, para luego realizar un gráfico en baja dimensión cuya interpretación se basa en las propiedades geométricas del producto escalar entre vectores fila (denominados marcadores fila) y vectores columna (denominados marcadores columna), de tal forma que dicho producto reproduzca aproximadamente cada elemento de la matriz de datos X.

Las dos factorizaciones Biplot más importantes propuestas por Gabriel (1971) fueron denominadas: GH-Biplot y JK-Biplot. El GH-Biplot consigue una alta calidad en la representación de las columnas (variables), mientras que el JK-Biplot logra una alta calidad en la representación de las filas (individuos). Por lo tanto, esta representación Biplot no es simultánea para ninguno de estos dos tipos de Biplots porque la bondad de ajuste no es la misma para las filas y para las columnas de la matriz de datos.

Galindo (1986) demuestra que, con una conveniente elección de los marcadores, es posible representar las filas y las columnas simultáneamente sobre un mismo sistema de coordenadas, con una alta calidad de representación para ambos marcadores. Galindo denomina a este tipo de Biplot, HJ-Biplot.

4.4.4. Métodos de rotación

El PCA se ha convertido en la técnica de análisis estadístico multivariante más empleada. Es por eso, como se ha mencionado, bien conocido uno de los problemas que presenta: la interpretación de las componentes principales.

Tras la ejecución de la técnica, la dificultad en la interpretación de las componentes resultantes ha llevado a la aplicación de técnicas de rotación sobre las componentes retenidas, con el fin de encontrar una estructura en los datos más simple que facilite su interpretación.

Para el estudio de una matriz de datos, el PCA es realmente útil cuando la interpretación de las PCs es sencilla. Sin embargo, esto no siempre se da, a pesar de que se alcance una dimensión reducida en k PCs. Además, la interpretación se complicará si en la matriz de datos de estudio el número de variables es muy grande.

Normalmente, las matrices de cargas resultantes de técnicas de reducción de la dimensionalidad no son fáciles de interpretar. A pesar de que en la práctica una forma de eliminar cargas es tomar como nulas aquellas cuyo valor absoluto queda por debajo de un umbral, el enfoque más antiguo para resolver este problema es el uso de las técnicas de rotación. Estas técnicas consisten en hacer girar los nuevos ejes, obtenidos tras la reducción de dimensionalidad, hasta conseguir aproximarlos lo máximo posible a las variables que cargan en ellos.

Inicialmente, las técnicas de rotación se desarrollaron en el Análisis Factorial (*Factorial Analysis*, FA).

Según Thurstone (1935), que desarrolló la teoría y método de las rotaciones factoriales para obtener la estructura más sencilla en el FA, la finalidad al aplicar una rotación sobre los ejes es encontrar una solución que siga el **principio de estructura simple.** La búsqueda de esta estructura simple se introdujo debiendo verificarse tres reglas deseables (Thurstone, 1935), que más tarde se completaron en (Thurstone, 1947) en las siguientes:

- Cada fila de la matriz factorial (de cargas) debe contener al menos un 0.
- Si existen r factores comunes, cada columna de la matriz factorial debe contener al menos r ceros.
- Para cada pareja de columnas de la matriz factorial, deberán existir varias variables cuyas cargas desaparezcan en una columna, pero no en el resto.
- Para cada pareja de columnas de la matriz factorial, una larga proporción de variables deberán tener sus cargas nulas en ambas columnas cuando haya 4 o más factores.
- Para cada pareja de columnas de la matriz factorial, deberá existir un número pequeño de variables con entradas no nulas en ambas columnas.

Jolliffe (2002) adaptó esta teoría al PCA, para hacer las componentes más interpretables.

Con todo ello, se consigue distinguir grupos de variables observadas, cada uno de los cuales está altamente correlacionado con distintas componentes. Examinando dichas características, se pueden encontrar rasgos comunes que identifiquen las nuevas variables latentes, desvelando así las relaciones entre las variables originales.

Este concepto de estructura simple en la matriz se intentó implementar en un gran número de métodos de rotación y se introdujeron métodos analíticos que suplían la desventaja de la subjetividad en el primer enfoque gráfico de las técnicas de rotación. Ahora bien, era obvia la dificultad de encontrar una única expresión matemática que englobase todas las características de estructura ideal del conjunto de datos tras un método de rotación. Por ello, el número de técnicas que se han ido desarrollando a lo largo del tiempo es tan grande como la dificultad de implementar en ellas todas las características aconsejadas por Thurstone (1947), pues ninguna es capaz de proporcionar una solución satisfactoria para cualquier tipo de problema.

Siguiendo la teoría de Abdi (2010), se distinguen dos tipos principales de rotación: **rotación ortogonal** (si se requiere que cada par de los nuevos ejes rotados sean ortogonales, puesto que se formarán con un ángulo recto entre ellos y **rotación oblicua** (cuando no son requeridos ejes ortogonales). La varianza total que presenta el modelo completo rotado es la misma que en el caso inicial, aunque cambie la matriz de cargas. Quedará redistribuida la inercia absorbida por los nuevos ejes rotados: no están ordenados de acuerdo con la información que contienen, cuantificada a través de su varianza.

Con la única finalidad de no hacer el estudio más extenso y dado que en la práctica son las rotaciones más utilizadas por sus propiedades favorables, en este trabajo únicamente se considerará el estudio de las técnicas de rotación ortogonales.

Rotaciones ortogonales

De manera resumida, los problemas de rotación encargados de la búsqueda de una estructura simple en los datos no son más que problemas de optimización con restricciones de ortogonalidad (u oblicuas) impuestas.

Las rotaciones ortogonales se caracterizan principalmente por su simplicidad, dado que las cargas representan las correlaciones entre los factores y las variables de partida, cosa que no ocurre en las rotaciones oblicuas. Su idea es maximizar la varianza de los

cuadrados de las cargas factoriales, logrando así que se dispersen los valores al máximo, aumentando los mayores y disminuyendo los más pequeños.

Dada una matriz de cargas V, se desea encontrar una matriz ortogonal Q tal que la nueva matriz de cargas $\hat{V} = VQ$ defina unas componentes que tengan una estructura más simple. La rotación ortogonal viene especificada por una matriz de rotación Q cuyas filas se corresponden con los ejes originales, y las columnas con los nuevos ejes rotados. El elemento $q_{m,n}$ de la matriz se corresponderá con el coseno del ángulo formado entre el eje original m y el nuevo eje rotado n; esto es, $q_{m,n} = \cos\theta_{m,n}$. La matriz Q es una matriz ortogonal, verificándose que $Q^T Q = I$.

Como se muestra en (Cuadras 2014) un criterio analítico considera la función:

Ecuación 9

$$G = \sum_{k=1}^{r} \sum_{k \neq j=1}^{r} \left[\sum_{i=1}^{p} v_{ij}^2 v_{ik}^2 - \frac{\gamma}{p} \sum_{i=1}^{p} v_{ij}^2 \sum_{i=1}^{p} v_{ik}^2 \right]$$

donde γ es un parámetro tal que $0 \leq \gamma \leq 1$.En función de cómo sea dicho γ, hay distintos criterios especialmente interesantes. De entre ellos, las principales rotaciones ortogonales son la rotación VARIMAX, QUARTIMAX y EQUIMAX.

Dado que la rotación VARIMAX es la más empleada en la práctica y la más conocida por sus propiedades, se recuerdan básicamente las propiedades de esta técnica para facilitar la comprensión del estudio de datos posterior.

Rotación VARIMAX (Kaiser, 1958)

La rotación VARIMAX (VARiance MAXimization) fue introducida por Kaiser (1958) en el Análisis Factorial. En ella se busca minimizar el número de variables que tienen saturaciones o cargas altas en cada factor o dimensión. Es decir, trata de que los nuevos ejes tengan unas pocas saturaciones altas y muchas casi nulas en las variables.

Este método actúa por columnas de la matriz de carga, maximizando la suma de las varianzas de las cargas factoriales dentro de cada factor (conocido como suma de simplicidades de los factores en el caso del PCA). El objetivo es buscar la existencia de factores con correlaciones altas con un número pequeño de variables y correlaciones nulas con el resto de ellas. Con todo ello, queda redistribuida la varianza de los factores. Se maximiza por tanto la dispersión de las cargas para cada factor por separado. Quizá por ello este sea el método más utilizado.

Si $\gamma = 1$ en la Ecuación 9, se habla de rotación Varimax:.

$$G = \sum_{k=1}^{r} \sum_{k \neq j=1}^{r} \left[\sum_{i=1}^{p} v_{ij}^2 v_{ik}^2 - \frac{1}{p} \sum_{i=1}^{p} v_{ij}^2 \sum_{i=1}^{p} v_{ik}^2 \right]$$

Esta cantidad será grande cuando existan pocas cargas de valores elevados y el resto cercanas a 0. Por el contrario, será pequeña cuando todas las cargas sean de magnitudes similares. El problema de la rotación VARIMAX se basa en encontrar una matriz Q ortogonal de tamaño $r\,x\,r$ de manera que dicha cantidad sea máxima (Trendafilov, 2014).

5. INTRODUCCIÓN A LOS MODELOS DE REGRESIÓN

Como se mencionó en la Introducción, el PCA se ha reformulado en varias ocasiones. La reformulación en base a la minimización del error de reconstrucción, permite definir una versión *sparse* del mismo. Para entender esta técnica, se hace necesario realizar un acercamiento a las bases teóricas que dan su origen: penalizaciones en los modelos de regresión.

5.1. Breve historia de los modelos de regresión

Desde el ajuste por mínimos cuadrados propuesto por Pearson (1901), los desarrollos que se relacionan estrechamente con el SPCA comienzan con la regresión *Ridge*. Esta fue dada a conocer con este nombre por Hoerl y Kennard (1970), aunque se cree que el primer acercamiento a esta regresión pudo ser planteado por Tikhonov (1943).

La penalización *Ridge* surge para superar la limitación de la regresión ordinaria por mínimos cuadrados para encontrar una solución única cuando hay presencia de colinealidad. Este método trata de contraer los coeficientes de la regresión hacia 0, penalizando la norma ℓ_2 de los coeficientes. Dado que la contracción no es exacta, su penalización no permite seleccionar variables.

Intentando paliar esto, se definió la regresión *Lasso* a través de programación cuadrática (Tibshirani 1996). Esta regresión implica una penalización en los coeficientes en términos de su norma ℓ_1, por lo que gana la ventaja de generar coeficientes exactamente iguales a cero. Esta característica mejora la interpretación del modelo y permite seleccionar claramente las variables que mejor explican el comportamiento del vector a explicar. Sin embargo, el número de variables seleccionadas por *Lasso* está limitado por el número de observaciones, es decir presenta limitaciones cuando el número de variables supera el número de observaciones, $p \gg n$. Este es el caso, por ejemplo, de los *microarrays,* donde hay miles de genes ($p > 1.000$) con menos de 100 muestras ($n < 100$). Para problemas de este tipo, *Lasso* puede seleccionar como mucho 100 genes ($p \leq n$), lo cual es claramente insatisfactorio en términos de interpretación. Fue entonces cuando Zou y Hastie (2005) desarrollaron la regresión *Elastic net,* combinando las penalizaciones *Ridge* y *Lasso* para superar las limitaciones de ambas, mientras conservaba sus propiedades favorables.

5.2. Regresión ordinaria

Considérese el modelo de regresión tradicional, con n observaciones y p variables explicativas o regresoras. Sea una matriz de datos de entrada X de dimensión $(n \times p)$, donde las variables explicativas del modelo son los vectores $\boldsymbol{x_j} = (x_{1j}, \dots, x_{nj})^T \in \mathbb{R}^n$ y $j \in \{1, \dots, p\}$,, asociados a las columnas de dicha matriz.

$$X = (x_1, x_2, \dots, x_p)$$

Sea $Y = (y_1, \dots, y_n)^T$ la matriz respuesta, con $y_i = (y_{i1}, \dots, y_{ip}) \in \mathbb{R}^p, i \in \{1, \dots, n\}$, que contiene los valores de las variables a explicar con el modelo de regresión lineal, a través de las variables explicativas.

Ahora bien, existen diferentes métodos para ajustar el modelo lineal al conjunto de datos dado, pero el más popular es el método de los **mínimos cuadrados** (*Ordinary Least Squares*, OLS). Con el modelo de regresión ordinaria por mínimos cuadrados se busca ajustar los datos de y_i mediante la combinación lineal de las variables x_j así:

$$y_i = \beta_o + \sum_{j=1}^{p} x_{ij}\beta_j$$

donde B es una matriz de coeficientes de tamaño pxp , dado que cada y_i es un p-vector: $B = (\beta_{\boldsymbol{1}}, \beta_{\boldsymbol{2}}, \dots, \beta_{\boldsymbol{p}})$ con $B_{\boldsymbol{j}} = (\beta_{1j}, \dots, \beta_{pj})^T$.

En forma matricial, cuando X es una matriz centrada y con norma 1, el valor Y estimado, $\hat{Y}$ es:

$$\hat{Y} = X\hat{B}$$

Como hemos dicho, el cálculo de los coeficientes de $\hat{\beta}$ tiene lugar a través del método de mínimos cuadrados, en el que los coeficientes estimados $\hat{\beta}$ se obtienen como los valores de β que minimizan la suma de cuadrados de los residuos:

$$RSS(\beta) = \|Y - \hat{Y}\|^2 = \sum_{i=1}^{n} (y_i - x_i\hat{\beta})^2$$

Con $\hat{\beta}$ el "candidato" para el parámetro β. El estimador $\hat{\beta}$ se denomina OLS de β.

En la Figura 4, se puede observar cómo se busca la función lineal de X que minimiza la suma de cuadrados de los residuos de Y.

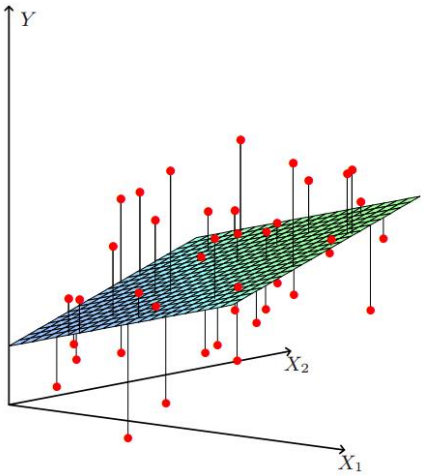

Figura 4 Mínimos cuadrados lineales ordinarios ajustados para $X \in R^2$.

Fuente: (Hastie et al. 2009)

La función $RSS(\beta)$ es una función cuadrática de los parámetros y, por tanto, su mínimo siempre existe, aunque puede no ser único. A partir de la notación matricial, es fácil caracterizar la solución (Hastie et al. 2009). Podemos reescribir:

$$RSS(\beta) = (Y - X\beta)^T(Y - X\beta)$$

Derivando respecto de β, se tiene:

$$\frac{\partial RSS}{\partial \beta} = -2X^T(Y - X\beta)$$

$$\frac{\partial^2 RSS}{\partial \beta \partial \beta^T} = -2X^T X$$

Asumiendo que X es de rango completo, y por tanto, $X^T X$ es definida positiva, establecemos la primera derivada igual a 0:

$$X^T(Y - X\beta) = 0$$

Con ello, se da lugar al siguiente teorema:

Teorema 1 *La función RSS posee un mínimo global único en $\hat{\beta}^{OLS}$ cuando $X^T X$ es no singular. Entonces, la solución del problema de regresión OLS está dado por*

$$\hat{\beta}^{OLS} = \underset{\beta}{\operatorname{argmin}} \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} x_i y_i = (X^T X)^{-1} X^T Y$$

La varianza de los coeficientes estimados mediante mínimos cuadrado es

$$Var(\hat{\beta}) = X^T X \sigma^2$$

donde σ^2 corresponde a la varianza de los errores del modelo, que se suponen independientes de X y definidos como $\varepsilon \sim N(0, \sigma^2)$.

Tras haber estimado $\hat{\beta}^{\mathrm{OLS}}$,

$$\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y = Hy$$

La matriz $H = X(X^T X)^{-1} X^T$ se conoce como matriz sombrero y sirve para definir los grados de libertad del modelo (*degrees of freedom*, df).Los grados de libertad del modelo están definidos como la traza de H:

$$\mathrm{tr}(H) = p = df_{OLS}$$

5.2.1. Selección del modelo

Como Hastie et al. (2009) mencionan, existen dos razones principales por las que normalmente no se está del todo satisfecho con las aproximaciones OLS.

- La exactitud de la predicción. Las estimaciones por mínimos cuadrados tienen a presentar un sesgo bajo, pero una gran varianza. Es por ello que, en ocasiones, la exactitud de la predicción puede verse mejorada con la contracción de algunos coeficientes hacia 0, obligando a hacerse nulos o prácticamente nulos. Se hace un sacrificio de un poco de sesgo para reducir la varianza de los valores predichos, y, con ello, se podría mejorar la exactitud de la precisión.
- Interpretación. Con un gran conjunto de variables explicativas, normalmente se desea escoger un conjunto menor de variables que recojan los mayores efectos. Se obtiene una imagen aproximada de la totalidad de la información, a cambio de perder pequeños detalles.

A continuación se describen algunos de los enfoques para realizar esta selección del subconjunto de variables, a partir de la regresión lineal. Más adelante, se discutirá el enfoque de la selección del modelo a través de los modelos de penalización.

Usualmente los problemas de regresión se resuelven seleccionando r variables explicativas, de las p variables, que generen el mínimo valor para la función $RSS\left(\hat{\beta}\right)$, pues no todos los coeficientes β_j asociados a cada una de las variables x_j resultan ser significativos para explicar Y. En otras palabras, la consideración de "demasiadas" variables en los modelos de regresión, plantea desafíos adicionales como la selección de variables relevantes en el análisis. Dicho subconjunto de variables se suele determinar por alguno de los siguientes métodos (Garside, 1965; Hastie et al., 2009):

- **Selección hacia adelante** (*Forward stepwise selection*, FST): Las r variables del subconjunto se incluyen secuencialmente al modelo, agregándose en función del coeficiente β_j más significativo, es decir, aquel que produzca la mayor mejora en el ajuste. Puede parecer que computacionalmente supone un coste alto en el momento en el que existan muchas variables explicativas candidatas; sin embargo, existen algoritmos para establecer rápidamente el siguiente candidato a añadirse al modelo. Aunque FST produce una secuencia anidada de modelos, lo cual puede parecer no óptimo, hay varias razones por las que este algoritmo puede ser preferible. Por ejemplo, computacionalmente hablando, siempre se puede calcular la secuencia de modelos, incluso cuando $p \gg n$.
- **Selección hacia atrás** (*Backward stepwise selection*, BST): Este método parte del modelo completo con todos los β_j. Las variables asociadas se eliminan secuencialmente en función de los coeficientes menos significativos, es decir seleccionando la variable explicativa que menos aporta al ajuste del modelo y eliminándola. Presenta la desventaja de que sólo puede utilizarse cuando $n > p$, mientras que FST puede ser empleado siempre. Aun así, en los casos factibles, sueles tener rendimientos similares.
- **El algoritmo *Branch y Bound* (BB)**: Consiste en una enumeración sistemática de las posibles soluciones del problema de regresión a través de la búsqueda del espacio de mejor ajuste. El conjunto de soluciones candidatas se considera como un árbol cuya raíz es el modelo completo. El algoritmo explora las ramas de este árbol, que representan subconjuntos soluciones factibles. Antes de enumerar las soluciones candidatas de una rama, la rama se compara con los límites superior e inferior estimados en la solución óptima. Se descarta si no se puede producir una solución mejor que la mejor encontrada por el algoritmo hasta el momento (Land & Doig, 1960).

5.3. Métodos de penalización

Debido a que estas técnicas clásicas de selección de variables, al mantener un subconjunto de las variables explicativas y desechando el resto, producen modelos interpretables que posiblemente tienen un menor error de predicción que el modelo completo. Sin embargo, debido a que son procesos discretos (las variables son o bien retenidas o bien descartadas) a menudo presentan una alta variabilidad, y por lo tanto no reducen el error de predicción del modelo completo ni producen estabilidad en la selección (Breiman, 1996; Hastie et al., 2009).

Es por ello que se introducen las técnicas de regularización o penalización: procedimientos continuos y menos variables, que se presentan como alternativas prometedoras. Estos métodos buscan convertir problemas *mal condicionados*, debido a la no unicidad de la solución o a una inexistente selección de variables, en problemas *bien condicionados.* Además, introducen el marco para tratar aquellos problemas en los *que* $p \gg n$, que tanto empiezan a darse en la actualidad.

Se sabe que en un problema de regresión ordinaria, los coeficientes de la regresión pueden darse de forma única como $\hat{\beta} = (X^T X)^{-1} X^T Y$, siempre y cuando exista $(X^T X)^{-1}$. Ahora bien, cuando $p \gg n$ esto no ocurre y el vector de coeficientes no es identificable sin supuestos o restricciones adicionales, ya que pueden existir infinitas soluciones del sistema de ecuaciones $X^T X\beta = X^T Y$. La regularización en estas técnicas permitirá disminuir el riesgo de sobreajuste mediante el control de la complejidad del modelo.

En el contexto de los modelos lineales, el problema de determinar la complejidad óptima del modelo se traslada de encontrar el número apropiado de funciones base (proceso discreto) a uno en el cual debe determinarse un valor adecuado de un parámetro de regularización λ (proceso continuo). Este parámetro de complejidad controlará la importancia dada a la penalización en el proceso de optimización. El objetivo en cada problema particular será encontrar un valor adecuado $0 < \lambda < \infty$. En la práctica esto suele hacerse mediante validación cruzada o bootstrap, con el propósito de minimizar una estimación del error esperado (Castro, 2012).

5.3.1. Regresión *Ridge*

Sea un modelo de regresión lineal con *n* observaciones y *p* variables explicativas donde $X = [X_1, \dots, X_p]$ está formado por las variables explicativas, con $x_j = (x_{1j}, \dots, x_{nj})^T$ e $Y = (y_1, \dots, y_n)^T$ es el vector respuesta.

Para formular la regresión *Ridge* será necesario añadir una restricción a este modelo lineal general.

Definición 1: *Se define la norma ℓ_2 de $\beta = (\beta_1, \dots, \beta_p)$ como:*

$$(||\beta||_2)^2 = \sum_{j=1}^{p} |\beta_j|^2 = \sum_{j=1}^{p} {\beta_j}^2$$

La regresión *Ridge* es un método de mínimos cuadrados penalizado que restringe la norma ℓ_2 de los coeficientes de la regresión ordinaria mediante el uso de un escalar λ no negativo.

Los coeficientes estimados $\hat{\beta}_{ridge}$ de la regresión son definidos como los valores de β que minimizan

$$\sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

Es decir, se obtienen resolviendo el siguiente problema de optimización:

Ecuación 10

$$\hat{\beta}_{ridge} = \underset{\beta}{\operatorname{argmin}} \, ||Y - \sum_{j=1}^{p} x_j\beta_j \,||^2 + \lambda \sum_{j=1}^{p} {\beta_j}^2$$

con $\lambda > 0$, parámetro de regularización que fuerza a los coeficientes a anularse.

Como se muestra en (Hastie et al., 2009) una forma equivalente de escribir el problema es:

Ecuación 11

$$\min_{\beta} \; ||Y - \sum_{j=1}^{p} X_j\beta_j \,||^2$$

$$\text{sujeto a} \sum_{j=1}^{p} {\beta_j}^2 \leq \lambda$$

Con $\lambda \geq 0$ que hace explícita la restricción de tamaño de los parámetros.

En esencia, la Ecuación 10 es la forma lagrangiana del problema de optimización con restricciones de Ecuación 11(Castro, 2012).

Cuando hay muchas variables correlacionadas en un modelo de regresión lineal, ya se sabe que los coeficientes pueden estar mal determinados y mostrar una alta varianza. Un coeficiente positivo en una variable puede ser cancelado por uno similar negativo sobre la variable con la que está correlacionada. Mediante la imposición de una restricción de tamaño de los coeficientes, como en (3.42), este problema se alivia.

Escribiendo el criterio de Ecuación 10 matricialmente,

$$RSS(\lambda) = (Y - X\beta)^T(Y - X\beta) + \lambda\beta^T\beta$$

se llega a:

Teorema 2 *La solución del problema de la regresión Ridge está dado por*

$$\hat{\beta}^{ridge} = (X^T X + \lambda I)^{-1} X^T y$$

donde I es la matriz identidad de tamaño pxp.

Corolario: *Cuando $\lambda \to 0, \hat{\beta}^{ridge} \to \hat{\beta}^{OLS}$.*

Corolario: *Cuando $\lambda \to \infty, \hat{\beta}^{ridge} = 0$.*

Corolario: *En el caso especial del diseño de una matriz ortogonal; es decir cuando X es un matriz de rango completo, con todas sus columnas independientes.*

$$\hat{\beta}_j^{ridge} = \frac{\hat{\beta}_j^{OLS}}{1 + \lambda}$$

- Esto ilustra que la característica esencial de la regresión *Ridge* es la contracción.
- La aplicación de la penalización de regresión *Ridge* tiene el efecto de contraer las estimaciones de los coeficientes hacia cero introduciendo sesgo pero reduciendo la varianza de la estimación.

Ya se ha mencionado que los estimadores de OLS no siempre existen si X no es de rango completo, es decir si $(X^T X)^{-1}$ no es invertible, y no hay solución única para $\hat{\beta}^{MCO}$. Este problema no ocurre en la regresión *Ridge*.

Nótese que con la elección de la penalización cuadrática $\beta^T\beta$, la solución de la regresión *Ridge* es nuevamente una función lineal de Y. La solución añade una constante positiva a la diagonal de $X^T X$ antes de calcular la inversa. Esto hace que, para cualquier matriz

X, la factorización $(X^T X + \lambda I)^{-1}$ siempre es invertible. Por lo tanto siempre existe una solución única para $\hat{\beta}^{ridge}$.

Teorema 3 *La varianza de la estimación de la regresión Ridge es*

$$Var(\hat{\beta}) = \sigma^2 W X^T X W$$

donde $W = (X^T X + \lambda I)^{-1}$

Teorema 4 *El sesgo (Bias) de la estimación de la regresión Ridge es*

$$Bias(\hat{\beta}) = -\lambda W \beta$$

Se puede demostrar que la varianza total $\left(\sum_j \mathrm{Var}(\hat{\beta}_j)\right)$ es monótona decreciente con respecto a λ, mientras que el sesgo total cuadrático $\left(\sum_j \mathrm{Bias}^2(\hat{\beta}_j)\right)$ es monótono creciente con respecto a λ.

Teorema 5 *Siempre existe un λ tal que el error cuadrático medio (MSE) de $\hat{\beta}^{ridge}$ es menor que el MSE de $\hat{\beta}^{OLS}$.*

Este es un resultado bastante sorprendente con implicaciones radicales. Incluso si el modelo ajustado es exactamente correcto y sigue exactamente la distribución esperada, siempre se puede obtener un estimador contraído cercano a cero.

Teorema 6 *Interpretación Bayesiana. Supóngase $\beta \sim N(0, \tau^2)$. Entonces la media de este β está dada por*

$$\left(X^T X + \frac{\sigma^2}{\tau^2} I\right)^{-1} X^T y$$

La interpretación geométrica que se hace de este tipo de regresión para el caso de dos dimensiones, es la que aparece en la Figura 5. La estimación *Ridge* contrae los coeficientes β_j^{MCO} hacia 0. Las elipses constituyen los contornos de la función de error de mínimos cuadrados y el área sólida azul la región de restricción $\beta_1^2 + \beta_2^2 \leq s^2$

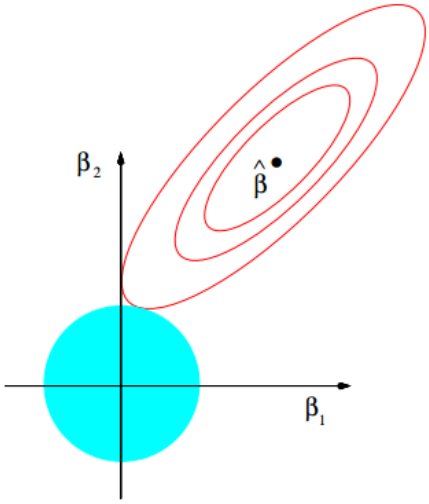

Figura 5 Estimación Ridge para el caso de dos dimensiones
(Hastie et al., 2009)*:*

Selección de λ

La selección del λ adecuado para el modelo se puede escoger mediante la aplicación de algún criterio de información, como el *Akaike Information Criterion,* AIC (Akaike, 1974) o el *Bayesian Information Criterion*, BIC (Kass & Raftery, 2012). Estos permiten elegir de entre varios modelos él de mejor ajuste a los datos, equilibrando los objetivos de parsimonia y entropía, ajustando el modelo con mejor vector respuesta y y con la menor cantidad de variables y error.

Para aplicar el AIC o el BIC al problema de la elección λ, se necesita la estimación de los grados de libertad. Dado que en la regresión ordinaria los grados de libertad del modelo estaban dados por $\mathrm{tr}(H) = p$, en la regresión *Ridge* también se puede encontrar una matriz sombrero H_{ridge} tal que $\hat{y} = H_{ridge}y$:

$$H_{ridge} = X(X^T X + \lambda I)^{-1} X^T$$

Análogamente, los grados de libertad para la $\mathrm{tr}(H_{ridge})$ se obtienen así:

$$df_{ridge} = \sum \frac{\lambda_i}{\lambda_i + \lambda}$$

donde $\{\lambda_i\}$ son los valores propios de $X^T X$

Se puede comprobar que df es una función decreciente de λ con $df = p$, cuando $\lambda = 0$, y $df = 0$, cuando $\lambda = \infty$.

Una vez cuantificados los grados de libertad del modelo de regresión *Ridge,* se puede calcular el AIC o el BIC y utilizarlos para guiar la elección de λ.

$$AIC = n\log(RSS) + 2df$$

$$BIC = n\log(RSS) + df\log(n)$$

Una forma alternativa de elegir λ sería viendo cómo de bien predicen las estimaciones basadas en $\hat{\beta}$, los casos reales de y.

Ahora bien, no es apropiado utilizar los datos dos veces dos veces, una para ajustar el modelo y otra para calibrar la capacidad de explicación. Lo ideal sería tener un conjunto de datos extra para la validación, pero obviamente la obtención de datos es costosa y esto raramente es posible. Una idea consiste en dividir el conjunto de datos en dos conjuntos, uno para ajustar $\hat{\beta}^{ridge}$ y el otro para evaluar cómo de bien $X\hat{\beta}^{ridge}$ predice las observaciones. El problema de esta solución es que rara vez tenemos tantos datos como para dividirlos en dos conjuntos, con el único fin de elegir λ.

Para resolver este problema se hace uso de la validación cruzada. La validación cruzada divide los datos en K pliegues, ajustando los datos sobre los $K-1$ pliegues restantes, y evaluando el riesgo en el pliegue que quedó fuera. Las opciones más habituales para K son 5, 10, y n.

Como por lo general llevar a cabo este procedimiento tiene un alto coste computacional, se suele recurrir a la validación cruzada generalizada, definida a través de:

$$GCV = \frac{1}{n}\sum_{i}\left(\frac{y_i - \hat{y}_i}{1 - \mathrm{tr}(H)/n}\right)^2$$

5.3.2. Regresión *Lasso*

Se ha visto cómo la regresión *Ridge* es capaz de reducir la variabilidad y mejorar la precisión de los modelos de regresión lineal, con beneficios más notables en presencia de multicolinealidad. Sin embargo, la regresión *Ridge* no es un método de selección de variables y no proporciona un modelo parsimonioso con pocos parámetros. Es por ello que surgió la regresión *Lasso.*

Definición 2: *Se define la norma* ℓ_1 *de* $\beta = (\beta_1, \dots, \beta_p)$ *como:*

$$(||\beta||_1)^1 = ||\beta||_1 = \sum_{j=1}^{p}|\beta_j|^1 = \sum_{j=1}^{p}|\beta_j|$$

La regresión *Lasso* es un método de mínimos cuadrados penalizado, que incluye una restricción en la norma ℓ_1 de los coeficientes de regresión.

Los coeficientes estimados $\hat{\beta}_{lasso}$ de la regresión son definidos como los valores de β que minimizan:

$$\sum_{i=1}^{n}\left(y_i - x_i^T\beta\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

Se obtienen resolviendo el siguiente problema de optimización:

Ecuación 12

$$\hat{\beta}_{lasso} = \underset{\beta}{\operatorname{argmin}} ||Y - \sum_{j=1}^{p} X_j\beta_j||^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

con $\lambda > 0$, parámetro de regularización, que fuerza a los coeficientes a anularse.

Este problema también se encuentra definido de la siguiente forma equivalente:

Ecuación 13

$$\min_{\beta}||Y - \sum_{j=1}^{p} X_j\beta_j||^2$$

$$sujeto\ a\ \sum_{j=1}^{p}|\beta_j| \leq s$$

La regresión *Lasso* se diferencia de la *Ridge* porque utiliza la penalización en la norma ℓ_1, en vez de en la ℓ_2. El cambio en la función de penalización es sutil, pero tiene un impacto dramático en el estimador resultante.

Al igual que la regresión *Ridge,* al penalizar los coeficientes también se contraen a cero. Sin embargo, a diferencia de la *Ridge*, en *Lasso* algunos de los coeficientes son encogidos hasta llegar a ser exactamente iguales cero. Este tipo de soluciones, con múltiples valores idénticamente cero, se dice que son soluciones dispersas (*sparse*). Este tipo de penalización se convierte en una especie de selección variable continua precisamente por ello.

El estimador resultante fue nombrado *Lasso* por sus siglas en inglés *LeastAbsolute Shrinkage and Selection Operator*. Su propio nombre ya proporciona una idea de las propiedades del estimador.

Lasso contrae continuamente sus coeficientes a 0, y logra la precisión de su predicción a través de la compensación del sesgo y de la varianza. Debido a la naturaleza de la penalización ℓ_1, algunos coeficientes se hacen nulos si λ es suficientemente grande. Por lo tanto, *Lasso* produce simultáneamente un modelo tanto preciso como *sparse,* convirtiéndolo en un método favorable de selección de variables.

En cuanto a la interpretación geométrica de esta técnica, la Figura 6 muestra la forma de actuar de la técnica en dos dimensiones. La solución se establece en el primer punto donde los contornos elípticos (contornos de la función de error de mínimos cuadrados) se encuentran con la región de restricción, representada por el cuadrado: $|\beta_1| + |\beta_2| \leq s$. A diferencia de *Ridge*, la solución ocurre habitualmente en los vértices del cuadrado donde alguno de los β_j es igual a 0. Cuando el número de variables $p > 2$, el cuadrado se convertirá en un hipercubo con mayor cantidad de vértices y, por tanto, mayor oportunidad para que los parámetros estimados sean nulos.

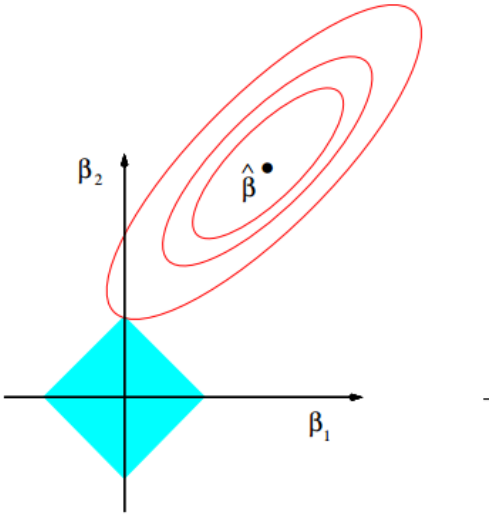

Figura 6 Estimación Lasso para el caso de dos dimensiones
(Hastie et al., 2009)*:*

Aunque *Lasso* presenta soluciones exitosas en muchas situaciones, tiene ciertas limitaciones (Tibshirani, 1996):

- En el caso en que $p > n$, *Lasso* selecciona como mucho n variables, por la naturaleza del problema de optimización convexa. Esto parece una limitación para un método de selección de variables. Es más, *Lasso* no está bien definido cuando el valor frontera (s) para la norma ℓ_1 es menor que un cierto valor.

- Si existe un grupo de variables con correlaciones por parejas muy altas, *Lasso* tiende a seleccionar una única variable del grupo, sin tener cuidado con cuál de ellas se queda.
- Para las situaciones habituales en que $n > p$, si existen altas correlaciones entre regresoras, se ha demostrado empíricamente que el rendimiento de la predicción *Ridge* es mejor que el de *Lasso.*

Solución

Debido a cómo es la penalización *Lasso*, esta hace que la solución sea no linear en y_i, por lo que la función RSS no es diferenciable y carece de una forma cerrada en general, a diferencia de cómo ocurría con *Ridge*.

Encontrar la solución *Lasso* se convierte en un problema de programación cuadrática. Por lo tanto se definen operadores de umbralización diferentes para determinar los coeficientes.

Se define el operador de umbralización suave (*soft-thresholding operator, S*)

$$S(z,\lambda) = \begin{cases} z-\lambda & \text{si } z > \lambda \\ 0 & \text{si } |z| \leq \lambda \\ z+\lambda & \text{si } z < \lambda \end{cases}$$

Por otro lado, se determina el operador de umbralización fuerte (*Hard-thresholding operator, h*), que se diferencia del operador de umbralización suave en que este último es continuo mientras que el operador fuerte no lo es. En el caso ortonormal, la selección del mejor subconjunto es equivalente a la umbralización fuerte.

$$h(z,\lambda) = \begin{cases} z & \text{si } |z| > \lambda \\ 0 & \text{si } |z| \leq \lambda \end{cases}$$

Dado que ambas umbralizaciones son simples funciones de las soluciones de mínimos cuadrados para el subconjunto con r variables, los coeficientes $\hat{\beta}_j$ pueden determinarse en *Lasso* como

$$\hat{\beta} = S\left(\hat{\beta}_J^{OLS}, \lambda\right)_j$$

$$\hat{\beta}_j = H\left(\hat{\beta}_J^{OLS}, \lambda\right)$$

Mientras que en la regresión *Ridge* los coeficientes eran estimados como

$$\hat{\beta}_j = \hat{\beta}_J^{OLS}/(1+\lambda).$$

Selección de λ

A diferencia de la regresión *Ridge*, como la regresión *Lasso* no es un estimador lineal, no existe una matriz H_{Lasso} tal que $\hat{y} = H_{Lasso}y$. Es por ello que definir los grados de libertad para escoger es un tanto diferente.

Para ello se pueden utilizar los coeficientes no nulos en el modelo de regresión para cuantificar los grado de libertad y poder implementar el AIC, el BIC o la validación cruzada general (*General Cross Validation*, GCV) como criterio de selección de λ. Aunque se suele usar e AIC o el BIC, estos poseen aproximaciones poco fiables y es preferible seleccionar λ a través de la validación cruzada.

5.3.3. Elastic net

La técnica *Ridge* tiende a contraer los coeficientes de las variables correlacionadas en forma conjunta, de forma que posean coeficientes estimados similares, y tiende a contraer hacia 0 los coeficientes de la regresión. *Lasso* es indiferente en la elección entre un conjunto de variables fuertemente correlacionadas, ya que tiende a elegir una sola del conjunto y descartar las demás, y contrae hacia 0 los coeficientes de la regresión haciendo exactamente nulos algunos de ellos.

Se define entonces *Elastic net,* un modelo de regresión presentado por Zou y Hastie (2005), que combina las penalizaciones *Ridge* y *Lasso* para superar las limitaciones de ambas, mientras conserva sus propiedades favorables.

Para cualquier λ_1, λ_2 no negativo, esta técnica estima los coeficientes de la regresión $\hat{\beta}_{en}$ como sigue:

Ecuación 14

$$\hat{\beta}_{en} = (1 + \lambda_2)\left\{ \underset{\beta}{\operatorname{argmin}} \; ||y - \sum_{j=1}^{p} X_j \beta_j \,||^2 + \lambda_2 \sum_{j=1}^{p} \left|\beta_j\right|^2 + \lambda_1 \sum_{j=1}^{p} \left|\beta_j\right| \right\}$$

La penalización introducida por *Elastic net* es una combinación convexa de las penalizaciones *Ridge y Lasso*. El primer término $\lambda_2 \sum_{j=1}^{p} \left|\beta_j\right|^2$ promueve que variables altamente correlacionadas tengan coeficientes similares y el segundo $\lambda_1 \sum_{j=1}^{p} \left|\beta_j\right|$, soluciones *sparse*. Obviamente, *Lasso* es un caso especial de *Elastic net* cuando $\lambda_2 = 0$.

Para resolver el problema de solución única cuando $p \gg n$, se elige $\lambda_2 > 0$. *Elastic net* puede incluir potencialmente todas las variables en el modelo ajustado, por lo que elimina la limitación particular de *Lasso*.

La Figura 7 muestra cómo el área sólida formada por las restricciones de la técnica se presenta como una combinación de las anteriores.

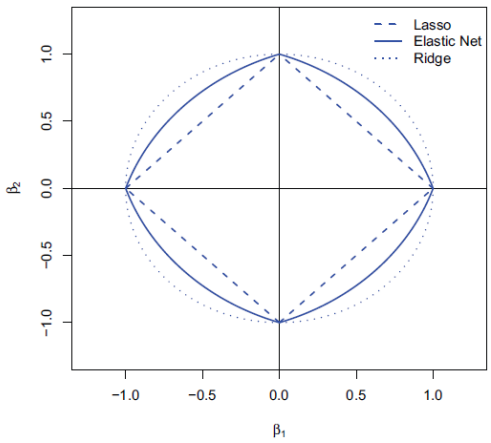

Figura 7. Interpretación gráfica de Elastic net
(Castro, 2012)

6. ANÁLISIS DE COMPONENTES PRINCIPALES *SPARSE* COMO TÉCNICA DE REGRESIÓN PENALIZADA

Como bien se viene diciendo a lo largo del documento, las componentes principales presentan la desventaja de ser combinación lineal de todas las variables originales. Es por ello que se pretende buscar métodos que produzcan cargas nulas, facilitando así la comprensión de dichas nuevas variables.

Se presenta aquí uno de dichos métodos: el Análisis de Componentes Principales *Sparse,* cuya finalidad es conseguir gran parte de los coeficientes de la matriz de cargas nulos.

Se analiza el PCA inicialmente desde la descomposición en valores singulares de la matriz de correlación de los datos de entrada, continuando a través de su expresión como problema de regresión.

Una vez presentadas las técnicas de regularización sobre los problemas de regresión, y reescrito el PCA como uno de ellos, se seguirán los pasos que permiten la definición de esta nueva técnica tan provechosa. En lugar de los coeficientes de regresión que anteriormente eran contraídos hacia 0, ahora serán las cargas del PCA las que pretendan ajustarse a 0.

Para ello se hace uso de la imposición de las restricciones vistas, sobre las cargas del PCA, para finalizar con la obtención de cargas *sparse,* que dan lugar al SPCA.

6.1. Taxonomía de los modelos

El *Sparse* PCA puede considerarse como un PCA modificado y al igual que este, puede definirse a través de los enfoques vistos en la sección 0.

Al igual que ocurre en todas las técnicas desarrolladas en la estadística multivariante, pueden desarrollarse modificaciones respecto a la técnica como tal, dando lugar a distintos algoritmos que resuelven el problema, enfocándolos desde distintos puntos de vista y atendiendo a distintas restricciones. Esto hace que cada algoritmo tenga propiedades ventajosas que otros no tienen, y al igual ocurriría con sus desventajas.

Todo ello es aplicable al caso del SPCA, técnica para la cual surgen distintas formas de enfocar la resolución del problema de optimización planteado (bien sea un problema de

optimización que pretende capturar la máxima información posible, o hacer mínimo el error de construcción cometido), dado que ningún algoritmo proporciona soluciones óptimas y computacionalmente eficientes para todos los casos posibles de solución.

6.1.1. Formas del modelo

Existen distintas formas de enfocar el problema de optimización, atendiendo a la definición de la función objetivo y de las restricciones a imponer.

Si el nuevo vector y se puede estimar como una función de X, es decir, $\hat{y} = f\,(x)$, entonces dicha función puede penalizarse para hallar una solución donde lo coeficientes sean lo más cercanos a cero posible.

En la literatura académica, existen varias formas del modelo de PCA. Por ejemplo, Wright (2011) propone la siguiente clasificación:

- Forma ponderada: En esta forma se puede minimizar el error de ajuste o maximizar la varianza pero la penalización a los coeficientes se encuentra directamente en la función objetivo:

$$\min\,(f\,(x)\,+\,\lambda\,\|v\|_1),\ \text{para}\ \lambda > 0$$

- Forma limitada: En esta forma se puede minimizar el error de ajuste o maximizar la varianza pero la penalización a los coeficientes se encuentra definida como una restricción a la función objetivo:

$$\min f(x)$$
$$\text{sujeto a}\ \|V\|_1 \leq \lambda$$

- Forma de función con limitaciones: En esta forma se minimiza la matriz X pero en función de y:

$$\min X$$
$$\text{sujeta a}\ f\,(x) \leq \bar{f}.$$

6.1.2. Resumen de los modelos SPCA

Trendafilov (2014) realizó una revisión de los principales enfoques para la interpretación de los resultados del análisis de componentes principales durante los últimos 50 a 60 años, desde el enfoque de maximización de varianza hasta el enfoque moderno de regresión con penalizaciones donde se obtienen soluciones directamente interpretables. Como se mencionó al inicio del documento el enfoque de maximización de la varianza ha sido desarrollado con más detalle desde el Análisis Factorial. Por lo tanto, sólo en

esta sección y con el fin de presentar los avances en este tipo de formulación, el vector de cargas a es equivalente al vector de cargas v del PCA.

Los métodos de rotación no pueden producir las cargas *sparse,* ya que están diseñados para simplificar las cargas preservando al mismo tiempo el porcentaje ya elegido de la varianza. Por lo tanto, parece razonable flexibilizar este requisito, poniendo más énfasis en la simplicidad que en la maximización de la varianza.

Como se verá a continuación, el autor comienza con la modificación de Jolliffe y Uddin (2000) a las PCs originales para satisfacer el criterio VARIMAX, preservando el decrecimiento de la proporción de varianza explicada. Este algoritmo, conocido como SCoTLASS, es uno de los más influyentes en el SPCA. Los autores resuelven el siguiente problema:

Ecuación 15

$$\max_{a^T a=1} a^T (X^T X) a + \tau \mathcal{V}(a)$$

La Ecuación 14 es la primera formulación penalizada del PCA diseñada para producir directamente cargas interpretables de las componentes: de a. El método se llama SCoT (Simplified Component Technique) y la función criterio VARIMAX $\mathcal{V}(a)$ se usa como una restricción para inducir coeficientes o cargas con poca densidad (*sparseness).* Sin embargo, las cargas resultantes contienen varios valores que, aunque pequeños, son distintos de cero, es decir, que todavía no son *sparse*. Entonces, Jolliffe, Trendafilov, y Uddin (2003) modificaron las PCs originales para satisfacer adicionalmente la restricción *Lasso*, que dirigía muchas cargas a ser exactamente ceros. Este método, llamado SCoTLASS (Simplified Component Technique subject to LASSo), se convirtió en el primero en producir cargas *sparse* en componentes.

Se enumeran a continuación las definiciones del *Sparse* PCA más influyentes y los métodos para su solución, todos ellos de actual desarrollo. El primer grupo importante de métodos se basa en un tipo de reformulación penalizada del PCA. Sean X y R la matriz de datos y la matriz de correlación de la muestra.

- El algoritmo SCoTLASS (Jolliffe et al., 2003;Trendafilov & Jolliffe, 2006) es un procedimiento que obtiene las cargas *sparse* imponiendo directamente una restricción ℓ_1 al problema original del PCA para maximizar sucesivamente la varianza. Encuentra el $i-ésimo$ vector de cargas *sparse* como una solución al siguiente problema:

$$\max_{\substack{\|a\|_2=1 \; y \; \|a\|_1 \leq \tau \\ a \perp \{a_1,\ldots,a_{i-1}\}}} a^T R a$$

Con $R = X^T X$ y donde $a_1, \ldots, a_{i-1}$ son los vectores asociados a las cargas de componentes *sparse*. La condición $\|a\|_1 \leq \tau$, para algún parámetro de ajuste τ, es conocida como restricción *Lasso*. SCoTLASS no presenta una orientación clara sobre cómo elegir un valor apropiado para τ, aunque un τ lo suficientemente pequeño produce algunas cargas nulas. Los autores definen que podrían probarse varios valores de τ, pero esto generaría un alto coste computacional, dado que SCoTLASS no es un problema de optimización convexa, haciendo su solución poco práctica (Trendafilov, 2014).

- Otro enfoque para obtener componentes *Sparse* está basado en los límites espectrales de las sub-matrices de la matriz de correlación R. La idea es identificar el subconjunto m de variables que explican la máxima varianza entre todos los posibles subconjuntos de tamaño m y reemplaza las cargas del resto $p - m$ variables por ceros. Cadima y Jolliffe (2001) emplearon este método para la selección de variables en el análisis discriminante lineal de Fisher (LDA). Moghaddam et al. (2006) desarrollaron posteriormente la idea y la usaron para construir un algoritmo para *Sparse* PCA y LDA sujeto a restricciones de cardinalidad.
- sPCA-rSVD: Shen y Huang (2008) resuelven el siguiente problema.

$$\min_{a,b,\|b\|_2=1} \|X - ba^T\|_F^2 + \mathrm{P}_\lambda(a)$$

Donde $a \in \mathbb{R}^p$, $b \in \mathbb{R}^n$ y $\mathrm{P}_\lambda(a)$ es el termino particular de penalización. La solución a da las cargas de la $j - ésima$ componente *sparse*.

- Witten et al. (2009) proponen un algoritmo general para *sparse* SDV.

$$\min_{a,b} b^T X a$$
$$s.a \; \|a\|_2^2 \leq 1, \|b\|_2^2 \leq 1, \mathrm{P}_1(a) \leq \tau_1, \mathrm{P}_2(b) \leq \tau_2$$

Esta formulación generaliza y simplifica las definiciones de las *Sparse* PCA dadas por SCoTLASS, SPCA y SPCA-RSVD, y se puede aplicar para el Análisis de Correlación Canónica *Sparse* (CCA)

- Qi, Luo, y Zhao (2013) construyen componentes *sparse* usando la suma ponderada de las normas ℓ_1 y ℓ_2, así: $\|a\|_\lambda^2 = (1-\lambda)\|a\|_2^2 + \lambda\|a\|_1^2$. Se forma similar al SCoTLASS, el $j - ésimo$ vector de cargas de las componentes *sparse* es encontrado como una solución del siguiente problema de optimización.

$$\min_{\|a\|_2=1} \frac{a^T R a}{\|a\|_{\lambda_i}^2}$$
$$a \perp \{a_1, \dots, a_{i-1}\}$$

Donde $a_1, \dots, a_{j-1}$ son los vectores asociados a las cargas de componentes *sparse*. En contraste para *Lasso* y la mayoría de los otros métodos, el conjunto de restricciones es estrictamente convexo para $0 \leq \lambda < 1$. La solución de este problema encuentra un nuevo tipo de umbralización. Las cargas de las componentes *sparse* $A = \{a_1, \dots, a_{i-1}\}$ son ortonormales, $A^T A = I_r$, pero las componentes están correlacionadas y $A^T R A$ es no diagonal. Por eso se suele considerar el siguiente problema modificado

$$\min_{\|a\|_2=1} \frac{a^T R a}{\|a\|_{\lambda_i}^2}$$
$$Ra \perp \{a_1, \dots, a_{i-1}\}$$

Los autores obtienen componentes *sparse* no correlacionadas, es decir, $A^T R A$ diagonal. Sin embargo, las saturaciones en las componentes *sparse* no son ortonormales, es decir, $A^T A \neq I_r$. Es pertinente tener en cuenta que la penalización como una mezcla de las normas ℓ_1 y ℓ_2 ya ha sido considerado para el tipo de regresión de los problemas por Friedlander y Tseng (2007).

Estas son algunas de las formulaciones existentes para lograr el objetivo *sparse,* dada una matriz de datos. Sin embargo, nos centraremos en la siguiente formulación, planteada por Zou et al. (2006):

- SPCA: Zou et al. (2006) transforman el PCA en una regresión estándar resolviendo el siguiente problema para un vector auxiliar $b \in \mathbb{R}^p$:

$$\min_{a,b} \|X - ab^T X\|_F^2 + \lambda_1 \|a\|_2^2 + \lambda_2 \|a\|_1$$
$$s.a \ \|b\|_2 = 1$$

La solución a da para $i - ésimo$ vector las cargas de las componentes *sparse*. Witten, Tibshirani y Hastie (2009) reescribieron de forma equivalente el PCA de la siguiente forma:

$$\min_{a,b} \|X - ab^T X\|_F^2 + \lambda_1 \|a\|_2^2 + \lambda_2 \|a\|_1$$
$$s.a \ \|a\|_2^2 \leq 1, \|a\|_1 \leq \tau, \|b\|_2 = 1$$

Este método de resolución se analizará más detenidamente en la siguiente sección.

6.2. Modelo de minimización del error

De entre todos los algoritmos vistos para la resolución de esta técnica, en este apartado se presenta la definición matemática de las componentes principales *Sparse* desde el enfoque de minimización del error de reconstrucción del nuevo espacio con respecto al original (Zou et al., 2006). Tras ir mostrando un enfoque diferente para modificar el PCA, se mostrará cómo el PCA puede "reescribirse" exactamente como un problema de regresión a pasos. Tras introducir la penalización *Ridge*, se incluirá la penalización *Lasso,* cambiando la regresión *Ridge* por una regresión *Elastic net*.

6.2.1. Aproximación a las componentes *sparse*

Como se muestra en (Zou et al., 2006), se comenzará con un enfoque del PCA como un problema de regresión simple. Como cada componente principal es una combinación lineal de p variables, sus cargas pueden obtenerse mediante la regresión de la componente principal sobre las p variables.

Teorema 7 Para cada i, se denota la $i-ésima$ componente principal como $y_i = u_i d_{ii}$. Sea λ_2 el escalar positivo utilizado por penalización ℓ_2 de la regresión *Ridge y las estimaciones* $\hat{\beta}_{ridge}$ *Ridge dadas por* :

Ecuación 16

$$\hat{\beta}_{ridge} = \underset{\beta}{argmin}\, ||Y_i - X\beta||^2 + \lambda_2||\beta||^2$$

Tomando $\hat{v} = \frac{\hat{\beta}_{ridge}}{||\hat{\beta}_{ridge}||}$, *entonces* $\hat{v} = V_i$.

Este teorema muestra la conexión entre el PCA y el método de regresión.

Obviamente, cuando $n > p$ y X es una matriz de rango completo, la Ecuación 16 no requiere un λ_2 positivo. Además, cuando $p > n$ y $\lambda_2 = 0$, la regresión múltiple ordinaria no tiene una solución única, que sea exactamente V_i. Lo mismo ocurre cuando $n > p$ y X no es una matriz de rango máximo. Sin embargo, el PCA siempre proporciona una solución única en todas las situaciones. Como se muestra en la Ecuación 16, esta indeterminación se elimina con la penalización R*idge* positiva ($\lambda_2||\beta||^2$). También se puede demostrar que, después de la normalización, los coeficientes son independientes de λ_2; por lo tanto, la penalización *Ridge* no es usada para penalizar los coeficientes de la regresión, pero sí para garantizar la reconstrucción de las componentes principales (Zou et al., 2006).

Ahora, se añade la penalización ℓ_1, de la regresión *Lasso*, a la Ecuación 16 y se considera el siguiente problema de optimización:

Ecuación 17

$$\beta = \underset{\beta}{argmin} \left|\left|Y_i - X\beta\right|\right|^2 + \lambda_2\left|\left|\beta\right|\right|^2 + \lambda_1\left|\left|\beta\right|\right|_1$$

Donde $\left|\left|\beta\right|\right|_1 = \sum_{j=1}^{p}|\beta_j|$ es la norma ℓ_1 de β, $\widehat{V_i} = \frac{\widehat{\beta}}{||\widehat{\beta}||}$ una aproximación de V_i y $X\widehat{V_i}$ la i-ésima componente principal aproximada.

Zou y Hastie (2005) llamaron a la Ecuación 17 *naive Elastic net* que difiere del modelo *Elastic net* en un factor de escala $(1 + \lambda_2)$. Ellos demostraron que si se utilizan los coeficientes normalizados, el factor de escala no afecta a $\widehat{V_i}$.

Claramente, un λ_1 suficientemente grande proporciona un β *sparse* y, por tanto, una $\widehat{V_i}$ *sparse.*

La Ecuación 17 puede ser resuelta usando el algoritmo LARS-EN (Zou & Hastie, 2005), el cual permite elegir flexiblemente una aproximación *sparse* para las componente principales. Dado un λ_2 fijo, se resuelve el problema de minimización para todo λ_1.

<u>Criterio de solución SPCA</u>

Ahora bien, realmente la Ecuación 16 depende de los resultados del PCA y no es una alternativa real. Sin embargo, puede ser utilizado en un análisis exploratorio en el que la primera etapa sea un análisis PCA, para después utilizar la Ecuación 17 para encontrar aproximaciones *sparse* adecuadas.

A continuación se define el criterio de solución del SPCA desde los vectores fila de la matriz X, x_i.

Teorema 8 *Considerando las primeras r componentes principales. Y definiendo las matrices* $A_{pxr} = [\alpha_1, \dots, \alpha_k]$ *y sea* $B_{pxr} = [\beta_1, \dots, \beta_r]$*. Para cualquier* $\lambda_2 > 0$*, sea:*

Ecuación 18

$$(\hat{A}, \hat{B}) = \underset{A,B}{\operatorname{argmin}} \sum_{i=1}^{n} ||x_i - AB^T\, x_i||^2 + \lambda_2 \sum_{j=1}^{r} ||\beta_j||^2$$
$$\text{s. a. } A^T A = I_{rxr}$$

Entonces, $\widehat{\beta_i} \propto V_j$ *para* $j = 1,2, \dots, r$*.*

Con el Teorema 8 efectivamente se transforma el problema del análisis de componentes principales en un problema de regresión. El elemento crítico es la función objetivo $\sum_{i=1}^{n} ||x_i - AB^T x_i||^2$. Si se restringe que $B = A$, entonces:

Ecuación 19

$$\sum_{i=1}^{n} ||x_i - AB^T x_i||^2 = \sum_{i=1}^{n} ||x_i - AA^T x_i||^2$$

Que al minimizar bajo la restricción ortonormal en A, se trata exactamente de los primeros r vectores de cargas de las componentes principales ordinarias (del análisis del PCA ordinario).

La formulación anterior fue desarrollada por Hastie y Tibshirani (2001). El Teorema 8 muestra que se puede tener un PCA exacto relajando la restricción $B = A$ y sumando el término de penalización *Ridge*. Como se mostrará después, estas generalizaciones nos permiten modificar flexiblemente el PCA.

Llevando adelante la conexión entre PCA y la regresión, se usa el enfoque *Lasso* para producir cargas ("coeficientes de regresión") *sparse*. Para ello, se suma la penalización *Lasso* a la Ecuación 18 y se considera el siguiente problema de optimización:

Ecuación 20

$$(\hat{A}, \hat{B}) = \underset{A,B}{\operatorname{argmin}} \sum_{i=1}^{n} ||x_i - AB^T x_i||^2 + \lambda_2 \sum_{j=1}^{r} ||\beta_j||^2 + \sum_{j=1}^{r} \lambda_{1,j} ||\beta_j||_1$$
$$\text{sujeto a } A^T A = I_{rxr}$$

Mientras que se utiliza el mismo λ_2 para todas las r componentes, se permiten diferentes $\lambda_{1,j}$ para penalizar las cargas de las distintas PCs.

Nuevamente, si $p > n$, se requiere un λ_2 positivo con el fin de obtener un análisis de componentes principales exacto cuando la penalización *Lasso* se anula ($\lambda_{1,j} = 0$).

Es la Ecuación 20 la que de aquí en adelante recibirá el nombre de **criterio SPCA:**

$$(\hat{A}, \hat{B}) = \underset{A,B}{\operatorname{argmin}} \sum_{i=1}^{n} ||x_i - AB^T x_i||^2 + \lambda_2 \sum_{j=1}^{r} ||\beta_j||^2 + \sum_{j=1}^{r} \lambda_{1,j} ||\beta_j||_1$$
$$\text{sujeto a } A^T A = I_{rxr}$$

Téngase en cuenta que los pasos considerados en la obtención del criterio SPCA son los mismos que se siguen a la hora de resolver el problema. El primer paso es determinar el parámetro de regularización λ_2 correspondiente a la parte de la penalización *Ridge.* Una vez determinado y fijado, se pasa al análisis de los parámetros $\lambda_{1,j}$ (uno diferente para cada componente por cómo es la penalización *Lasso*).

Solución numérica

Zou et al. (2006) proponen dos formas para minimizar el criterio SPCA de la Ecuación 20, así:

B dada A: Para cada *j*, se define $y_j^* = \boldsymbol{X}\alpha_j$. Se sabe que $\hat{B} = [\hat{\beta}_1, \dots, \hat{\beta}_r]$, donde cada $\hat{\beta}_j$ es un estimador *Elastic net*:

Ecuación 21

$$\hat{\beta}_j = \underset{\beta_j}{argmin}\, ||y_j^* - X\beta_j||^2 + \lambda_2||\beta_j||^2 + \lambda_{1,j}||\beta_j||_1$$

A dada B: Por otro lado, si se fija B, la parte de penalización en la Ecuación 20 se puede ignorar y tratar de optimizar lo siguiente:

$$\text{argmin} \sum_{i=1}^{n} ||x_i - AB^T x_i||^2 = ||X - XBA^T||^2$$
$$\text{s.a. } A^T A = I_{rxr}$$

Donde la solución a este problema se obtiene por una forma de rango reducido de la rotación *Procrustes,*dada en el teorema 4 que viene a continuación.

Se computa la descomposición en valores singulares:

Ecuación 22

$$(XX^T)B = UDV^T$$

y se toma $\hat{A} = UV^T$.

Teorema 9 *Reduced Rank Procrustes Rotation. Sean M_{nxp} y N_{nxr} dos matrices. Se considera el siguiente problema de minimización restringida:*

Ecuación 23

$$\hat{A} = \underset{A}{\text{argmin}} \left||M - NA^T|\right|^2$$

$$\text{s.a. } A^T A = I_{rxr}$$

Suponiendo que la SVD de $M^T N$ *es* $UDV^T, entonces\ \hat{A} = UV^T.$

En la rotación Procrustes original (Mardia, Kent, & Bibby, 1979) N tiene el mismo tamaño que M.

Cabe señalar que para resolver la Ecuación 21 y la Ecuación 22, sólo se necesita conocer la matriz $X^T X$, porque:

Ecuación 24

$$||y^*_j - X\beta_j||^2 + \lambda_2||\beta_j||^2 + \lambda_{1,j}||\beta_j||_1 = \left(\alpha_j - \beta_j\right)^T X^T X\left(\alpha_j - \beta_j\right) + \lambda_2||\beta_j||^2 + \lambda_{1,j}||\beta_j||_1$$

Si la matriz de covarianzas muestral de X se define como $\Sigma = \frac{1}{n}X^T X$, se puede reemplazar $X^T X$ por Σ en la Ecuación 24. Si X está estandarizada de antemano, entonces se utilizará la matriz de covarianzas muestral, que es mejor cuando las escalas de las variables son diferentes.

El siguiente algoritmo resume los pasos descritos anteriormente para resolver el criterio SPCA propuesto.

Algoritmo 1. SPCA General

1. Se toma inicialmente A como las cargas de las r primeras componentes principales ordinarias: $A = V[,1:r]$.
2. Dada una A fija, $A = [\alpha_1, \dots, \alpha_r]$, resolver el problema *Elastic net* para $j = 1, \dots, r$ siguiente:
$$\beta_j = \arg\min_{\beta}\ \left(\alpha_j - \beta\right)^T X^T X\left(\alpha_j - \beta\right) + \lambda_2\left|||\beta|\right|^2 + \lambda_{1,j}\left|||\beta|\right|_1$$
3. Para una B fija, $B = [\beta_1, \dots, \beta_r]$, se computa la SVD de $X^T XB = UDV^T$ y se toma $A = UV^T$.
4. Se repiten los pasos 2-3 hasta alcanzar algún criterio de convergencia.
5. Normalizar: $\hat{V}_j = \frac{\beta_j}{||\beta_j||}, j = 1, \dots, r$

Zou et al. (2006) mencionan que la evidencia empírica sugiere que la solución del algoritmo anterior no cambia mucho cuando λ_1 varía. Para $n > p$, la elección por defecto de λ_2 es 0. Prácticamente, λ_2 es elegido para ser un número entero positivo pequeño, para evitar problemas de colinealidad en X.

En principio se pueden probar varias combinaciones de $\{\lambda_{1,j}\}$ para encontrar una buena elección de los parámetros de ajuste. De la Ecuación 17 se puede implementar el

algoritmo LARS-EN que proporciona de manera eficaz una secuencia de aproximaciones *sparse* para cada componente principal y los correspondientes valores de $\lambda_{1,j}$. Por tanto, se puede escoger un $\lambda_{1,j}$ que de un buen compromiso entre varianza y cantidad de cargas nulas. Cuando hay que ceder entre una u otra, se da una mayor prioridad a la varianza (Zou & Hastie, 2005).

Varianza Total Ajustada

En el problema de análisis de componentes principales ordinarias se garantiza tanto que las componentes principales sean no correlacionadas como que sus cargas sean ortogonales, pues dada $\hat{S} = X^T X$, se satisface que $V^T R V$ es diagonal y que $V^T V = I_r$ y.

En el caso de las componentes principales *sparse* sólo se fuerza a que las PC sean ortogonales y no se impone explícitamente que las PC sean incorrelacionadas (Jolliffe et al., 2003; Zou et al., 2006)

Sean $\hat{Y}$ las componentes principales sparse estimadas. Usualmente, la varianza total explicada por las componentes principales es calculada como la $tr(Y^T Y)$. Sin embargo, si están correlacionadas como es el caso de las componentes sparse $\hat{Y}$, no es correcto medir la varianza de esta forma porque se estaría recogiendo tanto la varianza propia de la componente como la compartida con las demás varias veces, generando una varianza sobreestimada de la real.

Es decir, si una componente principal $\widehat{\boldsymbol{Y}}_r$ está correlacionada con sus anteriores $\widehat{\boldsymbol{Y}}_j, j \in \{1, \dots, r-1\}$, entonces su varianza contiene contribuciones de dichas componentes anteriores. Estas contribuciones no deben ser incluidas en la varianza total, debido a la presencia de las componentes anteriores $\widehat{\boldsymbol{Y}}_j$ en ella.

Se puede usar la proyección de $\widehat{\boldsymbol{Y}}_k$ sobre el conjunto $\{\widehat{\boldsymbol{Y}}_j\}_1^{r-1}$ para eliminar la dependencia lineal entre las componentes correlacionadas.

Entonces, se puede definir el residuo, $\widehat{\boldsymbol{Y}}_{j \cdot 1,\dots,j-1}$, como la diferencia de $\hat{Y}_j$ menos su proyección sobre $\{\widehat{\boldsymbol{Y}}_i\}_1^{j-1}$.

$$\widehat{\boldsymbol{Y}}_{j \cdot 1,\dots,j-1} = \widehat{\boldsymbol{Y}}_{\boldsymbol{j}} - H_{\boldsymbol{1,\dots,j-1}} \widehat{\boldsymbol{Y}}_{\boldsymbol{j}}$$

donde $\boldsymbol{H}_{1,\dots,j-1}$ es la matriz de proyección sobre $\{\widehat{\boldsymbol{Y}}_i\}_1^{j-1}$.

De esta forma, la varianza ajustada de $\widehat{\boldsymbol{Y}}_j$ es $||\widehat{\boldsymbol{Y}}_{j \cdot 1,\dots,j-1}||^2$ y la varianza total explicada se define como $\sum_{j=1}^{r} ||\hat{Y}_{j \cdot 1,\dots,j-1}||^2$. En el caso en que las componentes principales *sparse* estimadas $\widehat{\boldsymbol{Y}}_j$ sean no correlacionadas, esta fórmula coincide con la $tr(\hat{Y}^T \hat{Y})$.

Usando la descomposición QR, también se puede calcular la varianza ajustada fácilmente. Supóngase la descomposición $\widehat{\boldsymbol{Y}} = \boldsymbol{QR}$, donde **Q** es ortonormal y **R** es triangular superior. Entonces:

$$||\widehat{Y}_{j\cdot 1,\dots,j-1}||^2 = R_{jj}^2$$

Por tanto, la varianza total explicada es igual a:

$$\sum_{j=1}^{r} \boldsymbol{R}_{jj}^2$$

SPCA para datos con $p >> n$

Para datos en los que $p >> n$, por ejemplo, expresión génica en arrays, el número de variables (genes) es normalmente mucho más grande que el número de muestras (por ejemplo: $n = 10.000\ y\ p = 100$). El algoritmo general SPCA se adapta a esta situación usando un λ_1 positivo. Sin embargo, el coste computacional es elevado cuando se requiere un gran número de cargas no nulas. Es deseable, simplificar el algoritmo general SPCA para impulsar el cálculo.

Obsérvese que el teorema 3 es válido $\forall \lambda_1 > 0$, así que en principio se puede utilizar cualquier λ_1 positivo. Surge una solución "ahorrativa" cuando $\lambda_1 \to \infty$. Precisamente, se tiene el siguiente teorema:

Teorema 10 Sean $\hat{V}_j(\lambda_2) = \frac{\widehat{\beta_j}}{||\widehat{\beta_j}||}$ $(j = 1, \dots, r)$ las cargas derivadas de la Ecuación 20. Sea $(\hat{A}, \hat{B})$ la solución del problema de optimización:

Ecuación 25

$$\left(\hat{A}, \hat{B}\right) = \arg\min_{A,B} - 2tr(A^T X^T X B) + \sum_{j=1}^{r} \left|\left|\beta_j\right|\right|^2 + \sum_{j=1}^{r} \lambda_{1,i} \left|\left|\beta_j\right|\right|_1$$

$$\text{s.a. } A^T A = I_{rxr}$$

Cuando $\lambda_1 \to \infty, \hat{V}_j(\lambda_2) \to \frac{\beta_j}{\left|\left|\widehat{\beta_j}\right|\right|}$.

Se puede utilizar el mismo algoritmo de la solución numérica para resolver la Ecuación 25, donde sólo hay que reemplazar el problema general de *Elastic net* con su caso especial ($\lambda_2 = \infty$). Nótese que dada A fija,

Ecuación 26

$$\widehat{\beta_j} = \arg\min_{\beta_j} - 2\alpha_j^T (X^T X)\beta_j + \left|\left|\beta_j\right|\right|^2 + \lambda_{1,j} \left|\left|\beta_j\right|\right|_1$$

Que tiene la solución en forma explícita dada en la Ecuación 27.

Algoritmo SPCA para datos con $p \gg n$

Se reemplaza el paso 2 en el algoritmo SPCA general con:

2.- Para $j = 1, \dots, r$,

Ecuación 27

$$\beta_j = \left(\left|\alpha_j^T X^T X\right| - \frac{\lambda_{1,j}}{2}\right)_+ \operatorname{sign}\left(\alpha_j^T X^T X\right)$$

Esta operación se conoce como Umbralización suave *"Soft-Thresholding"*.

7. APLICACIÓN PRÁCTICA

7.1. Software disponible

En esta sección se presentan los resultados de la aplicación práctica de la teoría expuesta anteriormente. El programa utilizado para resolver el SPCA es R (Team, 2015).

Inicialmente se presentan los datos a analizar: una primera matriz conocida como *Pitprops,* que trata cortes de madera y una segunda que realiza un estudio sobre yogures. Sobre estos datos se realizó un análisis clásico de componentes principales, seguido de la rotación *VARIMAX*. Todo ello con el fin de identificar las limitaciones de ambas técnicas e introducir las bondades de las versiones *sparse* para seleccionar variables, extraer características relevantes, reducir la dimensión y mejorar la interpretación de los mismos.

Las matrices de ejemplos a tratar parten de puntos diferentes. La matriz *Pitprops* utilizada en el estudio es la matriz de correlaciones entre las variables, debido al desconocimiento de la matriz original, mientras que el estudio de los yogures si se realiza a partir de los datos originales tomados para el estudio.

Dado que el estudio de *Pitprops* parte de la matriz de correlaciones, el primer acercamiento al PCA se hace desde la función de R *pca()* del paquete *spca*. El primer acercamiento al PCA de la matriz de yogures se hace desde la función de R *prcomp()*, la cual calcula componentes principales por medio de la Descomposición en Valores Singulares (SVD). El método de solución pretende reducir la dimensión de una matriz de datos por medio de la proyección de las variables observadas en las direcciones de máxima varianza. La función *prcomp()* devuelve como solución la raíz cuadrada de los valores propios de la matriz de correlación y entrega la matriz de cargas de cada variable en una componente dada.

Una vez obtenidos los resultados del PCA, se utiliza la función *Varimax()* de R, para llevar acabo la rotación VARIMAX sobre ellos.

A pesar de que el PCA y la rotación VARIMAX aplicadas permitieron reducir la dimensión de la matriz y ayudaron a la interpretación de las componentes principales, existen casos en los que esta interpretación no es tan clara y en los que podría interesar

al investigador reducir la cantidad de variables que carga en cada PC con el fin de encontrar un significado de cada componente más claro.

Es por ello que se recurre a la aplicación de la técnica SPCA sobre la matriz de datos en estudio.

Los resultados se obtuvieron mediante la ejecución de funciones del software R (2015), incluidas en el paquete *elasticnet* (Zou & Hastie, 2012) para calcular componentes *sparse*; es decir, componentes principales con cargas nulas y cardinalidad baja (entendiendo por cardinalidad el número de cargas no nulas en cada componente), de forma que sean combinación lineal de sólo unas pocas variables.

El paquete de R *elasticnet* fue desarrollado por Zou & Hastie (2012) y proporciona dos funciones para calcular componentes principales *sparse*: la primera es *spca()* que se usa para los casos cuando $n < p$ y la segunda es *arrayspc()* que se puede aplicar cuando $p \gg n$.

Recordamos que tal y como se ha visto en esta técnica, para la aplicación de ésta es necesario calcular por un lado el parámetro de complejidad λ respectivo a la penalización *Ridge* y los distintos parámetros de complejidad para cada una de las penalizaciones *Lasso* en cada componente.

Los autores comprobaron que con un valor de $1e^{-6}$ para el parámetro de complejidad de la penalización *Ridge,* se satisface el fin de esta restricción, que es generar cargas de las variables en las componentes principales cercanas a cero y con valores de cargas similares en aquellas variables fuertemente correlacionadas. Por lo tanto, en esta aplicación práctica sólo se definirán los valores de los parámetros de complejidad asociados a la penalización *Lasso* para cada componente.

A pesar de que los resultados en este trabajo se obtienen a través de este paquete acorde a la teoría desarrollada, este no es el único desarrollado en R para el cálculo de componentes principales *sparse.* A pesar de que en este trabajo no se presentan los resultados obtenidos a través de ellas, otras de las funciones disponibles son las incluidas en el paquete *spca* (Merola, 2014). Específicamente*, spca()* es una función alternativa para solucionar el Criterio SPCA descrito anteriormente (Zou et al., 2006). Se denomina como alternativa porque esta función además de minimizar el error de reconstrucción mediante un vector de escalares, que contiene los valores respectivos para las penalizaciones *Ridge* y *Lasso*, puede también definir un vector de cardinalidades, es decir, estipular un vector cuyos elementos determinen el número de

cargas diferentes de cero en cada componente. Se permite así al usuario imponer dicha cardinalidad.

Este paquete calcula las componentes principales *sparse* a través del método de mínimos cuadrados (LS SPCA) pudiendo emplear dos algoritmos diferentes de solución: *Branch and Bound* (BB) y *Backward Elimination* (BE). El primero de ellos tiene lugar a través de la función *spcabb(),* y sus soluciones se obtienen de la aproximación por mínimos cuadrados de la matriz de datos mediante una matriz de correlación (o covarianza) utilizando el algoritmo de solución *Branch and Bound*. Sin embaro, la búsqueda BB es computacionalmente exigente. Es por ello que para los problemas grandes $(n > 50)$ se recomienda usar otro algoritmo de solución, dado por función *spcabe().* Esta se encarga de calcular componentes LS SPCA mediante el recorte de forma iterativa de pequeñas cargas en cada eje.

Es claro que, aun así, R contiene otros paquetes disponibles para el cálculo tanto del PCA cómo del SPCA (Tabla 1).

Tabla 1 Paquetes disponibles en el programa R para PCA y SPCA

Paquete	Descripción	Enlace
bootSVD	Análisis de Componentes Principales con Bootstrap para altas dimensiones	http://cran.r-project.org/web/packages/bootSVD/index.html
bpca	Biplot de Datos Multivariantes basado en el Análisis de Componentes Principales	http://cran.r-project.org/web/packages/bpca/index.html
cpca	Methods to perform Common Principal Component Analysis (CPCA)	http://cran.r-project.org/web/packages/cpca/index.html
eigenprcomp	Computa intervalos de confianza para el PCA ordinario	http://cran.r-project.org/web/packages/eigenprcomp/index.html
FastHCS	Algoritmo Robusto para el Análisis de Componentes Principales	http://cran.r-project.org/web/packages/FastHCS/index.html
fpca	MLE restringido para PCA Funcional	http://cran.r-project.org/web/packages/fpca/index.html
gPCA	Componentes Principales Guiadas	http://cran.r-project.org/web/packages/gPCA/index.html
GPCSIV	Componentes Principales generalizadas para variables de intervalo simbólico	http://cran.r-project.org/web/packages/GPCSIV/index.html
jackstraw	Jackstraw no paramétrico para Análisis de Componentes Principales	http://cran.r-project.org/web/packages/jackstraw/index.html
lpc	Componentes Principales con *Lasso*	http://cran.r-project.org/web/packages/lpc/index.html
missMDA	Manejo de valores perdidos con/en el análisis de datos multivariantes (métodos de componentes principales)	http://cran.r-project.org/web/packages/missMDA/index.html
onlinePCA	Análisis de Componentes Principales Online	http://cran.r-project.org/web/packages/onlinePCA/index.html

Paquete	Descripción	Enlace
paran	Test de Horn's de Componentes Principales/Factores	http://cran.r-project.org/web/packages/paran/index.html
pcaBootPlot	Creación de gráficos 2-D de Componentes Principales a través de bottstrap	http://cran.r-project.org/web/packages/pcaBootPlot/index.html
PCGSE	Componentes Principales para el enriquecimiento de conjunto de genes	http://cran.r-project.org/web/packages/PCGSE/index.html
plotpc	Histogramas de Componentes Principales a través de diagramas de dispersión	http://cran.r-project.org/web/packages/plotpc/index.html
pls	Mínimos Cuadrados Parciales y regresión de Componentes Principales	http://cran.r-project.org/web/packages/pls/index.html
sGPCA	Análisis de Componentes Principales *Sparse* Generalizado	http://cran.r-project.org/web/packages/sGPCA/index.html
SpatPCA	Regularized Principal Component Analysis for Spatial Data	http://cran.r-project.org/web/packages/SpatPCA/index.html
Spca	Análisis de Componentes Principales *Sparse*	http://cran.r-project.org/web/packages/spca/index.html
Spcr	Regresión de Componentes Principales *Sparse*	http://cran.r-project.org/web/packages/spcr/index.html
Superpc	Componentes Principales Supervisadas	http://cran.r-project.org/web/packages/superpc/index.html

Fuente:(Team, 2015)

7.2. Datos *Pitprops*: Apoyos de madera

La matriz de datos a utilizar se conoce como *Pitprops*. Fue introducida por Jeffers (1967) y se trata de una matriz que recoge la información de 180 apoyos de madera (observaciones) que dan soporte estructural a las minas subterráneas. Sobre ellas se han medido una serie de 13 variables que miden características físicas asociadas a la resistencia de estos cortes de madera, como son el diámetro de la base, la longitud del apoyo y la cantidad de anillos.

A pesar de que se conoce la estructura inicial de los datos, únicamente queda disponible para su uso la matriz de correlaciones muestral entre las 13 variables mencionadas Jeffers (1967). Este es un ejemplo clásico que muestra la dificultad de analizar una matriz de datos cuando las variables están correlacionadas entre si y el problema para interpretar las componentes principales.

Debido a ciertas particularidades estos datos han sido citados por diferentes autores como (Jolliffe et al., 2003; Koch, 2013; Hui Zou et al., 2006) para enseñar la dificultad de analizar una matriz de datos cuando existe alta correlación entre las variables y el problema para interpretar las componentes principales como variables nuevas que surgen de la combinación lineal de las observadas. Estos datos se consideran como un ejemplo clásico para ilustrar de forma clara los beneficios de cada mejora a la técnica de análisis de componentes principales y conforman un claro ejemplo de conjunto con $n \gg p$, siendo n el número de observaciones y p el número de variables.

El fin práctico de este caso de estudio es determinar las características, o combinación de variables, que hacen que los apoyos de madera cortados sean lo suficientemente fuertes como para su uso en las minas. El diseño experimental implementado por Jeffers (1967) se basó en escoger al azar distintos apoyos de una población definida de árboles, teniendo en cuenta el tamaño, la clase y la región geográfica. La decisión se tomará en base a si los apoyos son suficientemente fuertes para su uso en las minas.

- DIAMSUP: El diámetro superior de la hélice en pulgadas.
- LONGHE: La longitud de la hélice en pulgadas.
- PHUMHE: El contenido de humedad de la hélice, expresado como un porcentaje del peso seco.
- GRAPRU: La gravedad específica de la madera en el momento de la prueba.
- GRAHOR: La gravedad específica de la madera en el horno.
- ANISUP: El número de anillos anuales en la parte superior de la hélice.
- ANIBAS: El número de anillos anuales en la base de la hélice.

- ARCMAX: El arco máximo en pulgadas.
- DISMAX: La distancia del punto de máxima arco desde la parte superior de la hélice en pulgadas.
- VERNUD: El número de verticilos por nudo.
- LONGAP: La longitud de apoyo claro a partir de la parte superior de la hélice en pulgadas.
- NUDOS: El número medio de nudos por verticilo.
- DIAMNUD: El diámetro promedio de los nudos en pulgadas.

7.2.1. Análisis de las correlaciones

En la Tabla 2 se enseña la correlación entre cada par de variables, donde se puede observar que en 3 pares existe una asociación mayor de 0.8 lo que podría entenderse como una relación lineal alta entre las variables implicadas, en 8 pares existe una asociación mayor de 0.5 pero menor de 0.8 lo que podría entenderse como una relación lineal media entre las variables implicadas, en 44 pares se da una asociación mayor de 0.1 pero menor de 0.5 lo que podría entenderse como una relación lineal baja entre las variables implicadas y en los 23 pares restantes se detectó una asociación casi nula con una relación lineal menor de 0.1.

Tabla 2 Matriz de correlaciones apoyos de madera

Variable	Diamsup	Longhe	Phumhe	Grapru	Grahor	Anisup	Anibas	Arcmax	Dismax	Vernud	Longap	Nudos	Diamnud
Diamsup	1												
Longhe	0.954	1											
Phumhe	0.364	0.297	1										
Grapru	0.342	0.284	0.882	1									
Grahor	-0.129	-0.118	-0.148	0.22	1								
Anisup	0.313	0.291	0.153	0.381	0.364	1							
Anibas	0.496	0.503	-0.029	0.174	0.296	0.813	1						
Arcmax	0.424	0.419	-0.054	-0.059	0.004	0.09	0.372	1					
Dismax	0.592	0.648	0.125	0.137	-0.039	0.211	0.465	0.482	1				
Vernud	0.545	0.569	-0.081	-0.014	0.037	0.274	0.679	0.557	0.526	1			
Longap	0.084	0.076	0.162	0.097	-0.091	-0.036	-0.113	0.061	0.085	-0.319	1		
Nudos	-0.019	-0.036	0.22	0.169	-0.145	0.024	-0.232	-0.357	-0.127	-0.368	0.029	1	
Diamnud	0.134	0.144	0.126	0.015	-0.208	-0.329	-0.424	-0.202	-0.076	-0.291	0.007	0.184	1

Como es sabido, en la estadística Invariante como la regresión lineal ordinaria existen dificultades para ajustar modelos en presencia de relaciones lineal altas entre las variables, colinealidad, dado que entre los supuestos de estos modelos se tiene la hipótesis de que las variables son independientes entre sí, lo cual que comúnmente no sucede en la práctica. Además, el hecho de que la correlación sea muy baja o casi nula como sucede en este ejemplo, no garantiza que ambas variables no tengan otro tipo de asociación, ya que el coeficiente de correlación sólo está midiendo el grado de relación lineal.

Por lo tanto el Análisis de Componentes Principales surge como una técnica multivariante para aprovechar las relaciones existentes entre las variables, evitando perder información y poder así explicar de mejor forma la variabilidad de los datos.

7.2.2. Análisis de Componentes Principales

Como se ha dicho, el primer acercamiento al PCA se hace desde la función de R *prcomp()*, la cual calcula componentes principales por medio de la Descomposición en Valores Singulares (SVD), en este caso, de la matriz de Correlación centrada por columnas.

Valores propios

A partir de la SVD se obtiene los valores singulares de la Tabla 3, donde se puede observar que hay cuatro valores mayores que uno dando a entender que con los cuatro primeros ejes, o componentes principales, asociados a los primeros planos de proyección generados los vectores propios de la descomposición sería suficiente para explicar la variabilidad de los datos. Sin embargo para evitar la pérdida de información asociados a los valores propios cercanos a uno, autores como Jeffers (1967), Zou, Hastie, & Tibshirani (2006) y Trendafilov (2014) basaron su estudio en la interpretación las seis primeras componentes principales. Además, La proporción de varianza explicada por cada componente principal decrece y a partir de la sexta componente la varianza explicada por cada eje no aumenta significativamente. Por ello, y a lo largo de este trabajo, el número de dimensiones elegidas para representar la información será seis, reteniendo un 87% de la información total.

Gráficamente puede analizarse esta información a través del Scree Plot (Figura 8) donde se representa el valor propio respectivo a cada componente principal. O a través de la Figura 9, que representa el porcentaje de varianza explicada acumulada por las componentes principales. Puede observarse cómo reteniendo cuatro dimensiones ya se

obtendría una buena imagen de los datos originales, pues se estaría reteniendo un 73.7% de la varianza total.

Tabla 3 Valores propios y varianza explicada por el PCA

Componente Principal	Valores Propios	% Varianza Explicada	% Varianza Acumulada
PC1	4,2186	32,5	32,5
PC2	2,3781	18,3	50,7
PC3	1,8782	14,4	65,2
PC4	1,1094	8,5	73,7
PC5	0,9100	7	80,7
PC6	0,8154	6,3	87
PC7	0,5763	4,4	91,4
PC8	0,4396	3,4	94,8
PC9	0,3527	2,7	97,5
PC10	0,1908	1,5	99
PC11	0,0506	0,4	99,4
PC12	0,0415	0,3	99,7
PC13	0,0387	0,3	100

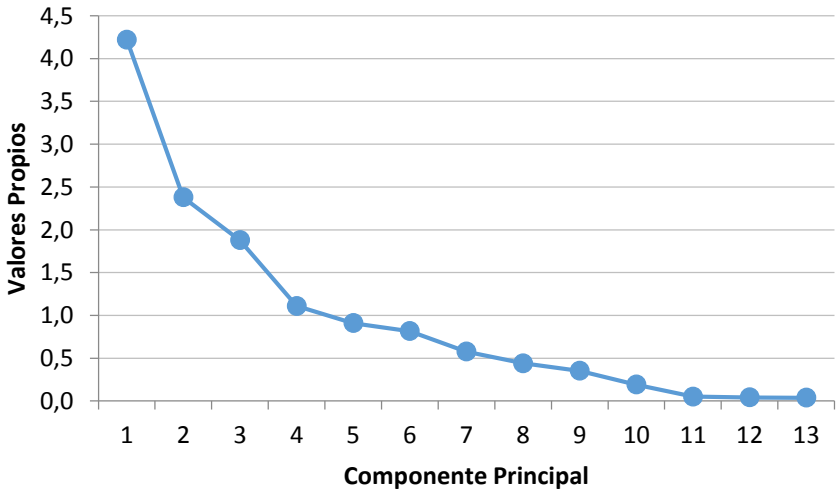

Figura 8 Scree Plot-Pitprops

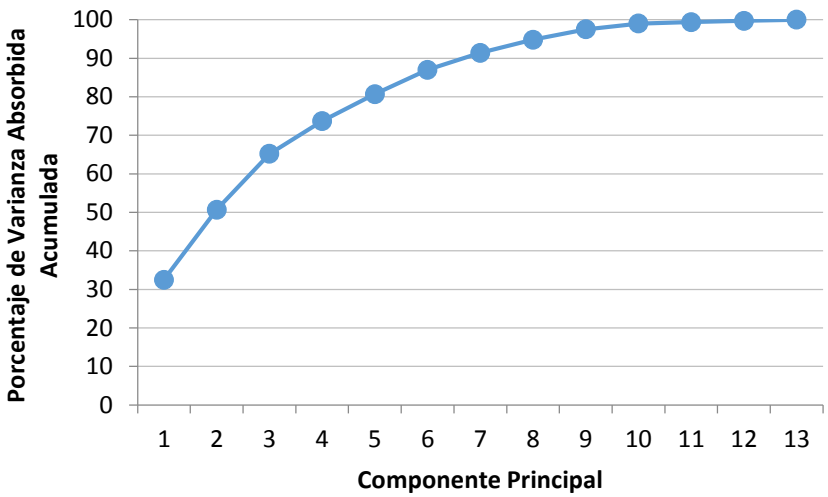

Figura 9 Porcentaje de Varianza Absorbida por cada componente principal

En relación con la cantidad de información que cada componente representa de una variable específica, se tienen las cargas de la Tabla 4, donde se resaltan las variables que tienen una carga superior al 20% en cada eje. En las componentes 1, 5 y 6 se destacan nueve variables, seis en la PC2, tres en la PC3, tres en la PC4.

En la componente 1, 2 y 3 destacan 7 variables, 4 en PC4, 6 en PC5 y 3 en PC6.

La matriz de cargas muestra la dificultad que los autores mencionaban en la interpretación de esta matriz.

Tabla 4 Cargas PCA

	PC1	PC2	PC3	PC4	PC5	PC6
Diamsup	**0.4038**	**0.2179**	0.2073	0.0912	0.0826	0.1198
Longhe	**0.4055**	0.1861	0.235	0.1027	0.1128	0.1629
Phumhe	0.1244	**0.5406**	-0.1415	-0.0784	**-0.3498**	**-0.2759**
Grapru	0.1732	**0.4556**	**-0.3524**	-0.0548	**-0.3558**	-0.054
Grahor	0.0575	-0.1707	**-0.4812**	-0.0491	-0.1761	**0.6256**
Anisup	**0.2844**	-0.0142	**-0.4753**	0.0634	**0.3158**	0.0523
Anibas	**0.3998**	-0.1896	**-0.2531**	0.065	**0.2151**	0.0027
Arcmax	**0.2936**	-0.1892	**0.2431**	**-0.2855**	-0.1853	-0.0551
Dismax	**0.3566**	0.0171	**0.2076**	-0.0967	0.1061	0.0342
Vernud	**0.3789**	**-0.2485**	0.1188	0.205	-0.1564	-0.1731
Longap	-0.0111	**0.2053**	0.0705	**-0.8037**	**0.343**	0.1753
Nudos	-0.1151	**0.3432**	-0.092	**0.3008**	**0.6004**	-0.1698
Diamnud	-0.1125	**0.3085**	**0.3261**	**0.3034**	-0.0799	**0.6263**

Tal como se enseñó en la Tabla 4, en la Figura 10 se ilustra cómo las variables que se asocian con valores positivos en la PC1 son longhe, diamsup y anibas, y con valores negativos, nudos y diamnud. Asimismo puede decirse que el comportamiento de la variable longap no es bien representado por la componente principal.

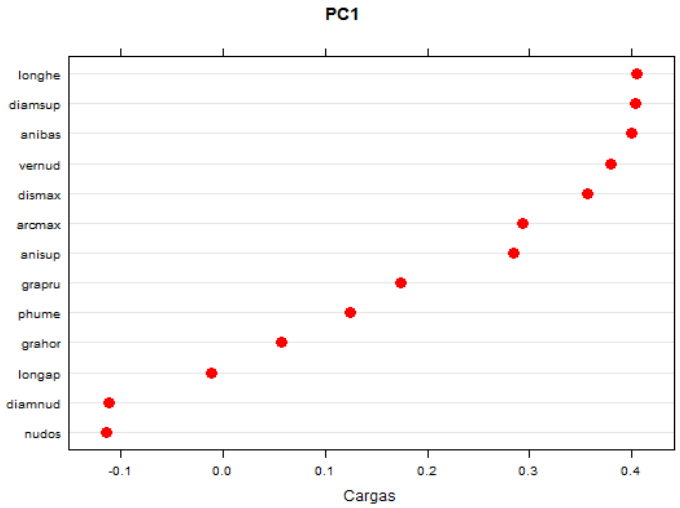

Figura 10 Cargas de las variables en la PC1

De igual forma, en la Figura 11 se ilustra cómo los valores positivos en la PC2 se asocian a las variables grapru y phumhe, las cuales no estaba bien representada en el eje anterior dado que se supone que la PC2 es ortogonal a la PC1. Los valores negativos en la PC2 se identifican con variables como grahor, arcmax, anibas y vernud. Asimismo puede decirse que el comportamiento de la variable grahor no está bien representado ni en componente principal uno ni en la dos, dado que como se enseñó en la Tabla 4, esta variable carga principalmente en las componentes PC3 y PC6.

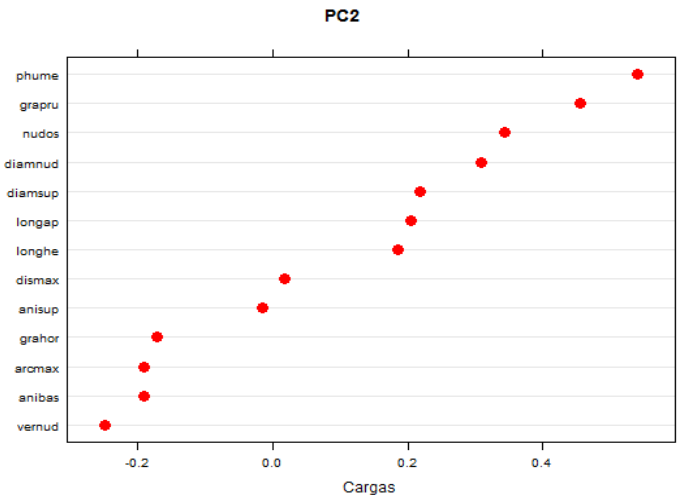

Figura 11 Cargas de las variables en la PC2

Ya estas figuras vislumbran la dificultad de interpretación de las componentes principales, pues no existen variables claras de las que dependa la interpretación de una componente principal, por tener todas magnitudes pequeñas y similares, la mayoría con valor en el intervalo $(-0.2, 0.4)$ y pocas de ellas nulas. Esto dificulta la elección de un umbral por parte del usuario por debajo del cual considerar toda carga como nula.

7.2.3. Rotación VARIMAX

Se examinan los resultados obtenidos tras aplicar la rotación VARIMAX del programa R sobre la matriz de datos *Pitprops*.

Como ya se ha mencionado, la rotación VARIMAX (Kaiser 1958) se encarga de rotar los factores obtenidos forzando a que unas cargas se aproximen más a uno y las otras a cero. Todo ello para facilitar su interpretación.

Al realizar esta rotación a las cargas de la Tabla 4, se obtienen las cargas rotadas de la Tabla 5. Se puede observar que, en comparación con la Tabla 4, las cargas de las variables que estaban poco representadas en una componente específica, se hicieron mucho más pequeñas, siendo algunas casi iguales a cero. Las variables que poseían cargas altas, han obtenido magnitudes mayores en este caso, como es el caso de la componente PC6. Sin embargo, aun así, la interpretación de las variables continúa siendo complicada, pues las componentes principales, aun eliminando subjetivamente aquellas variables con cargas por debajo del 0.2, siguen siendo combinación lineal de un número alto de variables.

Además, gracias a la rotación, se observa aún mejor como las últimas componentes poseen variables más importantes, con mayores magnitudes, que las primeras, y esto ha aumentado en parte en la aplicación de la rotación. Por ejemplo, mientras que tras el PCA 3 variables en la PC6 tenían realmente importancia, en este caso el número de variables importantes, determinantes de la componente son 6.

Tabla 5 Cargas rotadas

	PC1	PC2	PC3	PC4	PC5	PC6
Diamsup	**0,4038**	0,1244	-0,2844	0,0111	-0,1151	-0,1125
Longhe	**0,2179**	**0,5406**	0,0142	-0,2053	**0,3432**	**0,3085**
Phumhe	**0,2073**	-0,1415	**0,4753**	-0,0705	-0,092	**0,3261**
Grapru	0,0912	-0,0784	-0,0634	**0,8037**	**0,3008**	**0,3034**
Grahor	0,0826	**-0,3498**	**-0,3158**	**-0,343**	**0,6004**	-0,0799
Anisup	0,1198	**-0,2759**	-0,0523	-0,1753	-0,1698	**0,6263**

	PC1	PC2	PC3	PC4	PC5	PC6
Anibas	0,1108	0,022	**-0,2962**	-0,1415	**-0,5387**	0,1577
Arcmax	0,1361	0,0021	-0,1472	-0,0067	**0,2103**	0,1144
Dismax	**0,3329**	-0,0782	**0,4109**	-0,1502	0,0799	**-0,3839**
Vernud	**0,3081**	-0,0608	-0,1012	**0,3369**	-0,1942	**-0,3279**
Longap	0,0047	-0,1173	**-0,5371**	-0,0484	-0,0471	-0,0447
Nudos	**0,3916**	**0,5266**	-0,0798	-0,0047	0,0023	0,0131
Diamnud	**0,572**	**-0,4077**	0,0567	-0,0074	-0,004	0,0094

7.2.4. Elastic net

Como se mencionó el Criterio SPCA se puede resolver mediante la imposición de un vector de cardinalidades adecuado para cada conjunto de datos. Para determinar el número de cargas óptimo para el caso de los apoyos de manera se recurre a los resultados del PCA ordinario, donde se determina la cardinalidad de cada componente sparse como la cantidad de cargas con magnitudes superiores a 0.3 del PCA ordinario, obteniendo los resultados de la Tabla 6. En la misma tabla se puede ver que la cantidad total de varianza explicada por las componentes principales *sparse* es del 72.3%. También, se puede observar que las variables diamsup, longhe, dismax cargan principalmente en la PC 1 y que en la segunda componente se encuentran las variables phume y arcmax.

Tabla 6 Solución Sparse para apoyos de madera con vector de cardinalidad

	PC1	PC2	PC3	PC4	PC5	PC6
Diamsup	-0.6	0.026	0	0	0	0
Longhe	-0.624	0	0	0	0	0
Phume	0	0.692	0	0	0	0
Grapru	0	0.721	0	0	0	0
Grahor	0.081	0	0.571	0	0	0.057
Anisup	0	0	0.526	0	0	0
Anibas	0	0	0.63	0	0	0
Arcmax	-0.103	-0.006	0	0	0.497	0
Dismax	-0.456	0	0	0	0	0
Vernud	-0.163	0	0	0.368	0.117	0
Longap	0	0	0	-0.93	0	0
Nudos	0	0	0	0	-0.86	0
Diamnud	0	0	0	0	0	0.998

	PC1	PC2	PC3	PC4	PC5	PC6
% varianza	22.30%	13.50%	13.70%	8.40%	8.40%	6%
Cardinalidad	6	4	3	2	3	2

Sin embargo como el interés real del Criterio SPCA es la definición de los valores para las penalizaciones *Ridge* y *lasso* se genera otra solución para el mimo conjunto de datos pero con un vector de escalares igual a (0.06, 0.16, 0.1, 0.5, 0.5, 0.5), los cuales fueron escogidos por Zou et al., (2006) como los óptimos para la penalización lasso en este ejemplo. Recordar que en la función que resuelve el Criterio SPCA el valor de la penalización Ridge ya viene fijado por defecto en $1e^{-6}$.

En la Tabla 7 se puede ver que la cantidad total de varianza explicada por las componentes principales *spars*e es del 75.7%, ligeramente superior a la solución anterior. También, se puede observar que las variables con mayor carga en las componentes principales son similares a la solución con el vector de cardinalidades, sin embargo se observa una mejora en la interpretación de las últimas tres componentes principales donde solamente carga una variable.

Tabla 7 Solución Sparse para apoyos de madera con vector de penalización

	PC1	PC2	PC3	PC4	PC5	PC6
Diamsup	-0.477	0	0	0	0	0
Longhe	-0.476	0	0	0	0	0
Phume	0	0.785	0	0	0	0
Grapru	0	0.619	0	0	0	0
Grahor	0.177	0	0.641	0	0	0
Anisup	0	0	0.589	0	0	0
Anibas	-0.25	0	0.492	0	0	0
Arcmax	-0.344	-0.021	0	0	0	0
Dismax	-0.416	0	0	0	0	0
Vernud	-0.4	0	0	0	0	0
Longap	0	0	0	-1	0	0
Nudos	0	0.013	0	0	-1	0
Diamnud	0	0	-0.016	0	0	1
% varianza	28.00%	14.00%	13.30%	7.40%	6.80%	6%
Cardinalidad	7	4	4	1	1	1

En la Figura 12 se puede ver gráficamente como la solución *sparse* le indica claramente al investigador que con un primer plano se puede describir el comportamiento de la resistencia de los apoyos de madera interpretando que el eje uno viene dado por la combinación del diámetro superior, la longitud de la hélice y la distancia del punto de máxima arco desde la parte superior de la hélice y el eje dos agrupa al porcentaje de humedad y el arco máximo del apoyo.

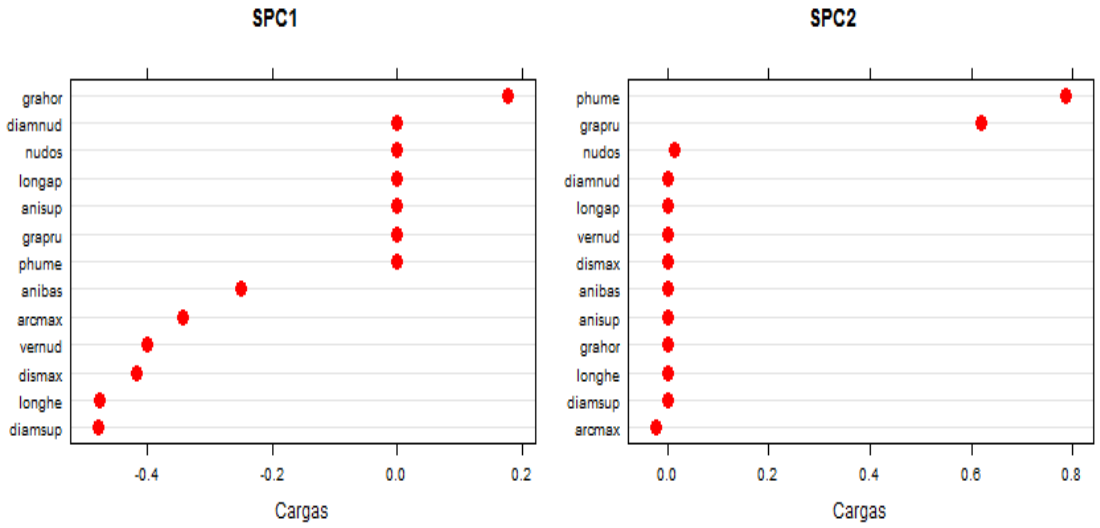

Figura 12 Componentes sparse 1 y 2 para los apoyos de madera

7.3. Datos: Yogures

La matriz de datos empleada para el análisis de las técnicas trata un estudio de 38 yogures, para cada uno de los cuales se analiza una serie de 30 compuestos propios de ellos, que incluye entre otros el contenido calórico del mismo, o sus nutrientes principales. Estos datos forman un claro ejemplo de conjunto con $n > p$, siendo n el número de observaciones y p el número de variables.

La matriz mencionada recoge datos provenientes de la base de datos BEDCA (AESAN, 2010). La Base de Datos Española de Composición de Alimentos, publicada por la Red BEDCA del Ministerio de Ciencia e Innovación, surge bajo la coordinación y financiación de la Agencia Española de Seguridad Alimentaria y Nutrición del Ministerio de Sanidad, Servicios Sociales e Igualdad. Los valores de composición de alimentos recogidos en esta base de datos han sido obtenidos de distintas fuentes que incluyen laboratorios, industria alimentaria y publicaciones científicas o valores directamente calculados.

La matriz de datos en estudio es un ejemplo que muestra la dificultad de analizar una matriz de datos cuando las variables están correlacionadas entre sí y el problema para interpretar las componentes principales.

Este ejemplo se basa en las observaciones realizadas a distintos tipos de yogur. El estudio se realiza tomando una muestra de 38 individuos que se corresponden con 38 tipos de maíz, y 30 variables cuantitativas medidas sobre cada una de las variedades, que se corresponden con características de composición propios de cada uno de los lácteos.

Los tipos de yogur sometidos a análisis son las variedades: **Yogur griego, Yogur líquido, Yogur líquido aromatizado, Yogur líquido con frutas, Yogur líquido entero, con cereales, Yogur líquido natural azucarado, Yogur búlgaro, Yogur desnatado, aromatizado, Yogur desnatado con cereales, Yogur desnatado con cereales, manzana y ciruela, Yogur desnatado con cereza y frambuesa, Yogur desnatado con ciruela, albaricoque y fibra, Yogur desnatado con fresa, grosella y fibra, Yogur desnatado con frutas, Yogur desnatado con frutas del bosque, Yogur desnatado con frutas tropicales, Yogur desnatado con manzana, Yogur desnatado con melocotón y maracuyá, Yogur desnatado con melocotón, frambuesa y fibra, Yogur desnatado con piña y pomelo, Yogur desnatado sabor natural, Yogur desnatado sabor natural azucarado, Yogur desnatado sabor vainilla, Yogur enriquecido con frutas, Yogur enriquecido natural, Yogur enriquecido natural azucarado, Yogur enriquecido natural con nata, Yogur enriquecido con sabor, Yogur entero con cereales y fresas, Yogur entero con fresas, Yogur entero con frutas del bosque, Yogur líquido desnatado natural, Yogur líquido entero con fresas, Yogur líquido entero con frutas, Yogur líquido entero sabor fresa, Yogur líquido entero sabor fresa y plátano, Yogur líquido entero sabor frutas del bosque y Yogur líquido entero sabor piña y coco.**

Los compuestos analizados sobre cada uno de los tipos de maíz son características referentes a la composición de los lácteos y relacionadas con las propiedades de cada uno de ellos: **energía total en Kcal, grasa, proteínas, agua, fibra, carbohidratos, ácidos grasos monoinsaturados y poliinsaturados, ácidos grasos saturados totales, colesterol, Vitamina A, Vitamina D, Vitamina E, ácido fólico (vitamina B-9), niacina (vitamina B-3), riboflavina (Vitamina B-2), tiamina (Vitamina B-1), Vitamina B-12, Vitamina B-6, Vitamina C, calcio, hierro, potasio, magnesio, sodio, fósforo, ioduro, selenio, zinc**. En ellas se recoge información de forma cuantitativa.

- Energia_kcal: Energía Total (Kcal).

- Grasa: Total de grasa (lípidos totales) (g).
- Proteinas: Total de proteínas (g).
- Agua: Humedad (g).
- Fibra: Total de fibra (dietética) (g).
- Carbohidratos: Cantidad de carbohidratos (g).
- AGrasos_mono: Ácidos grasos monoinsaturados totales (g).
- AGrasos_poli: Ácidos grasos poliinsaturados totales (g).
- AGrasos_Saturados: Ácidos grasos saturados totales (g).
- Colesterol: Nivel de colesterol (g).
- VitA: Cantidad de vitamina A equivalentes de retinol de actividades de retinos y carotenoides (ug).
- VitD: Cantidad de vitamina D (ug).
- VitE: Cantidad de vitamina E equivalentes de alfa tocoferol de actividades de vitámeros E (mg).
- VitB9: Folato total (ug).
- VitB3: Cantidad de equivalentes de niacina totales (mg).
- VitB2: Cantidad de riboflavina (mg).
- VitB1: Cantidad de tiamina (mg).
- VitB12: Cantidad de vitamina B-12 (ug).
- VitB6: Vitamina B-6 total (mg).
- VitC: Cantidad de Vitamina C (ácido ascórbico) (mg).
- Calcio: Calcio total (mg).
- Hierro: Hierro total (mg).
- Potasio: Potasio total (mg).
- Magnesio: Magnesio total (mg).
- Sodio: Sodio total (mg).
- Fósforo: Fósforo total (mg).
- Ioduro: Ioduro total (ug).
- Selenio: Selenio total (ug).
- Zinc: Zinc total (mg).

La matriz de datos que se conforma es una matriz de dimensión $38x30$, con tipos de yogur en filas y compuestos en columnas.

7.3.1. Análisis de las correlaciones

El coeficiente de correlación de Pearson es una medida que permite detectar la relación lineal entre dos variables aleatorias cuantitativas, independiente de la escala de medida de las variables (Pearson, 1926).

A partir de la Tabla 8, donde se muestra las correlaciones entre cada par de variables, se puede observar que en 16 pares de variables existe una asociación mayor de 0.8 lo que podría entenderse como una correlación alta entre las variables implicadas. En 44 pares de variables existe una asociación mayor de 0.5 pero menor de 0.8 lo que podría entenderse como una correlación media entre las variables implicadas y 258 pares tienen una asociación mayor de 0.1, pero menor de 0.5 lo que podría entenderse como una correlación baja entre las variables implicadas. En 102 pares de variables se detectó una asociación casi nula con un coeficiente de correlación menor de 0.1.

Tabla 8 Matriz de correlaciones - Yogures

	Energia_kcal	Grasa	Proteinas	Agua	Fibra	Carbohidratos	AGrasos_mono	AGrasos_poli	AGrasos_Saturados	Colesterol	VitA	VitD	VitE	VitB9	VitB3	VitB2	VitB1	VitB12	VitB6	VitC	Calcio	Hierro	Potasio	Magnesio	Sodio	Fosforo	Ioduro	Selenio	Zinc
Energia_kcal	1.00																												
Grasa	0.66	1.00																											
Proteinas	0.14	0.29	1.00																										
Agua	-0.91	-0.37	-0.05	1.00																									
Fibra	-0.54	-0.55	-0.09	0.31	1.00																								
Carbohidratos	0.83	0.14	-0.16	-0.92	-0.31	1.00																							
AGrasos_mono	0.67	0.97	0.35	-0.41	-0.51	0.16	1.00																						
AGrasos_poli	0.56	0.81	0.36	-0.36	-0.25	0.12	0.90	1.00																					
AGrasos_Saturados	0.65	0.99	0.30	-0.36	-0.51	0.12	0.98	0.84	1.00																				
Colesterol	0.72	0.74	0.16	-0.54	-0.45	0.41	0.72	0.56	0.73	1.00																			
VitA	-0.02	0.09	0.21	0.16	0.24	-0.12	0.06	0.17	0.10	-0.01	1.00																		
VitD	-0.08	-0.06	-0.08	0.18	0.13	-0.05	-0.18	-0.21	-0.08	0.02	0.55	1.00																	
VitE	-0.16	-0.19	0.08	0.20	0.47	-0.09	-0.20	-0.07	-0.18	-0.19	0.84	0.66	1.00																
VitB9	0.16	0.15	0.10	-0.08	-0.12	0.10	0.23	0.32	0.16	0.36	0.01	-0.30	-0.22	1.00															
VitB3	0.29	0.40	0.27	-0.07	-0.35	0.08	0.26	0.15	0.37	0.29	0.45	0.36	0.24	-0.17	1.00														
VitB2	0.13	0.50	0.62	0.02	-0.08	-0.25	0.51	0.58	0.52	0.07	0.18	-0.23	-0.07	0.00	0.38	1.00													
VitB1	0.24	0.24	0.02	-0.06	-0.55	0.16	0.13	-0.03	0.19	0.40	-0.01	0.01	-0.11	0.27	0.52	0.04	1.00												
VitB12	-0.37	-0.34	0.05	0.41	-0.03	-0.26	-0.43	-0.54	-0.35	-0.35	0.07	0.18	0.04	-0.23	0.26	0.07	0.27	1.00											
VitB6	-0.58	-0.55	-0.04	0.30	0.62	-0.37	-0.54	-0.37	-0.53	-0.56	-0.24	-0.27	-0.23	-0.11	-0.34	0.04	-0.44	0.21	1.00										
VitC	-0.23	-0.36	-0.02	0.05	0.26	-0.05	-0.35	-0.23	-0.36	-0.26	-0.09	-0.03	0.02	-0.51	-0.05	0.02	-0.29	-0.06	0.32	1.00									
Calcio	-0.26	0.02	0.21	0.33	-0.09	-0.37	-0.03	-0.02	0.02	-0.17	0.08	-0.13	-0.06	-0.17	0.27	0.50	0.37	0.58	0.15	0.18	1.00								
Hierro	-0.01	0.15	-0.03	0.01	0.45	-0.11	0.18	0.34	0.18	0.04	0.22	-0.02	0.26	0.14	-0.08	0.21	-0.11	-0.28	0.07	-0.17	0.01	1.00							
Potasio	-0.32	-0.20	0.08	0.32	-0.06	-0.28	-0.24	-0.26	-0.20	-0.25	-0.05	-0.13	-0.19	-0.03	0.20	0.25	0.35	0.68	0.31	0.13	0.83	-0.04	1.00						
Magnesio	-0.30	-0.26	0.06	0.22	0.11	-0.22	-0.32	-0.34	-0.26	-0.33	-0.07	-0.08	-0.06	-0.37	0.05	0.15	0.07	0.48	0.36	0.33	0.54	0.18	0.55	1.00					
Sodio	-0.30	-0.01	0.21	0.37	0.06	-0.41	-0.04	0.02	0.01	-0.25	0.17	-0.13	0.00	-0.06	0.28	0.57	0.23	0.53	0.23	0.12	0.85	0.21	0.84	0.60	1.00				
Fosforo	0.24	0.34	0.65	-0.17	-0.43	-0.01	0.39	0.30	0.33	0.26	-0.34	-0.50	-0.48	0.22	0.03	0.40	0.21	-0.01	-0.04	-0.07	0.20	-0.24	0.13	0.07	0.11	1.00			
Ioduro	-0.12	-0.05	0.00	0.15	0.10	-0.11	-0.10	-0.15	-0.05	0.27	0.06	0.10	0.04	0.27	-0.06	-0.03	0.36	0.40	0.01	-0.12	0.42	0.11	0.39	0.19	0.29	0.00	1.00		
Selenio	0.28	0.30	0.03	-0.06	-0.39	0.16	0.13	-0.09	0.26	0.44	0.39	0.49	0.24	0.01	0.68	0.03	0.67	0.41	-0.47	-0.26	0.24	-0.17	0.21	0.09	0.14	-0.09	0.50	1.00	
Zinc	-0.12	0.14	0.33	0.28	-0.22	-0.29	0.13	0.08	0.15	0.08	0.14	-0.08	-0.02	0.16	0.18	0.44	0.39	0.58	-0.13	-0.11	0.76	-0.04	0.63	0.26	0.67	0.30	0.67	0.39	1.00

7.3.2. Análisis de Componentes Principales

El primer acercamiento al PCA se hace por medio de la Descomposición en Valores Singulares de la matriz de Correlación centrada por columnas.

Valores propios

De dicha descomposición se obtiene el vector de valores propios (Tabla 9). Existen siete valores propios mayores que uno dando a entender que con los siete primeros ejes, o componentes principales, sería suficiente para explicar la variablidad de los datos. En cuanto a la varianza absorbida por las componentes principales, ya con 5 componentes se retendría un 75.21% de la varianza total, pero con la retención de 7 dimensiones se absorbe un 84.83%, porcentaje adecuado para explicar una alta proporción de información. Por ello, y a lo largo de este trabajo, el número de dimensiones elegidas para representar la información será siete.

Tabla 9 Varianza explicada por cada componente

Componente Principal	***Valores Propios***	***% Varianza Explicada***	***%Varianza Acumulada***
PC1	7.5345	25.98%	25.98%
PC2	5.5732	19.22%	45.20%
PC3	3.5606	12.28%	57.48%
PC4	2.9890	10.31%	67.78%
PC5	2.0106	6.93%	74.72%
PC6	1.5773	5.44%	80.16%
PC7	1.2216	4.21%	**84.37%**
PC8	0.9372	3.23%	87.60%
PC9	0.7266	2.51%	90.11%
PC10	0.6907	2.38%	92.49%
PC11	0.6302	2.17%	94.66%
PC12	0.3521	1.21%	95.87%
PC13	0.2887	1%	96.87%
PC14	0.2230	0.77%	97.64%

Componente Principal	*Valores Propios*	*% Varianza Explicada*	*%Varianza Acumulada*
PC15	0.1628	0.56%	98.20%
PC16	0.1196	0.41%	98.61%
PC17	0.1137	0.39%	99.00%
PC18	0.1016	0.35%	99.35%

Esta información de decisión acerca del número de componentes óptimas queda respaldada por el Scree-Plot (Figura 13) y por el gráfico de varianza acumulada por las componentes principales (Figura 14).

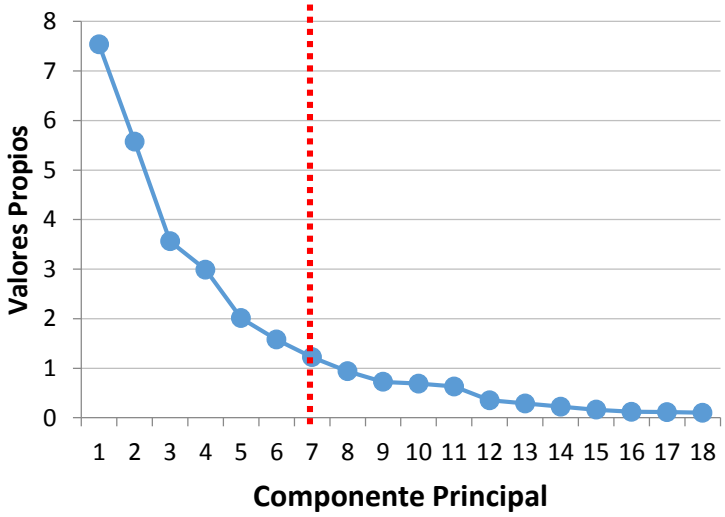

Figura 13 Gráfico de sedimentación o Scree-Plot

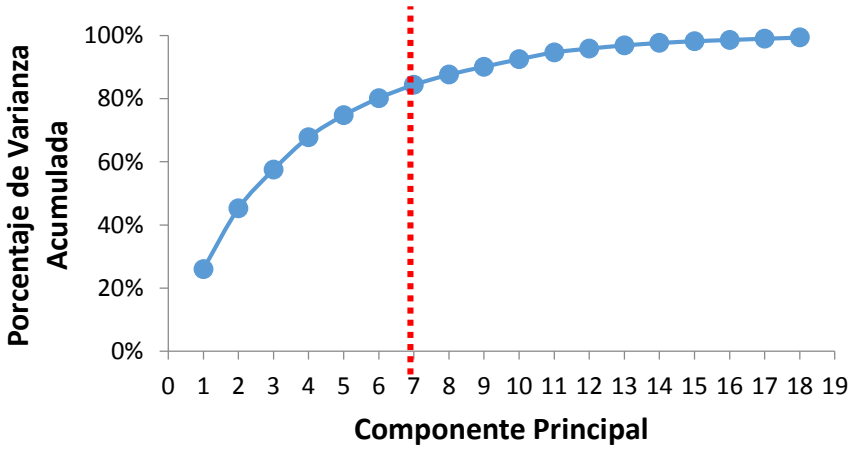

Figura 14 Porcentaje de varianza acumulada por componentes

En relación con la cantidad de información que cada componente representa de una variable específica, se tienen las cargas de la Tabla 10, donde se resaltan las variables que tienen una carga superior al 20% en cada eje. En la componente 1 se destacan 9 variables, 10 en PC2, 7 en PC3, 8 en PC4,4 en PC5,6 en PC6 y 8 en PC7. Sin embargo, hay que destacar que la carga más alta se encuentra en la PC6 con un valor de 0.5768, sin la existencia de ninguna carga mayor a esta en ninguna componente.

Ya en estos primeros resultados se observan las dificultades existentes para la interpretación de cada una de las componentes, pues puede observarse que, prácticamente la mayoría de las cargas no son mayores a 0.5, con lo que no se identifican variables claramente importantes en las PCs. Cada una de las PCs es combinación lineal de todas las variables originales y, siguiendo el método más utilizado en la práctica (haciendo exactamente nulas las cargas cercanas a 0), el número de variables originales que se combinan para dar lugar a las PCs continúa siendo alto. El número de cargas con una magnitud por debajo de 0.1 no es excesivamente bajo (10, 12, 16, 11, 15, 15 y 14, respectivamente, en cada componente), lo que unido al hecho de que prácticamente la mayor parte de la magnitud de las cargas se encuentra en un intervalo entre 0.2 y 0.35, clarifica la gran dificultad de interpretación de cada una de las componentes, pues no existen variables realmente relevantes en ninguna de las componentes (recuérdese que la carga más alta se identificaba con un valor de 0.5768 en la sexta componente).

Se observa en esta matriz cómo el método *ad hoc* de convertir en 0 cargas por debajo de un cierto umbral, no es tan claro cómo en otras matrices de datos, pues además se obviarían cargas con un valor muy cercano a las que se dejarían en el estudio, en el que podrían haber entrado fácilmente de manera subjetiva para otro investigador.

Tabla 10 Cargas PCA al 30%

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Energia_Kcal	**-0.3171**	-0.0508	-0.0392	0.1186	-0.1570	0.2407	-0.1149
Grasa	**-0.3252**	-0.0982	0.0229	-0.1223	-0.0234	-0.0474	0.2260
Proteinas	-0.0996	-0.1683	0.1134	-0.1987	-0.1320	-0.2006	**-0.5682**
Agua	0.2348	-0.1337	-0.0286	-0.1584	0.1992	**-0.3853**	0.2251
Fibra	0.2251	0.1088	-0.0246	**-0.3039**	0.1195	0.2514	-0.1827
Carbohidratos	-0.1759	-0.1547	-0.0819	0.2652	-0.1679	**0.3755**	-0.2271

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
AGrasos_mono	**-0.3307**	-0.0664	0.0874	-0.1576	0.0090	-0.0377	0.1423
AGrasos_poli	-0.2845	-0.0320	0.1207	-0.2870	0.0440	0.0620	0.0406
AGrasos_Saturados	**-0.3223**	-0.0953	0.0359	-0.1482	-0.0105	-0.0330	0.2220
Colesterol	**-0.3039**	-0.0266	-0.0667	0.0876	0.1197	0.1084	-0.0169
VitA	-0.0107	-0.0833	**-0.3575**	**-0.3195**	-0.0081	0.0034	-0.2039
VitD	0.0391	0.0087	**-0.4467**	-0.0882	-0.0750	-0.1096	0.0228
VitE	0.0698	0.0153	**-0.3962**	-0.2944	-0.0293	0.0353	-0.2225
VitB9	-0.1202	-0.0055	0.0818	0.0468	0.5310	0.0071	-0.2436
VitB3	-0.1185	-0.2072	-0.2522	-0.0182	-0.2910	-0.0774	0.0431
VitB2	-0.1085	-0.2489	0.1703	-0.2941	-0.1603	-0.0060	-0.0769
VitB1	-0.1097	-0.2246	-0.1403	0.2904	0.1051	0.0026	0.0610
VitB12	0.1731	-0.2712	-0.1025	0.1808	-0.0432	-0.0767	-0.0541
VitB6	0.2457	0.0182	0.2440	-0.0721	-0.0067	0.1249	-0.0833
VitC	0.1443	0.0280	0.08756	-0.0285	**-0.4109**	0.1456	-0.0626
Calcio	0.0807	**-0.3717**	0.0755	-0.0030	-0.0736	0.1074	0.0791
Hierro	-0.0149	-0.0034	0.0036	**-0.3368**	0.2200	**0.4617**	0.1772
Potasio	0.1351	**-0.3286**	0.0816	0.1191	-0.0139	0.1472	0.0578
Magnesio	0.1628	-0.2010	0.0778	0.0289	-0.2120	0.2851	0.1026
Sodio	0.0971	**-0.3527**	0.0859	-0.1114	-0.0439	0.1790	0.1035
Fósforo	-0.13578	-0.1366	**0.3138**	0.0862	-0.0605	-0.2240	**-0.3478**
Ioduro	0.0360	-0.2122	-0.1009	0.1283	**0.3808**	0.2381	-0.1673
Selenio	-0.0995	-0.2152	**-0.3651**	0.1795	0.0059	0.0139	0.0011
Zinc	0.0003	**-0.3638**	0.0153	0.0351	0.1692	-0.0040	-0.1357

Gráficamente, pueden observarse en las Figura 15, Figura 16, Figura 17 y Figura 18 los valores de las cargas en cada una de las componentes. Se vislumbra de manera más sencilla la dificultad en la interpretación de las componentes, pues la mayor parte de estas tienen una magnitud de entre 0 y 0.300. Además, puede observarse cómo variables como VitD queda bien representada en la componente principal 3, como la

Tabla 10 mostraba que era donde principalmente cargaba, pero no queda bien representada en el eje anterior (PC2), con una carga prácticamente nula.

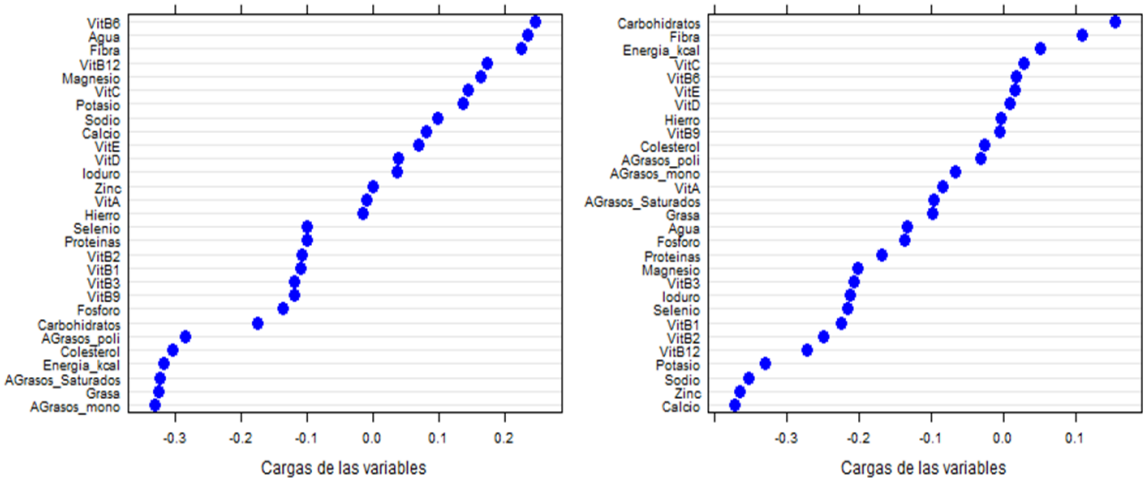

Figura 15 Cargas de las variables para la PC1 (izquierda) y PC2 (derecha)

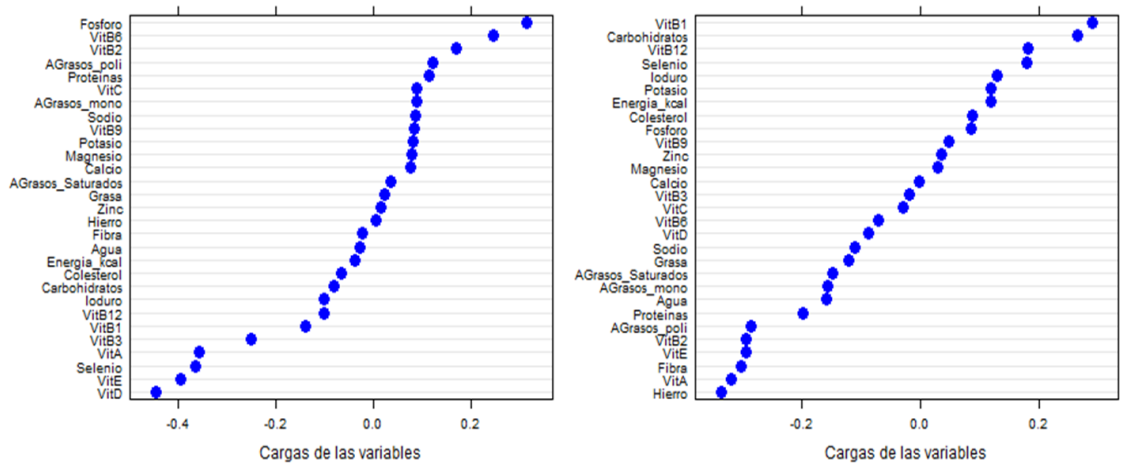

Figura 16 Cargas de las variables para la PC3 (izquierda) y PC4 (derecha)

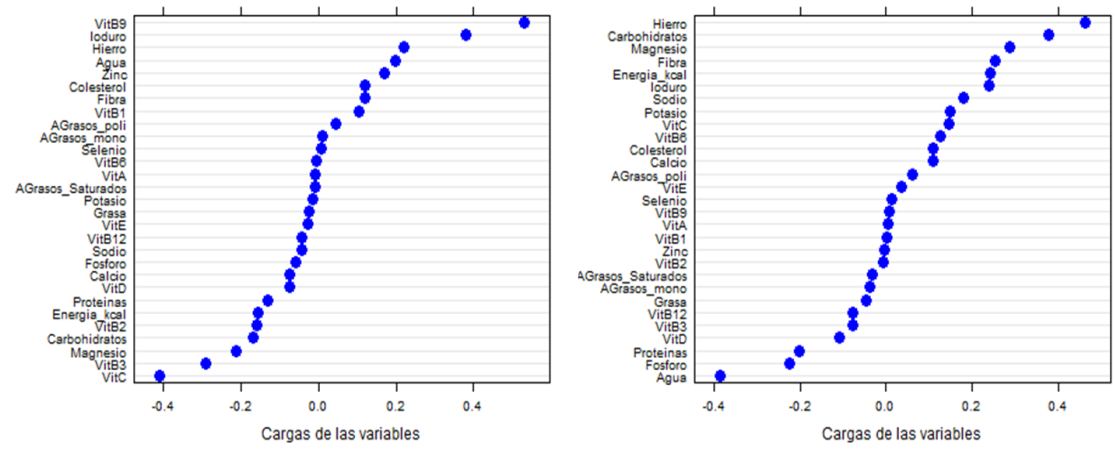

Figura 17 Cargas de las variables para la PC5 (izquierda) y PC6 (derecha)

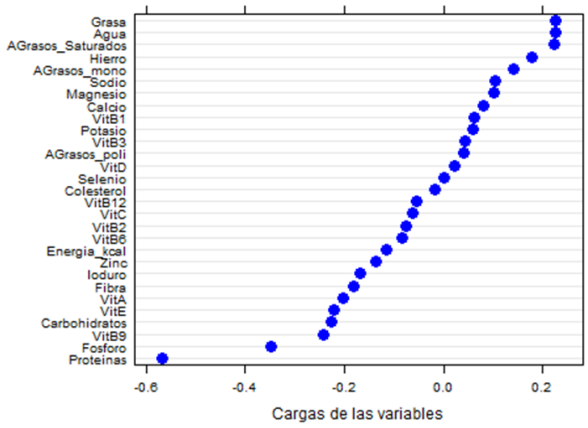

Figura 18 Cargas de las variables para la PC7

7.3.3. Rotación VARIMAX

Se examinan los resultados obtenidos tras aplicar la rotación VARIMAX del programa R sobre la matriz de datos de yogur.

Como ya se ha mencionado, la rotación VARIMAX se encarga de rotar los factores obtenidos forzando a que unas cargas se aproximen más a uno y las otras a cero. Todo ello para facilitar su interpretación.

Al realizar esta rotación a las cargas de la Tabla 10, con la función *Varimax(),* se obtienen las cargas rotadas de las Figura 19y Figura 20. Se puede observar que, en comparación con las cargas de las PCs del análisis original que se pueden observar en las mismas figuras, las cargas de las variables que estaban poco representadas en una componente específica, se hicieron mucho más pequeñas, siendo algunas casi iguales a cero.

En estas figuras se observa fácilmente la forma de actuar de la rotación VARIMAX sobre una matriz de cargas dadas. En todas las componentes principales, pero sobre todo en la PC1, PC2, PC4 y PC7 de manera mucho más clara, como la rotación pretende separar las cargas, haciendo que algunas tengan valores más grandes y otras mucho más cercanas a 0. Se observa este último efecto en las gráficas, pues el número de cargas cercanas a 0 se ha visto aumentado.

Aun así, sigue siendo clara la dificultad de interpretación de las componentes resultantes. Esto es así porque la rotación no consigue mostrar un menor número de cargas relevantes, con magnitudes más grandes que las obtenidas en el PCA ordinario, en cada componente principal. Esto provoca que, aun acercando algunas variables a 0,el número de variables que conforman cada componente principal con mayores

valores sigue siendo alto y, no sólo eso, si no que con magnitudes no superiores a 0.4 en ninguna de las nuevas variables. Esto provoca que no haya variables con un efecto primordial que de un gran sentido a cada componente.

Por ello, sigue presente la necesidad de encontrar un método que proporcione un menor número de cargas relevantes en cada componente, con mayor magnitud, y de manera que el resto sean insignificantes en su creación.

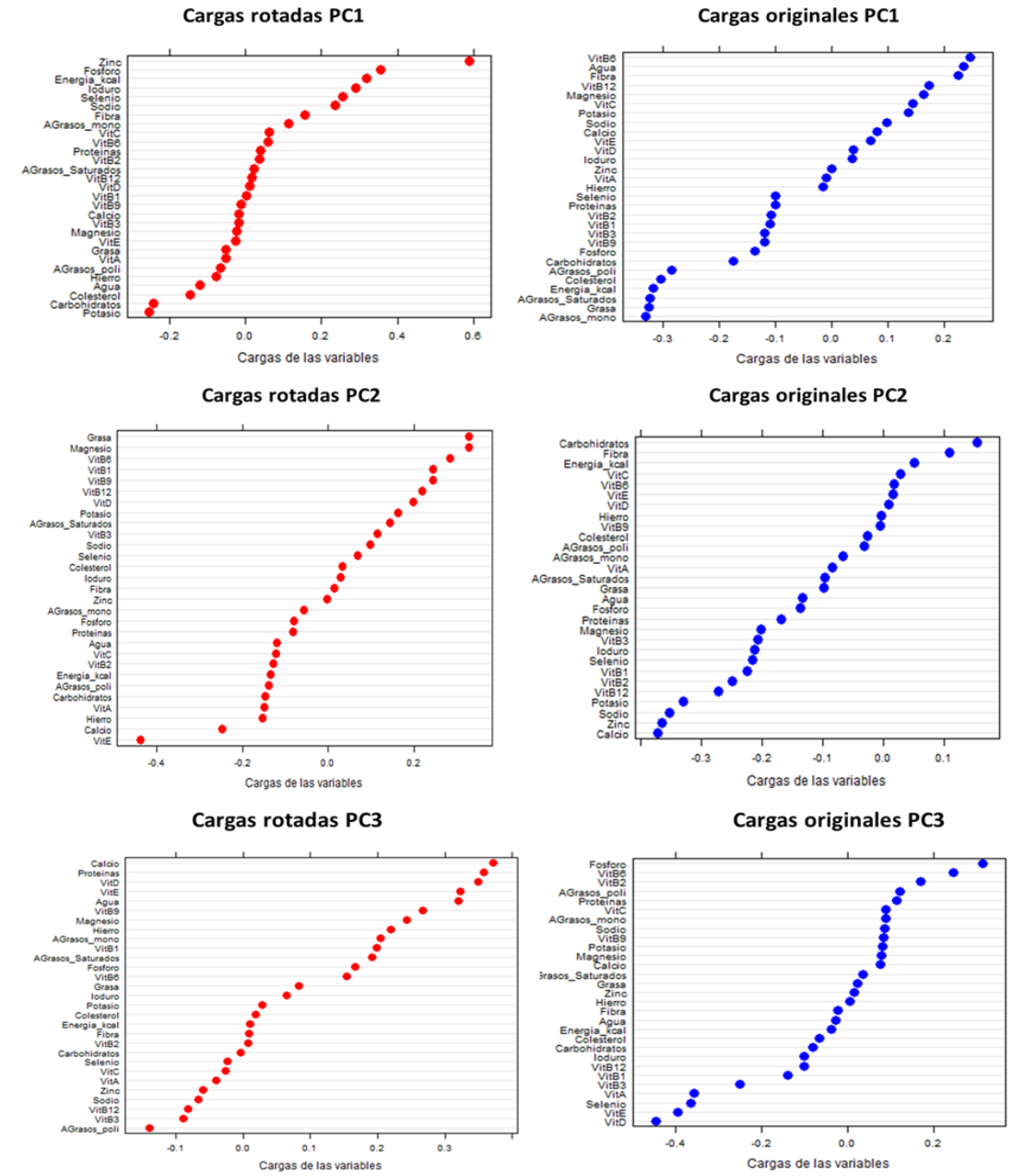

Figura 19 Comparación de las cargas obtenidas tras la rotación VARIMAX (rojo) y las obtenidas en el PCA ordinario (azul) en las componentes principales 1, 2 y 3

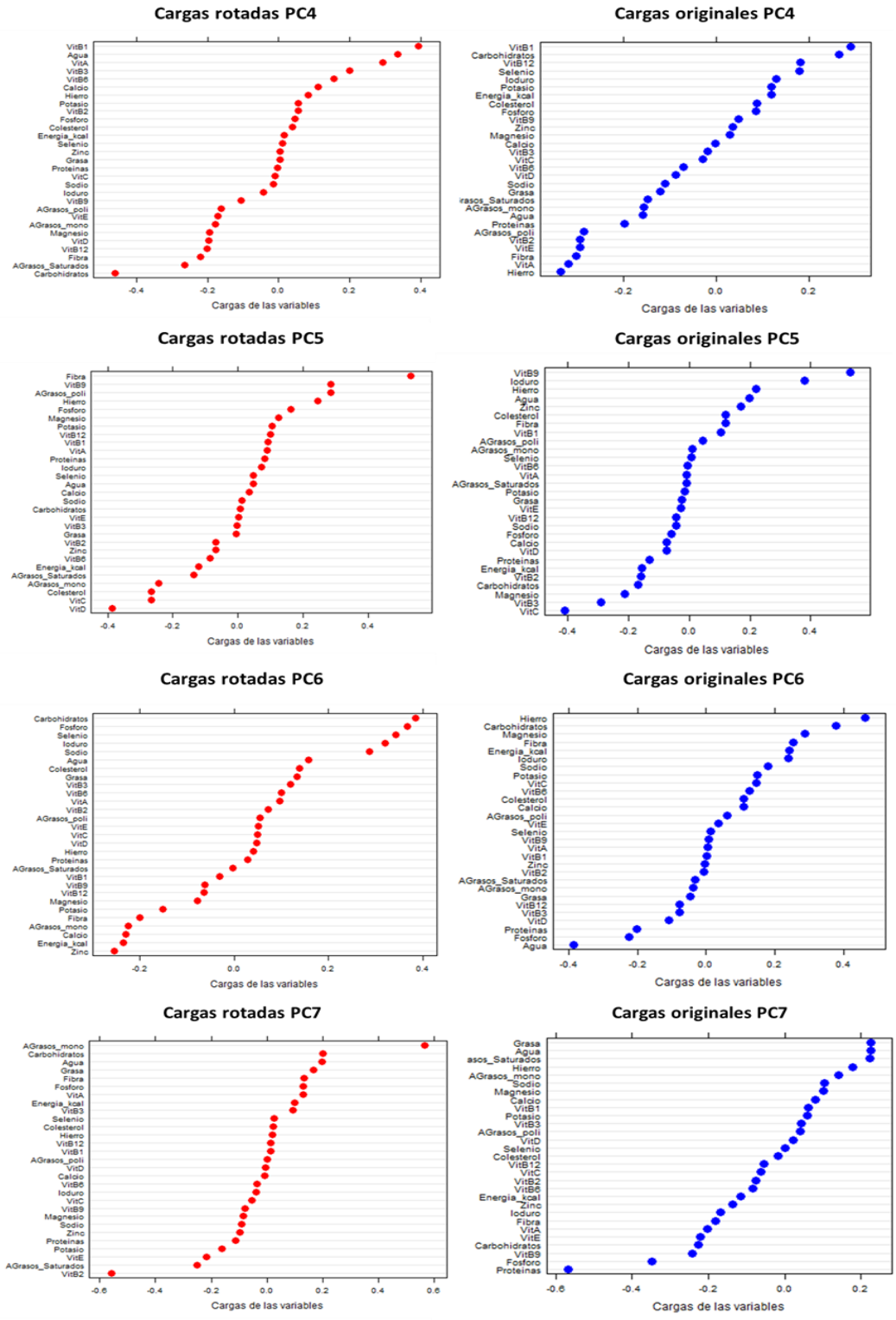

Figura 20 Comparación de las cargas obtenidas tras la rotación VARIMAX (rojo) y las obtenidas en el PCA ordinario (azul) en las componentes principales 4, 5, 6 y 7

7.3.4. Elastic net

De forma similar al caso de los apoyos de madera para este ejemplo también se ejecuta el Análisis de Componentes *Sparse* mediante la función *spca* del paquete de R *elasticnet* (Zou & Hastie, 2012).

La primera solución del Criterio SPCA se realiza mediante la imposición de un vector de cardinalidades determinado por el número de cargas óptimo para el caso de los apoyos de manera se recurre a los resultados del PCA ordinario, donde se determina la cardinalidad de cada componente *sparse* como la cantidad de cargas superiores a $|0.3|$ del PCA ordinario, obteniendo los resultados de la Tabla 11.

En la misma Tabla se puede ver que la cantidad total de varianza explicada por las componentes principales *sparse* es del 41.1%. La primera componente reúne a las variables que miden la cantidad de grasa y ácidos grasos de los yogures, la segunda componente agrupa las variables miden los minerales como calcio, potasio, magnesio y sodio. La tercera componente es la combinación de las vitaminas A, D E, B13 y el selenio. En la cuarta componente se encuentra la cantidad de carbohidratos, aunque en esta componente también carga la Vitamina B9 su valor es relativamente bajo.

Tabla 11 Solución Sparse para yogures con vector de cardinalidad

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Energia_kcal	0	0	0	0	0	0.513	0
Grasa	-0.155	0	0	0	0	0	0
Proteinas	0	0	0	0	0	0	-0.895
Agua	0	0	0	0	0	-0.313	0
Fibra	0	0	0	-0.999	0	0	0
Carbohidratos	0	0	0	0	0	0.733	0
AGrasos_mono	-0.431	0	0	0	0	0	0
AGrasos_poli	-0.686	0	0	0	0	0	0
AGrasos_Saturados	-0.565	0	0	0	0	0	0
Colesterol	0	0	0	0	0	0.318	0
VitA	0	0	-0.291	0	0	0	0
VitD	0	0	-0.406	0	0	0	0
VitE	0	0	-0.474	0	0	0	0
VitB9	0	0	0	0	0.619	0	0
VitB3	0	0	-0.213	0	0	0	0
VitB2	0	0	0	0	0	0	0

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
VitB1	0	0	0	0.045	0	0	0
VitB12	0	0	0	0	0	0	0
VitB6	0	0	0	0	0	0	0
VitC	0	0	0	0	-0.25	0	0
Calcio	0	-0.8	0	0	0	0	0
Hierro	0	0	0	0	0	0	0
Potasio	0	-0.323	0	0	0	0	0
Magnesio	0	-0.105	0	0	0	0	0
Sodio	0	-0.495	0	0	0	0	0
Fosforo	0	0	0	0	0	0	-0.447
Ioduro	0	0	0	0	0.744	0	0
Selenio	0	0	-0.693	0	0	0	0
Zinc	0	0	0	0	0	0	0
% de varianza	10.8%	8.8%	8.8%	2.9%	4.8%	5.7%	4.1%
Cardinalidad	4	4	5	2	3	4	2

Aunque la solución anterior proporciona una interpretación un poco más clara para un investigador sobre la composición de los yogures, a continuación se intentará solucionar el Criterio SPCA mediante la selección de los valores de la penalización *Lasso*, dado que la función *spca()* usa por valor de defecto $1e^{-6}$ para la penalización *Ridge*.

La selección de dichos valores debe lograr un equilibrio entre el porcentaje de varianza explicada y el número de cargas diferente de cero para cada componente en el caso específico de los yogures. Por lo tanto, se propone una función que inicialmente se genera una matriz de *landas* con 12 filas y 7 columnas con valores aleatorios entre 0 y 2. Después se ejecutan 1008 ensayos de SPCA generados por un vector de penalización que se construye como la combinación de 12 filas por 7 columnas variadas 12 veces. Con el fin de encontrar el valor adecuado de la penalización *Lasso* que logre equilibrar los objetivos de porcentaje de varianza y la cardinalidad de cada ensayo. Con la que se obtienen los siguientes gráficos y se escoge el vector (0.4, 0.8, 0.48, 0.55, 0.5, 0.7, 0.65) para resolver el criterio SPCA.

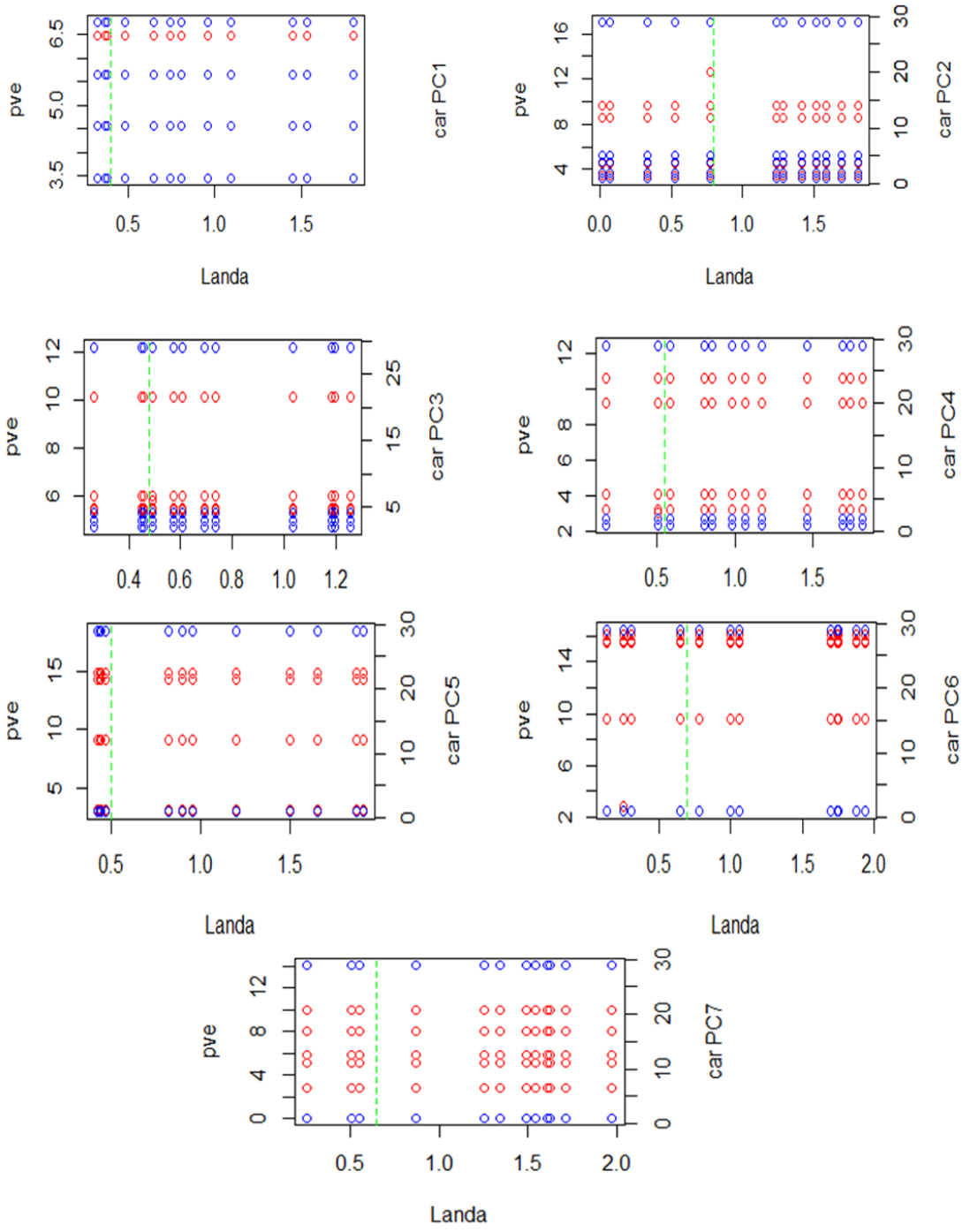

Figura 21 Gráficos conjuntos de varianza explicada y cardinalidad para cada componente en función de landa para su elección

En la Tabla 12 se puede ver que la cantidad total de varianza explicada por las componentes principales *sparse* es del 36.8%, la interpretación del caso de la composición de los yogures se hace más simple partiendo identificar grupo de variables importantes como grasas, minerales y vitaminas.

Tabla 12 Solución Sparse para yogures con vector de penalización

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Energia_kcal	-0.183	0	0	0	0	0	0
Grasa	-0.166	0	0	0	0	0	0
Proteinas	0	0	0	0	0	0	-1
Agua	0	0	0	0	0	-0.199	0
Fibra	0	0	0	-0.473	0	0	0
Carbohidratos	0	0	0	0	0	0.98	0
AGrasos_mono	-0.928	0	0	0	0	0	0
AGrasos_poli	-0.27	0	0	0	0	0	0
AGrasos_Saturados	0	0	0	0	0	0	0
Colesterol	0	0	0	0	0	0	0
VitA	0	0	-0.346	0	0	0	0
VitD	0	0	-0.273	0	0	0	0
VitE	0	0	-0.864	0	0	0	0
VitB9	0	0	0	0	0.763	0	0
VitB3	0	0	0	0	0	0	0
VitB2	0	0	0	0	0	0	0
VitB1	0	0	0	0.197	0	0	0
VitB12	0.072	0	0	0	0	0	0
VitB6	0	0	0	0	0	0	0
VitC	0	0	0	0	-0.135	0	0
Calcio	0	-0.594	0	0	0	0	0
Hierro	0	0	0	0	0	0	0
Potasio	0	0	0	0	0	0	0
Magnesio	0	-0.014	0	0	0	0	0
Sodio	0	-0.804	0	0	0	0	0
Fosforo	0	0	0.245	0	0	0	0
Ioduro	0	0	0	0	0.632	0	0
Selenio	0	0	0	0.859	0	0	0
Zinc	0	0	0	0	0	0	0
% Varianza	7.70%	6.30%	7.50%	5%	4.30%	3.20%	2.80%
Cardinalidad	5	3	4	3	3	2	1

Entonces puede observarse como el SPCA mejora la interpretación de las componentes principales, reduciendo el número de cargas diferentes de cero, permitiendo la selección de variables importantes. Lo cual aplica para el caso de los Yogures donde en la Figura

18 y la Figura 22 se puede ver cómo las componentes *sparse* generan una interpretación correcta del problema de composición de los mismos.

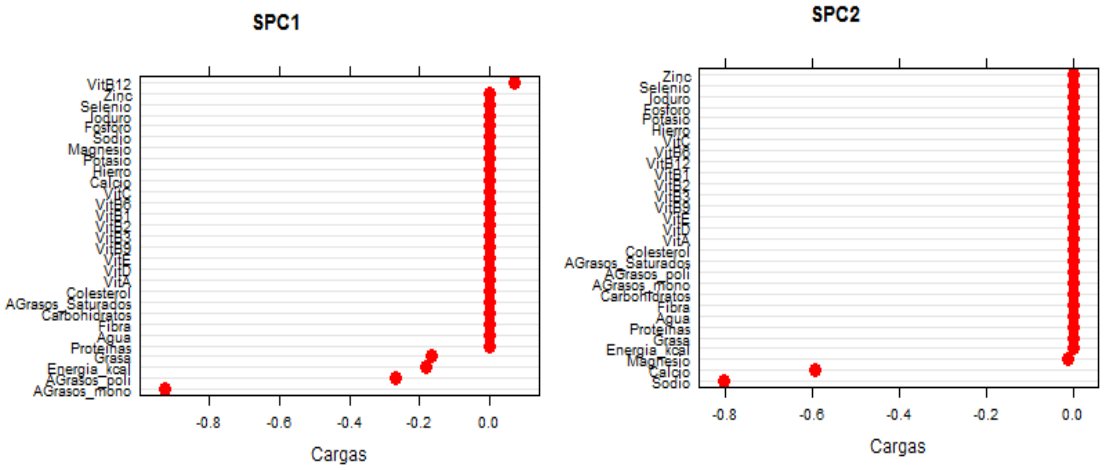

Figura 22 Componentes Sparse 1 y 2 de la composición de los yogures

8. DISCUSIÓN FINAL

El Análisis de Componentes Principales es una técnica versátil capaz de proporcionar una visión general de datos multivariantes complejos, aplicada en múltiples áreas.

Ahora bien, esta técnica es realmente útil cuando los resultados son fácilmente interpretables. Sin embargo, esto no siempre es así, y, es entonces, cuando la técnica presenta su mayor deficiencia. El PCA sufre de la dificultad de interpretación de las componentes principales, surgidas como combinación lineal de todas y cada una de las variables de la matriz de datos original. Esto se agrava en el caso de matrices de datos de grandes dimensiones y, aún más, cuando $p \gg n$.

Con el objetivo de resolver este problema, han surgido distintas técnicas a lo largo de los años: técnicas de rotación, umbralización, o técnicas de reducción del valor de las variables, pero todas ellas cuestionables, puesto que sus resultados, facilitan la interpretación, pero llevan consigo la subjetividad humana.

Es entonces cuando surge el Análisis de Componentes Principales *Sparse* como forma de superar la debilidad de la mayor parte de estos métodos, con el fin de obtener una mejor interpretación. Este método tiene la finalidad de proporcionar una gran cantidad de magnitudes de la matriz de cargas *sparse;* es decir, nulas, haciendo que cada componente principal sea combinación lineal de un número muy pequeño de variables originales. Se consigue con ello, la finalidad que el usuario busca, por ejemplo, con las técnicas de rotación, en las que en la mayor parte de los casos las cargas no son exactamente nulas, a pesar de que el usuario así lo pretenda.

Desde un punto de vista práctico, un buen método *Sparse* debe, como mínimo, cumplir las siguientes propiedades:

- Sin ningún tipo de restricción, el método debe reducirse al PCA.
- Debe ser computacionalmente eficiente, tanto para pequeñas como grandes matrices de datos.
- Se debe evitar la mala identificación de variables importantes.

La formulación del PCA como un problema de mínimos cuadrados, ha facilitado la modificación del comportamiento de la técnica mediante la adición de nuevos términos, desarrollados para problemas de regresión y aplicables en el caso del PCA una vez que este se ha enunciado como tal.

Al igual que en el caso del PCA, dos grandes vertientes permiten el cálculo de componentes principales *sparse.* Por un lado, algoritmos diseñados para la búsqueda de componentes principales *sparse* de forma que se maximice la varianza absorbida por ellas. Por otro, la búsqueda de estas partiendo de un problema de minimización del error.

Centrándonos en este último punto de vista, se ha estudiado la obtención de componentes principales *sparse*, entendiendo el problema del PCA, como un problema de mínimos cuadrados sujeto a la penalización de la regresión *Elastic net.*

Sobre un problema de regresión, la penalización *Ridge* permitía obtener cargas cercanas a 0, aunque no exactamente nulas, por lo que esta penalización no es una solución óptima para el objetivo buscado.

Por otro lado, la penalización *Lasso* permitía obtener cargas nulas, consiguiendo así una técnica de selección de variables, objetivo de las *Sparse*.

Con ello, el criterio ya proporcionaba soluciones óptimas para matrices de datos dadas. Sin embargo, este modelo falla cuando $p \gg n$, así como en el caso de variables correlacionadas, sin proporcionar cargas similares a un conjunto de variables relacionadas. Queriendo ir más allá para superar esta limitación, dada la necesidad de estudiar grandes masas de datos, (como puede ser el caso de microarrays en los que el número de variables (unas 10000) supera con creces el número de observaciones (100)) se ha aplicado sobre el problema inicial otra penalización: *Elastic net*, que, presentándose como una combinación convexa de las penalizaciones Ridge y Lasso, supera los inconvenientes de cada una de ellas y conserva sus propiedades favorables.

El SPCA como técnica de regresión penalizada con Elastic net, se convierte en un método de selección de variables óptimo para el desarrollo de un algoritmo eficiente para datos en los que $p \gg n$. Esta técnica disfruta de ventajas en diversos aspectos, incluyendo la eficiencia computacional, la alta cantidad de varianza explicada y la habilidad de identificar variables importantes.

Al igual que el PCA, esta nueva técnica se convierte en un pilar central de investigación, que puede proporcionar resultados realmente interesantes para la comunidad científica, en diversas áreas, al igual que su progenitora. La investigación en el tema está totalmente abierta y con mucho futuro por delante.

Durante los últimos diez años, desde que apareció el primer documento sobre el *Sparse* PCA, se han publicado una cantidad enorme de trabajos sobre el tema. Sin embargo, los métodos existentes para obtener cargas *sparse* en componentes, al igual que ha

ocurrido con otras técnicas, son muchos y muy diversos, y su número aumenta constantemente. Sobre la base de un número de comparaciones principalmente empíricos (como hacen referencia Journée y Nesterov (2010), Lu y Zhang (2012), Richtárik, Takáč, Ahipaşaoğlu, Richt, y Tak (2012) y Sriperumbudur et al (2011)) se puede concluir que el método GPower de Journée y Nesterov (2010) es probablemente el método más rápido y más versátil disponible para *Sparse* PCA. Su algoritmo adopta la penalización ℓ_0 que resulta más eficiente que ℓ_1, (Journée & Nesterov, 2010). El rendimiento GPower es seguido de cerca por el método propuesto por Zou et al. (2006) y presentado en este trabajo.

La existencia de tantos y diversos algoritmos de aplicación que pretenden suplir los fallos de sus predecesores, abre futuras líneas de investigación cuya finalidad sea encontrar un algoritmo realmente eficiente, con buenos resultados, fácilmente utilizable por el usuario y con la rapidez computacional necesaria para aplicarlo a grandes masas de datos.

Autores como Trendafilov (2014) proponen otras vertientes cómo profundizar en el entendimiento de los algoritmos que preservan explícitamente la condición de ortonormalidad. Restricción requerida, entre otras, para dejar claros los beneficios de este tipo de soluciones con respecto a la coherencia en altas dimensiones.

Las autoras del presente trabajo consideran que sería interesante presentar un estudio que busque aclarar las discrepancias y similitudes en las investigaciones más relevantes del SPCA. También se considera pertinente profundizar en la interpretación de la representación Biplot de cargas *sparse* y el desarrollo de algoritmos SPCA en matrices de datos de tres vías.

9. CONCLUSIONES

- **1.-** La formulación del PCA como un problema de mínimos cuadrados permite mejorar la interpretación de la técnica mediante la adición de nuevos términos de penalización, desarrollados inicialmente para los problemas de regresión.

- **2.-** En un problema de regresión, la penalización *Ridge* permite obtener cargas cercanas a 0, aunque no exactamente nulas, además de proporcionar magnitudes similares a cargas de variables correlacionadas.

- **3.-** En un problema de regresión, la penalización *Lasso* permite obtener cargas nulas, consiguiendo así una selección de variables. Sin embargo, este modelo falla cuando $p \gg n$, así como en el caso de variables correlacionadas, sin proporcionar cargas similares a un conjunto de variables relacionadas.

- **4.-** En un problema de regresión, la penalización *Elastic net* surge como una combinación convexa de las penalizaciones Ridge y Lasso, superando los inconvenientes de cada una de ellas, mientras conserva sus propiedades favorables.

- **5.-** El Análisis de Componentes Principales *Sparse,* como modelo de regresión penalizada con *Elastic net,* para minimizar el error de reconstrucción, es un método de selección de variables, que permite el desarrollo de un algoritmo eficiente para datos multivariantes y datos en los que $p \gg n$.

- **6.-** El SPCA así presentado proporciona ventajas en la eficiencia computacional, en la alta cantidad de varianza explicada y en la habilidad de identificar variables importantes.

10. REFERENCIAS

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(4), 433–459.

AESAN. (2010). BEDCA Base de Datos Española de Composición de Alimentos.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Anderson, T. (1963). Asymptotic theory for principal component analysis. *Annals of Mathematical Statistics*.

Berge, J., & Kiers, H. (1997). Are all varieties of PCA the same? A reply to Cadima & Jolliffe. *British Journal of Mathematical and …*.

Breiman, L. (1996). Bagging predictors. *Machine Learning*.

Bro, R., & Smilde, A. K. (2014). Principal component analysis. *Analytical Methods*, *6*, 2812.

Bryant, E. H., & Atchley, W. R. (1975). *Multivariate Statistical Methods: Within Group Covariation*. Stroudsberg: Halsted Press.

Cadima, J. J. F. C. L., & Jolliffe, I. T. (2001). Variable selection and the interpretation of principal subspaces. *Journal of Agricultural, Biological, and …*, *6*(1), 62–79.

Castro, S. (2012). ANALISIS DE DATOS EN GRANDES DIMENSIONES. ESTIMACION Y SELECCION DE VARIABLES EN REGRESION.

Cuadras, C. M. (2014). Nuevos métodos de análisis multivariante.

Eckart, C., & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*.

ELSEIVER. (2015). SCOPUS.

Friedlander, M., & Tseng, P. (2007). Exact regularization of convex programs. *SIAM Journal on Optimization*.

Frisch, R. (1929). Correlation and scatter in statistical variables. *Nordic Statist*, *8*, 36–102.

Gabriel, K. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*.

Galindo Villardon, M. P. (1986). Una alternativa de representacion simultanea: HJ-Biplot. *Qüestiió. 1986, Vol. 10, Núm. 1*.

Girshick, M. (1939). On the sampling theory of roots of determinantal equations. *The Annals of Mathematical Statistics*.

Gower, J. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*.

Gower, J. (1995, July 1). A general theory of biplots. Oxford University Press.

Hastie, T. J. ., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction.

Hastie, T., & Tibshirani, R. (n.d.). Friedman (2001). The elements of statistical learning: data mining, inference and prediction. *Springer*.

Hoerl, A., & Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*(6), 417–441.

Jeffers, J. (1967). Two case studies in the application of principal component analysis. *Applied Statistics*.

Jeffers, J. N. R. (1967). in the Application Two Case Studies of Principal Component Analysist. *Journal of the Royal Statistical Society*, *16*(3), 225–236.

Jolliffe, I. T. (2002). Principal Component Analysis, Second Edition. *Encyclopedia of Statistics in Behavioral Science*, *30*(3), 487.

Jolliffe, I. T., Trendafilov, N. T., & Uddin, M. (2003). A modified principal component technique based on the LASSO. *Journal of Computational …*, *12*(3), 531–547.

Jolliffe, I. T., & Uddin, M. (2000). The simplified component technique: an alternative to rotated principal components. *Journal of Computational and …*, *9*(4), 689–710.

Journée, M., & Nesterov, Y. (2010). Generalized power method for sparse principal component analysis. *The Journal of Machine.*

Kaiser, H. H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, *23*(3), 187–200.

Kass, R. E., & Raftery, A. E. (2012). Bayes Factors. *Journal of the American Statistical Association*.

Koch, I. (2013). *Analysis of Multivariate and High-Dimensional Data*. Cambridge University Press.

Land, A., & Doig, A. (1960). An automatic method of solving discrete programming problems. *Econometrica: Journal of the Econometric Society*.

Lu, Z., & Zhang, Y. (2012). An augmented Lagrangian approach for sparse principal component analysis. *Mathematical Programming*.

Mardia, K., Kent, J., & Bibby, J. (1979). Multivariate analysis.

Merola, M. G. (2014). Package " spca ."

Moghaddam, B., Weiss, Y., Avidan, S., Moghadam, B., Weiss, Y., Avidan, S., … Avidan, S. (2006). Generalized Spectral Bounds for Sparse LDA. *Proceedings of the 23rd …*.

Nilsson, N. (1996). Introduction to machine learning. An early draft of a proposed textbook.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, *2*(6), 559 – 572.

Pearson, K. (1926). On the coefficient of racial likeness. *Biometrika*.

Qi, X., Luo, R., & Zhao, H. (2013). Sparse principal component analysis by choice of norm. *Journal of Multivariate Analysis*.

Rao, C. (1964). The use and interpretation of principal component analysis in applied research. *Sankhyā: The Indian Journal of Statistics, Series A*.

Richtárik, P., Takáč, M., Ahipaşaoğlu, S., Richt, P., & Tak, M. (2012). Alternating maximization: unifying framework for 8 sparse PCA formulations and efficient parallel codes. *arXiv Preprint arXiv:1212.4137*, 1–20.

Sanchez Mangos, A. (2012). Análisis de componentes principales: versiones dispersas y robustas al ruido impulsivo.

Shen, H., & Huang, J. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, *99*(6), 1015–1034.

Sriperumbudur, B., Torres, D., & Lanckriet, G. (2011). A majorization-minimization approach to the sparse generalized eigenvalue problem. *Machine Learning*.

Team, R. D. C. (2015). R.

Thurstone, L. (1931). Multiple factor analysis. *Psychological Review*.

Thurstone, L. (1935). The vectors of mind.

Thurstone, L. (1947). Multiple-factor analysis; a development and expansion of The Vectors of Mind.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society.*

Tikhonov, A. (1943). On the stability of inverse problems. *Dokl. Akad. Nauk SSSR*.

Trendafilov, N., & Jolliffe, I. T. (2006). Projected gradient approach to the numerical solution of the SCoTLASS. *Computational Statistics & Data Analysis*.

Trendafilov, N. T. (2014). From simple structure to sparse components: A review. *Computational Statistics*, *29*(3-4), 431–454.

Witten, D., Tibshirani, R., & Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*.

Wright, S. (2011). Gradient Algorithms for Regularized Optimization. *SPARS11*. Edinburgh, Scothland.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic-net. *Journal of the Royal Statistical Society*, *67*, 301–320.

Zou, H., & Hastie, T. (2012). CRAN - Package elasticnet. Retrieved June 23, 2015, from http://cran.r-project.org/web/packages/elasticnet/index.html

Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics*, *15*(2), 265–286.